ARO Report 89—1

TRANSACTIONS OF THE SIXTH ARMY

CONFERENCE ON APPLIED MATHEMATICS

AND COMPUTING

Sponsored by

The Army Mathematics Steering Committee

on behalf on

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

U.S. ARMY RESEARCH OFFICE

Report No. 89-1

February 1989

TRANSACTIONS OF THE SIXTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Host

The University of Colorado
Boulder, Colorado

31 May - 3 June 1988

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

FOREWORD

The first paragraph of a letter dated 13 July 1987 from Professor Jerry Bebernes to Dr. Jagdish Chandra stated, "The University of Colorado would like to host the 1988 Annual Army Conference on Applied Mathematics and Computing. Dave Kassoy and I would assist with the local arrangements and other details if requested." These yearly conferences are sponsored by the Army Mathematics Steering Committee (AMSC). On behalf of this Committee, its Chairman, Dr. Chandra, was pleased to accept this invitation to host the sixth meeting in this series, which was held on 31 May - 3 June 1988 in Boulder, Colorado. The Local Chairpersons, Drs. Bebernes and Kassoy, are to be commended on the very fine job they did, not only for the excellent visitor arrangements, but also for their help in selecting speakers and organizing special sessions.

This year the planned program of the conference consisted of three parts, namely: (a) Seven one hour invited addresses; (b) Thirty-two half hour solicited talks covering the following topics: Computational Solid and Structural Mechanics, Reactive and Compressible Flows, Symbolic Computing and Applications, and Parallel Computing; and (c) Thirty-two contributed papers. Most of the latter were presented by Army scientists and covered topics directly related to problems they face in their laboratories. During the course of the conference, these Army scientists had an opportunity to discuss problems with nationally known scientists. Some of these were the invited speakers who are listed below, together with the titles of their talks, but also, many others that appeared on the program or were members of the audience.

SPEAKER AND AFFILIATION	TITLE OF ADDRESS
Professor Thomas Kailath Stanford University	Some New Applications of Matrix Displacement Structures
Professor Ted Belytschko Northwestern University	Nonmonotonic Stress-Strain Laws: Bizarre Behavior and Its Repercussions on Numerical Solutions
Professor A. R. Kapila Rensselaer Polytechnic Institute	Recent Developments in the Theory of Compressible Reactive Flows
Professor Moss Sweedler Cornell University	Applicable Algebraic Methods

Professor Oliver A. McBryan University of Colorado	Promise vs. Performance for Massively Parallel Computers
Professor Robert B. Schnabel University of Colorado	New Sequential and Parallel Methods for Unconstrained Optimization
Professor Luc Tartar Carnegie-Mellon University	How to Describe Oscillations of Solutions of Nonlinear Partial Differential Equations

The members of the AMSC would like to express their thanks to the speakers and research scientists who participated in this meeting, and to all the attendees for supporting it with many stimulating questions. The AMSC is pleased to be able to publish in these Transactions many of the conference papers and thus to make available to the scientific community some of the research results presented at this meeting.

TABLE OF CONTENTS

SIXTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

University of Colorado, at Boulder, Colorado

31 May-3 June 1988

AGENDA

Tuesday, 31 May 1988

0745 - 1600 Registration - Hallway Outside ECCR 2-06

0815 - 0830 Opening Remarks - CR 2-06

0830 - 0930 General Session I - CR 2-06

Chairperson: Benjamin E. Cummings, U.S. Army Human Engineering Laboratory, Aberdeen Proving Ground, Maryland

Some New Applications of Matrix Displacement Structure

Thomas Kailath, Stanford University
Stanford, California

0930 - 1000 BREAK

1000 - 1200 Special Session 1 - Computational Solid and Structural Mechanics, Part 1. Part 2 will be held on Friday morning. CR 1-40

Chairperson and Organizer: Alexander Tessler, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Recent Development in High Performance of Plate and Shell Elements Based on the Free Formulation

C. A. Felippa, University of Colorado
Boulder, Colorado

Large Elasto-Plastic Deformations of Shells of Revolution

I. Fried, Boston University, Boston, Massachusetts, and A. R. Johnson, and A. Tessler U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Aspects of Edge Constraints in Shear-Deformable Plate and Shell Elements

A. Tessler, U.S. Army Materials Technology Laboratory
Watertown, Massachusetts

Finite Element Solution of Planar Elasto-Plastic and Viscoelastic Fracture Problems

A. E. Beagles, J. R. Walton, M. K. Warby, and J. R. Whiteman, Brunel University, Uxbridge, England

******

Tuesday (Continued)

1000 - 1200 Technical Session 1 - Fluid Mechanics - Part 1. Part 2 will be held in Technical Session 6. CR 1-46

Chairperson: Siegfried Lehnigk, U.S. Army Missile Command, Redstone Arsenal, Alabama

The Supersonic Steady State Riemann Problem

John Grove, New York University
New York, New York

Numerical Implication of Riemann Problem Theory for Fluid Dynamics

Ralph Menikoff, Los Alamos National Lab
Los Alamos, New Mexico

Mathematical Modeling of Sound Propagation in the Atmosphere Using the Parabolic Approximation

J. S. Robertson, U.S. Military Academy
West Point, New York, M. J. Jacobson and W. L. Siegmann
Rensselaer Polytechnic Institute, Troy, New York

Dynamics of a Fluid Contained in a Spinning, Coning Cylinder

Raymond Sedney and Nathan Gerber, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD and Philip Hall, Exeter University, Exeter, England

Shearing Flows for Non-Newtonian Fluids

D. S. Malkus, J. A. Nohel, and B. J. Plohr, University of Wisconsin, Madison, Wisconsin

Turbulent Behavior in Channel Flows

Kurt D. Fickie and John D. Kuzan, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD

******

1200 - 1330 LUNCH

1330 - 1530 Technical Session 2 - Mathematical Physics and Numerical Methods. CR 1-40

Chairperson: Miles Miller, Chemical R&D Center, Aberdeen Proving Ground, MD

Numerical Analysis of a Particular Set of Polynomial Equations

William F. Donovan, Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

Annulus-based Inclusion Testing for Multiple-Connected Sets

Terence M. Cronin, U.S. Army Center for Signals Warfare
Warrenton, Virginia

On the Positive Roots of an Equation Involving a Bessel Function

Siegfried H. Lehnigk, U.S. Missile Command
Redstone Arsenal, Alabama

Better Piecewise Linear Approximation

Royce Soanes, Benet Laboratories
Watervliet, New York

Relativistic Thermodynamics of Real Gases with Broken Internal Symmetry

Maxwell's Equations With Broken Internal Symmetries

Guage Theory of Atomic Processes

The Broken Symmetry of Space and Time in Bulk Matter and the Vacuum

Richard A. Weiss, U.S. Army Waterways Experiment Station
Vicksburg, Mississippi

Fixed Points of Expansive Mappings

Walter Egerland, U.S. Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

******

1330 - 1530 Technical Session 3 - Applied Mechanics. CR 1-46

Chairperson: John Vasilakis, Benet Weapons Laboratories
Watervliet, New York

Mixed Finite Element Analysis of Inelastic Shells

V. L. Bergmann and S. Mukherjee, Cornell University
Ithaca, New York

Nonlinear Problems in the Study of Water Movement in Frozen Soils

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

Design Analysis of Artillery Projectile Joints

Kristen Weight and Tien-Yu Tsui, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Incompressible Elements for Rubber Elasticity Using an Elimination Method

A. R. Johnson and C. J. Quigley, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Stress Distributions Near Microstructural Inhomogeneities

Dennis Tracey and Paul J. Perrone, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Finite Element Analysis of Swage Autofrettage Process

Peter C. T. Chen, Benet Laboratories
Watervliet, NY 12189-4050

******

1530 - 1600 BREAK

1600 - 1700 General Session II - CR 2-06

Chairperson: Dennis M. Tracey, U.S. Army Materials Technology Laboratory, Watertown, MA

Nonmonotonic Stress-Strain Laws: Bizarre Behavior and Its Repercussions on Numerical Solutions

Ted Belytschko, Northwestern University
Evanston, Illinois

Wednesday, 1 June 1988

0745 - 1600 Registration - Hallway Outside CR 2-06

0830 - 1030 Special Session 2 - Reactive and Compressible Flows - Part 1. Part 2 is in Special Session 4. CR 1-46

Chairperson and Organizer: Jerrold Bebernes, University of Colorado, Boulder, Colorado

Detonation Shock Dynamics

J. B. Bdzil and D. S. Stewart, Los Alamos National Laboratory, Los Alamos, New Mexico

The Confined Nondiffusive Thermal Explosion with Spatially Homogeneous Pressure Variation

J. Bebernes, University of Colorado
Boulder, Colorado

Steady State Solutions of the Karamoto-Sivashinsky Equation

William C. Troy, University of Pittsburgh
Pittsburgh, Pennsylvania

Counting the Number of Solutions in Reactive Flow Problems

E. Ash and K. Gustafson, University of Colorado
Boulder, Colorado

******

0830 - 1030 Special Session 3 - Symbolic Computation and Applications - Part 1. See Special Session 5 for Part 2. CR 2-06

Chairpersons and Organizers: M. A. Hussain, General Electric Corporate R & D Center, Schnectady, NY and Julian Wu Research Office, Research Triangle Park, NC

Decision Procedures for Solving Differential Equations in Closed Form

B. F. Caviness, National Science Foundation
Washington, DC - On Leave from the University of Delaware

An Integrated Approach of Scientific Computing

Paul S. Wang, Kent State University
Kent, Ohio

A Study of Symbolic Processing and Computational Aspects in Helicopter Dynamics

S. Ravichandran and G. Gaonkar, Florida Atlantic University, Boca Raton, FL, J. Nagbhushana, Indian Institute of Science, and T. S. Reddy, University of Toledo, Toledo, Ohio

Symbolic Computation - Hope, Reality, Serendipity

Clarence J. Maday, North Carolina State University
Raleigh, North Carolina

******

Wednesday (Continued)

1030 - 1100 BREAK

1100 - 1200 General Session III - CR 2-06

Chairperson:

Recent Developments in the Theory of Compressible Reactive Flows

A. R. Kapila, Rensselaer Polytechnic Institute
Troy, New York

1200 - 1330 LUNCH

1330 - 1550 Special Session 4 - Reactive and Compressible Flows - Part 2. CR 1-46

Chairperson and Organizer: Jerrold Bebernes, University of Colorado, Boulder, Colorado

Nonlinear Geometrical Acoustics

John K. Hunter, Colorado State University
Fort Collins, Colorado

Mass-Conserving Treatment of Accumulation Terms in Flows Through Unsaturated Soils

Myron B. Allen, University of Wyoming
Laramie, Wyoming

Phase-Change Problem for Hyperbolic Heat Transfer Model

Dening Li, University of Colorado
Boulder, Colorado

Improvements in the Calculation of Fire Spread in Large Compartments

K. C. Heaton, Defence Research Establishment Valcartier
Courcelette, P.Q., Canada

On the Numerical Solution of a System of Partial Differential Equations to Obtain the Wind from the Geopotential for Numerical Weather Prediction and on Related Mathematical Aspects

H. Baussus von Luetzow, U.S. Army Engineer Topographic Laboratories, Fort. Belvoir, Virginia

******

1330 - 1550 Special Session 5 - Symbolic Computing and Application Part 2. CR 2-06

Chairpersons and Organizers: M. A. Hussian, General Electric, Corporate R&D Center, Schenectady, NY and Julian Wu, U.S. Army Research Office, Research Triangle Park, NC

Computer Algebra Implementation of Lie Transforms for Hamiltonian Systems: Application to the Nonlinear Stability of L_4.

Vincent T. Coppola and Richard H. Rand, Cornell University Ithaca, New York

Symbolic Computation and Perturbation Methods Using Elliptic Functions

Vincent T. Coppola and Richard H. Rand, Cornell Univeristy Ithaca, New York

The Effective Use of Computer Algebra Systems

Joel S. Cohen, University of Denver Denver, Colorado

Using Macsyma in a Generalized Harmonic Balance Method for a Problem of Forced Nonlinear Oscillations

M. A. Hussain, General Electric R & D Center, Schenectady, NY, B. Noble, Brunell University, Umbridge, UK, and J. J. Wu, Army Research Office, Research Triangle Park, NC

Expression Swell Analysis of the Computation of Matrix Characteristic Polynomicals

Michael Wester, University of New Mexico Albuquerque, New Mexico

******

1550 - 1610 BREAK

1610 - 1710 General Session IV - CR 2-06

Chairperson: William Jackson, U.S. Army Tank-Automotive Command, Warren, Michigan

Applicable Algebraic Methods

Moss Sweedler, Cornell University Ithaca, New York

Thursday, 2 June 1988

0800 - 1600 Registration - Hallway Outside CR 2-06

0830 - 1030 Special Session 6 - Parallel Computing - Part 1. Part 2 is in Special Session 7. CR 2-06

Chairpersons and Organizers: Robert Schnabel, University of Colorado, Boulder, CO and Arthur Wouk, U.S. Army Research Office, Research Triangle Park, NC

Multiprocessor FFTs

William L. Biggs, University of Colorado
Denver, Colorado

State of the Art in Current Finite Elements Computations

Charbel Farhat, University of Colorado
Boulder, Colorado

Parallel Methods for Block Bordered Nonlinear Problems

Xiadog Zhang, Richard H. Byrd, and Robert B. Schnabel
University of Colorado, Boulder, Colorado

The HK Singular Value Decomposition

L. Magnus Ewerbring and Franklin T. Luk, Cornell University
Ithaca, New York

******

0830 - 1030 Technical Session 4 - Control. CR 1-46

Chairperson: Herbert Cohen, U.S. Army Material Systems Analysis Activity, Aberdeen Proving Ground, Maryland

Determining Confidence Factors for Expert Systems

Albert Nigrin, Duke University
Durham, North Carolina

The Use of Tomek Links in the Design of Piecewise Linear Classifiers

Jack Sklansky and Youngtae Park, University of California
Irvine, California

Fast and Stable Algorithms for Computing the Principal nth Root of a Complex Matrix and Their Applications to Mathematical Science and Control Systems

L. S. Shieh, University of Houston, Houston, Texas, R. E. Yates, U.S. Army Missile R&D Command, Redstone Arsenal, Alabama, and N. P. Coleman, U.S. Army Armament Center, Dover, New Jersey

Example of A Pursuit-Evasion Game with M Pursuers

Leszek S. Zaremba, Mathematical Reviews
Ann Arbor, Michigan

On the Dynamics of Feedback Systems with Quantized Outputs

David Delchamps, Cornell University
Ithaca, New York

A^G: A Heuristic Search Algorithm for OR Graphs

James W. Lark, III, and Chelsea C. White, III, University of Virginia, Charlottesville, Virginia

******

1030 - 1100 BREAK

1100 - 1200 General Session V - CR 2-06

Chairperson: Michael John Muuss, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Promise vs Performance for Massively Parallel Computers

Oliver A. McBryan, University of Colorado
Boulder, Colorado

1200 - 1330 LUNCH

1330 - 1530 Special Session 7 - Parallel Computing - Part 2. CR 2-06

Chairpersons and Organizers: Robert Schnabel, University of Colorada, Boulder, CO and Arthur Wouk, U.S. Army Research Office, Research Triangle Park, North Carolina

A QR Factorization Algorithm with Control Local Pivoting

Christian H. Bischof, Cornell University
Ithaca, New York

Thursday (Continued)

A Parallel Nonlinear-Jacobi Algorithm for Solving Nonlinear Equations

Rodrigo Fontecilla, University of Maryland
College Park, Maryland

A Performance Comparison of the Roy-Tracing Algorithm on Both Shared Memory and Network Distributed Parallel Computers

Michael John Muuss, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Experiments in Scientific Visualization

Michael J. Muuss and Phillip C. Dykstra, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD

******

1330 - 1530 Technical Session 5 - Stochastic Techniques. CR 1-46

Chairperson:

Deconvolution of Multidetector Systems

Carlos A. Berenstein and B. A. Taylor, University of Maryland, Greenbelt, Maryland

The Modeling of Weapon Dynamics Using Stochastic Mathematical Techniques

Anthony Baran and John Groff, U.S. Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, James Cantor, Steve Carchedi, and Bruce Gibbs, Business Technological Systems, Inc., Seabrook, Maryland, and Herbert Cohen, Army Materiel Systems Analysis Activity, Aberdeen Proving Ground, Maryland

Parallel Algorithms for Stochastic Simulations

Peter W. Glynn, Stanford University
Stanford, California

Covariance Analysis for Split Plot and Split Block Designs and Computer Packages

Walter T. Federer, Cornell University
Ithaca, New York

Comparison of Mean Statistics and Decision Making Using Maximum Likelihood Estimates as an Alternative to Hypothesis Testing

Nathanael Roman, U.S. Army Materiel Test and Evaluation Directorate, White Sands Missile Range, New Mexico

Thursday (Continued)

Numerical Solution of Stochastic Differential Equations

M. Sambandham, Atlanta University
Atlanta, Georgia

1530 - 1600 General Session VI - CR 2-06

Chairperson: Norman Coleman, U.S. Army Armament RD&E Center, Dover, New Jersey

New Sequential and Parallel Methods for Unconstrained Optimization

Robert B. Schnabel, Univerrsity of Colorado
Boulder, Colorado

Friday, 3 June 1988

0745 - 1100 Registration - Hallway Outside CR 2-06

0830 - 1030 Technical Session 6 - Fluid Mechanics - Part 2. CR 1-46

Chairperson: Raymond Sedney, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Helicopter Sound Scattering from Shear Layers Using Generalized Functions

A. Unal and C. Tung, U.S. Army Aviation Research and Technology Activity, Moffett Field, California

Inverse Source Modeling in Helicopter Acoustics

A. Unal and C. Tung, U.S. Army Aviation Research and Technology Activity, Moffett Field, California

Diagonal Implicit Multigrid Solution of the Three-Dimensional Euler Equations

David A. Caughey, Cornell University
Ithaca, New York

Adaptive Mesh Experiements for Time-Dependent Partial Differential Equations

David C. Arney, U.S. Military Academy, West Point, NY,
Joseph E. Flaherty, Benet Laboratories, Watervliet, NY and
Rupak Biswas, Rensselaer Polytechnic Institute, Troy, NY

Friday (Continued)

An Adaptive Mesh Method for Solving Blast Problems Using the Euler Equations

David C. Arney and Erin Misner, U.S. Military Academy, West Point, NY, and Garry Carofano, Benet Laboratories, Watervliet, NY

Computations of Transonic Flow Over a Projectile at Angle of Attack

Jubaraj Sahu, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

0830 - 1030 Special Session 8 - Computational Solid and Structural Mechanics - Part 2. CR 2-06

Chairperson and Organizer: Alexander Tessler, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

Recent Developments in Assumed Strain Shell Elements

K. C. Park, University of Colorado Boulder, Colorado

Oblique Impact of a Plate on a Rigid Target

Aaron D. Gupta, U.S. Army Ballistic Laboratory Aberdeen Proving Ground, Maryland

Nonlinear Analysis by a Mixed Quadrilateral Shell Element

T. Y. Chang, A. F. Saleeb, W. Graf, University of Akron Akron, Ohio

Accuracy of Finite Element Frequency Analysis of Thin Walled Cylinder

Joseph M. Santiago and Henry L. Wisniewski, U.S. Army Ballistic Laboratory, Aberdeen Proving Ground, Maryland

1030 - 1100 BREAK

1100 - 1200 General Session VII - CR 2-06

Chairperson: Jagdish Chandra, U.S. Army Research Office, Research Triangle Park, NC

How to Describe Oscillations of Solutions of Nonlinear Partial Differential Equations

Luc Tartar, Carnegie Mellon University Pittsburgh, Pennsylvania

1200 - 1215 ADJOURNMENT

Divide-and-Conquer Solutions of Least-Squares Problems for Matrices with Displacement Structure

J. Chun and T. Kailath †

Information Systems Laboratory
Stanford University
Stanford, CA 94305

ABSTRACT. A divide-and-conquer implementation of a generalized Schur algorithm enables us to solve various (exact and) least squares block-Toeplitz or Toeplitz-block systems of equations with $O(\alpha^3 n \log^2 n)$ operations, where the displacement rank α is a small constant (typically between 2 to 4 for scalar near-Toeplitz matrices) independent of the size of matrices.

1. Introduction.

In recent years, there has been considerable research on fast algorithms for the solution of linear systems of equations with Toeplitz matrices. The Levinson and Schur algorithms allow (recursive) solutions with $O(n^2)$ floating point operations (flops) for systems with $n \times n$ Toeplitz matrices.

In 1980. Brent *et al* [5] described a (nonrecursive) scheme for obtaining a solution with $O(n \log^2 n)$ flops. This was based on two ideas - the use of the Gohberg-Semencul formula [10], [11], [15] for the inverse of a Toeplitz matrix, and the use of divide-and-conquer (or doubling) techniques for computing (generators of) the Gohberg-Semencul formula.

Let $\mathbf{x}$ and $\mathbf{y}$ denote the first and last columns of $T^{-1} \in \mathbf{R}^{n \times n}$. Then if the first component of $\mathbf{x}$, say x_1, is nonzero, Gohberg and Semencul [11] showed that we could write

$$T^{-1} = \frac{1}{x_1}[L(\mathbf{x})L^T(\tilde{I}_n \mathbf{y}) - L(Z_n \mathbf{y})L^T(Z_n \tilde{I}_n \mathbf{x})] \tag{1}$$

where $\tilde{I}_n$ is the *reverse-identity matrix*, Z_n is the *shift matrix*,

$$\tilde{I}_n \equiv \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \quad Z_n \equiv \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & & \\ & & 1 & 0 \end{bmatrix},$$

and

$L(\mathbf{v})$ = a lower-triangular Toeplitz matrix with first column $\mathbf{v}$.

The significance of (1) in the present application is that the product of a vector and a lower- or upper-

† This work was supported in part by the U.S. Army Research Office under Contract DAAL03-86-K-0045, the SDIO/IST, managed by the Army Research Office under Contract DAAL03-87-K-0033, and the National Science Foundation under Grant MIP-21315-A2.

triangular Toeplitz matrix is equivalent to the convolution of two vectors, which can be done in a fast way using $O(n\log n)$ flops (see, e.g. [4]). This compares with the $O(n^2)$ operations required with the non-Toeplitz triangular matrix factor of T^{-1} (obtainable in $O(n^2)$ flops via the Levinson algorithm). Brent *et al* showed how to use divide-and-conquer techniques combined with a fast Euclidean algorithm (faster than the one in [1]) to obtain the vectors $\{\mathbf{x}, \mathbf{y}\}$ of the Gohberg-Semencul formula with $O(n\log^2 n)$ flops. Later Bitmead and Anderson [3] and Morf [19] used another approach, based on the displacement-rank properties of matrix Schur complements, to obtain similar results; while this approach allows for generalization to non-Toeplitz matrices (further discussed below), the hidden coefficient in their proposed $O(n\log^2 n)$ constructions turned out to be extremely large (see Sexton [23]). Later Musicus [20], Bruckstein and Kailath [6], de Hoog [9], Ammar and Gragg [2] used an approach based on the Schur (rather than Levinson) algorithm to obtain better coefficients; in particular, Ammar and Gragg made a detailed study and claimed an operation count of $8n\log^2 n$ flops. With this count, the new (called superfast in [2]) method for solving Toeplitz systems is better than the one based on the Levinson algorithm whenever $n > 256$. We should mention here that Schur-algorithm-based methods are natural in the context of transmission-line and layered-earth models, so it is not a surprise that similar techniques were also conceived in those fields - see Choate [7], McClary [18] and Bruckstein and Kailath [6]. A good source for background on the Levinson and Schur algorithms, transmission line models, displacement representations as mentioned and used in the present paper may be [12].

Our paper is in the spirit of the methods based on the Schur algorithm, but is more general without the drawback of large coefficient of the methods by Bitmead and Anderson or Morf. We can handle matrices such as $(T^TT)^{-1}$ and $(T^TT)^{-1}T^T$, where T may be a near-Toeplitz matrix including rectangular block-Toeplitz matrices and Toeplitz-block matrices; in particular, therefore, we can also obtain the *least-squares* solutions of over-determined Toeplitz and near-Toeplitz systems with $O(n\log^2 n)$ flops.

An outline of our approach is the following. For a matrix E,

$$E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix}, \qquad E_{1,1}, \text{ nonsingular,}$$

the Schur complement of $E_{1,1}$ in E is

$$S \equiv E_{1,1} - E_{2,1}E_{1,1}^{-1}E_{1,2}.$$

Notice that matrices such as

$$S_1 \equiv T^{-1}, \quad S_2 \equiv (T^TT)^{-1}, \quad S_3 \equiv (T^TT)^{-1}T^T \tag{2}$$

can be identified as the Schur complements of the following *extended matrices*,

$$E_1 = \begin{bmatrix} T & I \\ -I & O \end{bmatrix}, \quad E_2 = \begin{bmatrix} T^TT & I \\ -I & O \end{bmatrix}, \quad E_3 = \begin{bmatrix} T^TT & T^T \\ -I & O \end{bmatrix}. \tag{3}$$

Now the matrices E in (3) have the following (generalized) *displacement representation*, for suitably chosen matrices $\{F^f, F^b\}$,

$$E = \sum_{i=1}^{\alpha} K(\mathbf{x}_i, F^f)K^T(\mathbf{y}_i, F^b),$$

where $K(\mathbf{x}_i, F^f)$ and $K(\mathbf{y}_i, F^b)$ are lower triangular matrices whose j columns are $(F^f)^{(j-1)}\mathbf{x}_i$ and $(F^b)^{(j-1)}\mathbf{y}_i$, respectively. The smallest possible number α is called the *displacement rank* of E with respect to $\{F^f, F^b\}$. For an example, let T be an $m \times n$ scalar Toeplitz matrix, with $m \geq n$. Then the matrix E_2 has the displacement rank 4 with respect to $\{F, F\}$, where $F = \begin{bmatrix} Z_n & O \\ O & Z_n \end{bmatrix}$, and has a displacement representation [13],

$$\begin{bmatrix} T^T T & I \\ -I & O \end{bmatrix} = \sum_{i=1}^{2} K(\mathbf{y}_i, F) K^T(\mathbf{x}_i, F) - \sum_{i=3}^{4} K(\mathbf{y}_i, F) K^T(\mathbf{x}_i, F), \quad \mathbf{y}_i \equiv \begin{bmatrix} I_n & O \\ O & -I_n \end{bmatrix} \mathbf{x}_i. \tag{4a}$$

If we define $\mathbf{x}_i^T \equiv [\mathbf{w}_i^T, \mathbf{v}_i^T]$, note that the matrix $K(\mathbf{x}_i, F)$ in (4a) has the form

$$\begin{bmatrix} L(\mathbf{w}_i) & O \\ L(\mathbf{v}_i) & O \end{bmatrix} \in \mathbf{R}^{2n \times 2n}, \quad O \in \mathbf{R}^{n \times n}, \tag{4b}$$

where $L(\mathbf{w}_i)$ and $L(\mathbf{v}_i)$ are lower triangular Toeplitz matrices with first columns $\mathbf{w}_i$ and $\mathbf{v}_i$.

Given a displacement representation of E, we use a certain *generalized Schur algorithm* [8], [13] to successively compute displacement representations of the Schur complements of all the leading principal submatrices in E. For the above example, n steps of the generalized Schur algorithm will yield

$$\begin{bmatrix} O & O \\ O & (T^T T)^{-1} \end{bmatrix} = \sum_{i=1}^{2} K(\mathbf{u}_i, F) K^T(\mathbf{u}_i, F) - \sum_{i=3}^{4} K(\mathbf{u}_i, F) K^T(\mathbf{u}_i, F),$$

where the top n elements of $\mathbf{u}_i$ are zero. Therefore, if we denote the bottom n elements of $\mathbf{u}_i$ as $\mathbf{u}_{2,i}$, we can re-write

$$(T^T T)^{-1} = \sum_{i=1}^{2} L(\mathbf{u}_{2,i}) L^T(\mathbf{u}_{2,i}) - \sum_{i=3}^{4} L(\mathbf{u}_{2,i}) L^T(\mathbf{u}_{2,i}).$$

Now, the generalized Schur algorithm, which is a two-term polynomial recursion, can be implemented in a divide-and-conquer fashion with $O(\alpha^3 f(n) \log n)$ flops, where $f(n)$ denotes the number of operations for the multiplication of two polynomials. Therefore, if the multiplication of two polynomials is done again by divide-and-conquer, i.e., by using fast convolution algorithms, then the overall computation requires $O(\alpha^3 n \log^2 n)$ flops. We remark that the factor α^3 can be reduced to α if several convolutions can be performed in parallel. Once we have a displacement of the desired Schur complement S, the matrix-vector multiplication, $S\mathbf{b}$ can be done with $O(\alpha n \log n)$ flops using fast convolutions. As an example, we can obtain the least squares solution for

$$T\mathbf{x} = \mathbf{b}, \quad T \in \mathbf{R}^{m \times n}, \quad m \geq n$$

by

(i) Multiply $T^T\mathbf{b}$ using a fast convolution algorithm,

(ii) Obtain a displacement representation of $(T^T T)^{-1}$ using the divide-and-conquer version of generalized Schur algorithm,

(iii) Multiply $(T^T T)^{-1}(T^T\mathbf{b})$ using a fast convolution algorithm.

If we obtain displacement representation of $(T^TT)^{-1}T^T$ directly using E_3, then the step (i) would not be needed.

2. Generalized Schur Algorithm.

After a brief review of basic concepts and definitions, we shall present the generalized Schur algorithm in polynomial form.

Generators of Matrices.

Let F^f and F^b be nilpotent matrices. The matrix

$$\nabla_{(F^f, F^b)}A \equiv A - F^f A F^{bT}$$

is called the *displacement* of A with respect to the *displacement operators* $\{F^f, F^b\}$. Define the (F^f, F^b)-displacement rank of A as rank$[\nabla_{(F^f, F^b)}A]$. Any matrix pair $\{X, Y\}$ such that

$$\nabla_{(F^f, F^b)}A = XY^T, \quad X \equiv [\, x_1, x_2, \ldots, x_\alpha \,], \quad Y \equiv [\, y_1, y_2, \ldots, y_\alpha \,] \tag{5}$$

is called a *(vector form) generator* of A with respect to $\{F^f, F^b\}$. The generator will be said to have *length* α. If the length α is equal to the displacement rank of A, we say that the generator is *minimal*. A generator such as $Y = X\Sigma$, where Σ is a diagonal matrix with 1 or -1 along the diagonal, is called a *symmetric generator*.

The following Lemma [13], [14] establishes the connection between generators and displacement representations.

Lemma. Let E be an $m \times n$ matrix. If F^f and F^b are nilpotent, then $\nabla_{(F^f, F^b)}E = \sum_1^\alpha x_i y_i^T$ has the unique solution $E = \sum_1^\alpha K(x_i, F^f)K^T(y_i, F^b)$, where $K(x_i, F^f) \equiv [x_i, F^f x_i, \cdots, F^{f(n-1)}x_i]$, $K(y_i, F^b) \equiv [y_i, F^b y_i, \cdots, F^{b(n-1)}y_i]$.

Choice of Displacement Operators.

The generalized Schur algorithm operates with generators, and needs $O(\alpha mn)$ flops for sequential implementation and $O(\alpha^3 n \log^2 n)$ for divide-and-conquer implementation. Therefore, for a given matrix A, we should try to choose the displacement operators that give the smallest α. If the matrix A is an $n \times n$ Toeplitz matrix, the appropriate displacement operator F is Z_n, an $n \times n$ shift matrix. If A has some near-Toeplitz structure, then F would have forms such as

$$F = Z_n \oplus Z_m, \quad F = \bigoplus_{i=1}^{n} Z_{n_i}, \quad F = Z_n^\beta,$$

where $\oplus$ denotes the *direct sum*, $Z_n \oplus Z_m \equiv \begin{bmatrix} Z_n & O \\ O & Z_m \end{bmatrix}$, and $\bigoplus_{i=1}^{n}$ denotes the concatenated direct sum.

Example 1. Let $T = (t_{i-j})$ be an $m \times n$ pre- and post-windowed scalar Toeplitz matrix, i.e., $t_{ij} = 0$ if $j > i$ or $i > m - n + j$ with $m > n$. Then it is easy to check that the matrix $C = (c_{i-j}) \equiv T^TT$ is also a (unwindowed) Toeplitz matrix, and with respect to $\{Z_n \oplus Z_n, Z_n \oplus Z_m\}$, E_3 in (3) has a

generator $\{X, Y\}$ of length 2, where

$$\mathbf{x}_1 = [c_0, c_1, \cdots, c_n, -1, 0, \cdots, 0]^T / c_0^{1/2},$$

$$\mathbf{x}_2 = [0, c_1, \cdots, c_n, -1, 0, \cdots, 0]^T / c_0^{1/2},$$

$$\mathbf{y}_1 = [c_0, c_1, \cdots, c_n, t_0, t_1, \cdots t_{m-n}, 0, \cdots 0\ ]^T / c_0^{1/2},$$

$$\mathbf{y}_2 = -[0, c_1, \cdots, c_n, t_0, t_1, \cdots t_{m-n}, 0, \cdots 0\ ]^T / c_0^{1/2}. \qquad \square$$

Example 2. If T is a Toeplitz-block matrix, i.e.,

$$T = \begin{bmatrix} T_{1,1} & T_{1,2} & \cdot & T_{1,N} \\ T_{2,1} & T_{2,2} & \cdot & T_{2,N} \\ \cdot & \cdot & \cdot & \cdot \\ T_{M,1} & T_{M,2} & \cdot & T_{M,N} \end{bmatrix} \in \mathbf{R}^{m \times n}, \quad T_{i,j} = \text{scalar } m_i \times n_j \text{ Toeplitz matrix}, \tag{6}$$

then for the matrices E in (3), we choose [8], [13] the following displacement operators

$$E_1: \quad F^f = [\bigoplus_{i=1}^{M} Z_{m_i}] \oplus F_1, \quad F^b = [\bigoplus_{i=1}^{N} Z_{n_i}] \oplus F_1, \quad m = n, \tag{7a}$$

$$E_2: \quad F^f = [\bigoplus_{i=1}^{N} Z_{n_i}] \oplus F_1, \quad F^b = [\bigoplus_{i=1}^{N} Z_{n_i}] \oplus F_1, \quad m = n, \tag{7b}$$

$$E_3: \quad F^f = [\bigoplus_{i=1}^{N} Z_{n_i}] \oplus F_1, \quad F^b = [\bigoplus_{i=1}^{N} Z_{n_i}] \oplus [\bigoplus_{i=1}^{M} Z_{m_i}], \tag{7c}$$

where F_1 can be either† Z_n or $\bigoplus_{i=1}^{N} Z_{n_i}$.

Example 3. On the other hand, if the matrix T in (3) is a block-Toeplitz matrix with $\beta \times \beta$ blocks,

$$T = \begin{bmatrix} B_0 & B_{-1} & \cdot & B_{-N+1} \\ B_1 & B_0 & \cdot & B_{-N+2} \\ \cdot & \cdot & \cdot & \cdot \\ B_{M-1} & B_{M-2} & \cdot & B_{-N+M} \end{bmatrix} \in \mathbf{R}^{m \times n}, \quad B_k \in \mathbf{R}^{\beta \times \beta}, \quad m \equiv M\beta, \quad n \equiv N\beta. \tag{8}$$

then for the extended matrices E, we should choose [8] the displacement operators

$$F^f = Z_n^{\beta} \oplus Z_n^{\beta}, \qquad F^b = Z_n^{\beta} \oplus Z_m^{\beta}, \tag{9}$$

where, for E_1 we assumed that T is a square $n \times n$ matrix.

Generators of the above and other extended block-Toeplitz or Toeplitz-block matrices can be found in [8] and [13].

Polynomial Form of Generators.

† For the divide-and-conquer implementation, we prefer to choose $\bigoplus_{i=1}^{N} Z_{n_i}$; see the Remark in Sec 4.

In general, the displacement operators F^f and F^b for both extended block-Toeplitz matrices and extended Toeplitz-block matrices have the form,

$$F = \bigoplus_{i=1}^{N} Z_{n_i}^{\beta}, \quad n \equiv \sum_{i=1}^{N} n_i. \tag{10}$$

We shall say that the displacement operator F in (10) has N *sections*. One of the key operations in generalized Schur algorithms is matrix-vector multiplication, $F\mathbf{v}$, i.e, a *sectioned shift* operation. With the polynomial representation of vectors, the shift operation has a nice algebraic expression. For a given vector $\mathbf{v}$, let $v(z)$ denote the polynomial whose coefficient for the term z^i is the i+1st component of the vector, i.e.,

$$\mathbf{v} = [v_0, v_1, v_2, \cdots, v_{n-1}]^T \quad \rightarrow \quad v(z) = v_0 + v_1 z + v_2 z^2 + \cdots v_{n-1} z^{n-1}. \tag{11}$$

Then,

$$Z_n \mathbf{v} \equiv \mathbf{v}' = [0, v_0, v_1, \cdots, v_{n-2}]^T \quad \rightarrow \quad v(z)z \mod z^n.$$

In general, for the matrix whose displacement operator is the F in (10), let us define integers $\{\delta_i\}$ by

$$\delta_i = \sum_{k=1}^{i} n_k, \quad \delta_1 < \delta_2 < \cdots < \delta_N.$$

Let $v(z)$ and $\theta(z)$ be polynomials of degree less than or equal to $n-1$, and define the degree at most (n_i-1) polynomial, $v_i(z)$, by

$$v(z) = v_1(z) + z^{\delta_1} v_2(z) + z^{\delta_2} v_3(z) + \cdots + z^{\delta_{N-1}} v_{n_N}(z). \tag{12a}$$

Now the *(polynomial form) displacement operator* $\otimes_F$ is defined by the following operation,

$$v(z) \otimes_F \theta(z) \equiv r(z) \equiv r_1(z) + z^{\delta_1} r_2(z) + z^{\delta_2} r_3(z) + \cdots + z^{\delta_{N-1}} r_{n_N}(z), \tag{12b}$$

where

$$r_i(z) \equiv v_i(z)\theta(z^{\beta}) \mod z^{n_i}, \tag{12c}$$

i.e., $r_i(z)$ is the polynomial $v_i(z)\theta(z^{\beta})$ after chopping off the higher degree terms, so that $r_i(z)$ has the degree at most $(n_i - 1)$.

Let

$$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\alpha], \quad Y = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\alpha]$$

be a generator of a matrix A with respect to certain $\{F^f, F^b\}$, and let

$$\mathbf{x}_i \rightarrow x_i(z), \quad \mathbf{y}_i \rightarrow y_i(w).$$

Then we call the pair of polynomial vectors, $\{X(z), Y(w)\}$, where

$$X(z) \equiv [\, x_1(z), x_2(z), \cdots, x_\alpha(z) \,], \quad Y(w) \equiv [\, y_1(w), y_2(w), \cdots, y_\alpha(w) \,],$$

a (polynomial form) generator of A, with respect to *(polynomial from) displacement operator* $\{\otimes_{F^f}, \otimes_{F^b}\}$.

Example 1 (Continued). The matrix E_3 in (3) has a generator $\{X(z), Y(w)\}$ with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, where $F^f = Z_n \oplus Z_n$, $F^b = Z_n \oplus Z_m$, and

$$x_1(z) = [c_0 + c_1 z + \cdots + c_n z^n - z^{n+1}]c_0^{-1/2},$$

$$x_2(z) = [c_1 z + cz^2 + \cdots + c_n z^n - z^{n+1}]c_0^{-1/2},$$

$$y_1(w) = [c_0 + c_1 w + \cdots + c_n w^n + t_0 w^{n+1} + t_1 w^{n+2} + \cdots + t_{m-n} w^{m+1}]c_0^{-1/2},$$

$$y_2(w) = -[c_1 w + \cdots + c_n w^n + t_0 w^{n+1} + t_1 w^{n+2} + \cdots + t_{m-n} w^{m+1}]c_0^{-1/2}.$$

Also notice that

$$x_1(z) \otimes_{F^f} z = [c_0 z + c_1 z^2 + \cdots + c_{n-1} z^n - z^{n+2}]c_0^{-1/2},$$

$$y_1(w) \otimes_{F^b} w = [c_0 w + c_1 w^2 + \cdots + c_{n-1} w^n + t_0 w^{n+2} + t_1 w^{n+3} + \cdots + t_{m-n-1} w^{m+1}]c_0^{-1/2}. \quad \square$$

Next we note that for given vectors $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a}^T\mathbf{b} \neq 0$, we can always find [8] matrices Θ and Ψ such that

$$\mathbf{a}^T\Theta = [a_1', 0, 0, \cdots, 0], \quad \mathbf{b}^T\Psi = [b_1', 0, 0, \cdots, 0], \quad \Theta \cdot \Psi^T = I, \tag{13}$$

and therefore, $\mathbf{a}^T\mathbf{b} = a_1' b_1'$. We define polynomial matrices $\Theta(z)$ and $\Psi(w)$ by

$$\Theta(z) \equiv \Theta \begin{bmatrix} z & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \quad \Psi(w) \equiv \Psi \begin{bmatrix} w & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}. \tag{14}$$

We remark also that if $\mathbf{a} = \mathbf{b}$, then $\Psi(w) = \Theta(w)$, and if $\mathbf{b} = \Sigma\mathbf{a}$, where $\Sigma \equiv I_p \oplus -I_q$, then $\Psi(w) = \Theta(w)\Sigma$, so that we only need to find, and post-multiply by, $\Theta(z)$.

Generalized Schur Algorithm

Let a matrix E have a generator $\{X_0(z), Y_0(w)\}$ with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$. and define $E_{i,j}$ by

$$E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix} \in \mathbf{R}^{m \times n},$$

where $E_{1,1}$ is a $k \times k$ strongly nonsingular matrix, i.e., the one with all nonsingular leading submatrices. The k-step of the generalized Schur algorithm [8], [13] presented below in polynomial form gives a generator of the matrix,

$$\begin{bmatrix} O & O \\ O & S \end{bmatrix}, \quad S \equiv E_{2,2} - E_{2,1}E_{1,1}^{-1}E_{1,2} \in \mathbf{R}^{(m-k)\times(n-k)},$$

with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, or equivalently, a generator of S with respect to $\{\otimes_{\bar{F}^f}, \otimes_{\bar{F}^b}\}$, where $\bar{F}^f$ and $\bar{F}^b$ denote the trailing square submatrices of size $(m-k)$ and $(n-k)$ of F^f and F^b, respectively.

Algorithm (k-step Generalized Schur Algorithm)

Input: Generator of E, $\{X_0(z), Y_0(w)\}$; displacement operator $\{\otimes_{Ff}, \otimes_{Fb}\}$; Number of steps k.
Output: Generator of S $\{X_k(z), Y_k(w)\}$
Procedure GeneralizedSchur
begin
for $i := 0$ **to** $k-1$ **do begin**
$\mathbf{a}^T := [z^{-i}X_i(z)]_{z=0}$;
$\mathbf{b}^T := [z^{-i}Y_i(z)]_{z=0}$;
Find $\Theta_i(z)$ and $\Psi_i(w)$ to transform $\mathbf{a}^T$ and $\mathbf{b}^T$ such as (13);
$X_{i+1}(z) = X_i(z)\otimes_{F_f}\Theta_i(z)$; $Y_{i+1}(w) = Y_i(w)\otimes_{F_b}\Psi_i(w)$;
end
return $\{X_k(z), Y_k(w)\}$;
end

Remark. The polynomial vectors, $X_i(z)$ and $Y_i(w)$, have degrees $m-1$ and $n-1$ respectively, for all i. Each step eliminates the non-zero lowest degree term, and therefore the terms of $X_i(z)$ and $Y_i(w)$ whose degrees are less than z^i and w^i are zeros.

By applying the generalized Schur algorithm, one can obtain generators, or equivalently displacement representations for various interesting Schur complements.

3. Divide-and-Conquer Implementation.

The (sequential) k-step generalized Schur algorithm in Sec 2 can also be implemented efficiently using divide-and-conquer approach. We shall only explain how to find $X_k(z)$; essentially the same argument applies for $Y_k(w)$.

Let us define $\Theta_{p:q}(z)$ and $X_{p:q}(z)$ by

$$\Theta_{p:q}(z) \equiv \Theta_p(z)\Theta_{p+1}(z)\cdots\Theta_q(z),$$

$$X_{p:q}(z) \equiv X_{0:q}(z)\otimes_{Ff}\Theta_{0:p-1}(z), \quad X_{0:q}(z) \equiv X_0(z) \mod z^{q+1},$$

where $0 \le p \le q$. The polynomial matrix $\Theta_{p:q}(z)$ has a degree $q-p+1$. The polynomial vector $X_{p:q}(z)$ has degree q, and is obtained by dropping from $X_p(z)$ all terms of degree higher than z^q. Also note the useful properties,

$$[x(z)\otimes_F\theta_1(z)]\otimes_F\theta_2(z) = x(z)\otimes_F[\theta_1(z)\theta_2(z)],$$

$$[x_1(z)+x_2(z)]\otimes_F\theta(z) = [x_1(z)\otimes_F\theta(z)] + [x_2(z)\otimes_F\theta(z)].$$

These properties and the fact that $\Theta_{p:q}(z)$ is completely determined by $X_{p:q}(z)$ allow a divide-and-conquer implementation of the generalized Schur algorithm.

Given $X_{p:q}(z)$, we can compute $\Theta_{p:q}(z)$ as follows. If $p = q$, then we are successful, and compute $\Theta_{p:p}(z) = \Theta_p(z)$. Otherwise, we choose an appropriate† *division point* r such that $p < r < q$, and try to solve the smaller sub-problem of finding $\Theta_{p:r-1}(z)$, given $X_{p:r-1}(z)$. Once we know $\Theta_{p:r-1}(z)$, we can compute $X_{r:q}(z)$ by

† See Sec 4.

$$X_{r:q}(z) = X_{0:q}(z) \otimes_{F^f} \Theta_{0:r-1}(z) = [X_{0:q}(z) \otimes_{F^f} \Theta_{0:p-1}(z)] \otimes_{F^f} \Theta_{p:r-1}(z) \tag{15a}$$

$$= X_{p:q}(z) \otimes_{F^f} \Theta_{p:r-1}(z). \tag{15b}$$

Now we again try to find $\Theta_{r:q}(z)$ given $X_{r:q}(z)$. After we obtain $\Theta_{r:q}(z)$, we can combine the two results, $\Theta_{p:r-1}(z)$ and $\Theta_{r:q}(z)$, by multiplication,

$$\Theta_{p:q}(z) = \Theta_{p:r-1}(z)\Theta_{r:q}(z). \tag{16}$$

Programming details of the above *recursive generalized Schur algorithm* are shown in the Appendix.

The previous recursive description can be visualized nonrecursively using *trees* (see Fig 1 and 2). Each node in the tree is annotated with the *rules:* "find", "apply" and "combine",

$f_{p:p}$: Find $\Theta_{p:p}(z)$,

$a_{p:q}$: $X_{r:q}(z) := X_{p:q}(z) \otimes_F \Theta_{p:r-1}(z)$,

$c_{p:q}$: $\Theta_{p:q}(z) := \Theta_{p:r-1}(z)\Theta_{r:q}(z)$.

We traverse the tree in *post-order* (i.e., follow the order labeled on each node of the tree), and evaluate the rules.

Now, we shall consider two examples in detail.

Example 4. Pseudo-Inverse of pre and post windowed Toeplitz Matrices.

Consider the matrix E_3 in Example 1, where

$$T^T T = \begin{bmatrix} 16 & 8 & 4 & 1 \\ 8 & 16 & 8 & 4 \\ 4 & 8 & 16 & 8 \\ 1 & 4 & 8 & 16 \end{bmatrix}, \quad T^T = \begin{bmatrix} 3 & 2 & 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 3 & 2 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 3 & 2 & 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 3 & 2 & 1 & 1 & -1 \end{bmatrix}.$$

It is desired to find a displacement representation of $(T^T T)^{-1} T^T$. This can be done by the 4-step recursive generalized Schur algorithm. The input to the algorithm is a generator $\{X_0(z), Y_0(w)\}$ of

$$E_3 = \begin{bmatrix} T^T T & T^T \\ -I & O \end{bmatrix},$$

with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, where $F^f = Z_n \oplus Z_n$, $F^b = Z_n \oplus Z_m$. The output, $\{X_4(z), Y_4(w)\}$ is a generator of $(T^T T)^{-1} T^T$, with respect to $\{\otimes_{Z_n}, \otimes_{Z_m}\}$. The computational sequence is illustrated in Fig 1, where it was assumed that the division points were chosen successively by 2, 1 and 3.

(1). $f_{0:0}$: $\Theta_{0:0}(z) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$ because $X_{0:0}(z) = [4, 0]$

(2). $a_{0:1}$: $X_{1:1}(z) = X_{0:1}(z) \otimes_{F^f} \Theta_{0:0}(z) = [4 + 2z, 2z] \otimes_{F^f} \Theta_{0:0}(z) = [4z, -2z]$

(3). $f_{1:1}$: $\Theta_{1:1}(z) = \frac{2}{\sqrt{3}} \cdot \begin{bmatrix} 1 & +1/2 \\ -1/2 & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$

(4). $c_{0:1}$: $\Theta_{0:1}(z) = \Theta_{0:0}(z)\Theta_{1:1}(z) = \frac{2}{\sqrt{3}} \cdot \begin{bmatrix} z^2 & -z/2 \\ -z/2 & 1 \end{bmatrix}$

(5). $a_{0:3}$: $X_{2:3}(z) = X_{0:3}(z) \otimes_{F'} \Theta_{0:1}(z) = \frac{2}{\sqrt{3}} \cdot [3z^2 + 3z^3/2, -z^3/4]$

(6). $f_{2:2}$: $\Theta_{2:2}(z) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$ because $X_{2:2}(z) = \frac{2}{\sqrt{3}} \cdot [3z^2, 0]$

(7). $a_{2:3}$: $X_{3:3}(z) = X_{2:3}(z) \otimes_{F'} \Theta_{2:2}(z) = \frac{2}{\sqrt{3}} \cdot [3z^3, z^3/4]$

(8). $f_{3:3}$: $\Theta_{3:3}(z) = \frac{12}{\sqrt{143}} \cdot \begin{bmatrix} 1 & 1/12 \\ -1/12 & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$

(9). $c_{2:3}$: $\Theta_{2:3}(z) = \Theta_{2:2}(z)\Theta_{3:3}(z) = \frac{12}{\sqrt{143}} \cdot \begin{bmatrix} z & z/12 \\ 1/12 & 1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$

(10). $c_{0:3}$: $\Theta_{0:3}(z) = \Theta_{0:1}(z)\Theta_{2:3}(z) = \frac{24}{\sqrt{3}\sqrt{143}} \cdot \begin{bmatrix} z^4 - z^2/24 & z^3/12 - z/12 \\ -z^3/12 + z/12 & -z^2/24 + 1 \end{bmatrix}$

(11). $a_{0:7}$: $X_{4:7}(z) = [4 + 2z + z^2 + z^3/4 - z^4/4,\ 2z + z^2 + z^3/4 - z^4/4] \otimes_{F'} \Theta_{0:3}(z)$

$$= [(4 + 2z + z^2 + z^3/4,\ 2z + z^2 + z^3/4) - z^4(1/4,\ 1/4)] \otimes_{F'} \Theta_{0:3}(z)$$

$$= -z^4[(1/4,\ 1/4)\Theta_{0:3}(z) \bmod z^4]$$

$$= -\frac{6z^4}{\sqrt{3}\sqrt{143}}[z/12 - z^2/24 - z^3/2,\ \ 1 - z/2 - z^2/24 + z^3/12]$$

Because $T^T T$ is symmetric, $\Psi_{0:3}(w) = \Theta_{0:3}(w)\Sigma$, where $\Sigma = 1 \oplus -1$, and therefore,

$$Y_{4:13}(w) = [(4 + 2z + z^2 + z^3/4) + z^4(3/4 + z/2 + z^2/4 - z^4/4),$$

$$(2z + z^2 + z^3/4) + z^4(3/4 + z/2 + z^2/4 + z^3/4 - z^4/4)] \otimes_{F'} \Theta_{0:3}(w)\Sigma$$

$$= \frac{z^4 6}{\sqrt{3}\sqrt{143}}[1/4z + z^2/24 - 3z^3/2 + 49z^4/24 + 11z^5/8 + 13z^6/24 + 3z^7/2,$$

$$-3 - z/2 + z^2/8 - 2z^3/3 + 11z^4/8 - 13/24z^5 - z^6/8 - z^7/12].$$

Therefore,

$$(T^T T)^{-1} T^T = \gamma^2[L(\mathbf{x}_1)L^T(\mathbf{y}_1) + L(\mathbf{x}_2)L^T(\mathbf{y}_2)], \quad \gamma = \frac{6}{\sqrt{3}\sqrt{143}},$$

where $L(\mathbf{x}_i)$ and $L(\mathbf{y}_i)$ are the lower triangular Toeplitz matrices whose first columns are $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively, and

$$\mathbf{x}_1 = [0,\ -1/12,\ 1/24,\ 1/2]^T,$$

$$\mathbf{x}_2 = [-1,\ 1/2,\ 1/24,\ -1/12]^T,$$

$$\mathbf{y}_1 = [0,\ 1/4,\ 1/24,\ -3/2,\ 49/24,\ 11/8,\ 13/24,\ 3/2]^T,$$

$$\mathbf{y}_2 = [-3,\ -1/2,\ 1/8,\ -2/3,\ 11/8,\ -13/24,\ -1/8,\ 1/12]^T. \qquad \square$$

Remark 1. For a symmetric generator of length 2 with $\beta = 1$, the 2×2 polynomial matrix $\Theta(z)$ in (14) can have the form (hyperbolic reflection)

$$\Theta_i(z) = \begin{bmatrix} ch_i z & sh_i \\ -sh_i z & -ch_i \end{bmatrix}, \qquad ch_i^2 - sh_i^2 = 1.$$

Let

$$\Theta_{p:q}(z) \equiv \Theta_p(z)\Theta_{p+1}(z) \cdots \Theta_q(z) \equiv \begin{bmatrix} \Theta_{1,1}(z) & \Theta_{1,2}(z) \\ \Theta_{2,1}(z) & \Theta_{2,2}(z) \end{bmatrix}.$$

Then, by induction, one can easily prove that

$$z^{q-p+1}\Theta_{1,1}(z^{-1}) = (-1)^{q-p+1}\Theta_{2,2}(z), \quad z^{q-p+1}\Theta_{1,2}(z^{-1}) = (-1)^{q-p+1}\Theta_{2,1}(z).$$

Therefore, we need to compute and store only two entries of $\Theta_{p:q}(z)$.

Remark 2. For an unwindowed scalar Toeplitz matrix, the matrix E_2 in (3) has a displacement rank 4, whereas the matrix E_3 has a displacement rank 5. Therefore, it is more efficient to find a displacement representation of $(T^T T)^{-1}$ rather than of $(T^T T)^{-1}T^T$ when we solve Toeplitz least squares problems. With the notation in (4), the matrix E_2 for an unwindowed scalar Toeplitz matrix $T = (t_{i-j}) \in \mathbf{R}^{m \times n}$ $(m \geq n)$ has a generator [13],

$$\mathbf{w}_1 = T^T \mathbf{t}_1 / \|\mathbf{t}_1\|, \quad \mathbf{w}_2 = \mathbf{t}_2, \quad \mathbf{w}_3 = Z_n Z_n^T \mathbf{w}_1, \quad \mathbf{w}_4 = Z_n \mathbf{l},$$

$$\mathbf{t}_1 = [t_0, t_1, \cdots, t_{m-1}]^T, \quad \mathbf{t}_2 = [0, t_{-1}, \cdots, t_{1-n}]^T, \quad \mathbf{l} = [t_{m-1}, \cdots, t_{m-n}]^T$$

$$\mathbf{v}_1 = \mathbf{v}_3 = \mathbf{e}_1 / \|\mathbf{t}_1\|, \quad \mathbf{v}_2 = \mathbf{v}_4 = \mathbf{0},$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{e}_1$ is the vector with 1 at the first position, and zeros elsewhere.

Example 5. Displacement Representation for the inverse of a Sylvester Matrix.

Let T denote the following Sylvester matrix,

$$T \equiv \begin{bmatrix} 2 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 2 & 1 \\ 3 & 1 & 2 & 1 & 2 \\ 0 & 3 & 1 & 1 & 1 \\ 0 & 0 & 3 & 0 & 1 \end{bmatrix} \tag{17}$$

and suppose that it is desired to obtain a displacement representation of T^{-1}. Then the appropriate extended matrix is

$$E_1 = \begin{bmatrix} T & I \\ -I & O \end{bmatrix}, \tag{18}$$

and it is easy to see that the following $\{X_0(z), Y_0(w)\}$ is a generator of E_1 with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, where $F^f = Z_5 \oplus Z_5$, $F^b = Z_3 \oplus Z_2 \oplus Z_5$;

$$X_0(z) \equiv [x_1(z), x_2(z), x_3(z)], \quad Y_0(w) \equiv [y_1(w), y_2(w), y_3(w)]$$

$$x_1(z) = 2 + z + 3z^2 - z^5, \quad x_2(z) = 1 + 2z + z^2 + z^3 - z^8, \quad x_3(z) = 1, \tag{19a}$$

$$y_1(w) = 1, \quad y_2(w) = w^3, \quad y_3(w) = w^5 \tag{19b}$$

Now the 5-step recursive generalized Schur algorithm gives a desired generator of T^{-1}, with respect to $\{Z_5, Z_5\}$, and a possible computational sequence is shown in Fig 2, where the division points are chosen successively as 2, 1, 3 and 4.

(1). $f_{0:0}$: $\Theta_{0:0}(z) = \begin{bmatrix} z & -1/2 & -1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\Psi_{0:0}(w) = \begin{bmatrix} w & 0 & 0 \\ w/2 & 1 & 0 \\ w/2 & 0 & 1 \end{bmatrix}$

(2). $a_{0:1}$: $X_{1:1}(z) = [2z, 3z/2, -z/2]$, $Y_{1:1}(w) = [w, 0, 0]$

(3). $f_{1:1}$: $\Theta_{1:1}(z) = \begin{bmatrix} z & -3/4 & -1/4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\Psi_{1:1}(w) = \begin{bmatrix} w & 0 & 0 \\ 3w/4 & 1 & 0 \\ -w/4 & 0 & 1 \end{bmatrix}$

(4). $c_{0:1}$: $\Theta_{0:1}(z) = \begin{bmatrix} z^2 & -3z/4-1/2 & z/4-1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\Psi_{0:1}(w) = \begin{bmatrix} w^2 & 0 & 0 \\ w^2/2+3w/4 & 1 & 0 \\ w^2/2-w/4 & 0 & 1 \end{bmatrix}$

(5). $a_{0:4}$: $X_{2:4}(z) = [2z^2+z^3+3z^4, -5z^2/4-5z^3/4, -5z^2/4+3z^3/4]$

$$Y_{2:4}(w) = Y_{0:4}(w) \otimes_{F^b} \Psi_{0:1}(w)$$

$$= [(1, 0, 0)\Psi_{0:1}(w) \mod w^3] + w^3[(0, w, 0)\Psi_{0:1}(w) \mod w^2]$$

$$= [w^2+3w^4/4, w^3, 0]$$

(6). $f_{2:2}$: $\Theta_{2:2}(z) = \begin{bmatrix} z & 5/8 & 5/8 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $\Psi_{2:2}(w) = \begin{bmatrix} w & 0 & 0 \\ -5w/8 & 1 & 0 \\ -5w/8 & 0 & 1 \end{bmatrix}$

(7). $a_{2:4}$: $X_{3:4}(z) = [2z^3+z^4, -5z^3/8+15z^4/8, 11z^3/8+15z^4/8]$,

$$Y_{3:4}(w) = Y_{2:4}(w) \otimes_{F^b} \Psi_{2:2}(w)$$

$$= [(w^2, 0, 0)\Psi_{2:2}(w) \mod w^3] + w^3[(3w/4, 1, 0)\Psi_{2:2}(w) \mod w^2]$$

$$= [-5w^4/8, w^3, 0]$$

(8). $f_{3:3}$: $\Theta_{3:3}(z) = \begin{bmatrix} 0 & 1 & 0 \\ z & 16/5 & 11/5 \\ 0 & 0 & 1 \end{bmatrix}$, $\Psi_{3:3}(w) = \begin{bmatrix} -16w/5 & 1 & 0 \\ w & 0 & 0 \\ -11w/5 & 0 & 1 \end{bmatrix}$

(9). $a_{3:4}$: $X_{4:4}(z) = [-5z^4/8, 7z^4, 6z^4]$, $Y_{4:4}(w) = [w^4, -5w^4/8, 0]$

(10). $c_{4:4}$: $\Theta_{4:4}(z) = \begin{bmatrix} z/(2\sqrt{2}) & 28/(5\sqrt{2}) & 6/5 \\ -5z/(16\sqrt{2}) & 1/(2\sqrt{2}) & -3/4 \\ 0 & 0 & 1 \end{bmatrix}$ $\Psi_{4:4}(w) = \begin{bmatrix} w/(2\sqrt{2}) & 5/(16\sqrt{2}) & 0 \\ -28w/(5\sqrt{2}) & 1/(2\sqrt{2}) & 0 \\ -12\sqrt{2}w/5 & 0 & 1 \end{bmatrix}$

After evaluating, $c_{3:4}$, $c_{2:4}$ and $c_{0:4}$, we obtain $\Theta_{0:4}(z)$ and $\Psi_{0:4}(w)$, and finally

(14). $a_{0:9}$: $X_{0:9}(z) = [x_1(z), x_2(z), x_3(z)] \otimes_{F^f} \Theta_{0:4}(z)$

$$= z^5[(-1, -z^3, 0) \otimes_{F^f} \Theta_{0:4}(z)]$$

$$= z^5[(-1, -z^3, 0)\Theta_{0:4}(z) \mod z^5] = z^5[u_1(z), u_2(z), u_3(z)],$$

where

$$u_1(z) = -z/(2\sqrt{2}) - z^2/(2\sqrt{2}) + z^3/\sqrt{2} + z^4/\sqrt{2}$$

$$u_2(z) = 4/(5\sqrt{2}) + 4z/\sqrt{2} + 16z^2/(5\sqrt{2}) - 28z^3/(5\sqrt{2}) - 28z^4/(5\sqrt{2})$$

$$u_3(z) = 2/5 + z/5 + 2z^2/5 + z^3/5 - 6z^4/5.$$

$$Y_{0:9}(w) = [y_1(w), y_2(w), y_3(w)] \otimes_{F^b} \Psi_{0:4}(w)$$

$$= w^5[(0, 0, 1) \otimes_{F^b} \Psi_{0:4}(w)] = w^5[v_1(w), v_2(w), v_3(w)],$$

where

$$v_1(w) = -12\sqrt{2}w/5 + 12w^2/(5\sqrt{2}) + 12w^3/(5\sqrt{2}) - 12w^4/(5\sqrt{2}),$$

$$v_2(w) = -w/\sqrt{2} + w^2/(2\sqrt{2}) + w^3/(2\sqrt{2}) - w^4/(2\sqrt{2}),$$

$$v_3(w) = 1.$$

Therefore,

$$T^{-1} = L(\mathbf{u}_1)L^T(\mathbf{v}_1) + L(\mathbf{u}_2)L^T(\mathbf{v}_2) + L(\mathbf{u}_3)L^T(\mathbf{v}_3),$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the vectors whose jth component is the coefficient of z^{j-1} and w^{j-1} of $u_i(z)$ and $v_i(w)$, respectively.

□

Remark 1. If we had chosen the displacement operator $F^f = Z_5 \oplus Z_3 \oplus Z_2$, $F^b = Z_3 \oplus Z_2 \oplus Z_5$ for the matrix T in (17) we would have the same generator (19) for E_1, but the obtained generator of T^{-1} would be the one with respect to $\{Z_3 \oplus Z_2, Z_5\}$ rather than $\{Z_5, Z_5\}$. The displacement ranks of T^{-1} with respect to both of the displacement operators are 2, and the above procedure gives non-minimal generators of length 3.

Remark 2, The following extended matrix

$$\begin{bmatrix} T & \mathbf{b} \\ -I & 0 \end{bmatrix}, \quad T = \text{Sylvester matrix} \tag{20}$$

also has a displacement rank 3. One could as well obtain the solution $T^{-1}\mathbf{b}$ directly by applying recursive generalized Schur algorithm to (20); the last column of X, where $\{X, \mathbf{y}\}$ is the computed generator of $T^{-1}\mathbf{b}$ with respect to $\{Z_n, 1\}$, will be $T^{-1}\mathbf{b}$.

4. Polynomial Products with Fast Convolutions.

The product of two polynomials of degree d_1 and d_2 can be performed efficiently using $d \equiv d_1+d_2+1$ point fast cyclic convolution algorithms [4]. If d is a power of two, then a d-point fast cyclic convolution needs $O(d\log d)$ flops. If d is not a power of two, but a highly composite number, then the number of computations is close to $O(d\log d)$. Among others, fast Fourier transformations (FFT's) can be used for convolutions; Ammar and Gragg [2] carefully examined the use of FFT's for a doubling algorithm for square Toeplitz systems of equations. We shall only consider the subtle complications that arise in the recursive generalized Schur algorithm in this paper.

The polynomial matrix-matrix product of (16) needs α^3 of $q-p$ point cyclic convolutions. The polynomial vector-matrix product of (15b) has α^2 of scalar polynomial products of the form, $x(z)\otimes_{F^f}\theta(z)$, where $x(z)$ is a polynomial with nonzero terms of $z^p, z^{p+1}, \cdots, z^q$. Let us assume that

$$0 < \delta_1 < \cdots < \delta_l \le p < \delta_{l+1} < \cdots < \delta_s \le r < \delta_{s+1} < \cdots < \delta_t \le q < \delta_{t+1} < \cdots < \delta_N .$$

Then

$$x'(z) \equiv x(z)\otimes_{F^f}\theta(z) \tag{21a}$$

$$= [z^{\delta_l}x_{l+1}(z) + z^{\delta_{l+1}}x_{l+2}(z) + \cdots + z^{\delta_s}x_{s+1}(z) + \cdots + z^{\delta_t}x_{t+1}(z)]\otimes_{F^f}\theta(z) \tag{21b}$$

$$= [z^{\delta_l}x_{l+1}(z) + \cdots + z^{\delta_{s-1}}x_s(z)]\otimes_{F^f}\theta(z) \tag{22a}$$

$$+ z^{\delta_s}[x_{s+1}(z)\theta(z^\beta) \mod z^{n_{s+1}}] \tag{22b}$$

$$+ z^{\delta_{s+1}}[x_{s+2}(z)\theta(z^\beta) \mod z^{n_{s+2}}] \tag{22c}$$

$$\cdots$$

$$+ z^{\delta_t}[x_{t+1}(z)\theta(z^\beta) \mod z^{n_{t+1}}]. \tag{22d}$$

The terms in (22a) do not need to be computed because these terms will be summed to zeros after adding all the partial sums in the vector-matrix multiplication of (15b). Recall that $x_i(z)$ has degree n_i, and $\theta(z^\beta)$ has degree $\beta^{(q-p+1)}$. Therefore, the product $x_i(z)\theta(z^\beta)$ from (22b) to (22d) can be performed by

$2n_i+1$	point cyclic convolutions†	if degree$[\theta(z^\beta)] \ge$ degree$[x_i(z)]$,
$n_i+\beta^{(q-p+1)}+1$	point cyclic convolutions	if degree$[\theta(z^\beta)] <$ degree$[x_i(z)]$.

Remark. Notice two d/2 point convolutions take $cd\log(d/2)$ flops if one d point convolution takes $cd\log d$ flops. Therefore, the polynomial product (21) is more efficient for the displacement operator F^f with more sections, because such displacement operators break a long convolution into many smaller convolutions. Therefore, for a given matrix we prefer to choose a displacement operator with as many sections as possible, while keeping the displacement rank minimal.

If the dimensions of the matrix are powers of 2, then we can always choose the center division point, $r = \lceil (p+q)/2 \rceil$. This *balanced division* (or doubling) gives the least number of computations, in general. For this case, let $\eta \equiv p-q$, and $T(\eta)$ denote the number of computations for one recursion.

† The first and last terms (22b) and (22d) need smaller point convolutions.

Then

$$T(\eta) \le 2T(\eta/2) + W(\eta), \qquad W(\eta) \equiv O(\alpha^3 \eta \log \eta),$$

and therefore, one can show [1] that the k-step recursion takes

$$T(k) \le O(\alpha^3 k \log^2 k).$$

However, in most cases the doubling is not possible, and for such circumstances, the desirable choice of r is such that $r-p$ and $q-r+1$ are highly composite numbers (so that fast convolution algorithms can be applied efficiently), as well as r is close to $(q-p)/2$ (so as to achieve balancing).

Matrix-Vector Products using Displacement Representation.

The final step of finding solutions for linear equations is the matrix-vector multiplication $S\mathbf{b}$, given a displacement representation of $S \in \mathbf{R}^{m \times n}$,

$$S = \sum_{i=1}^{\alpha} K(\mathbf{x}_i, F^f) K^T(\mathbf{y}_i, F^b), \tag{23}$$

where the length α is a multiple of the block size β; $\alpha = \beta\delta$, and

$$F^f = \bigoplus_{i=1}^{M} Z_{m_i}^{\beta}, \quad F^b = \bigoplus_{i=1}^{N} Z_{n_i}^{\beta}.$$

The expression in (23) can be re-written as the *block displacement representation*

$$S = \sum_{i=1}^{\delta} K_{\beta}(X_i, F^f) K_{\beta}^T(Y_i, F^b), \quad X_i \in \mathbf{R}^{m \times \beta}, \quad Y_i \in \mathbf{R}^{n \times \beta}, \tag{24}$$

where

$$K_{\beta}(X_i, F^f) = [X_i, FX_i, F^{f\,2}X_i, \cdots F^{f\,[(m/\beta)-1]}X_i],$$

$$K_{\beta}(Y_i, F^b) = [Y_i, FY_i, F^{b\,2}Y_i, \cdots F^{b\,[(n/\beta)-1]}Y_i].$$

Furthermore, because F^f and F^b have M and N sections, respectively, we can write

$$K_{\beta}(X_i, F^f) = \begin{bmatrix} L_{\beta}(X_{1,i}, Z_{m_1}^{\beta}) & O \\ L_{\beta}(X_{2,i}, Z_{m_2}^{\beta}) & O \\ \vdots & \vdots \\ L_{\beta}(X_{M,i}, Z_{m_M}^{\beta}) & O \end{bmatrix}, \quad K_{\beta}(Y_i, F^b) = \begin{bmatrix} L_{\beta}(Y_{1,i}, Z_{n_1}^{\beta}) & O \\ L_{\beta}(Y_{2,i}, Z_{n_2}^{\beta}) & O \\ \vdots & \vdots \\ L_{\beta}(Y_{N,i}, Z_{n_N}^{\beta}) & O \end{bmatrix},$$

where $L_{\beta}(X, Z^{\beta})$ is the block lower triangular Toeplitz matrix with the first column block X. The matrix O denotes the null matrix with appropriate size such that $K_{\beta}(X_i, F^f)$ and $K_{\beta}(Y_i, F^b)$ are $m \times n$ and $n \times n$ matrices, respectively. The product $L_{\beta}(X, Z^{\beta})\mathbf{b}$ can be expressed as sum of β products of scalar lower triangular Toeplitz matrix and vectors. As an example,

$$\begin{bmatrix} a_0 & c_0 & & \\ a_1 & c_1 & & \\ a_2 & c_2 & a_0 & c_0 \\ a_3 & c_3 & a_1 & c_1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_0 & & & \\ a_1 & a_0 & & \\ a_2 & a_1 & a_0 & \\ a_3 & a_2 & a_1 & a_0 \end{bmatrix} \begin{bmatrix} b_0 \\ 0 \\ b_2 \\ 0 \end{bmatrix} + \begin{bmatrix} c_0 & & & \\ c_1 & c_0 & & \\ c_2 & c_1 & c_0 & \\ c_3 & c_2 & c_1 & c_0 \end{bmatrix} \begin{bmatrix} b_1 \\ 0 \\ b_3 \\ 0 \end{bmatrix}. \tag{25}$$

Now the multiplications in the right sides of (25) can be done by fast convolutions, and therefore, so does the multiplication $S\,\mathbf{b}$.

5. Concluding Remarks.

We have presented $O(\alpha^3 n \log^2 n)$ algorithms for the determination of exact and least squares solutions of linear systems with matrices having (generalized) displacement rank α. Such algorithms for exact solutions have been studied by several authors, most recently by Ammar and Gragg [2] for Toeplitz systems. They also made a very close study of the implementation of the convolution operation in an attempt to obtain the smallest coefficient; we have not attempted so close an analysis for the more general algorithm in our paper. Nor have we attempted a numerical error analysis of the algorithm; nevertheless one might hope that numerical refinements devised for the Schur algorithm (see e.g., Koltracht and Lancaster [16]) may be carried over to the divide-and-conquer framework as well.

APPENDIX

We shall summarize the explanation in Sec 3 using a recursive procedure. First, note that the polynomial $\Theta_{p:q}(z)$ (and $\Psi_{p:q}(z)$) has $q-p+2$ terms. The first column of $\Theta_{p:q}(z)$ has terms ranging from degree z to z^{q-p+1}, and the other columns have terms from 1 to z^{q-p}. Hence, by shifting the first column by one position, we can store $\Theta_{p:q}(z)$ and $\Psi_{p:q}(z)$ in the array "Poly" from p to q slots inclusive:

```
Poly: array [1..α, 1..α, 0..MAX-1] of record
    θ: coefficients;
    ψ: coefficients
end;
```

The computation of $\Theta_{p:q}(z)$ is sequential, i.e., once we compute $\Theta_{p:q}(z)$, we do not need to keep $\Theta_{p:r-1}(z)$, and therefore, the array "Poly" can be kept as a single global variable.

The polynomial vector $X_{p:q}(z)$ has $q-p+1$ terms, and therefore, can be stored in an array type GENERATORS:

```
type
    GENERATORS = array [1..α, 0..MAX-1] of record
        x: coefficient;
        y: coefficient
    end;
```

However, $X_{p:q}(z)$ cannot be kept as a global variable, and local copies should be maintained during each recursive call.

Now we can describe the recursive generalized Schur algorithm as follows.

Algorithm (Recursive k-step Generalized Schur Algorithm).

Input: Generator of E, $\{X_0(z), Y_0(w)\}$; displacement operator $\{\otimes_{F^f}, \otimes_{F^b}\}$; Number of steps, k.
Output: Generator of S, $\{X_k(z), Y_k(w)\}$;

```
procedure RecursiveSchur
    var
        G, LowerG: GENERATORS;
    begin
        Find(0, k-1, G);
        Apply(0, k, n, G, LowerG);
        return (LowerG)
    end
```

The procedure Find(p, q, G) computes $\Theta_{p:q}(z)$, and $\Psi_{p:q}(w)$ given $\{X_{p:q}(z), Y_{p:q}(w)\}$, and the procedure Apply(p, r, q, G, LowerG) returns LowerG = $\{X_{r:q}(z), Y_{r:q}(w)\}$ given G = $\{X_{p:q}(z), Y_{p:q}(w)\}$

```
procedure Find(p, q: index; G: GENERATORS);
    var
        r : index;
        G, LowerG: GENERATORS;
    begin
        if p = q then begin
            Compute Θ_{p:p}(z) and Ψ_{p:p}(w);
            return
        end
        r := appropriate integer close to ⌈(p+q)/2⌉;
        Find(p. r-1, G);
        Apply(p, r, q, G, LowerG);
        Find(r, q, LowerG);
        (* Use fast convolution for polynomial products *)
        Θ_{p:q}(z) := Θ_{p:r-1}(z)Θ_{r,q}(z);
        Ψ_{p:q}(w) := Ψ_{p:r-1}(w)Ψ_{r,q}(w)
    end

procedure Apply(p, r, q: index; G: GENERATORS; var LowerG: GENERATORS);
    begin
        (* Use fast convolution for polynomial products *)
        X_{r:q}(z) := X_{p:q}(z) ⊗_{F^f} Θ_{p:r-1}(z);
        Y_{r:q}(w) := Y_{p:q}(w) ⊗_{F^b} Ψ_{p:r-1}(w);
        LowerG := {X_{r:q}(z), Y_{r:q}(w)}
        return (LowerG);
    end
```

REFERENCES

[1]. A. Aho, J. Hopcroft and J. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley, Reading, MA, 1974. p. 305.

[2]. G. Ammar and W. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal., Appl., Vol. 9, No. 1, Jan, (1988), pp. 61-76.

[3]. G. Bitmead and B. D. O. Anderson, *Asymptotically fast solution of Toeplitz and related systems of linear equations*, Linear Algebra and its Appli., 34 (1980), pp. 103-116.

[4]. R. Blahut, *Fast algorithms for digital signal processing*, Addison-Wesley, Reading, MA, 1985.

[5]. R. Brent, F. Gustavson and D. Yun *Fast Solution of Toeplitz Systems of Equations and Computation of Pade Approximants* Journal of Algorithms, 1, (1980), pp. 259-295

[6]. A. Bruckstein and T. Kailath, *Doing inverse scattering the fast(est) way*, Technical Report, Stanford University, CA 94305, July, 1985.

[7]. W. Choate, *A fast algorithm for normal incidence seismograms*, Geophysics, vol. 47, No. 2, Feb., (1982), pp. 196-202.

[8]. J. Chun and T. Kailath, *Unified approach for matrix factorization using generalized Schur algorithm*, Pre-print, Stanford University, CA 94305, 1988.

[9]. F. De Hoog, *A new algorithm for solving Toeplitz systems of equations, Linear Algebra and its Appli., 88/89, (1987), pp. 122-138.*

[10]. I. Gohberg and I. Fel'dman, *Convolution equations and projection methods for their solutions*, Translations of Mathematical Monographs, vol. 41, Amer. Math. Soc., 1974.

[11]. I. Gohberg and A. Semencul, *On the inversion of finite Toeplitz matrices and their continuous analogs*, Mat. Issled., 2 (1972), pp. 201-233.

[12]. T. Kailath, *Signal processing applications of some moment problems*, Proceedings of Symposia in Applied Mathematics, vol. 37, 1987 pp. 71-109.

[13]. T. Kailath and J. Chun, *Generalized Gohberg-Semencul formulas for matrix inversion*, Proc. Symp. on Operator Theory and Applications, Calgary, Canada, Aug. 1988.

[14]. T. Kailath, S. Kung and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979) pp. 395-407. See also Bull. Amer. Math. Soc., 1 (1979), pp. 769-773.

[15]. T. Kailath, A. Vieira and M. Morf, *Inverses of Toeplitz operators, innovations, and orthogonal polynomial*, SIAM Review, vol. 20, No 1, Jan. (1978), pp. 106-119.

[16]. I. Koltracht and P. Lancaster, *Threshold algorithms for the prediction of reflection coefficients in a layered medium*, Geophysics, vol. 53, No. 7, Jul., (1988), pp. 908-919.

[17]. H. Lev-Ari and T. Kailath, *Triangular factorization of structured Hermitian matrices*, Operator Theory, Advances and Applications, vol. 18, Birkhauser, Boston, pp. 301-324, (1986).

[18]. W. McClary, *Fast seismic inversion*, Geophysics, vol. 48, No. 10, Oct., (1983), pp. 1371-1372.

[19]. M. Morf, *Doubling algorithms for Toeplitz and related equations*, Proceedings of the IEEE International Conf. on ASSP, Denver, (1980), pp. 954-959.

[20]. B. Musicus, *Levinson and fast Cholesky algorithms for Toeplitz and almost Toeplitz matrices*, Report, Research Lab. of Electronics, MIT, Cambridge, MA 1984.

[21]. I. Schur, *Über Potenzreihen, die im Innern des Einheitskreises beschrankt sind*, J. für die Reine und Angewandte Mathematik, 147 (1917), pp. 205-232.

[22]. I. Schur, *On power series which are bounded in the interior of the unit circle. I.* (English translation of [21]), Operator Theory, Advances and Applications, vol. 18, Birkhauser, Boston, pp. 31-60, (1985).

[23]. H. Sexton, M. Shensa and J. Speiser, *Remarks on a displacement-rank inversion method for Toeplitz systems*, Linear Algebra and its Appli., 45, (1982), pp. 127-130.

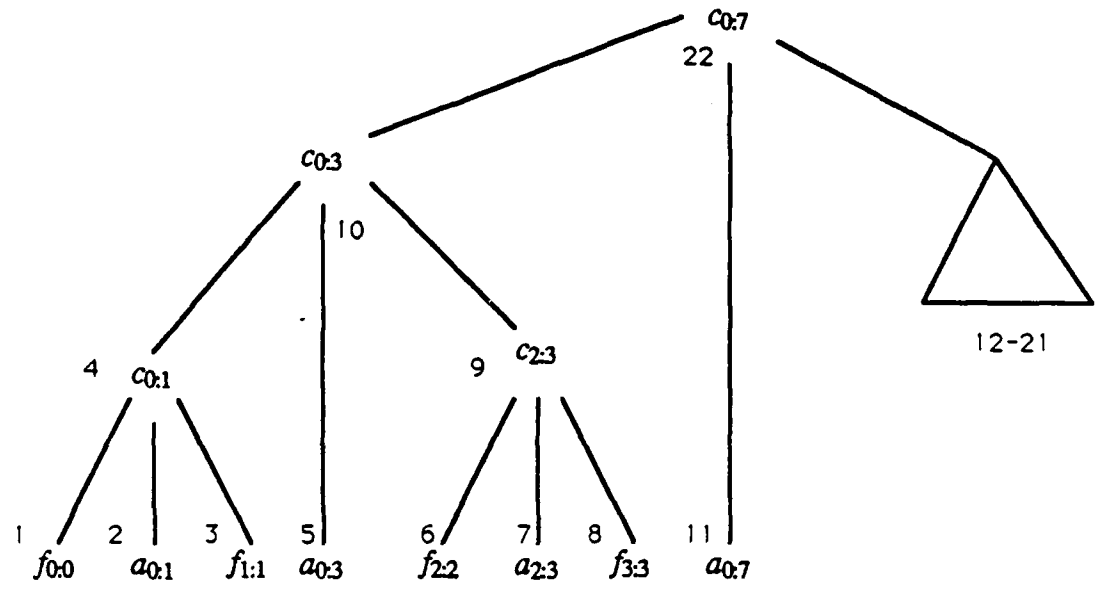


Fig 1. Sequence of Computations for Example 4.

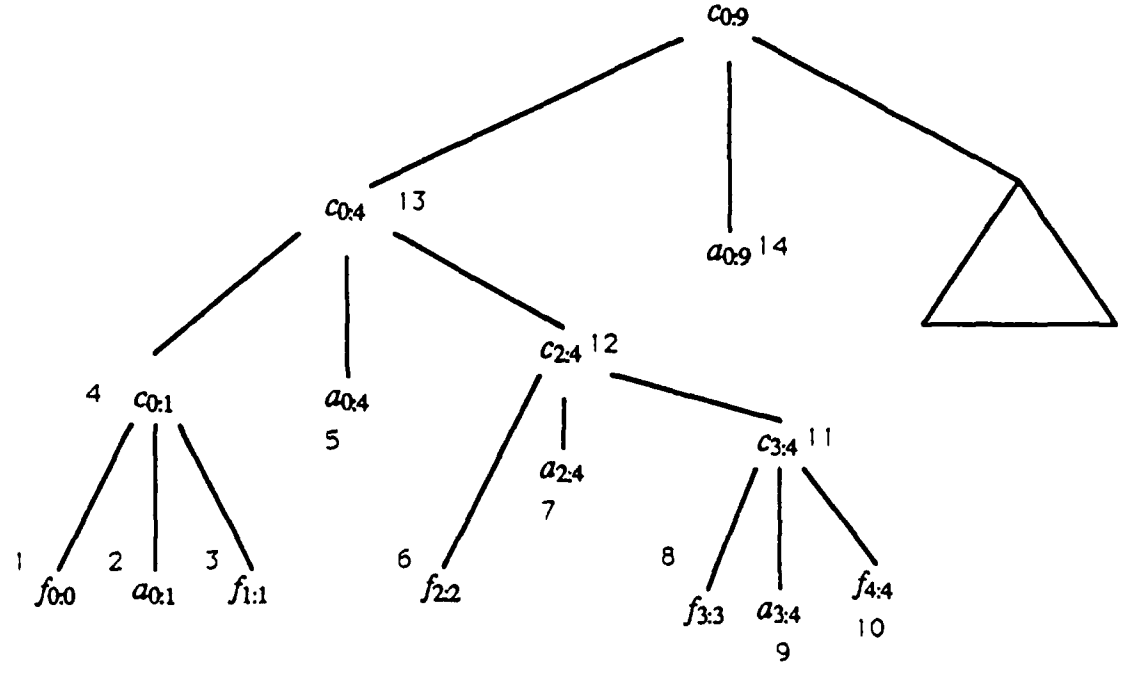


Fig 2. Sequence of Computations for Example 5.

RECENT DEVELOPMENTS IN HIGH-PERFORMANCE ELEMENTS BASED ON THE FREE FORMULATION

CARLOS A. FELIPPA

Department of Aerospace Engineering Sciences
and Center for Space Structures and Controls
University of Colorado
Boulder, Colorado 80309-0429, USA

ABSTRACT. The free formulation of Bergan and Nygård (1984) has enjoyed considerable success in the construction of high-performance finite elements for linear and nonlinear structural analysis. In its original form the formulation combines nonconforming internal displacement assumptions with a specialized version of the patch test. Recent developments in fitting this formulation within a variational framework are described, and extensions opened up by these developments discussed.

INTRODUCTION. The term *high-performance finite element* is used here to collectively identify elements that are developed to attain the following goal:

> *To deliver engineering accuracy with coarse, arbitrary meshes of simple elements*

The fulfillment of this goal gives rise to a myriad of requirements, which are to be addressed in higher or lesser degree during element development. Such requirements are listed in Table 1.

Some of these requirements are obvious. For example, *low distortion sensitivity* is an immediate consequence of trying to achieve satisfactory accuracy with *arbitrary* meshes. But other items in Table 1 require some explanation.

A key requirement is that the element be as *simple* as possible. It should be observed that this is in sharp contrast to trends of the late 1960s and 1970s that lauded *higher order elements* and culminated with the development of very complex models, including elements with nonphysical degrees of freedom. One primary source of this "backlash" is feedback from users of general-purpose finite element programs. As use of these programs expands to more engineers without deep knowledge of "what's inside the black box" the overwhelming preference in model construction is to select the "simplest elements that will do the job" that is available in the program.

Table 1. Target Requirements for High-Performance Elements

- Simple: few freedoms, all physical
- Frame invariant
- No locking
- Rank sufficient: no spurious modes
- Balanced stiffness: not too rigid, not too flexible
- Stresses as accurate as displacements
- Low distortion sensitivity
- Mixable with other elements
- Economical to form
- Easily extendible to nonlinear analysis

The *balanced stiffness* demand also deserves some comment. It follows from the goal of attaining reasonable accuracy with *coarse meshes.* This is illustrated in Figure 1, which shows a convergence study of a classical model problem: the bending of a simply-supported square plate under a concentrated central load. The mesh contains $N \times N$ elements over a plate quadrant. An "accuracy band" of $\pm 1\%$ is taken, somewhat arbitrarily, as representative of engineering accuracy for this rather simple problem. The convergence characteristics of several triangular elements are taken from the extensive study of Batoz, Bathe and Ho (1980). Although most elements converge, some are too stiff while others are too flexible, and generally do not enter the accuracy band until the mesh is fairly refined ($N \geq 8$). On the other hand, the results labelled 'FF', obtained with a plate element based on the free formulation (FF) discussed later, lie within the band for all meshes.

The balanced-stiffness requirement should *not* be misconstrued for fast asymptotic convergence for *fine* meshes. Simple elements cannot compete with higher-order elements in this regard. What is important is *how good are the results for coarse meshes.*

THEME AND TOOLS. Many researchers are presenly working to develop such elements. The common theme of the investigations is

Abandon the conventional displacement formulation

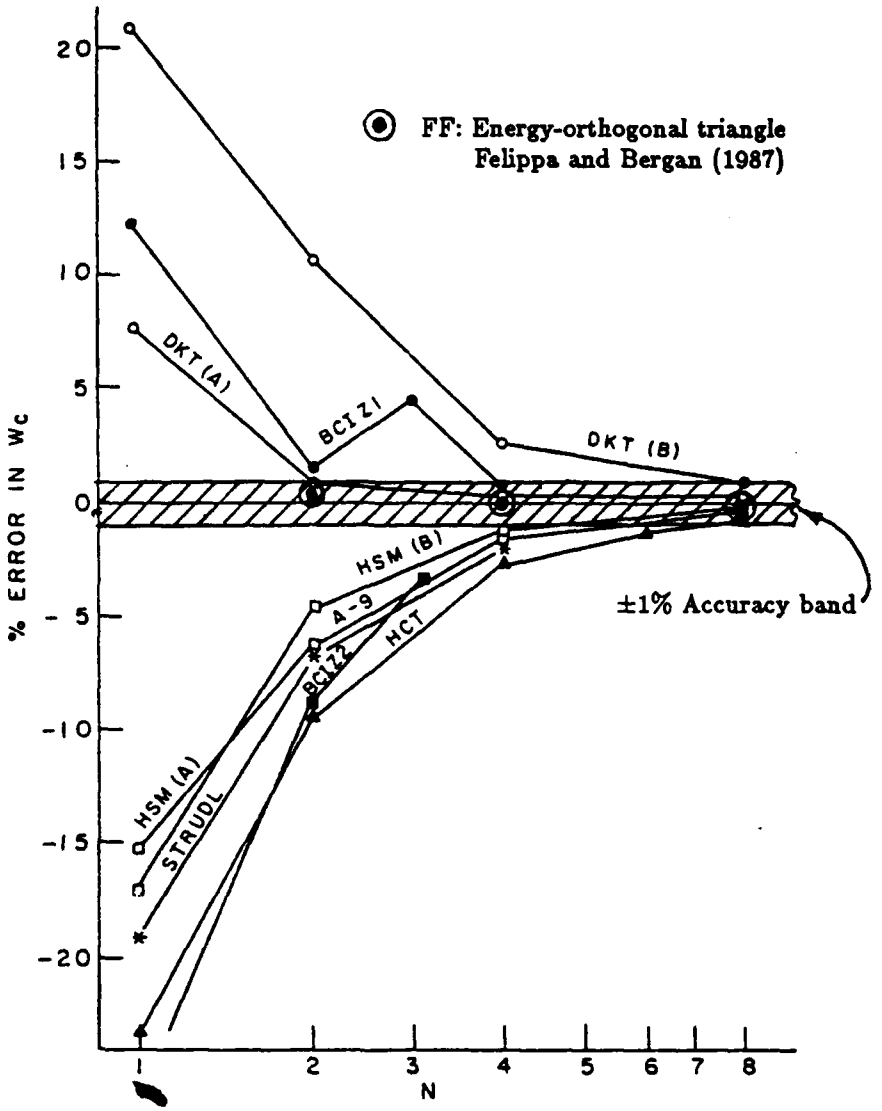


Figure 1. Convergence study of several plate bending finite elements as reported in Batoz *et.al.* (1980). The FF results are from Felippa and Bergan (1987).

Various tools used by these researchers in their quest for high-performance elements are listed in Table 2. It can be observed that many of these were introduced over 20 years ago. But it is only now that a concerted effort is made to combine several tools to forge superior products.

The present paper focuses on *one* of the possible approaches to the construction of high-performance elements. This approach is based on the free formulation (FF).

Table 2. Tools of the Trade

Technique	*Year introduced*
• Incompatible shape functions	1964
• Patch test	1965
• Mixed and hybrid variational principles	1965
• Projectors	1967
• Reduced and selective integration	1969
• Assumed strains	1970
• Energy balancing	1974
• Limit differential equations	1982

THE FREE FORMULATION. In the early 1980s Bergan and Nygård developed the free formulation (FF) for the construction of displacement-based, incompatible finite elements. This work, published in Bergan and Nygård (1984), consolidated a decade of research of Bergan and coworkers at Trondheim, milestones of which may be found in Bergan and Hanssen (1976), Hanssen *et.al.* (1979) and Bergan (1980). The products of this research have been finite elements of high performance, especially for plates and shells. Linear applications are reported in the aforementioned papers as well as in Bergan and Wang (1984), Bergan and Felippa (1985) and Felippa and Bergan (1987); whereas nonlinear applications are presented in Bergan and Nygård (1985, 1988) and Nygård (1986).

The basic concept is that the element stiffness matrix can be decomposed into two parts:

$$\mathbf{K} = \mathbf{K}_b + \mathbf{K}_h \tag{1}$$

where

$\mathbf{K}_b$ the basic stiffness matrix, which is constructed for *convergence.*

$\mathbf{K}_h$ the *higher-order* stiffness matrix, which is constructed for *stability* and *accuracy.*

The decomposition (1) may be interpreted at the assembled or master-stiffness equation level as the force decomposition

$$\mathbf{K}^A = (\mathbf{K}_b^A + \mathbf{K}_h^A)\mathbf{v} = \mathbf{f}_b^A + \mathbf{f}_h^A = \mathbf{f}^A \tag{2}$$

where $\mathbf{v}$ and $\mathbf{f}^A$ are the vectors of nodal displacements and assembled nodal forces, respectively. A FF postulate is that as the mesh size decreases and the solution converges, $\mathbf{K}_b\mathbf{v}$

dominates.

The original FF was based on nonconforming displacement assumptions, the principle of virtual work and a specialized form of Irons' patch test that Bergan and Hanssen (1976) called the individual element test. The basic and higher order stiffness are constructed in largely independent fashion by following the procedures outlined below.

CONSTRUCTION OF BASIC STIFFNESS MATRIX. The main steps are outlined below in "recipe" form; for justification the reader is referred to the references listed above.

Step 1. Assume a constant stress, $\boldsymbol{\sigma}$, inside the element.

Step 2. Assume boundary displacements, $\mathbf{d}$, over the element boundary B. This field is described in terms of element node displacements $\mathbf{v}$ as

$$\mathbf{d} = \mathbf{V}\mathbf{v} \tag{3}$$

where $\mathbf{V}$ is an array of boundary shape functions. The boundary motion must satisfy interelement continuity, and contain rigid-body and constant-strain motions exactly.

Step 3. Construct the "lumping" matrix

$$\mathbf{L} = \oint_B \mathbf{V}.\mathbf{n}\, dB \tag{4}$$

that consistently "lumps" the boundary tractions $\boldsymbol{\sigma}.\mathbf{n}$ associated with $\boldsymbol{\sigma}$, into element node forces, $\mathbf{f}$, conjugate to $\mathbf{v}$. That is, $\mathbf{f} = \mathbf{L}\boldsymbol{\sigma}$.

Step 4. The basic stiffness matrix is

$$\mathbf{K}_b = \frac{1}{v}\mathbf{L}\mathbf{E}\mathbf{L}^T \tag{5}$$

where $\mathbf{E}$ is the stress-strain constitutive matrix, assumed to be constant over the element, and $v = \int_V dV$ denotes the element volume.

CONSTRUCTION OF HIGHER-ORDER STIFFNESS MATRIX. Again the key steps are outlined below in "how to do it" form.

Step 1. The same compatible boundary displacements used in constructing $\mathbf{K}_b$ are assumed:

$$\mathbf{d} = \mathbf{V}\mathbf{v} \tag{6}$$

Step 2. Assume an *internal displacement field* over the element volume V:

$$u = \mathbf{N}\mathbf{q} = \underbrace{\mathbf{N}_r\mathbf{q}_r}_{\textit{rigid motion}} + \underbrace{\mathbf{N}_c\mathbf{q}_c}_{\textit{constant-strain}} + \underbrace{\mathbf{N}_h\mathbf{q}_h}_{\textit{higher-order}} \tag{7}$$

where array $\mathbf{N}$ collects shape functions and $\mathbf{q}$ collects generalized coordinates. This assumption satisfies the following conditions:

(a) linear independence with respect to $\mathbf{v}$,

(b) the dimension of vectors $\mathbf{q}$ and $\mathbf{v}$ are the same,

(c) the rigid motions and constant-strain fields are complete,

(d) (optional but recommended) the higher-order displacements are energy orthogonal to the constant-strain displacements.

The associated internal strains are:

$$\mathbf{e} = \mathbf{B}\mathbf{u} = \mathbf{e}_c + \mathbf{e}_h = \mathbf{B}_c\mathbf{q}_c + \mathbf{B}_h\mathbf{q}_h \tag{8}$$

since the rigid-body strains, $\mathbf{B}_r\mathbf{q}_r$, must vanish.

Step 3. Construct the square nonsingular transformation

$$\mathbf{v} = \mathbf{G}\mathbf{q} \tag{9}$$

which inverted gives:

$$\mathbf{q} = \begin{Bmatrix} \mathbf{q}_r \\ \mathbf{q}_c \\ \mathbf{q}_h \end{Bmatrix} = \mathbf{H}\mathbf{v} = \begin{bmatrix} \mathbf{H}_r \\ \mathbf{H}_c \\ \mathbf{H}_h \end{bmatrix} \mathbf{v} \tag{10}$$

Step 4. The higher-order stiffness matrix is

$$\mathbf{K}_h = \mathbf{H}_h^T\mathbf{K}_{qh}\mathbf{H}_h, \qquad \text{where} \qquad \mathbf{K}_{qh} = \int_V \mathbf{B}_h^T\mathbf{C}\mathbf{B}_h\, dV \tag{11}$$

$\mathbf{K}_{qh}$ is the generalized stiffness in terms of the $\mathbf{q}$ coordinates.

Table 3. Elements Developed with FF

Type	*Shape*	*Dofs*
Kirchhoff plates	Triangles (several)	9
	Quadrilaterals	12
Membrane with drilling freedoms	Triangle	9
	Quadrilateral	12
Shells	Triangle	18
	Quadrilateral	24

SCALING THE HIGHER-ORDER STIFFNESS. In more recent work (see Bergan and Felippa (1985) and following papers) the concept of *scaling the higher order stiffness* was introduced. A *one parameter* scaling generalizes (1) to

$$\mathbf{K} = \mathbf{K}_b + (1-\gamma)\mathbf{K}_h \tag{12}$$

where $\gamma < 1$ is a scalar. If $\gamma = 0$ one recovers (1), but higher accuracy for coarse meshes may be obtained by adjusting the value of γ. (This value may vary from element to element.) Multiparameter scaling is discussed in Felippa and Bergan (1987) for a specific plate bending element.

APPLICATIONS. Table 3 lists elements that have been developed using the FF as of this writing. The major code in which these elements have been implemented is FENRIS, developed in collaboration between the Norwegian Institute of Technology (NIT) at Trondheim, SINTEC and Der Norske Veritas; see Bergan et. al. (1984). FENRIS has been primarily used in the analysis of nonlinear marine structures such as offshore drilling platforms. Table 4 lists the major application problems to which these elements have been applied.

Table 4. Application Problems

- Linear plate/shell analysis
- Geometrically nonlinear plate/shell analysis (corotational formulation)
- Materially nonlinear plates and shells

VARIATIONAL FORMULATION. An intriguing question has been: does the FF fit in a variational framework? This was partly answered by Bergan and Felippa (1985), who showed that the basic stiffness part was equivalent to a constant-stress hybrid element. But persistent efforts by the author to encompass the higher order stiffness within a hybrid variational principle were unsuccessful until the development of parametrized mixed-hybrid functionals in Felippa (1988a, 1988b). With the help of these more general functionals it is possible to show that the FF is a very special type of mixed-hybrid element which does not fit within the classical Hellinger-Reissner principle. In retrospect the classification of FF elements as hybrids is not surprising. Under mild conditions studied in Felippa (1988c), hybrid elements satisfy Irons' patch test *a priori*, and the FF development has been founded on that premise.

To encompass the FF within the hybrid framework, the following assumptions must be invoked.

Assumption 1. A non-standard hybrid functional, identified as Π_γ^d in Felippa (1988b), is constructed. This functional depends linearly on a parameter γ. This parameter "interpolates" between the minimum potential energy functional and the Hellinger-Reissner functional, which are obtained for $\gamma = 0$ and $\gamma = 1$, respectively.

Assumption 2. Three fields are assumed over each element:

(a) a constant stress field,

(b) an internal displacement field $\mathbf{u}$ defined by n_q generalized coordinates collected in vector $\mathbf{q}$, and

(c) a boundary displacement field $\mathbf{d}$ defined by n_v nodal displacements collected in vector $\mathbf{v}$. Both $\mathbf{d}$ and $\mathbf{u}$ must represent rigid body motions and constant strain states exactly.

Assumption 3. The number of generalized coordinates, n_q, equals the number of nodal displacements, n_v, and the square transformation matrix $\mathbf{G}$ relating $\mathbf{v} = \mathbf{G}\mathbf{q}$ is *nonsingular*.

The last two assumptions are precisely those invoked in the construction of $\mathbf{K}_h$ as discussed previously. The first one defines the variational principle and accounts for the higher-order stiffness scaling.

In Felippa (1988b) it is shown that substituting the finite element expansions into Π_γ^d, rendering the functional stationary with respect to the degrees of freedom, and eliminating both internal fields by a combination of static condensation and kinematic constraints, leads to the scaled FF stiffness equations (12) in terms of the nodal displacements $\mathbf{v}$. The parameter γ appears as a coefficient of the higher order stiffness. These stiffness equations can be readily implemented into any displacement-based finite element code.

CONCLUDING REMARKS. Why is the FF variational formulation deemed useful? There are several reasons:

1. It explains the behavior of FF elements as regards convergence, stability and accuracy.

2. It opens up the door to extensions that are not obvious from a physical standpoint. Two such extensions involve: retaining higher order stress fields, and allowing more internal displacement modes that nodal displacements, that is, the dimension of vector $\mathbf{q}$ in (7) exceeds that of $\mathbf{v}$ in (6). These extensions are studied in Felippa (1988c).

3. Supplies foundations for local error estimation and adaptive mesh refinement.

4. Facilitates the construction of "designer elements" needed for applications such as stress, stability and vibrations of advanced laminate-composite structures. Such elements may combine the three ingredients of internal statics, internal kinematics and boundary kinematics in harmonious synergy to satisfy special behavior requirements.

ACKNOWLEDGEMENTS. The preparation of this paper was jointly supported by the Office of Naval Research under Contract N0001486-C-0082, and by the Naval Research Laboratory under Grant N00014-87-K-2018.

REFERENCES

Batoz, J.L., Bathe, K.-J. and Ho, L.W., 1980, "A Study of Three-Node Triangular Plate Bending Elements," *Int. J. Num. Meth. Engrg.*, **15**, pp. 1771–1812 (1980)

Bergan, P. G. and Hanssen, L., 1976, "A new approach for deriving 'good' finite elements," MAFELAP II Conference, Brunel University, 1975, in *The Mathematics of Finite Elements and Applications - Volume II*, ed. by J. R. Whiteman, Academic Press, London

Bergan, P. G., 1980, "Finite Elements Based on Energy Orthogonal Functions," *Int. J. Num. Meth. Engrg.*, **15**, pp. 1141–1555

Bergan, P. G. and Nygård, M. K., 1984, "Finite elements with increased freedom in choosing shape functions," *Int. J. Num. Meth. Engrg.*, **20**, pp. 643–664

Bergan, P. G. and Wang X., 1984, "Quadrilateral plate bending elements with shear deformations," *Computer & Structures*, **19**, pp. 25–34

Bergan, P. G. *et. al.*, 1984, FENRIS Manuals, Theory - Program Outline - Data Input, NTH, SINTEF and A. S. Veritas, Hovik, Norway

Bergan, P. G. and Felippa, C. A., 1985, "A Triangular Membrane Element with Rotational Degrees of Freedom," *Computer Methods in Applied Mechanics & Engineering*, **50**, pp. 25–69

Bergan, P. G. and Nygård, M. K., 1985, "Nonlinear Shell Analysis Using Free Formulation Finite Elements," *Proc. Europe-US Symposium on Finite Element Methods for Nonlinear Problems*, held at Trondheim, Norway, August 1985, Springer-Verlag, Berlin

Bergan, P. G. and Nygård, M. K., 1988, "Free Formulation Elements Applied to Stability of Shells," *Proc. Int. Conf. on Computational Engineering Science — ICES88*, April 10-14, 1988, Atlanta, Georgia

Felippa, C. A. and Bergan, P. G., 1987, "A Triangular Plate Bending Element Based on an Energy-Orthogonal Free Formulation," *Computer Methods in Applied Mechanics & Engineering*, **61**, pp. 129–160

Felippa, C. A., 1988a, "Parametrized Multifield Variational Principles in Elasticity: I. Mixed Functionals," submitted to *Communications in Applied Numerical Methods*

Felippa, C. A., 1988b, "Parametrized Multifield Variational Principles in Elasticity: II. Hybrid Functionals and the Free Formulation," submitted to *Communications in Applied Numerical Methods*

Felippa, C. A., 1988c, "The Extended Free Formulation of Finite Elements in Linear Elasticity," submitted to *Journal of Applied Mechanics*

Hanssen, L., Syvertsen, T. G. and Bergan, P. G., 1979, "Stiffness Derivation Based on Element Convergence Requirements," MAFELAP III Conference, Brunel University, 1978, in *The Mathematics of Finite Elements and Applications - Vol III*, ed. by J. R. Whiteman, Academic Press, London

Nygård, M. K., 1986, *The Free Formulation for Nonlinear Finite Element Analysis with Applications to Shells*, Dr. Ing. Dissertation, Report 86-1, Division of Structural Mechanics, The Norwegian Institute of Technology, Trondheim, Norway

Nonlinear Elasto-Plastic Finite Element Analysis of the Thin Shell of Revolution

Isaac Fried

Boston University
Department of Mathematics
Boston, Mass. 02215

Abstract

Reversible plasticity is modeled with a stress strain law having a distinct yield point. Expressions are derived for the element tangent stiffness matrix and load vector of a largely deformed cubic-cubic shell element by discrete integration. Numerical experiments are carried out using the orthogonal trajectory technique to trace sequences of equilibrium configurations for high loads.

1. Introduction

Very thin shells share a numerical troublesome property with nearly incompressible elasticity [1,2] that their stiffness matrix seriously decline in condition as the thickness is reduced. Enforcement of the emerging condition of C^1 continuity (incompressibility in elastomers) creates an imbalanced elastic energy expression consisting of two out of proportion terms. The corresponding stiffness matrix consists in such cases of the linear combination of two matrices with widely separated coefficients, and consequently with an eigenvalue spectrum consisting of two groups largely shifted. The lowest eigenvalue of the global stiffness matrix is related to the fundamental frequency of the elastic system and is therefore only slightly affected by the large parameter, but the largest eigenvalue of the global stiffness matrix grows without bound causing the decline in conditioning.

Direct solution methods for the linear, or linearized, stiffness equation are operationally unaffected by ill condition but the convergence of the Newton-Raphson method, the fundamental solution procedure for all nonlinear finite element equilibrium problems, is measurably influenced by out of balance elastic energy expressions. Also, the mere storage of such ill conditioned algebraic systems is an immediate cause for round-off errors and loss of numerical significance.

Iterative methods for the solution of the linearized stiffness equation such as conjugate gradients are most attractive for the solution of the very large discrete systems set up by finite elements, but unlike direct methods they do show great operational sensitivity to the spectrum span of the global stiffness matrix and may very well lose all convergence properties in the presence of ill conditioning, rendering them useless.

A possible way out of the numerical instability of the imbalanced energy of nearly incompressible elasticity and thin shells is to use a multiparameter technique whereby a well conditioned system is set up for a shell of not excessive thickness and is comfortably and accurately solved to provide an initial guess for a next, less well conditioned system. Extrapolation to the limit over the disturbing parameter may be carried out to accomplish the computations to the needed accuracy. This technique is important but we shall not deal with it in the present paper. It deserves a separate discussion.

The issue of plasticity is more central to our present discussion. Irreversible plasticity is extremely expensive and computationally complicated in being dissipating and incremental. Such plastic formulation should be reserved in our opinion only to such elasto-plastic problems where a clear phenomenological merit is established for irreversibility. Otherwise, reversible plasticity, namely an analytic nonlinear constitutive law with a distinct yield point [3], and drastically reduced (even to zero) elastic modulus, should be amply adequate and sufficiently revealing, both physically and mathematically.

The paper is devoted to the incorporation of these three computational elements to create a computational procedure [4] for the large elasto-plastic deformation of thin shells of revolution: (1) The creation by discrete numerical integration of a cubic-cubic element stiffness matrix and load vector for a largely displaced and highly strained shell of revolution. (2) The incorporation into the shell element program of reversible plasticity. (3) The adap-

tation of the orthogonal trajectory technique to trace load displacement branches; and (4) The performance of numerical computation to validate the theory and show its practicality.

2. Load-displacement tracking

It is an integral part of any nonlinear finite element program. Newton-Raphson techniques are the most widely used solution procedures for the nonlinear stiffness equation. They are fast and usually reliable as long as the sought equilibrium configuration is far from being a critical, or turning point. To account for such singularities we need modify the Newton-Raphson iterations to include variation of both displacement and load.

To fix ideas we shall first present the orthogonal trajectory method for nonlinear continuation for the single, implicit equilibrium curve

$$r(x, \lambda) = 0 \tag{1}$$

in which x is displacement and λ load. To trace λ vs. x we need to compute close pairs x, λ that satisfy eq. (1).

Tracing the x, λ curve consists of the two distinct stages of *prediction* and *correction*. Predictor is the stage in which we move from a previously established equilibrium point A to a new guess at point B, usually a distance s on the tangent line at A as in Fig. 1(a).

If $A(x_0, \lambda_0)$ is an equilibrium point and $B(x_1, \lambda_1)$ is on the tangent so that $\overline{AB} = s$, then linearization of $r(x_0 + \delta_x, \lambda_0 + \delta_\lambda) = 0$, $\delta x = x_1 - x_0$, $\delta\lambda = \lambda_1 - \lambda_0$ produces

$$r_0 + r_0' \delta x + \dot{r}_0 \delta\lambda = 0, \quad \delta x^2 + \delta\lambda^2 = s^2 \tag{2}$$

where r_0' is dr/dx at A and $\dot{r}_0$ is $dr/d\lambda$ at A. Hence

$$\lambda_1 = \lambda_0 \pm \frac{s r'}{(\dot{r}^2 + r'^2)^{1/2}}, \; x_1 = x_0 \pm \frac{s \dot{r}}{(\dot{r}^2 + r'^2)^{1/2}} \tag{3}$$

where the choice of sign determines the direction, and where the subscript zero is omitted for typographical brevity.

In vector form we write the total potential energy as $\pi(x, \lambda)$ and have the system of stiffness equations as $r(x, \lambda) = \partial\pi/\partial x$, the gradient of r. Linearization is here of the form

$$\left[\frac{\partial r}{\partial x}\right](x - x_0) + (\lambda - \lambda_0)\frac{\partial r}{\partial \lambda} = 0 \tag{4}$$

where $K = \partial r/\partial x$ is the global stiffness *matrix*, and where $p = \partial r/\partial\lambda$ is the load vector. Equation (4) is a vector equation. With the constraint on the traveled distance

$$(x_1 - x_0)^T (x_1 - x_0) + (\lambda - \lambda_0)^2 = s^2 \tag{5}$$

linearization leads to

$$\lambda_1 = \lambda_0 + s(1 + p_0^T K_0^{-1} K_0^{-1} p_0)^{-\frac{1}{2}}, \; x_1 = x_0 - \delta\lambda_0 K_0^{-1} p_0 \tag{6}$$

that completes the prediction.

Starting with the predicted initial guess (x_0, λ_0) we set out to approach the equilibrium curve. This can be done with a Newton-Raphson iterative method in which: (1) The load λ is constant as in Fig. 1(b). The failing of such iterative scheme near a limit point is clear. (2) The load λ is linearly related to the displacement x, as in Fig. 1(c). In case the linear constraint misses the equilibrium curve convergence is not achieved by this technique neither. (3) The load λ is related to the displacement x by the condition that the the iterated points lie on a circle of radius s having its center at the previous equilibrium point, as in Fig. 1(d). (4) The load λ and displacement x are constrained to be an orthogonal trajectory to the equilibrium curve, as in Fig. 2.

In the orthogonal trajectory accession technique [5] the ultimate correction step is orthogonal to the equilibrium curve. Load and displacement enter symmetrically in this algorithm which is therefore indifferent to critical points and turning points.

Analytically we add to the linearized equation $r + r'\delta x + \dot{r}\ \delta\lambda = 0$ the orthogonality condition

$$d\lambda = \dot{r}(r')^{-1}dx \tag{7}$$

and obtain the corrections

$$\delta\lambda = -\frac{r\dot{r}}{r'^2 + \dot{r}^2},\ \delta x = -\frac{rr'}{r'^2 + \dot{r}^2} \tag{8}$$

in which the right hand sides include computed values only, and where $\delta\lambda = \lambda_1 - \lambda_0$ and $\delta x = x_1 - x_0$.

In the multidimensional case x stands for the displacement *vector* so that $r(x, \lambda) = 0$ is a set of nonlinear stiffness equations. Here $r' = K$ is the global tangent, displacement dependent, global stiffness matrix, and $\dot{r} = p$ is the global nonlinear load vector. Linearization is here of the form $r + K\delta x + p\delta\lambda = 0$, or $r + Kdx + pd\lambda = 0$, since we are dealing with *differentials* rather than *differences*, and the orthogonality conditions assume the matrix vector form

$$\begin{aligned} d\lambda &= \dot{r}(r')^{-1}dx = p^T K^{-1}dx \\ dx &= -K^{-1}(r + d\lambda p) \end{aligned} \tag{9}$$

producing finally the corrections

$$d\lambda = \frac{p^T K^{-1}K^{-1}r}{1 + p^T K^{-1}K^{-1}p},\ dx = -K^{-1}(r + d\lambda p) \tag{10}$$

Figures 3 shows the orthogonal accession algorithm applied to the equilibrium equation $r = 8x(1 - x) - \lambda = 0$ with a variety of starting points of the form $(0, \lambda)$. Never does the algorithm fail, and it shows a penchant to hit the limit point of the curve. The limit point is a special attractor for the method. Notice that if the initial guess lies on a *normal* to the equilibrium point, then convergence is in *one* step. Otherwise, it is not easy to find an initial guess requiring more than six steps to convergence.

Nonlinear elastic equilibrium problems are replete with critical points and a modified Newton-Raphson method must be used on these problems to successfully pass limit and turning points. All the present computations were done with this algorithm.

3. Reversible plasticity

All we need to model reversible plasticity is a stress-strain law that exhibits a bilinear behavior with a clearly defined yield point. It should also include a parameter to adjust the slope in the plastic range. We suggest to relate the stress σ to the strain ϵ by

$$\sigma = \alpha\epsilon\left[\frac{1+(\epsilon/\beta)^2}{1+(\epsilon/\beta)^4}\right] \tag{11}$$

in which α, β and p are parameters. To obtain the modulus of elasticity for the law in eq. (11) we differentiate σ with respect to ϵ, $d\sigma/d\epsilon = E$. Figures 4 and 5 show $\sigma = \sigma(\epsilon)$ and $\sigma' = \sigma'(\epsilon)$ as a function of the parameter p. As p increases law (11) approaches describing perfect plasticity. Notice in Fig. 5 the existence of an accurate yield point at $\sigma/\beta = 1$.

4. Cubic-Cubic shell of revolution element

As for the plastica, also here [6,7] we may write the stiffness matrix as for the elastic shell, except that we have to bear in mind that the elastic modulus E is strain dependent.

Let $r = r(\eta)$ and $z = z(\eta)$ be the parametric equations of the generating curve for the *deformed* sell. Referred to the Cartesian coordinate system oxyz

$$x = r(\eta)\cos\theta,\ y = r(\eta)\sin\theta,\ z = z(\eta) \tag{12}$$

where, obviously, $x^2 + y^2 = r^2$. It is helpful to introduce the angle ϕ measured between the positive r-axis and the tangent to the generating middle curve. The position vector p and unit normal vector n to a point on the middle *surface* becomes with ϕ

$$p = (r\cos\theta, r\sin\theta, z)^T, n = (\sin\phi\cos\theta, \sin\phi\sin\theta, -\cos\phi)^T \tag{13}$$

In the same way the position vector to a material point on n at a distance ζ from the middle surface is

$$q = r + \zeta n,\ dq = dr + d\zeta n + \zeta dn \tag{14}$$

since dr is on the tangent plane, and $n^T n = 1$, we have that $n^T dr = 0$, $n^T dn = 0$, and an arc element $ds^2 = dq^T dq$ becomes

$$ds^2 = dr^T dr + \zeta^2 dn^T dn + d\zeta^2 + 2\zeta dr^T dn \tag{15}$$

in which

$$dr = \frac{\partial r}{\partial \eta}d\eta + \frac{\partial r}{\partial \theta}d\theta,\ dn = \frac{\partial n}{\partial \phi}d\phi + \frac{\partial n}{\partial \theta}d\theta \tag{16}$$

Because the η =const. and θ =const. curves are orthogonal

$$(\frac{\partial r}{\partial \eta})^T(\frac{\partial r}{\partial \theta}) = 0,\ (\frac{\partial n}{\partial \phi})^T(\frac{\partial n}{\partial \theta}) = 0$$

(17)

$$(\frac{\partial r}{\partial \eta})^T(\frac{\partial n}{\partial \theta}) = 0,\ (\frac{\partial n}{\partial \phi})^T(\frac{\partial r}{\partial \theta}) = 0$$

and we are left with

$$\begin{aligned} dr^T dr &= \alpha^2 d\eta^2 + r^2 d\theta^2 \\ dn^T dn &= \phi'^2 d\eta^2 + \sin^2\phi d\theta^2 \\ dr^T dn &= \alpha\phi' d\eta^2 + r\sin\phi d\theta^2 \end{aligned} \tag{18}$$

where prime denotes differentiation with respect to η, and where

$$\sin\phi = \frac{z'}{\alpha},\ \cos\phi = \frac{r'}{\alpha}$$

$$\phi' = \frac{z''r' - z'r''}{\alpha^2},\ \alpha = (r'^2 + z'^2)^{1/2} \tag{19}$$

Finally

$$ds^2 = (\alpha + \zeta\phi')^2 d\eta^2 + (r + \zeta\sin\phi)^2 d\theta^2 + d\zeta^2 \tag{20}$$

written for the undeformed shell as

$$ds_0^2 = (\alpha_0 + \zeta\phi_0^2)d\eta^2 + (r_0 + \zeta\sin\phi_0)^2 + d\zeta^2 \tag{21}$$

under the simplification $\zeta = \zeta_0$.

Strain is obtained from the ration of the two quadratic forms ds^2 and ds_0^2 as

$$\epsilon_1(\zeta) = \frac{(\alpha - \alpha_0) + \zeta(\phi' - \phi_0')}{\alpha_0 + \zeta\phi_0'}, \epsilon_2(\zeta) = \frac{(r - r_0) + \zeta(\sin\phi - \sin\phi_0)}{r_0 + \zeta\sin\phi_0} \tag{22}$$

Integration of the elastic energy with respect to ζ yields, after some obvious simplifications

$$E = \pi E(\epsilon,\nu)\epsilon[t(\epsilon_1^2 + 2\nu\epsilon_1\epsilon_2 + \epsilon_2^2) + \frac{1}{12} + 3(\kappa_1 + 2\nu\kappa_1\kappa_2 + \kappa_2^2)\alpha_0 r_0 d\eta \tag{23}$$

in which ν is the Poisson ratio, that is left independent of the strain, and where t is the shell thickness. Also

$$\begin{aligned} \epsilon_1 &= \frac{\alpha}{\alpha_0} - 1,\ \epsilon_2 = \frac{r}{r_0} - 1 \\ \kappa_1 &= \frac{\phi' - \phi_0'}{\alpha_0},\ \kappa_2 = \frac{\sin\phi - \sin\phi_0}{r_0} \end{aligned} \tag{24}$$

Equations (11) and (23) form the basis for the derivation of the element data for the shell. But before getting on with the matrix and vector derivation of the sell element we need briefly consider some conventions for differentiation with respect to a vector.

Let $f(x) = f(x_1, x_2, \ldots, x_n)$ be a scalar function of the vector argument x. We define the differentiation of f with respect to the vector x as

$$\frac{\partial f}{\partial x} = f' = \frac{\partial f}{\partial x_i}, \ \frac{\partial^2 f}{\partial x^2} = f'' = \frac{\partial^2 f}{\partial x_i, \partial x_i} \tag{25}$$

Notice that f' is a vector and f'' is a symmetric matrix.

Obviously,

$$(f+g)' = f' + g', \ (cf)' = cf', \ (fg)' = gf' + fg', \tag{26}$$

but

$$(gf')' = gf'' + f'g'^T,$$

where

$$f'g'^T = \frac{\partial f}{\partial x_j}\frac{\partial g}{\partial x_j} \tag{27}$$

is a nonsymmetric matrix. The matrix

$$(fg)'' = gf'' + fg'' + f'g'^T + g'f'^T \tag{28}$$

is symmetric. Now we have all that is needed for the discrete integration of the total potential energy and the formation of the element data.

Recall that the parametric equation of the generating curve for the shell is $r = r(\eta)$, $z = z(\eta)$. Let ζ measure arc length along the generator so that $\alpha_0 = 1$ and $d\eta = ds$. The finite element extends between s_1 and $s_1 + h$ so that we may write $s = s_1 + h\sigma$, $0 \le \sigma \le 1$, and $ds = h\, d\sigma$. If prime denotes differentiation with respect to s and dot differentiation with respect to σ then $(\ )' = h^{-1}(\ )^{\cdot}$ and $(\ )'' = h^{-2}(\ )^{\cdot\cdot}$.

To have a cubic-cubic C^1 element we choose the nodal values vector

$$u_e = (r_1, \dot{r}_1, z_1, \dot{z}_1, r_2, \dot{r}_2, z_2, \dot{z}_2)^T \tag{29}$$

and interpolate r and z with

$$r = u_e^T \phi, \ z = u_e^T \psi, \tag{30}$$

where ϕ and ψ are the shape function vectors

$$\phi = (\phi_1, \phi_2, 0, 0, \phi_3, \phi_4, 0, 0)^T, \ \psi = (0, 0, \phi_1, \phi_2, 0, 0, \phi_3, \phi_4)^T,$$

with

$$\phi_1 = 1 - 3\xi^2 + 2\xi^3, \ \phi_2 = \xi - 2\xi^2 + \xi^3, \ \phi_3 = 3\xi^2 - 2\xi^3, \ \phi_4 = -\xi^2 + \xi^3. \tag{31}$$

We integrate the total potential energy by sampling it at the three Gauss points

$$\xi_1 = \frac{1}{10}(5 - \sqrt{15}), \ \xi_2 = \frac{1}{2}, \ \xi_3 = \frac{1}{10}(5 + \sqrt{15}), \tag{32}$$

and weights

$$w_1 = w_3 = \frac{5}{18},\ w_2 = \frac{8}{18}. \tag{33}$$

where ϕ_j and ψ_j shortly stand for $\phi(\xi_j)$ and $\psi(\xi_j)$ and

$$\begin{aligned}
\phi_{1,3} &= \frac{1}{100}(50 \pm 12\sqrt{15}, 5 \pm \sqrt{15}, 0, 0, 50 \mp 12\sqrt{15}, -5 \pm \sqrt{15}, 0, 0)^T, \\
\phi_2 &= \frac{1}{8}(4, 1, 0, 0, 4, -1, 0, 0)^T, \\
\dot{\phi}_{1,3} &= \frac{1}{10}(-6, 2 \pm \sqrt{15}, 0, 0, 6, 2 \mp \sqrt{15}, 0, 0)^T, \\
\dot{\phi}_2 &= \frac{1}{4}(-6, -1, 0, 0, 6, -1, 0, 0), \\
\ddot{\phi}_{1,3} &= \frac{1}{5}(\mp 6\sqrt{15}, -5 \mp 3\sqrt{15}, 0, 0, \mp 6\sqrt{15}, 5 \mp 3\sqrt{15}, 0, 0)^T, \\
\ddot{\phi}_2 &= (0, -1, 0, 0, 0, 1, 0, 0)^T, \\
\psi_{1,3} &= \frac{1}{100}(0, 0, 50 \pm 12\sqrt{15}, 5 \pm \sqrt{15}, 0, 0, 50 \mp 12\sqrt{15}, -5 \pm \sqrt{15})^T, \\
\psi_2 &= \frac{1}{8}(0, 0, 4, 1, 0, 0, 4, -1)^T, \\
\dot{\psi}_{1,3} &= \frac{1}{10}(0, 0, -6, 2 \pm \sqrt{15}, 0, 0, 6, 2 \mp \sqrt{15})^T, \\
\dot{\psi}_2 &= \frac{1}{4}(0, 0, -6, -1, -, -, 6, -1)^T, \\
\ddot{\psi}_{1,3} &= \frac{1}{5}(0, 0, \mp 6\sqrt{15}, 5 \mp 3\sqrt{15}, 0, 0, \pm 6\sqrt{15}, 5 \mp 3\sqrt{15})^T, \\
\ddot{\psi}_2 &= (0, 0, 0, -1, 0, 0, 0, 1).
\end{aligned}$$

The upper sign of $\sqrt{15}$ is for Gauss point 1 and the lower for Gauss point 3.

We derive the element data from the energy expression and write for the eth element

$$E_e^* = \frac{1}{2}h \sum_{j=1}^{3} w_j r_{0j}[c(\epsilon_{1j}^2 + 2\nu\epsilon_{1j}\epsilon_{2j} + \epsilon_{2j}^2) + (\kappa_{1j}^2 + 2\nu\kappa_{1j}\kappa_{2j} + \kappa_{2j}^2)], \tag{35}$$

where j refers to the jth Gauss point, and

$$c = \frac{12}{t^2},\ E^* = \frac{6}{\pi E t^3}E(\epsilon). \tag{36}$$

From the definitions for the element gradient and matrix

$$g_e = \frac{\partial E_e^*}{\partial u_e},\ k_e = \frac{\partial^2 E_e}{\partial u_e^2}, \tag{37}$$

we have that

$$g_e = h \sum_{j=1}^{3} w_j r_{0j}\{c[\epsilon_{1j}\epsilon'_{ij} + \nu(\epsilon_{1j}\epsilon'_{2j} + \epsilon'_{ij}\epsilon_{2j}) + \epsilon_{2j}\epsilon'_{2j}] + \kappa_{1j}\kappa'_{ij} + \nu(\kappa_{1j}\kappa'_{2j} + \kappa_{2j}\kappa'_{1j}) + \kappa_{2j}\kappa'_{2j}\} \tag{38}$$

and

$$k_e = h\sum_{j=1}^{3} w_j r_{0j}\{c[\epsilon'_{ij}\epsilon'^{T}_{1j} + \epsilon_{1j}\epsilon''_{1j} + \nu(\epsilon'_{1j}\epsilon'^{T}_{2j} + \epsilon_{1j}\epsilon''_{2j} + \epsilon'_{2j}\epsilon'^{T}_{ij} + \epsilon_{2j}\epsilon''_{ij})] + \kappa'_{1j}\kappa'^{T}_{1j} + \kappa_{1j}\kappa''_{1j} + \kappa'_{2j}\kappa'^{T}_{2j} + \kappa_{2j}\kappa''_{2j} + \nu(\kappa'_{1j}\kappa'^{T}_{2j} + \kappa_{1j}\kappa''_{2j} + \kappa_{2j}\kappa''_{1j} + \kappa'_{2j}\kappa'^{T}_{1j})\}, \tag{39}$$

where $(\ )' = \partial/\partial u_e$.

To shorten the notation we introduce

$$f = \ddot{z}\dot{r} = \dot{z}\ddot{r},\ g = \dot{r}^2 + \dot{z}^2, \tag{40}$$

so that

$$\epsilon_1 = h^{-1}g^{1/2} - 1,\ \kappa_1 = h^{-1}g^{-1}f \tag{41}$$
$$\epsilon_2 = r_0^{-1}r - 1,\ \kappa_2 = r_0^{-1}g^{-1/2}\dot{z},$$

and

$$f' = \dot{r}\vec{\psi} + \dot{\phi}\ddot{z} - \ddot{r}\dot{\phi} - \dot{z}\ddot{\phi},\ g' = 2(\dot{r}\dot{\phi} + \dot{z}\dot{\psi}),$$
$$f'' = \ddot{\psi}\dot{\phi}^T + \dot{\phi}\ddot{\psi}^T - \dot{\psi}\ddot{\phi}^T - \ddot{\phi}\dot{\psi}^T,\ g'' = 2(\dot{\phi}\dot{\phi}^T + \dot{\psi}\dot{\psi}^T). \tag{42}$$

Next we write

$$\epsilon'_1 = \frac{1}{2}h^{-1}g^{-1/2}g',\ \kappa'_1 = h^{-1}(g^{-1}f' - g^{-2}fg'),$$
$$\epsilon'_2 = r_0^{-1}\phi,\ \kappa'_2 = r_0^{-1}(g^{-1/2}\dot{\psi} - \frac{1}{2}g^{-3/2}\dot{z}g'), \tag{43}$$

and

$$\epsilon''_1 = \frac{1}{2}h^{-1}(g^{-1/2}g'' - \frac{1}{2}g^{-3/2}g'g'^{T}),\ \epsilon''_2 = 0,$$

$$\kappa''_1 = h^{-1}[2g^{-3}fg'g'^{T} + g^{-1}f'' - g^{-2}fg'' - g^{-2}(f'g'^{T} + g'f'^{T})],$$
$$\kappa''_2 = r_0^{-1}[\frac{3}{4}g^{-3/2}\dot{z}g'g'^{T} - \frac{1}{2}g^{-3/2}(g'\dot{\psi}^T + \dot{\psi}g'^{T}) - \frac{1}{2}g^{-3/2}\dot{z}g''], \tag{44}$$

which is all we need to program the element gradient and stiffness matrix.

To account for pressure p and point force F we add their potential π^*,

$$\pi^* = p^*\frac{1}{t}\int r^2 z' ds - F^* z \tag{45}$$

where

$$p^* = \frac{6}{Et^2}p,\ F^* = \frac{6}{\pi Et^3}F \tag{46}$$

and have that

$$g_e = p^*\frac{1}{t}\sum_{j=1}^{3} w_j(2r_j\dot{z}_j\phi_j + r_j^2\dot{\psi}_j) \tag{47}$$

and

$$k_e = \frac{2p^*}{t} \sum_{j=1}^{3} w_j [\dot{z}\phi_j \phi_j^T + r_j(\phi_j \dot{\psi}_j^T + \dot{\psi}_j \phi_j^T)], \tag{48}$$

5. Numerical computations

An extensive number of numerical tests were carried out to test the working of the element formulation, material modeling, and tracking routine. Some representative examples will be discussed here. Apart from the most obvious conclusion that our program worked correctly under the most adverse numerical circumstances, we also observe that with the thin shell of revolution plasticity is not as interesting as elasticity. The clearest manifestation of yield happened to be the creation of a plastic hinge at a latitude of excessive bending. Most examples shown here are, therefore, of nearly pure elastic nature.

The nonlinear behavior of the highly deformed thin spherical cap is highly complex, and its finite element computation should be considered a significant numerical feat. In the following examples s denotes the step size in the tracking predictor, t the shell thickness (its radius being 1), Ne the number of elements, and θ_0 the angle between the r axis and the tangent to the generator at $z = 0$ ($\theta_0 = 90^o$ means a complete sphere while $\theta_0 = 180^o$ means a flat plate.) All discretizations are done with $Ne = 7$.

Figure 6 shows the inversion of a spherical cap ($t = 0.002, \theta_0 = 3\pi/8$) by an apex force for a step size of $s = 0.5$. At first the shell exhibits considerable stiffness observed by the close equilibrium configuration, but a point is reached at which the dent becomes nonstable and snaps through to the inverted form through a wavy pattern. It is remarkable that orthogonal tracking handled this transition smoothly.

Figure 7 refers to the same cap, except for a different edge condition. We observe that the fixed rim condition has a considerable stiffening effect on the shell near the transition point.

Figure 8 represents a thinner, and hence more flexible, shell. Transition occurs here earlier and the relatively large $s = 0.5$ caused the program to jump between far apart equilibrium states.

A larger step size causes earlier inversion as seen in Fig. 9. We repeat showing these examples (Fig. 10) to demonstrate the robustness of the algorithm that takes a distant initial guess for the equilibrium configuration to successful convergence.

The interest of Fig. 11 lies in the fact that the tracking algorithm landed on an equilibrium configuration that is *mathematically correct but physically impossible.* There is no provision in the algorithm to tell it that the shell may not loop. In any event, the shell became so stiff by the loop that the algorithm could not get out of it and the program was aborted.

To discern in Fig. 12 which of the equilibrium configurations is fictitious and which is not requires some discrimination and more numerical evidence. What is interesting is that the program *actually inverted the cap.* Loading started at the lower half and ends in the upper

half of the shell. Some of the bending modes appear to be accompanied by considerable stretching but in the presence of such large displacements and high strains it might be well possible that the shell prefers to stretch than to bend.

References

[1] I. Fried and A.R. Johnson, Nonlinear computation of axisymmetric solid rubber deformation *CMAME* **67** (1988), 241-253.

[2] I. Fried, and A.R. Johnson, Condition of the finite element stiffness matrix of highly irregular triangular grids. Fifth Army Conference on Applied Mathematics and Computing, West Point, N.Y. June 1987.

[3] I. Fried, Large-deflection computation of the plastica. *CMAME* **49** (1985), 163-173.

[4] I. Fried, A.R. Johnson and A. Tessler, Large Plasto-plastic deformation of shells of revolution. Sixth Army Conference on Applied Mathematics and Computing, Boulder, Colorado, 31 May-3 June 1988.

[5] I. Fried, Orthogonal trajectory accession to the non linear equilibrium curve, *CMAME* **47** (1984), 283-297.

[6] I. Fried, Nonlinear finite element analysis of the thin shell of revolution, *CMAME* **48** (1985), 283-299.

[7] I. Fried, Stability and equilibrium of the straight and curved elastica-finite element computation, *CMAME* **28** (1981), 49-61.

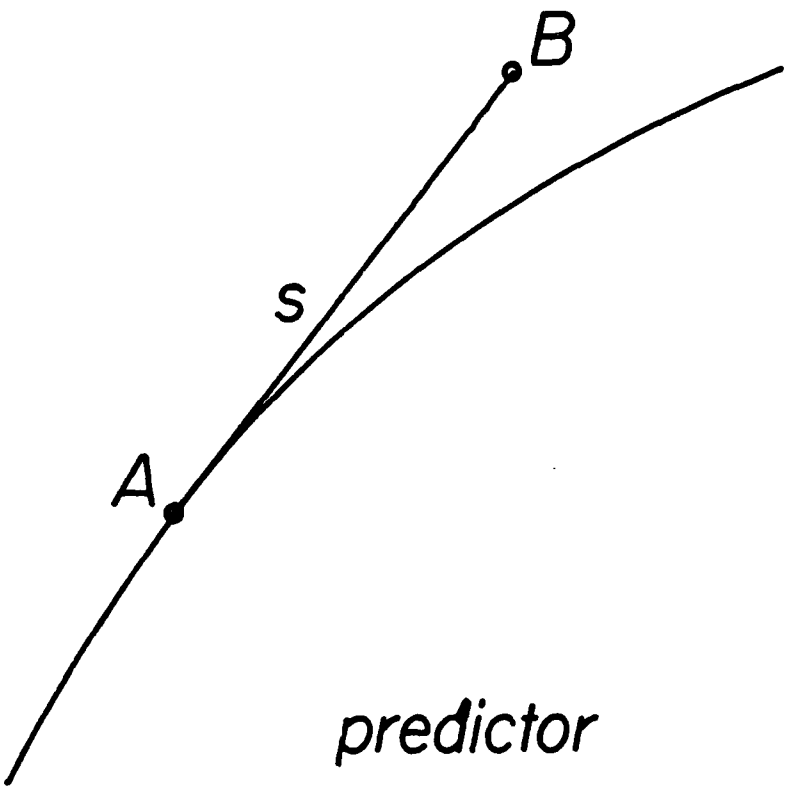


Figure 1.a Tangent line predictor.

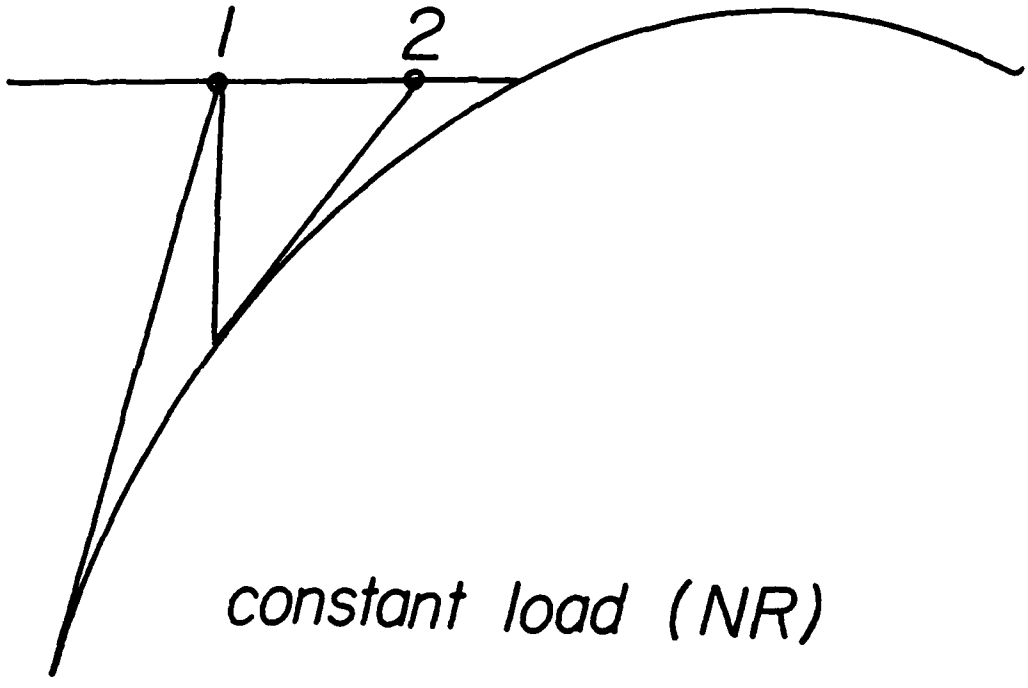


Figure 1.b Constant load corrector.

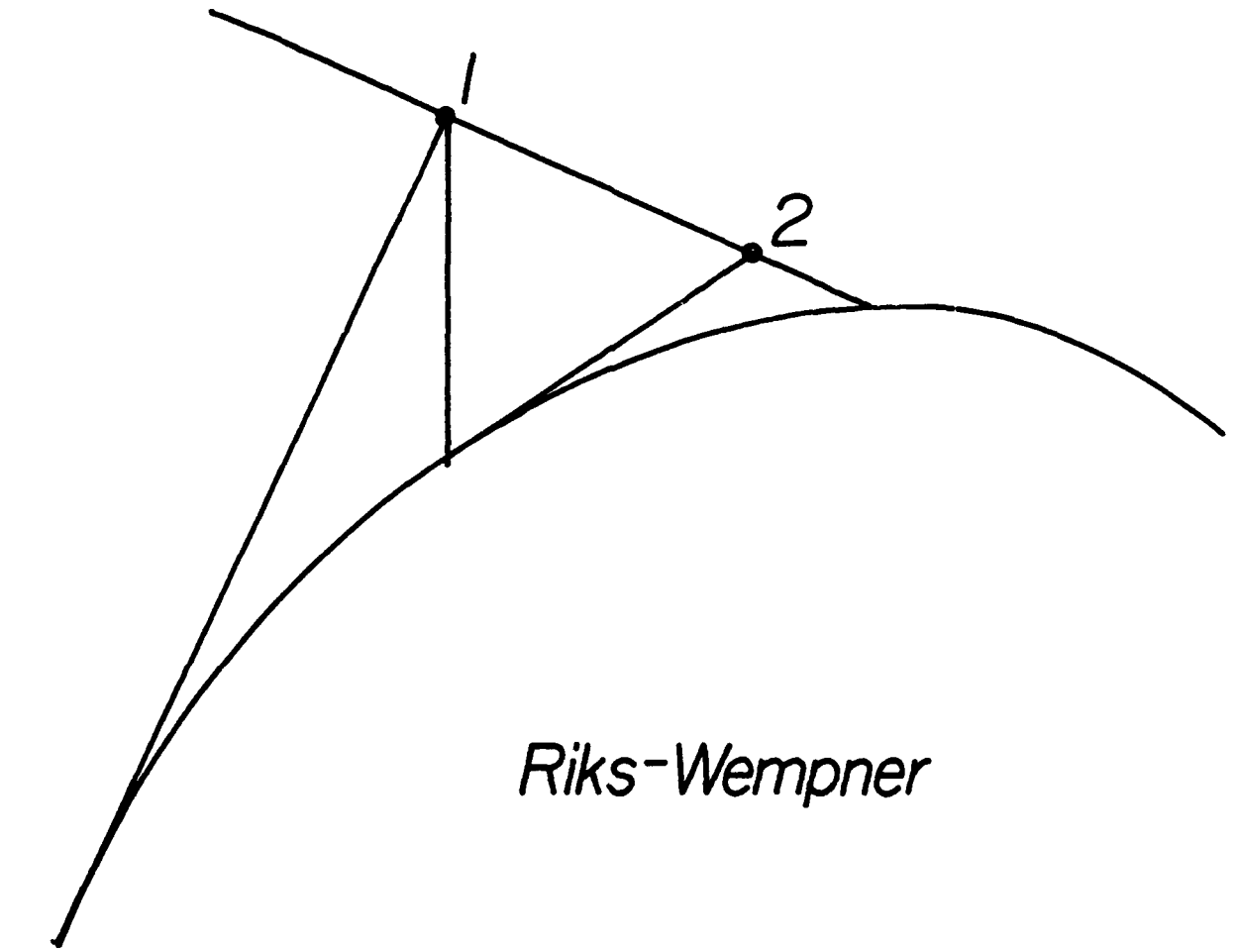


Figure 1.c Linear corrector.

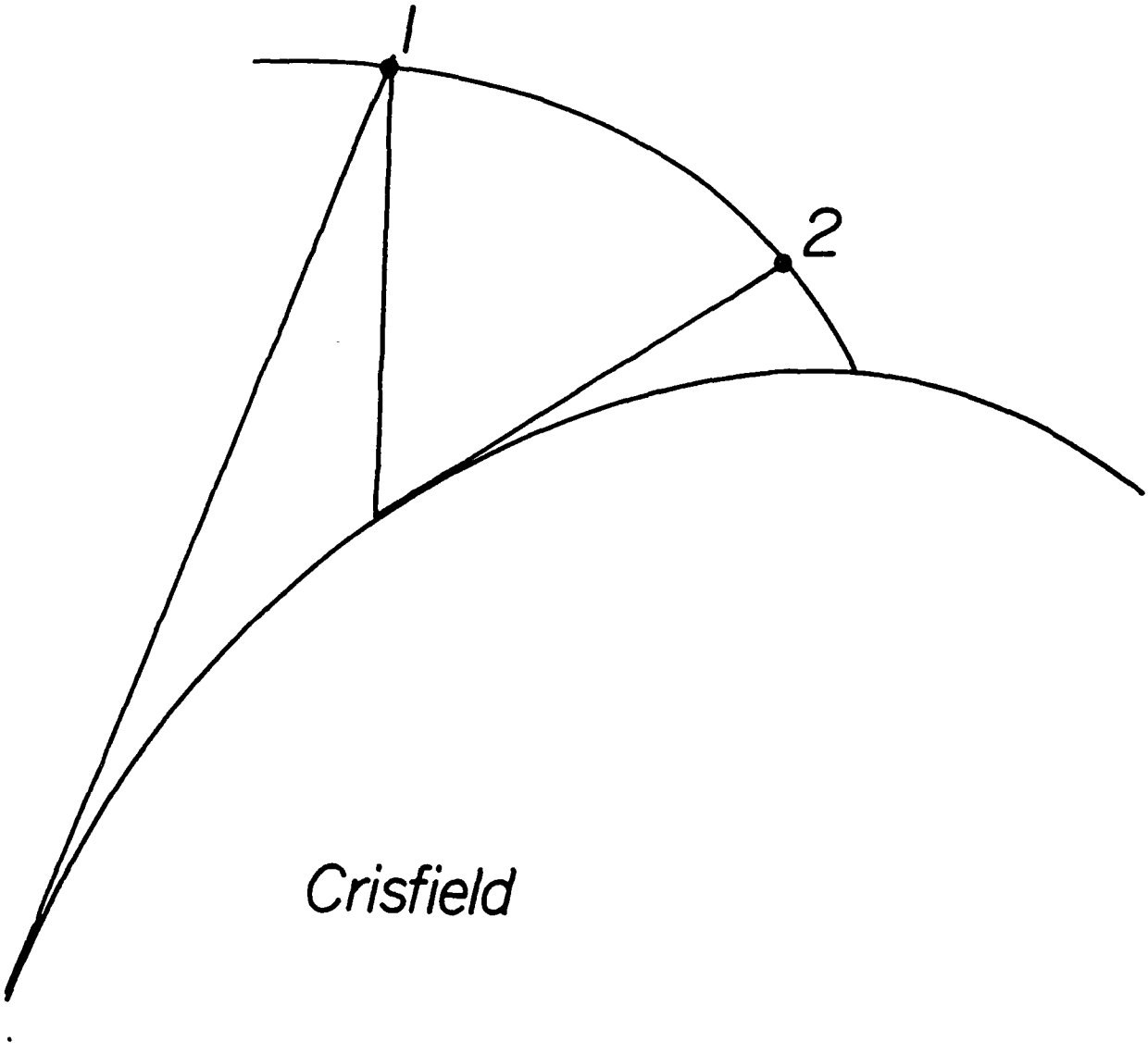


Figure 1.d Circular corrector.

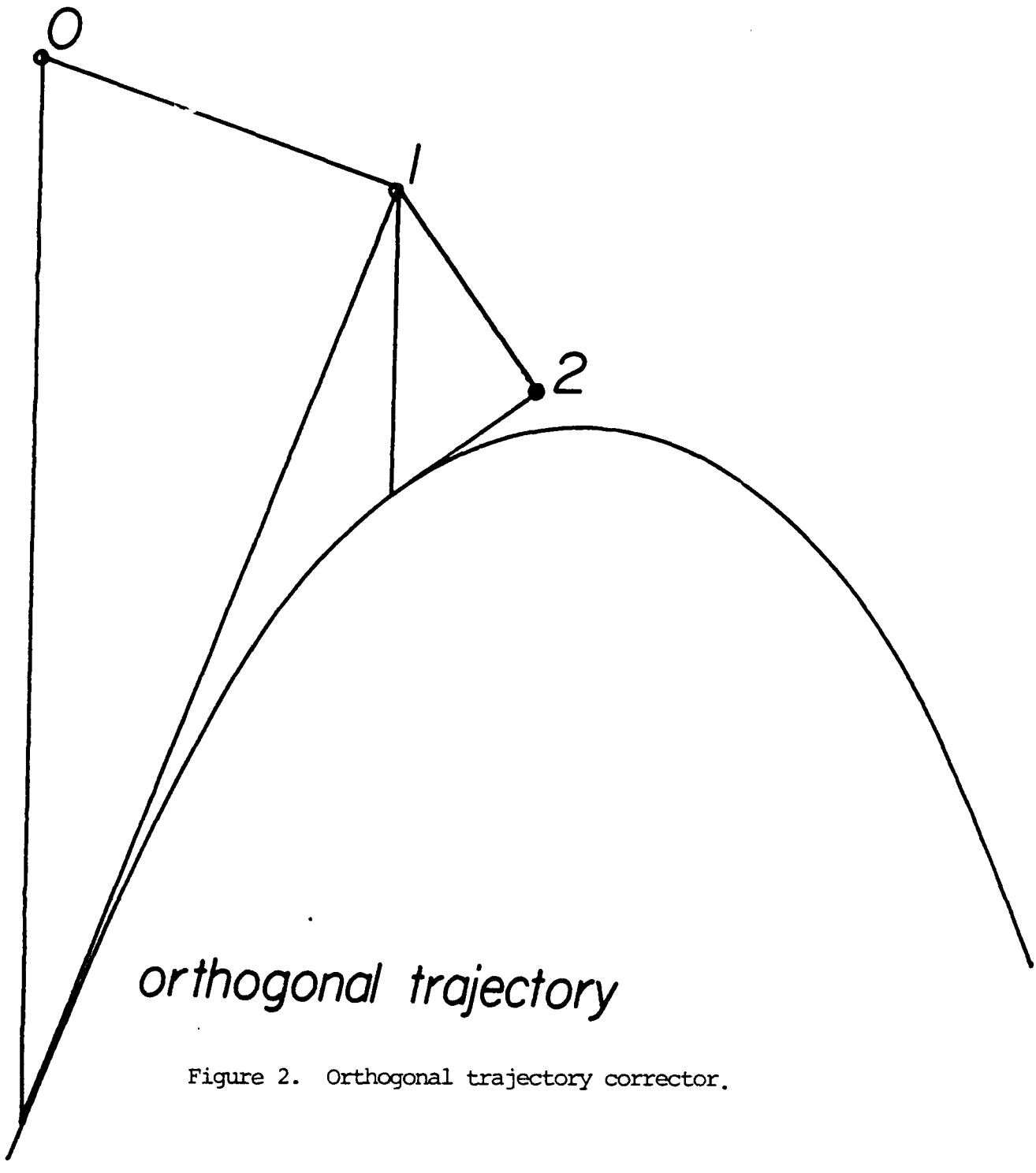


Figure 2. Orthogonal trajectory corrector.

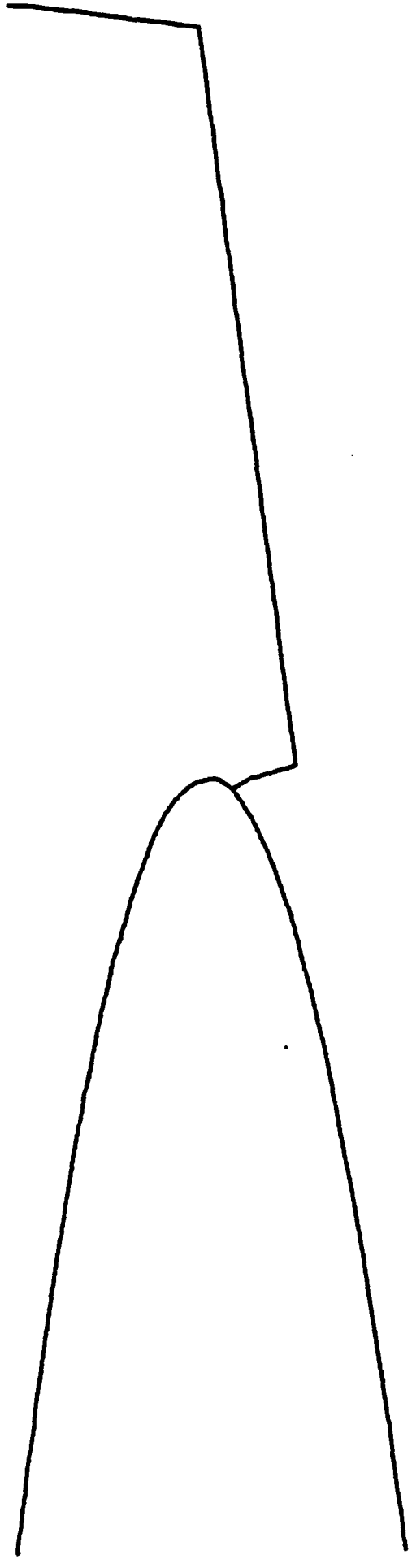


Figure 3.a Orthogonal accession.

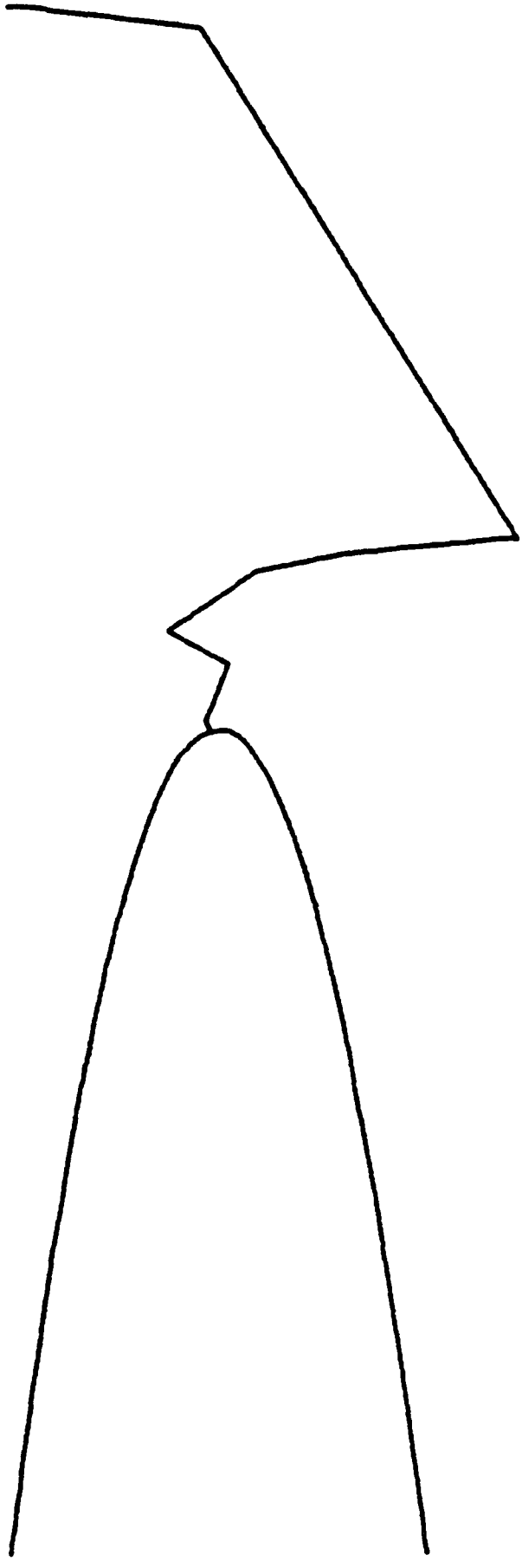


Figure 3.b Orthogonal accession.

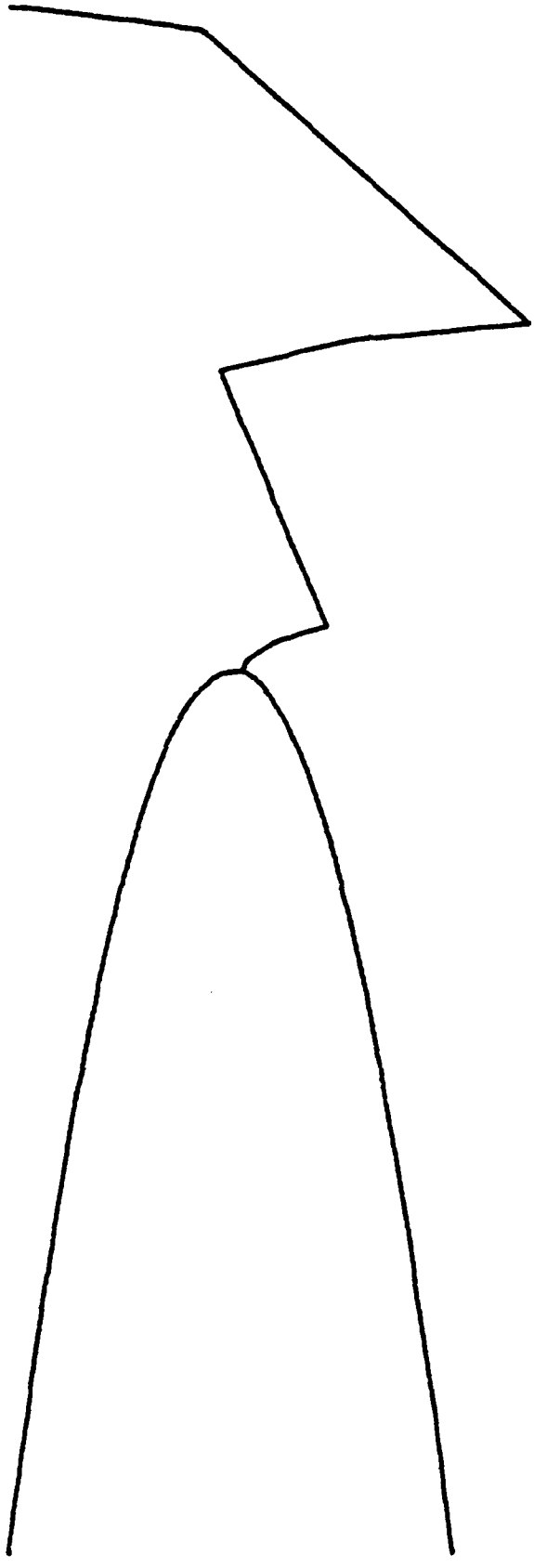


Figure 3.c Orthogonal accession.

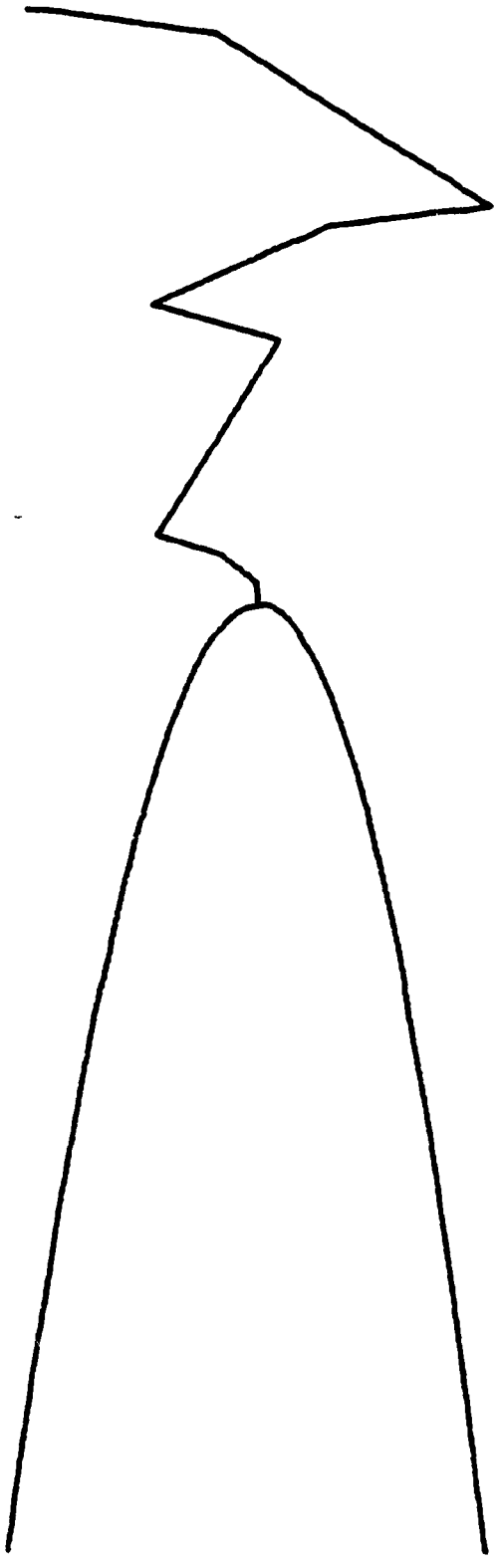


Figure 3.d Orthogonal accession.

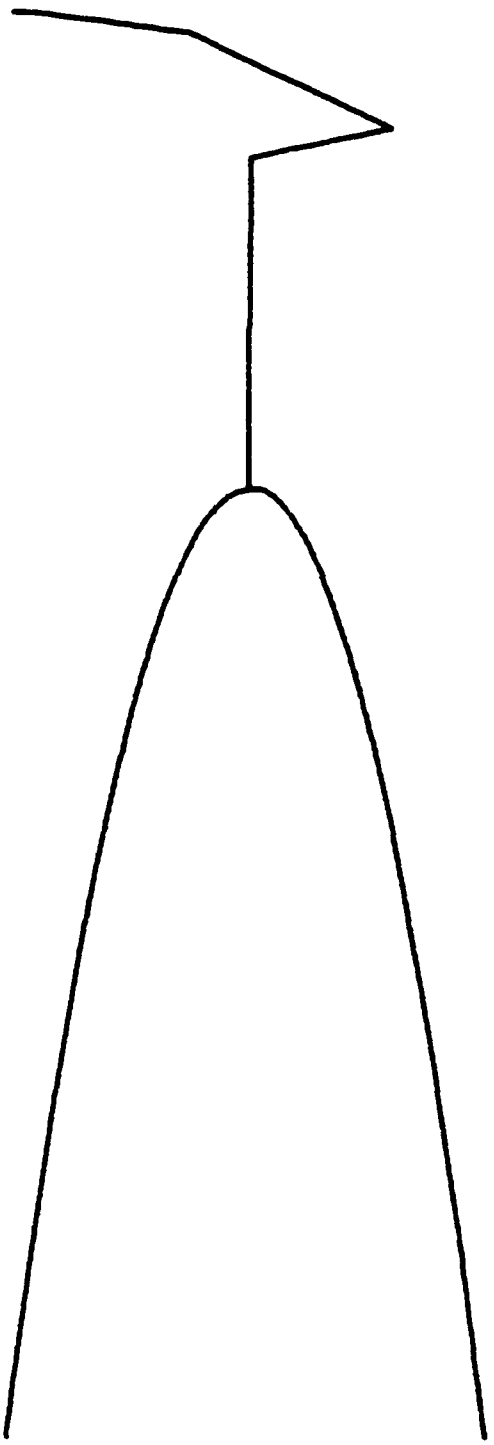


Figure 3.e Orthogonal accession.

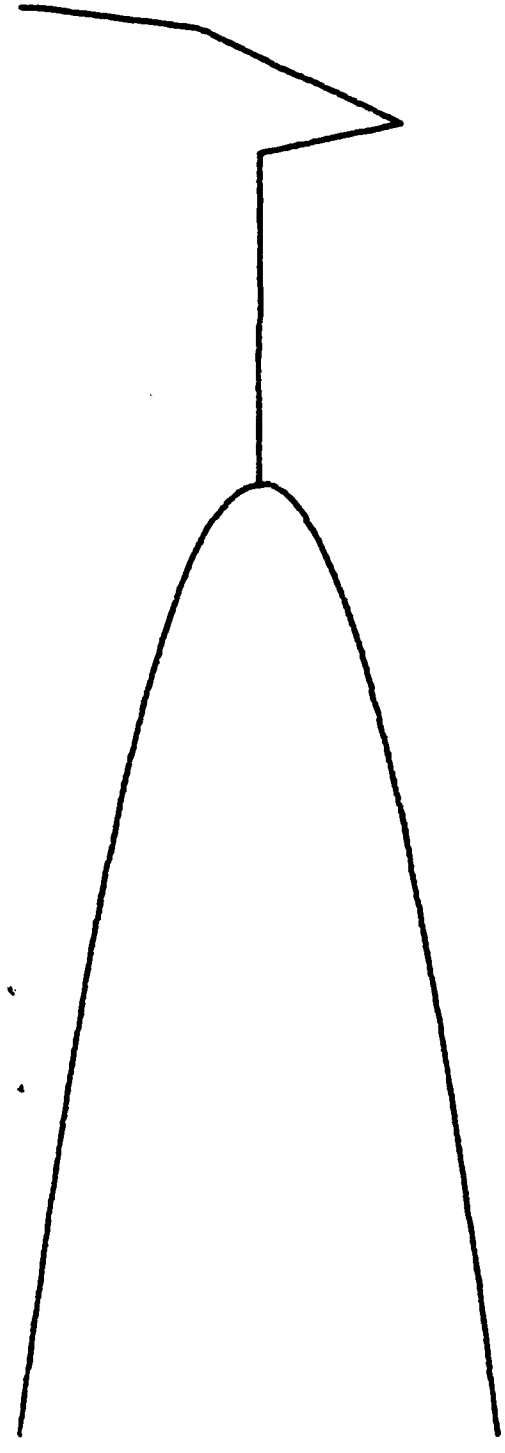


Figure 3.f Orthogonal accession.

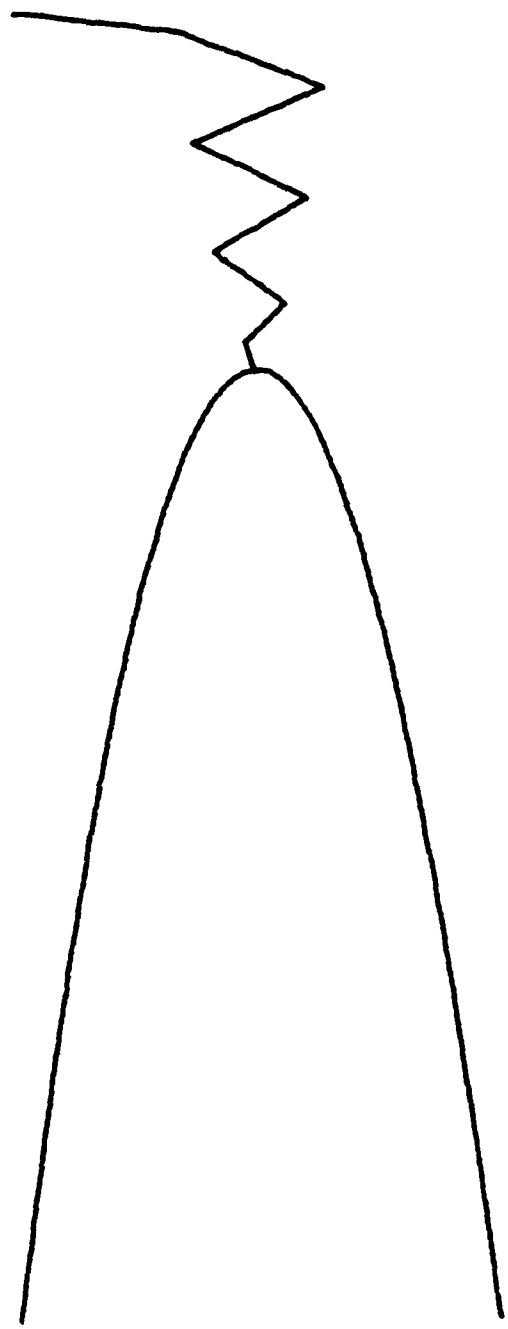


Figure 3.g Orthogonal accession.

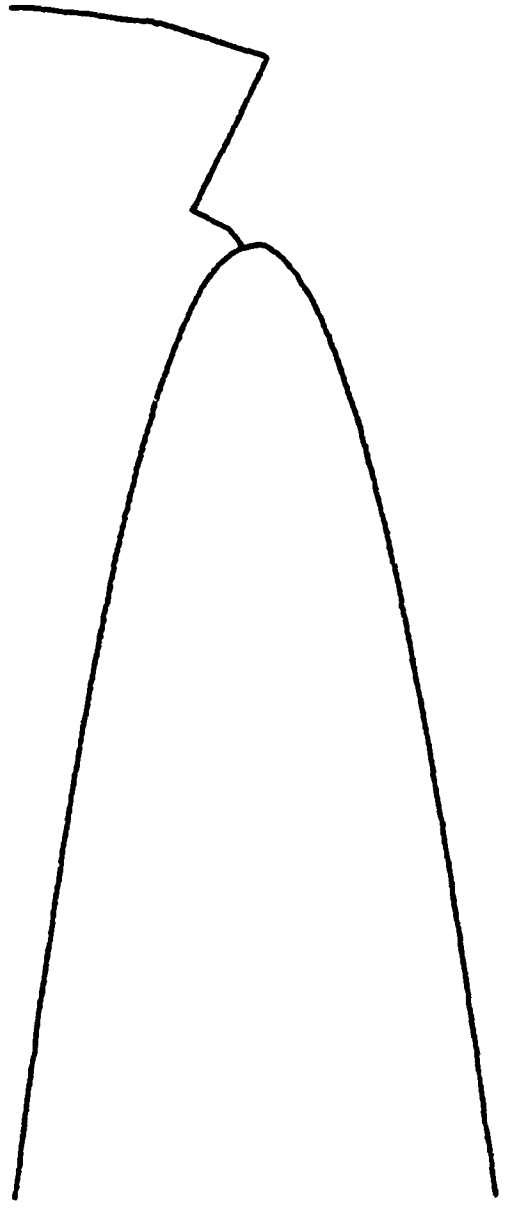
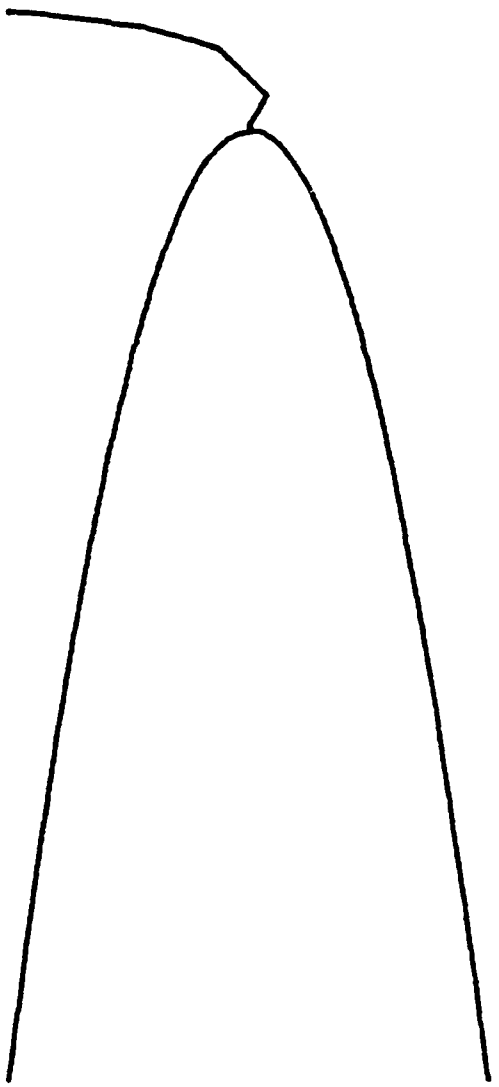


Figure 3.h Orthogonal accession.



3.i Orthogonal accession.

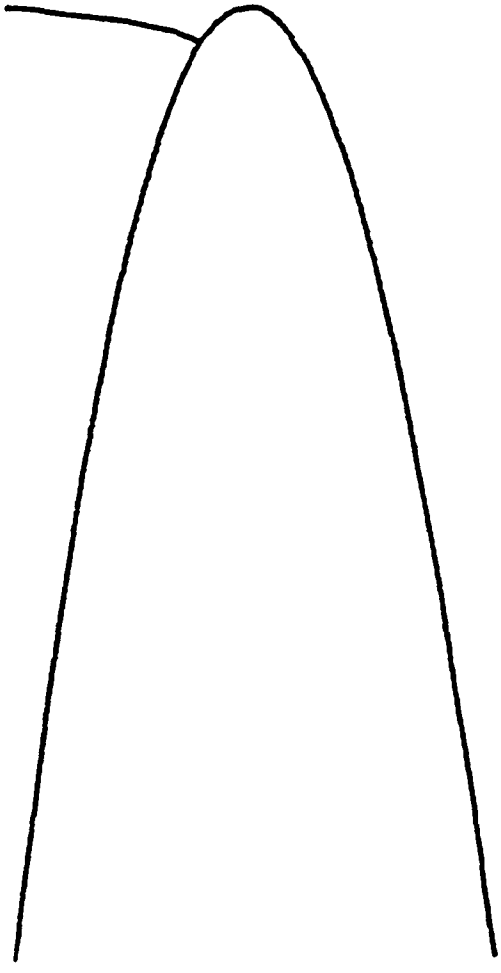


Figure 3.j Orthogonal accession.

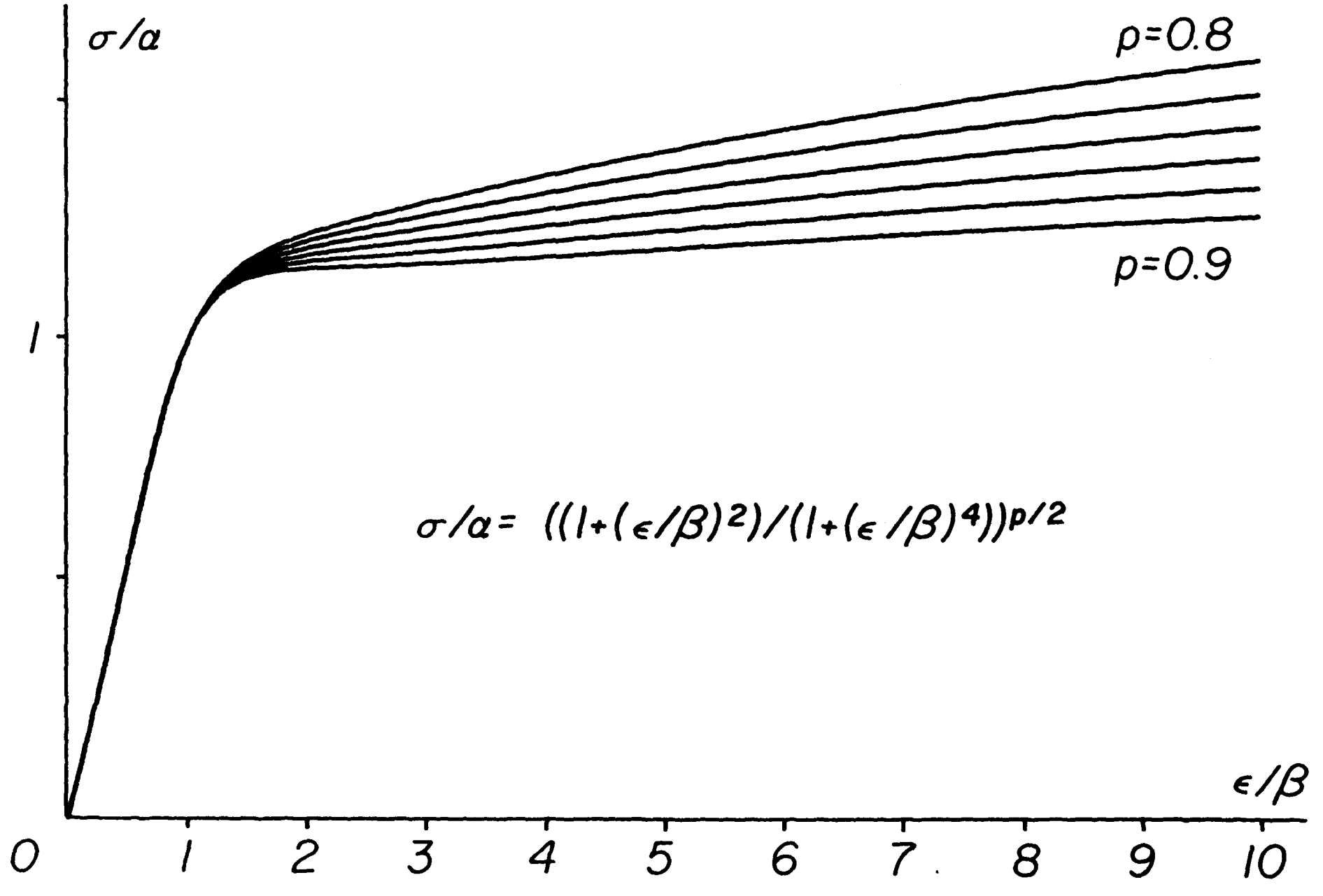


Figure 4. Stress σ vs. strain ε with distinct yield point.

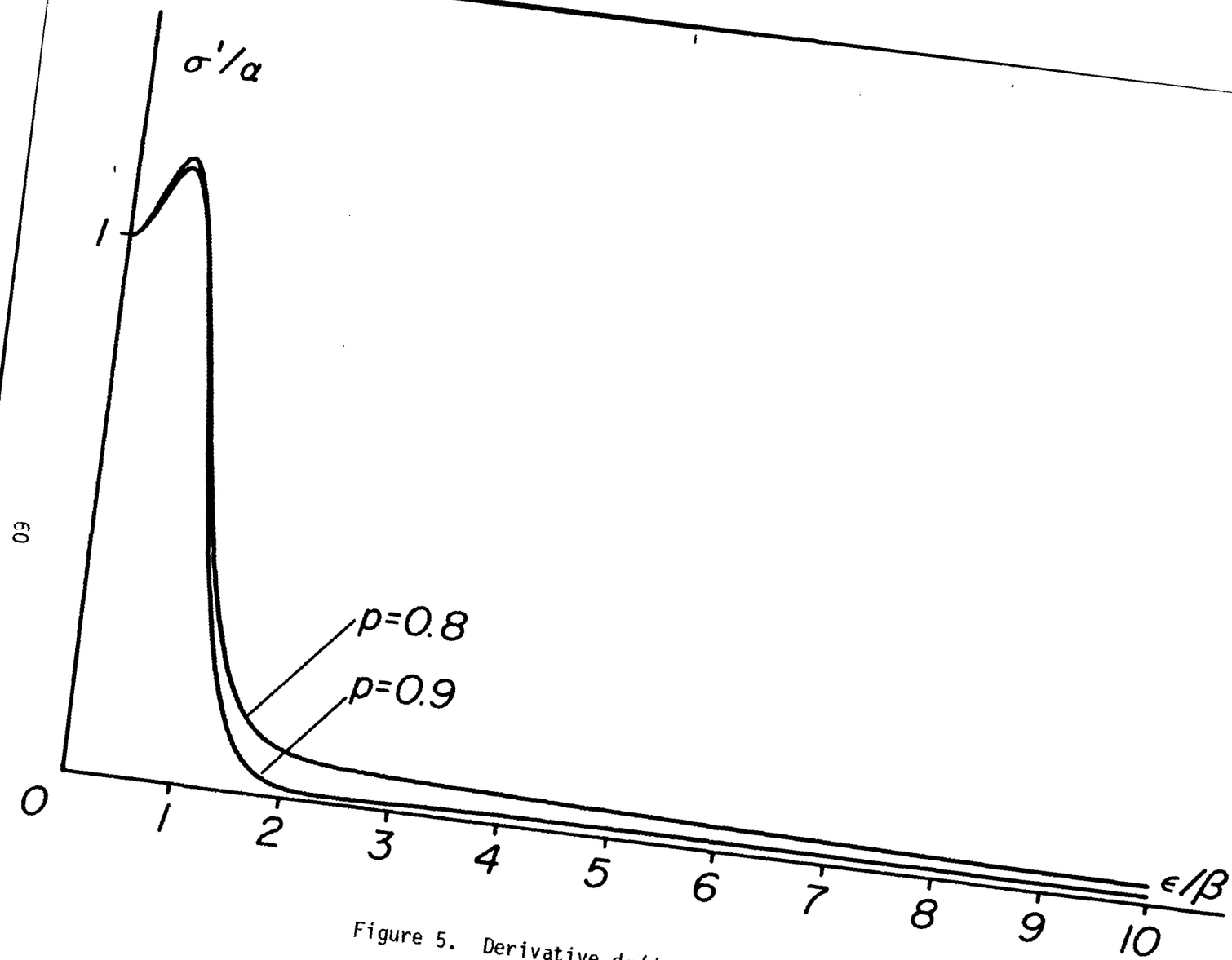


Figure 5. Derivative $d\sigma/d\varepsilon$ vs. ε for Figure 4.

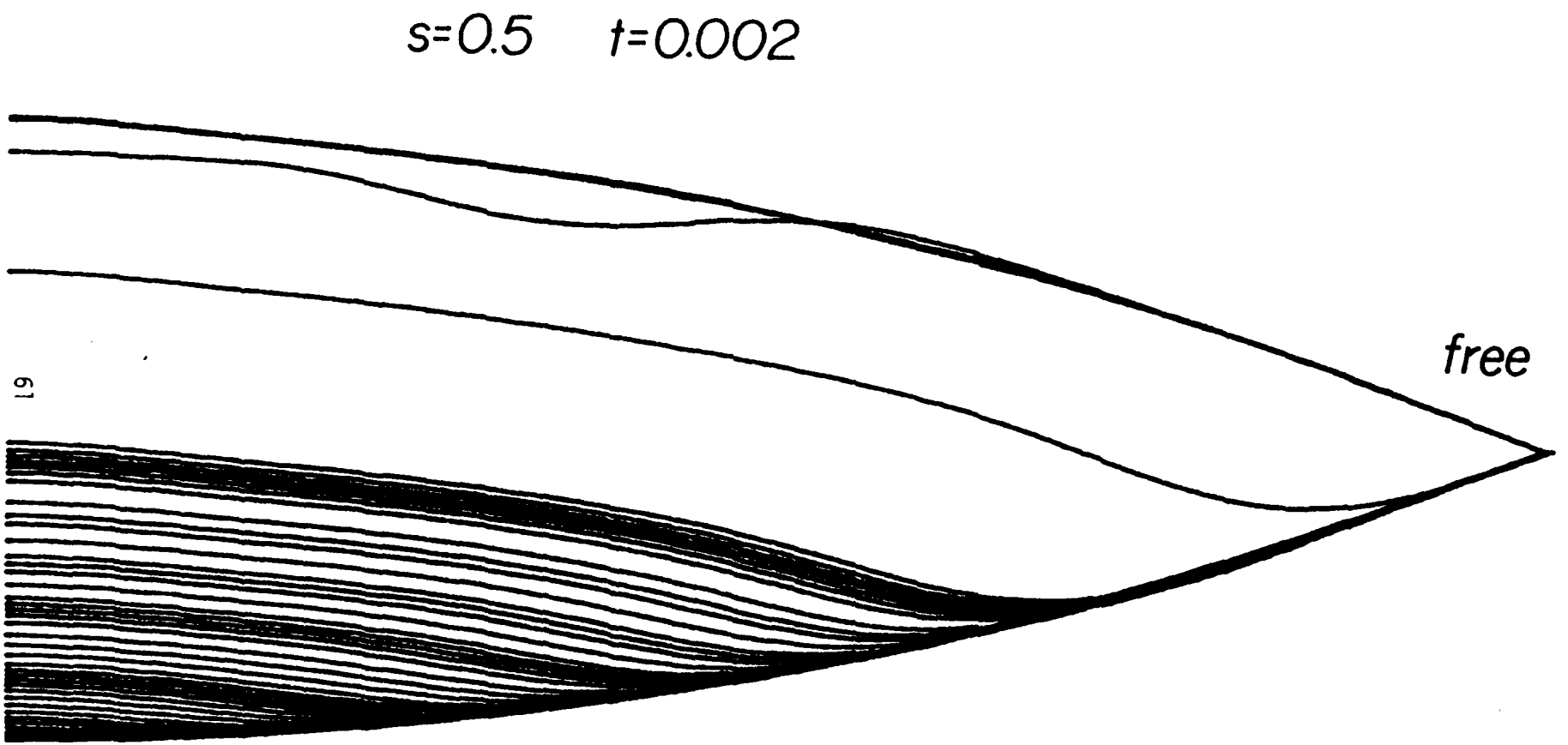


Figure 6. Inverted spherical cap.

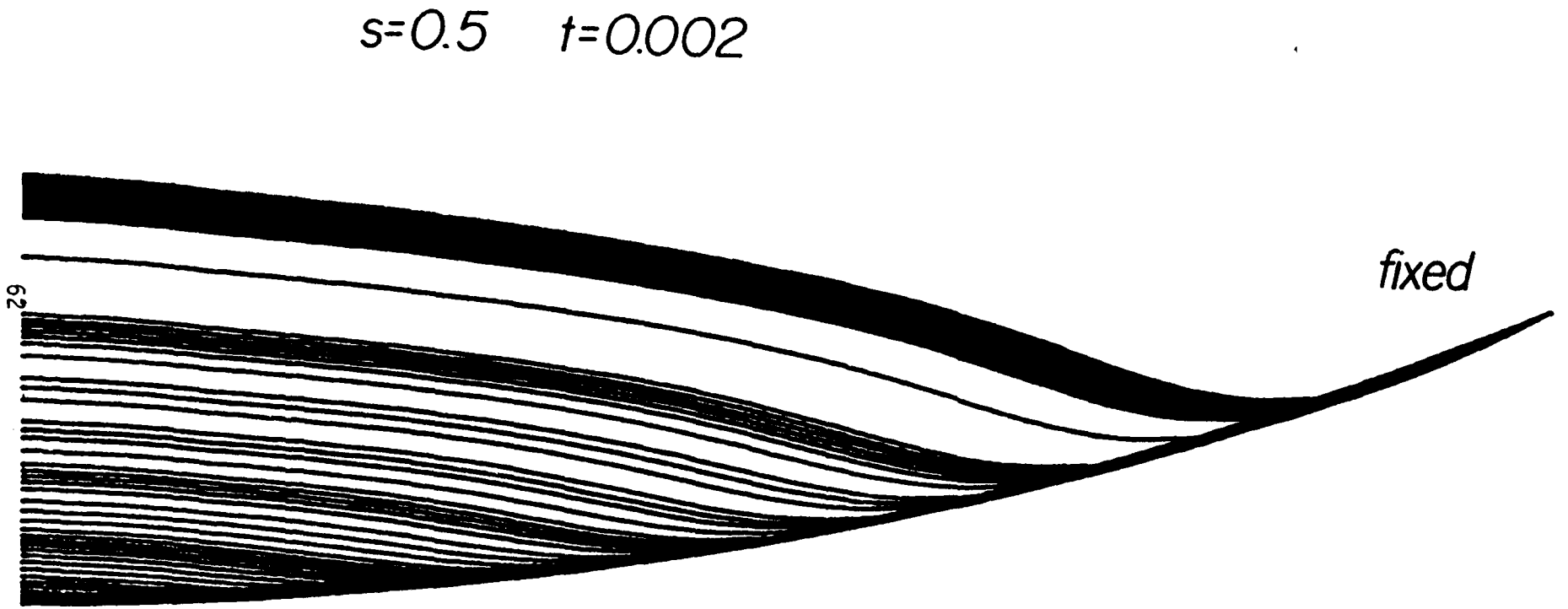


Figure 7. Inverted spherical cap.

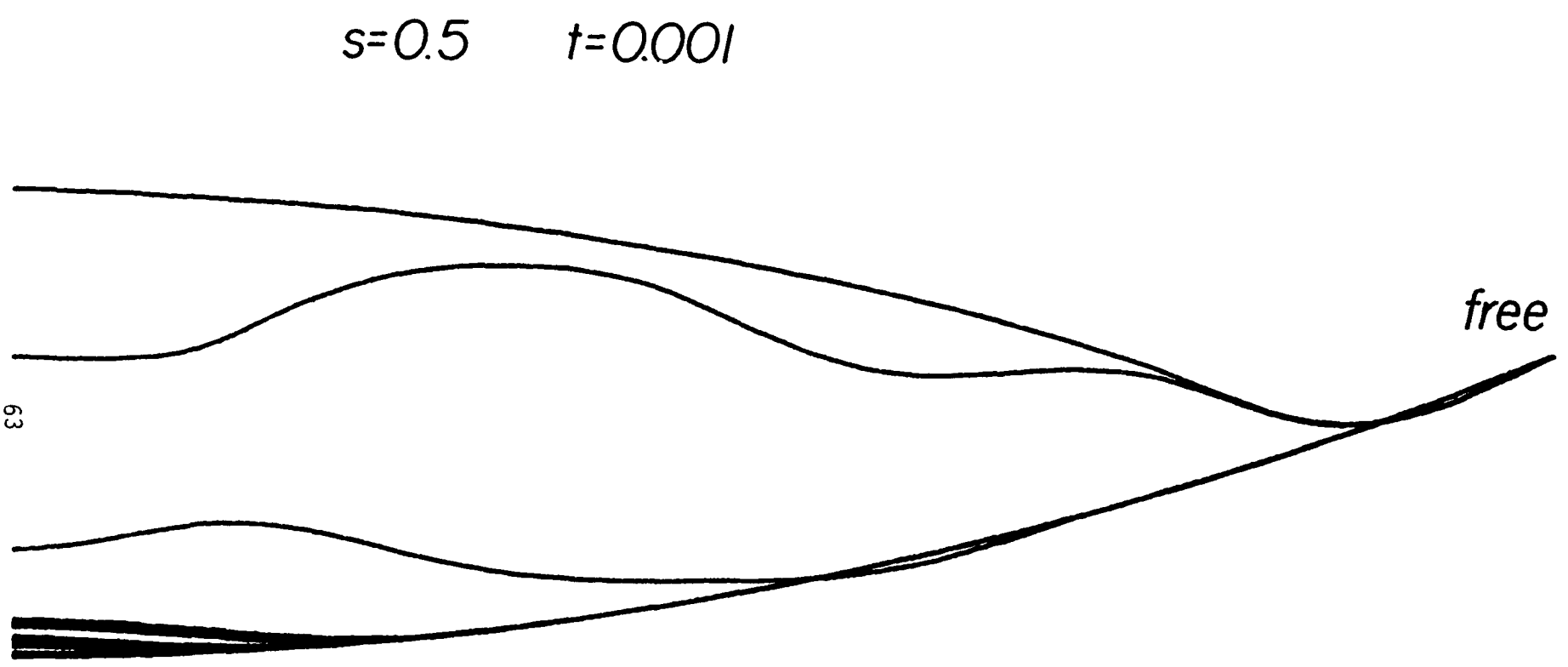


Figure 8. Inverted spherical cap.

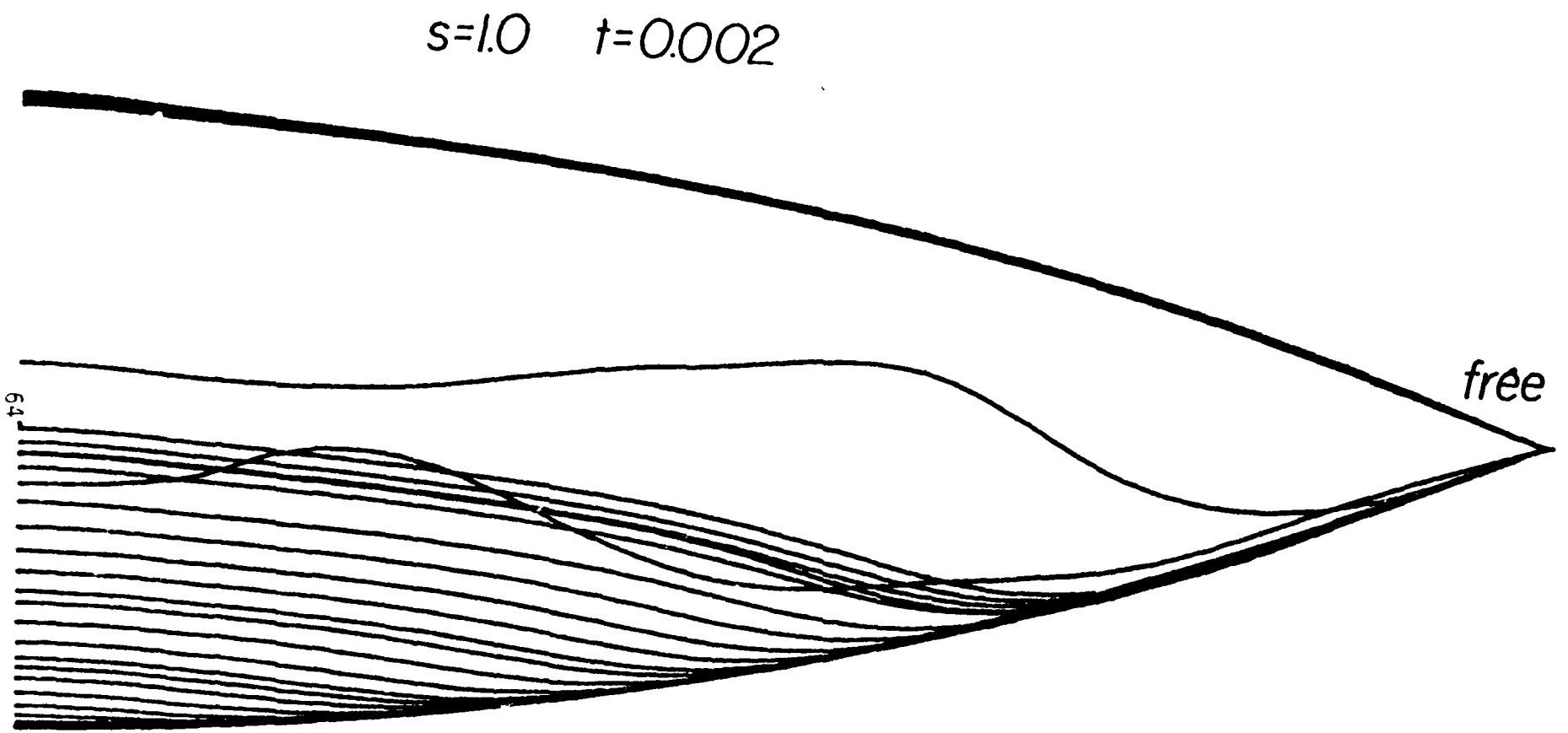


Figure 9. Inverted spherical cap.

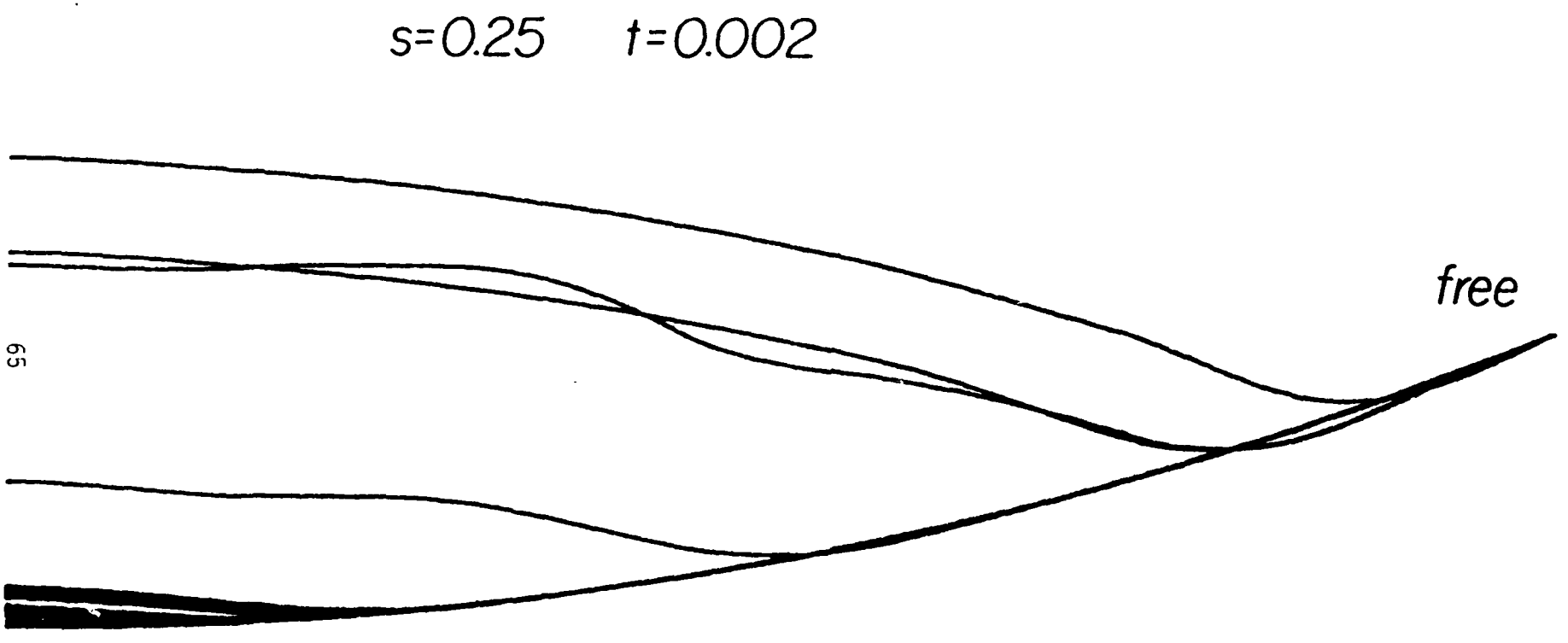


Figure 10. Inverted spherical cap.

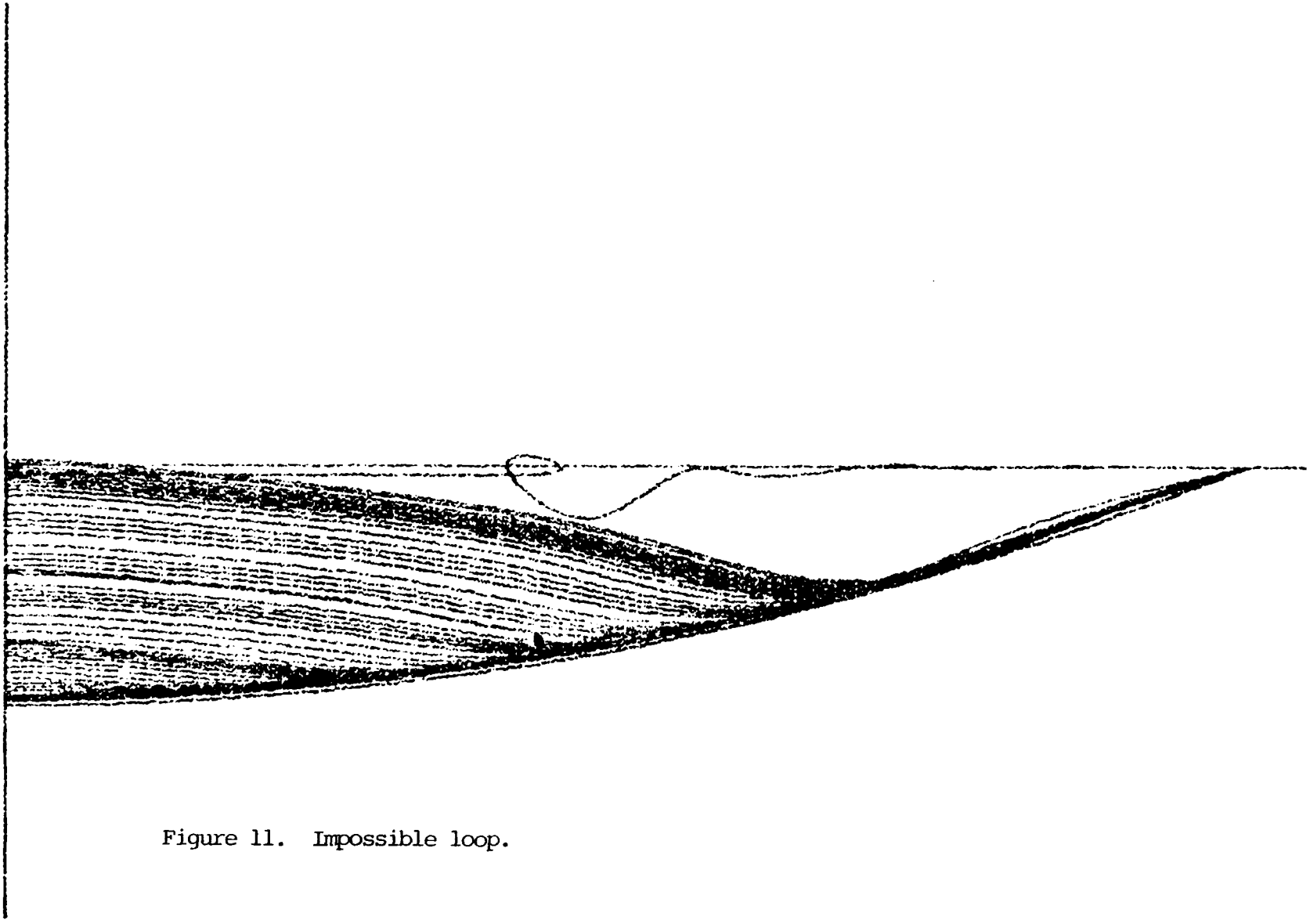


Figure 11. Impossible loop.

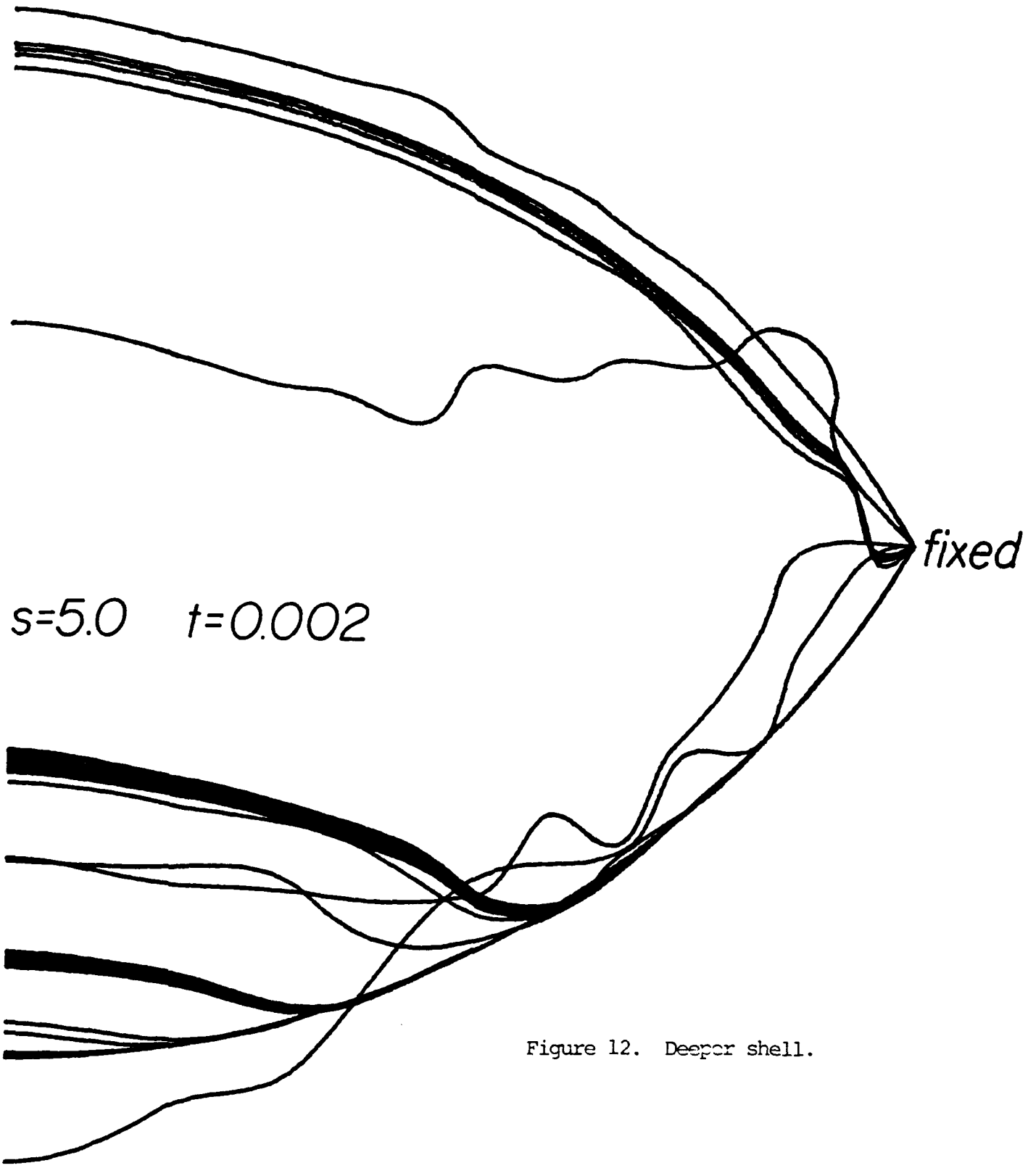


Figure 12. Deeper shell.

ASPECTS OF EDGE CONSTRAINTS IN SHEAR-DEFORMABLE PLATE AND SHELL ELEMENTS

Alexander TESSLER
Mechanics and Structures Branch
U.S. Army Materials Technology Laboratory
Watertown, Massachusetts 02172, U.S.A.

ABSTRACT. The method of explicit edge constraints for generating simple, consistent and efficient shear-deformable displacement bending elements is discussed. Particular attention is focused on the derivation of a highly desirable three-node shallowly curved shell element. Shell theory and finite element approximation issues are discussed in detail. Several numerical studies are carried out which demonstrate the effectiveness of the constraint methodology.

I. INTRODUCTION. The search for "optimal" shell finite elements has been underway for nearly two decades. In recent years it has further accelerated in light of significant progress in the technology of shear-deformable C^0 bending elements (e.g., [1-19]). Although the main obstacles for these developments, known as shear and membrane locking phenomena, have been addressed extensively and several remedial schemes have been proposed, a viable three-node doubly curved shear-deformable element, which is the most desirable element for general shell analysis, has not yet been developed. The purpose of this effort is to derive such an element.

We base our finite element derivation upon Reissner-Mindlin plate theory which will constitute the bending part of the element. To account for the membrane deformations and the membrane-bending coupling associated with the shell-element curvatures, we shall resort to Marguerre's shallow shell equations. Shallow shell elements of this type specialized to the axisymmetric response proved effective in discretizing shallow as well as deep shell structures [12]. The major advantage of this analytic approach over general shell formulations (e.g., [5,18]) is its inherent simplicity. Herein, the displacements and stress resultants are attributed to the element reference plane. Consequently, integrations are carried out across the reference plane rather than the curved surface as in the general shell elements.

According to Reissner-Mindlin theory [21-23], the strain-displacement relations can be expressed as:

$$\kappa = \{\kappa_{xx},\ \kappa_{yy},\ \kappa_{xy}\}^T = \mathbf{L}_1\theta \tag{1}$$

$$\gamma = \{\gamma_{xz},\ \gamma_{yz}\}^T = \mathbf{L}_2 w + \mathbf{I}\theta \tag{2}$$

where κ and γ are respectively the curvature and transverse shear strain vectors, θ is the bending rotation vector

$$\theta^T = \{\theta_y, \theta_x\} \tag{3}$$

with θ_x and θ_y denoting the bending rotations about the x and y axes, respectively, w is the transverse displacement (refer to Fig. 1), and the superscript T denotes transpose; $\mathbf{L}_1$ and $\mathbf{L}_2$ denote the linear strain-displacement operators, and $\mathbf{I}$ is an identity matrix:

$$\mathbf{L}_1 = \begin{bmatrix} \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \end{bmatrix}, \quad \mathbf{L}_2 = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4}$$

The Marguerre membrane strain-displacement relations for a <u>thin</u> shallow shell have the form [24]:

$$\epsilon = \mathbf{L}_1\mathbf{u} + \mathbf{L}_1(\xi)\mathbf{L}_2 w \tag{5}$$

with

$$\mathbf{u}^T = \{u, v\} \tag{6}$$

where u and v are the membrane displacements in the x and y coordinate directions, respectively; and $\xi=\xi(x,y)$ is the initial height of the shallow shell.

One important aspect, which in previous attempts to merge the two theories has not been addressed [9-12], is the conceptual difference in the transverse displacement variables appearing in (2) and (5). In (2), w is a weighted average transverse displacement across the thickness, whereas in (5) w represents the midsurface transverse displacement. The former variable comes into play due to the inclusion of shear

deformation in Reissner-Mindlin theory; the latter one is a consequence of the Kirchhoff thin-regime assumption, which neglects shear deformation. Utilizing (2), the Kirchhoff thinness constraint reads:

$$\mathbf{L}_2 w = - \mathbf{I}\theta \tag{7}$$

Replacing (7) into (5) yields the Marguerre membrane strains consistent with the Reissner-Mindlin strains:

$$\boldsymbol{\epsilon} = \mathbf{L}_1 \mathbf{u} - \mathbf{L}_1(\xi)\theta \tag{8}$$

The stress resultants, which are attributed to the reference plane of the shell, are related to the strains through the constitutive law:

$$\mathbf{N} = \{N_{xx},\ N_{yy},\ N_{xy}\}^T = \mathbf{A}\boldsymbol{\epsilon} \tag{9}$$

$$\mathbf{M} = \{M_{xx},\ M_{yy},\ M_{xy}\}^T = \mathbf{D}\boldsymbol{\kappa} \tag{10}$$

$$\mathbf{Q} = \{Q_x,\ Q_y\}^T \qquad = \mathbf{G}\boldsymbol{\gamma} \tag{11}$$

where **A**, **D** and **G** are respectively the membrane, bending and transverse shear constitutive matrices. The principle of virtual work can then be employed to derive the finite element stiffness equilibrium equations:

$$\iint_A (\mathbf{N}^T\delta\boldsymbol{\epsilon} + \mathbf{M}^T\delta\boldsymbol{\kappa} + \mathbf{Q}^T\delta\boldsymbol{\gamma} - q\delta w)\, dA = 0 \tag{12}$$

where q is the distributed transverse loading, A is the reference plane area, and δ denotes the variational operator.

II. FINITE ELEMENT ISSUES. The development of effective curved shear-deformable shell elements is severely hampered by the "locking phenomena" (extreme stiffening), reflecting the inability of the shell to bend without stretching ("membrane locking") and transverse shearing ("shear locking"). The two phenomena are directly link to the penalized strain energy which, in its nondimensional form, can be expressed as:

$$U(\kappa,\gamma,\varepsilon) = U_b(\kappa) + \alpha_s U_s(\gamma) + \alpha_m U_m(\varepsilon) \tag{13}$$

in which $U_b(\kappa)$, $U_s(\gamma)$, and $U_m(\varepsilon)$ denote the nondimensional bending, transverse shear, and membrane energy integrals; and α_s and α_m are the nondimensional shear and membrane penalty parameters, respectively. Note that $\alpha_s=O(\lambda^2/t^2)$ and $\alpha_m=O((\beta\lambda)^2/t^2)$, where λ and β are respectively some characteristic span and slope of the shallow shell [11,12]. As the shell thickness, t, diminishes to zero, both α_s and α_m approach infinity, thereby enforcing the vanishing shear and membrane strains:

$$\begin{aligned} &\mathbf{L}_2 w \to -\mathbf{I}\theta \quad \text{(Kirchhoff constraints)} && \text{(a)} \\ &\mathbf{L}_1 \mathbf{u} \to \mathbf{L}_1(\xi)\theta \quad \text{(Membrane inextensibility constraints)} && \text{(b)} \end{aligned} \tag{14}$$

The particular appeal of this theory is that the variational statement (12) requires a class of C^0 continuous approximations for the w, **u**, and θ fields (since their highest spatial derivative in (12) is of order one) and, therefore, simple shape functions can be used. On the other hand, constraints (14), when imposed at the element level, pose severe limitations on the kinematic freedom attainable by each element.

A consistent resolution of this deficiency for a successful discretization of the theory is twofold: (i) redefine the penalty parameters to allow relaxation of (14) at the element level; (ii) implement appropriate interpolation schemes to best accommodate (14). The two complementary approaches have shown to be effective and produced a series of efficient and reliable bending elements [6-8,11-14].

(i) REVIEW OF PENALTY RELAXATION CONCEPT. The first approach deals with an introduction of a parametric device in the variational statement for the purpose of relaxing enforcement of penalty constraints at the element level.

Concurrently with the element displacement approximations defined as w^h, $\mathbf{u}^h$, and θ^h, we also approximate the constitutive matrices **A** and **G**, incorporating appropriate "penalty relaxation" parameters for the element:

$$\mathbf{N}^h = \phi_m^2 \, \mathbf{A}\boldsymbol{\varepsilon}^h, \quad \mathbf{Q}^h = \phi_s^2 \, \mathbf{G}\boldsymbol{\gamma}^h, \quad \mathbf{M}^h = \mathbf{D}\boldsymbol{\kappa}^h \tag{15}$$

where the element strains are

$$\boldsymbol{\varepsilon}^h = \mathbf{L}_1\mathbf{u}^h - \mathbf{L}_1(\xi)\boldsymbol{\theta}^h, \quad \boldsymbol{\gamma}^h = \mathbf{L}_2\mathbf{w}^h + \mathbf{I}\boldsymbol{\theta}^h, \quad \boldsymbol{\kappa}^h = \mathbf{L}_1\boldsymbol{\theta}^h \tag{16}$$

and the penalty relaxation parameters are nondimensional positive quantities of the form:

$$\phi_i^2 = (1 + C_i\alpha_i)^{-1} \quad (i = m,s) \tag{17}$$

where C_i are positive element constants, and α_i are element analytic penalty parameters of the order $\alpha_s = O(h^2/t^2)$ and $\alpha_m = O((\beta^h h)^2/t^2)$, where h and β^h are respectively some characteristic span and slope of the element. The corresponding principle of virtual work for a single element approximation takes the form:

$$\iint_{A^e} [(\mathbf{N}^h)^T\delta\boldsymbol{\varepsilon}^h + (\mathbf{M}^h)^T\delta\boldsymbol{\kappa}^h + (\mathbf{Q}^h)^T\delta\boldsymbol{\gamma}^h - q\delta w^h] \, dA = 0 \tag{18}$$

where integration extends over the element reference plane with A^e denoting the element reference area. The resulting element strain energy appears in the basic form of (13), except that all quantities are superscribed with h (i.e., element approximations); however, the element penalty parameters take a fundamentally different form:

$$\alpha_i^h = \alpha_i/(1 + \alpha_i) \quad (i = m,s) \tag{19}$$

These α_i^h penalties relax enforcement of (14) as $t \to 0$ and thus alleviate possible spurious constraining. Note, however, as the kinematic approximations improve with the h-refinement, α_i^h approach their analytic values α_i, thus ensuring convergence to the "true" solution both in the constitutive and kinematic sense [4,8,12].

(ii) ANISOPARAMETRIC INTERPOLATION SCHEME. The other fundamental means for improving element behavior is to devise appropriate interpolation schemes which best accommodate the requirements of (14). Such interpo-

lations, termed anisoparametric [13], employ distinctly different degree polynomials for w, $\boldsymbol{\theta}$, and **u** to reflect the differences in the order of the differential operators $\mathbf{L}_2$ and **I** in (14a) and, likewise, $\mathbf{L}_1$ and $\mathbf{L}_1(\xi)$ in (14b). The specific aim is to <u>design out</u> the unwanted "spurious" constraint equations arising from (14) [14].

To represent the bending part of the shell element, we adopt the 3-node anisoparametric plate element [13-14], in which θ_x and θ_y are interpolated linearly, while w is represented by a complete quadratic polynomial; throughout the formulation, area-parametric coordinates $\zeta=(\zeta_1,\zeta_2,\zeta_3)$ are used as a basis for all interpolations (refer to Fig. 2):

$$\theta_I^h = \mathbf{N}^{(1)}\boldsymbol{\theta}_I^h \quad (I=x,y), \qquad w^h = \mathbf{N}^{(2)}\mathbf{w}^h \tag{20}$$

where $\mathbf{N}^{(1)}$ and $\mathbf{N}^{(2)}$ are the row vectors of linear and quadratic shape functions, respectively, and

$$(\boldsymbol{\theta}_I^h)^T = \{\theta_{Ij}^h\}, \quad (\mathbf{w}^h)^T = \{w_k^h\} \quad (I=x,y;\ j=1,2,3;\ k=1,\ldots,6) \tag{21}$$

are the vectors of nodal dof.

Adopting the shell element of constant curvature (i.e., interpolating $\xi^h(\zeta)$ parabolically), constraints (14b) necessitate a complete 10-term cubic polynomial for the u and v displacements:

$$u^h = \mathbf{N}^{(3)}\mathbf{u}^h, \qquad v^h = \mathbf{N}^{(3)}\mathbf{v}^h \tag{22}$$

where $\mathbf{N}^{(3)}$ is a row vector of cubic shape functions, and

$$(\mathbf{u}^h)^T = \{u_k^h\}, \quad (\mathbf{v}^h)^T = \{v_k^h\} \quad (k=1,\ldots,10) \tag{23}$$

are the vectors of nodal dof.

Evidently, the anisoparametric interpolations produce the same polynomial representation for the left- and right-hand sides of the constraint equations (14) — the condition that is paramount to enhancing element behavior in the vanishing thickness regime.

<u>(a) Edge Shear Constraints.</u> Although the initial w^h rests on six w^h dof (i.e., three corner and three mid-edge dof), a kinematically

consistent elimination of the mid-edge dof is possible a priori to the element stiffness derivation. To obtain a 3-node pattern, w^h can be constrained by the one-dimensional edge constraints:

$$\gamma^{(k)}_{sz,_s} = \frac{\partial}{\partial s}\Big[\, w^h(s),_s + \theta^h_n(s) \,\Big]^{(k)} = 0 \quad (k=1,2,3) \tag{24}$$

where s denotes a coordinate running along the k^{th} edge of the triangular element reference plane; and $\theta^h{}_n(s)$ is the tangential edge rotation which is related to $\theta^h{}_x(s)$ and $\theta^h{}_y(s)$ via an orthogonal transformation. From (24), there result three decoupled equations in terms of the mid-edge w^h dof, which give rise to the constraints:

$$\mathbf{w}^h_c = \mathbf{W}\,\mathbf{w}^h + \mathbf{W}_x\boldsymbol{\theta}^h_x + \mathbf{W}_y\boldsymbol{\theta}^h_y \tag{25}$$

where $\mathbf{W}_q$ are 3x3 transformation matrices, and

$$(\mathbf{w}^h_c)^T = \{w^h_{j+3}\}, \quad (\mathbf{w}^h)^T = \{w^h_j\} \quad (j=1,2,3) \tag{26}$$

Upon substituting (25) into (20), we obtain a 3-node interpolation for the transverse displacement.

(b) Edge Membrane constraints. In the manner analogous to the above dof reduction for w^h, one-dimensional edge constraints can be devised to condense out the intra-edge u^h and v^h dof. The following constraint equations provide four edge-compatible relations for each edge:

$$\frac{\partial^p}{\partial s^p}\begin{bmatrix} \varepsilon_s(s) \\ \gamma_{sn}(s) \end{bmatrix}^{(k)} = \frac{\partial^p}{\partial s^p}\begin{bmatrix} \underline{u}^h_{,s} - \xi^h_s\,\theta^h_n \\ \underline{v}^h_{,s} - \xi^h_n\,\theta^h_n - \xi^h_s\,\theta^h_s \end{bmatrix}^{(k)} \equiv \mathbf{0} \quad (k=1,2,3;\ p=1,2) \tag{27}$$

where $\underline{u}^h(s)$ and $\underline{v}^h(s)$ are cubic displacement fields along and normal to the k-th edge, respectively, and

$$\xi_q^h \equiv \xi^h(\zeta),_q \Big|_{\zeta_k = 0} \qquad (q= s,n;\ k=1,2,3) \tag{28}$$

are the k-th edge slopes.

By the use of appropriate orthogonal transformations, (27) are expressed in terms of the shell element variables of interest, namely, u^h, v^h, θ^h_x and θ^h_y dof, and algebraically solved for the intra-edge u^h and v^h dof:

$$\begin{aligned} \mathbf{u}_c^h &= \mathbf{U}\, \mathbf{u}^h + \mathbf{U}_x \boldsymbol{\theta}_x^h + \mathbf{U}_y \boldsymbol{\theta}_y^h \\ \mathbf{v}_c^h &= \mathbf{V}\, \mathbf{v}^h + \mathbf{V}_x \boldsymbol{\theta}_x^h + \mathbf{V}_y \boldsymbol{\theta}_y^h \end{aligned} \tag{29}$$

where

$$(\mathbf{u}_c^h)^T = \{u_i^h\}, \quad (\mathbf{v}_c^h)^T = \{v_i^h\} \quad (i=4,\ldots,9) \tag{30}$$

and $\mathbf{U}_q$ and $\mathbf{V}_q$ are 6x3 transformation matrices. Equations (29) are substituted into the initial interpolations (22) to give the constrained fields for the membrane displacements in terms of the corner-node dof and two centroidal dof. The latter dof are condensed-out statically after the formation of the element stiffness matrix and consistent load vector. Consequently, a 3-node, 15 dof element pattern is achieved.

Note that the edge constraint procedures just described preserve the original polynomial order of the constrained variables (w^h, u^h and v^h); moreover, one can show that the constrained fields are fully compatible across element edges, and they allow for rigid-body motion without straining. For further details on this procedure and for the explicit form of the shape functions, the interested reader is referred to [13,14,20].

The remainder of the formulation follows standard finite element procedures. Application of the virtual work statement (18), while performing exact integration throughout, yields the element stiffness equations. The issue of the rotational variable normal to the reference plane, θ^h_z, needed to avoid mathematical singularities in the global coordinates, produces three additional dof for the element (e.g., see [10]).

III. NUMERICAL EXAMPLES. An important step in completing the relaxation methodology of Section II is to obtain appropriate α_i parameters and the values for C_i (i=m,s). Herein, we adopt the approach developed in [13], where α_i are defined as:

$$\alpha_i \equiv \Sigma\, k_i^{\theta} / \Sigma\, k_b^{\theta} \quad (i=s,m; \quad b \text{ - bending}) \tag{31}$$

in which k^{θ}_i and k^{θ}_b denote the element diagonal stiffness coefficients associated with θ^h_x and θ^h_y dof for the unrelaxed case, i.e., $\phi^2_i = 1$. As far as the "optimal" values for C_s and C_m, these are determined from numerical testing. The shear relaxation constant, $C_s=2$, has already been established to ensure free of locking plate-element behavior [13]; $C_m=1$ was chosen from the numerical results of the present study.

The present element was critically tested on four challenging shell problems, where two of its versions were employed: (a) the element with both the shear and membrane relaxations ($C_s=2$, $C_m=1$), labeled "MIN3sm", (b) the element with the shear relaxation only ($C_s=2$, $C_m=0$), labeled "MIN3s". Our findings are summarized as follows.

(i) Test of Rigid-Body Motion. A spectral analysis was performed on the element stiffness matrix for the flat, singly curved, and doubly curved element geometry, to check MIN3's ability to move as a rigid body without incurring any straining. Under all conditions tested, there resulted six requisite zero eigenvalues associated with rigid body motion.

(ii) Clamped Circular Arch. A simple test of both membrane inextensibility and shearless deformation is a clamped, thin circular arch under a tip bending moment (Fig.3). An additional modeling difficulty is that the arch is rather narrow, hence the element aspect ratios are large. At all discretization levels, exact values for the stress resultants are obtained in each element (i.e., $M_z=M$, with all forces vanishing). Figure 3 depicts a convergence study of the tip bending rotation, which is also a direct measure of the strain energy for this problem. Note that MIN3s exhibits considerable membrane stiffening. Clearly, MIN3sm is a superior performer, yielding highly accurate results even under coarse discretizations.

(iii) Pinched Cylinder with Free Ends. The free-ended cylindrical shell subjected to two radial forces 180 degrees apart (Fig. 4) is a widely used test problem to establish how well a singly curved shell element can represent inextensinal bending. As $t/R \to 0$, pure inextensional state of deformation is attained in the cylinder.

Figure 4 shows convergence studies of the deflection under the load for the moderately thin (R/t=50) and very thin (R/t=2000) cylinders. The present results are compared with the exact solution and those of four reduced integration quadrilateral elements (for details on these quadrilaterals, refer to [1]). Both MIN3s and MIN3sm exhibit excellent behavior, with MIN3s being somewhat stiffer than MIN3sm.

(iv) Pinched Hemisphere. A thin hemispherical shell under self-equilibrating radial forces (Fig. 5) is in the state of near extensional bending, having large rigid-body rotations in the deformed configuration. This problem is a challenging test for doubly curved elements (e.g., see [25]).

The convergence study for the deflection under the load is depicted in Fig. 5, where the results of nine quadrilateral elements, examined in [5], were included for comparison. Here again, MIN3sm evolved among the best performing elements, while MIN3s exhibited some excessive membrane stiffening.

In summary, we conclude that our three-node shallow shell element (MIN3sm) is an excellent candidate for general shell analysis — it is theoretically sound, has the simplest nodal/dof pattern, possesses six rigid-body modes, and is devoid of both membrane and shear locking.

REFERENCES.

1. T. J. R. Hughes, The Finite Element Method: Linear Static and Dynamic Finite Element Analysis, Chapter 5, Prentice-Hall, N.J., (1987).

2. T. J. R. Hughes and E. Hinton (eds.), Finite Element Methods for Plate and Shell Structures, Vol. 1: Element Technology, Pineridge Press, Swansea, U.K., (1986).

3. T. Belytschko, "A review of recent developments in plate and shell elements," in Computational Mechanics - Advances and Trends (ed. A. K. Noor), ASME, AMD-Vol.75, N.Y., 217-231 (1986)

4. A. Tessler, "Shear-deformable bending elements with penalty relaxation," in Finite Element Methods for Plate and Shell Structures, Vol. 1: Element Technology, (eds. T. J. R. Hughes and E. Hinton), Pineridge Press, Swansea, U.K., 266-290 (1986).

5. G. M. Stanley, "Continuum-Based Shell Elements," Ph.D. Dissertation, Stanford University (1985).

6. A. Tessler and S. B. Dong, "On a hierarchy of conforming Timoshenko beam elements," Computers and Structures, 14, 335-344 (1981).

7. A. Tessler, "An efficient, conforming axisymmetric shell element including transverse shear and rotary inertia," Computers and Structures, 15, 567-574 (1982).

8. 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, 311-335 (1983).

9. H. Stolarski and T. Belytschko, "Shear and membrane locking in curved C^0 elements," Computer Meth. Appl. Mech. Engrg., 41, 279-296 (1983).

10. M. A. Crisfield, Finite Elements and Solution Procedures for Structural Analysis, Vol.1: Linear analysis, Pineridge Press, Swansea, U.K. (1986).

11. A. Tessler and L. Spiridigliozzi, "Curved beam elements with penalty relaxation," Int. J. Numer. Meth. Engng., 23, 2245-2262 (1986).

12. A. Tessler and L. Spiridigliozzi, " Resolving membrane and shear locking phenomena in curved shear-deformable axisymmetric shell elements," Int. J. Num. Meth. Engrg., 26, 1071-1086 (1988).

13. A. Tessler and T. J. R. Hughes, "A three-node Mindlin plate element with improved transverse shear," Comput. Meths. Appl. Mech. Engrg., 50, 71-101 (1985).

14. A. Tessler, "A priori identification of shear locking and stiffening in triangular Mindlin elements," Comput. Meths. Appl. Mech. Engrg., 53, 183-200 (1985).

15. C. Ramesh Babu and G. Prathap, "A field consistent two-noded curved axisymmetric shell element," Int. J. Numer. Meth. Engng., 23, 1245-1261 (1986).

16. I. Fried, A. Johnson and A. Tessler, "Minimal-degree thin triangular plate and shell bending finite elements of order two and four," Comput. Meths. Appl. Mech. Engrg., 56, 283-307 (1986).

17. E. Hinton and H. C. Huang, "A family of quadrilateral Mindlin plate elements with substitute shear strain fields," Computers and Structures, 33, 409-431 (1986).

18. K. J. Bathe and E. N. Dvorkin, "A formulation of general shell elements - the use of mixed interpolation of tensorial components," Int. J. Num. Meth. Engrg., 22, 697-722 (1986).

19. A. F. Saleeb, T. Y. Chang and S. Yingyeunyong, "A mixed formulation of C^0-linear triangular plate/shell element — the role of edge shear constraints," Int. J. Num. Meth. Engrg., 26, 1101-1128 (1988).

20. A. Tessler, "A C^0-anisoparametric three-node shallow shell element for general shell analysis," Army Materials Technology Laboratory, Technical Report (to appear).

21. E. Reissner, "On the Theory of Bending of Elastic Plates," Journal of Mathematics and Physics, 23, 184-191 (1944).

22. E. Reissner, "The Effects of Transverse Shear Deformation on the Bending of Elastic Plates," Journal of Applied Mechanics 12, 69-77 (1945).

23. R. D. Mindlin, "Influence of Rotatory Inertia and Shear on Flexural Motions of Isotropic, Elastic Plates," Journal of Applied Mechanics, 18, 31-38 (1951).

24. K. Marguerre, "Zur Theorie der gekrummten Platte grosser Formenderung," Proc. 5th Internat. Congress of Applied Mechanics, 693-701 (1938).

25. R. H. MacNeal and R. L. Harder, "A proposed standard set of problems to test finite element accuarcy," in Proc. of AIAA Conf. on Structures and Structural Dynamics, Palm Springs, CA (1984).

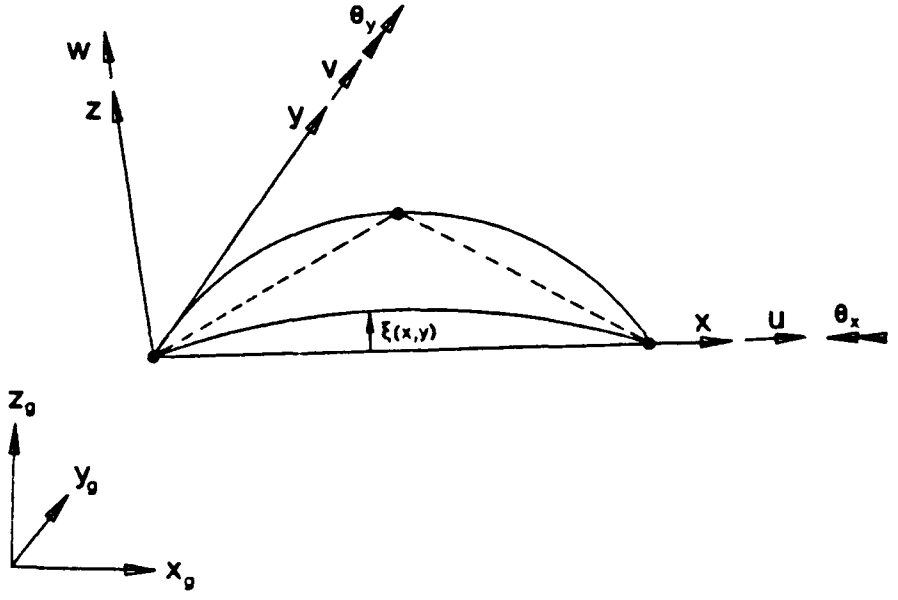


Fig. 1. Shallow Shell Notation.

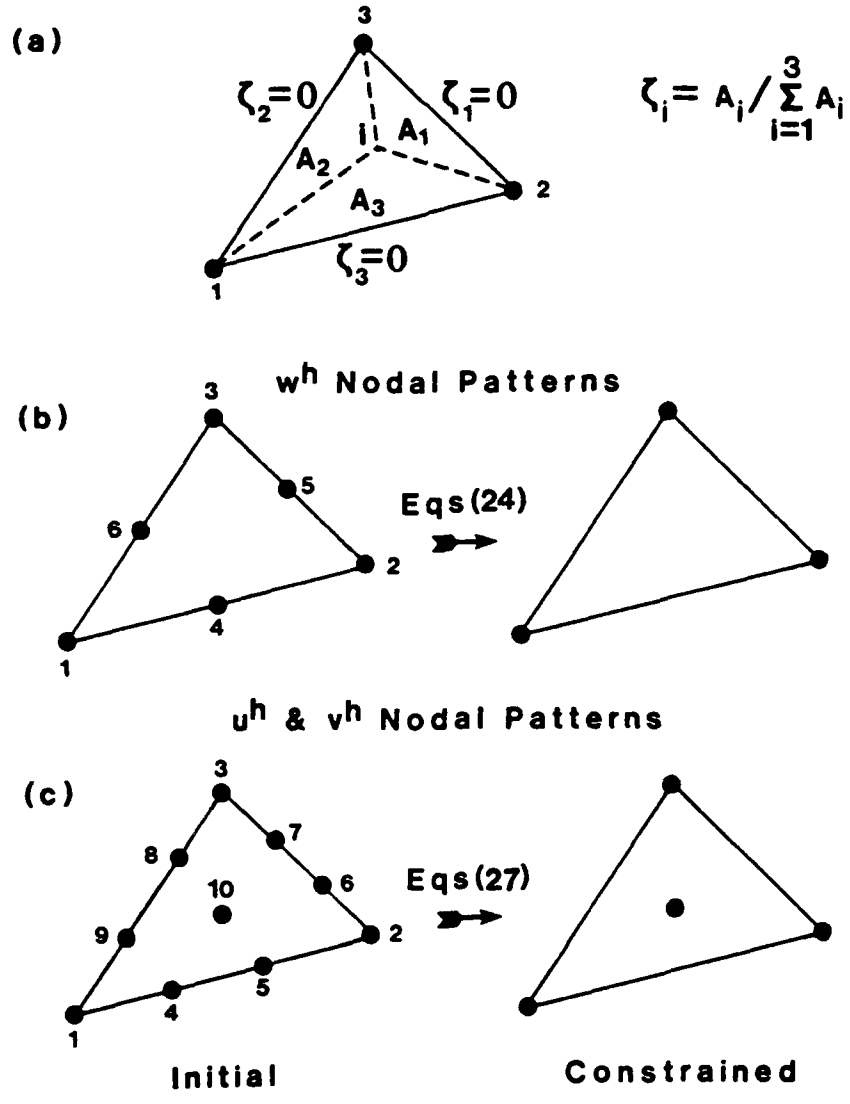


Fig. 2. (a) Area-Parametric Coordinates,
(b) w^h Initial and Constrained Nodal Patterns,
(c) u^h and v^h Initial and Constrained Nodal Patterns.

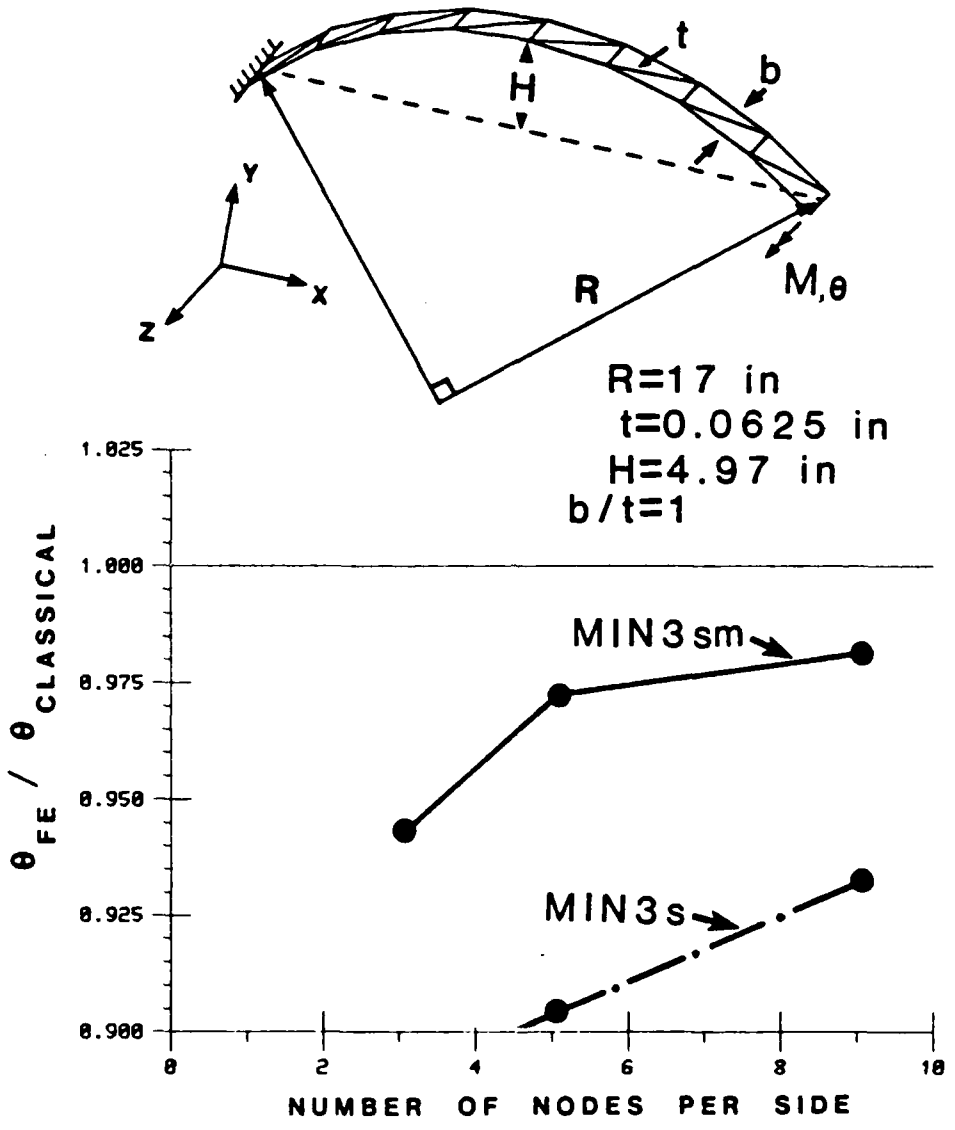
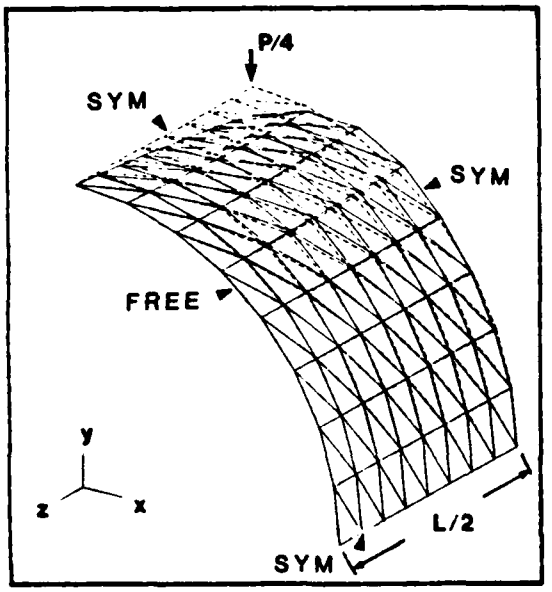
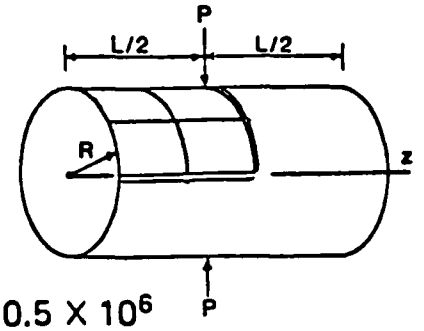


Fig. 3. Cantilevered Arch Under Tip Moment.





$E = 10.5 \times 10^6$
$\nu = 0.3125$
$R = 4.953$
$L = 10.35$

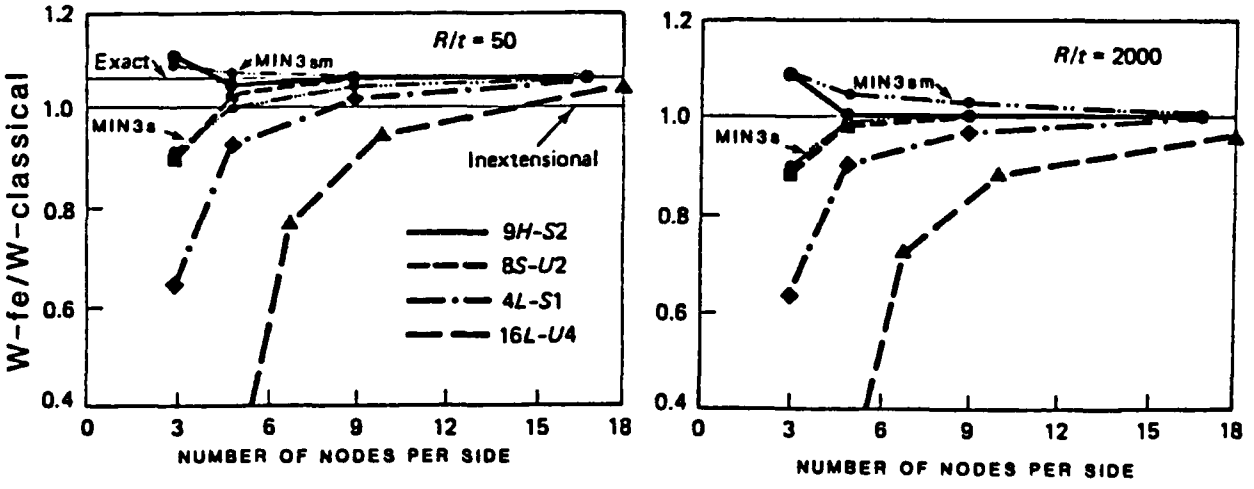


Fig. 4. Pinched Cylinder with Free Ends.

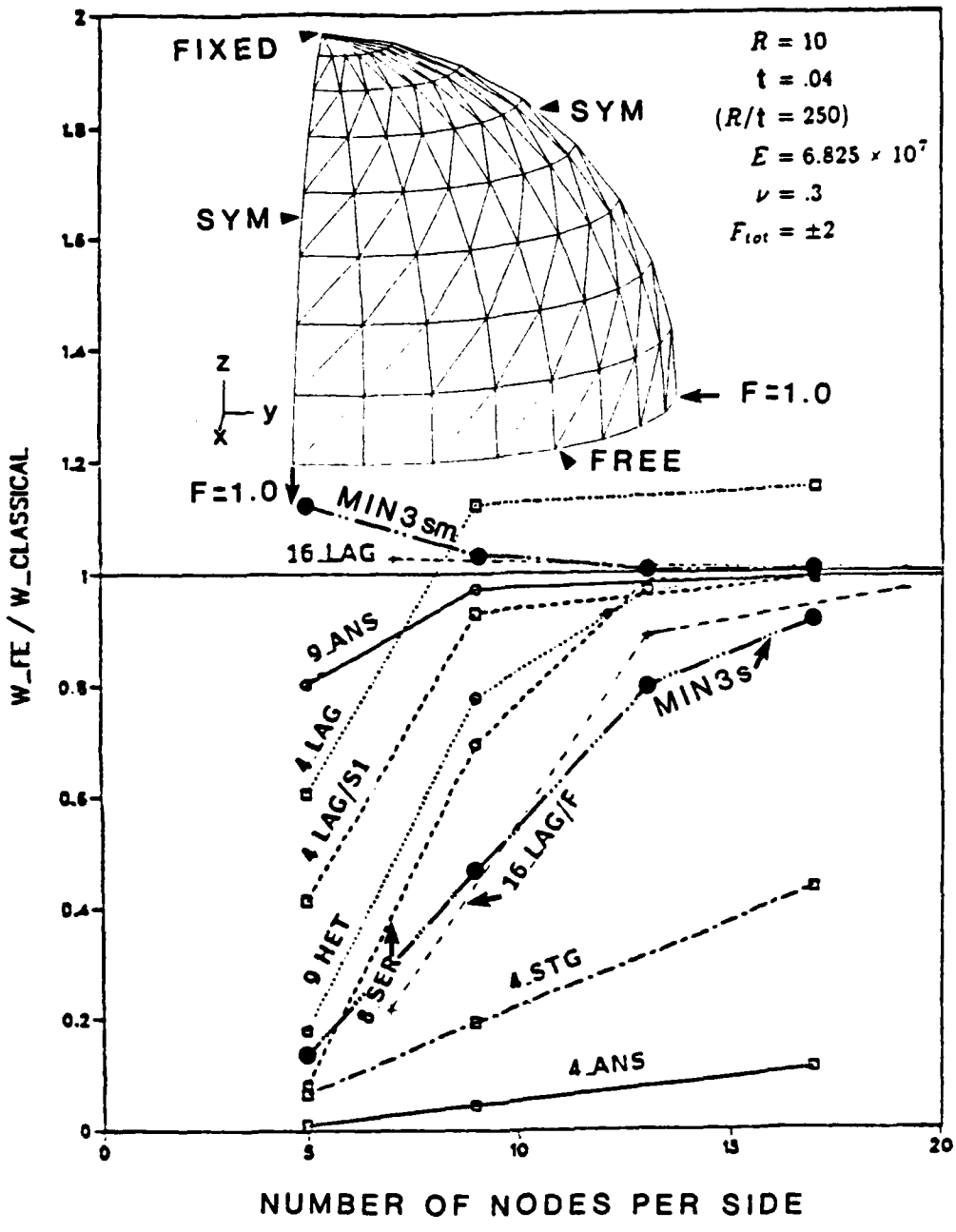


Fig. 5. Pinched Hemisphere.

SOME NUMERICAL RESULTS IN NONLINEAR AND VISCOELASTIC FRACTURE

A.E. Beagles*, J.R. Walton+, M.K. Warby* and J,R. Whiteman*

*BICOM, Institute of Computational Mathematics
Brunel University, Uxbridge, England.
+Department of Mathemtics, Texas A & M University,
College Station, Texas, U.S.A.

This work was undertaken whilst J.R. Walton was a visitor in BICOM for the 1987-88 academic year. This visit was financed in part by the Science and Engineering Research Council under Grant No. GR/E/40538 and this support is gratefully acknowledged.

ABSTRACT

Finite element methods are used to approximate parameters which characterise fracture properties of solids containing stationary cracks under conditions of (a) elasto-plastic and (b) viscoelastic deformation.

1. FINITE ELEMENT METHODS FOR PLASTICITY AND VISCOELASTICITY

1.1 Introduction

In this paper we are concerned with the finite element approximation of parameters which characterise fracture properties (a) for stationary cracks in materials which exhibit elasto-plastic deformation and (b) for stationary cracks in materials which exhibit viscoelastic deformation. For these problems the first task is to define satisfactory mathematical models of the deformation and of a fracture parameter, after which the finite element method can be applied so that the deformation and the fracture parameters can be approximated. This field is currently the subject of intensive research, as is evident from the succession of conference proceedings on nonlinear computational solid mechanics and fracture mechanics which are appearing, and the work reported on here is as yet only at a preliminary stage.

Only planar problems are considered here and the approach to the discretisation of both the elasto-plastic and viscoelastic problems is via the stress equilibrium equation of continuum mechanics and the constitutive relations relevant to each context. The Galerkin method is applied in each case.

A J-type path integral is employed for the case of fracture in the elasto-plastic case. This is introduced and approximated in Section 2, after which results of some numerical experiments for an elastic perfectly plastic problem are presented. The limitations of this approach to nonlinear fracture are discussed. A similar approach, but also involving crack opening displacement (COD), is taken in Section 3 for viscoelastic fracture, where the concept of a failure zone is discussed and an algorithm for the finite element analysis of a fracture problem involving such a zone at the tip of a stationary crack is described. Again some results are presented.

1.2 Equilibrium Problem and Galerkin Approximation

We consider a two-dimensional solid defined in a region $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega \equiv \partial\Omega_C \cup \partial\Omega_T$. The displacement at any point $\mathbf{x} \equiv (x_1,x_2)^T$ of Ω (the reference configuration) is denoted by $\mathbf{u} \equiv (u_1,u_2)^T$, whilst the stress and strain tensor components are denoted respectively by σ_{ij} and ϵ_{ij}. The

deformation of the body under the action of external forces $\mathbf{f} \equiv (f_1, f_2)^T$ and boundary tractions $\mathbf{g} \equiv (g_1, g_2)^T$ satisfies the equilibrium equation

$$(1.1) \qquad \sum_{j=1}^{2} \frac{\partial \sigma_{ij}(\mathbf{x})}{\partial x_j} + f_i(\mathbf{x}) = 0 \ , \ \mathbf{x} \in \Omega \ , \ i = 1,2 \ ,$$

together with boundary conditions

$$(1.2) \qquad \mathbf{u}(\mathbf{x}) = \mathbf{0} \ , \quad \mathbf{x} \in \partial\Omega_c$$

$$(1.3) \qquad \sum_{j=1}^{2} \sigma_{ij}(\mathbf{x}).n_j = g_i(\mathbf{x}) \ , \ \mathbf{x} \in \partial\Omega_T \ , \ i = 1,2 \ .$$

The finite element method is applied to problem (1.1)-(1.3) via a weak formulation and the Galerkin technique. This is obtained by first taking the scalar product of (1.1) with a test vector function $\mathbf{v} \in V$, where

$$V \equiv \left\{ \mathbf{v} : \mathbf{v} \in (H^1(\Omega))^2 \ , \ v_i\big|_{\partial\Omega_c} = 0 \ , \ i = 1,2 \right\}$$

is the space of admissible vectors, and then integrating by parts. Thus in the weak problem we seek $\mathbf{u} \in V$ such that

$$(1.4) \qquad \int_\Omega \boldsymbol{\sigma}^T.\boldsymbol{\varepsilon}(\mathbf{v})d\mathbf{x} - \int_\Omega \mathbf{f}^T.\mathbf{v}\ d\mathbf{x} - \int_{\partial\Omega_T} \mathbf{g}^T.\mathbf{v}\ ds = 0 \ , \quad \forall \ \ \mathbf{v} \in V \ ,$$

where in (1.4) the strain tensor components ε_{ij} are defined by

$$(1.5) \qquad \varepsilon_{ij}(\mathbf{v}) \equiv \frac{1}{2}\left[\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right] \ , \ i,j = 1,2 \ ,$$

the vectors $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ are given by $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \varepsilon_{22}, 2\varepsilon_{12})^T$, $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})^T$, and the displacement $\mathbf{u}$ is involved in (1.4) through $\boldsymbol{\sigma}$ via an appropriate constitutive relation.

For the application of the finite element method the region Ω is partitioned into elements $\Omega = \bigcup_e \Omega^e$. A finite dimensional space $S^h \subset V$ consisting of piecewise polynomial functions defined over the partition is set up and the Galerkin problem approximating (1.4) is that of finding $\mathbf{u}_h \in S^h$ such that

$$(1.6) \qquad \int_\Omega \boldsymbol{\sigma}_h^T.\boldsymbol{\varepsilon}(\mathbf{v}_h)d\mathbf{x} - \int_\Omega \mathbf{f}^T.\mathbf{v}_h\ d\mathbf{x} - \int_{\partial\Omega_T} \mathbf{g}^T.\mathbf{v}_h\ ds = 0 \quad \forall\ \mathbf{v}_h \in S^h$$

where, in a similar manner to (1.4), the approximation $\mathbf{u}_h$ to the displacement $\mathbf{u}$ is involved in (1.6) through $\boldsymbol{\sigma}_h$ via an approximation to an appropriate constitutive relation.

Each component of the approximating vector $\mathbf{u}_h(\mathbf{x})$ is defined in terms of basis functions $N_k(\mathbf{x})$ for the n nodes of Ω so that, in terms of point evaluations U_k of $\mathbf{u}_h(\mathbf{x})$ at the nodes $k = 1,2,\ldots,n$,

(1.7) $$u_h(\mathbf{x}) = \mathbf{N}(\mathbf{x})\mathbf{U} ,$$

where $\mathbf{N}(\mathbf{x}) = [\mathbf{N}_1(\mathbf{x}),\mathbf{N}_2(\mathbf{x}),\ldots,\mathbf{N}_n(\mathbf{x})]$ with $\mathbf{N}_k(\mathbf{x}) = N_k(x)\mathbf{I}_2$ and $\mathbf{I}_2$ is the 2×2 unit matrix. If we define the approximate strain

(1.8) $$\boldsymbol{\varepsilon}_h \equiv \boldsymbol{\varepsilon}(u_h(\mathbf{x})) \equiv \mathbf{B}\,\mathbf{U} ,$$

in the usual way, see e.g. Zienkiewicz [13], we now require a constitutive relation between $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$.

For the case of an isotropic material and *linear elasticity* this is

(1.9) $$\boldsymbol{\sigma}(\mathbf{u}(\mathbf{x})) = \mathbf{D}\,\boldsymbol{\varepsilon}(\mathbf{u}(\mathbf{x}))$$

where $\mathbf{D}$ is the 3×3 matrix arising from Hooke's law. The matrix $\mathbf{D}$ depends on the Lamé coefficients λ and μ of the material. Using (1.9) with (1.6), and having taken $\mathbf{v}_h$ in turn to be each column of $\mathbf{N}$, we obtain the linear equation system

(1.10) $$\int_\Omega (\mathbf{B}^T\,\mathbf{D}\,\mathbf{B}\,d\mathbf{x})\,\mathbf{U} - \int_\Omega \mathbf{f}^T.\mathbf{N}\,d\mathbf{x} - \int_{\partial\Omega^T} \mathbf{g}^T.\mathbf{N}\,ds = \mathbf{0} ,$$

which when solved produces $\mathbf{U}$ and hence $\mathbf{u}_h(\mathbf{x})$.

2. ELASTO-PLASTIC PROBLEM AND NONLINEAR FRACTURE

2.1 Elasto-plastic Mathematical Model

In order to provide a mathematical model for the case where the material of the solid exhibits an elasto-plastic response, we have to set up a model of the constitutive relationship between stress and strain appropriate to the nonlinear post-yield plastic case.

We adopt here the incremental (flow theory) of plasticity and apply the loading incrementally. Thus in (1.1)-(1.3) we consider increments $d\boldsymbol{\sigma}$, $d\boldsymbol{\varepsilon}$, $d\mathbf{u}$, respectively of stress, strain and displacement, which result from increments of loading $d\mathbf{f}$ and $d\mathbf{g}$. The displacement $\mathbf{u}$ is now a function not only of space but also of the current load. We therefore introduce a load factor t (fraction of total load), so that $\mathbf{u} = \mathbf{u}(\mathbf{x},t)$. In the usual manner, see e.g. Owen and Hinton [7], Harrison, Ward and Whiteman [2], the level of stress at which plastic deformation takes place is determined by a yield criterion, based on a yield function F,

(2.1) $$F(\boldsymbol{\sigma},k) = f(\boldsymbol{\sigma}) - k \leq 0 ,$$

where f is the equivalent stress function and k varies during plastic deformation so that $k = f(\boldsymbol{\sigma})$ and $F(\boldsymbol{\sigma},k) = 0$. For any load increment, after initial yielding, it is assumed that the increment of strain can be written as the sum of elastic and plastic components so that

(2.2) $$d\boldsymbol{\varepsilon} = d\boldsymbol{\varepsilon}_e + d\boldsymbol{\varepsilon}_p ,$$

where $d\boldsymbol{\varepsilon}_e$ is related to $d\boldsymbol{\sigma}$ by the $\mathbf{D}$ matrix of (1.9). The plastic flow of the material is governed by a flow rule which, for associative plasticity,

relates the increment of plastic strain to the gradient of the yield function, so that

$$d\boldsymbol{\varepsilon}_p = d\lambda \frac{\partial F}{\partial \boldsymbol{\sigma}} , \tag{2.3}$$

where $d\lambda$ is the plastic multiplier. It is related to the k of (2.2) through a hardening rule, $d\lambda = A dk$. When a state of plastic flow exists stresses must remain on the yield surface so that

$$dF = \left(\frac{\partial F}{\partial \boldsymbol{\sigma}}\right)^T d\boldsymbol{\sigma} + \frac{\partial F}{\partial k} dk = 0$$

and hence

$$\mathbf{a}^T d\boldsymbol{\sigma} - A d\lambda = 0 , \tag{2.4}$$

where the flow vector $\mathbf{a}$ is defined by

$$\mathbf{a} = \frac{\partial F}{\partial \boldsymbol{\sigma}} .$$

We thus obtain from (2.1)-(2.4) the relation

$$d\boldsymbol{\varepsilon} = \left(\mathbf{D}^{-1} + \frac{\mathbf{a}\mathbf{a}^T}{A}\right) d\boldsymbol{\sigma} ,$$

from which we obtain

$$d\boldsymbol{\sigma} = \left(\mathbf{D} - \frac{\mathbf{D}\mathbf{a}\mathbf{a}^T\mathbf{D}}{A + \mathbf{a}^T\mathbf{D}\mathbf{a}}\right) d\boldsymbol{\varepsilon} \tag{2.5}$$

$$\equiv \mathbf{D}_{ep} d\boldsymbol{\varepsilon} .$$

where $\mathbf{D}_{ep} \equiv \mathbf{D}_{ep}(\boldsymbol{\sigma},k)$ is the elasto-plastic constitutive matrix. Thus, for the load increment in the post yield state, we have a nonlinear constitutive relation.

If the Galerkin technique is applied to the elasto-plastic problem in a specific incremental load step, the resulting (*nonlinear*) global equation system corresponding to (1.10) is

$$\int_\Omega (\mathbf{B}^T \mathbf{D}_p \mathbf{B}\, d\mathbf{x})\, d\mathbf{U} - \int_\Omega d\mathbf{f}^T.\mathbf{N}\, d\mathbf{x} - \int_{\partial\Omega_T} d\mathbf{g}^T.\mathbf{N}\, ds = \mathbf{0} , \tag{2.6}$$

where $\mathbf{D}_p \equiv \mathbf{D}_{ep}$ during yielding and $\mathbf{D}_p \equiv \mathbf{D}$ otherwise. When yield takes place the system (2.6) is nonlinear and is solved iteratively within the load step; the two most used methods for this are the *initial stiffness* and the *tangent stiffness* methods. If we define

$$\mathbf{K}(\boldsymbol{\sigma},k) \equiv \int_\Omega \mathbf{B}^T \mathbf{D}_p \mathbf{B}\, d\mathbf{x} , \tag{2.7}$$

then the iteration for solving (2.6), with general iteration step i, is as follows:

Step 1 Set $i = 1$, $d\mathbf{U}_1 = \mathbf{0}$, and take $\boldsymbol{\sigma}_1$ and k_1 to be their final values from the previous load step. Define

$$\mathbf{R}_1^T \equiv \int_\Omega d\mathbf{f}^T.\mathbf{N}\, d\mathbf{x} + \int_{\partial\Omega^T} d\mathbf{g}^T.\mathbf{N}\, ds \ .$$

Step 2 Set $d\mathbf{U}_{i+1} = d\mathbf{U}_i + \mathbf{H}^{-1}\mathbf{R}_i$.

Calculate $\boldsymbol{\sigma}_{i+1}$, k_{i+1}, A_{i+1} and $\mathbf{R}_{i+1} \equiv \mathbf{R}_1 - \mathbf{K}(\boldsymbol{\sigma}_{i+1}, k_{i+1})d\mathbf{U}_{i+1}$.

Step 3 For some tolerance ε if $\left|\mathbf{R}_{i+1}\right| > \varepsilon\left|\mathbf{R}_1\right|$ then set $i := i+1$ and repeat 2. Otherwise set $d\mathbf{U} := d\mathbf{U}_{i+1}$ and stop the iteration.

In Step 2 the matrix H can be taken as $\mathbf{K}(\mathbf{0},0)$ giving the initial stiffness method, or as the matrix $\mathbf{K}(\boldsymbol{\sigma}_i, k_i)$ from the previous iteration step giving the tangent stiffness method. The values of $\boldsymbol{\sigma}_{i+1}$ and k_{i+1} are calculated using (2.5), (1.5) and the hardening rule. Currently this initial value problem is integrated using the explicit forward Euler scheme. The value obtained is then scaled to lie on the yield surface $F(\boldsymbol{\sigma},k) = 0$.

2.2 J-Integral for Elasto-Plastic Fracture

The path independent J-integral of Rice [8] can be used in linear elastic fracture to obtain the stress intensity factor. For a Mode I problem with a crack having faces parallel to the x_1-axis, J is defined as, see [8],

$$J \equiv \int_\Gamma \left[W\, dx_2 - \sum_{i=1}^{2} T_i \frac{\partial u_i}{\partial x_1}\, ds \right] , \tag{2.8}$$

where Γ is a contour running anticlockwise from the lower to the upper crack faces enclosing the crack tip, W is the strain energy density, the T_i are tractions in the outward normal direction to Γ and ds is the increment of arc length.

The application of a similar J_p-integral in elasto-plastic fracture is motivated by the work of Huchinson [3], Rice and Rosengren [9] on the forms of near tip HRR stress and strain fields in power-law hardening materials based on the ***deformation*** theory of plasticity. These forms indicate that J_p plays in elasto-plastic fracture the same role as that of J in linear elastic fracture. However, as the mathematical theory of plasticity developed above is for incremental plasticity, our use of J_p in this context needs some justification. The basis for this is that under monotonic loading conditions incremental plasticity and deformation plasticity produce similar results, so that a secondary quantity J_p derived from either model will also be similar.

The J_p integral is calculated using (2.8) and noting that W now has both elastic and plastic components W_e and W_p so that

$$W = W_e + W_p \tag{2.9}$$

with

$$W_e \equiv \frac{1}{2}\,\sigma^T \varepsilon_e \ ,$$

and

$$W_p \equiv \int_0^{\lambda} f(\sigma)d\lambda \ ;$$

note that with the von Mises yield criterion ($f(\sigma) \equiv \left(\frac{3}{2}\sigma_{ij}\sigma_{ij} - \frac{1}{2}\sigma_{ii}\sigma_{jj}\right)^{\frac{1}{2}}$) $W_p = \int_0^{\bar{\varepsilon}_p} \bar{\sigma}\, d\bar{\varepsilon}_p$, where $\bar{\sigma}$ and $\bar{\varepsilon}_p$ are respectively the effective stress and effective plastic strain, see [6].

The method of approximating $(J_p)^{(k)}$ in the k^{th} load step of the deformation, with $(J_p)_h^{(k)}$ calculated from finite element approximations $\mathbf{u}_h(\mathbf{x})$, $\sigma_h(\mathbf{x})$ and $\varepsilon_h(\mathbf{x})$, derived as in Section 1.2, is a straightforward discretisation of (2.8) using calculated values at the Gauss quadrature points in the numerical integration and the splitting (2.9).

2.3 Mode I Elasto-Plastic Fracture Problem

A two-dimensional plane stress Mode I elasto-plastic fracture problem with centre crack, see Fig. 1, has been modelled using the techniques of Sections 1.2 and 1.3, assuming a von Mises yield condition.

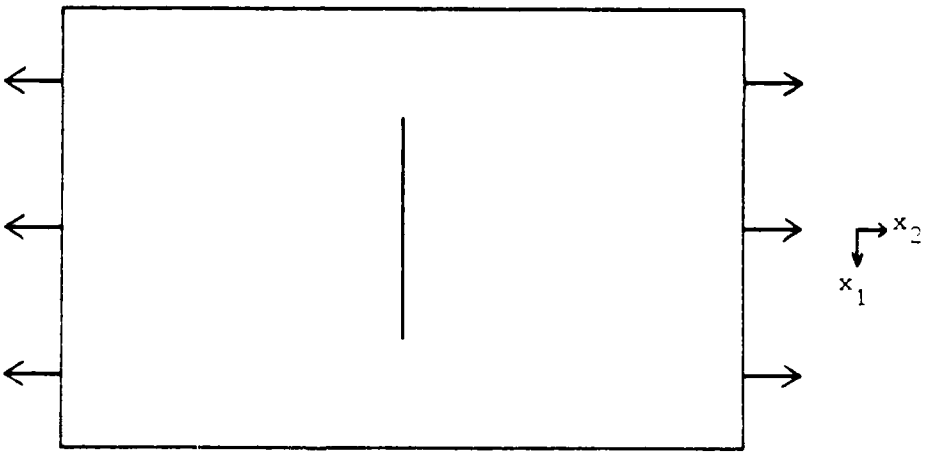


Fig. 1

This model problem has been treated by Owen and Fawkes [6]. An elastic perfectly plastic material has been assumed so that in (2.4), and subsequently, A = 0. The width of the region is 2/5ths of the length and the crack length is 2/5ths of the width, there is a normal tensile load of 100 units on each end, the Young's modulus E = 10,000, Poisson's ratio ν = 0.3 and the uniaxial yield stress σ_y = 100.

A basic finite element mesh based on 8-node quadrilateral elements is put over the quarter region as shown in Fig. 2; this mesh is refined locally around the crack trip to investigate the effect that this has on near crack tip $(J_p)_h$ values. Contours Γ surrounding the crack tip are then defined, the top half of some of these is shown in Fig. 3 and they pass

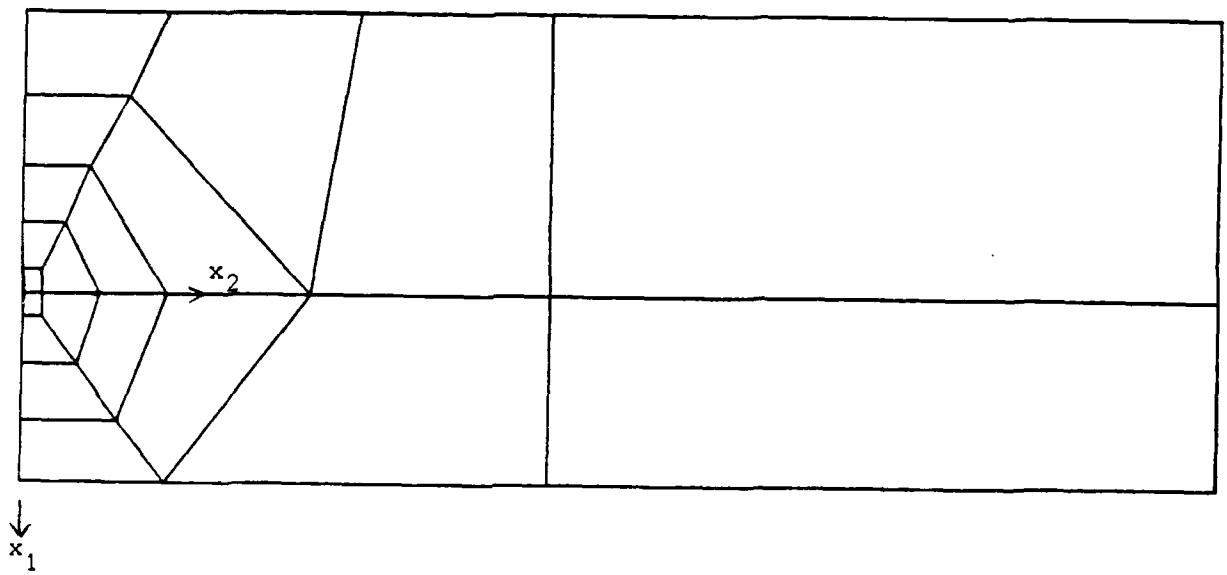


Fig. 2

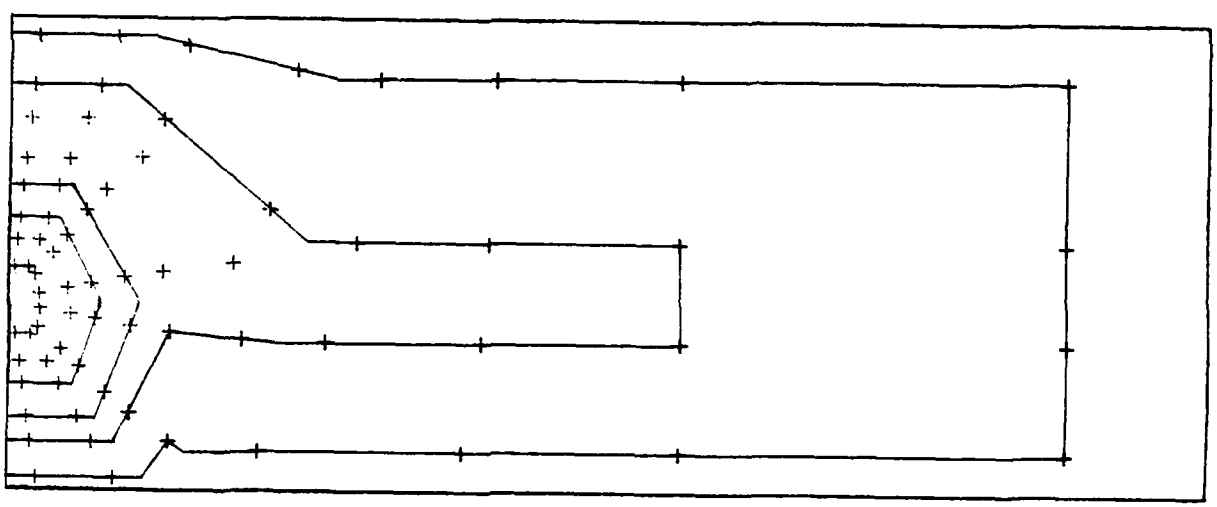


Fig. 3

through the 2 × 2 Gauss points for each element.

It is found that the calculated values $(J_p)_h^{(k)}$ of the $J_p^{(k)}$ integral are the same, within each load step, irrespective of whether the contours pass through zones of elastic or plastic deformation. The only exceptions for any attempted level of refinement are those calculated on contours passing through elements which have the crack tip as a node.

We conjecture that this effect is due to the poor accuracy of the approximating $\mathbf{u}_h$ in these elements at the crack tip; it should be emphasised that in these elements no effort has been made to model the form of the singularity. In this near-tip "failure zone" it is believed that constitutive laws of the above type break down, so that further modelling is necessary to represent the behaviour in this zone. We consider that outside the failure zone the calculated $(J_p)_h$ being virtually constant, are good values and thus can be used in a fracture criterion.

3. VISCOELASTIC FRACTURE WITH A FAILURE ZONE

3.1 Viscoelastic Model and Finite Element Discretization

We consider now viscoelastic materials which have the property that the displacement $\mathbf{u} = \mathbf{u}(\mathbf{x},t)$ at point $\mathbf{x} \in \Omega$ and time t depends on the previous history at that point; i.e. $\mathbf{u}(\mathbf{x},\tau)$, $\tau < t$. The weak problem at each time t relating to equilibrium equations (1.1)-(1.3) for a general viscoelastic material is then

$$(3.1)\quad \int_\Omega \boldsymbol{\sigma}(\mathbf{u}(\mathbf{x},\tau);\tau\le t)^T\boldsymbol{\varepsilon}(\mathbf{v})d\mathbf{x} - \int_\Omega \mathbf{f}(t)^T\mathbf{v}d\mathbf{x} - \int_{\partial\Omega_T} \mathbf{g}(t)^T\mathbf{v}ds = 0 \ , \ \forall \ \mathbf{v} \in V \ ,$$

where the test space V remains as in Section 1.2, i.e. $\mathbf{v} \in V$ does not involve time. We limit discussion here to linear viscoelastic materials in which the constitutive equation has the form

$$(3.2)\qquad \boldsymbol{\sigma}(\mathbf{x},t) = \int_{-\infty}^{t} \mathbf{D}(t-\tau)\dot{\boldsymbol{\varepsilon}}(\mathbf{x},\tau)d\tau \ ,$$

where $\mathbf{D}$ is the stress relaxation matrix. We further restrict discussion to problems for which there is no deformation for time $\tau < 0$, i.e. $\mathbf{u}(\mathbf{x},\tau) = \mathbf{0}$, $\tau < 0$ which implies that $\boldsymbol{\varepsilon}(\mathbf{x},\tau) = 0$, $\tau < 0$ and thus the lower limit of the time integral in (3.2) can be replaced by 0.

In discretising (3.1)-(3.2) using the Galerkin technique at time t we approximate $\mathbf{u}(\mathbf{x},t)$ by $\mathbf{u}_h(\mathbf{x},t)$ where

$$(3.3)\qquad \mathbf{u}_h(\mathbf{x},t) = \mathbf{N}(\mathbf{x})\mathbf{U}(t) \ ,$$

with $\mathbf{U}(t)$ denoting the vector of displacement nodal variables at time t. Thus the approximate stress is given by

$$(3.4)\qquad \boldsymbol{\sigma}_h(\mathbf{x},t) = \int_0^t \mathbf{D}(t-\tau)\mathbf{B}\,\dot{\mathbf{U}}(\tau)d\tau \ ,$$

and the discrete form of (3.1) is, at time t, the linear integro-differential equation system

$$(3.5)\qquad \int_0^t \left(\int_\Omega \mathbf{B}^T \mathbf{D}(t-\tau)\mathbf{B}\,d\mathbf{x} \right) \dot{\mathbf{U}}(\tau)\,d\tau - \int_\Omega \mathbf{f}(t)^T \mathbf{N}\,d\mathbf{x} - \int_{\partial\Omega_T} \mathbf{g}(t)^T \mathbf{N}\,ds = 0 \ .$$

We discretise (3.5) in time by taking time levels t_j, $j = 0,1,2,\ldots$ and for $t_{j-1} < \tau < t$ approximating $\dot{\mathbf{U}}(\tau)$ by $(\mathbf{U}(t_j)-\mathbf{U}(t_{j-1}))/(t_j-t_{j-1})$. This gives

$$(3.6)\qquad \left(\int_\Omega \mathbf{B}^T \bar{\mathbf{D}}\, \mathbf{B}\, d\mathbf{x} \right) (\mathbf{U}(t_j)-\mathbf{U}(t_{j-1})) = \int_\Omega \mathbf{f}(t)^T \mathbf{N}\, d\mathbf{x} + \int_{\partial\Omega} \mathbf{g}(t)^T \mathbf{N} ds + \mathbf{P}(t_j)$$

where

$$\bar{\mathbf{D}} = \left(\int_{t_{j-1}}^{t_j} \mathbf{D}(t_j-\tau)\,d\tau \right) / (t_j-t_{j-1})$$

and

$$\mathbf{P}(t_j) = \sum_{q=1}^{j-1} \int_\Omega \mathbf{B}^T \left(\int_{t_{q-1}}^{t_q} \mathbf{D}(t_j-\tau)\,d\tau / (t_q-t_{q-1}) \right) \mathbf{B}\, d\mathbf{x} (\mathbf{U}(t_q)-\mathbf{U}(t_{q-1})) \ .$$

For a general form of $\mathbf{D}$ the computation of $\mathbf{P}(t_j)$ involves the solution at all previous time steps. We however restrict attention to materials with constant Poisson's ratio and matrices $\mathbf{D}$ which can be expressed in terms of decaying exponentials, i.e.

$$(3.7)\qquad \mathbf{D}(s) = \varphi(s)\mathbf{D}(0) \ , \ \varphi(s) = \sum_1^M C_k e^{-\alpha_k s} \ , \ \alpha_k \geq 0 \ ,$$

where φ is the stress relaxation function. In this way, at each time level t_j, $\mathbf{P}(t_j)$ can be computed from the M vectors $\boldsymbol{\gamma}(\alpha_k, t_{j-1})$, $k = 1(1)M$ where $\boldsymbol{\gamma}(.,.)$ is defined by

$$(3.8)\qquad \boldsymbol{\gamma}(\alpha, t_q) = \int_0^{t_q} e^{-\alpha(t_q-\tau)} \dot{\mathbf{U}}(\tau)\,d\tau \ ,$$

This greatly simplifies the numerical algorithm. Also, because of the assumption of constant Poisson's ratio, the problem is in a form where correspondence principles can be used, see e.g. Schapery [11].

3.2 Viscoelastic Fracture

For a cracked viscoelastic body we now consider the choice of realistic fracture criteria for the onset of crack propagation and the approximation of such criteria using the numerical scheme outlined in Section 3.1. For viscoelastic materials this involves the introduction of a Barenblatt type failure zone about the crack tip, see Barenblatt [1], as was first considered by Knauss [5] and Schapery [10]. This arises in order to attempt to describe the physics of the process about the crack tip, i.e. to model the cohesive forces and the region of localised damage which occurs about the tip, and because criteria which are not based on such a failure zone are found, in the quasi static viscoelastic case, to give crack growth predictions which are greatly at variance with experimental observation.

In our model the Barenblatt failure zone is mathematically a small interval of length a_f behind the crack tip on which cohesive stresses σ_f

act in order to cancel the stress singularity produced at the crack tip by the external loads on the body. We assume that these cohesive stresses are constant on each crack face as indicated in Fig. 4. A physically motivated

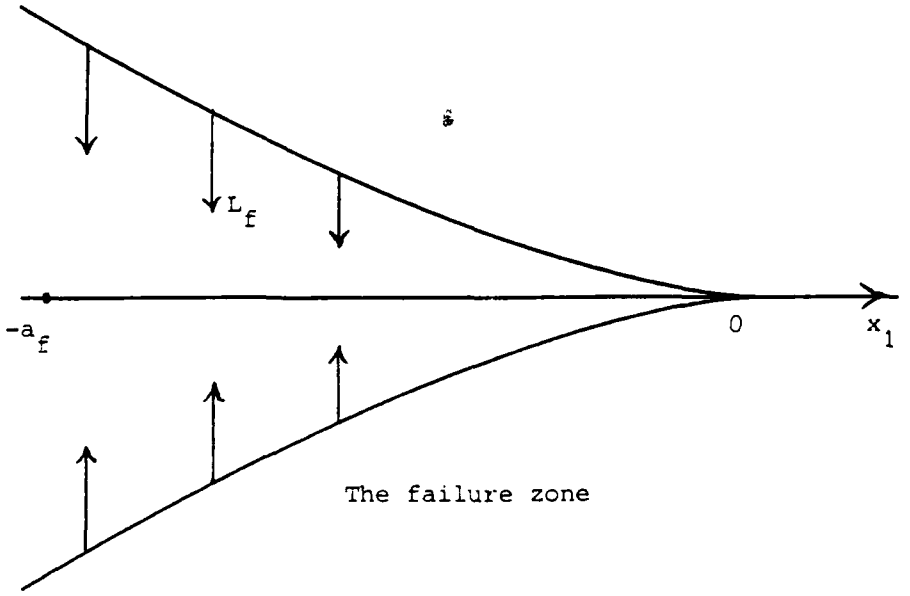


Fig. 4

fracture criterion that we consider is then based upon a critical value for the work input to the failure zone, i.e. when the work done by the cohesive stresses in this zone exceeds the critical value we predict that the crack will propagate. Since this work is given by $2L_f u_2(-a_f,0+,t)$ for the problem indicated in Fig. 4, the criterion can equivalently be expressed in terms of a critical crack opening displacement (COD). In the case of an elastic material this criterion is equivalent to the traditional stress intensity factory based criterion, see e.g. Kanninen and Popelar [4,p63], but is qualitatively and quantitatively a different criterion in the viscoelastic case.

There are several numerical difficulties arising from the inclusion of the failure zone. Firstly, because the zone is *small* relative to the rest of the domain, local refinement of the finite element mesh is required about the tip. Also, the equation required to determine a_f in order to cancel the stress singularity is non-linear and thus an iterative scheme must be developed. The algorithm for doing this will be described fully in Walton, Warby and Whiteman [12]. We give here only a brief description of the method in the context of a Mode I fracture problem in a region and with external loading as in Fig. 1, the load $L_e(t)$ now being time dependent. The constraint boundary conditions are not time dependent. As indicated in the previous discussion we also assume that loads of magnitude L_f act normal to the crack faces in the failure zone as in Fig. 4.

In order to calculate a_f we make use of the linearity of the model in terms of displacement and consider two Mode I problems of the form as above, (i) with only the external loading $L_e(t)$, (ii) with only the failure zone loading L_f. If $K_e = K_e(t)$ and $K_f = K_f(a_f(t))$ denote respectively the stress intensity factors of these problems then an

equation that we can use to determine $a_f = a_f(t)$ is given by

$$K_e(t) + K_f(a_f(t)) = 0 \ . \tag{3.9}$$

To determine K_e and K_f we use a correspondence principle of Schapery [11] which relates the solution of the viscoelastic problem to the solution of reference elastic problems with the same geometry, loadings and boundary conditions. More specifically it proves that the stress in the viscoelastic case is the same as the stress in the reference elastic problem at t. Hence, since a J-integral of the form (1.18) is related to the square of the stress intensity factor in the elastic case, we can replace (3.9) by

$$J_e^R(t) - J_f^R(a_f,t) = 0 \ , \tag{3.10}$$

where J_e^R and J_f^R denote respectively the J-integrals due to loads L_e and L_f applied separately. To approximate (3.10) numerically, there are two roughly equivalent approaches, either we solve the reference elastic problem at each time step and calculate the viscoelastic displacement, when required, by a time convolution with the creep function, or we solve the viscoelastic problem as in Section 3.1 and calculate the reference elastic displacement required in (3.10) by a time convolution with the relaxation function. We adopt the latter approach here. Our numerical algorithm then involves the following: For $t = t_j$, $j = 1,2,\ldots$ solve (3.10) for $a_f = a_f(t_j)$ by Newton's method, where each step of Newton's method involves the solution of the finite element equations. Then evaluate the crack opening displacement (COD) given by $u_2(-a_f,0+,t_j)$ and compare with the critical value. If the COD exceeds the critical value then determine a_f and t_{cr}, $t_{j-1} < t_{cr} < t_j$ by Newton's method so that (3.10) and

$$u_2(-a_f,0+,t_{cr}) = \text{critical COD} \tag{3.11}$$

are satisfied. Our value of t_{cr} is then our prediction for the time at which the crack will propagate.

The above problem is solved with a normalised relaxation function $\varphi(t) = (1+9e^{-t})/10$ which corresponds to the normalised creep function $\psi(t) = 10 - 9e^{0.1t}$. We now non-dimensionalise the Lamé parameters λ_0 and μ_0 relating to D(0) of (3.7) by setting $\mu_0 = 1$ and taking Poisson's ratio $\nu = 0.49$ so that Young's modulus $E = 2.98$ and $\lambda_0 = \nu E/(1+\nu)(1-2\nu) = 49$. The failure load $L_f = 10^{-1}$ and the external load L_e is applied at the three different rates

$$L_e(t) = 10^{-3}t, \ 10^{-2}t \text{ and } 10^{-1}t \ .$$

We assume that Ω is a square of size 2 and consider a selection of crack lengths between 0.25 and 1.50. A typical mesh consisting of 8 noded quadrilaterals is shown in Fig. 5. The values of a_f, t_{cr} and $L_e(t_{cr})$ are given in Table 1. These show the dependence on loading rate. If a stress intensity factor criterion is instead used then this would give a crack propagation criterion independent of loading rate, since K_e^R depends only on L_e, and thus would ignore creep effects. Hence it is clear that there can be considerable differences between J-based and COD-based fracture criteria.

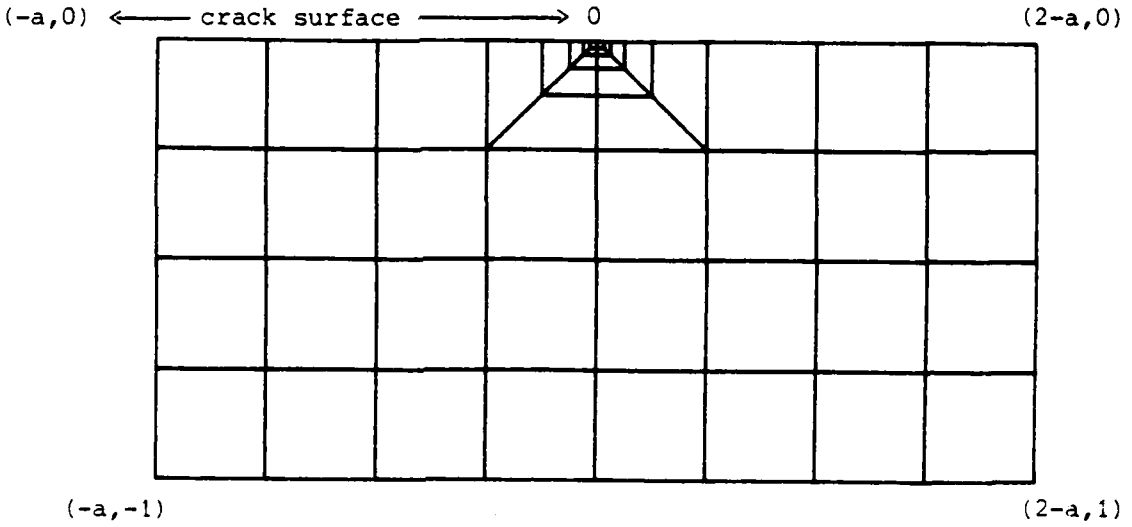


Fig. 5 Finite element mesh

Table 1

a = crack length, a_f = failure zone length
t_{cr} = time of critical COD (for the different loading rates)
$L_e(t_{cr}) = 10^{-3}t_{cr}$, $10^{-2}t_{cr}$ or $10^{-1}t_{cr}$ is the critical load
0.05 is the critical COD

a	a_f	t_{cr}	$L_e(t_{cr})$
0.25	0.138	6.245(-1)	6.245(-2)
	0.052	3.442	3.442(-2)
	0.022	2.044(1)	2.044(-2)
0.5	0.129	3.269(-1)	3.269(-2)
	0.065	2.131	2.131(-2)
	0.024	1.216(1)	1.216(-2)
1.0	0.127	1.309(-1)	1.309(-2)
	0.086	9.962(-1)	9.962(-3)
	0.035	5.860	5.860(-3)
1.5	0.120	4.293(-2)	4.293(-3)
	0.101	3.758(-1)	3.758(-3)
	0.055	2.408	2.408(-3)

Future Work

The next phase of this work is to simulate, using finite elements, the problem of a moving crack in a viscoelastic body. The aim of such work being to determine the conditions under which the crack will propagate stably or unstably. Details of this work are to be contained in [12].

References

1. Barenblatt, G.I., The mathematical theory of equilibrium cracks in brittle fracture. pp.55-129 of **Advances in Applied Mechanics, Vol.VII.** Academic Press, New York, 1962.
2. Harrison, D., Ward, T.J.W. and Whiteman, J.R., Finite element analysis of plates with nonlinear properties. **Comp. Meth. Appl. Mech. Eng. 37,** 1019-1034, 1982.
3. Hutchinson, J.W., Singular behaviour at the end of a tensile crack in a hardening material. **J. Mech. Phys. Solids 16,** 13-31, 1968.
4. Kanninen, M.F. and Popelar, C.H., **Advanced Fracture Mechanics.** Oxford University Press, 1985.
5. Knauss, W.G., On the steady propagation of a crack in a viscoelastic sheet: Experiments and analysis. pp.501-541 of W.H. Kausch and R. Jaffee (eds.), **Deformation and Fracture of High Polymers.** Plenum Press, London, 1973.
6. Owen, D.R.J. and Fawkes, A.J., **Fracture Mechanics.** Pineridge Press, Swansea, 1983.
7. Owen, D.R.J. and Hinton, E., **Finite Elements in Plasticity: Theory and Practice.** Pineridge Press, Swansea, 1980.
8. Rice, J.R., A path independent integral and the approximate analysis of strain concentration by notches and cracks. **J. Appl. Mech. 34,** 379-386, 1968.
9. Rice, J.R. and Rosengren, G.F., Plane strain deformation near a crack tip in a power law hardening material. **J. Mech. Phys. Solids 16,** 1-12, 1968.
10. Schapery, R.A., A theory of crack initiation and growth in viscoelastic media. I. Theoretical development. **Int. J. Frac. 11,** 141-159, 1975.
11. Schapery, R.A., Correspondence principles and a generalised J integral for large deformation and fracture analysis of viscoelastic media. **Int. J. Frac. 25,** 195-223, 1984.
12. Walton, J.R., Warby, M.K. and Whiteman, J.R., A finite element model of crack growth in a finite body in the context of Mode I linear viscoelastic fracture. **BICOM Technical Report,** Brunel University, Uxbridge, England, 1988. (In preparation)
13. Zienkiewicz, O.C., **The Finite Element Method, 3rd Edition.** McGraw-Hill, New York, 1977.

Anomalous Waves in Shock Wave-Fluid Interface Interactions

John W. Grove [1]

Courant Institute of Mathematical Sciences
New York University
New York, New York 10012

ABSTRACT

The interaction of a planar shock wave with a small amplitude fluid interface is characterized by the production of diffracted wave patterns that correspond to Galilean transforms of slowly varying perturbations of stationary waves. These waves are described by solutions to Riemann problems for the steady state Euler equations. When the amplitude of the interface is not small or the geometry of the two waves is changing, bifurcations in the solution occur. This article will analyze such a bifurcation for a shock wave in a dense material diffracting into a lighter material. The small amplitude case produces reflected Prandtl-Meyer rarefaction waves, and the bifurcations that occur at larger amplitudes can be interpreted as a two dimensional analogue of a rarefaction wave overtaking a shock. This analysis is incorporated into a front tracking code and provides a high quality description of the interacting waves.

1. Introduction

The interaction of a shock wave with a fluid interface can be subdivided into three regimes. These are a period of collision, a small amplitude linear growth regime, and the long time non-linear growth of instabilities in the interface.

The collision stage is characterized by the production of complicated diffracted wave patterns [Jahn, 1956; Abd-El-Fattah *et. al.*, 1976; Abd-El-Fattah & Henderson, 1978a; Abd-El-Fattah & Henderson, 1978b; Grove, 1989], and is extremely non-linear. In what is called the case of regular diffraction, these waves are perturbations of Galilean transforms of solutions to a Riemann problem for a steady flow with supersonic data. An asymptotic analysis [Grove, 1989] of the wave curves for the stationary Euler equations shows that the regular case occurs provided the angle between the two incoming waves is small. This observation means that the irregular wave patterns observed for larger amplitude interactions can be studied as two dimensional bifurcations of regular solutions.

1. Supported in part by the Army Research Office, grant DAAG29-85-0188.

The second period of the interaction [Richtmyer, 1960] corresponds to the growth of unstable modes in an impulsively loaded material interface that is a small amplitude perturbation of a planar interface. During this regime the flow is an approximate solution for the linearized Euler equations and the interface instabilities grow at an exponential rate.

The third stage concerns the long term growth of the unstable interface [Youngs, 1984; Mikaelian, 1986; Grove, 1989]. The flow here is characterized by the competition and mixing of unstable modes and leads to an eventual chaotic behavior of the material surface.

The main thrust of this article is to study the bifurcation behavior of a regular shock-contact diffraction pattern where the reflected wave is a Prandtl-Meyer rarefaction. Such waves occur in shock-fluid interface interactions where the shock is incident in the denser fluid. The bifurcation occurs when changes in the incident shock strength or in the geometries of the interacting curves cause the state behind the incident shock to become subsonic. When this happens the reflected Prandtl-Meyer wave overtakes the incident shock from behind. The overtaking of the incident shock by the Prandtl-Meyer wave both dampens and bends the incident shock, with corresponding effects in the transmitted shock wave. Two classes of examples of such interactions will be described; a planar shock wave in water diffracting through an air bubble, and a cylindrically expanding shock wave diffracting through a planar air-water interface. Previous work [Grove, 1989], considered the modeling using the method of front tracking of regular diffraction cases, and a major goal of this present work is to be able to track these waves beyond the "regular" regime. A first order analysis of this wave bifurcation is incorporated into the front tracking algorithm and gives an accurate description of the colliding waves throughout their interaction.

2. The Supersonic Steady State Riemann Problem

The equations of motion for an inviscid, non-heat conducting fluid are given by the well known Euler equations:

$$\partial_t \rho + \nabla\cdot(\rho\vec{q}) = 0, \tag{2.1.1}$$

$$\partial_t(\rho\vec{q}) + \nabla\cdot(\rho\vec{q}\otimes\vec{q}) + \nabla P = \rho\vec{g}, \tag{2.1.2}$$

$$\partial_t(\rho\mathcal{E}) + \nabla\cdot((\rho\mathcal{E} + P)\vec{q}) = \rho\vec{q}\cdot\vec{g}. \tag{2.1.3}$$

Here $\vec{q}$ is the particle velocity, ρ is the mass density, P is the thermodynamic pressure, $\vec{g}$ is the constant gravitational acceleration vector, $\mathcal{E} = \frac{|\vec{q}|^2}{2} + E$ is the specific total energy, and E is the specific internal energy. These equations express the conservation of mass, momentum, and energy respectively. Assuming non-reactive equilibrium thermodynamics, this system is closed by a thermodynamic equation of state

$$E = E(V, S) \tag{2.2}$$

where $E(V, S)$ is a convex function of the specific volume $V = 1/\rho$ and the specific entropy S. $E(V, S)$ satisfies the first law of thermodynamics:

$$TdS = dE + PdV, \tag{2.3}$$

where T is the absolute temperature. Equation (2.3) implies that

$$P = -\frac{\partial E}{\partial V}. \tag{2.4}$$

In practice S is eliminated from equations (2.2) and (2.4), and thermodynamics of the fluid is described by an incomplete equation of state

$$P = P(E, V). \tag{2.5}$$

The key to modeling shock-wave fluid interface interactions using a front tracking method is the notion of a two dimensional elementary wave [Glimm *et. al.*, 1985]. Elementary waves describe the downstream scattering of a pair of interacting waves, and are calculated by the solution of a Riemann problem for the steady state Euler equations, where the stream direction serves as a time like axis. Briefly, for a shock-contact interaction, the state behind the incoming shock wave and the state on the side of the material interface opposite to the incoming shock serve as data for a downstream directed Riemann problem. Since the actual collision of the two waves occurs over a short interval in time, gravity can be neglected in the analysis.

Restricting (2.1) to time independent planar flow with $\vec{g} = 0$, we obtain a system of four conservation laws. This system is hyperbolic for supersonic flow. For most single phase flows [Thompson, 1971], and certainly for the simple analytic equations of state used in the numerical simulations below, this system has two genuinely non-linear eigenvalues that correspond to the propagation of sound waves, and two linearly degenerate modes that travel with particle velocity.

Both the pressure and the polar velocity angle are partial Riemann invariants for the linearly degenerate fields, so the Riemann problem for this system can be solved by finding the intersection of the wave curves for the two genuinely non-linear fields in the pressure-flow angle phase plane. This is analogous to the solution to the Riemann problem for time dependent one dimensional flow [Menikoff, 1988], with some important differences. The Hugoniot locus portion of these wave curves are the well known shock polars [Courant & Friedrichs, 1948], and extend into the subsonic region where the flow ceases to be hyperbolic. The shock polars also form closed and bounded loops. These two facts lead to a loss of existence or uniqueness for the solution to the general steady state Riemann problem with supersonic data. The non-uniqueness is addressed by choosing the supersonic solution whenever it exists [Henderson, 1966]. It can be shown that under suitable conditions on the equation of state, at most one such solution is possible. Also a linear stability analysis [Henderson & Atkinson, 1976] shows that when both a supersonic and subsonic solution exist, only the supersonic solution is stable. The non-existence of the solution must be resolved by allowing it to become time dependent. This leads to the development of many irregular diffraction patterns [Jahn, 1956; Abd-El-Fattah &

Henderson, 1978a; Abd-El-Fattah & Henderson, 1978b]. The next section will describe one such irregular wave whose structure corresponds to the overtaking of an oblique shock wave by a Prandtl-Meyer wave.

3. The Anomalous Reflection Wave

Figure 3.1 illustrates the collision of a planar shock in water with an air bubble. When the shock first reaches the bubble, the two waves are tangent and regular diffraction patterns (called diffraction nodes) are produced at the points of collision between the two waves. To leading order, the flow near a point of diffraction is described by the solution to a supersonic steady state Riemann problem as mentioned above. Here, the interaction produces a reflected Prandtl-Meyer wave. Figure 3.2 shows the set of wave curves used for the solution shown in figure 3.2.b. It is important to note that the existence of such a solution depends on the flow being supersonic behind the incident shock wave in a reference frame where the node is at rest. Because of the large difference in the compressibility of the two fluids (for this model at constant temperature air is about 15,000 times as compressible as water) as long as the flow behind the incident shock remains supersonic, the waves produced by this interaction are prevented from interacting with the incident waves and the flow downstream from this point remains nearly self similar in a neighborhood of the node.

If β denotes the instantaneous angle between the incident shock and the bubble, then the Mach number behind the incident shock is given by

$$M_1 = \frac{m}{\rho_1 c_1}\left(1 + \frac{\rho_1^2}{\rho_0^2}\cot^2\beta\right)^{1/2}, \tag{3.1}$$

where the subscripts 0 and 1 refer to the states ahead and behind the incident shock respectively, and $m^2 = \frac{P_1 - P_0}{V_0 - V_1}$ is the mass flux across the shock. For most equations of state [Menikoff & Plohr, 1989], $m < \rho_1 c_1$, so the flow behind the shock will be subsonic if β is sufficiently close to $\frac{\pi}{2}$. Thus a transition into an irregular wave pattern occurs before the shock reaches the equator of the bubble.

When the flow behind the incident shock becomes subsonic, the leading edge of the reflected Prandtl-Meyer wave begins to overtake the incident shock wave from behind. This process will dampen the incident shock producing additional curvature in the incident shock wave. These effects will be transmitted into the outgoing waves as well. The portion of the reflected rarefaction that overtakes the shock in a given period of time is the amount that is enough to restore the flow immediately behind the point of collision of the two waves to a sonic flow. Since the sound speed in water is much higher than the sound speed in air, the flow in the air inside the bubble remains supersonic, and the transmitted wave continues to propagate downstream from the node. This prevents the formation of precursor type waves such as those described in [Abd-El-Fattah & Henderson, 1978b].

As the interaction proceeds and the bubble interface continues to diverge away from the incident shock, more and more of the rarefaction fan spreads out onto the incident shock wave, leading to the formation of a structure that is analogous to a Mach reflection with a non-centered reflected Prandtl-Meyer wave, see figure 3.3. Eventually, enough of the rarefaction overtakes the incident shock so that the flow near the trailing edge of the Prandtl-Meyer wave becomes nearly sonic. When this happens, the trailing edge of the rarefaction wave is almost parallel to the incident shock and to leading order the flow near the point of shock diffraction is a one dimensional unsteady flow with a rarefaction wave overtaking a shock wave from behind. Thus in finite time, the entire reflected rarefaction wave overtakes the incident shock. The Mach node corresponds to the spread out wave where the non-centered rarefaction meets the incident shock, and the Mach stem corresponds to the portion of the incident shock from the trailing edge of the rarefaction wave to the fluid interface. For weak incident shocks the reflected rarefaction is of about the same strength as the incident wave, and this "anomalous reflection stem" is a sound wave.

There is experimental evidence of these anomalous waves. In particular Jahn [Jahn, 1956] figure 14g shows such a wave for the oblique diffraction of a planar shock through a thin membrane separating two gases.

The structure of this anomalous wave will be described in more detail in a coming paper [Grove & Menikoff, 1988].

4. The Tracking of the Anomalous Reflection Wave

The qualitative discussion of the anomalous reflection in the previous section can be incorporated into a front tracking code to give an enhanced resolution of the interaction.

Previous work [Grove, 1989] described the tracking of a regular shock-contact diffraction node. When a shock-contact diffraction node is identified at a given time step with time increment dt, the pair of incoming interacting waves (the incident shock and material interface) are first propagated independently ignoring their interaction. The intersection between the two propagated waves is found and gives the time updated position p_0 of the diffraction node. The displacement of the node position divided by dt provides the node velocity and the Galilean transformation for the flow near the node into a frame where the node is stationary. If the state behind the incident shock is supersonic, it together with the state on the opposite side of the material interface provide data for a supersonic steady state Riemann problem, whose solution determines the outgoing waves. The outgoing tracked waves are then modified to incorporate this solution.

A bifurcation to an anomalous reflection is detected when the state behind the incident shock is subsonic in the frame of the node and the reflected wave from the previous time step is a Prandtl-Meyer wave. The first step in modeling this bifurcation is to propagate the leading edge of the reflected wave onto the incident shock. This is done as before, by finding the point of intersection p_1 between the two

propagated curves. If (as is often the case) the reflected wave is untracked, it is recovered by calculating the characteristic through the old node position corresponding to the state behind the incident shock. This characteristic makes the Mach angle $A_1 = \arcsin 1/M_1$ with the stream line through the node. It is assumed that the bifurcation occurs during the time step so the Mach number M_1 at the beginning of the time step is greater than or equal to one, and A is real. The leading edge of the reflected rarefaction moves with sound speed in the direction normal to the characteristic. If the leading edge of the Prandtl-Meyer wave is tracked, it is disconnected from the original diffraction node and a new node (called a cross node [Glimm *et. al.*, 1985]) corresponding to the oblique overtaking of a characteristic (zero strength shock wave) with a shock wave of the same family is installed at p_1.

The next step is to determine the states and position of the point of shock diffraction after the bifurcation. As the rarefaction expands onto the incident wave, the incident shock near the material interface is weakened and curves into the contact. The interaction slows down the node until the flow behind the incident shock at the node is sonic. Thus it suffices to compute a corrected node velocity or equivalently a corrected propagated node position that takes into account the shock-rarefaction interaction. Once this corrected position is determined, the flow downstream from the node is computed as for a regular diffraction.

For each number s, let $p(s)$ be the point on the propagated material interface that is located a distance s from p_0 when measured along the curve, the positive direction being oriented away from the node into the region ahead of the incident shock. Let $\beta(s)$ be the angle between the tangent vector to the material interface at $p(s)$ and the directed line segment between the points $p(s)$ and p_1. See figure 4.1. Let $\vec{v}(s)$ be the node velocity found by moving the diffraction node to position $p(s)$, and let $\vec{q}(s)$ be the velocity of the flow ahead of the incident shock in the frame that moves with velocity $\vec{v}(s)$ with respect to the computational lab frame. The mass flux across the incident shock that makes an angle $\beta(s)$ with the upstream material interface is given by

$$m(s) = \rho_0 |\vec{q}(s)| \sin\beta(s). \tag{4.1}$$

Given $m(s)$ and the state ahead of the incident shock, the state behind the shock and hence its Mach number $M(s)$ can be found. The new node position is given by $p(s^*)$, where

$$M(s^*) = 1. \tag{4.2}$$

Finally, the state behind the incident shock with mass flux $m(s^*)$ together with the state on the opposite side of the contact are used as data for a steady state Riemann problem whose solution supplies the states and angles of the transmitted shock, the trailing edge of the reflected rarefaction, and the downstream material interface.

The subsequent propagation of the anomalous reflection node is performed in the same way. The bifurcation simply repeats itself as more of the reflected rarefaction propagates up the incident shock. The leading edge of the reflected rarefaction

wave that connects to the diffraction node is not tracked after the first bifurcation.

The secondary bifurcations that occur when the trailing edge of the rarefaction overtakes the incident shock are detected in a couple of ways. If the incident shock is sufficiently weak, ie the normal shock Mach number is close to 1, then it is possible for the numerically calculated upstream Mach number to be less than one. Physically of course the state ahead of the incident shock is always supersonic, but if it is nearly sonic, such numerical undershoot may occur. When this happens, the construction described above must be modified. The tracked trailing rarefaction edge is disengaged from the diffraction node and installed in a new overtake node found by intersecting the propagated characteristic with the ahead shock. The residual shock strength for the portion of the incident shock behind the rarefaction wave is small. The diffraction node at the material interface reduces to the degenerate case of a sonic signal diffracting through a material interface, and the induced downstream waves are also sound waves. The second way in which the secondary bifurcation is detected occurs when the trailing edge of the rarefaction overtakes the shock. Here a new intersection between the incident shock and the trailing edge characteristic is produced. Again the tracked characteristic is disengaged from the diffraction node and a new overtake node is installed at the point of intersection. Here, the residual shock strength behind the rarefaction is positive. The diffraction at the material interface is non-trivial and will produce an additional expansion wave behind the original one. Most often this new expansion wave is not tracked.

Some remarks about the amount to tracking of these diffraction nodes seems to be pertinent at this point. The secondary bifurcations described in the previous paragraph need only be explicitly dealt with when the edges of the reflected Prandtl-Meyer wave are tracked. The current algorithm assumes that at a minimum the two interacting incoming waves are tracked. At this extreme none of the outgoing waves are tracked but are captured by the interior solver that is coupled to the front tracking method. In such a case the bifurcations occur automatically and the algorithm is much simpler. More commonly, the material interface separates different fluids so that the change in equation of state across this interface requires the tracking of the downstream portion of the material interface as well. Furthermore any capturing method will spread the captured wave over several grid zones thus reducing the resolution of the two dimensional wave. This spreading will be particularly pronounced at expansive waves such as Prandtl-Meyer waves [Glimm *et. al.*, 1987]. Also, unless the capturing algorithm is carefully designed, instabilities in the finite difference approximation can destroy the accuracy of the solution near the node. This is especially the case for stiff materials such as water. Tracking these waves seems to considerably reduce these problems. It also allows the use of much coarser grid, which is important when the diffraction occurs in a small but important zone of a larger simulation and the entire region of diffraction extends over only a fraction of a grid block. The point of these remarks is that the amount of tracking is problem dependent, and a compromise can be made between the increased accuracy and stability of front tracking, and the simplicity of a capturing algorithm.

5. Numerical Examples

Figure 5.1 shows a series of frames documenting the collision of a 10 Kbar shock wave with a bubble of air in water. The states ahead of the incident shock are at one atmosphere pressure and standard temperature. Under these conditions, water is about 1000 times as dense as air. During the initial stage of the interaction regular diffraction patterns are produced. By real time 4.5 μsec an anomalous reflection has formed, and by 10 μsec the trailing edge of the rarefaction has also overtaken the incident shock. It is interesting to note that this interaction causes the bubble to collapse into itself. Longer simulations (not available at the time of this writing) show the bubble splitting in two (in three dimensions it forms a torus) with the resulting production of vorticity. Very long time simulations are expected to show the bubbles going into oscillation as they are overcompressed and then expand. This over compression and expansion is important in the transfer of energy as a shock passes through a bubbly fluid. The first diffraction considerably dampens the shock, and some of the energy will eventually be returned to the shock wave in the form of compression waves generated by the expanding bubble. One goal of this research is to be able to perform simulations of such long term behavior that develop on time scales orders of magnitude greater than the shock diffraction itself.

Figure 5.2 shows the diffraction of an expanding underwater shock wave through the water's surface. The problem is initialized by placing at 2 meters below the water's surface the center of a 10 Kbar cylindrically expanding shock wave of radius 1 meter. Inside the cylindrical shock is a bubble of hot dense air. The initial conditions outside the expanding shock are ambient at one atmosphere pressure and normal temperature. In this simulation, the fluids are subject to a gravitational acceleration of $1g$. The reflected Prandtl-Meyer wave is untracked. The pressure contour plots show that by 6 msec an anomalous reflection has developed. Another interesting feature of this problem is the acceleration of the bubble inside the shock wave by the reflected rarefaction wave. This causes the bubble to rise much faster than it would under just gravity. As the bubble reaches the surface it begins to expand into the atmosphere. The expansion leads to the formation of a kink in the transmitted shock wave between the region ahead of the surfacing bubble, and the rest of the wave. This kink is an untracked version of another elementary wave (cross node) where two oblique shocks collide.

The two fluids in the simulations described above are modeled by what is called a stiffened polytropic equation of state [Harlow & Amsden, 1971; Plohr, 1988], where the pressure is given by

$$P = (\gamma - 1)\rho E - \gamma P_\infty. \tag{5.1}$$

If $P_\infty = 0$ this reduces to the more common polytropic equation of state. The values used for the EOS parameters are $\gamma_{air} = 1.4$, $P_\infty\big|_{air} = 0$, $\gamma_{water} = 7.0$, $P_\infty\big|_{water} = 3000$ atm.

6. Conclusion and Open Questions

The diffraction of a shock wave in water through an air-water interface produces an interesting wave that is an analogue of a Mach reflection with a non-centered rarefaction. A first order description of this wave can be incorporated into a front tracking algorithm to provide a high quality resolution of the wave on a given grid. This allows an accurate initialization of the shocked interface whose long term unstable structure can then be studied.

There are several interesting mathematical questions associated with the anomalous reflection wave discussed in this paper. One would be to provide an analytical asymptotic description of the interaction between the two waves. This would include decay estimates for the incident shock strength as it is overtaken by the Prandtl-Meyer wave. A related question would be to provide a higher order description of the interaction by taking into account the reflected waves produced when the rarefaction overtakes the shock. This would include an attempt to give a detailed description of the flow in the immediate wake of the anomalous reflection. This wake region bounded by the leading edge reflected sound wave behind the overtake node where the Prandtl-Meyer fan overtakes the shock, the anomalous "Mach" stem, and the material interface is analogous to the Mach bubble produced in ordinary Mach reflection.

The author would like to thank Dr. Ralph Menikoff for many helpful discussions of this work, and Prof. James Glimm for his encouragement and physical insight into these problems.

References

Abd-El-Fattah *et. al.*, 1976.

Abd-El-Fattah, A. M., L. F. Henderson, and A. Lozzi, "Precursor Shock Waves at a Slow-Fast Gas Interface," *J. Fluid Mech.*, vol. 76, p. 157, 1976.

Abd-El-Fattah & Henderson, 1978a.

Abd-El-Fattah, A. M. and L. F. Henderson, "Shock Waves at a Fast-Slow Gas Interface," *J. Fluid Mech.*, vol. 86, p. 15, 1978.

Abd-El-Fattah & Henderson, 1978b.

Abd-El-Fattah, A. M. and L. F. Henderson, "Shock Waves at a Slow-Fast Gas Interface," *J. Fluid Mech.*, vol. 89, p. 79, 1978.

Courant & Friedrichs, 1948.

Courant, R. and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, pp. 294-317, Springer Verlag, New York, 1948.

Glimm *et. al.*, 1985.

Glimm, J., C. Klingenberg, O. McBryan, B. Plohr, D. H. Sharp, and S. Yaniv, "Front Tracking and Two Dimensional Riemann Problems," *Adv. in Appl. Math.*, vol. 6, pp. 259-290, 1985.

Glimm *et. al.*, 1987.

Glimm, J., J. Grove, and X. L. Li, "Three Remarks on the Front Tracking Method," *Proceedings of the conference in Taormina Sicily*, 1987.

Grove & Menikoff, 1988.

Grove, J. W. and R. Menikoff, "The Anomalous Reflection of a Shock Wave through a Material Interface," *in preparation*, 1988.

Grove, 1989.

Grove, J. W., "The Interaction of Shock Waves with Fluid Interfaces," *Adv. Appl. Math.*, vol. To Appear, 1989.

Harlow & Amsden, 1971.

Harlow, F. H. and A. A. Amsden, *Fluid Dynamics*, LA-4700, Los Alamos National Laboratory, Los Alamos, 1971. Available from National Technical Information Service U.S. Dept of Commerce.

Henderson, 1966.

Henderson, L. F., "The Refraction of a Plane Shock Wave at a Gas Interface," *J. Fluid Mech.*, vol. 26, p. 607, 1966.

Henderson & Atkinson, 1976.

Henderson, L. F. and J. D. Atkinson, "Multi-valued Solution of Steady-State Supersonic Flow. Part 1. Linear Analysis," *J. Fluid Mech.*, vol. 75, pp. 751-764, 1976.

Jahn, 1956.

Jahn, R. G., "The Refraction of Shock Waves at a Gaseous Interface," *J. Fluid Mech.*, vol. 1, pp. 457-489, 1956.

Menikoff, 1988.

Menikoff, R., "Analogies Between Riemann Problem for 1-D Fluid Dynamics and 2-D Steady Supersonic Flow," *AMS preceedings*, 1988.

Menikoff & Plohr, 1989.

Menikoff, R. and B. Plohr, "Riemann Problem for Fluid Flow of Real Materials," *Rev. Mod. Phys.*, To Appear 1989.

Mikaelian, 1986.

Mikaelian, Karnig O., "Turbulent Energy at Accelerating and Shocked Interfaces," *Lawrence Livermore National Labratory Preprint UCRL-93977*, Jan. 10, 1986.

Plohr, 1988.

Plohr, B., "Shockless Acceleration of Thin Plates Modeled by a Tracked Random Choice Method," *AIAA J.*, 1988. To appear

Richtmyer, 1960.

Richtmyer, R. D., "Taylor Instability in Shock Acceleration of Compressible Fluids," *Comm. Pure and Appl. Math*, vol. 13, pp. 297-319, 1960.

Thompson, 1971.

Thompson, Philip A., "A Fundamental Derivative in Gasdynamics," *Phys.*

Fluids, vol. 14, no. 9, pp. 1843-1849, Sept. 1971.

Youngs, 1984.

Youngs, David L., "Numerical Simulation of Turbulent Mixing by Rayleigh-Taylor Instability," *Physica*, vol. 12D, pp. 32-34, 1984.

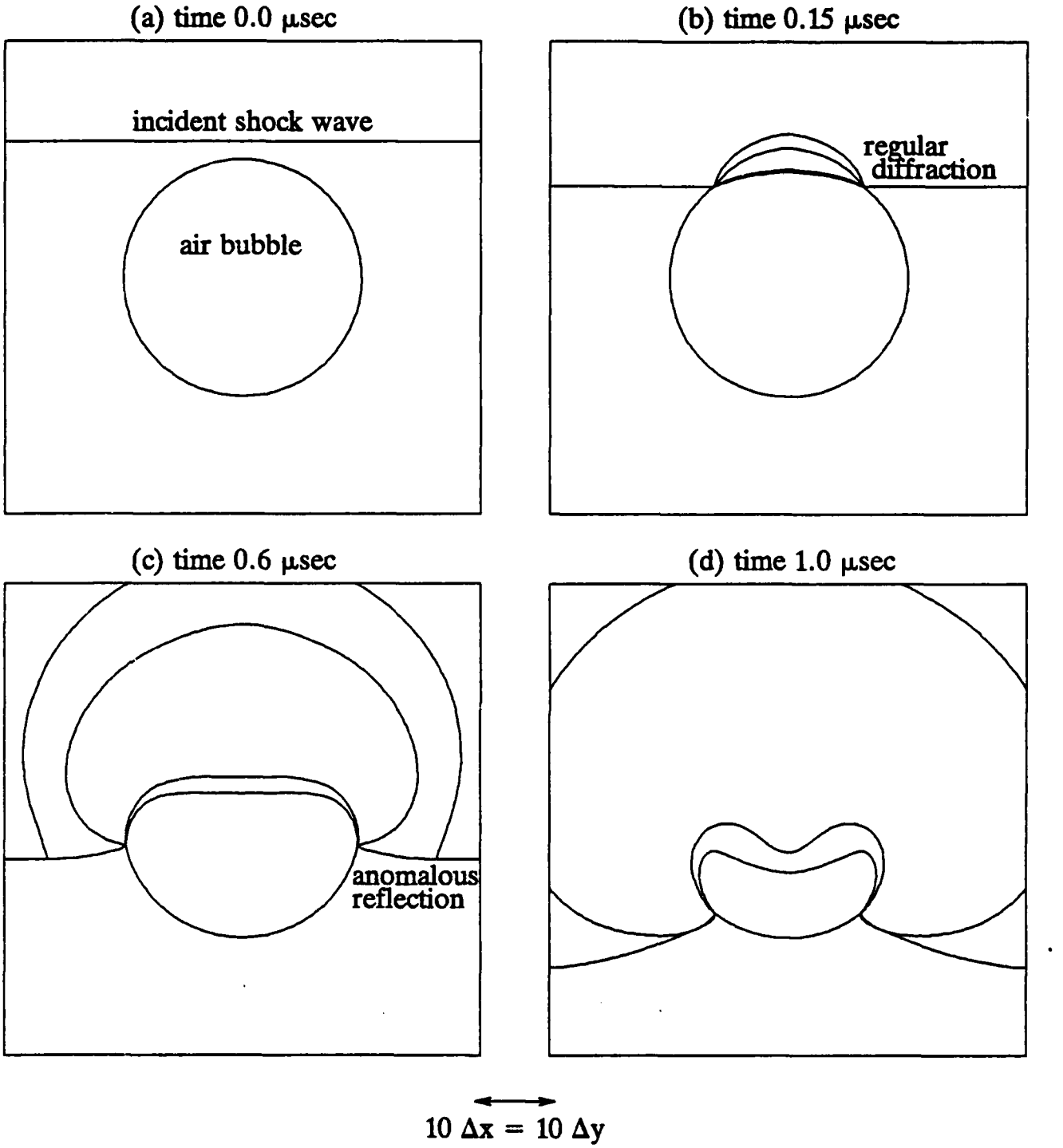


Figure 3.1. The collision of a shock wave in water with an air bubble. The fluids ahead of the shock are at normal conditions of 1 atm. pressure, with the density of water 1 g/cc and air 0.0012 g/cc. The pressure behind the incident shock is 10 Kbar with a shocked water density of 1.195 g/cc. The grid is 60x60.

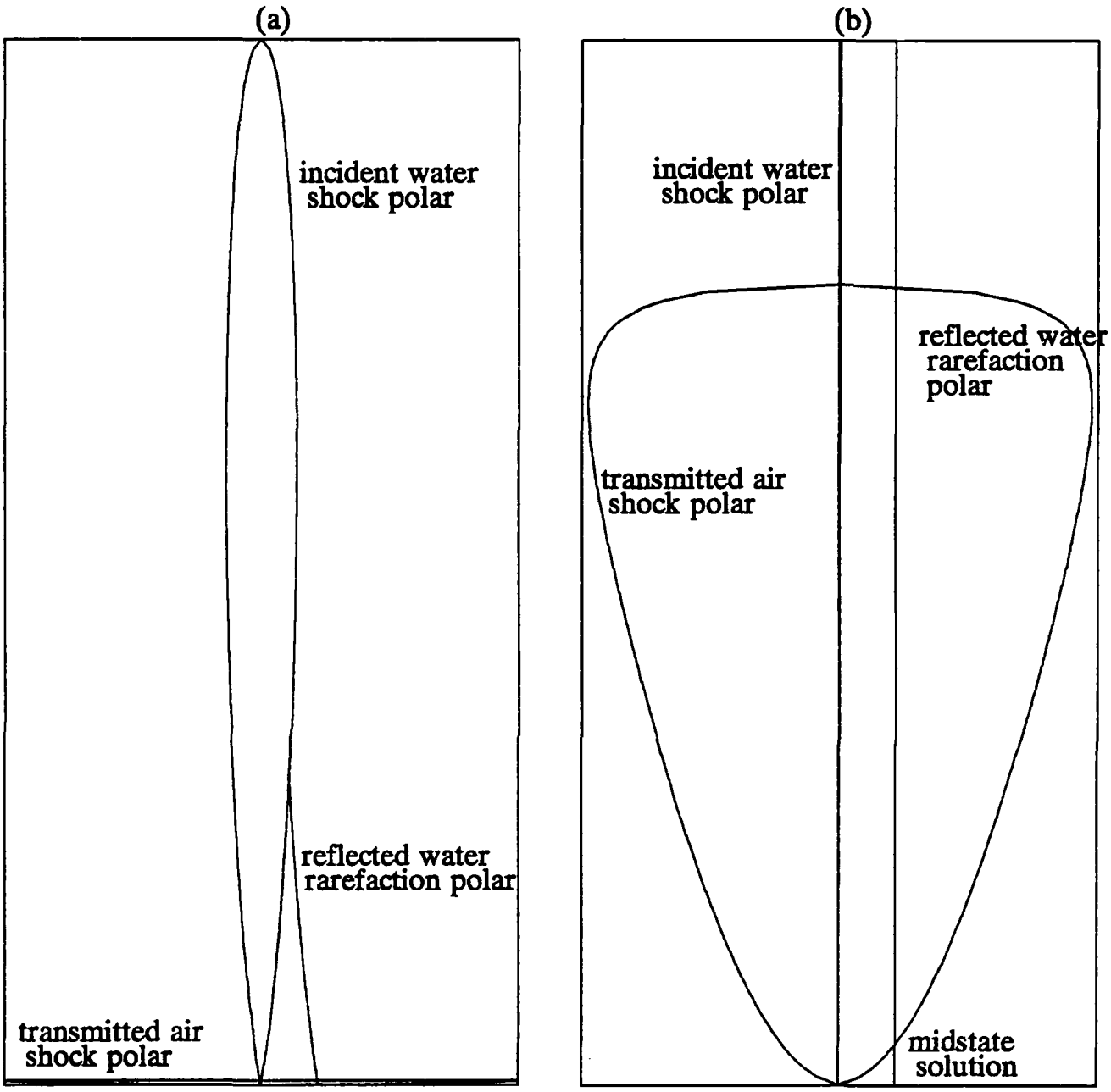


Figure 3.2. The wave curves (shock polars) used in the solution of the steady state Riemann problem in figure 3.1.b. Note that the shock polar for the transmitted shock in air is much lower and wider than the corresponding incident shock polar for water. Figure 3.2.b contains a detail from the lower section of figure 3.2.a that shows the air shock polar clearly. The Mach numbers for the incident and transmitted shocks are 2.7 and 11.4 respectively. The pressure at the midstate solution is 8.8 bars.

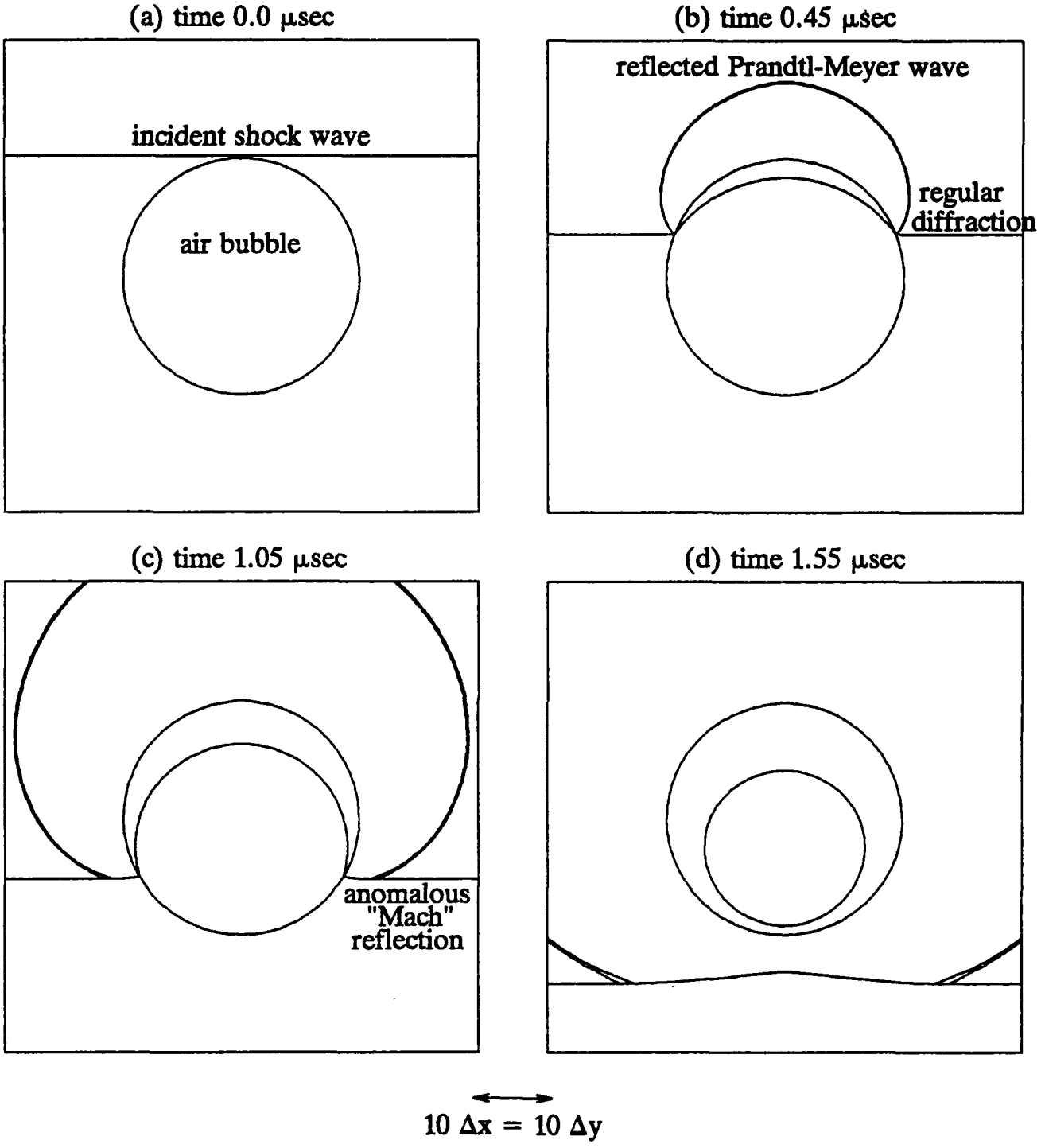


Figure 3.3. The production of an anomalous "Mach" reflection. A shock wave with behind pressure of 100 bars (Mach number 1.09) in water is incident on a bubble of air. The upstream states are ambient at 1 atm. and standard densities. By time 1 μsec the trailing edge of the reflected Prandtl-Meyer wave has overtaken the incident shock producing an analogue to an ordinary Mach reflection where the reflected wave is a non-centered rarefaction. The grid is 60x60.

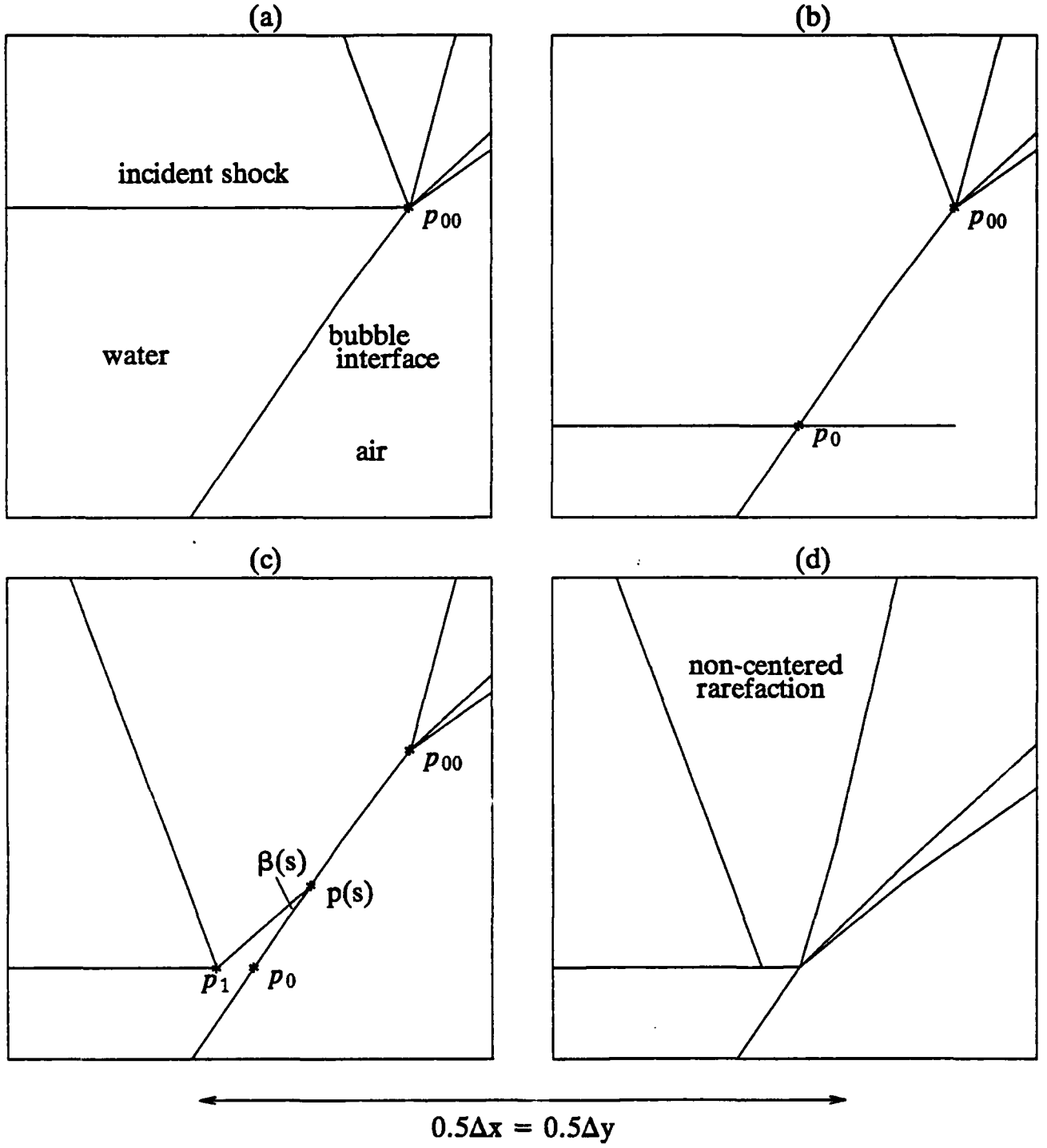


Figure 4.1. A diffraction node initially at p_{00} bifurcates into an anomalous reflection. The predicted new node position at p_0 yields a Mach number of 0.984 behind the incident shock. The leading edge of the reflected Prandtl-Meyer wave breaks away from the diffraction node to form an overtake node at p_1. The propagated position of the diffraction node is adjusted by a distance $s^* = 3.59 \times 10^{-5}$ to return the flow to sonic behind the node.

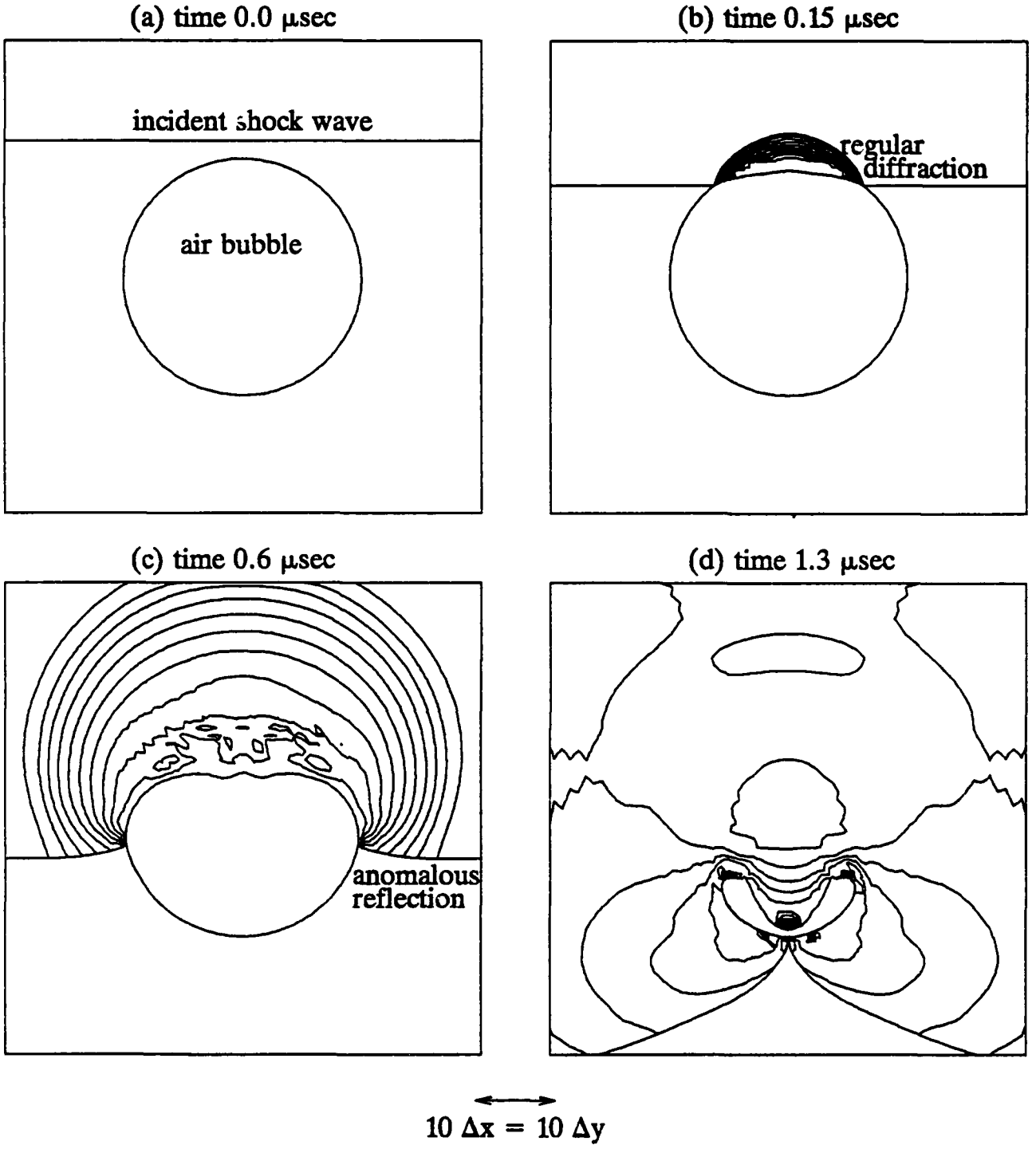


Figure 5.1. Log(1 + pressure) contours for the collision of a shock wave in water with an air bubble. The fluids ahead of the shock are at normal conditions of 1 atm. pressure, with the density of water 1 g/cc and air 0.0012 g/cc. The pressure behind the incident shock is 10 Kbar with a shocked water density of 1.195 g/cc. The grid is 60x60.

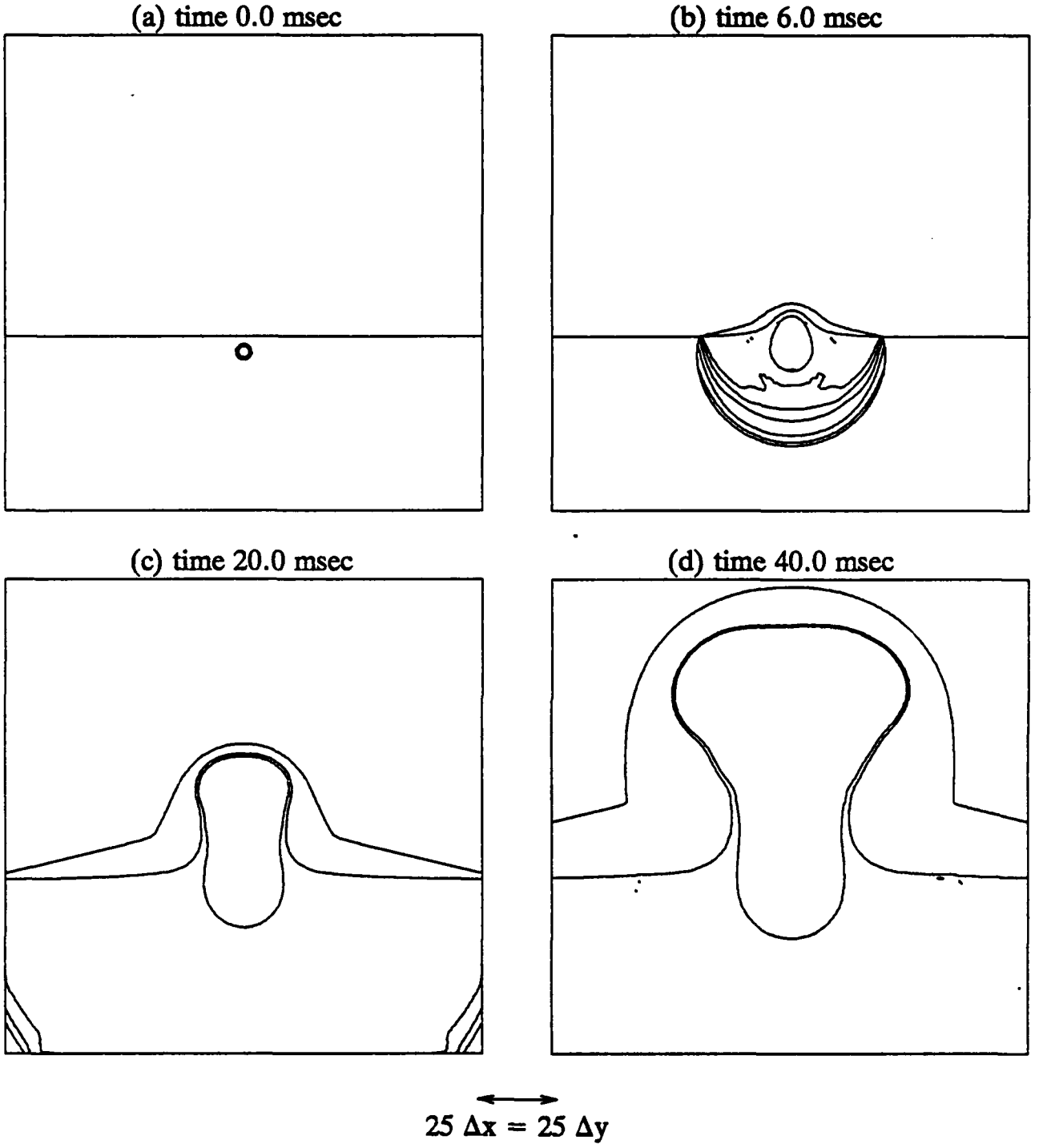


Figure 5.2. An underwater expanding shock wave diffracting through the water's surface. An expanding shock wave with an internal pressure of 10 Kbars and initial radius of 1 meter is installed at a depth of 2 meters below the water's surface. The external conditions are ambient at one atmosphere pressure and normal densities for the air and water. The boundary conditions are constant Dirichlet at the initial ambient values. The grid is 150x150.

August 11, 1988

Numerical Implication of Riemann Problem Theory for Fluid Dynamics *

Ralph Menikoff

Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545

Abstract

The Riemann problem plays an important role in understanding the wave structure of fluid flow. It is also a crucial step in some numerical algorithms for accurately and efficiently computing fluid flow; Godunov method, random choice method, and front tracking method. The standard wave structure consists of shock and rarefaction waves. Due to physical effects such as phase transitions, which often are indistinguishable from numerical errors in an equation of state, anomalous waves may occur; "rarefaction shocks", split waves, and composites. The anomalous waves may appear in numerical calculations as waves smeared out by either too much artificial viscosity or insufficient resolution. In addition, the equation of state may lead to instabilities of fluid flow. Since these anomalous effects due to the equation of state occur for the continuum equations, they can be expected to occur for all computational algorithms. The equation of state may be characterized by three dimensionless variables: the adiabatic exponent γ, the Grüneisen coefficient Γ, and the fundamental derivative $\mathcal{G}$. The fluid flow anomalies occur when inequalities relating these variables are violated.

1. Introduction

Fluid dynamics is governed by a first-order hyperbolic system of conservation laws [Courant & Friedrichs, 1948]. The wave structure in 1–D is determined by the Riemann problem; initial value problem with two piecewise constant states. In addition to qualitatively understanding fluid flow, the Riemann problem is crucial to achieving high accuracy in several numerical algorithms for the computation of fluid flow; Godunov method [Godunov, 1959] and its descendants [Leer, 1979; Colella & Woodward, 1984], random choice method [Glimm, 1965; Chorin, 1976], and front tracking method [Chern, *et al.*, 1985].

The flux function for the fluid dynamic PDEs depends on the equation of state (EOS). This in turn determines the wave curve used to solve the Riemann problem. For the standard case, in which the isentropes are convex, the wave curve consists of the usual shock and rarefaction waves. Near phase transitions the isentropes may not be convex and anomalous waves occur; "rarefaction shocks", split waves and composites. Furthermore, the constraints of thermodynamics on the EOS are not sufficient to obtain reasonable fluid flow; uniqueness and stability.

* Supported by the U. S. Department of Energy

For numerical calculations the EOS should be regarded as input data. Simulations of fluid flow with any numerical algorithm should approximate the continuum solution to the PDEs for fluid dynamics. Many numerical algorithms tacitly assume the EOS has the standard wave structure and may have difficulties for applications in which phase transitions are important and anomalous waves occur. In addition, numerical errors in calculating equations of state and the use of equations of state outside their range of validity may lead to anomalous behavior in situations when they should not occur. Therefore, numerical simulations may have qualitatively or physically incorrect solutions as a consequence of the EOS. Some numerical difficulties may be avoided by preprocessing equations of state to determine the regions in state space for which unphysical flow occurs. Then for applications in which the flow enters these abnormal regions the EOS should be corrected rather then artificially modifying the numerical algorithm to compensate for deficiencies in the EOS.

We begin in §2 by defining notation and stating the fluid equations. The important variable for characterizing the EOS are defined. In §3 the theory described in [Menikoff & Plohr, 1989] on the effect of the EOS on the wave structure for the PDEs is briefly summarized. The numerical implications of the wave structure are described in §4. Conclusions are stated in §5.

2. Mathematical Formulation

2.1 Fluid Equations

We consider an ideal fluid in which viscosity and heat conduction may be neglected. The fluid motion is governed by the equation for the conservation of mass, momentum and energy. In conservation form the fluid dynamic equations are [Courant & Friedrichs, 1948; Landau & Liftshitz, 1959]

$$\partial_t q_j + \partial_{x_i} F_{ij} = 0$$

where

$$\vec{q} = \begin{pmatrix} \rho \\ \rho\vec{u} \\ \rho\mathcal{E} \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} \rho\vec{u} \\ \rho\vec{u}\vec{u} + P\mathbf{I} \\ \rho\mathcal{E}\vec{u} + P\vec{u} \end{pmatrix}$$

and

ρ = Density,

$\vec{u}$ = Particle Velocity,

$\mathcal{E} = E + \frac{1}{2}u^2$ = Total Specific Energy,

E = Specific Internal Energy,

P = Pressure.

The material properties enter the dynamic equations through an incomplete equation of state $P(V, E)$, where $V = 1/\rho$ is the specific volume. The flux function depends on the EOS. Consequently, the EOS determines the wave structure of fluid flow.

2.2 Equation of State

The equation of state may be characterized by the behavior of its isentropes in the P–V plane. Even though the entropy, S, and temperature, T, may not be uniquely defined for an incomplete EOS an isentrope (constant S) is determined by the ODE, $dE = -PdV$. Three variable play an important role in determining the wave structure.

1. The adiabatic exponent
$$\gamma = -\left.\frac{\partial \ln P}{\partial \ln V}\right|_S$$
is the negative slope of the isentrope in the $\ln P$–$\ln V$ plane. It is a dimensionless sound speed, $c^2 = \gamma PV$.
2. The Grüneisen coefficient
$$\Gamma = V\left.\frac{\partial P}{\partial E}\right|_V$$
is a measure of the spacing of the isentropes in the $\ln P$–$\ln V$ plane.
3. The Fundamental Derivative
$$\mathcal{G} = -\tfrac{1}{2}V\frac{\partial^2 P/\partial V^2|_S}{\partial P/\partial V|_S}$$
is a measure of the convexity of the isentropes in the $\ln P$–$\ln V$ plane.

Thermodynamic stability requires that $\gamma \geq 0$. This implies the fluid equations are hyperbolic. The standard case assumes the isentropes are convex, $\mathcal{G} > 0$. When $\Gamma > 0$ the isentropes do not intersect.

3. Wave Curve

The wave curve is the locus of states that may be joined to a fixed initial state by a scale-invariant solution of the PDEs. For fluid dynamics the wave curve consists of two connected branches corresponding to left and right facing (sound) waves. There is also a linear degenerate contact wave which follows the particle trajectories. The solution to the Riemann problem in 1–D is determined by the intersection in the P–u plane of the left wave curve from the left state and the right wave curve from the right state.

In general, scale-invariant solutions are composites of two types of elementary wave; continuous solutions called rarefactions or simple waves, and discontinuous solutions called shock waves. Shocks are determined by algebraic equations, the Hugoniot jump conditions. Simple waves correspond to isentropes and are determined by integrating an ODE along a characteristic. For consistency it is important that the characteristic velocity, $u \pm c$, be monotonic. The variation of the Lagrangian wave speed ρc is given by

$$\mathcal{G} = \left.\frac{\partial \rho c}{c\partial \rho}\right|_S$$

When $\mathcal{G} > 0$ the sound modes are linearly non-degenerate. Compressive waves steepen to form shocks and expansive waves spread out to form rarefactions. When $\mathcal{G} < 0$ the nature of the waves reverses; shock waves are expansive and "rarefaction waves" are compressive.

Finally, when $\mathcal{G}$ changes sign, scale-invariant solutions include composite waves consisting of rarefactions and sonic shocks propagating as a single entity.

3.1 Standard Case

The standard case assumes $\mathcal{G} > 0$ and

$$\gamma + \frac{PV}{2E} \geq \Gamma \, .$$

Here we assume the origin of energy is chosen such that $E \to 0$ as $V \to \infty$. The wave curve satisfying the entropy condition consists of rarefactions in expansion and shocks in compression. The second condition implies the wave curve in the P–u plane is monotonic. It is needed to ensure the Riemann problem has a unique solution [Smith, 1979].

3.2 Anomalous Waves

When a material undergoes a phase transition there is a jump in the sound speed, $c_{\text{mixed}} < c_{\text{pure}}$. As a result when an isentrope in the P-V plane passes through the saturation boundary it suffers a kink, a discontinuity in the slope. Consequently, $\mathcal{G}$ contains a δ-function singularity. Depending on the sign of the kink the isentrope may become non-convex. When the isentrope is non-convex, some single shocks are unstable and split into 2 shocks of the same family. Similarly, a convex kink results in split rarefactions.

For some materials, near a phase transition $\mathcal{G} < 0$ and smooth. In this case the wave speed is not monotonic and some single shocks and rarefactions are unstable resulting in composite waves. Replacing unstable shocks and rarefactions with split waves and composites results in a unique continuous wave curve [Wendroff, 1972]. This corresponds to the extended entropy condition of Liu [1975] and Oleĭnik [1959]. The wave curve of stable scale-invariant states may then be used with the standard construction to solve the Riemann problem.

3.3 Shock Instability

In 2–D a new mode of instability is possible [Fowles, 1981; Kontorovich, 1957; Majda & Rosales, 1983]. When the condition

$$\gamma \geq \Gamma + 1$$

is violated some shocks are unstable to the development of transverse waves. This is qualitatively similar to what is observed experimentally for detonations [Fickett & Davis, 1979]. Another result is that on the shock polar, P at the sonic point is greater than P at the maximum turning angle. As a consequence an unusual wave pattern is possible, a Mach configuration with a reflected rarefaction. On the other hand when shocks are 2–D stable the 1–D Riemann problem has a unique solution.

4. Numerical Implications

The equation of state is input data for numerical calculations. It is not surprising that an incorrect EOS will result in an unphysical simulation of the fluid flow. Two important question are: 1. Given an EOS how to determine if it is physically reasonable; and 2. What

are the symptoms in a numerical simulation of an unphysical EOS. One of the interesting aspects of these problems is that in a simulation the fluid flow samples only a region of state space. Thus an EOS may be adequate for one application and unacceptable for another. In general it is important to determine the domain of validity of an EOS.

For applications with real materials the correct EOS may be experimentally determined. However, accurate EOS experiments are difficult to perform and expensive. Data is available in only a limited range of state space. As a consequence for an EOS one resorts to extrapolating data, semi-empirical fits, and theoretical models. A minimal requirement on any EOS is thermodynamic consistency and stability. This is difficult to check for only the incomplete part of the equations of state needed to simulate ideal fluid flow. It is also insufficient to guarantee the behavior of fluid flow observed experimentally. Simple conditions on the incomplete EOS are based on obtaining the correct wave structure and shock stability. This at least is sufficient to obtain qualitatively correct fluid flow, though the simulation may still be quantitatively inaccurate.

Similar difficulties may occur when the domain of an equation of state is extended by extrapolating empirical fits outside their range of validity or patching together theoretical models. Several conditions on the EOS should be checked. Consider the isentropes in the P–V plane. (1) If the slope of an isentrope becomes positive then $c^2 < 0$ and the fluid equations are no longer hyperbolic. This implies the adiabatic compressibility is negative and results in numerical instabilities; compressing a fluid element lowers its pressure which causes the fluid element to further compress until either the flow leaves the elliptic region of state space or catastrophic failure occurs. (2) If the slopes of the isentropes have kinks, which are physically indistinguishable from a phase transition, then split shocks and rarefactions may develop. (3) If the isentropes are not convex then the extended entropy needs to be imposed. Because of the non-uniqueness different numerical algorithms may give different solution depending for example on the form of artificial viscosity used to impose the entropy condition. In addition, "rarefaction shocks" are possible. Algorithms which rely on artificial viscosity to smear out shocks but only use it in compression will be unstable. Composites consisting of sonic shocks followed by compression waves which do not steepen because $\mathcal{G} < 0$ may have the appearance of shocks smeared out by too much artificial viscosity or too little mesh resolution. (4) If $c^2 > 0$ and $\mathcal{G} > 0$ but $\gamma + \frac{PV}{2E} \geq \Gamma$ is violated then 1–D fluid dynamics does not have a unique solution. Numerical calculations may not converge under mesh refinement or be unusually sensitive to initial data. (5) If $\gamma \geq \Gamma + 1$ is violated then shocks are 2–D unstable. In numerical calculations transverse waves would develop along shock waves. In extreme cases diverging shocks speed up and converging shocks slow down resulting in a ripple instability of shock waves. (6) If $2\gamma > \Gamma$ is violated then the shock Hugoniot may have disconnected branches giving rise to further non-uniqueness of fluid flow.

It should be emphasised that even simple analytic seemingly reasonable EOS can give rise to anomalous and unphysical fluid flow. One example is analyzed in detail by [Menikoff & Plohr, 1989]. This consists of a Grüneisen EOS with constant Γ and a linear u_s–u_p fit for the principle shock Hugoniot.

Even when anomalous waves do not occur other aspects of the EOS may be important. Some numerical algorithms use approximate Riemann solvers for efficiency. Typically ap-

proximate Riemann solvers tacitly assume the isentropes are convex and γ is slowly varying, or the wave curve in the P–u plane is convex. Conditions on the EOS for monotonicity of thermodynamic variables along the wave curve have been worked out in [Menikoff & Plohr, 1989].

5. Conclusions

Equations of state have the potential for causing difficulties in numerical simulations of fluid flow. This may take the form of qualitatively incorrect wave structure or instabilities of shock waves. When non-convexity of an isentrope $\mathcal{G} < 0$ is physical, such as may occur near a phase transition, the numerical algorithm must be capable of producing "rarefaction shocks", and the stable split or composite waves rather than the unstable elementary waves. Problems due to physically inadequate EOS or numerical errors in EOS can be avoided by preprocessing the equation of state to determine the regions of state space with anomalous properties. This entails checking when one of the inequalities $c^2 > 0$, $\mathcal{G} > 0$, or $\gamma \geq \Gamma + 1$ is violated. If in a simulation the flow enters an anomalous region of state space then the EOS needs to be corrected.

There are several important open questions. Analysis of numerical algorithms have assumed the fluid equations are genuinely non-linear with a unique solution determined by an entropy condition, see *e.g.*, [Harten, *et al.*, 1983]. The analysis needs to be generalized to the non-convex case. Some algorithms (for example typical Godunov method) for efficiency use approximate Riemann solvers. This is reasonable when the method averages over a cell and does not use the full detail of the Riemann solution. The properties of an approximate Riemann solver and the artificial viscosity required to obtain the correct wave structure for general EOS needs further study. For other algorithms, such as the front tracking method, which requires a good Riemann solver either better equations of state with accurate derivatives or robust methods of utilizing the EOS are needed.

Acknowledgements

This article is the outgrowth of a collaboration with Prof. B. Plohr in which we analyzed the Riemann problem for fluid dynamics [Menikoff & Plohr, 1989].

References

Chern, I-L., J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, 1985, "*Front Tracking for Gas Dynamics*," J. Comp. Phys. **62**, pp. 83–110.

Chorin, A., 1976, "*Random Choice Solutions of Hyperbolic Systems*," J. Comp. Phys. **22**, pp. 517–533.

Colella, P. and P. Woodward, 1984, "*The Piecewise Parabolic Method (PPM) for Gas-Dynamical Simulation*," J. Comp. Phys. **54**, p. 174.

Courant, R. and K. Friedrichs, 1948, *Supersonic Flow and Shock Waves* (Interscience, New York).

Fickett, W. and W. Davis, 1979, *Detonation* (University of California Press, Berkeley).

Fowles, G., 1981, "*Stimulated and Spontaneous Emission of Acoustic Waves from Shock Fronts*," Phys. Fluids **24**, pp. 220–227.

Glimm, J., 1965, "*Solutions in the Large for Nonlinear Hyperbolic Systems of Equations*," Comm. Pure Appl. Math. **XVIII**, pp. 697–715.

Godunov, S., 1959, "*A Difference Method for Numerical Calculation of Discontinuous Solutions of the Equations of Hydrodynamics*," Mat. Sb. **47**, pp. 271–306.

Harten, A., P. D. Lax, and B. van Leer, 1983, "*On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws*," SIAM Review **25**, pp. 35–61.

Kontorovich, V., 1957, "*Concerning the Stability of Shock Waves*," Soviet Phys.-JETP **6**, pp. 1179–1180.

Landau, L. and E. Lifshitz, 1959, *Fluid Mechanics* (Addison-Wesley, Reading, Mass.).

Leer, B. van, 1979, "*Towards the Ultimate Conservative Difference Scheme V. A Second Order Sequel to Godunov's Method*," J. Comp. Phys. **32**, pp. 101–136.

Liu, T.-P., 1975, "*The Riemann Problem for General Systems of Conservation Laws*," J. Diff. Eqs. **18**, pp. 218–234.

Majda, A. and R. Rosales, 1983, "*A Theory for Spontaneous Mach Stem Formation in Reacting Shock Fronts: I. The Basic Perturbation Analysis*," SIAM J. Appl. Math. **43**, pp. 1310–1334.

Menikoff, R. and B. Plohr, 1989, "*Riemann Problem for Fluid Flow of Real Materials*," Revs. Mod. Phys. **(to be published)**.

Oleĭnik, O., 1959, "*Uniqueness and Stability of the Generalized Solution of the Cauchy Problem for a Quasi-linear Equation*," Usp. Math. Nauk. **14**, pp. 165–170. English transl. in Amer. Math. Soc. Transl., Ser. 2, **33** (1964), pp. 285–290.

Smith, R., 1979, "*The Riemann Problem in Gas Dynamics*," Trans. Am. Math. Soc. **249**, pp. 1–50.

Wendroff, B., 1972, "*The Riemann Problem for Materials with Non-Convex Equations of State: I Isentropic Flow; II General Flow*," J. Math. Anal. and Appl. **38**, pp. 454–466; 640–658.

MATHEMATICAL MODELING OF SOUND PROPAGATION IN THE ATMOSPHERE USING THE PARABOLIC APPROXIMATION

J.S. Robertson
Department of Mathematics
United States Military Academy
West Point, New York 10996-1786
and
M.J. Jacobson and W.L. Siegmann
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

ABSTRACT. There is a significant level of interest in the analytical and numerical modeling of lower-frequency atmospheric acoustic propagation in battlefield environments. Ray-based models, because of their frequency limitations, do not always give an adequate prediction of quantities such as sound pressure or intensity levels. However, the parabolic approximation method, widely used in ocean acoustics, and often more accurate than ray models for frequencies of interest, can be applied to acoustic propagation in the atmosphere. We discuss appropriate physical and asymptotic conditions under which this model is valid. Modifications of an existing implicit finite-difference implementation for computing solutions to the parabolic approximation are discussed. In addition, we present calculations of acoustic intensity levels in a windy atmosphere and contrast the results with those of ray theory.

1. INTRODUCTION. The propagation of low frequency sound through the earth's atmosphere over long distances is a problem with numerous applications. In many instances, acoustic propagation occurs in environments which may be characterized by winds, atmospheric turbulence, extremes of weather, and other natural and man-made atmospheric variations, as well as irregular topography and terrain structure. These environmental variations are typically range- as well as height-dependent, and can profoundly affect the behavior of sound waves. Geometrical acoustics, or classical ray theory, is one approach that has been commonly used to study atmospheric acoustics. Unfortunately, the approximations under which the ray equations hold are valid only for sufficiently high source frequencies. At lower frequencies where diffraction effects are especially important, the use of other mathematical models can provide more accurate and useful results.

2. PARABOLIC APPROXIMATION METHOD. An alternative approach to low-frequency propagation modeling is known as the parabolic approximation method (PAM). This method, originally developed for studies of tropospheric radio wave propagation,[1] exploits characteristic features of the propagation medium associated with the formation of a

waveguide. Atmospheric acoustic waveguides can be formed by certain meteorological conditions either with or without boundary interaction. Within such a waveguide, sound waves may propagate to relatively large distances with significant amplitudes. The parabolic approximation has been successfully applied to a broad variety of problems in ocean acoustics,[2] where many features occur that are analogous to those in atmospheric acoustics.

Let $p(r,z)$ be the acoustic pressure caused by the presence of a point source in a stratified, moving atmosphere, where r and z denote the range and height in cylindrical coordinates. We will confine our attention to a vertical plane containing the source, and parallel to the wind motion. In addition, we assume that the the sound speed is independent of azimuth. We consequently deal with two-dimensional sound propagation. The time-independent wave field, denoted as A, is obtained by assuming that the source is harmonic with frequency f, so that $p = A\exp(2\pi i f t)$. It can be shown that A satisfies the reduced wave equation

$$\nabla^2 A + k_0^2 n^2 A + 2ik_0^2 n^2 \frac{u_0}{c_0} A_r - 2i\frac{k_0}{c_0}\frac{du_0}{dz} A_{zr} = 0, \tag{1}$$

where c_0 is a reference sound speed, $k_0 = 2\pi f/c_0$ is a reference wave number, $c(r,z)$ is the sound speed, $n(r,z) = c_0/c(r,z)$ is the index of refraction, and $u_0(z)$ is the wind speed. Furthermore, it can be shown that away from the source, the quantity A takes on the asymptotic form

$$A = \psi\frac{e^{ik_0 r}}{\sqrt{r}}. \tag{2}$$

Equation (2) is an essential feature of the parabolic approximation when the quantity ψ is related to the slow-scale (i.e. many wavelengths) variation in the acoustic pressure. Furthermore, through careful scaling and asymptotic arguments, it can also be shown that ψ satisfies a family of parabolic equations (PEs). Details of the complete derivation of this family of PEs in an inhomogeneous moving medium can be found in Refs. 3 and 4. For the numerical examples considered in the next section, the appropriate member of this family is given by

$$2ik_0\psi_r + \psi_{zz} + k_0^2(\tilde{n}^2 - 1)\psi = 0, \tag{3}$$

where

$$\tilde{n} = c_0/\tilde{c}, \tag{4}$$

with

$$\tilde{c} = c + u_0. \tag{5}$$

The quantity $\tilde{c}$ is is called the *effective sound speed profile* (ESSP).

Several numerical algorithms for solving Eq. 3 have been found useful and are widely implemented. One employs implicit finite differences,[5] using a Crank-Nicolson scheme to march the solution forward in range. This method is well-suited for many propagation situations, for example, those involving irregularly-shaped boundaries. From this algorithm, we determine ψ, then $p(r,z) = \psi(r,z)\frac{e^{ik_0 r}}{\sqrt{r}}$ is the complex-valued pressure field, and finally relative intensity I, defined as

$$I = 20\log_{10}\left|\frac{p(r,z)}{p_{ref}}\right|, \tag{6}$$

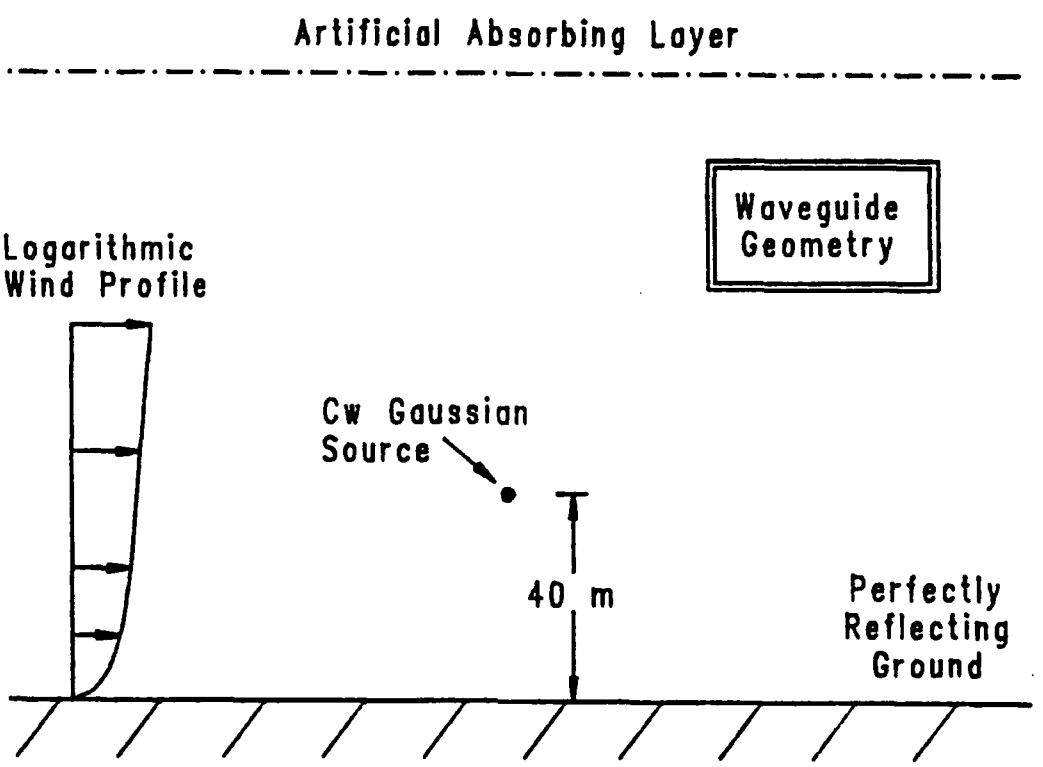


Figure 1: An atmospheric sound channel.

where p_{ref} is the pressure at 1 m from the source. The quantity I will be computed and discussed in the next section.

3. NUMERICAL EXAMPLES Figure 1 depicts an idealized atmospheric acoustic waveguide. We note here that this waveguide is similar to one used as a model in Ref. 6, a study of the downwind propagation of low frequency noise from a wind turbine at a test site in Wyoming. A cw sound source is located 40 m above a horizontal, perfectly-reflecting ground surface. The air is assumed to be isospeed with $c_0 = 330$ ms^{-1}. The atmosphere moves within the indicated plane with a logarithmic velocity profile, a modeling assumption often used for the vertical structure of winds:

$$u_0 = K v_f \ln\left(1 + \frac{z}{z_0}\right), \tag{7}$$

where $K = 2.5$, $v_f = 0.64$ ms^{-1}, and $z_0 = 0.1$ m. As shown, the channel is bounded above by a horizontal, artificial, pressure-release surface of height h, beneath which is an artificial absorbing layer. This absorbing layer is designed to eliminate reflections that would otherwise occur from the pressure-release surface at the top of the waveguide. This combination boundary model is widely used to simulate bottom boundaries in ocean acoustics, and, modified by us to function as a surface model, is a feature of the numerical implementation which we use for our calculations.

A detailed picture of the sound field from a source with frequency 10 Hz when no wind is present can be seen in Fig. 2. Relative intensity is displayed as level curves between a

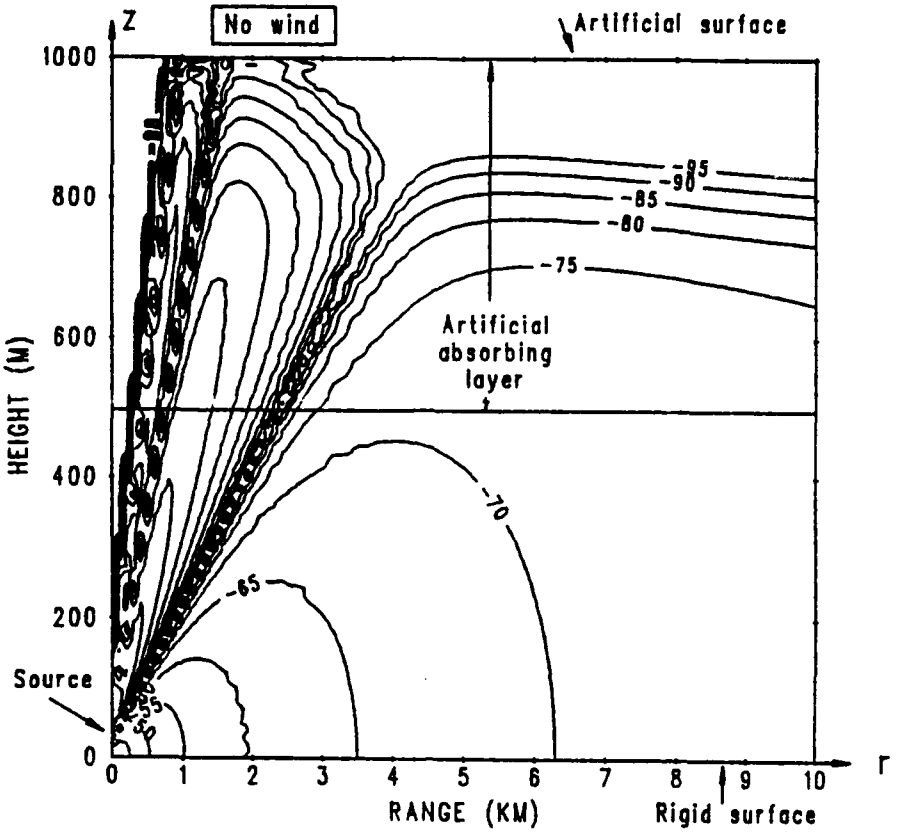


Figure 2: Level curves of relative intensity in the $r-z$ plane. No wind is present.

rigid ground and an artificial surface at height 1000 m. The maximum range shown is 10 km. The artificial absorbing layer is 500 m thick, beginning at height 500 m and extending vertically to the artificial surface. The intensity within the artificial absorbing layer is shown for completeness, but we emphasize that the significant portion of the solution is located near the vicinity of the ground surface. Note the regular way in which intensity decreases for increasing range. It can be shown that this corresponds to spherical spreading of the sound, which is entirely expected from a point source in a homogeneous motionless atmosphere.

In the next illustration, a logarithmic wind profile is present with a maximum wind speed of 14 ms^{-1}. The level curves of intensity downwind from the source are shown in Fig. 3. The boundary locations and layer thickness is as in Fig. 2. Note that a substantial change in the intensity pattern has occurred near the ground surface. Intensity is seen to decrease much more slowly with increasing range. In fact, this intensity pattern can be shown to correspond roughly to cylindrical spreading of the sound. This effect is caused by the direction (and magnitude) of the wind. For sufficiently high frequencies, geometrical acoustics predicts that ray paths would be curved toward the ground, with many rays repeatedly striking the ground. At the source frequency of 10 Hz used here, ray theory is no longer applicable. Nonetheless, the wind still serves to focus the sound near the ground, so that the intensity there is substantially larger than that in the no-wind case discussed above.

When the wind direction is reversed, an entirely different result is encountered. Figure 4 depicts level curves of intensity for the channel region upwind from the source. Note that the intensity decreases more rapidly near the ground surface for increasing range when compared

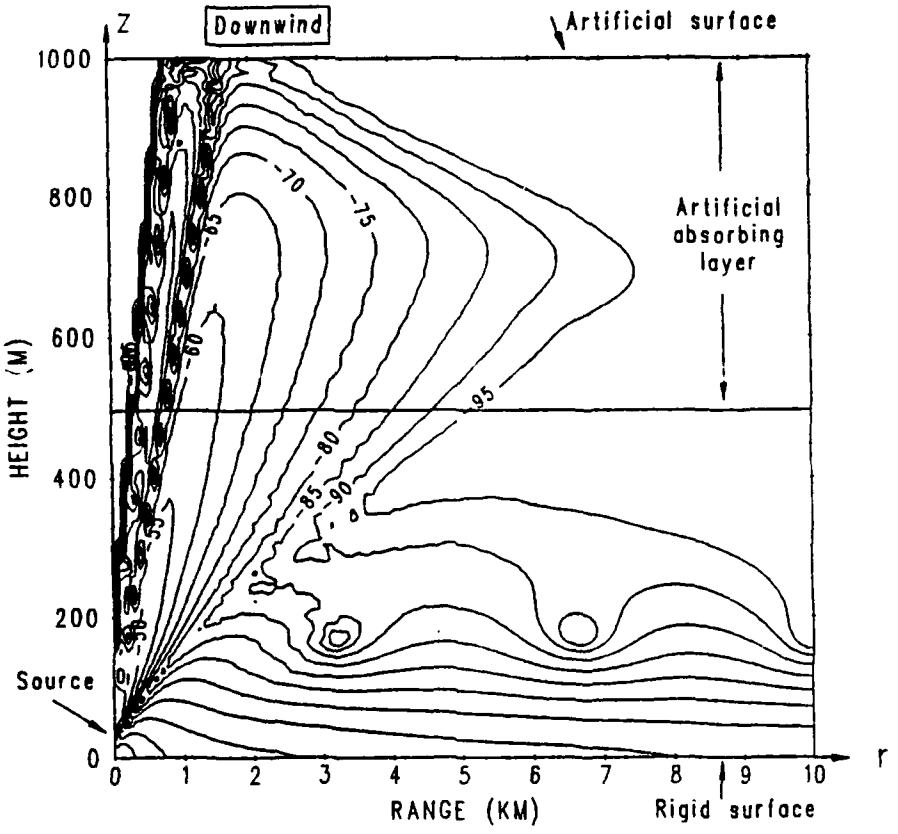


Figure 3: Level curves of relative intensity in the $r-z$ plane downwind from the source.

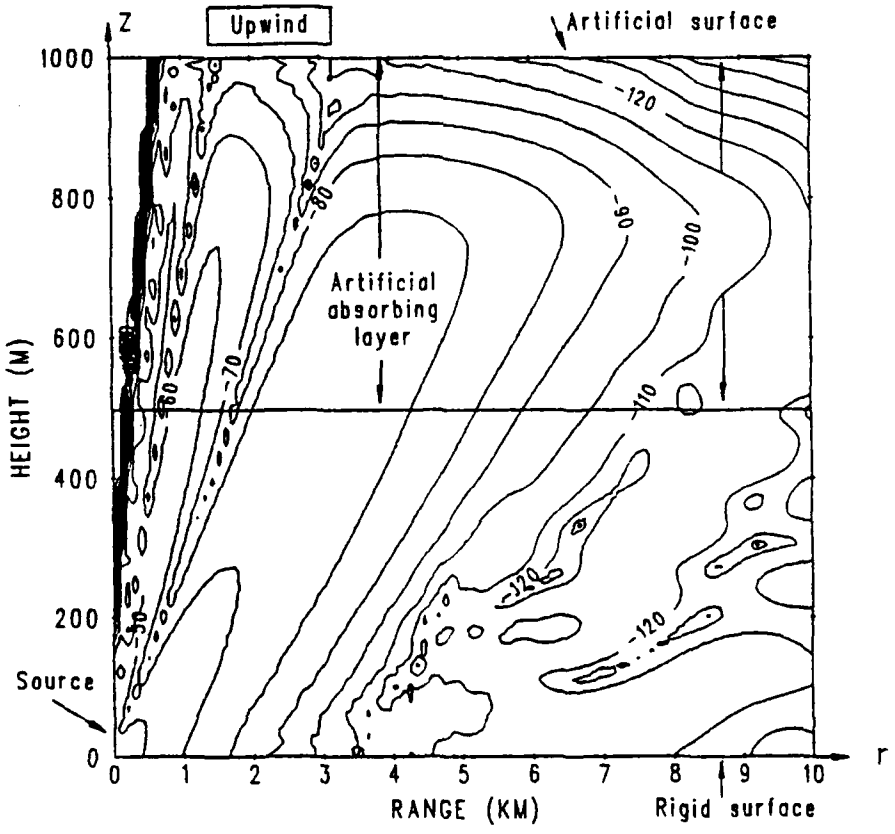


Figure 4: Level curves of relative intensity in the $r-z$ plane upwind from the source.

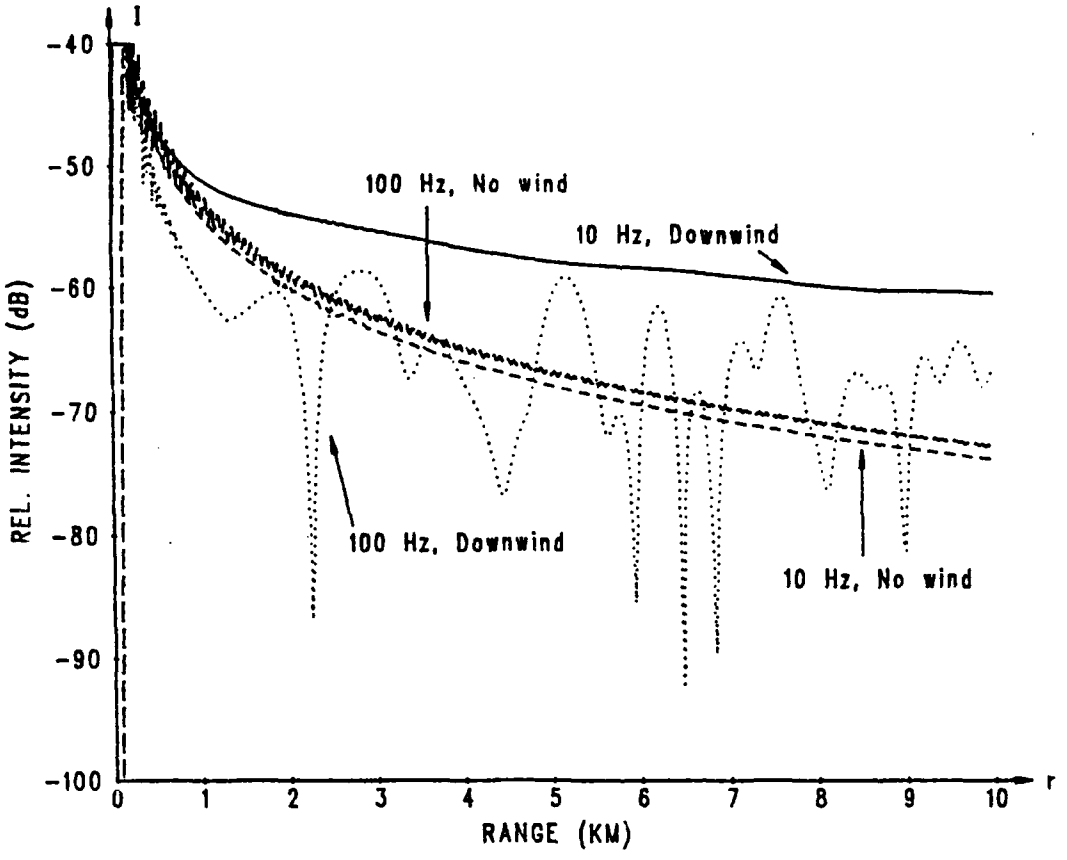


Figure 5: Relative intensity I versus range r at a receiver on the ground. No wind is present.

to the no-wind case shown in Fig. 2. This decrease is stronger than spherical spreading. In fact, geometrical acoustics predicts the existence of a shadow *zone* (a "zone of silence") beginning immediately upwind from the source. As we noted before, the ray model is not applicable at our source frequency. In fact, sound energy can diffract into the shadow zone, so that some sound can be detected at the ground. Naturally, the intensity tends to be reduced when compared to the no-wind channel. We also note that this effect at low frequencies has been detected experimentally,[6] suggesting that a low-frequency propagation model such as the parabolic approximation can be used to interpret experimental data.

Next, we compare calculations done at two frequencies and two wind conditions. Figure 5 displays relative intensity versus range for four different cases. The receiver is located on the ground surface. First, note the calculation for the 10 Hz no-wind case and the 10 Hz downwind case. These two curves illustrate clearly the differences than can occur in the presence of a wind. Even more striking is a comparison between calculations performed with and without wind but at a source frequency of 100 Hz. At this higher frequency, relative intensity would be expected to exhibit more "ray-like" behavior. In the absence of a wind, the high frequency curve is very nearly identical to the low frequency curve. At 100 Hz, the downwind curve is seen to exhibit a strong oscillatory behavior, a consequence of interference effects resulting from multi-mode propagation. This does not occur in the low frequency case, which exhibits no interference pattern at all, again suggesting the predictive power of a low-frequency propagation model such as the parabolic approximation.

4. SUMMARY. We have briefly described the utility of the parabolic approximation for predicting the relative intensity of sound propagating in a moving atmosphere. After sketching the development of this model, we applied it to an idealized atmospheric waveguide containing a steady height-dependent wind. We found that the parabolic approximation yields computed acoustic fields for low-frequency sources which qualitatively differ from those predicted by ray theory. We emphasize that significantly more complicated environments, which might include irregular ground topography, penetrable ground surfaces, and range-dependent sound speed conditions in the medium, can be handled by this numerical implementation.

ACKNOWLEDGEMENT. This work was partially supported by Code 1125, Office of Naval Research. We wish to thank Mr. Arnold Mueller, Applied Acoustics Branch, NASA Langley Research Center, for stimulating our interest in this work.

REFERENCES

[1] M. Leontovich and V. Fock, Zh. Eks. Teor. Fiz. **16**, 557-573 (1946).

[2] W.F. Ames and D. Lee, "Current development in the treatment of ocean acoustic problems," Appl. Num. Math. **3**, 24–47 (1987).

[3] J.S. Robertson, W.L. Siegmann, and M.J. Jacobson, "Current and current shear effects in the parabolic approximation for underwater sound channels," J. Acoust. Soc. Am. **77**, 1768–1780 (1985).

[4] J.S. Robertson, W.L. Siegmann, and M.J. Jacobson, "Acoustical effects of ocean current shear structures in the parabolic approximation," J. Acoust. Soc. Am. **82**, 559–573 (1987).

[5] D. Lee and S.T. McDaniel, "Ocean acoustic propagation by finite difference methods," Comp. & Maths. with Appls. **14**, 305–423 (1987).

[6] J.A. Hawkins, "Application of ray theory to propagation of low frequency noise from wind turbines," NASA Langley Res. Ctr., Langley, VA (1987), CR-178367

TIME-DEPENDENT SHEAR FLOW OF A NON-NEWTONIAN FLUID *

David S. Malkus [1]
John A. Nohel [2]
Bradley J. Plohr [3]

Center for the Mathematical Sciences
University of Wisconsin–Madison
Madison, WI 53706

Abstract

Viscoelastic materials with fading memory, *e.g.*, polymers, suspensions, and emulsions, exhibit behavior that is intermediate between the nonlinear hyperbolic response of purely elastic materials and the strongly diffusive, parabolic response of viscous fluids. Many popular numerical methods used in the computation of (supposedly steady) viscoelastic fluid flows appear to fail in physically relevant regions of parameter space and thus do not capture important phenomena. It is found that a key to a satisfactory explanation of significant non-Newtonian phenomena is to study the fully dynamic governing system of equations. We present results obtained using three classes of numerical methods that accurately represent the dynamics, and we discuss analytical results for related models. We reproduce experimental results on non-Newtonian "spurt" for shearing flow through a slit die and other related phenomena associated with the non-monotone constitutive relation of the shear-stress *vs.* shear strain-rate. We conclude that our results provide a physically reasonable explanation of spurt, hysteresis, and shape memory. Moreover, experiments are suggested to verify our approach.

1. Introduction

Viscoelastic materials with fading memory, *e.g.*, polymers, suspensions, and emulsions, exhibit behavior that is intermediate between the nonlinear hyperbolic response of purely elastic materials and the strongly diffusive, parabolic response of viscous fluids. They incorporate a subtle dissipative mechanism induced by effects of the fading memory. The understanding of the equations of motion coupled with various constitutive assumptions at the mathematical level is crucial for modeling, design of algorithms and computation

* Supported by the U. S. Army Research Office under Grant DAAL03-87-K-0036, the National Science Foundation under Grants DMS-8712058 and DMS-8620303, and the Air Force Office of Scientific Research under Grants AFOSR-87-0191 and AFOSR-85-0141.

[1] Also Department of Engineering Mechanics.

[2] Also Department of Mathematics.

[3] Also Computer Sciences Department.

of particular problems. Shear flows of viscoelastic fluids exhibit a variety of interesting physical phenomena of importance, for example, in polymer processing. We have been intrigued by the fact that many numerical methods used in the computation of (supposedly steady) viscoelastic fluid flows appear to fail in physically relevant regions of parameter space and thus do not capture important phenomena. One such phenomenon is "spurt," the occurrence of which in shear flows of non-Newtonian fluids through a capillary has been confirmed by careful experiments (Vinogradov, *et al.* [18]). The understanding of this and related phenomena has proved to be of surprising physical, mathematical, and computational interest.

Our goals in this study are:

1. To understand the physical model: How do the computed solutions correspond to the molecular or continuum model on which they are based? Can the character of these solutions serve to validate the physical model or suggest improvements in it?
2. To understand the physical consequences of the model: Do the solutions obtained make physical sense? Do solutions that have mathematically interesting character correspond to observed phenomena? Do they predict behavior that should be studied in the laboratory? What solutions to the problem are relevant to processing and design?
3. To understand qualitative properties of the mathematical model: the global existence and uniqueness of solutions, dependence on data, regularity and asymptotic behavior of solutions for large time, approach to steady states, etc.
4. To design numerical methods that account for the mathematics and reproduce the physics.

The outline of this paper is as follows: In §2, we discuss the modelling of spurt and related physical phenomena in capillary flow as a fully time-dependent one-dimensional flow through a slit die, using the Johnson-Segalman differential constitutive relation. In §3, we derive a one-dimensional initial boundary-value problem for shearing flows through a slit die, starting with the 3-D equations. In §4, we present mathematical results for the governing system and for related model problems that capture some of the key phenomena. In §5, we describe three numerical methods and present a variety of results of physical interest obtained using them, including a comparison with experimental data for the spurt phenomenon. In §6. we discuss our conclusions.

2. Physical Phenomena

Interesting phenomena have been observed by Vinogradov, *et al.* [18] in the flow of viscoelastic fluids (monodisperse polyisoprenes) through capillaries. They found that the volumetric flow rate increased dramatically at a critical stress that was independent of molecular weight. This phenomenon, which is called "spurt", had been overlooked or dismissed by rheologists because no plausible mechanism to explain it in the context of steady flows was known. Spurt was lumped together with instabilities such as "slip," "apparent slip," and "melt fracture," which are poorly understood. While regarded as anomalous, these instabilities can severely disrupt polymer processes; they can be avoided in practice only with *ad hoc* engineering expedients. The mechanisms of such phenomena are not understood primarily because the governing equations are analytically intractable

and because popular numerical methods for steady flows fail to capture these dramatic non-Newtonian effects.

Several explanations have been offered for the spurt phenomenon [2, 4, 9, 12]. Their common feature is that the shear stress in steady flow does not vary monotonically with shear strain rate, (as illustrated in Fig. 2, below). These explanations have been rejected by many rheologists as being somehow unphysical. We believe that this criticism is unfounded because it is based on intuition derived from generalized Newtonian models of non-Newtonian fluids.

A key to satisfactory explanation of the spurt phenomenon is the dynamical behavior of the governing equations. While there is a great variety of constitutive models for viscoelastic fluids, the dynamical behavior for many is inaccessible. In this paper, we model the spurt phenomenon using the Johnson-Segalman model [7] as constitutive relation. The latter correctly models the spurt phenomenon, and yet is sufficiently simple to be understood through a combination of analysis, asymptotics, and numerical simulation.

We study idealized flow through a narrow slit die. Assuming that the driving pressure is transmitted instantaneously, the three-dimensional flow may be approximated by a one-dimensional problem. Our analytical and numerical results show that flow in a slit die reflects the essential features observed for capillaries. We believe that this is because the spurt phenomenon depends solely on material properties and the smallest physical dimension of the problem.

A non-monotone stress–strain-rate relation of the kind that causes the spurt phenomenon arises when the fluid behavior is characterized by multiple relaxation times. Interpretation of small-amplitude oscillatory shear data [18] indicates that the relaxation times are widely spaced. Formal asymptotic analysis [10] of the dynamics shows that the effects of the smallest relaxation time are mimicked by a Newtonian viscosity term. For simplicity, we study the Johnson-Segalman model with a single relaxation time and added Newtonian viscosity.

3. Mathematical Formulation

The motion of a fluid under incompressible and isothermal conditions is governed by the balance of linear momentum

$$\rho\left[\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right] = \nabla\cdot\mathbf{S}\ . \tag{3.1}$$

Here, ρ is the fluid density, $\mathbf{v}$ is the particle velocity, and $\mathbf{S}$ is the stress tensor. The response characteristics of the fluid are embodied in the constitutive relation for the stress. For viscoelastic fluids with fading memory, these relations specify the stress as a functional of the deformation history of the fluid. Many sophisticated constitutive models have been devised; see Ref. [1] for a survey. In the present work, we focus on the Johnson-Segalman model [7] as a prototype for general constitutive models. This model accounts for non-affine deformation of Gaussian networks by introducing a slip parameter a, $-1 \le a \le 1$, leading to a nonlinear generalization of the classical Maxwell model.

To specify this constitutive relation, we decompose the stress as

$$\mathbf{S} = -p\mathbf{I} + 2\eta\mathbf{D} + \Sigma\ . \tag{3.2}$$

In this equation, p is an isotropic pressure (which is determined from the incompressibility constraint), η is the coefficient of Newtonian viscosity, and Σ is the non-Newtonian extra stress. Also, we let $\mathbf{D} := \frac{1}{2}\left[\nabla\mathbf{v} + (\nabla\mathbf{v})^T\right]$ and $\mathbf{\Omega} := \frac{1}{2}\left[\nabla\mathbf{v} - (\nabla\mathbf{v})^T\right]$ be the symmetric and antisymmetric parts of the velocity gradient $\nabla\mathbf{v}$, which has components $(\nabla\mathbf{v})^i{}_j := \partial v^i/\partial x^j$. The extra stress is specified by the differential constitutive law

$$\overset{*}{\Sigma} = 2\mu\mathbf{D} - \lambda\Sigma \ , \tag{3.3}$$

where

$$\overset{*}{\Sigma} := \frac{\partial\Sigma}{\partial t} + \mathbf{v}\cdot\nabla\Sigma + \Sigma[\mathbf{\Omega} - a\mathbf{D}] + [\mathbf{\Omega} - a\mathbf{D}]^T\Sigma \tag{3.4}$$

is the objective time derivative of Σ with parameter a. The parameter μ is an elastic shear modulus, and λ is a relaxation rate.

Constitutive relations such as Eq. 3.3 exhibit a mixture of elastic and viscous behavior. This may be seen heuristically as follows. In the long relaxation-time limit, $\lambda \to 0$, Eq. 3.3 shows that an objective time derivative of Σ is proportional to the deformation rate: $\overset{*}{\Sigma} \sim 2\mu\mathbf{D}$. This is characteristic of elastic behavior, and leads to the interpretation of μ as a shear modulus. By contrast, when $\lambda, \mu \to \infty$ with μ/λ fixed, $\Sigma \sim 2(\mu/\lambda)\mathbf{D}$; thus, the model displays viscous behavior with μ/λ being the Newtonian shear viscosity coefficient.

Essential properties of the constitutive relation are exhibited in simple planar shear flow. With the flow aligned along the y-axis (see Fig. 1), the flow variables are independent of y. Therefore, the velocity field is $\mathbf{v} = (0, v(x,t))$, and the balance of mass is automatically satisfied. Furthermore, the components of the extra stress tensor Σ may be written $\Sigma^{xx} = \gamma(x,t)$, $\Sigma^{xy} = \Sigma^{yx} = \sigma(x,t)$, and $\Sigma^{yy} = \tau(x,t)$, while the pressure takes the form $p = p_0(x,t) - f(t)y$, f being the pressure gradient driving the flow. In these terms, Eqs. 3.3 become

$$\gamma_t + (1-a)\sigma v_x = -\lambda\gamma \ , \tag{3.5a}$$

$$\sigma_t - \left[\tfrac{1}{2}(1+a)\gamma - \tfrac{1}{2}(1-a)\tau + \mu\right] v_x = -\lambda\sigma \ , \tag{3.5b}$$

$$\tau_t - (1+a)\sigma v_x = -\lambda\tau \ . \tag{3.5c}$$

Introducing the variables $Z := \frac{1}{2}(1+a)\gamma - \frac{1}{2}(1-a)\tau$ and $W := -\frac{1}{2}(1+a)\gamma - \frac{1}{2}(1-a)\tau$, Eqs. 3.5 simplify to

$$\sigma_t - (Z+\mu)v_x = -\lambda\sigma \ , \tag{3.6a}$$

$$Z_t + (1-a^2)\sigma v_x = -\lambda Z \ , \tag{3.6b}$$

$$W_t = -\lambda W \ . \tag{3.6c}$$

Because W must remain finite as $t \to -\infty$, $W \equiv 0$, and the last equation may be omitted. As a result, $Z = -\frac{1}{2}(1-a^2)(\tau - \gamma)$, where $\Sigma^{yy} - \Sigma^{xx} = \tau - \gamma$ is the principal normal stress difference.

Combining the constitutive law 3.6 with the balance of linear momentum 3.1, we are led to the system of equations

$$\rho v_t - \sigma_x = \eta v_{xx} + f \ , \tag{3.7a}$$

$$\sigma_t - (Z+\mu)v_x = -\lambda\sigma \ , \tag{3.7b}$$

$$Z_t + (1-a^2)\sigma v_x = -\lambda Z \ . \tag{3.7c}$$

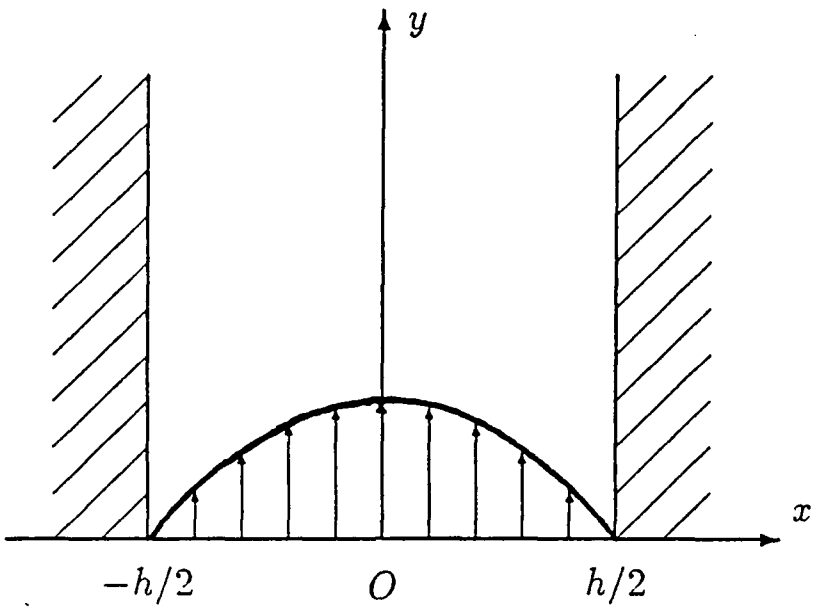


Fig. 1: Shear flow through a slit-die.

In this paper, we study shear flow between two parallel plates, located at $x = \pm h/2$. By symmetry, we need only consider the flow on the interval $[-h/2, 0]$. The no-slip condition at the plate implies the boundary condition $v(-h/2, t) = 0$, while symmetry imposes that $v_x(0, t) = 0$. We also prescribe initial values for v, σ, and Z, which must be compatible with the boundary conditions. To conform with the symmetry, we require that $\sigma(0, 0) = 0$; then, according to Eq. 3.7b, $\sigma(0, t) = 0$ for all time.

To eliminate unnecessary parameters, we scale distance by h, time by λ^{-1}, and stresses σ and Z by μ. Furthermore, if we replace σ, v, and f by $\hat{\sigma} := (1-a^2)^{1/2}\sigma$, $\hat{v} := (1-a^2)^{1/2}v$, and $\hat{f} := (1-a^2)^{1/2}f$ respectively, then the parameter a disappears from Eqs. 3.7. Since no confusion will arise, we omit the caret. The dimensionless parameters are $\alpha := \rho h^2 \lambda^2/\mu$ and $\epsilon := \eta\lambda/\mu$. Consequently, we study the initial-boundary-value problem for the system

$$\begin{aligned} \alpha\, v_t - \sigma_x &= \epsilon\, v_{xx} + f\ , \\ \sigma_t - (Z+1)v_x &= -\sigma\ , \\ Z_t + \sigma v_x &= -Z\ , \end{aligned} \tag{JS}$$

on the interval $[-1/2, 0]$, with boundary conditions

$$v(-1/2, t) = 0 \quad \text{and} \quad v_x(0, t) = 0 \tag{BC}$$

and initial conditions

$$v(x,0) = v_0(x)\ , \quad \sigma(x,0) = \sigma_0(x)\ , \quad \text{and} \quad Z(x,0) = Z_0(x)\ , \qquad (IC)$$

where $v_0(-1/2) = 0$, $v_0'(0) = 0$ and $\sigma_0(0) = 0$.

The steady-state solutions of (JS), when the forcing term f is a constant $\overline{f}$, play an important role in our discussion. Such a solution, denoted by $\overline{v}$, $\overline{\sigma}$, and $\overline{Z}$, is given as follows. The stress components $\overline{\sigma}$ and $\overline{Z}$ are related to the velocity gradient $\overline{v}_x$ (which, in dimensionless units, is the Deborah number) through

$$\overline{\sigma} = \frac{\overline{v}_x}{1+\overline{v}_x^2} \qquad (3.8)$$

and

$$\overline{Z} + 1 = \frac{1}{1+\overline{v}_x^2}\ . \qquad (3.9)$$

Therefore, the total steady shear stress, which is defined by $T := \sigma + \epsilon v_x$, takes the form

$$\overline{T}(\overline{v}_x) = \frac{\overline{v}_x}{1+\overline{v}_x^2} + \epsilon\overline{v}_x\ . \qquad (3.10)$$

In this manner, a non-monotone relation between shear stress and strain rate, shown in Fig. 2, derives naturally in the Johnson-Segalman model.

The total steady shear stress satisfies

$$\overline{T}(\overline{v}_x) + \overline{f}x = 0 \quad \text{for} \quad x \in [-\tfrac{1}{2}, 0]\ , \qquad (3.11)$$

so that the velocity gradient may be expressed in terms of x. However, because the function $\overline{T}$ of Eq. 3.10 is not monotone, $\overline{v}_x$ may take up to three distinct values for any given x. The steady velocity profile, shown in Fig. 3, is obtained by integrating $\overline{v}_x$ and using the boundary condition $\overline{v}(-1/2) = 0$. Notice that $\overline{v}_x$ may suffer jump discontinuities, resulting in kinks in the velocity profile (as at the point x_* in Fig. 3).

Traditionally, a non-monotone relation between stress and strain rate is regarded as a defect of the constitutive law. This conclusion is based on intuition appropriate for generalized Newtonian models of non-Newtonian fluids. Shear flow for such a fluid is governed by the single equation

$$\rho v_t - [\eta(v_x)v_x]_x = f\ , \qquad (3.12)$$

corresponding to having a viscosity coefficient η that depends on strain-rate. In a flow regime where $\eta(v_x)v_x$ decreases with strain rate v_x, however, Eq. 3.12 has the character of a backward heat equation, which suffers from the Hadamard instability. Therefore, for generalized Newtonian fluids, $\eta(v_x)v_x$ must increase with v_x in a physically stable steady solution.

The system (JS) has the same steady solutions as a generalized Newtonian fluid with $\eta(v_x)v_x = \overline{T}(v_x)$, so one might think that it exhibits the same instability in regions where $\overline{T}$ decreases. This conclusion is not warranted, however, because the system (JS) maintains its evolutionary character when $\epsilon > 0$.

4. Mathematical Results

Several mathematical results are known for the system (JS); we refer to Refs. [6, 17, 3, 16, 14, 4] for further discussions and additional references.

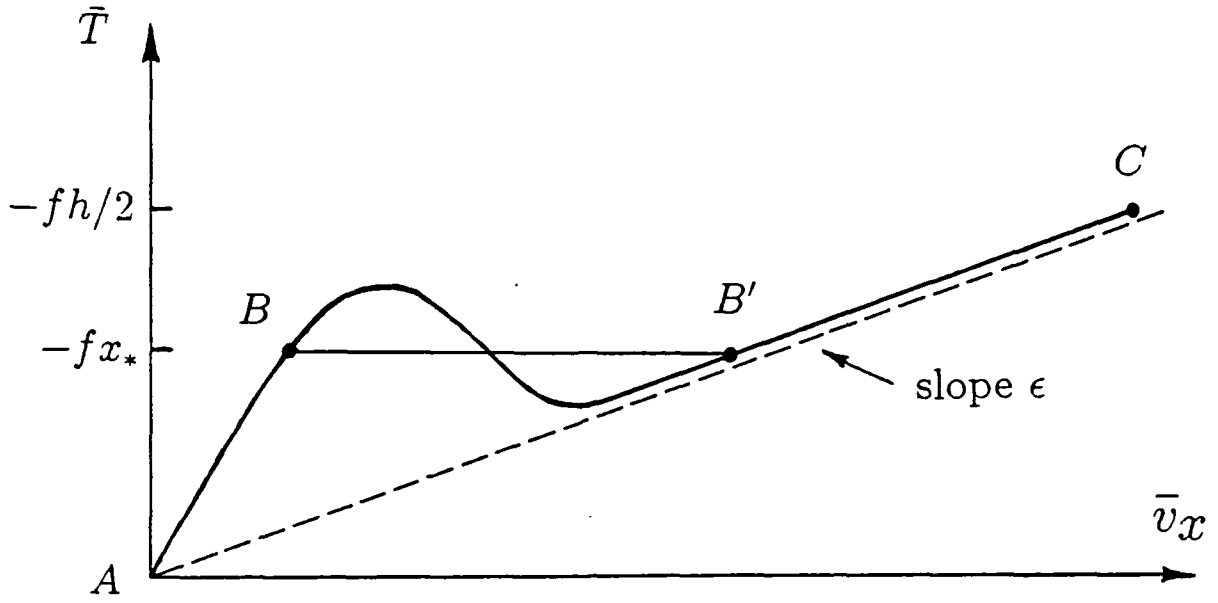


Fig. 2: Total steady shear stress $\overline{T}$ *vs.* shear strain rate $\overline{v}_x$ for steady flow.

(1) When the viscosity parameter $\epsilon = 0$, the quasi-linear system (JS) is strictly hyperbolic provided that $Z + 1 > 0$. In this case, the wave speeds are $\pm[(Z+1)/\alpha]^{1/2}$ and zero. If, on the other hand, $Z + 1$ becomes negative, then (JS) with $\epsilon = 0$ undergoes a change of type and loses its evolutionary character. Joseph, Renardy, and Saut [6] have associated this change of type with certain fluid instabilities.

(2) Let $\epsilon = 0$ and $f = 0$; assume that the initial data are smooth and lie in the hyperbolic region. If the data have sufficiently small variation, then a unique classical solution of (JS) exists globally in time. Moreover, the solution decays to zero as $t \to \infty$. This can be proved using the energy methods discussed in Ref. [17].

On the other hand, if the data have sufficiently large variation, then the classical solution blows up within finite time: $|v_x|$, $|\sigma_x|$, and $|z_x|$ approach infinity as t approaches a finite critical time. This is proved in Ref. [17] using the method of characteristics.

Thus, the fading memory acts as a weak dissipative mechanism: the source terms in the equations serve to counteract the formation of singularities from sufficiently smooth data. When discontinuities do form, system (JS) is no longer valid: the products of distributions Zv_x and σv_x are ill-defined. (See the discussion under (4), below.)

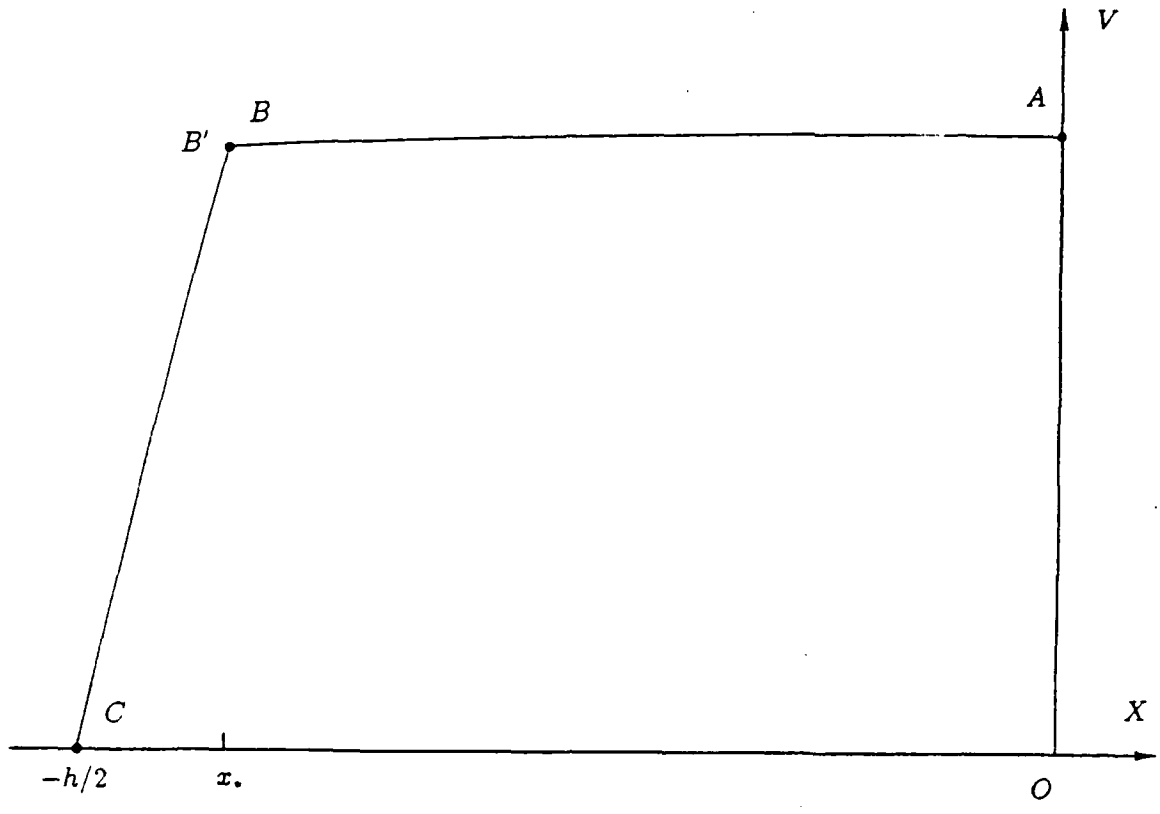


Fig. 3: Velocity profile for steady flow.

(3) If $\epsilon > 0$, the system (JS) is evolutionary, but it cannot be classified according to type. Recently, Guillopé and Saut [3] established the global existence of solutions of (JS) for planar Couette and Poiseuille flow with data of arbitrary size. They also studied the asymptotic (Lyapunov) stability of steady states in the Couette case.

(4) It is important to observe that (JS) is not in conservation form. The evolution of a Johnson-Segalman fluid is, in fact, governed by physical conservation laws [7]. A conservative formulation of (JS) must be used when solutions are discontinuous.

Following Plohr [16], we introduce the "elastic part" τ of the shear strain and the "entropy" variable z through the relations

$$\sigma := z \sin \tau \ , \tag{4.1a}$$

$$Z + 1 := z \cos \tau \ . \tag{4.1b}$$

Then system (JS) is transformed into the equivalent system

$$\begin{aligned} \tau_t - v_x &= -z^{-1} \sin \tau \ , \\ \alpha v_t - [\sigma(\tau, z) + \epsilon v_x]_x &= f \ , \\ z_t &= -(z - \cos \tau) \ , \end{aligned} \tag{C}$$

which is in conservative (*i.e.*, divergence) form. Furthermore, if the internal energy $\mathcal{E}$ is defined by

$$\alpha\,\mathcal{E} := 1 - z\cos\tau\ , \tag{4.2}$$

the energy is dissipated according to the equation

$$\alpha\left[\tfrac{1}{2}v^2 + \mathcal{E}(\tau,z)\right]_t - \left\{[\sigma(\tau,z) + \epsilon\, v_x]\, v\right\}_x = vf - \alpha\,\mathcal{E}(\tau,z) - \epsilon(v_x)^2\ . \tag{4.3}$$

The conservative formulation (C) of (JS) is used in one of the numerical methods discussed in §5.

(5) More detailed analytical results are obtained by simplifying the system (JS). A model system that incorporates several qualitative features of (JS) is obtained by freezing Z at its equilibrium value: $Z + 1 = 1/(1 + v_x^2)$. Defining $g(v_x) := v_x/(1 + v_x^2)$, system (JS) becomes

$$\begin{aligned} \alpha\, v_t - \sigma_x &= \epsilon\, v_{xx} + f\ , \\ \sigma_t - g(v_x) &= -\sigma\ . \end{aligned} \tag{M}$$

More generally, g may be any smooth, odd function. The boundary and initial conditions for v and σ are the same as in (BC) and (IC). We assume that $\epsilon > 0$ and that f is the constant $\overline{f}$. The function g is related to the steady stress–strain-rate relation through $\overline{T}(\overline{v}_x) = g(\overline{v}_x) + \epsilon\overline{v}_x$. A steady solution of (M) satisfies $\overline{\sigma} = g(\overline{v}_x)$ and $\overline{T}(\overline{v}_x) + \overline{f}x = 0$, just as for the system (JS).

Nohel, Pego, and Tzavaras [14] have shown that the global classical solution v, σ of (M), (BC), (IC) has the following properties.

(a) With $S := \sigma + \epsilon\, v_x + \overline{f}x$, $S(x,t) \to 0$ as $t \to \infty$, uniformly for $x \in [-1/2, 0]$.

(b) There exists a steady state $\overline{v}$, $\overline{\sigma}$ such that for each $x \in [-1/2, 0]$, $v(x,t) \to \overline{v}(x)$, $v_x(x,t) \to \overline{v}_x(x)$, and $\sigma(x,t) \to \overline{\sigma}(x)$ as $t \to \infty$. We emphasize that the steady velocity gradient $\overline{v}_x$ and stress $\overline{\sigma}$ may be discontinuous (as in Fig. 2).

(c) Let $\overline{v}$, $\overline{\sigma}$ be a steady state such that

$$\overline{T}'(\overline{v}_x) = g'(\overline{v}_x) + \epsilon \geq \text{const.} > 0\ . \tag{4.4}$$

(Referring to Fig. 2, inequality 4.4 precludes top and bottom jumping and excludes the region where $\overline{T}(\overline{v}_x)$ decreases.) Consider a union $\mathcal{U}$ of small subintervals of $-\frac{1}{2} < x < 0$ that are centered at points where $\overline{v}_x$ and $\overline{\sigma}$ are discontinuous. Let smooth initial data be chosen such that $|S(x,0)|$ is sufficiently small except in $\mathcal{U}$. Then the solution of (M) converges to the steady state $\overline{v}$, $\overline{\sigma}$ on the complement of $\mathcal{U}$. Moreover, the measure of $\mathcal{U}$ can be made arbitrarily small by choosing $|S(x,0)|$ small enough. In this sense, steady states are stable (even if $\overline{v}_x$ and $\overline{\sigma}$ are discontinuous).

The numerical results discussed in §5 suggest that similar results hold for the system (JS). Proofs for (JS) are under investigation.

(6) The model problem (M) was studied also by Hunter and Slemrod [4]. In their construction of the model, the steady-state relation $\overline{\sigma} = g(\overline{v}_x)$ between the stress and strain rate is chosen to be $g_{HS}(v_x) := \sigma_{HS}(v_x) - \epsilon v_x$, where the graph of the function σ_{HS} resembles Fig. 2 (but is independent of ϵ). Hunter and Slemrod base their analysis on the conservation laws

$$w_t - u_x = 0 \,, \tag{4.5a}$$

$$\alpha\, u_t - \sigma_{HS}(w)_x = \epsilon\, u_{xx} - \alpha u \tag{4.5b}$$

for the acceleration $u = v_t$ and the strain rate $w = v_x$. Therefore, jumps in the strain rate v_x are seen to correspond to steady shock waves for the system 4.5 with $\epsilon = 0$. Based on a local dynamical analysis of shock structure for small ϵ, the centerline velocity is shown to exhibit hysteresis under quasi-static cycling of the pressure gradient. (This same behavior is observed in the numerical simulation of the system (JS); see §5.) We emphasize, however, that this analysis cannot be applied to the model problem (M) as derived from the Johnson-Segalman system (JS) because the function $g_{JS}(v_x) = v_x/(1 + v_x^2)$ decays to zero at high strain-rate.

5. Numerical Results

To study the dynamics of system (JS), we developed several different numerical methods; each has its advantages for certain ranges of physical parameters. Calculations with these methods produce similar qualitative and quantitative results.

(1) *Solid Mechanics Formulation:* In this approach, the system (JS) is regarded as governing the extensional motion of an elastic-plastic bar. The first equation is momentum balance, in which the parabolic term adds viscous "stiffness damping." The remaining equations are incremental constitutive relations for the stress. The stiffness of the material is reflected in the wave speed $[(Z + 1)/\alpha]^{1/2}$. We have observed that the wave speed is diminished under loading, so that the material exhibits plastic softening. (See also Ref. [16] for an interpretation of (JS) as governing a viscoplastic material.)

We have solved the system (JS) numerically using a method motivated by solid mechanics. The momentum equation is cast in Galerkin weak form, with the velocity approximated as piecewise linear and the stress components as piecewise constant. With the time derivative discretized using a trapezoidal approximation, and the shear stress determined through a semi-implicit treatment of its evolution equation, the Galerkin equation is solved for the velocity. Then the stress components are updated using an implicit form of the constitutive equations; further details can be found in Ref. [8]. The stability of this method has been analyzed for the system (JS) with Z frozen [11]: the method is stable provided that $Z + 1 > -\epsilon$ and the time step is restricted by $\Delta t < 2/\lambda$ in Eq. 3.7 *i.e.*, $\Delta t < 2$ in (JS). For $Z + 1 \leq -\epsilon$, the linearized equations and the method are unstable.

(2) *Parabolic Formulation:* Recall that the total stress is defined to be $T = \sigma + \epsilon v_x$. Introducing T as an independent variable for $\epsilon > 0$, the system (JS) is replaced by:

$$T_t = \frac{\epsilon}{\alpha} T_{xx} + (Z + 1)\left(\frac{T - \sigma}{\epsilon}\right) - \sigma \,, \tag{5.1a}$$

$$\sigma_t = (Z+1)\left(\frac{T-\sigma}{\epsilon}\right) - \sigma\,, \qquad 5.1b$$

$$Z_t = -\sigma\left(\frac{T-\sigma}{\epsilon}\right) - Z\,. \qquad 5.1c$$

The boundary conditions are $T_x(-1/2,t) = -f$ and $T(0,t) = 0$. The velocity profile may be reconstructed by integrating $(T-\sigma)/\epsilon$.

The system 5.1 has the form of a linear heat equation forced by a nonlinear heat source that is governed by two auxiliary ordinary differential equations. To solve this system numerically, we discretize the parabolic term in 5.1a implicitly while treating the remaining forcing terms explicitly. Time integration is performed using a stiff ODE solver.

We remark that system 5.1 is convenient also for studying existence and regularity of solutions of (JS).

(3) *Conservative Formulation*: The system (JS) is equivalent to the system (C); therefore, it may be studied from the viewpoint of conservation laws. In Ref. [16], we have determined completely the structure of scale-invariant nonlinear waves for (C) when $\epsilon = 0$. Such a wave consists of a sequence of elementary scale-invariant waves, either centered discontinuities or rarefaction waves, connecting constant states on the left and right. Discontinuities are required to satisfy Liu's generalization of Oleĭnik's entropy condition, which guarantees that energy is dissipated (*cf.* Eq. 4.3). This admissibility condition is equivalent to requiring shock waves to have viscous profiles: admissible shock waves arise as limits of traveling-wave solutions of (C) as $\epsilon \to 0$. Our analysis follows the techniques for general systems of conservation laws discussed in Refs. [13] and [5].

With the structure of scale-invariant waves known, Riemann initial-value problems may be solved. We have written a computer program that solves Riemann problems, and have incorporated it into the Glimm-Chorin random choice method. This method solves the Cauchy problem without introducing artificial Newtonian viscosity. We refer to Ref. [15] for a detailed discussion.

As our first numerical experiment, we simulated system (C) with $\epsilon = 0$ using the random choice method. The channel width was chosen so that $\alpha = 1$. The flow was initially in the classical steady state corresponding to the critical pressure gradient $f_{\text{crit}} = 1$; then the pressure gradient is increased abruptly to the super-critical value $1.2 f_{\text{crit}}$.

The result is shown in Fig. 4. The fluid velocity v is plotted *vs.* position x at successive time intervals; generally the velocity increases with time. During the early stages of the experiment, the flow settled into a quasi-steady state. This latency effect is especially evident in a plot of the centerline velocity as a function of time, and it is more pronounced when the channel width (*i.e.*, h — hence also α) is smaller. Eventually, however, a thin layer develops at the plate in which the velocity rises to a value that is nearly constant across the channel. For practical purposes, the fluid has broken free from the plate and is accelerating uniformly under the applied pressure gradient; thus the fluid "slips." We do not claim to have developed a new theory of wall slip at this point, though this phenomenon has been associated with non-monotone constitutive relations by others [2, 9]. If this connection is to be explored more deeply in the future, it is worth noting the success of the random

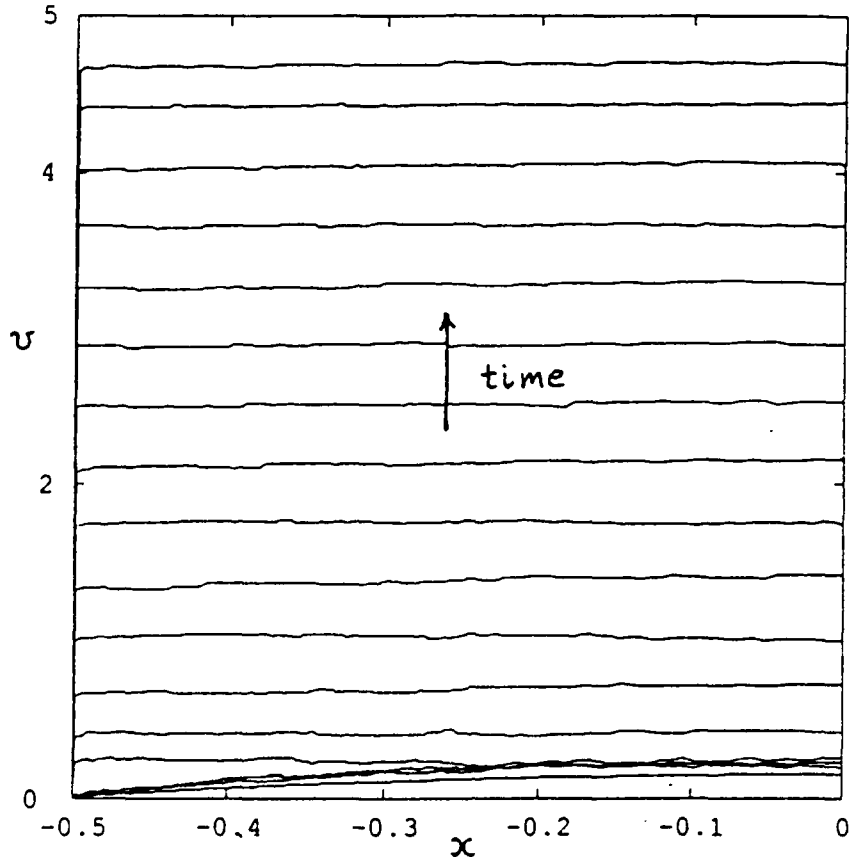


Fig. 4: Onset of slip for a fluid without Newtonian viscosity.

choice method in the post-critical, $\epsilon = 0$ regime; it is the only one of our methods that can compute in this range.

The same experiment was performed for system (C) with a small, but nonzero, Newtonian viscosity coefficient ϵ. Fig. 5 shows the results for $\epsilon = 0.01$, as calculated using the Lax-Wendroff method with Tyler artificial viscosity. What results is evidently a different phenomenon, in which the shorter relaxation response (here modelled by Newtonian viscosity) of the fluid arrests the acceleration in a layer near the wall. Now the layer is much thicker, with its outer boundary corresponding to a discontinuity in the strain rate v_x. The solution approaches a steady state in which v_x is discontinuous but the total stress $T = \sigma(v_x) + \epsilon v_x$ is continuous. The steady state has the same layer thickness as predicted analytically, but the centerline velocity is 20% too high; this is because the centerline velocity is extremely sensitive to the slope of the velocity profile in the layer, which is affected by the artificial viscosity in the numerical method. the layer formation is the key to our interpretation of the spurt phenomenon.

More extensive experiments were performed using the solid mechanics algorithm. For example, the calculation of Fig. 5 was repeated using this method and a graded mesh of 160 elements; the same layer thickness as shown in Fig. 5 was obtained, and the centerline velocity of the long-time solution differed from the analytic prediction by about only 1%. We

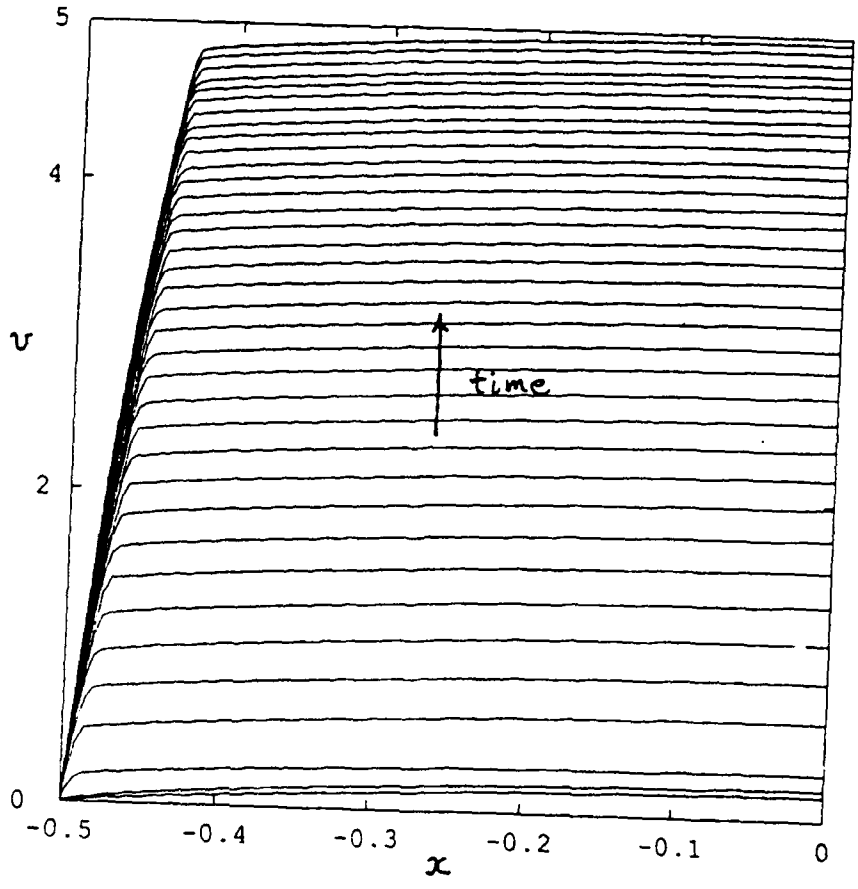


Fig. 5: Onset of spurt for a fluid with Newtonian viscosity.

also used this method to simulate experiments of Vinogradov, *et al.*with polyisoprene [18]. For these calculations, α and ϵ were chosen to correspond to the measurements of Ref. [18]. The dimensionless number α reflects the relative importance of inertial effects compared to elastic effects, and ϵ reflects the relative importance of the dominant relaxation time compared to the secondary relaxation time. As discussed in §2, we use a zero relaxation time to model the effect of a second relaxation time very much shorter than λ^{-1}. When this is done, ϵ is the ratio of viscosities given in §3. We emphasize that, although the ϵ-term in system (JS) appears formally as a Newtonian or solvent viscosity [1], there is no solvent involved in Vinogradov's materials. The samples are labelled PI-1 through PI-8, odered by increasing molecular weight, M. The following features of Vinogradov's polyisoprene samples were used to determine the physical constants:

1. The elastic modulus, μ, is independent of the molecular weight, M.
2. The contribution to the zero shear viscosity from the dominant relaxation time, $\mu_0 = \mu/\lambda$ varies over nearly two orders of magnitude, due to variation in relaxation time, λ^{-1}, with M.
3. PI-1 and PI-2 do not exhibit spurt; there is a critical M below which the material will not spurt.
4. For samples PI-3–8, the observed critical stress is not a function of M.

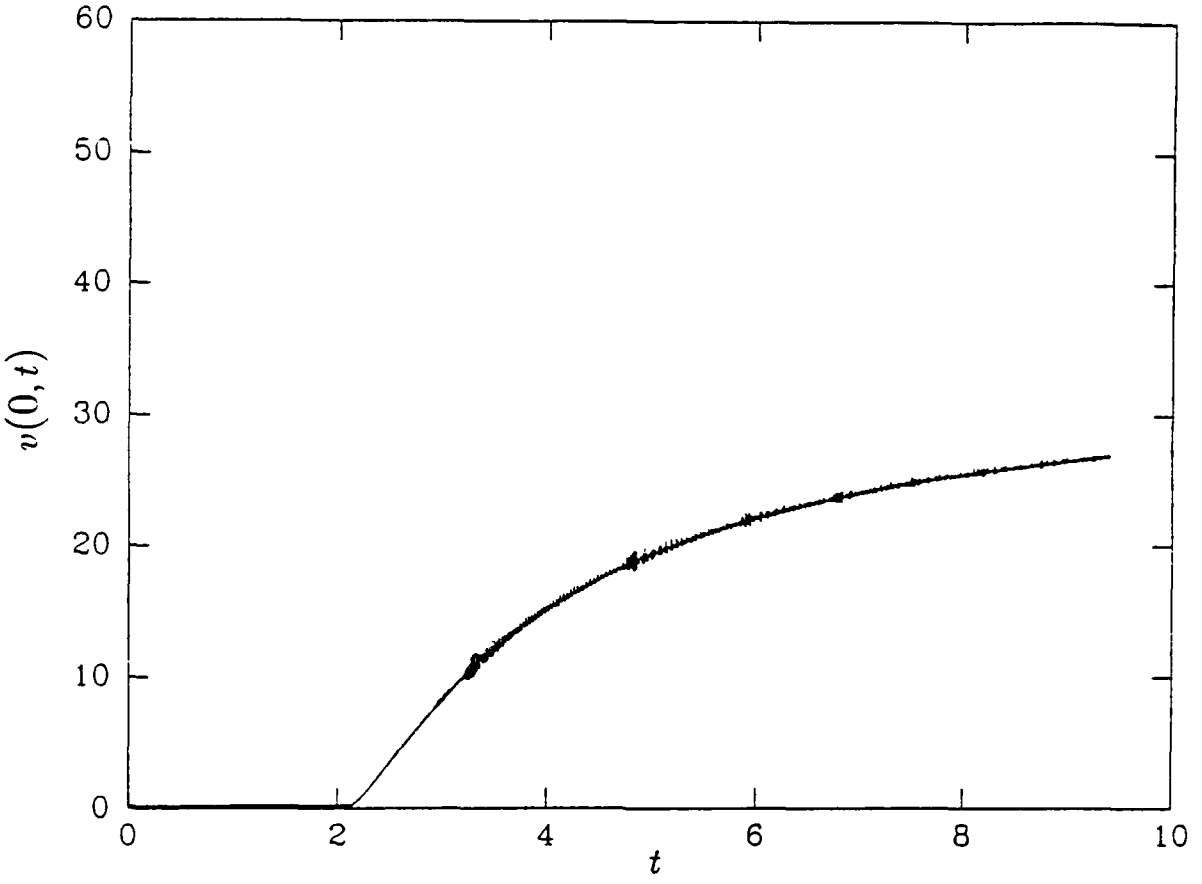


Fig. 6: Centerline velocity *vs.* time.

These observations and the presumption that the secondary relaxation time and its associated viscosity are independent of M, lead to a set of values of α and ϵ that decreases with M. These values are readily obtainable from our definitions in §3 and the dimensional information given in Ref. [8], where further details on parameter estimation may be found. The results are shown in Figs. 6–8.

Fig. 6 shows the evolution of the spurt process in time; centerline velocity is plotted *vs.* time for values of α and ϵ of sample PI-7 with $f = 1.2$. The simulations were carried out using zero initial data. The spatial discretization was a graded mesh with smaller elements near the wall, consisting of 640 elements. The maximum velocity in Fig. 6 is scaled by a Newtonian viscous response with viscosity ϵ that happens on such a short time-scale as not to be distinguishable from the $t = 0$ axis. The period of time from start-up to the onset of spurt at $t = 2.17$ in nondimensional units is the latency period in which a quasi-steady flow exists. Rescaling with appropriate dimensions gives the prediction of a latency time of 346 *sec.* for sample PI-7. The spurt process in Fig. 6 has not been carried out for a long enough time to achieve a very nearly steady state; numerical simulations which run for about 5 more nondimensional time units would be required. Thus we predict that the whole dynamic process takes on the order of forty minutes to unfold for this sample. We have run a sequence of such simulations for each of the eight samples, allowing a sufficiently long

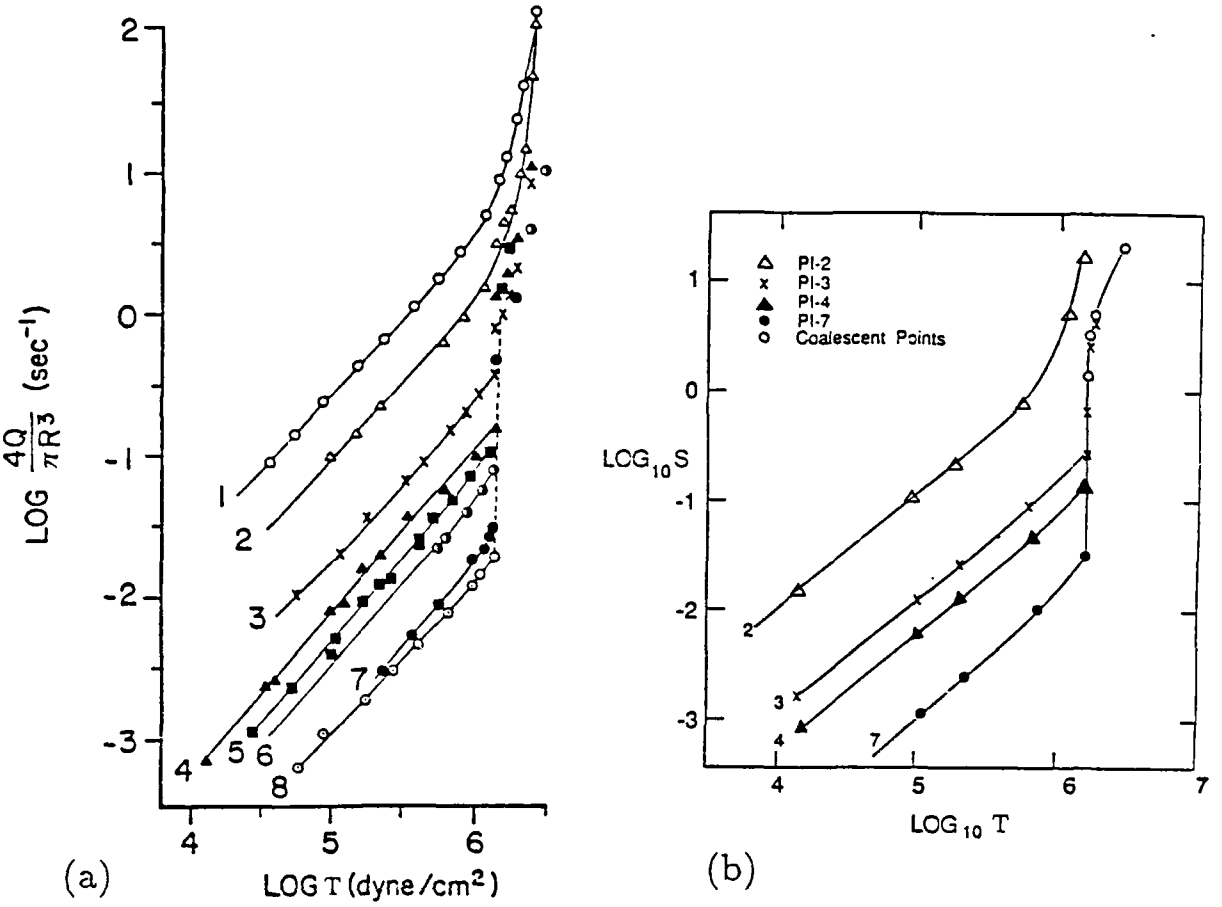


Fig. 7: Volumetric flow rate *vs.* effective shear stress: (a) experiment [18]; (b) numerical calculation [8]. Note that the horizontal scale of this panel matches that of panel (a), but the vertical scale does not.

time to obtain essentially steady solutions. Fig. 7 shows the results of these simulations compared to the data reported in Ref. [18]; volumetric flow rate, normalized in such a way that it has units comparable to shear rate [8], is plotted *vs.* T at the die or capillary wall. The value of T can be deduced by knowing the pressure drop and using the relation of Eq. 3.11.

Fig. 8 shows the result of simulating a loading sequence in which the pressure gradient f is increased in small steps, allowing sufficient time between steps to achieve steady flow [8]. The loading sequence is followed by a similar unloading sequence, in which the driving gradient is decreased in steps. The initial step used zero initial data, and succeeding steps used the results of the previous steps as initial data. The resulting hysteresis loop exhibits features similar to those observed by Hunter and Slemrod in their model [4] (see system 4.5 above), which they called "shape memory." The width of the hysteresis loop at the bottom can be related directly to the molecular weight of the sample [8].

We have performed careful numerical experiments to test the validity of the results we report here. One of the questions we sought to resolve involves the oscillations evident

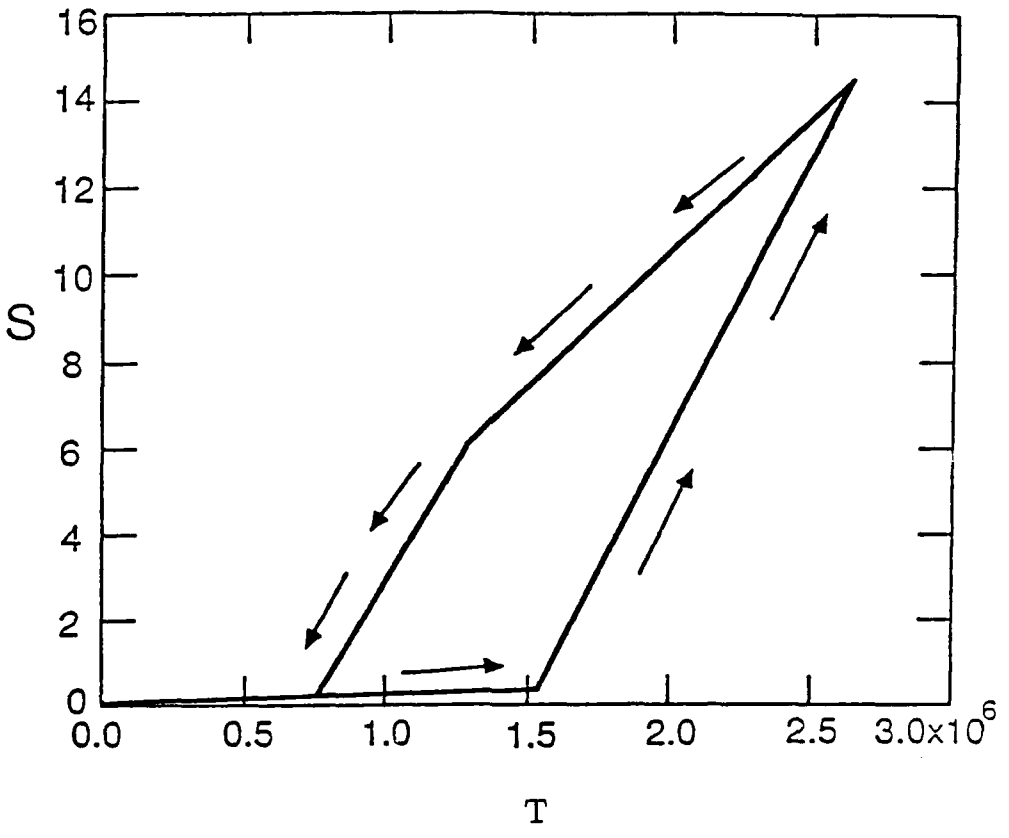


Fig. 8: Hysteresis under cyclic loading.

in Fig. 6 during the spurt process. In Ref. [8], results were reported on meshes much cruder than the one used to compute the results of Fig. 6; the oscillations were larger in amplitude which did not diminish with refinement of time step. Fig. 6 shows that these oscillations diminish with refinement of the grid size, and we are led to conclude that the oscillations reported in Ref. [8] are induced by spatial discretization error. This conclusion is reinforced by inspection of Fig. 6; the larger oscillations occur at later times, when the layer boundary moves toward the interior of the die where the elements of the graded mesh are larger. Eventually, these larger oscillations are damped, as they are using meshes consisting entirely of larger elements. Our mesh refinement studies lead us to infer that crude spatial resolution can lead to spurious oscillations in spurt dynamics that oscillate about the correct mean value and lead to accurately represented steady states. These conclusions have been confirmed by reproducing the results just described using the parabolic formulation (system 5.1). The mesh was refined to 3072 equal-sized cells; it was found that there is a weak stability condition relating time step to cell size. If this condition is violated, the spurt appears to occur prematurely with the parabolic method on fine grids; however, when the time step is refined on the finest grid, the results obtained with the solid mechanics method and parabolic method agree to at least graphical accuracy, and both give virtually the same estimate of latency time.

In the region of parameter space characteristic of Vinogradov's data, much can be deduced about the features of system (JS) without recourse to computed results. The deductions which follow were, however, guided by detailed study of the results of numerical simulation. First, since α is so small (of order 10^{-12}), Eq. 3.11 holds virtually instantaneously and for all time. Thus in system 5.1, the first equation may be eliminated, and T becomes a parameter whose value at any point in the die is given by Eq.(3.11). The resulting system of two ODEs can be analyzed completely by a phase-plane analysis [10] that shows a single attractor for $0 < T \leq \frac{1}{2}$, giving pre-critical solutions not involving spurt. If $\frac{1}{2} < T \leq 1$ and $\epsilon < \frac{1}{8}$, there are two stable attractors: one not involving spurt and one in which spurt has taken place. Latency can be interpreted as the time during which the system stays near the first attractor, which we call the "latent attractor." When $T > 1$ and $\epsilon < \frac{1}{8}$ there is no latent attractor; this result can be confirmed numerically. Furthermore, for all fluids PI-3–8 that exhibit spurt, $\epsilon \ll 1$. An asymptotic expansion of the solution of the two ODEs in system 5.1 for small ϵ gives quantitative estimates of dynamic behavior during latency and of the resulting asymptotic steady states [10]. This asymptotic analysis can also predict flow rates in steady states (and thus reproduce Fig. 7(b)) and predict the shape memory in hysteresis. It is easy to see that Z satisfies the ODE

$$Z_t = -\frac{Z^2 + Z + T^2}{Z + 1} \tag{5.2}$$

at zero order in ϵ (i.e., to $O(\epsilon)$ accuracy) near the latent attractor. The latency time is the time during which σ retains a relatively constant value of approximately T and grows only at first order in ϵ, while Z grows at zero order, goverened by Eq. 5.2. An initial value of $Z = 0$ at $t = 0$ for Eq. 5.2 is not appropriate because the asymptotic expansion leading to that equation is only valid near the latent attractor; however, an early-time expansion can be developed that is valid during the initial Newtonian response alluded to in connection with Fig. 6 [10]. On this time scale (which is of order ϵ) σ_t and Z_t are $O(\epsilon^{-1})$ while σ and Z are $O(1)$. This leads to

$$Z \approx (1 - T^2)^{\frac{1}{2}} - 1 \quad \text{at} \quad t \approx 0 \tag{5.3}$$

When Eq. 5.3 is used as an initial condition for Eq. 5.2 at $t = 0$, the latency time is estimated as the time at which $Z = -1$. This may be calculated directly by solving Eq. 5.2 for $t = t(Z)$, see [10]. The result for the case of Fig. 6, where $|T| = f/2 = 0.6$, is a prediction of a latency time of 366 *sec.*, which compares very favorably to the value of 346 *sec.* obtained from the full simulation.

In Ref. [8], several possible experiments are suggested that could verify the interpretation of spurt put forward here; the key experiment suggested is the verification of the molecular-weight dependence of the widest point of the hysteresis loop of Fig. 8. We remark that the shape of this loop is a key feature of "shape memory" in that the loop always opens from the point at which unloading begins. This occurs as the solutions proceed from "top-jumping" in Fig. 2 through intermediate convexifications of the curve to "bottom-jumping" at the point where a discontinuity in slope can be seen in the back part of the loop in Fig. 8; this is in distinct contrast to the interpretation of Ref. [12], where

bottom-jumping is always the rule for steady spurt solutions, and portions of the hysteresis loop are retraced during unloading. To these experimentally verifiable signatures of our model, the current analysis allows us to add more: When α and ϵ are sufficiently small, as they are with PI-3–8, latency time should be rather easy to measure, since a very slow flow with comparatively little throughput can persist for many minutes before dramatic growth occurs. We predict that latency can only occur for samples with ϵ sufficiently small ($\frac{1}{8}$ or less for J-S, but the precise number may be model-sensitive). For J-S, it can only occur for stresses in the range $\frac{1}{2} < T < 1$. It should scale with λ^{-1} at fixed T and obey Eqs. (5.2) and (5.3) approximately.

6. Conclusions

Well-posed dynamical problems based on non-monotone constitutive relations need not be unphysical. In fact, our Johnson-Segalman model provides a relatively simple example which accurately describes spurt. Other models, based on more sophisticated molecular theory, appear to have similar features [8] which require further investigation. In addition to reproducing spurt, our approach leads to results which suggest new experiments, as discussed in §5. Our numerical approaches to the fully time-dependent problem for a J-S fluid at a high Deborah number avoid the "high Weissenberg/Deborah number problem" at least in 1-D. We are currently investigating generalizations of our approaches to multi-D problems of physical interest.

Acknowledgements: We thank R. W. Kolkka, W. G. Pritchard, and A. E. Tzavaras for many helpful discussions. We are also grateful to M. Yao for his help with the figures.

References

1. R. Bird, R. Armstrong and O. Hassager, *Dynamics of Polymeric Liquids*, John Wiley and Sons, New York, 1987.

2. M. Doi and S. Edwards, "Dynamics of Concentrated Polymer Systems," *J. Chem. Soc. Faraday* **74** (1978), pp. 1789–1832.

3. C. Guillopé and J.-C. Saut, "Global Existence and One-Dimensional Nonlinear Stability of Shearing Motions of Viscoelastic Fluids of Oldroyd Type," *Math. Mod. Num. Anal.*, 1988. To appear.

4. J. Hunter and M. Slemrod, "Viscoelastic Fluid Flow Exhibiting Hysteretic Phase Changes," *Phys. Fluids* **26** (1983), pp. 2345–2351.

5. E. Isaacson, D. Marchesin and B. Plohr, "Construction of Nonlinear Waves for Conservation Laws," in preparation, 1988.

6. D. Joseph, M. Renardy, and J.-C. Saut, "Hyperbolicity and Change of Type in the Flow of Viscoelastic Fluids," *Arch. Rat. Mech. Anal.* **87** (1985), pp. 213–251.

7. M. Johnson and D. Segalman, "A Model for Viscoelastic Fluid Behavior which Allows Non-Affine Deformation," *J. Non-Newtonian Fluid Mech.* **2** (1977), pp. 255–270.

8. R. Kolkka, D. Malkus, M. Hansen, G. Ierley and R. Worthing, "Spurt Phenomena of the Johnson-Segalman Fluid and Related Models," *J. Non-Newtonian Fluid Mech.*, 1988. In press.

9. Y.-H. Lin, "Explanation for Slip-Stick Melt Fracture in Terms of Molecular Dynamics in Polymer Melts," *J. Rheol.* **29** (1985), pp. 609–637.

10. D. Malkus, J. Nohel, B. Plohr, "Phase Plane and Asymptotic Analysis of Spurt Phenomena," in preparation, 1988.

11. D. Malkus Y.-C. Tsai, "Stability Analysis of Implicit-Explicit Time Integration for Viscoelastic Flow," in preparation, 1988.

12. T. McLeish and R. Ball, "A Molecular Approach to the Spurt Effect in Polymer Melt Flow," *J. Polymer Sci.* **24** (1986), pp. 1735–1745.

13. R. Menikoff and B. Plohr, "The Riemann Problem for Fluid Flow of Real Materials," *Rev. Mod. Phys.*, 1989. To appear.

14. J. Nohel, R. Pego, and A. Tzavaras, "Stability of Discontinuous Shearing Motions of Non-Newtonian Fluids," in preparation, 1988.

15. B. Plohr, "Shockless Acceleration of Thin Plates Modeled by a Tracked Random Choice Method," *AIAA J.* **26** (1988), pp. 470–478.

16. B. Plohr, "Instabilities in Shear Flow of Viscoelastic Fluids with Fading Memory," in *Workshop on Partial Differential Equations and Continuum Models of Phase Transitions (Nice, 1988)*, ed. D. Serre, Springer–Verlag, New York, 1988. Lecture Notes in Mathematics.

17. M. Renardy, W. Hrusa and J. Nohel, *Mathematical Problems in Viscoelasticity*, Pitman Monographs and Surveys in Pure and Applied Mathematics, Vol. 35, Longman Scientific & Technical, Essex, England, 1987.

18. G. Vinogradov, A. Malkin, Yu. Yanovskii, E. Borisenkova, B. Yarlykov and G. Berezhnaya. "Viscoelastic Properties and Flow of Narrow Distribution Polybutadienes and Polyisoprenes," *J. Polymer Sci., Part A-2* **10** (1972), pp. 1061–1084.

POLYNOMINAL DEFINITION OF DISCRETE FIELD POINT OF MAP OF DIFFUSION EQUATION - PART II

William F. Donovan
Mechanics and Structures Branch
Interior Ballistics Division
Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

ABSTRACT

The one-dimensional diffusion equation, $\frac{\partial T}{\partial t} = a \frac{\partial^2 T}{\partial x^2}$, is transposed into algebraic expression (Part 1) as

$$T(N,P) = \frac{\phi}{2^h 2^j} (A - Bm + Cm^2 + \cdots \neq m^k)$$

where A,B,C, etc. are whole numbers. It is developed that correlations in sequenced A,B,C, etc. coeffcients are Pascal Triangle and/or arithematic square terms. Furthermore, these terms are expressible as Pokhammer relations.

INTRODUCTION

Part 1 of this report indicated a novel solution to the basic diffusion equation of Physics where the field boundary extends from zero to positive infinity. The nodal points of the field net are identified as terminating polynomials with the numerators of the coefficients found first by deduction - for the lower orders - and then by extrapolation. Part II considers the numerical analysis employed to complete the entire set of tables.

PROCEDURE

A. Direct Differences

Formulation of the polynomial form of the discrete solutions, Eq. (1), of the diffusion equation from the Schmidt plot geometry is described in References (1) and (2) and reflects a progressive trigonometric construction where the degree and term extension increases with time and decreases with distance, time and distance referring to the unsteady heat flow application.

$$T(N,P) = \frac{\phi}{2^h 2^j}\left(A - A_1 m + A_2 m^2 - A_3 m^3 + A_4 m^4 - \cdots \mp A_k m^k \right) \quad (1)$$

where m is the independent variable,

N is a distance index,

P is a time index,

$$h = \frac{(P+N)-2-\left|\sin (P+N)\pi/2\right|}{2},$$

j = (k - term exponent of m), the individual term denominator exponent, and $\phi = m(T_0 - T_1 0)$, with

$A, A_1, A_2, \ldots\ldots A_k$ the numerical coefficients of the interior terms of the equation.

The numerators of each term of the "nodal" equations, are uniquely related to adjacent time and distance term coefficients of the same degrees. This relation, originally found accidentally, is correlated by the table of differences shown in Table 1 & 1-A where the boxed vertical sequence; 17548, 25147, 35401, 49024 and 66868, is established by the reduction of the Schmidt plot through the trigonometric analysis. Step-wise right moving subtraction generates a column of residual zeroes, an adjoining column of ones, and a digital sequence identified as "IV" in Table 1. Reducing this column "IV" to zero vertically then allows a corresponding determination of the particular values of the entire matrix from inspection of the biased rows I, II and III.

1. William F. Donovan, "Determination of Heat Transfer Coefficient in a Gun Barrel from Experimental Data," Memorandum Report BRL-MR-3428, January 1985 (AD# A151815)

2. William F. Donovan, "Polynomial Definition of Discrete Field Points of Map of Diffusion Equation, Part I," Memorandum Report BRL-MR-3649.

Table 1. Difference Progression for Degree Zero Numerators

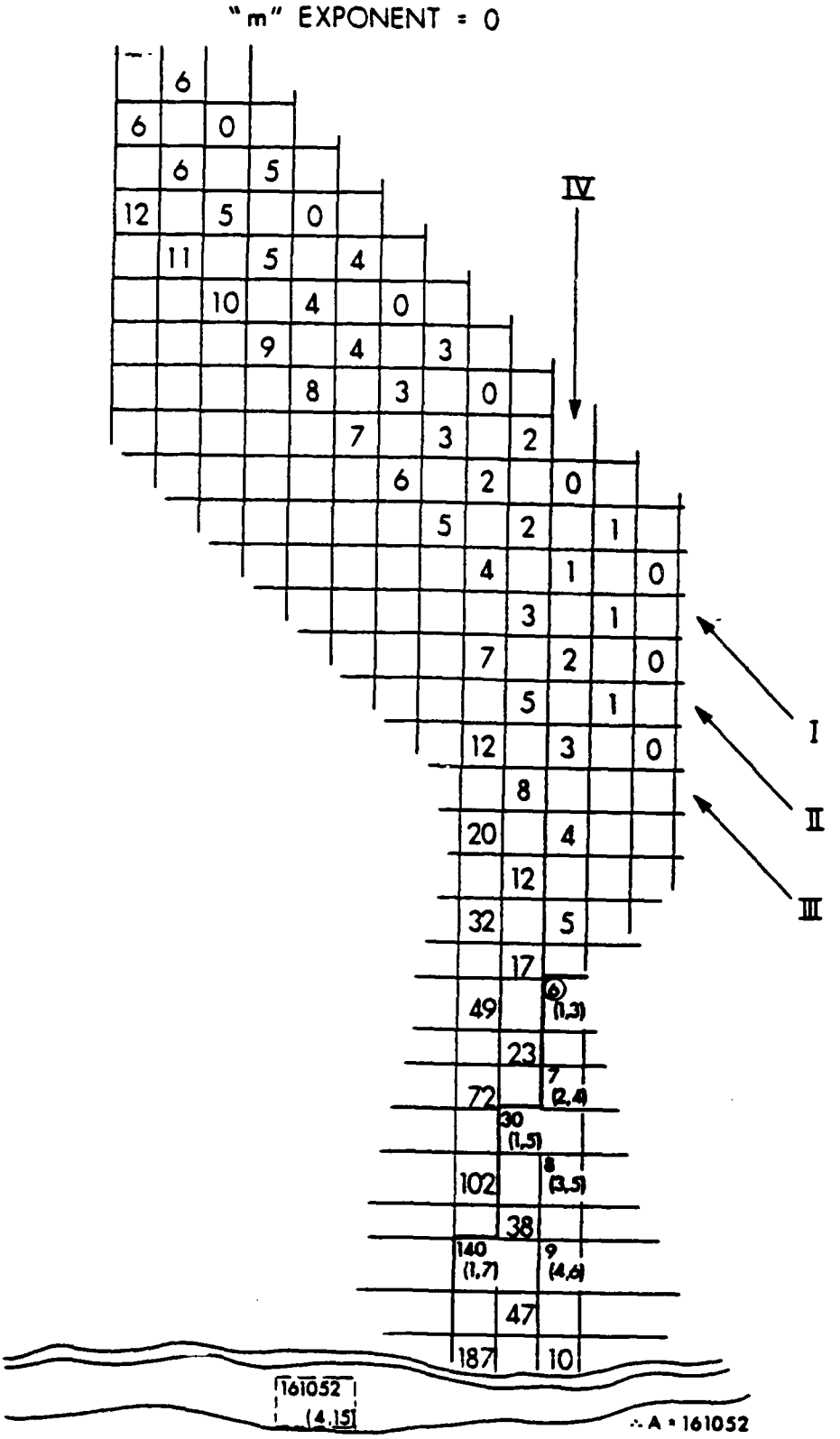


Table 1-a (continuation of Table 1)

					630 (1,9)		57 (4,8)			
				1898		244 (3,9)		11 (6,8)		0
			5282		874 (2,10)		68 (5,9)		1 (8,8)	
		13866		2772 (1,11)		312 (4,10)		12 (7,9)		0
			8054		1186 (3,11)		80 (6,10)		1 (9,9)	
		21920		3958 (2,12)		392 (5,11)		13 (8,10)		0
			12012 (1,13)		1578 (4,12)		93 (7,11)		1 (10,10)	
		33932		5536 (3,13)		485 (6,12)		14 (9,11)		0
		90683		17548 (2,14)		2063 (5,13)		107 (8,12)		1 (11,11)
	232009		51480 (1,15)		7599 (4,14)		592 (7,13)		15 (10,12)	0
572312		142163		25147 (3,15)		2655 (6,14)		122 (9,13)		1 (12,12)
	374172		76627 (2,16)		10254 (5,15)		714 (8,14)		16 (11,13)	0
946484		213790 (1,17)		35401 (4,16)		3369 (7,15)		138 (10,14)		1 (13,13)
	592962		112028 (3,17)		13623 (6,16)		852 (9,15)		17 (12,14)	0
1539446		330818 (2,18)		49024 (5,17)		4221 (8,16)		155 (11,15)		1 (14,14)
	923780 (1,19)		161052 (4,18)		17844 (7,17)		1007 (10,16)		18 3,15)	0
2463226		491870 (3,19)		66868 (6,18)		5228 (9,17)		173 (12,16)		1 (15,15)
						1180 (11,17)		19 (14,16)		0
							192 (13,17)			

B. Summing Progression

Rewriting Table 1 as Table 2, where the biased rows are horizontal and the zero column is left-justified instead of right-justified, shows that the sum of any two adjacent column values of any row gives the value of the next row entry directly under the right-wise addendum. The sequence of Row I, regardless of the degree of the term represented by the matrix, always starts with zero and then maintains alternate zeroes to infinity. Each of the difference tables corresponding to a given "m" exponent (Eq. (1)) can be resolved similarly, except that each first row is carried in a unique progression. This progression is repeated without the interspersed zeroes within the matrix.

C. Pascal Triangles

The classical Pascal triangle, Table 3, can be formed by simple addition where each term is the sum of the two previous superior terms and individual entries are represented by the binomial coefficient, $\binom{z}{w} = \frac{z!}{w!(z-w)!}$, where z and w represent the row and column of a particular coefficient of a binomial expansion. A modification of the Pascal triangle is found by writing the diagonals as rows which then generates the "arithmetic square", Table 4, also known historically.[3]

It is precisely these progressions alternating with zeroes, which comprise the first rows of the individual "summing progressions" shown as Table 2. In this case the digital enumeration of the rows indicates the degree of the "m" term.

D. Pockhammer's Symbol

Each row of Table 4 can be examined by finite differencing to establish, via Gregory-Newton,[4] a definitive polynomial expansion extending to infinity. Furthermore, the algebraic equation can be simplified to a factorial form known as Pockhammer's Symbol or as a π factorial. Appendix A presents an example of such a development for "m" degrees zero through three. The complete arithmetic square, Table 2, can then be written as

$$f(v) = \frac{1}{(r+1)!} \ (v)_r \qquad (2)$$

where

r = degree of m,
v = column value

and f(v) = row value of Table 2. From Table 2, a complete construction of Table 1 follows.

---

[3] N.Ya. Vilenkin, Combinatorics, Academic Press, New York & London, 1971, pp 90-94.

[4]. Spiegel, M. R., Theory and Problems of Finite Differences and Finite Difference Equations, Schaum's Outline Series in Mathematics, McGraw-Hill Book Company, New York, etc. 1971, p.p. 36-44.

Table 2. Summing Progression for Degree zero Numerators

"m" EXPONENT = 0

	COLUMN →																
ROW↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	1	0	2	0	3	0	4	0	5	0	6	0	7	0	8	0
2	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	0	1	3	5	7	9	11	13	15	17	19	21	23	25	27	29	31
5	0	1	4	8	12	16	20	24	28	32	36	40	44	48	52	56	60
6	0	1	5	12	20	28	36	44	52	60	68	76	84	92	100	108	116
7	0	1	6	17	32	48	64	80	96	112	128	144	160	176	192	208	224
8	0	1	7	23	49	80	112	144	176	208	240	272	304	336	368	400	432
9	0	1	8	30	72	129	192	256	320	384	448	512	576	640	704	768	832
10	0	1	9	38	102	201	321	448	576	704	832	960	1084	1216	1344	1472	1600

Table 3. Pascal Triangle

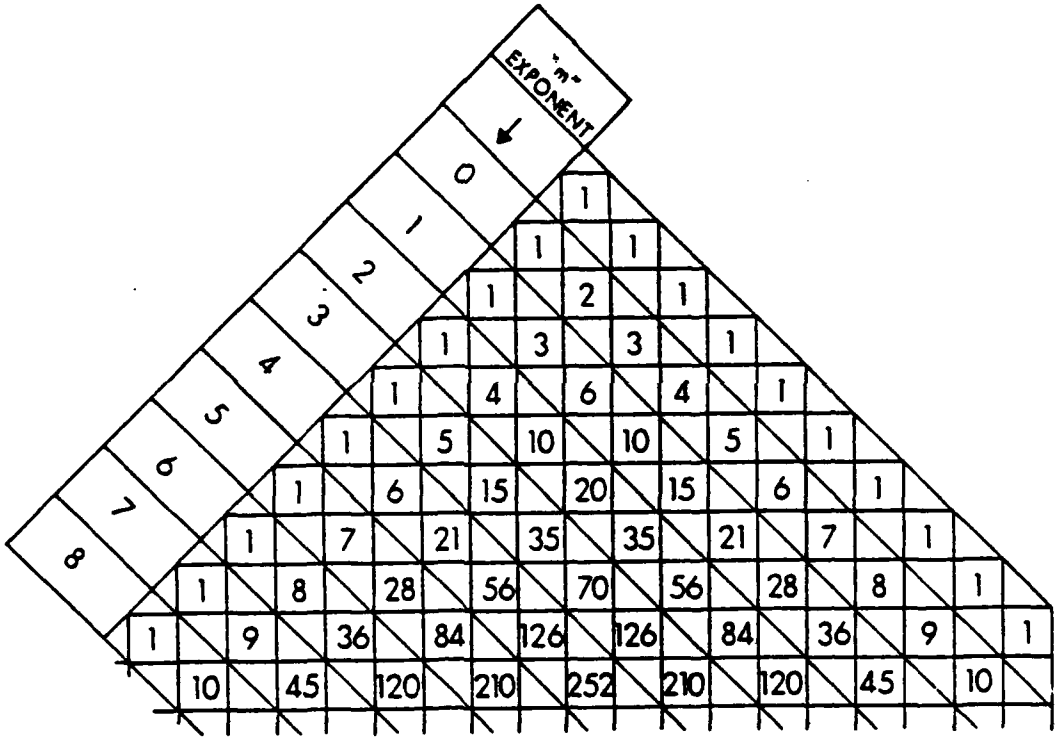


Table 4. Arithmetic Square

"m" EXPONENT ↓	BASELINE SEQUENCE →										
	1	1	1	1	1	1	1	1	1	1	1
0	1	2	3	4	5	6	7	8	9	10	11
1	1	3	6	10	15	21	28	36	45	55	66
2	1	4	10	20	35	56	84	120	165	220	286
3	1	5	15	35	70	126	210	330	495	715	1001
4	1	6	21	56	126	252	462	792	1287	2002	3003
5	1	7	28	84	210	462	924	1716	3003	5005	8008
6	1	8	36	120	330	792	1716	3432	6435	10010	18018
7	1	9	45	165	495	1287	3003	6435	12870	22880	40898

REFERENCES

1. William F. Donovan, "Determination of Heat Transfer Coefficient in a Gun Barrel from Experimental Data," Memorandum Report BRL-MR-3428, January 1985 (AD#

2. William F. Donovan, "Polynomial Definition of Discrete Field Points of Map Diffusion Equation, Part I," Memorandum Report BRL-MR-3649

3. N.Ya. Vilenkin, Combinatorics, Academic Press, New York and Long, 1971, p.p. 90-94.

4. Spiegel, M.R., "Theory and Problems of Finite Differences and Finite Difference Equations, Schaum's Outline Series in Mathematics, McGraw-Hill Book Company, N.Y., etc., 1971, p.p. 36-44.

Summary

A straight-forward method (arithmetic squares) is described to permit the numerical construction of the differencing tables of Part I of this report. The derivation through the Pascal triangle and correspondence to Pockhammer's Symbol notation is demonstrated.

APPENDIX A

Determination of Pockhammer's Notation

A discussion of the Gregory-Newton analysis is presented in Reference 4. It consists of determining a polynomial expression to represent a progressive sequence of numbers. In the present application, it is used to examine the base row development of Table 4.

Given a unit stepping difference in a counting reference, v, and a matched sequence f(y);

$$f(y) = f(v) + \frac{\Delta f(v)}{1!} y^{(1)} + \frac{\Delta^2 f(v)}{2!} y^{(2)} + \frac{\Delta^3 f(v)}{3!} y^{(3)} + \cdots\cdots$$

where

v is the step level,

f(y) is the dependent variable,

$\Delta^m f(v)$ are the diagonal values of the difference table, and

$y^{(0)} = 1$

$y^{(1)} = y$

$y^{(2)} = y(y-1)$

$y^{(3)} = y(y-1)\ (y-2) \cdots$

etc.

For the case of "m" degree zero where f(y) is from Table 4:

v	f(y)	f(v)	Δf(v)
0	1		
		1	
1	2		0
		1	
2	3		0
		1	
3	4		0
		1	
4	5		0
		1	
5	6		0
		1	
6	7		0
		1	
7	8		0
		1	
8	9		0
		1	
9	10		0
		1	
10	11		0
		1	
11	12		0

$$f(y) = f(v) + \frac{\Delta f(v)}{1!}\ y^{(1)} \cdots\cdots$$

$$= 1 + v + 0 \cdots\cdots$$

$$= (v + 1)$$

For m degree 1:

v	f(y)	f(v)	Δf(v)	$\Delta^2 f(v)$
0	1			
		2		
1	3		1	
		3		0
2	6		1	
		4		0
3	10		1	
		5		0
4	15		1	

$$f(y) = f(v) + \frac{\Delta f(v)}{1!}\ y^{(1)} + \frac{\Delta^2 f(v)}{2!}\ y^{(2)} + \cdots\cdots$$

$$= 1 + 2v + \frac{1}{2} v(v-1) + 0 + \cdots\cdots$$

$$= \frac{1}{2} (v^2 + 3v + 2)$$

$$= \frac{1}{2} (v + 1) (v + 2)$$

For m degree 2

v	f(y)	f(v)	$\Delta f(v)$	$\Delta^2 f(v)$	$\Delta^3 f(v)$
0	1				
		3			
1	4		3		
		6		1	
2	10		4		0
		10		1	
3	20		5		0
		15		1	
4	35		6		
		21			
5	56				

$$= 1 + 3v + \frac{3}{2} (v)(v-1) + \frac{1}{6} (v) (v-1) (v-2) + 0 \cdots\cdots$$

$$= 1 + 3v + \frac{3}{2} (v^2 - v) + \frac{v}{6} (v^2 - 3v + 2)$$

$$= \frac{1}{6} (v^3 + 6v^2 + 11v + 6)$$

$$= \frac{1}{6} (v + 1) (v + 2) (v + 3)$$

The pattern continues so that:

Degree of "m" = r	f(y)
0	$(v + 1)$
1	$\frac{1}{2} (v + 1) (v + 2)$
2	$\frac{1}{6} (v + 1) (v + 2) (v + 3)$
3	$\frac{1}{24} (v + 1) (v + 2) (v + 3) (v + 4)$
4	$\frac{1}{120} (v + 1) (v + 2) (v + 3) (v + 4) (v + 5)$

$$5 \qquad \frac{1}{720} (v + 1)(v + 2)(v + 3)(v + 4)(v + 5)(v + 6)$$

and the general expression is

$$f(y) = \frac{1}{(r + 1)!} (v)_{r+1}$$

where $(v)_{r+1} = (v + 1)(v + 2)(v + 3) \ldots . (v + r + 1)$.

With respect to the original time index, P, of the diffusion equation polynomial; $v = P-1$ by Table 3, and

$$(v)_{r+1} = (P - 1 + 1)(P - 1 + 2) \ldots . (P + r)$$

$$= P (P + 1)(P + 2)(P + 3) \ldots .(P + r)$$

so that

$$f(y) = \frac{1}{(r + 1)!} (P)_r$$

which is known as Pockhammer's Symbol.*

*Gravio A. Korn, Mathematical Handbook for Scientists and Engineers, McGraw-Hill Book Co., Inc., New York, etc., 1961.

List of Symbols

h	exponent of 2 in external denominator
j	exponent of 2 in each term denominator
k	exponent of "m" in final term
m	independent variable
r	degree of "m"
v	column value of stepping sequence
f(v)	row value of stepping sequence
w	inferior component of binomial coefficient
z	superior component of binomial coefficient
A_1, A_2, A_3, ...	numerical coefficients
N	distance index
P	time index
T	dependent variable
ϕ	external numerator

Annulus-based Inclusion Testing for Multiply-Connected Sets

Terence M. Cronin
US Army CECOM Center for Signals Warfare
Vint Hill Farms Station
Warrenton VA 22186

Abstract: A new data structure is introduced, as a vehicle to test for metrical inclusion of an arbitrary point within a potentially multiply-connected closed curve. From graph theory and topology, we know that if a line drawn from a point through a simply-connected closed contour non-tangentially intersects the contour an even number of times, then the point is on the exterior; otherwise it lies within the interior (the parity algorithm). This theoretical result is powerful, but fails for multiply-connected curves. It also does not provide information about either the distance or direction from a point to a contour. A proposed solution incorporates a new structure called the *inner annulus*, which is computed with a corresponding generator function, using as input the digital representation of a closed Jordan curve. The annulus can be viewed as the set of points which are 4-connected to the inside edge of the contour. It is algorithmically generated by traversing the inside edge of the contour in a counterclockwise fashion, and collecting the pixels visited, until the start point is seen again. During this process, the interior of the contour is *always to the left*. Once the annulus is constructed, a test is made to determine if an arbitrary coordinate is nearer the original contour or its inner annulus, which determines respectively whether the point is exterior or interior to the contour. The same technique may be applied to any holes contained in the contour, so that multiply-connected sets are accommodated. The computational complexity of the algorithm is analyzed in terms of time, space, and preprocessing requirements. A conjecture is posed, asserting the length of the inner annulus in terms of the length and shape of the original contour. An attempt to prove the conjecture has yielded a formal characterization of the shape of a contour, in terms of the convexities and concavities exhibited by the boundary of the contour.

Introduction: High level map reasoning is a problem area which lacks a complete mathematical formalization. This paper describes a new tool, termed *annulus-based inclusion testing*, built upon a foundation of topology and geometry. The tool addresses only one small portion of the spatial reasoning problem: that of metrical interior/exterior region discrimination for multiply-connected contours. Any holes contained in the parent contour are cross-referenced by a graph-theoretic structure called a *feature orientation lattice*. Another tool under development, called *equidistance loci-reduction*, is designed to rapidly render the contour nearest an arbitrary map coordinate, as well as the relative direction and distance to the coordinate [C1]. The flow of logic is as follows. Loci-reduction serves as a filter to rapidly pinpoint the nearest topological feature to an arbitrary point. If the feature is a closed contour, annulus-based inclusion testing decides whether the point lies within the boundary of the feature, and if so, *where* inside. The feature orientation lattice provides subordination and orientation relationships among the parent contour and any contours which it may contain. Figure 1 illustrates the way in which the tools are used as feeder technologies for automated map reasoning.

1.0 Problem Definition: Metrical Inclusion.

Given a digitized representation of a possibly multiply-connected closed Jordan curve and an arbitrary point, develop a reliable, computationally efficient technique to decide whether the point is interior or

exterior to the curve, while at the same time providing the respective distance and direction from the curve and any contours which it may contain.

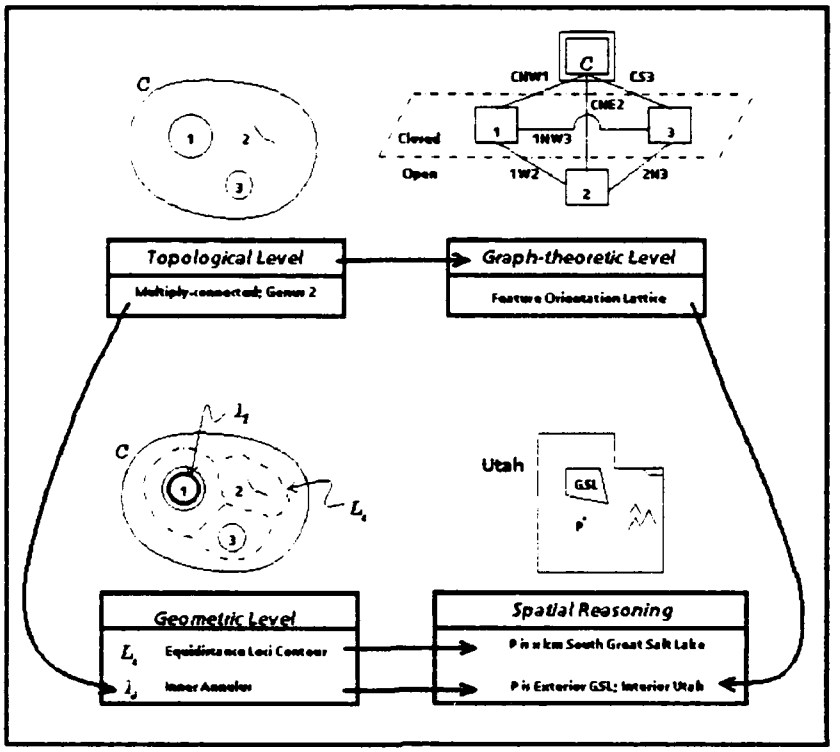


Figure 1. Mathematical Tools to Support Automated Map Understanding

1.1 Distance Calculations on a Binary Map.

There are several distance metrics discussed in the digital image processing literature. A good compendium of these is contained in [G4]. The distance used here is the city-block distance, also referred to as the d_4 distance [R2]. This metric is used for point-to-point, point-to-contour, and contour-to-contour distance measurements. The discussion which follows refers to trace contours; a trace contour is the boundary of a spatial feature represented on a binary map.

Definition 1.1.1. Distance Between Two Points.

Let $p_1 = (x_1,y_1)$ and $p_2 = (x_2,y_2)$ be arbitrary points. Then the **d_4 distance between p_1 and p_2** is defined:

$$d_4[p_1,p_2] = |x_1 - x_2| + |y_1 - y_2|.$$

Definition 1.1.2. Distance from a Point to a Contour.

Let T_j be a trace contour, and (x,y) an arbitrary coordinate. Then the **distance from (x,y) to T_j** is defined:

$$d_4[(x,y), T_j] = \min \{ |x_i - x| + |y_i - y| \}, \forall \text{ points } (x_i,y_i) \in T_j.$$

Definition 1.1.3. Distance Between Two Contours.

Let T_i and T_j be trace contours. Then the **distance from T_i to T_j** is defined:

$$d_4[T_i, T_j] = \min \{d_4[(x,y), T_j]\} \; \forall \text{ points } (x,y) \in T_i.$$

1.2 Computing the Relative Direction of an Arbitrary Point From a Contour.

Let (x_1,y_1) and (x_2,y_2) be arbitrary points. Let $\Delta x = x_1 - x_2$ and $\Delta y = y_1 - y_2$. Then the **relative direction of (x_1,y_1) from (x_2,y_2), denoted dir[(x_1,y_1),(x_2,y_2)]**, is defined piecemeal by the following function:

Conditions	dir[(x_1,y_1), (x_2,y_2)]
$\Delta x > 0$ and $\Delta y > 0$	NE
$\Delta x > 0$ and $\Delta y = 0$	E
$\Delta x > 0$ and $\Delta y < 0$	SE
$\Delta x = 0$ and $\Delta y < 0$	S
$\Delta x < 0$ and $\Delta y < 0$	SW
$\Delta x < 0$ and $\Delta y = 0$	W
$\Delta x < 0$ and $\Delta y > 0$	NW
$\Delta x = 0$ and $\Delta y > 0$	N

Definition 1.2. The **relative direction of a point (x,y) from a contour T, denoted dir[(x,y),T], is defined:**

$$\text{dir}[(x,y),T] = \text{dir}\,[(x,y), (x_c,y_c) \mid (x_c,y_c) \in T;\; D_4[(x,y),T] = D_4[(x,y),(x_c,y_c)]]$$

2.0 Automated Discrimination of the Interior and Exterior of a Closed Contour.

2.1 Other Approaches to the Problem.

Deciding if a point lies inside or outside a closed contour is a problem for which implementation is non-trivial, especially when the caveat is added that the contour may be multiply-connected. There are other techniques which address elements of Problem Definition 1.0. However, these approaches are of limited utility. One such technique is the odd-even count contour-crossing technique, also known as the parity algorithm [S1]. A potentially powerful approach to the problem, this technique "draws a line" from the coordinate in question "through" a contour, and counts the number of times the contour is crossed. The technique answers in the affirmative if the number of crossings is odd, and in the negative otherwise. This is an example of a technique which avails itself of elegant theoretical results from graph theory and topology, but for which implementation is fraught with error. The problems with the technique are evident from the words which are double-quoted. First, an implementation must accommodate drawing a line in the appropriate direction to "appropriately" intersect the contour. Digitized contours of general complexity are frequently convoluted in such a way that a true crossing is not detected, or a false crossing is counted as a valid one. Secondly, if a contour is multiply-connected, it is possible for a point lying within one of the "holes" to be considered outside the parent contour, which is a theoretically correct response, but is a failure from a pragmatic stance. This can have serious ramifications in practice - for example, a boat on Great Salt Lake would not be contained within the state of Utah. Another shortcoming is that the decision relies upon a simplistic count, and returns no metrical information about distance or direction from contours.

2.2 The Common Sense Logic Behind the Inner Annulus Discriminator Technique.

Intuitively, the inner annulus of a closed contour is a set of points adjoining the inner edge of the contour. The annulus is chosen to lie inside the contour because it is more computationally efficient than constructing it outside (an inner track is shorter). The sole purpose of generating the inner annulus is to create a memory-efficient computing technique to differentiate between the interior and exterior of the contour. Simply stated, a check is made to determine if the point is nearer the contour or its inner annulus. If it is nearer the annulus, it is decided that the point is inside the contour; otherwise, it is outside (Figure 2).

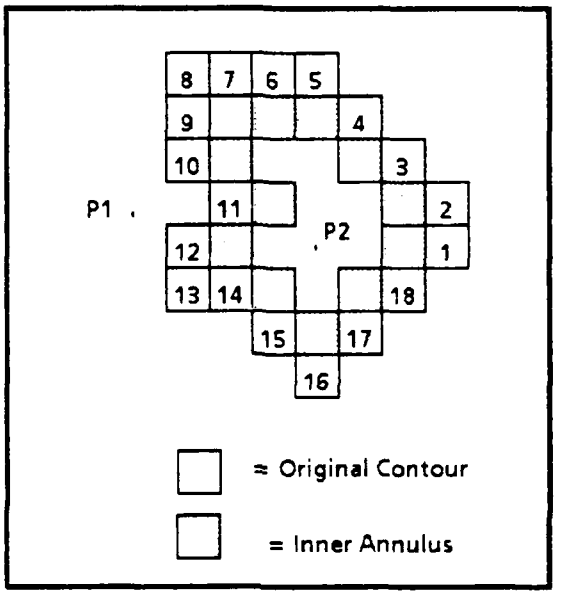


Figure 2. A graphic to illustrate the utility of the inner annulus of a closed contour. Point P2 is nearer the annulus and therefore inside; conversely, P1 is outside.

2.3 How Multiply-Connected Contours are Accommodated by the Inner Annulus Technique.

A hole contained in a closed contour is itself a closed contour; therefore it possesses an inner annulus. Equidistance loci reduction (see discussion at 1.0) tells us if a point is nearer a multiply-connected parent contour or one of the holes it contains. If the point is nearer the parent contour, a test is made to determine if the point is nearer the contour or its inner annulus, to assert exterior, or interior respectively. If the point is nearer a hole, the test is performed on the hole and its inner annulus, to assert exterior or interior to the hole, respectively. *But interior to the hole means exterior to the parent contour*, from the definition of multiply-connected set. In practice, this paradox is circumvented by resorting to the feature orientation lattice, within which the topology of a multiply-connected set is embedded.

2.4 Automated Generation of the Inner Annulus of a Closed Contour.

Definition 2.4a. An **inner annulus generator function** I_g is a function with domain the closed contour T_j, and range I_j defined as follows:

a) Starting at an arbitrary point, order Tj in a *counterclockwise* direction and call the result Tj(n), where n is the length of Tj.

b) For i from 2 to n + 1, let $T_j(i) = (x_i, y_i)$, $T_j(i+1) = (x_{i+1}, y_{i+1})$, and $T_j(i-1) = (x_{i-1}, y_{i-1})$, $T_j(n+1) = T_j(1)$. Furthermore, let $\Delta x = x_i - x_{i-1}$ and $\Delta y = y_i - y_{i-1}$; $\Delta xnew = x_{i+1} - x_i$ and $\Delta ynew = y_{i+1} - y_i$.

Conditions	*Element of Ij Produced*
F1. $\Delta x > 0$ and $\Delta y \geq 0$	Insert $(x_{i-1}, y_{i-1} + 1)$
SC1. [and $\Delta y = 0$, $\Delta xnew \geq 0$, $\Delta ynew < 0$]	[Insert $(x_{i-1} + 1, y_{i-1} + 1)$]
SC2. [and $\Delta y > 0$, $\Delta xnew > 0$, $\Delta ynew < 0$]	[Insert $(x_{i-1} + 1, y_{i-1} + 2)$]
*ST**. [and $\Delta y = 0$, $\Delta xnew = 0$, $\Delta ynew > 0$]	[Remove (x_{i-1}, y_{i-1})]
F2. $\Delta x \geq 0$ and $\Delta y < 0$	Insert $(x_{i-1} + 1, y_{i-1})$
SC1. [and $\Delta x = 0$, $\Delta xnew < 0$, $\Delta ynew \leq 0$]	[Insert $(x_{i-1} + 1, y_{i-1} - 1)$]
SC2. [and $\Delta x > 0$, $\Delta xnew < 0$, $\Delta ynew < 0$]	[Insert $(x_{i-1} + 2, y_{i-1} - 1)$]
*ST**. [and $\Delta x = 0$, $\Delta xnew > 0$, $\Delta ynew = 0$]	[Remove (x_{i-1}, y_{i-1})]
F3. $\Delta x \leq 0$ and $\Delta y > 0$	Insert $(x_{i-1} - 1, y_{i-1})$
SC1. [and $\Delta x = 0$, $\Delta xnew > 0$, $\Delta ynew \geq 0$]	[Insert $(x_{i-1} - 1, y_{i-1} + 1)$]
SC2. [and $\Delta x < 0$, $\Delta xnew > 0$, $\Delta ynew > 0$]	[Insert $(x_{i-1} - 2, y_{i-1} + 1)$]
*ST**. [and $\Delta x = 0$, $\Delta xnew < 0$, $\Delta ynew = 0$]	[Remove (x_{i-1}, y_{i-1})]
F4. $\Delta x < 0$ and $\Delta y \leq 0$	Insert $(x_{i-1}, y_{i-1} - 1)$
SC1. [and $\Delta y = 0$, $\Delta xnew \leq 0$, $\Delta ynew > 0$]	[Insert $(x_{i-1} - 1, y_{i-1} - 1)$]
SC2. [and $\Delta y < 0$, $\Delta xnew < 0$, $\Delta ynew > 0$]	[Insert $(x_{i-1} - 1, y_{i-1} - 2)$]
*ST**. [and $\Delta y = 0$, $\Delta xnew = 0$, $\Delta ynew < 0$]	[Remove (x_{i-1}, y_{i-1})]

F1-F4 are first order operators.

S_{C1}, S_{C2}, and S_{T^*} are second order operators.

Definition 2.4b. The **inner annulus** I_j is the set of points produced by inner annulus generator function I_g operating on closed trace contour T_j.

2.5 The Interior-Exterior Discrimination Function.

Once the inner annulus of a closed curve is generated, the discrimination process is a simple comparison of distance to the curve with distance to its annulus.

Definition 2.5. A point (x, y) is said to be on the **exterior** of closed contour T_j, with inner annulus I_j iff $d_4[(x,y), T_j] < d_4[(x,y), I_j]$. Otherwise, (x, y) is said to be on the **interior** of T_j.

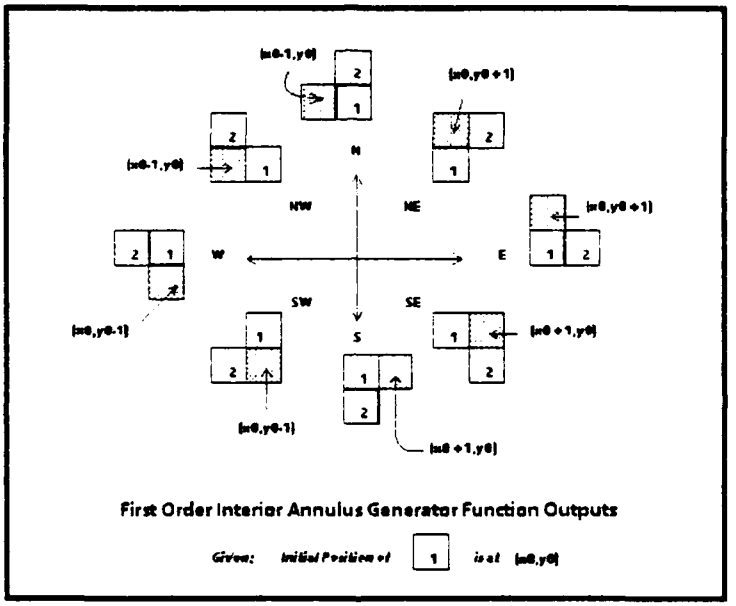


Figure 3. The first-order inner annulus generator operators. In actuality, there are four distinct outputs, since the following pairs of operators produce the same points: NE-E, SE-S, SW-W, and NW-N.

2.6 Concave Points Require Forceful Introduction: the Set C.

There are situations when the first order inner annulus generator operators are not sufficient to produce a continuous annulus - examples are the concavity situations in Figure 4 (a), (b), and (c). In such cases, to insure continuity, it is necessary to invoke second-order concavity operators to add points to the inner annulus. In case (4b), the discriminator commits an error unless the first order outputs are supplemented: if point III is not added to the annulus, then it is closer by one pixel to the original contour, resulting in the erroneous decision that it lies outside.

For a given trace contour, the set C is defined to be the set of all points introduced by the second order concavity operators.

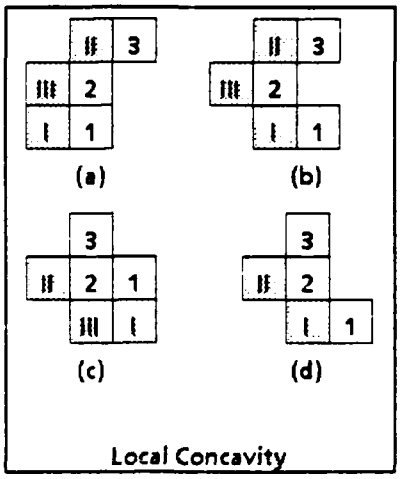


Figure 4. Sources of local concavity, three of which require special second order operators to introduce additional pixels. To ensure a continuous inner annulus, behaviors a,b, and c require that a third element (III) be attached to the inner annulus after generation of elements I and II.

2.7 Intersections with the Original Contour: the Set T.*

The possibility exists that extraneous elements belonging to the original closed contour may be introduced by the first order operators of the inner annulus generator function. Just as concavities in a closed contour mandate that second order operators be invoked to introduce points forsaken by the first order operators, local convexities require second order treatment to remove misbegotten points (Figure 5c). It will be shown that all such points are produced by the *convex corner triplets* of the contour, but first it is necessary to formally define what is meant by a *convex corner triplet*.

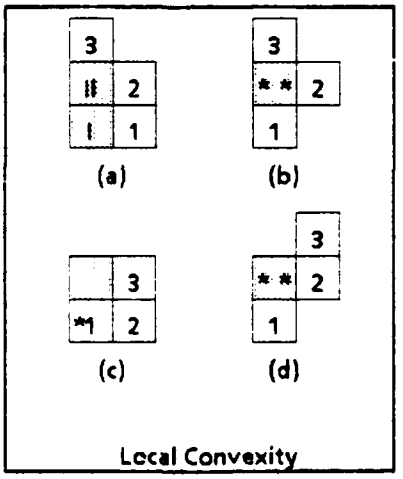


Figure 5. Sources of local convexity, three of which require special second order operators to remove misbegotten pixels. Three of the sources generate problematic annulus elements: behaviors b and d generate the same element (**) twice, whereas behavior c generates an element (1) on the original contour. Note that the figures are

respective mirror images through the vertical axis of those shown for local concavity at figure 3.

Definition 2.7.1. Let (x_{i-1},y_{i-1}), (x_i,y_i) and (x_{i+1}, y_{i+1}) be three consecutive elements of a counterclockwise-ordered contour T_j. Let $\Delta x = x_i - x_{i-1}$ and $\Delta y = y_i - y_{i-1}$, $\Delta xnew = x_{i+1} - x_i$ and $\Delta ynew = y_{i+1} - y_i$. If one of the following conditions is true, then $\{(x_{i-1},y_{i-1}), (x_i,y_i), (x_{i+1}, y_{i+1})\}$ is said to be a **convex corner triplet** of contour T_j.

Δx	Δy	$\Delta xnew$	$\Delta ynew$
>0	$=0$	$=0$	>0
$=0$	>0	<0	$=0$
<0	$=0$	$=0$	<0
$=0$	<0	>0	$=0$

The set T^* is the set of all convex corner triplets contained in a closed contour.

Theorem 2.7. An inner annulus generator function I_g produces a point on the original contour if and only if the point is the first element of a convex corner triplet.

The proof consists of two parts:

i.) If a point generated by an inner annulus generator function I_g is on the original contour, then the point must be the first element of a convex corner triplet.

Proof (by enumeration): from a given contour point (x_s,y_s), there are 8 possible directions in which to proceed, 4 diagonal (to the D-connected pixels) and 4 non-diagonal (to the 4-connected pixels). Each of these moves in turn may proceed to one of five points (without violating the condition that a contour contains no loops). Thus, there are 4*5 + 4*5 = 40 possible triplets emanating from the start point. The search space may be reduced by exploiting the fact that orthogonal rotations preserve triplet shape. Since the 4 diagonal moves are rotational variants of each other, as are the 4 non-diagonal moves, without loss of generality a move to the NE is selected as the first diagonal move, and a move to the N as the first non-diagonal move, resulting in a reduction of the search space from 40 to 10. From the start point (x_s,y_s), the diagonal (NE) move generates the point $(x_s + 1,y_s + 1)$ and the non-diagonal (N) move generates $(x_s,y_s + 1)$. From each of these points in turn there are 5 possible moves, as enumerated in the following table:

	D-connected (NE move)		4-connected (N move)	
	Original	*I_g-Generated*	*Original*	*I_g-Generated*
I.	(x_s,y_s)	-	(x_s,y_s)	-
II.	$(x_s + 1,y_s + 1)$	$(x_s,y_s + 1)$	$(x_s,y_s + 1)$	(x_s-1,y_s)
III. a)	$(x_s + 2,y_s + 2)$	$(x_s + 1,y_s + 2)$	$(x_s,y_s + 2)$	$(x_s-1,y_s + 1)$

b)	$(x_S,y_S + 2)$	$(x_S,y_S + 1)$	$(x_S-1,y_S + 2)$	$(x_S-1,y_S + 1)$
c)	$(x_S + 1,y_S + 2)$	$(x_S,y_S + 1)$	$(x_S-1,y_S + 1)$	(x_S,y_S)
d)	$(x_S + 2,y_S + 1)$	$(x_S + 1,y_S + 2)$	$(x_S + 1,y_S + 2)$	$\underline{(x_S,y_S + 2)}$
e)	$(x_S + 2,y_S)$	$(x_S + 2,y_S + 1)$	$(x_S + 1,y_S + 1)$	$(x_S,y_S + 2)$

Backtracking reveals that the underscored point is the only point contained in the original contour. Furthermore, it is the *first element* of the convex corner triplet $\{(x_S,y_S),(x_S,y_S + 1),(x_S-1,y_S + 1)\}$. In similar fashion, it can be shown that the 4-connected (non-diagonal) moves E, S and W produce generator elements which are the first points of convex corner triplets lying on the original contour. This completes part i) of the proof.

ii.) If a point is the first element of a convex corner triplet, then it is assigned by the inner annulus generator function Ig to the original contour.

Proof: Without loss of generality, let the convex corner triplet consist of the points (x_S,y_S), $(x_S,y_S + 1)$, and $(x_S-1,y_S + 1)$, which is a north move followed by a west move. Then, by definition of the inner annulus generator function, the first pair of points generates the interior point (x_S-1,y_S), whereas the second pair generates (x_S,y_S), which is the first point of the convex corner triplet we started with. In similar fashion, it follows that the other three orientations of a convex corner triplet yield intersections with the original contour. This completes the proof of Theorem 2.3.

2.8 Points Multiply Visited by the Inner Annulus Generator: the Duplicate Set $D^{(k)}$.

It is possible for a point to be produced more than once during the generation of the inner annulus. Examples are shown in the local convexity illustration, Figure 5(b) and (d). Let $D^{(k)}$ denote the set of points visited at least k times during the generation of the inner annulus. When computing the length of the inner annulus, it is necessary to subtract the number of points which are visited at least twice, at least three times, and at least four times, since each such point need be counted only once in the length calculation.

The set $D^{(k)}$ is the set of points produced multiple times after generating the inner annulus.

3.0 The Bottom Line: How Much Memory Does the Inner Annulus Consume?

The first order generator function Ig assigns points on a one-to-one basis from the original contour to the inner annulus. There are n such mappings. The second order concavity operators add points missed by the first order operators; there are |C| such points. The second order convexity operators remove points which are elements of the original contour; there are |T*| of these. Finally, there are points which are generated multiple times; there are $|D^{(2)}| + |D^{(3)}| + |D^{(4)}|$ of these.

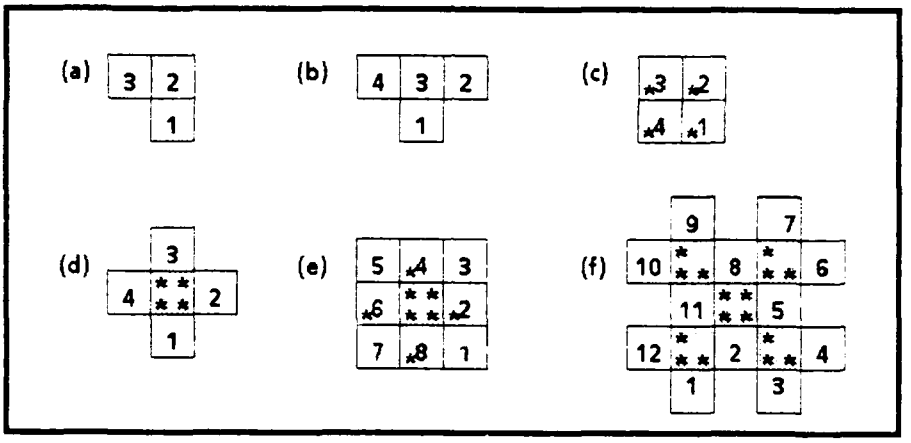
Conjecture 3.0. The general expression for the length of the inner annulus Ij of a closed contour Tj is:

$$|Ij| = n + |C| - |T^*| - \sum_{k=2}^{4} |D^{(k)}|,$$

where
- n = the length of contour Tj
- C = the set of local concavities contained in Tj
- T* = the set of convex corner triplets contained in Tj
- D(k) = the set of points produced at least k times by generator Ig.

4.0 Examples of Automated Inner Annulus Generation.

This section provides several examples of the inner annulus generator function operating on contours especially chosen to exhibit peculiar behavior. Also, Conjecture 3.0 is validated for the examples.



Contour?	n	+	|C|	- |T*|	- |D(2)|	- |D(3)|	- |D(4)|	=	|I|
a) No									
b) No									
c) Yes	4		0	4	0	0	0		0
d) Yes	4		0	0	1	1	1		1
e) Yes	8		0	4	1	1	1		1
f) Yes	12		4	0	5	5	1		5

Figure 6. Diagrams (a) and (b) are not contours because each violates a condition of the definition (the first has length less than 4, and the second contains a loop from point 1 to point 3). Figures (c) and (d) are the shortest contours possible. The inner annulus of (c) is the null set; whereas that of (d) is a singleton set. Figure (e) is interesting because the inner annulus generator produces four elements of the original contour, while it also produces the same interior element four times. Figure (f) is presented to demonstrate multiple local concave behavior.

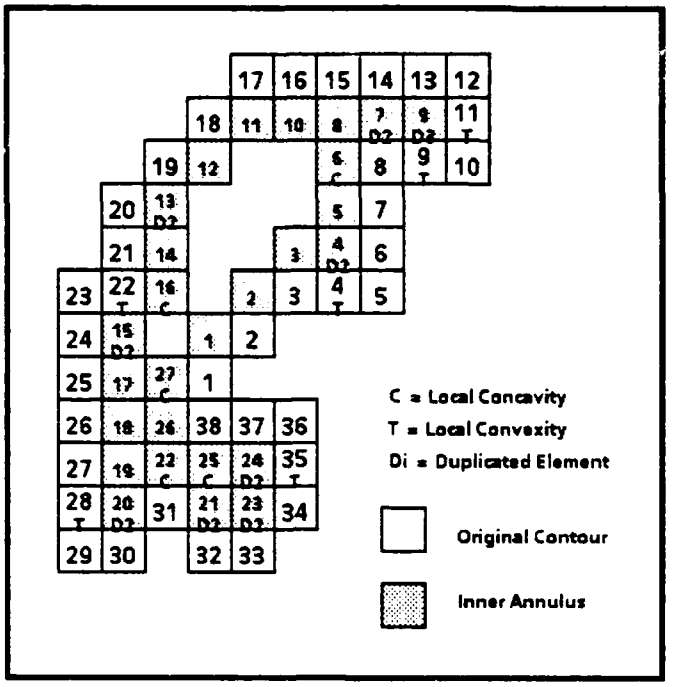


Figure 7. The generation of the inner annulus for a sample contour of length 38. The annulus elements are numbered in the order they are produced by the generator function. Pixels marked with a T are elements of the original contour produced by convex corner triplets; there are 6 such cases. Pixels marked with a C are produced by second-order local concavity operators; there are 5 such instances. Those marked with a D are multiply assigned by the first order generator function; there are 9 assigned at least twice, and one assigned at least three times. By Conjecture 3.0, the length of the annulus is predicted to be 38 + 5 - 6 - 9 - 1 = 27, which is the result obtained in practice.

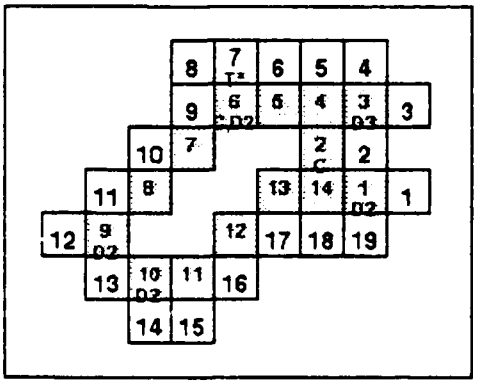


Figure 8. The generation of the inner annulus of a closed contour of length n = 19. In this example, C = {2,6}, T^* = {7}, $D^{(2)}$ = {1,3,6,9,10}, $D^{(3)}$ = {3}, and $D^{(4)}$ = { }. Conjecture 3.0 predicts that the length of the inner annulus is $n + |C| - |T^*| - |D^{(2)}| - |D^{(3)}| - |D^{(4)}| = 19 + 2 - 1 - 5 - 1 - 0 = 14$.

5.0 Complexity Considerations.

Recall that the inner annulus is the set of pixels 4-connected to the inner edge of a closed contour, and that an inclusion decision is made by comparing the distance to the contour with the distance to the annulus. For implementation purposes, it is important to note that the nearest contour element is actually *4-connected* to the nearest annulus element, which produces an efficient algorithm. If the original contour is of length n pixels, it can be sorted in n* log n preprocessing time, and uses space of order n. Once the sort is done, a run time distance calculation to the annulus can be performed in time log n + c, where log n is the time required for binary search, and c is the constant time required to compute the 4-connected annulus element.

When comparing the annulus technique to competitive techniques, recall that other techniques do not succeed for multiply-connected sets, nor do they return the distance and direction to the contour. This latter information is important for many real-world applications; for example, temporal reasoning with two-dimensional maps. Therefore, although other contour inclusion techniques may seem to be competitive in time complexity, the annulus technique is actually rendering a richer source of information in the same amount of time.

6.0 Conclusions.

A function which tests for *metrical inclusion* of an arbitrary point within a multiply-connected closed contour has been introduced. Metrical inclusion means that the distance and direction to the nearest point of the contour is rendered along with the inclusion decision. The technique exploits a new data structure called the *inner annulus*, which is automatically computed with a corresponding generator function, using as input the digital representation of a closed Jordan curve. It is shown that the annulus technique provides a richer source of information than the technique for simply-connected sets based on an even or odd count of contour crossings, since the latter technique does not return either distance or direction along with the inclusion decision. A conjecture is posed concerning the length of the inner annulus in terms of the length and shape of the original contour. An attempt to prove the conjecture has yielded several results pertaining to the convex and concave behavior of the parent contour. The complexity of the optimized inclusion algorithm is shown to be of order log n execution time, requires order n space, and consumes n * log n preprocessing time.

Acknowledgements

The research benefited from technical discussions with Gerald Andersen, Richard Antony, Christopher Bogart, Thomas Garvey, Henry Kyburg, James Mulligan, and Azriel Rosenfeld.

Bibliography

[B1] Barrow, H.G. et. al. "Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching", Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge MA, 1977.

[C1] Cronin, Terence M., "Loci-reduced Spatial Discrimination", US Army Center for Signals Warfare Technical Report, 1988.

[F1] Fischler, Martin A. "Interactive Aids for Cartography and Photo Interpretation", SRI International Technical Report, October 1979.

[F2] Foley, James. D., and Andries Van Dam. Fundamentals of Interactive Computer Graphics, Addison-Wesley, Reading MA, 1983.

[G1] Garvey, Thomas D. "Evidential Reasoning for Geographic Evaluation for Helicopter Route Planning", IEEE Transactions on Geoscience and Remote Sensing, May 1987.

[G2] Garvey, Thomas D. "Perceptual Strategies for Purposive Vision", SRI International Technical Note 117, September 1976.

[G3] Garvey, Thomas D., and Martin A. Fischler. "Machine-Intelligence-Based Multisensor ESM System", SRI International Technical Report AFAL-TR-79-1162, March 1979.

[G4] Gonzalez, Rafael C., and Paul Wintz. Digital Image Processing, Second Edition, Addison-Wesley, Reading MA, 1987.

[L1] Leyton, Michael. "Constraint-Theorems On the Prototypification of Shape", Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia PA, 1986.

[M1] Maragos, Petros A., and Ronald W. Schafer. "Morphological Skeleton Representation and Coding of Binary Images", IEEE Transactions on Acoustics, Speech, and Signal Processing, VOL. ASSP-34, No. 5, October 1986.

[M2] Mulder, Jan A. "Using Discrimination Graphs to Represent Visual Interpretations That are Hypothetical and Ambiguous", Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA, 1985.

[P1] Preparata, Franco P., and Michael Ian Shamos. Computational Geometry, Springer-Verlag, New York NY, 1985.

[R1] Rey, William J.J. Introduction to Robust and Quasi-Robust Statistical Methods, Springer-Verlag, Berlin GDR, 1983.

[R2] Rosenfeld, Azriel, and A.C. Kak. Digital Picture Processing, Vol. II, Second Edition. Academic Press, NY, 1982.

[S1] Saund, Eric. "Abstraction and Representation of Continuous Variables in Connectionist Networks", Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia PA, 1986.

[S2] Sedgewick, Robert. Algorithms, Addison-Wesley, Reading MA, 1983.

[W1] Weiss, Isaac. "3-D Shape Representation by Contours", Proceedings of the Ninth International Joint Conference on Artificial Intelligence, Los Angeles, CA, 1985.

On the Positive Roots of an Equation Involving a Bessel Function

Seigfried H. Lehnigk

Research Directorate
Research, Development, and Engineering Center
U.S. Army Missile Command
Redstone Arsenal, AL 35898-5248

Abstract

It is being shown that, depending on the parameters $A \in R$, $B > 0$, $q < -1$, the equation $(-Br^2+A)I_q(r) + r\, I_q(r) = 0$ has either one positive simple root, or two positive simple roots, or one positive double root, or no positive roots at all. The equation is related to the parabolic generalized Feller equation and is of statistical significance.

1980 Mathematics Subject Classification (1985 Revision).

Primary 33A40. Secondary 33K15, 62F10.

The full paper will appear in Vol. 32 (1989), Proc. Edin. Math. Soc.

UNIFORM ERROR BOUND MESHES IN PIECEWISE LINEAR INTERPOLATION

Royce W. Soanes
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

ABSTRACT. The classical bound on the error in linear interpolation of function f on interval (a,b) is

$$\frac{1}{8}\,(b-a)^2\|f''\|_{(a,b)}$$

where

$$\|f''\|_{(a,b)} = \max_{a\leq x\leq b} |f''(x)|$$

We will show how to obtain a mesh x for which

$$(x_{i+1}-x_i)^2\|f''\|_{(x_i,x_{i+1})} = \text{constant (very nearly)} \quad 1\leq i<n$$

The solution of this problem is important for the purpose of providing accurate functional data in tabular form for use in numerically controlled manufacturing machines (CAM).

GOOD MESHES. deBoor [1] has supplied us with a simple method for generating good meshes. His idea is to make the classical error bound roughly constant:

$$\frac{1}{8}\,(x_{i+1}-x_i)^2\|f''\|_{(x_i,x_{i+1})} = \text{constant} \quad 1\leq i<n$$

or

$$(x_{i+1}-x_i)\|f''\|^{\frac{1}{2}}_{(x_i,x_{i+1})} = \text{constant}$$

or

$$\int_{x_i}^{x_{i+1}} \|f''\|^{\frac{1}{2}}_{(x_i,x_{i+1})}dx = c$$

As n becomes large while $|f''(x)| > 0$, neighboring x's will get close together. This partially justifies the following asymptotic approximation of the norm of f":

$$\|f''\|^{\frac{1}{2}}_{(x_i,x_{i+1})} \sim |f''(x)|^{\frac{1}{2}} = g(x)$$

So we solve the simpler problem

$$\int_{x_i}^{x_{i+1}} g(x)dx = c$$

If we define $G(x) = \int_{x_1}^{x} g(t)dt$, we have

$$G(x_i) = \int_{x_1}^{x_i} g(t)dt = (i-1)c$$

and

$$G(x_n) = (n-1)c$$

Therefore

$$\frac{G(x_i)}{G(x_n)} = \frac{i-1}{n-1}$$

and we finally have deBoor's method for generating good meshes for piecewise linear interpolation

$$x_i = G^{-1}(\frac{i-1}{n-1}\ G(x_n))$$

where

$$G(x) = \int_{x_1}^{x} g(t)dt$$

and

$$g(t) = |\ f''(t)|^{\frac{1}{2}}$$

In practice, we typically have only a positive, continuous, piecewise linear estimate of g over some mesh u. We will denote this estimate of g by v. G as defined by

$$G(x) = \int_{u_1}^{x} v(t)dt \quad u_1 \leq x \leq u_m$$

would then be piecewise quadratic and invertible in the following manner:

$$G^{-1}(G^*) = x^* = u_i + \frac{2(G^*-G_i)}{v_i + \sqrt{D}}$$

where

$$G_1 = 0 \quad , \quad G_{j+1} = G_j + (u_{j+1}-u_j)(v_j+v_{j+1})/2 \qquad 1 \leq j < m$$

$$G_i \leq G^* \leq G_{i+1} ,$$

$$\rho = (G^*-G_i)/(G_{i+1}-G_i) ,$$

and

$$D = (1-\rho)v_i^2 + \rho v_{i+1}^2$$

Unfortunately, good meshes are not always quite as good as we might like them to be. Specifically, the lengths of the longer subintervals are always overestimated because the asymptotic approximation to the norm of f" is least valid for these longer subintervals. This, of course, leads to larger error bounds on the longer subintervals. The error bounds on the shorter subintervals are always pleasingly uniform because the asymptotic approximation is most valid for these shorter subintervals. In addition, it is easy to prove that for $f(x) = x^p$ ($p > 2$, $0 \leq x \leq 1$), the largest error bound on a good mesh is exactly equal to the largest error bound on a uniform mesh ($x_{i+1}-x_i = \text{const}$)!

BETTER MESHES. The following problem is correctly described in [1] as being quite difficult to solve in general:

Find n-2 x's (x_1 and x_n fixed) such that

$$(x_{i+1}-x_i)\|f''\|^{\frac{1}{2}}_{(x_i,x_{i+1})} = (x_{i+1}-x_i)\|g\|_{(x_i,x_{i+1})} = c \qquad \text{for } 1 \leq i < n$$

Even if we knew what c was, solving

$$(x_{i+1}-x_i)\|g\|_{(x_i,x_{i+1})} = c$$

for x_{i+1} given x_i would still be quite difficult in general.

The interesting thing is that this problem is not difficult to solve if we substitute v for g! In addition, if v is a very good approximation to g (with m >> n), we will get a virtually constant error bound for the entire mesh. At any rate, irrespective of the accuracy of v, we will be doing the best we can under the circumstances to solve the original problem.

MESH FUNCTION $\mu(c)$. For the correct value of c and for the correct mesh x,

$$(x_{i+1}-x_i)\|v\|_{(x_i,x_{i+1})} = c$$

or

$$\int_{x_i}^{x_{i+1}} \|v\|_{(x_i,x_{i+1})}dx = c$$

Defining the piecewise constant function γ by

$$\gamma(x) = \|v\|_{(x_i,x_{i+1})} \quad (x_i \leq x \leq x_{i+1})$$

we obviously have

$$\gamma(x) \geq v(x) \text{ for all } x$$

Now, since

$$\int_{x_i}^{x_{i+1}} \gamma(x)dx = c$$

we have

$$\int_{x_1}^{x_n} \gamma(x)dx = (n-1)c$$

but $\gamma(x) \geq v(x) \geq 0$, so

$$\int_{x_1}^{x_n} \gamma(x)dx \geq \int_{u_1}^{u_m} v(x)dx$$

and

$$(n-1)c \geq \int_{u_1}^{u_m} v(x)dx$$

Therefore, a lower bound on the correct value of c is given by

$$\frac{1}{n-1} \int_{u_1}^{u_m} v(x)dx$$

For an incorrect value of $c(\bar{c})$, we define $\bar{\gamma}$ and $\bar{x}$:

$$\bar{\gamma}(x) = \|v\|_{(\bar{x}_i, \bar{x}_{i+1})} \quad (\bar{x}_i \leq x \leq \bar{x}_{i+1})$$

where

$$\int_{\bar{x}_i}^{\bar{x}_{i+1}} \bar{\gamma}(x)dx = \bar{c}$$

Therefore

$$\int_{x_1}^{\bar{x}_{\nu+1}} \bar{\gamma}(x)dx = (\nu-1)\bar{c} + \int_{\bar{x}_\nu}^{\bar{x}_{\nu+1}} \bar{\gamma}(x)dx$$

where ν is the number of subintervals over which $\bar{\gamma}$ is defined, $\bar{x}_{\nu+1} = u_m$, and

$$\int_{\bar{x}_\nu}^{\bar{x}_{\nu+1}} \bar{\gamma}(x)dx \leq \bar{c}$$

We also have

$$(\nu-1)\bar{c} = \int_{x_1}^{\bar{x}_\nu} \bar{\gamma}(x)dx$$

therefore,

$$\nu = 1 + \frac{1}{c} \int_{x_1}^{\bar{x}_\nu} \bar{\gamma}(x)dx$$

Now since

$$\int_{x_1}^{\bar{x}_\nu} \bar{\gamma}(x)dx$$

is bounded below by

$$\int_{u_1}^{\bar{x}_\nu} v(x)dx$$

and above by

$$(u_m-u_1)\|v\|_{(u_1,u_m)}$$

ν is large for small $\bar{c}$ and small for large $\bar{c}$. Now, for the correct value of c, we want

$$\int_{x_1}^{\bar{x}_{\nu+1}} \bar{\gamma}(x)dx = \int_{x_1}^{x_n} \gamma(x)dx$$

or

$$(\nu-1)c + \int_{\bar{x}_\nu}^{u_m} \bar{\gamma}(x)dx = (n-1)c$$

We therefore define the mesh function μ by

$$\mu(c) = (\nu-n)c + (u_m-\bar{x}_\nu)\|v\|_{(\bar{x}_\nu,u_m)}$$

For small c, ν will be large and μ will be positive. For large c, ν will be small and μ will be negative. For the correct value of c, μ will be zero. But may there be more than one zero of μ? For the correct value of c, we will have

$$(u_m-\bar{x}_\nu)\|v\|_{(\bar{x}_\nu,u_m)} = c$$

and

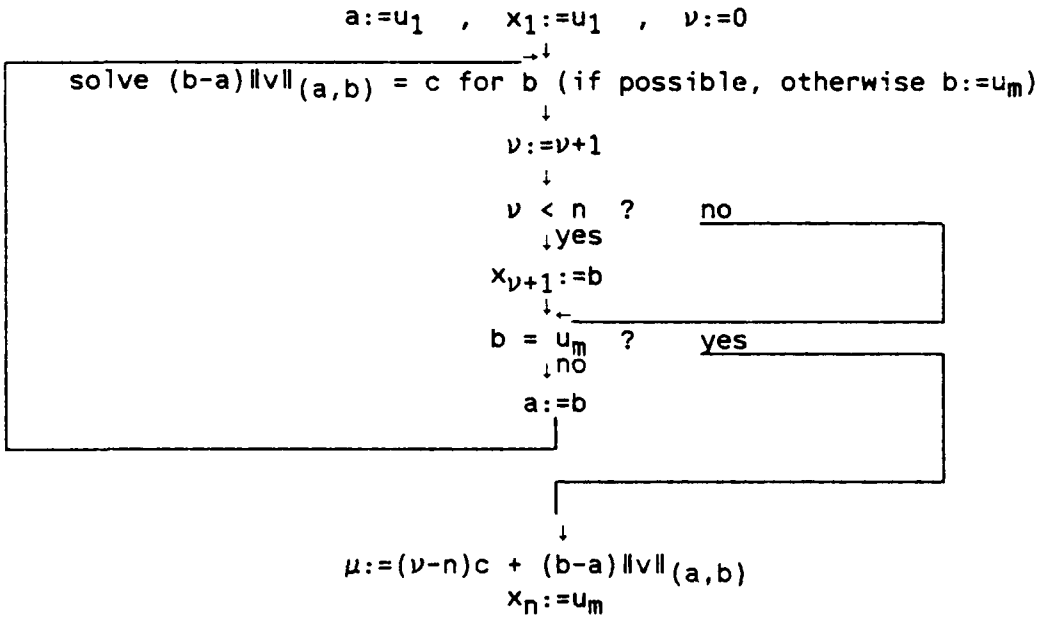
$$\nu = n-1$$

For a slightly smaller value of c, $\bar{x}_\nu$ will be very close to u_m and we will have $\nu = n$. This means that the entire (nearly zero) contribution to μ will be made by

$$(u_m-\bar{x}_\nu)\|v\|_{(\bar{x}_\nu,u_m)}$$

Now suppose that the original function f has a perfectly linear region on the right and v is identically equal to zero over a finite interval. The norm of v will therefore be zero and further slight changes of c will still not change ν and will still maintain a zero norm for v. We therefore conclude that μ can be identically equal to zero for some finite interval to the left of the correct c. This is not really a problem, however, since we only need to be sure to compute the rightmost zero of μ.

We now show the algorithm for computing $\mu(c)$. To compute $\mu(c)$:



When μ has been evaluated for the last time, near the correct value of c, we shall have collected the correct mesh x.

SOLVING $(b-a)\|v\|_{(a,b)} = c$ FOR b. This nonlinear equation turns out to be quite easy to solve noniteratively due to the piecewise linearity of v.

Given a, i, and c such that $u_i \leq a < u_{i+1}$, we want to find b such that $(b-a)\|v\|_{(a,b)} = c$. We set $j = i+1$ initially and increment j as necessary until

$$M_r(u_j-a) > c$$

and

$$M_\ell(u_{j-1}-a) \leq c$$

where

$$M_\ell = v(a) \quad \text{if } j = i+1$$

$$M_\ell = \text{Max}\{v(a), v_{i+1}, \ldots v_{j-1}\} \quad \text{if } j > i+1$$

and

$$M_r = \text{Max}\{M_\ell, v_j\}$$

Therefore

$$M_\ell = \|v\|_{(a,u_{j-1})}$$

$$M_r = \|v\|_{(a,u_j)}$$

and b will lie somewhere between u_{j-1} and u_j.

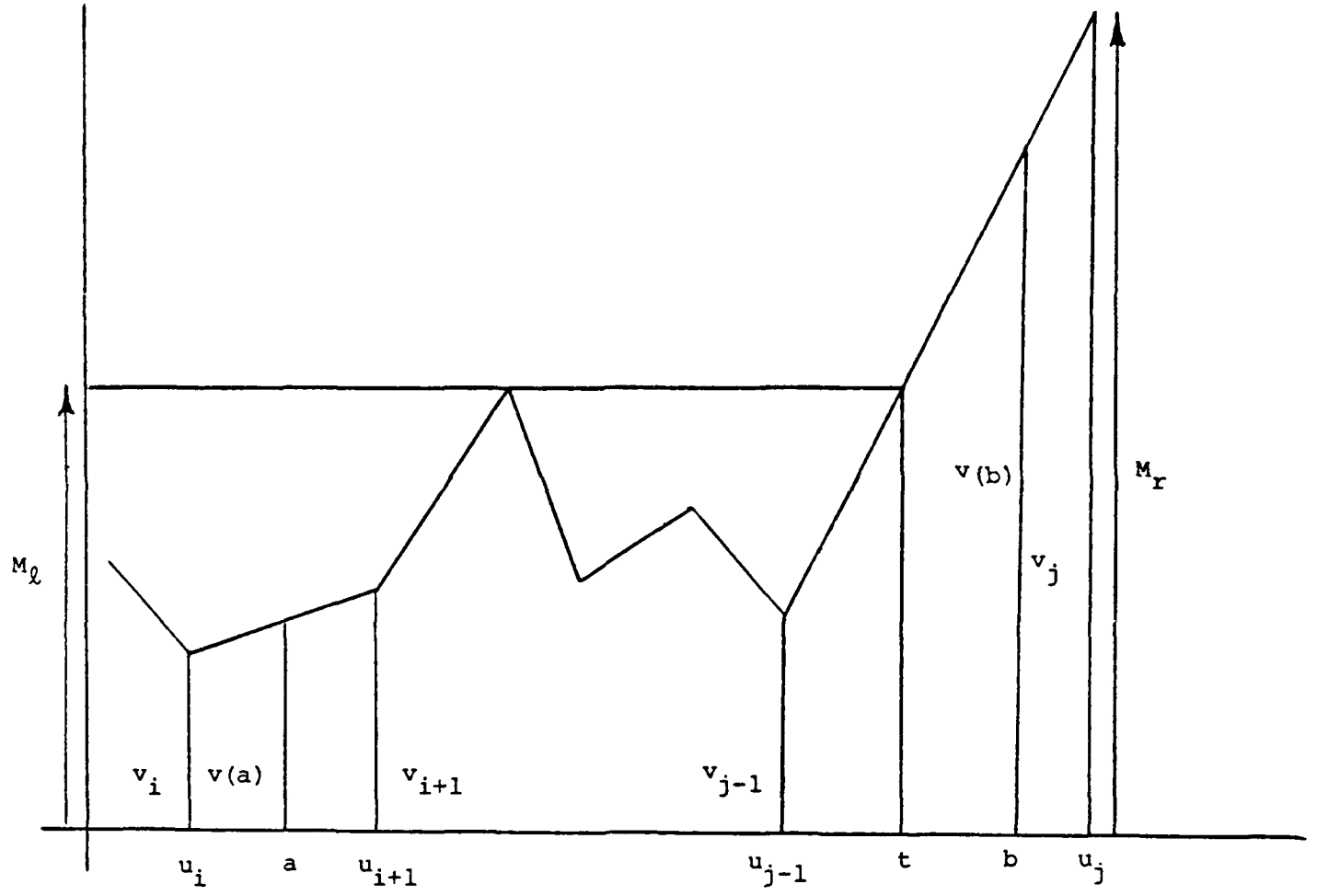


Figure 1. Diagrammatic reference for solving $(b-a)\|v\|_{(a,b)} = c$ for b.

If $M_r = M_\ell$, we want

$$(b-a)M_\ell = c$$

therefore $b = a + c/M_\ell$.

If $M_r > M_\ell$, we first need to compute the transition point t. Let

$$s = (v_j - v_{j-1})/(u_j - u_{j-1})$$

$$= (M_r - M_\ell)/(u_j - t)$$

Therefore,

$$t = u_j - (M_r - M_\ell)/s$$

If $M_\ell(t-a) \geq c$, b must lie to the left of t, so we want

$$(b-a)M_\ell = c$$

Again we have trivially that

$$b = a + c/M_\ell$$

Now if $M_\ell(t-a) < c$, then b must lie between t and u_j. We therefore want

$$(b-a)v(b) = c$$

but

$$\frac{M_r - v(b)}{u_j - b} = s$$

Therefore,

$$v(b) = M_r - s(u_j - b)$$

and

$$(b-a)(M_r - s(u_j - b)) = c$$

$$= (b-a)(M_r - s(u_j - a + a - b))$$

$$= (b-a)(M_r - s(u_j - a) + s(b-a))$$

We therefore have the following quadratic equation for b-a:

$$s(b-a)^2 + (M_r - s(u_j - a))(b-a) - c = 0$$

Letting $k = M_r - s(u_j - a)$, we have

$$b - a = \frac{-k \pm \sqrt{k^2 + 4sc}}{2s}$$

We need the plus sign in order to get $b - a > 0$ irrespective of the sign of k. We therefore have

$$b = a + \frac{\sqrt{k^2+4sc} - k}{2s} \qquad \text{for } k < 0$$

and for $k > 0$, we rationalize to get

$$b = a + \frac{2c}{k + \sqrt{k^2+4sc}}$$

Thus we see that the computational complexity of solving $(b-a)\|v\|_{(a,b)} = c$ for b is nearly trivial indeed, making the "better mesh" algorithm quite efficient.

SOME COMPUTATIONAL RESULTS. All the following examples have been obtained using a uniform preliminary mesh of size m and exact evaluation of f". The first two figures (Figures 2 and 3) show a virtually constant error bound pattern for the two functions x^{100} and $(1-x)^{100}$. These two functions have very flat regions to the left and right, respectively, and their respective meshes are naturally mirror images of each other. The mesh functions for these two cases are quite different, however, as indicated by Figures 4 and 5. The $(1-x)^{100}$ mesh function behaves as it does due to the large flat section to the right. Note the lack of monotonicity further to the left of the correct value of c and the way μ becomes and remains monotonic as c is approached from the left. Also note that the correct values of c are the same for both cases, as expected.

Figures 6 through 9 compare good and better meshes for the test function $x^{10}(1-x)^{20}$. Note the larger error bounds for the longer subintervals in the good mesh. Figures 10 and 11 indicate the effect of reducing m and the accuracy of v.

Figures 12 through 15 compare good and better meshes for a larger value of n. Figure 16 shows the effect of small m. Note that there is hardly any discernible difference between the better meshes for large and small m. Figures 17 and 18 show the equality of the largest bounds on uniform and good meshes, respectively. Figure 19 shows the better mesh bounds for this case.

The code contained in the Appendix, written in SALOME, computes better meshes. Reference [2] may be consulted for interpretation of the code, but it should be sufficient to know that IF and FI delimit conditional statements, DO and OD delimit looping statements, comma in a conditional means "then," semicolon in a conditional means "else," and sharp signs (#) delimit loop exit conditions.

REFERENCES

1. deBoor, C., A Practical Guide to Splines, Springer Verlag, New York, 1978.

2. Soanes, R. W., "Salome, A Structured And Logically Minimal Ensemble of Programming Constructs," Proceedings of the 1982 Army Numerical Analysis and Computers Conference, ARO Report 82-3, Research Triangle Park, NC, August 1982; also Technical Report ARLCB-TR-82021, Benet Weapons Laboratory, Watervliet, NY, July 1982.

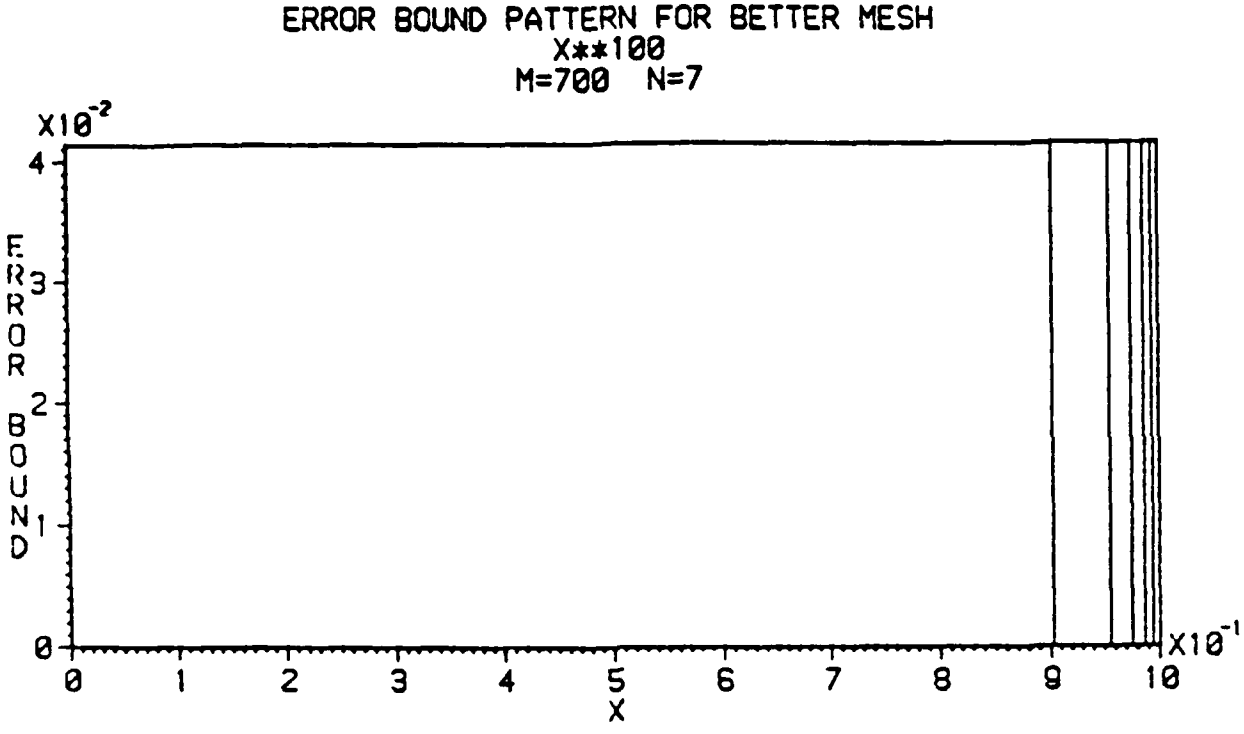


Figure 2

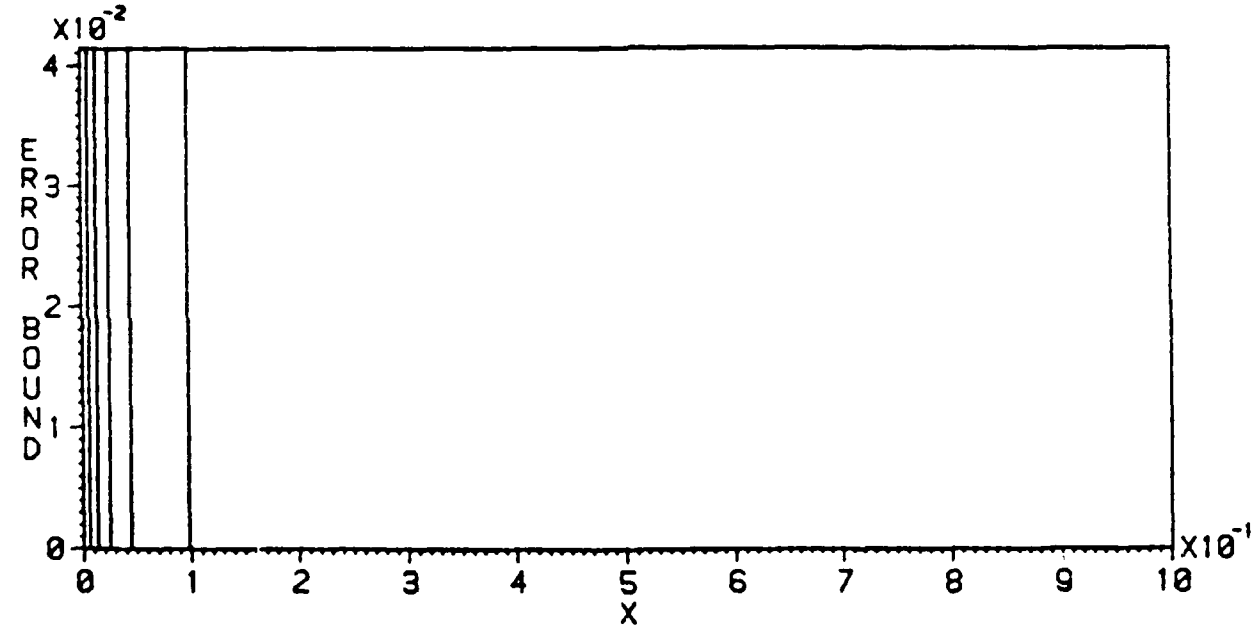


Figure 3

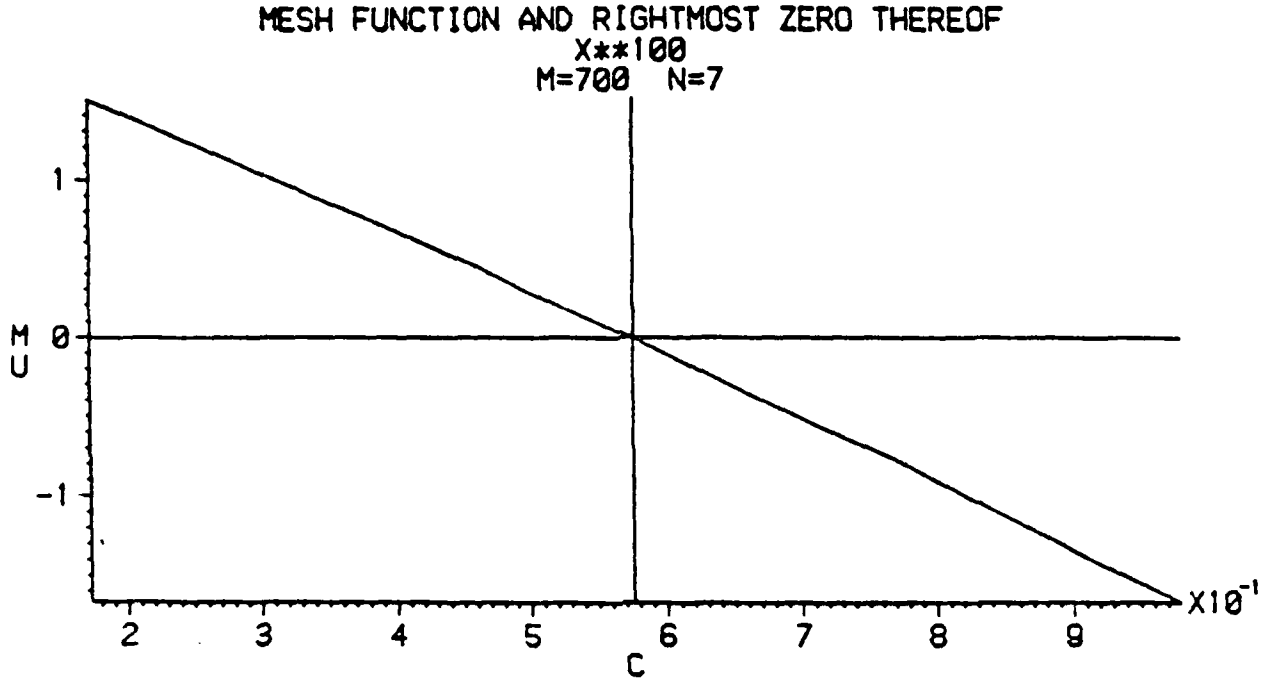


Figure 4

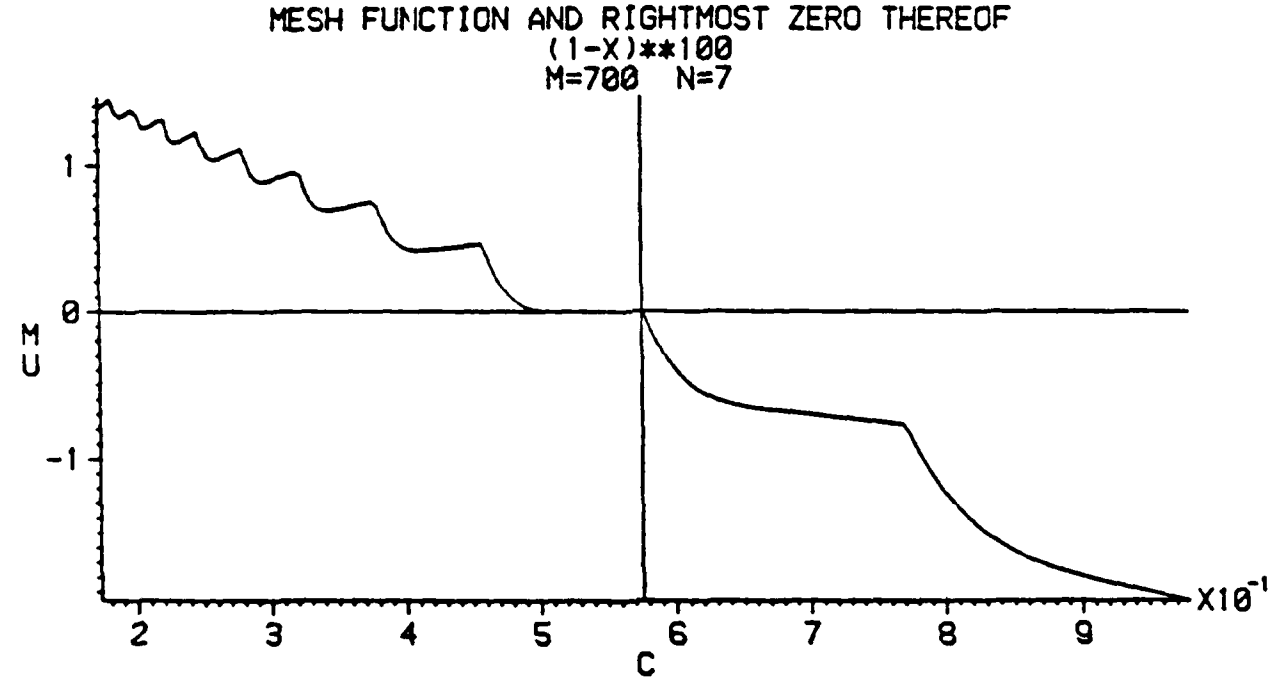


Figure 5

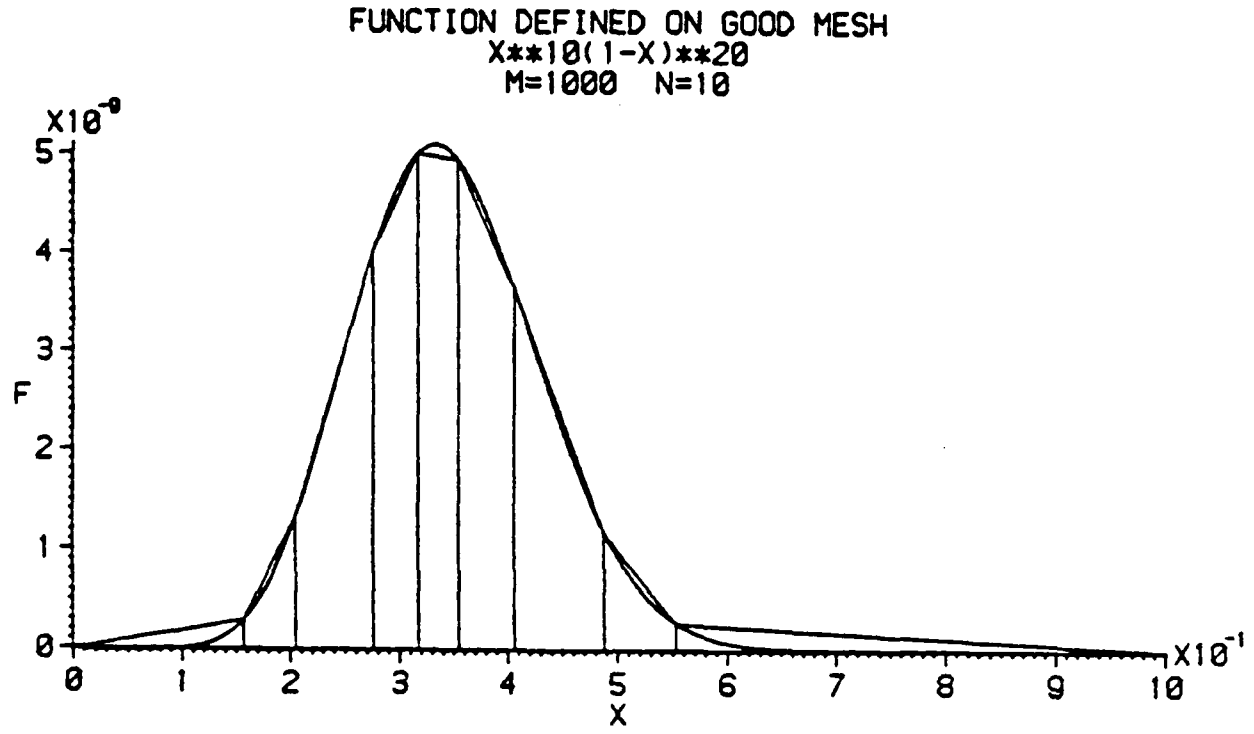


Figure 6

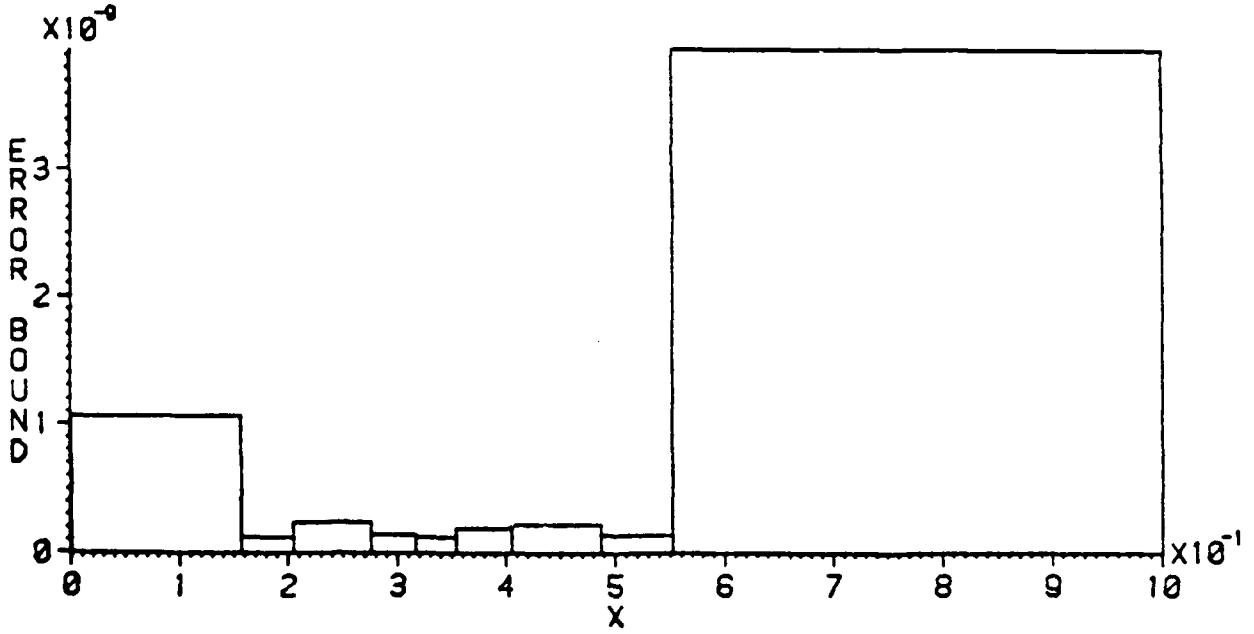


Figure 7

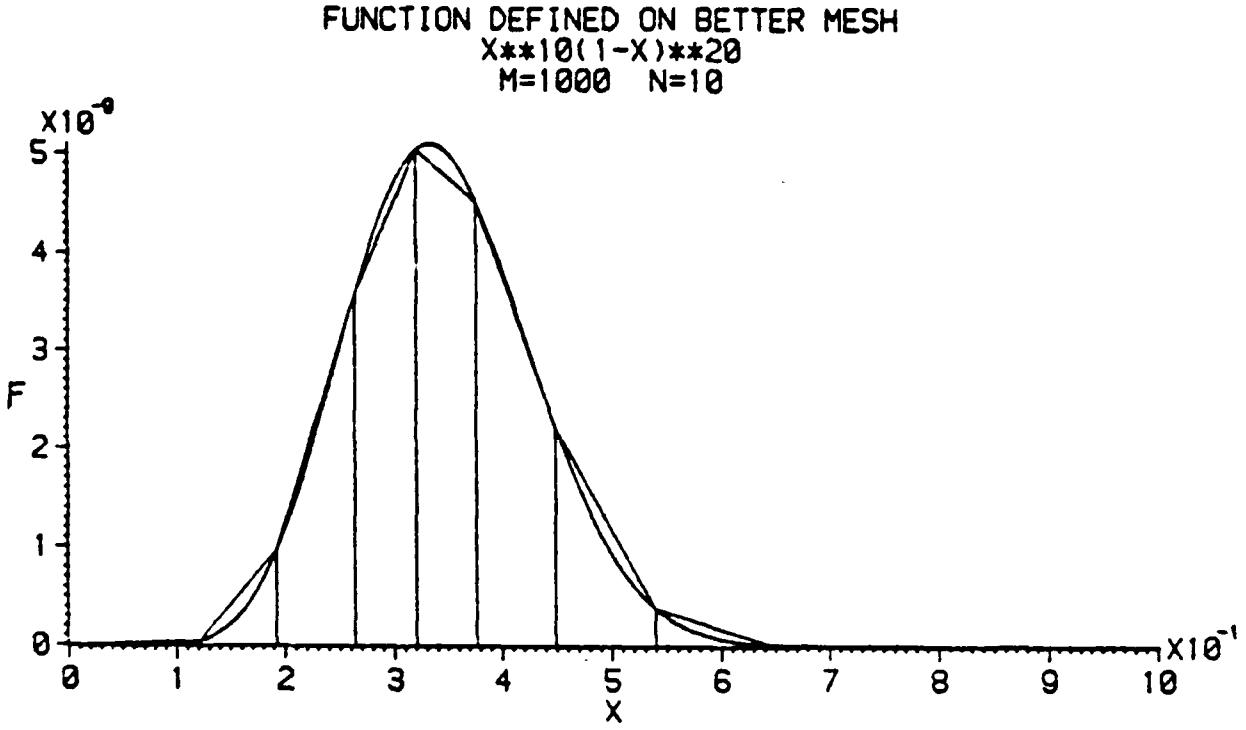


Figure 8

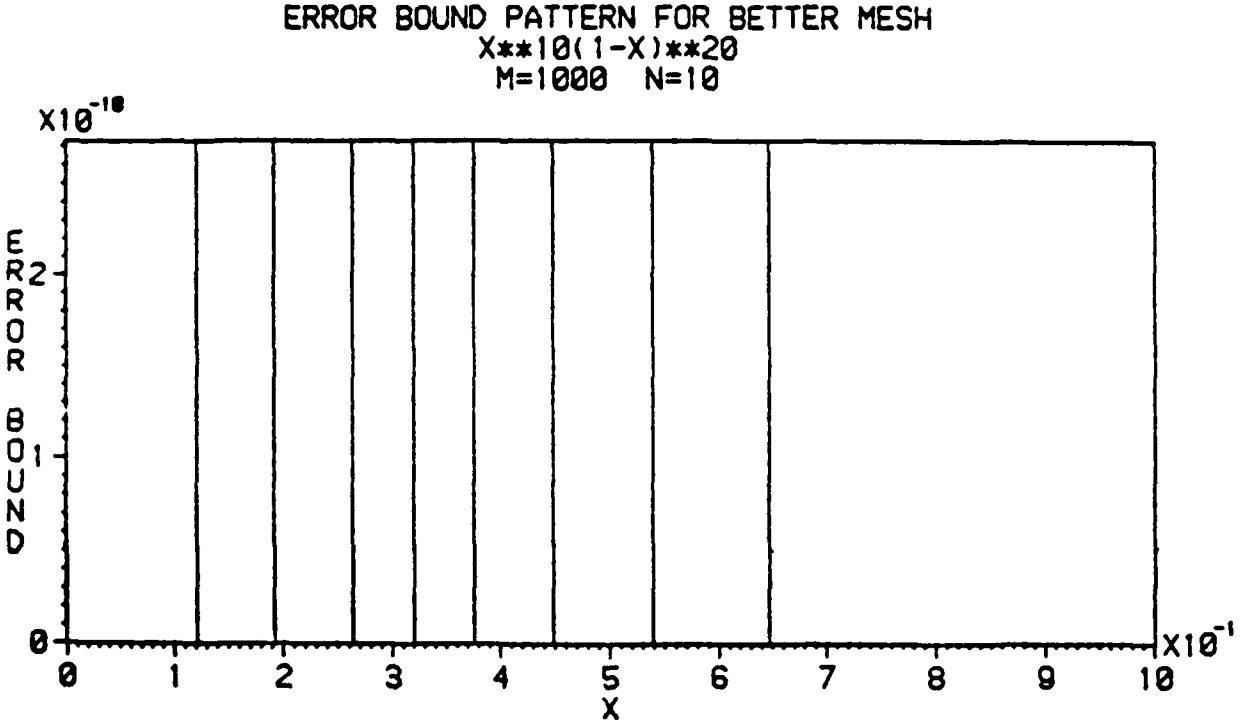


Figure 9

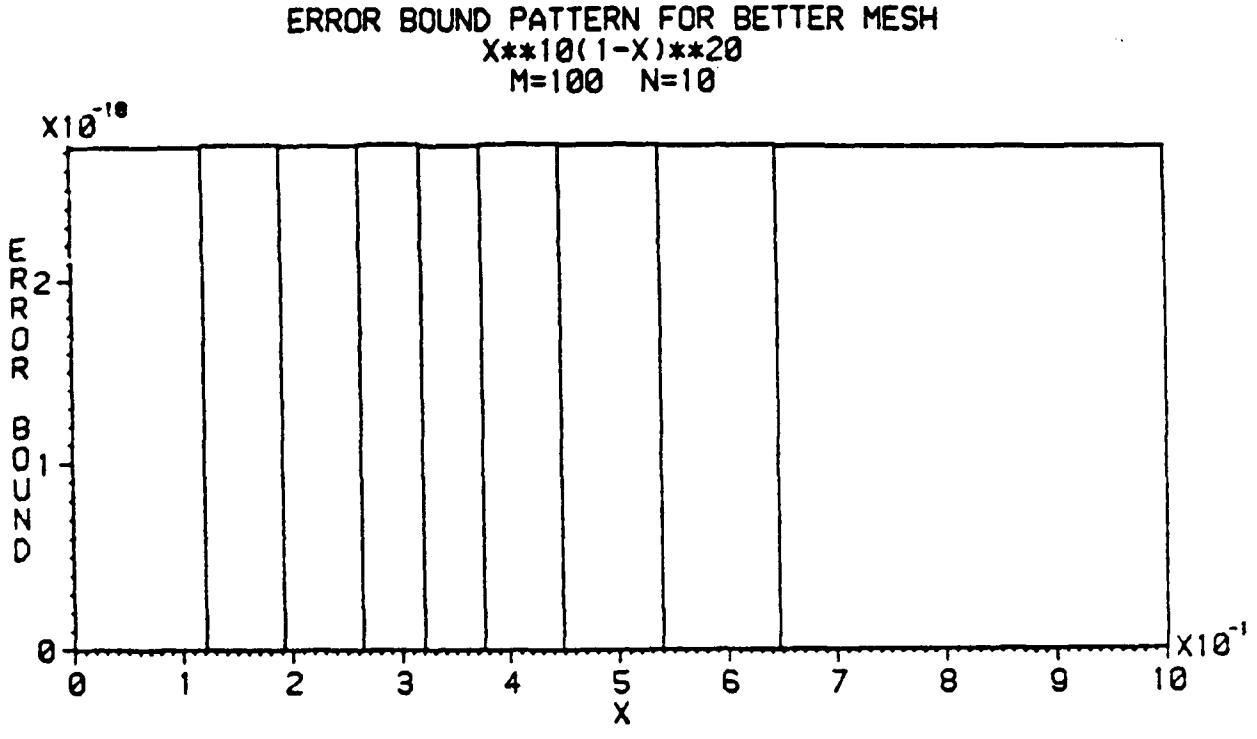


Figure 10

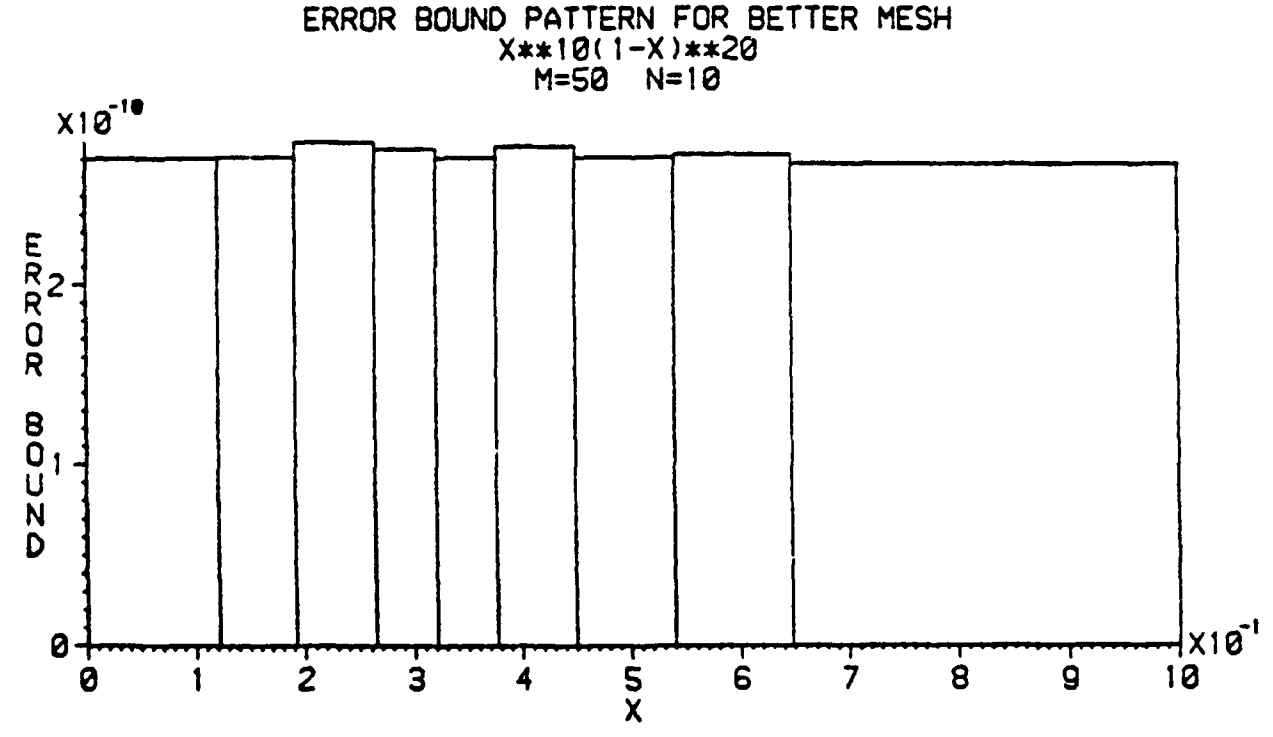


Figure 11

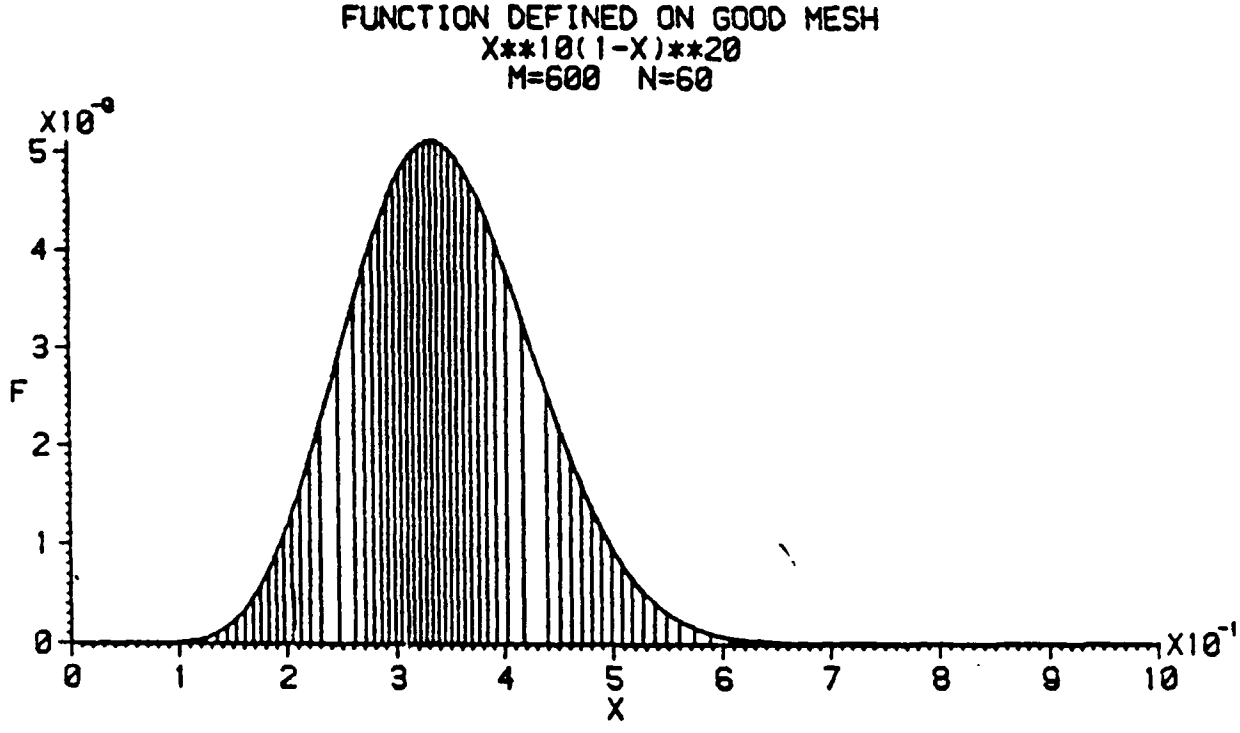


Figure 12

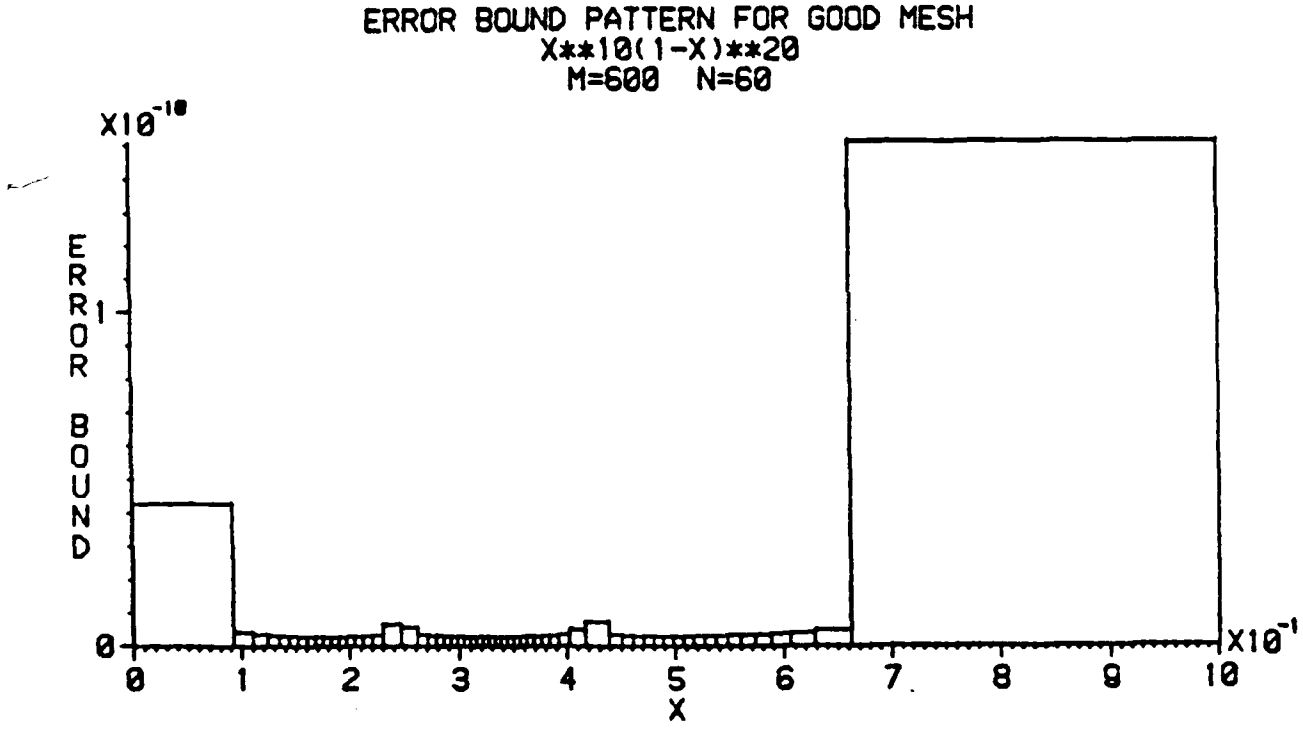


Figure 13

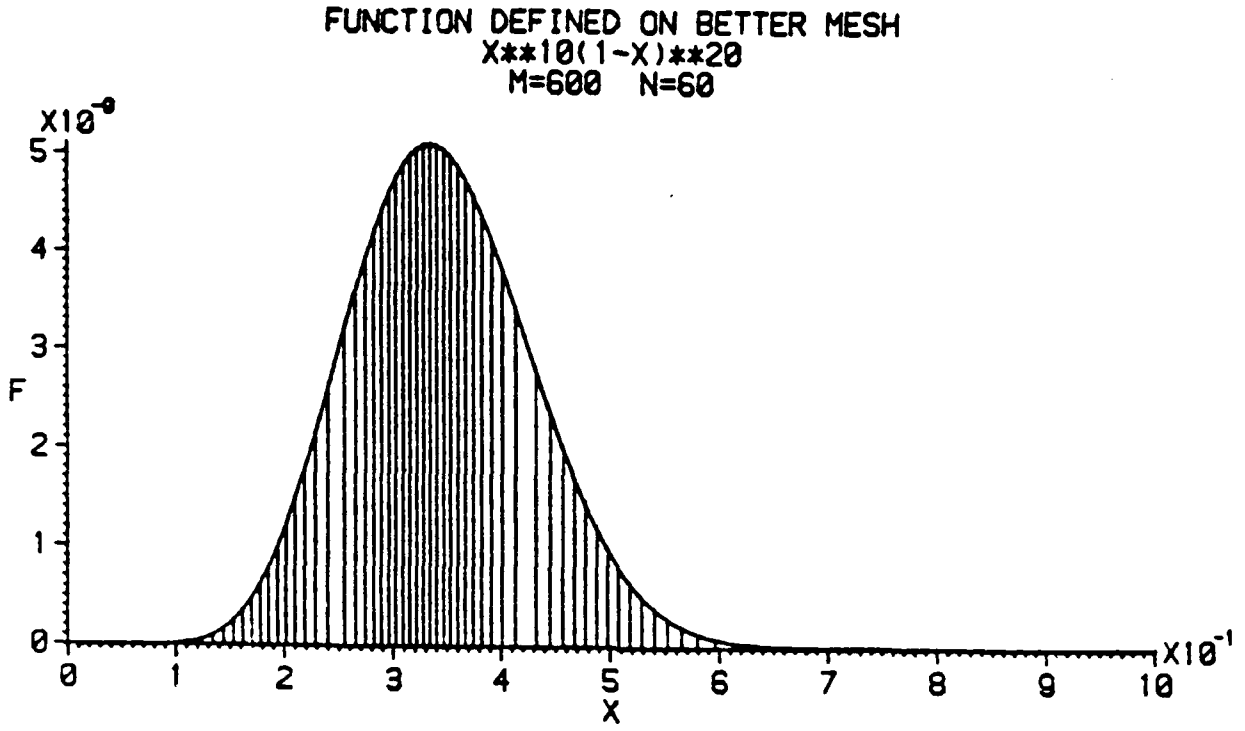


Figure 14

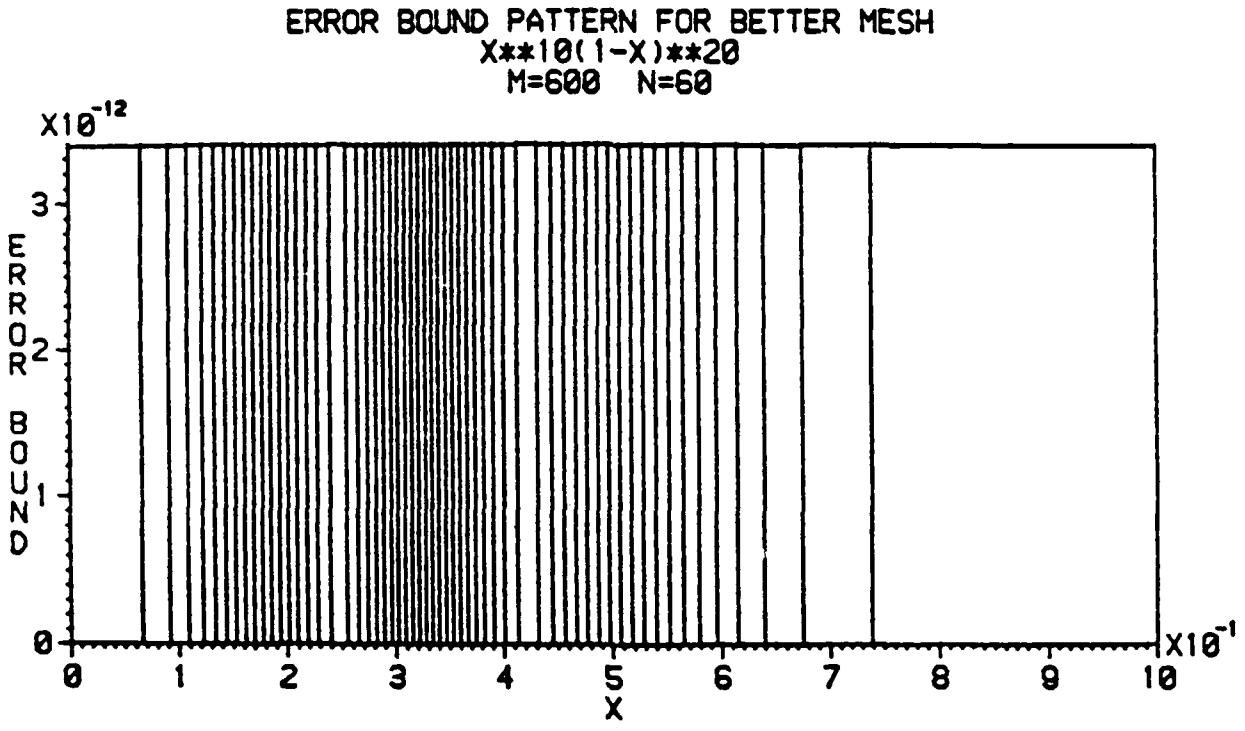


Figure 15

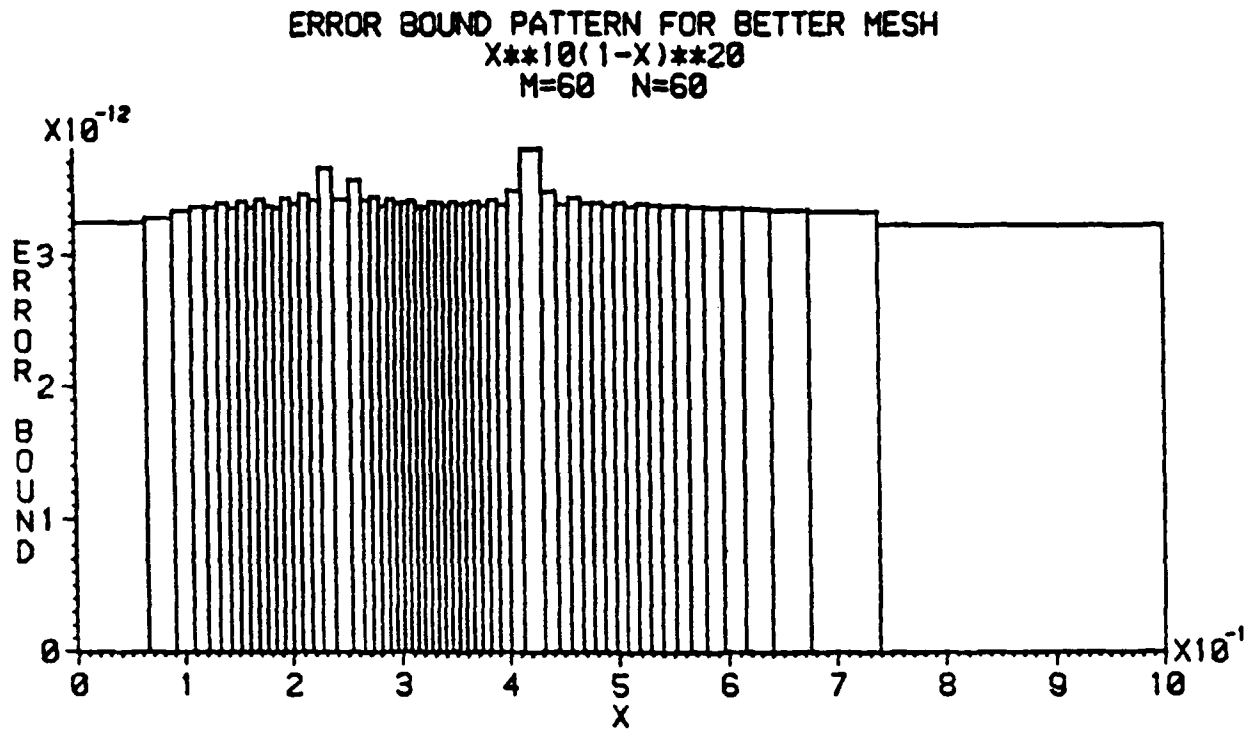


Figure 16

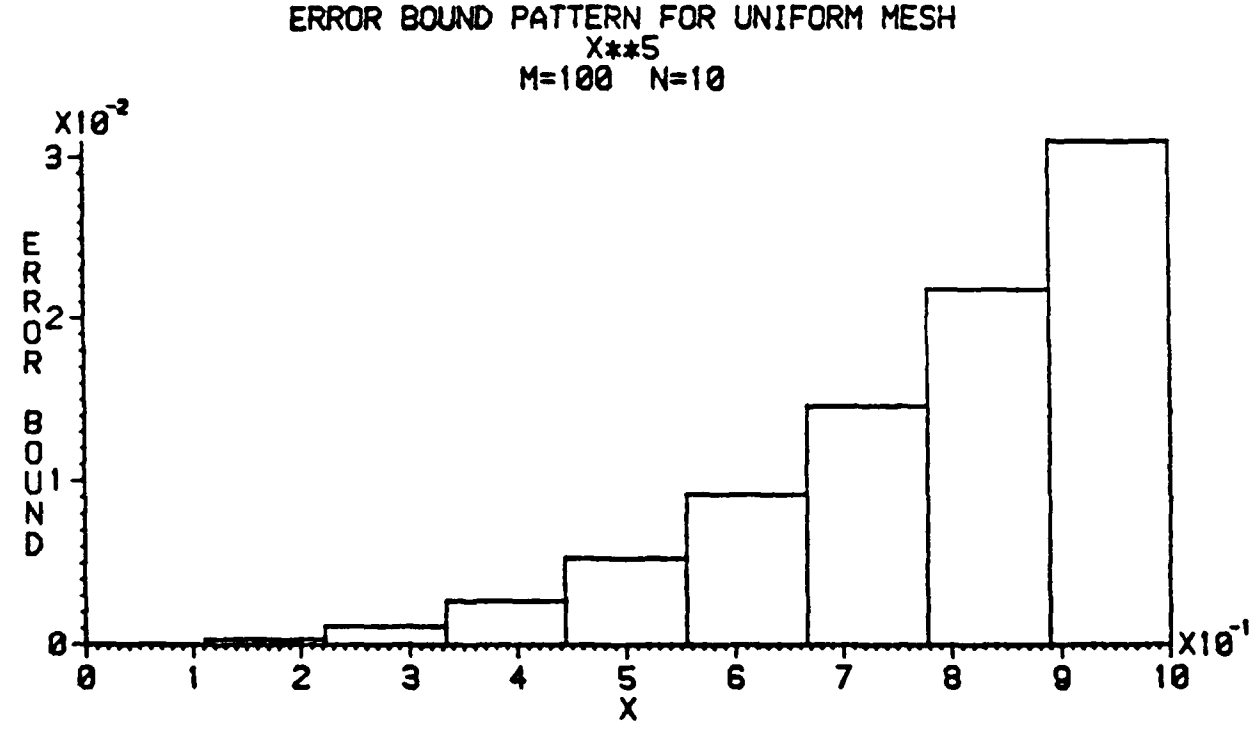


Figure 17

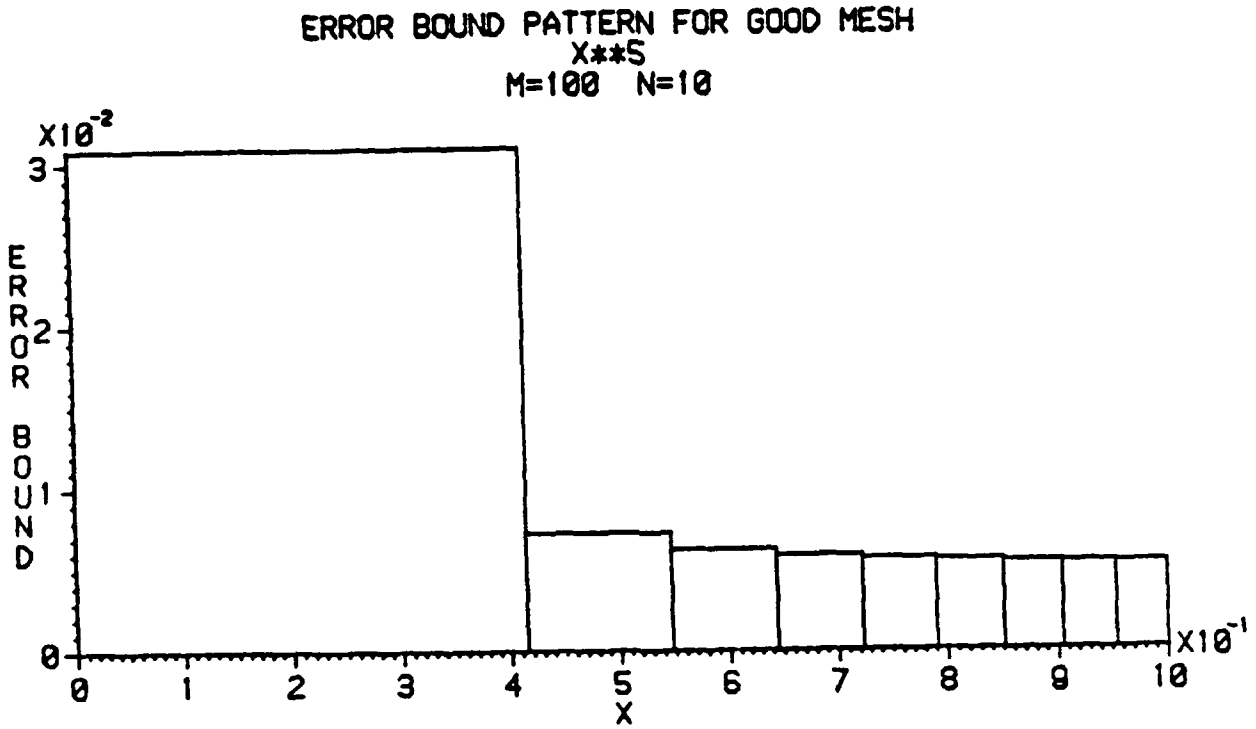


Figure 18

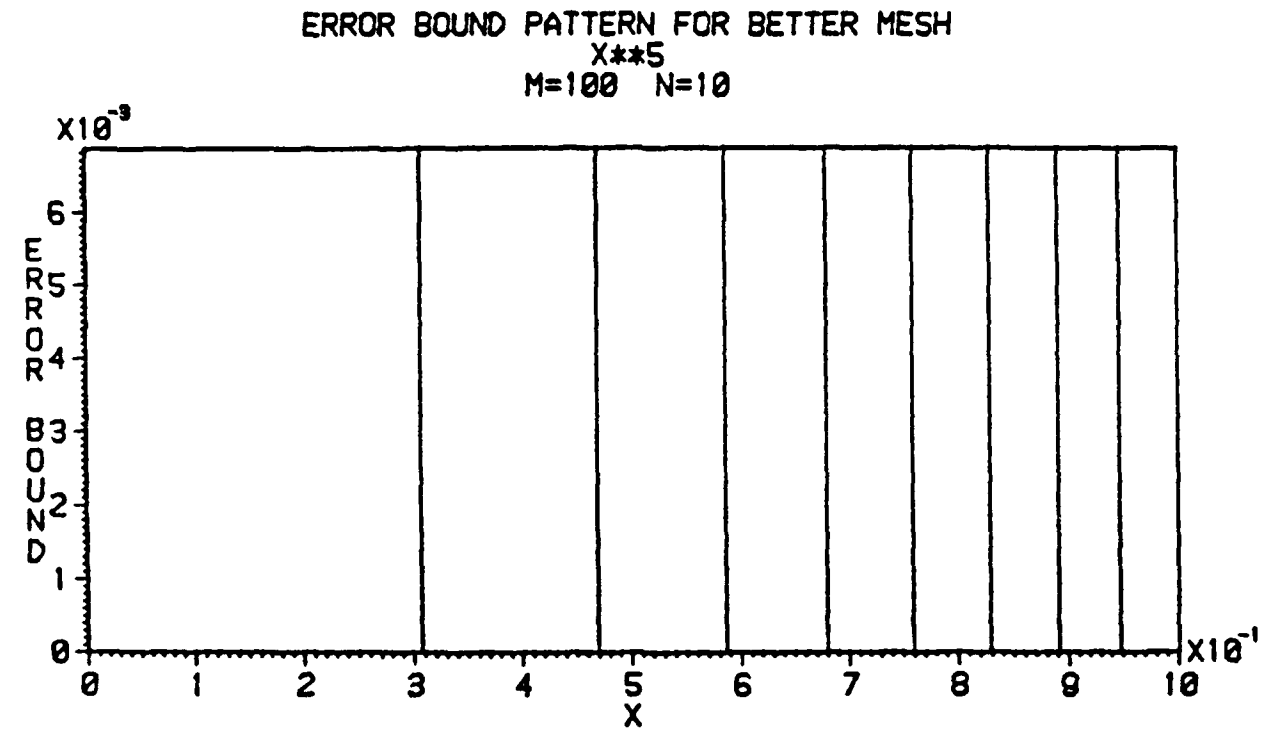


Figure 19

```
SUB UEBMSH ( M U V N X RELERR ) -- ----------UEBMSH
-- FOR UNIFORM ERROR BOUND MESH, SOLVE FMU(C)=0 BY BISECTION
(--
M=NO. OF ABSCISSAS IN PRELIMINARY MESH
U=ABSCISSAS IN PRELIMINARY MESH
V=APPROXIMATION TO SQUARE ROOT OF ABSOLUTE VALUE OF
  SECOND DERIVATIVE ON PRELIMINARY MESH
N=NO. OF POINTS IN FINAL MESH
X=ABSCISSAS IN FINAL MESH
RELERR=RELATIVE ERROR IN COMPUTING C
--)
DP U 1 , V 1 , X 1 , RELERR .
DP C1 , C2 , FMU , FC1 , FC2 , ABSERR , AE , C , FC , CC .
I=0 C1=0.0D0
DO I=I+1 # I >= M #
   C1=C1+(U(I+1)-U(I))*(V(I)+V(I+1))/2.0D0 OD
C1=C1/(N-1)
FC1=FMU(C1,M,U,V,N,X)
@ FC1 > 0.0D0 @
C2=1.3D0*C1
FC2=FMU(C2,M,U,V,N,X)
DO # FC2 < 0.0D0 #
   IF FC2 > 0.0D0 , C1=C2 FI
   C2=1.3D0*C2 FC2=FMU(C2,M,U,V,N,X) OD
ABSERR=RELERR*C1
AE=2.0D0*ABSERR
DO # AE < ABSERR #
   C=(C1+C2)/2.0D0
   FC=FMU(C,M,U,V,N,X)
   IF FC >= 0.0D0 , C1=C ; C2=C FI
   AE=C2-C1 OD
RET
==> GETC ( CC ) -- ----------GETC
-- GET CONSTANT
CC=C
RET END
DPFUN FMU ( C M U V N X ) -- ----------FMU
-- MESH FUNCTION OF C TO BE ZEROED TO GET UNIFORM ERROR BOUND MESH
DP FMU , C , U 1 , V 1 , X 1 .
DP A , VNORM , B .
I=1 A=U(1) X(1)=U(1) NU=0
DO FINDB ( A I C M U V VNORM J B ) NU=NU+1
   IF NU < N , X(NU+1)=B FI
   # B = U(M) #
   A=B I=J-1 OD
FMU=(NU-N)*C+(B-A)*VNORM
X(N)=U(M)
RET END
```

```
SUB FIND8 ( A I C M U V VNORM J B ) -- ------------------------FIND8
-- SOLVE (B-A)*SUPNORM(V OVER INTERVAL (A,B))=C FOR B
DP A , C , U 1 , V 1 , VNORM , B , VNORML , VNORMR .
OP S , T , E , DSQRT .
@ A >= U(I) @ @ A < U(I+1) @
J=I+1
S=(V(I+1)-V(I))/(U(I+1)-U(I))
VNORML=V(I)+S*(A-U(I))
-- FIND ROUGH LOCATION OF B
DO IF VNORML < V(J) , VNORMR=V(J) ; VNORMR=VNORML FI
   # VNORMR*(U(J)-A) > C #
   IF J = M , B=U(M) VNORM=VNORMR RET FI
   J=J+1 VNORML=VNORMR OD
-- FIND PRECISE LOCATION OF B
-- TAKE CARE OF TRIVIAL CASE
IF VNORML = VNORMR , B=A+C/VNORML VNORM=VNORML RET FI
-- COMPUTE TRANSITION POINT
S=(V(J)-V(J-1))/(U(J)-U(J-1))
T=U(J)-(VNORMR-VNORML)/S
-- TAKE CARE OF OTHER TRIVIAL CASE
IF (T-A)*VNORML >= C , B=A+C/VNORML VNORM=VNORML RET FI
-- TAKE CARE OF NONTRIVIAL CASE
E=VNORMR-S*(U(J)-A)
IF E >= 0.0D0 , B=A+2.0D0*C/(E+DSQRT(E**2+4.0D0*S*C)) ;
                B=A+(DSQRT(E**2+4.0D0*S*C)-E)/(2.0D0*S) FI
VNORM=V(J-1)+S*(B-U(J-1))
RET END
```

RELATIVISTIC THERMODYNAMICS OF REAL GASES WITH BROKEN INTERNAL SYMMETRY

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. The relativistic state equation for real gases with broken internal symmetry is developed. This is done by solving the complex form of a relativistic trace equation for the virial state equation of the real gases. The resulting solution affects only the third and higher virial coefficients. The complex relativistic third virial coefficient is given by a solution of two coupled differential equations. An approximate solution is found which is valid in the high temperature region, and an expression for the internal phase angle for the third virial coefficient is obtained. From this it is possible to develop expressions for the internal phase angles of the pressure, internal energy, entropy, enthalpy, and free energy of real gases that exhibit broken internal symmetries. Mixtures of interacting gases with broken internal symmetry are suggested to exhibit an interference phenomenon whereby the total pressure and internal energy will oscillate slightly in magnitude as the density of the system is increased. Accurate high temperature state equations of real gases are important for the description of the equilibrium configurations of gaseous stars, and for the description of nuclear explosions in the atmosphere.

1. INTRODUCTION. Spontaneously broken symmetry is a common phenomenon in physics because it appears in such diverse situations as ferromagnetism, superconductivity, weak interactions, and the vacuum screening currents that produce the asymmetric vacuum.[1-4] It is associated with a phase difference between a free particle in a potential and a particle in a coherent state of particles which forms due to some special system of forces. In superconductivity it is the Cooper pairs of electrons that form a self-coherent system which violates gauge invariance. In a similar fashion the vacuum state is thought to exhibit spontaneous symmetry breaking due to the presence of a Higgs scalar field which has a nonzero value for a minimum potential energy.

In relativistic thermodynamics a similar broken symmetry has been suggested to exist in the renormalized state equations of solids and quantum liquids.[5] This broken symmetry is associated with intrinsic phase angles of the thermodynamic state functions such as internal energy and pressure, and is due to the interaction of spacetime with bulk matter and the vacuum. The internal phase angles of the coherent state must be considered when applying the first and second laws of thermodynamics because the differentials of the state functions, such as the entropy and internal energy, must include a rotation in internal space.[5] This affects the measured state equation of thermodynamic systems such as solids, liquids and gases. This paper considers the vacuum induced broken symmetry of the state functions of the real gases. The broken symmetry appears in the third and higher order virial coefficients of the state equation for real gases.

The effects of the Minkowski metric of spacetime on the equation of state of bulk matter was originally described by the solutions of the scalar trace equation[6]

$$U + T\left(\frac{dU}{dT}\right)_{PV} - 3V\,\frac{d}{dV}(PV)_{U} = U^{a} + T\left(\frac{dU^{a}}{dT}\right)_{p^{a}V} \tag{1}$$

where U = relativistic (renormalized) internal energy, P = relativistic pressure, T = absolute temperature, V = volume of substance, and U^a and P^a = corresponding nonrelativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic (unrenormalized) calculations. It has been suggested that the spacetime induced broken symmetry effects on bulk matter can be represented by the following complex number trace equation[5]

$$\bar{U} + T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} - 3V\,\frac{d}{dV}(\bar{P}V)_{\bar{U}} = U^{a} + T\left(\frac{dU^{a}}{dT}\right)_{p^{a}V} \tag{2}$$

whose solution yields complex numbers, with internal phase angles, for the state functions of relativistic thermodynamics. Complex number solutions arise only in those cases for which there are nonzero gauge terms of the form[5,6]

$$T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} \neq 0 \tag{3}$$

$$\frac{d}{dV}(\bar{P}V)_{\bar{U}} \neq 0 \tag{4}$$

This includes interacting systems and the general case of the noninteracting relativistic Fermi gas.

The unrenormalized pressure and internal energy density for the real gases are given by[7,8]

$$P^{a} = nR^{a}T[1 + nB^{a}(T) + n^{2}C^{a}(T) + n^{3}D^{a}(T) + \cdots] \tag{5}$$

$$E^{a} = nR^{a}T\left[\frac{3}{2} - nT\,\frac{\partial B^{a}}{\partial T} - \frac{1}{2}\,n^{2}T\,\frac{\partial C^{a}}{\partial T} - \frac{1}{3}\,n^{3}T\,\frac{\partial D^{a}}{\partial T} - \cdots\right] \tag{6}$$

where

$$n = N/V = 1/\tilde{V} \tag{7}$$

where N = number of moles, $\tilde{V}$ = molar volume; R^a, $B^a(T)$, $C^a(T)$ and $D^a(T)$ = nonrelativistic values of the gas constant, second virial coefficient, third virial coefficient, and fourth virial coefficient respectively. The correspond-

ing renormalized pressure and energy density that are obtained from the solutions of the scalar trace equation (1) are written as[6,9]

$$P = nRT[1 + nB(T) + n^2C(T) + n^3D(T) + \cdots] \tag{8}$$

$$E = nRT[\frac{3}{2} - nT\frac{\partial B}{\partial T} - \frac{1}{2}n^2T\frac{\partial C}{\partial T} - \frac{1}{3}n^3T\frac{\partial D}{\partial T} - \cdots] \tag{9}$$

where

$$R = R^a \tag{9A}$$

$$B(T) = B^a(T) \tag{10}$$

$$C(T) = C^a(T) - 3[B^a(T)]^2 \ell n\ \psi^a \tag{11}$$

$$\psi^a = \frac{T}{T_R}\left|\frac{B^a(T)}{B^a(T_R)}\right|^{2/3} = \frac{T}{T_{CR}}\left|\frac{B^a(T)}{B^a(T_{CR})}\right|^{2/3} \tag{12}$$

and where T_R = relativity temperature constant, and T_{CR} = conjugate relativity temperature constant. An expression for the renormalized fourth virial coefficient D(T) has not been obtained.

The solution to the complex number trace equation (2) for the real gases is written as

$$\bar{P} = Pe^{j\theta P} = nRT[1 + n\bar{B}(T) + n^2\bar{C}(T) + n^3\bar{D}(T) + \cdots] \tag{13}$$

$$\bar{E} = Ee^{j\theta E} = nRT[\frac{3}{2} - nT\frac{\partial \bar{B}}{\partial T} - \frac{1}{2}n^2T\frac{\partial \bar{C}}{\partial T} - \frac{1}{3}n^3T\frac{\partial \bar{D}}{\partial T} - \cdots] \tag{14}$$

where

$$\bar{B} = Be^{j\theta B} \tag{15}$$

$$\bar{C} = Ce^{j\theta C} \tag{16}$$

$$\bar{D} = De^{j\theta D} \tag{16A}$$

are to be determined from a solution of the trace equation (2). Only $\bar{B}$ and $\bar{C}$ are obtained in this paper, and in fact $\bar{C}$ is obtained only through a high temperature approximation. It is the real parts of the complex number virial coefficients, pressure and internal energy that are the measured quantities.

The determination of P and θ_P for real gases with broken internal symmetry is important because the state equation of real gases enters the physical description of such diverse situations as the equilibrium configuration of stars and the latent heat associated with the gas-liquid phase transition. Consider for instance the Clausius-Clapyron equation for a real gas with broken internal symmetry[10-13]

$$\bar{\ell} = T(v_2 - v_1)\frac{d\bar{P}}{dT} \tag{17}$$

where $\bar{\ell}$ = complex number latent heat of vaporization, v_2 = specific volume of vapor, v_1 = specific volume of liquid, and $d\bar{P}/dT$ = slope of the vapor pressure curve. The complex number latent heat of vaporization can be written as

$$\bar{\ell} = \ell e^{j\theta_\ell} \tag{18}$$

where ℓ = magnitude of latent heat, and θ_ℓ = internal phase angle of the latent heat of vaporization. Equation (17) can then be written as the following three equations

$$\ell = T(v_2 - v_1)\sqrt{(\partial P/\partial T)_v^2 + (P\,\partial\theta_P/\partial T)_v^2} \tag{19}$$

$$\theta_\ell = \theta_P + \beta_{P,T} \tag{20}$$

$$\tan\beta_{P,T} = P\frac{\partial\theta_P/\partial T}{\partial P/\partial T} \tag{21}$$

Note that if $\theta_P = 0$ the standard form of the Clausius-Clapyron equation is regained. The measured value of the latent heat is $\ell\cos\theta_\ell$.

Accurate state equations for the real gases, including broken symmetry effects, are important for stellar structure calculations because the internal phase angle of the radial coordinate is related to the internal phase angle of the pressure. Therefore the complex values of the third and higher virial coefficients of the real gases will play an important role in stellar equilibrium calculations. In addition, it has been suggested that the third virial coefficient of the real gases can be utilized in the design of a gravitational wave detector.[9,14]

This paper calculates the internal phase angles of the second and third virial coefficients of the real gases. The phase angle of the second virial coefficient is determined to be equal to zero. The evaluation of the internal phase angle of the third virial coefficient is more complicated, and has been determined only in the regions of high temperature. The internal phase angles of the pressure, internal energy, entropy, enthalpy, and free energy are then

calculated in terms of the virial coefficients and their internal phases. The heat capacity for a real gas with broken internal symmetry is then evaluated. Finally, a general discussion of the interference effects expected to occur in mixtures of asymmetric gases is given.

2. THIRD VIRIAL COEFFICIENT FOR ASYMMETRIC REAL GASES. This section uses the complex number trace equation (2) to solve for the renormalized values of the second and third virial coefficients of the real gases with broken internal symmetry. It has been shown that the real number trace equation (1) does not change the value of the second virial coefficient as shown in equation (10).[6] In a similar fashion it is easy to show by substituting equations (13) and (14) into the complex number trace equation (2) that the second virial coefficient satisfies the following relation[6]

$$\frac{1}{\bar{\beta}}\frac{d\bar{\beta}}{dT} = \frac{1}{\beta^a}\frac{d\beta^a}{dT} \tag{22}$$

where

$$\bar{\beta} = RT\bar{B} \tag{23}$$

$$\beta^a = RTB^a \tag{24}$$

Equations (22) through (24) imply that

$$\bar{B} = Be^{j\theta_B} = B_R + jB_I = B^a = \text{real number} \tag{25}$$

which gives

$$B_I = 0 \tag{26}$$

$$\theta_B = 0 \tag{27}$$

$$B_R = B^a \tag{28}$$

Therefore the relativsitic value of the second virial coefficient, as determined from a solution of equation (2), is a real number which is equal to the unrenormalized value of the second virial coefficient.

The calculation of the complex number values of the third virial coefficients follows in a more complicated fashion from a solution of equation (2) for the real gases. An expedient way of doing the caluclation is to make use of the results for real gases that have been obtained for the relativistic form of the third virial coefficient from a solution of the scalar relativistic trace equation (1).[6] First define a function $\bar{f}(T)$ which is given by

$$RT\bar{C}(T) = RTC^{a}(T) + \bar{f}(T) \tag{29}$$

then substituting equations (13), (14) and (29) into the trace equation (2) gives the following equation for $\bar{f}(T)$[6]

$$\frac{d\bar{f}}{dT} + \frac{F}{T}\bar{f} = G \tag{30}$$

where

$$F = 1 - 2\frac{T}{\beta^{a}}\frac{d\beta^{a}}{dT} \tag{31}$$

$$G = \frac{3\beta^{a}}{RT^{2}}\left[R/c_{V}^{Ia}(\beta^{a} - T\,d\beta^{a}/dT) - \beta^{a}\right] \tag{32}$$

$$= -\frac{(\beta^{a})^{2}}{RT^{2}}\left(1 + 2\frac{T}{\beta^{a}}\frac{d\beta^{a}}{dT}\right)$$

where $c_{V}^{Ia} = 3R/2$. Equations (30) through (32) are of the same form that has already been obtained in Reference 6 for the scalar relativistic trace equation (1) except $\bar{f}(T)$ is a complex number.

The complex number $\bar{f}$ is written as follows

$$\bar{f} = -fe^{j\theta f} \tag{33}$$

$$f_{R} = -f\cos\theta_{f} \tag{34}$$

$$f_{I} = -f\sin\theta_{f} \tag{35}$$

with $f > 0$, and where f and θ_{f} are to be determined. Note that the use of the function f is different from that in Reference 6. In the present paper f is used as a magnitude of a complex number and is always positive. The choice of the negative sign in equation (33) is made so that θ_{f} is small and does not contain the π associated with negative real and imaginary values of $\bar{f}$ in the reions of high temperature. Placing equation (33) into equation (30) gives

$$e^{j\theta f}\left(\frac{df}{dT} + jf\frac{d\theta_{f}}{dT} + \frac{F}{T}f\right) = -G \tag{36}$$

Taking the real and imaginary parts of equation (36) yields

$$\cos \theta_f \left(\frac{df}{dT} + \frac{F}{T} f\right) - \sin \theta_f \, f \frac{d\theta_f}{dT} = - G \tag{37}$$

$$\sin \theta_f \left(\frac{df}{dT} + \frac{F}{T} f\right) + \cos \theta_f \, f \frac{d\theta_f}{dT} = 0 \tag{38}$$

Equations (37) and (38) are the two required equations for determining f and θ_f . Equations (37) and (38) can be rewritten as

$$\tan \theta_f = - \frac{f \, d\theta_f/dT}{\frac{df}{dT} + \frac{F}{T} f} \tag{39}$$

$$\left(\frac{df}{dT} + \frac{F}{T} f\right)^2 + \left(f \, d\theta_f/dT\right)^2 = G^2 \tag{40}$$

Equations (37) and (38) are difficult to solve without some approximations. In this paper a solution of equations (37) and (38) is obtained that is valid only for high temperatures.

An approximate solution for equations (37) and (38) can be obtained by assuming that θ_f is small so that

$$\sin \theta_f \sim \theta_f \sim 0 \tag{41}$$

$$\cos \theta_f \sim 1 \tag{42}$$

$$f = f_o \tag{43}$$

Substituting equations (41) through (43) into equations (37) and (38) yields

$$\frac{df_o}{dT} + \frac{F}{T} f_o = - G \tag{44}$$

$$- \theta_f G + f_o \frac{d\theta_f}{dT} = 0 \tag{45}$$

where all second order terms of θ_f are dropped. It has already been shown that the solution to equation (44) is[6]

$$f_o = 3RT(B^a)^2 \, \ell n \, \psi^a \tag{46}$$

so that from equations (33) through (35) it follows that

$$\bar{f} \sim - 3RT(B^a)^2 \ell n\, \psi^a\, e^{j\theta_f} \tag{47}$$

$$f_R \sim - 3RT(B^a)^2 \ell n\, \psi^a \cos \theta_f \tag{48}$$

$$f_I \sim - 3RT(B^a)^2 \ell n\, \psi^a \sin \theta_f \tag{49}$$

This solution is valid only when $f_o > 0$ since in fact f is the magnitude of a complex number. Therefore the assumption that θ_f is small and that $f_o > 0$ restricts the validity of the approximate solution to the regions of high temperature. For this case $f_R < 0$ and $f_I \lesseqgtr 0$. Combining equations (29) and (47) gives the following approximate relationships for the renormalized third virial coefficients

$$\bar{C}(T) = C^a(T) - 3(B^a)^2 \ell n\, \psi^a\, e^{j\theta_f} \tag{50}$$

$$C_R(T) = C^a(T) - 3(B^a)^2 \ell n\, \psi^a \cos \theta_f \tag{51}$$

$$C_I(T) = - 3(B^a)^2 \ell n\, \psi^a \sin \theta_f \tag{52}$$

It is the real value of the third virial coefficient C_R that is measured from experimental pressure versus volume curves at constant temperature.

It remains only to obtain θ_f from a solution of equation (45) which can be rewritten as

$$d\theta_f/\theta_f = G/f_o\, dT \tag{53}$$

$$\ell n|\theta_f| = \int G/f_o\, dT \tag{54}$$

Combining equations (24), (32) and (46) gives

$$G = - R(B^a)^2 \left(3 + 2 \frac{T}{B^a} \frac{dB^a}{dT}\right) \tag{55}$$

$$G/f_o = - \left(\frac{1}{T} + \frac{2}{3B^a} \frac{dB^a}{dT}\right) / \ell n\, \psi^a \tag{56}$$

But from equation (12) it follows that

$$\frac{d\ \ell n\ \psi^a}{dT} = \frac{1}{T} + \frac{2}{3B^a}\frac{dB^a}{dT} \tag{57}$$

Combining equations (54), (56) and (57) gives

$$\ell n|\theta_f| = -\int \frac{1}{\ell n\ \psi^a}\frac{d\ \ell n\ \psi^a}{dT}\,dT = -\ell n(\ell n\ \psi^a) + \ell n\ b \tag{58}$$

and therefore

$$\theta_f = \pm b/\ell n\ \psi^a \tag{59}$$

where b = constant whose value can only be determined from the full solution of equations (37) and (38). The solution in equation (59) is valid only when θ_f is a small number, and in general this limits the application of equation (59) to the regions of high temperature. Note that equations (37) and (38) are unchanged for $\theta_f \rightarrow -\theta_f$, so that either $\pm\theta_f$ are valid solutions, and therefore these equations exhibit degeneracy.

The relativistic third virial coefficient can be obtained from equations (50) through (52) as

$$\bar{C} = Ce^{j\theta C} = C_R + jC_I \tag{60}$$

$$C_R = C\cos\theta_C \tag{61}$$

$$C_I = C\sin\theta_C \tag{62}$$

and therefore

$$\tan\theta_C = C_I/C_R \tag{63}$$

$$\sin\theta_C = C_I/C \tag{64}$$

$$\cos\theta_C = C_R/C \tag{65}$$

where C_R and C_I are given by equations (51) and (52) and where the magnitude of the third virial coefficient is given by

$$C^2 = C_R^2 + C_I^2 \tag{66}$$

$$= [C^a - 3(B^a)^2 \ell n\, \psi^a]^2 + 6C^a(B^a)^2 \ell n\, \psi^a\, (1 - \cos\theta_f)$$

when a single phase θ_f of the gas is present. At high temperatures $C_R < 0$, and either $C_I < 0$ for positive θ_f, or $C_I > 0$ for $\theta_f < 0$. For $C_I < 0$ it follows from equations (63) through (65) that $\theta_C = \pi + \theta_C'$ and θ_C is in the third quadrant, while for $C_I > 0$ it follows that $\theta_C = \pi - \theta_C'$ and θ_C is in the second quadrant. Therefore equations (37) and (38) have degenerate solutions, and the real gases can appear in two states corresponding to $\pm\theta_f$. If a real gas is in a state which is a mixture of fraction α with $\theta_f > 0$ and fraction $(1 - \alpha)$ with $\theta_f < 0$ it follows

$$C_I(T) = -3(2\alpha - 1)(B^a)^2 \ell n\, \psi^a \sin|\theta_f| \tag{52A}$$

Therefore for equal mixtures of both phases $C_I = 0$ and $\theta_C = \pi$. The function $C_R(T)$ is determined in static pressure versus volume measurements at constant temperature, and is not affected by the sign of the phase function θ_f. Finally it should be noted that the fourth virial coefficient can be written as

$$\bar{D} = De^{j\theta_D} \tag{67}$$

but no calculations to determine D and θ_D have been done.

3. INTERNAL PHASE ANGLES OF THERMODYNAMIC FUNCTIONS. This section considers the calculation of the pressure, internal energy, entropy, enthalpy, and free energy of real gases with broken internal symmetry. The relativistic pressure and internal energy density for a broken symmetry real gas are given in equations (13) and (14). The corresponding expressions for the entropy density, enthalpy density and free energy density are given for asymmetric real gases as a generalization of the standard results in the literature.[8] The basic thermodynamic functions for a broken symmetry real gas are then given as follows

A. Pressure.

The pressure is written as

$$\bar{P} = Pe^{j\theta_P} = nRT(1 + nB + n^2\bar{C} + n^3\bar{D} + \cdots) \tag{68}$$

or in component form

$$P_R = nRT(1 + nB + n^2C \cos \theta_C + n^3D \cos \theta_D + \cdots) \tag{69}$$

$$P_I = nRT(n^2C \sin \theta_C + n^3D \sin \theta_D + \cdots) \tag{70}$$

$$\tan \theta_P = \frac{n^2(C \sin \theta_C + nD \sin \theta_D + \cdots)}{(1 + nB + n^2C \cos \theta_C + n^3D \cos \theta_D + \cdots)} \tag{71}$$

$$\sim n^2C \sin \theta_C \qquad \text{low density} \tag{72}$$

The magnitude of the pressure is given by

$$P = (P_R^2 + P_I^2)^{1/2} \tag{73}$$

It is the real part of the pressure P_R that is measured in laboratory experiments. For $\theta_f > 0$ and $\theta_C = \pi + \theta_C'$ it follows from equations (69) through (72) that $\theta_P < 0$ in the regions of high temperature, but for the case $\theta_f < 0$ and $\theta_C = \pi - \theta_C'$ it follows that $\theta_P > 0$. For an equal mixture of states with $\theta_f > 0$ and $\theta_f < 0$ it follows that $\theta_C = \pi$ and $\theta_P = 0$ neglecting the effects of the fourth and higher virial coefficients.

B. Internal Energy Density.

The energy density for an asymmetric real gas is written as

$$\bar{E} = Ee^{j\theta_E} = nRT\left(\frac{3}{2} - nT\frac{\partial B}{\partial T} - \frac{1}{2}n^2T\frac{\partial \bar{C}}{\partial T} - \frac{1}{3}n^3T\frac{\partial \bar{D}}{\partial T} - \cdots\right)$$

Using equations (60) and (67) in equation (74) gives

$$E_R = nRT\left[\frac{3}{2} - nT\frac{\partial B}{\partial T} - \frac{1}{2}n^2I_C \cos(\theta_C + \beta_{C,T}) - \frac{1}{3}n^3I_D \cos(\theta_D + \beta_{D,T}) - \cdots\right] \tag{75}$$

$$E_I = -n^3RT\left[\frac{1}{2}I_C \sin(\theta_C + \beta_{C,T}) + \frac{1}{3}nI_D \sin(\theta_D + \beta_{D,T}) + \cdots\right] \tag{76}$$

$$\tan \theta_E = E_I/E_R \tag{77}$$

$$\sim -\frac{1}{3}n^2I_C \sin(\theta_C + \beta_{C,T}) \qquad \text{low density} \tag{78}$$

where

$$I_C = \sqrt{(T\,\partial C/\partial T)^2 + (CT\,\partial\theta_C/\partial T)^2} \tag{79}$$

$$I_D = \sqrt{(T\,\partial D/\partial T)^2 + (DT\,\partial\theta_D/\partial T)^2} \tag{80}$$

$$\tan\beta_{C,T} = C\,\frac{\partial\theta_C/\partial T}{\partial C/\partial T} \tag{81}$$

$$\tan\beta_{D,T} = D\,\frac{\partial\theta_D/\partial T}{\partial D/\partial T} \tag{82}$$

The magnitude of the internal energy density is given by

$$E = \sqrt{E_R^2 + E_I^2} \tag{83}$$

It is E_R that is measured in the laboratory

C. Entropy.

The entropy density for an asymmetric real gas is written as

$$\bar{s} = se^{j\theta s} = -\,nR[\ell n(nRT) + n(B + T\,\frac{\partial B}{\partial T}) + \frac{1}{2}\,n^2(\bar{C} + T\,\frac{\partial \bar{C}}{\partial T}) \tag{84}$$

$$+ \frac{1}{3}\,n^3(\bar{D} + T\,\frac{\partial \bar{D}}{\partial T}) + \cdots\,]$$

Using equations (60) and (67) in equation (84) gives

$$s_R = -\,nR[\ell n(nRT) + n(B + T\,\frac{\partial B}{\partial T}) + \frac{1}{2}\,n^2 J_C\cos(\theta_C + \alpha_{C,T}) \tag{85A}$$

$$+ \frac{1}{3}\,n^3 J_D\cos(\theta_D + \alpha_{D,T}) + \cdots\,]$$

$$s_I = -\,n^3R[\,\frac{1}{2}\,J_C\sin(\theta_C + \alpha_{C,T}) + \frac{1}{3}\,nJ_D\sin(\theta_D + \alpha_{D,T}) + \cdots\,] \tag{85B}$$

$$\tan\theta_s = s_I/s_R \tag{86}$$

$$\sim \frac{1}{2}\,n^2 J_C\,\frac{\sin(\theta_C + \alpha_{C,T})}{\ell n(nRT)} \qquad \text{low density} \tag{87}$$

where

$$K_C = \sqrt{(2C - T\,\partial C/\partial T)^2 + (CT\,\partial\theta_C/\partial T)^2} \tag{98}$$

$$K_D = \sqrt{(3D - T\,\partial D/\partial T)^2 + (DT\,\partial\theta_D/\partial T)^2} \tag{99}$$

$$\tan\,\eta_C = \frac{CT\,\partial\theta_C/\partial T}{2C - T\,\partial C/\partial T} \tag{100}$$

$$\tan\,\eta_D = \frac{DT\,\partial\theta_D/\partial T}{3D - T\,\partial D/\partial T} \tag{101}$$

The magnitude of the enthalpy density is given by

$$h = \sqrt{h_R^2 + h_I^2} \tag{102}$$

The value of h_R is obtained from laboratory measurements.

E. Free Energy Density.

The complex number free energy density for an asymmetric real gas is given by

$$\bar{a} = ae^{j\theta_a} = nRT[\tfrac{3}{2} + \ell n(nRT) + nB + \tfrac{1}{2}n^2\bar{C} + \tfrac{1}{3}n^3\bar{D} + \cdots] \tag{103}$$

The real and imaginary parts of equation (103) are

$$a_R = nRT[\tfrac{3}{2} + \ell n(nRT) + nB + \tfrac{1}{2}n^2C\cos\theta_C + \tfrac{1}{3}n^3D\cos\theta_D + \cdots] \tag{104}$$

$$a_I = n^3RT(\tfrac{1}{2}C\sin\theta_C + \tfrac{1}{3}nD\sin\theta_D + \cdots) \tag{105}$$

$$\tan\theta_a = a_I/a_R \tag{105A}$$

$$\sim \tfrac{1}{2}n^2C\sin\theta_C/[\tfrac{3}{2} + \ell n(nRT)] \quad \text{low density} \tag{106}$$

The real part of the free energy density is a physically measurable quantity.

4. HEAT CAPACITY AND GRÜNEISEN PARAMETER. The calculation of the heat capacity and Grüneisen parameter for a real gas with broken internal symmetries

where nRT > 1 and where

$$J_C = \sqrt{(C + T\,\partial C/\partial T)^2 + (CT\,\partial\theta_C/\theta T)^2} \tag{88}$$

$$J_D = \sqrt{(D + T\,\partial D/\partial T)^2 + (DT\,\partial\theta_D/\partial T)^2} \tag{89}$$

$$\tan \alpha_{C,T} = \frac{CT\,\partial\theta_C/\partial T}{C + T\,\partial C/\partial T} \tag{90}$$

$$\tan \alpha_{D,T} = \frac{DT\,\partial\theta_D/\partial T}{D + T\,\partial D/\partial T} \tag{91}$$

The magnitude of the entropy per unit volume is given by

$$s = \sqrt{s_R^2 + s_I^2} \tag{92}$$

The real quantity is the entropy density measured in the laboratory.

D. Enthalpy.

The enthalpy density of an asymmetric real gas is given by

$$\bar{h} = he^{j\theta h} = nRT\left[\frac{5}{2} + n\left(B - T\frac{\partial B}{\partial T}\right) + \frac{1}{2}n^2\left(2\bar{C} - T\frac{\partial \bar{C}}{\partial T}\right)\right. \tag{93}$$

$$\left. + \frac{1}{3}n^3\left(3\bar{D} - T\frac{\partial \bar{D}}{\partial T}\right) + \cdots\right]$$

Placing equations (60) and (67) into equation (93) gives

$$h_R = nRT\left[\frac{5}{2} + n\left(B - T\frac{\partial B}{\partial T}\right) + \frac{1}{2}n^2K_C\cos(\theta_C - \eta_C)\right. \tag{94}$$

$$\left. + \frac{1}{3}n^3K_D\cos(\theta_D - \eta_D) + \cdots\right]$$

$$h_I = n^3RT\left[\frac{1}{2}K_C\sin(\theta_C - \eta_C) + \frac{1}{3}nK_D\sin(\theta_D - \eta_D) + \cdots\right] \tag{95}$$

$$\tan\theta_h = h_I/h_R \tag{96}$$

$$\sim \frac{1}{5}n^2K_C\sin(\theta_C - \eta_C) \qquad \text{low density} \tag{97}$$

is performed in this section. The complex number molar heat capacity is obtained from equation (79) to be[8]

$$\bar{C}_V = C_V e^{j\theta CV} = C_{VR} + jC_{VI} = R\left(\frac{3}{2} - nC_{V1} - \frac{1}{2}n^2\bar{C}_{V2} - \cdots\right) \tag{108}$$

where

$$C_{V1} = T^2\frac{\partial^2 B}{\partial T^2} + 2T\frac{\partial B}{\partial T} \tag{109}$$

$$\bar{C}_{V2} = T^2\frac{\partial^2 \bar{C}}{\partial T^2} + 2T\frac{\partial \bar{C}}{\partial T} = C_{V2R} + jC_{V2I} \tag{110}$$

Substituting equation (60) into equation (110) gives

$$\bar{C}_{V2} = I_{2C}e^{j\theta 2C} + 2I_C e^{j\theta 1C} \tag{111}$$

$$C_{V2R} = I_{2C}\cos\theta_{2C} + 2I_C\cos\theta_{1C} \tag{112}$$

$$C_{V2I} = I_{2C}\sin\theta_{2C} + 2I_C\sin\theta_{1C} \tag{113}$$

where I_C is given by equation (79) and

$$I_{2C} = \sqrt{L_{2C}^2 + M_{2C}^2} \tag{114}$$

$$L_{2C} = T^2\frac{\partial^2 C}{\partial T^2} - C\left(T\frac{\partial\theta_C}{\partial T}\right)^2 \tag{114A}$$

$$M_{2C} = CT^2\frac{\partial^2\theta_C}{\partial T^2} + 2\left(T\frac{\partial C}{\partial T}\right)\left(T\frac{\partial\theta_C}{\partial T}\right) \tag{114B}$$

$$\theta_{2C} = \theta_C + \gamma_{C,T} \tag{115}$$

$$\tan\gamma_{C,T} = M_{2C}/L_{2C} \tag{116}$$

$$\theta_{1C} = \theta_C + \beta_{C,T} \tag{117}$$

where $\beta_{C,T}$ is given by equation (81). Then it follows from equation (108) that

$$C_{VR} = R\left(\frac{3}{2} - nC_{V1} - \frac{1}{2} n^2 C_{V2R} - \cdots\right) \tag{118}$$

$$C_{VI} = -\frac{1}{2} n^2 R C_{V2I} \tag{119}$$

$$\tan \theta_{CV} = C_{VI}/C_{VR} \tag{120}$$

$$\nu - \frac{1}{3} n^2 (I_{2C} \sin \theta_{2C} + 2I_C \sin \theta_{1C}) \tag{121}$$

where equation (121) is valid for low densities.

The Grüneisen parameter for an asymmetric real gas can be written as[9]

$$\bar{\gamma} = \gamma e^{j\theta_\gamma} = \gamma_R + j\gamma_I = \frac{2}{3}(1 + n\gamma_1 + n^2\bar{\gamma}_2 + \cdots) \tag{122}$$

where

$$\gamma_1 = f_1 + g_1 \tag{123}$$

$$\bar{\gamma}_2 = \bar{f}_2 + \bar{g}_2 + f_1 g_1 \tag{124}$$

$$f_1 = T\frac{\partial B}{\partial T} + B \tag{125}$$

$$g_1 = \frac{2}{3}\left(T^2 \frac{\partial^2 B}{\partial T^2} + 2T \frac{\partial B}{\partial T}\right) \tag{126}$$

$$\bar{f}_2 = T\frac{\partial \bar{C}}{\partial T} + \bar{C}$$

$$\bar{g}_2 = \frac{1}{3}\left(T^2 \frac{\partial^2 \bar{C}}{\partial T^2} + 2T \frac{\partial \bar{C}}{\partial T}\right) + \frac{4}{9}\left(T^2 \frac{\partial^2 B}{\partial T^2} + 2T \frac{\partial B}{\partial T}\right)^2 \tag{128}$$

From equation (122) it follows that

$$\gamma_R = \frac{2}{3}(1 + n\gamma_1 + n^2\gamma_{2R} + \cdots) \tag{129}$$

$$\gamma_I = \frac{2}{3} n^2 \gamma_{2I} + \cdots \tag{130}$$

where

$$\gamma_{2R} = f_{2R} + g_{2R} + f_1 g_1 \tag{131}$$

$$\gamma_{2I} = f_{2I} + g_{2I} \tag{132}$$

$$f_{2R} = J_C \cos(\theta_C + \alpha_{C,T}) \tag{133}$$

$$f_{2I} = J_C \sin(\theta_C + \alpha_{C,T}) \tag{134}$$

$$g_{2R} = \frac{1}{3}(I_{2C} \cos\theta_{2C} + 2I_C \cos\theta_{1C}) + \frac{4}{9}(T^2 \frac{\partial^2 B}{\partial T^2} + 2T \frac{\partial B}{\partial T})^2 \tag{135}$$

$$g_{2I} = \frac{1}{3}(I_{2C} \sin\theta_{2C} + 2I_C \sin\theta_{1C}) \tag{136}$$

where J_C and $\alpha_{C,T}$ are given by equations (88) and (90) respectively, I_C and θ_{1C} are given by equations (79) and (117) respectively, and where I_{2C} and θ_{2C} are given by equations (114) and (115) respectively.

5. SUPERPOSITION OF THERMODYNAMIC FUNCTIONS. Consider a mixture of two interacting gases with broken internal symmetries. There will be interactions between the two species of gas as well as self interactions within each species. Therefore the total pressure and internal energy is written as

$$\bar{P} = \bar{P}_1 + \bar{P}_2 + \bar{P}_{12} \tag{137}$$

$$\bar{U} = \bar{U}_1 + \bar{U}_2 + \bar{U}_{12} \tag{138}$$

where $\bar{P}_{12}$ and $\bar{U}_{12}$ are the interspecies interaction pressure and internal energy respectively. For asymmetric real gases the terms in equations (137) and (138) are written as

$$\bar{P} = Pe^{j\theta P} \qquad \bar{P}_1 = P_1 e^{j\theta P1} \tag{139}$$

$$\bar{P}_2 = P_2 e^{j\theta P2} \qquad \bar{P}_{12} = P_{12} e^{j\theta P12} \tag{140}$$

$$\bar{U} = Ue^{j\theta U} \qquad \bar{U}_1 = U_1 e^{j\theta U1} \tag{141}$$

$$\bar{U}_2 = U_2 e^{j\theta U2} \qquad \bar{U}_{12} = U_{12} e^{j\theta U12} \tag{142}$$

Equations (137) and (138) can be written in component form as follows

$$P \cos \theta_P = P_1 \cos \theta_{P1} + P_2 \cos \theta_{P2} + P_{12} \cos \theta_{P12} \tag{143}$$

$$P \sin \theta_P = P_1 \sin \theta_{P1} + P_2 \sin \theta_{P2} + P_{12} \sin \theta_{P12} \tag{144}$$

$$U \cos \theta_U = U_1 \cos \theta_{U1} + U_2 \cos \theta_{U2} + U_{12} \cos \theta_{U12} \tag{145}$$

$$U \sin \theta_U = U_1 \sin \theta_{U1} + U_2 \sin \theta_{U2} + u_{12} \sin \theta_{U12} \tag{146}$$

From equations (143) through (146) it follows that

$$\tan \theta_P = \frac{P_1 \sin \theta_{P1} + P_2 \sin \theta_{P2} + P_{12} \sin \theta_{P12}}{P_1 \cos \theta_{P1} + P_2 \cos \theta_{P2} + P_{12} \cos \theta_{P12}} \tag{147}$$

$$\tan \theta_U = \frac{U_1 \sin \theta_{U1} + U_2 \sin \theta_{U2} + U_{12} \sin \theta_{U12}}{U_1 \cos \theta_{U1} + U_2 \cos \theta_{U2} + U_{12} \cos \theta_{U12}} \tag{148}$$

$$P = (\psi_D + \psi_I)^{1/2} \tag{149}$$

$$U = (\phi_D + \phi_I)^{1/2} \tag{150}$$

where

$$\psi_D = P_1^2 + P_2^2 + P_{12}^2 \tag{151}$$

$$\psi_I = 2P_1P_2 \cos (\theta_{P1} - \theta_{P2}) + 2P_1P_{12} \cos (\theta_{P1} - \theta_{P12}) + 2P_2P_{12} \cos (\theta_{P2} - \theta_{P12}) \tag{152}$$

$$\phi_D = U_1^2 + U_2^2 + U_{12}^2 \tag{153}$$

$$\phi_I = 2U_1U_2 \cos (\theta_{U1} - \theta_{U2}) + 2U_1U_{12} \cos (\theta_{U1} - \theta_{U12}) + 2U_2U_{12} \cos (\theta_{U2} - \theta_{U12}) \tag{154}$$

Therefore mixtures of two asymmetric real gases should exhibit interference in regard to the component pressures and internal energies, and the magnitude

of the total pressure and internal energy should exhibit a small oscillation as the density of the interacting mixture is increased. This is also true of the measured pressure and internal energy which can respectively be written as $P \cos \theta_P$ and $U \cos \theta_U$.

6. CONCLUSIONS. Real gases are expected to exhibit broken internal symmetries that manifest themselves as internal phase angles associated with the thermodynamic functions such as pressure, internal energy and entropy. These internal phase angles arise from the third and higher virial coefficients and are due to renormalization effects associated with the interaction of matter with spacetime as described by equation (2). The ideal gas and the second virial coefficient of an interacting gas are unaffected by spacetime interactions. The phase angle associated with the third virial coefficient can be determined from the solution of two simultaneous first order differential equations. These equations are generally not easy to solve analytically, but yield a simple solution for the high temperature regions of the real gas where the phase angle of the third virial coefficient is small. The virial form of the state equation for real gases allows the phase angles associated with the pressure, internal energy, entropy, enthalpy, and free energy to be calculated in terms of density and temperature. The existence of internal phase angles for the thermodynamic state functions suggests that mixtures of real gases will produce an interference phenomenon wherein the total pressure and internal energy will oscillate slightly as the density of the system is increased.

It can also be conjectured that parabolic waves of the form[15,16]

$$\frac{\partial \theta_U}{\partial \bar{t}} = f(\theta_U) + D_U \frac{\partial^2 \theta_U}{\partial \bar{x}^2} \tag{155}$$

$$\frac{\partial \theta_P}{\partial \bar{t}} = f(\theta_P) + D_P \frac{\partial^2 \theta_P}{\partial \bar{x}^2} \tag{156}$$

can exist in asymmetric gases and liquids as well as in asymmetric solids and quantum liquids, and these wave motions may have interesting applications to thermodynamics, hydrodynamics and chemical and biological cycles. The internal phase angles of the pressure and other thermodynamic functions of the real gases are also expected to play an important role in the determination of the equilibrium configuration of stars and planets. This is true because the internal phase angles of the radial coordinates in a star are determined by the internal phase angle of the pressure. Therefore the complex number values of the third and higher virial coefficients are intimately involved in stellar equilibrium calculations.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Aitchison, I. J. R. and Hey, A. J. G., Gauge Theories in Particle Physics, Adam Hilger Ltd., Bristol, United Kingdom, 1982.

2. O'Raifeartaigh, L., Group Structure of Gauge Theories, Cambridge University Press, London, 1986.

3. Bailin, D., Weak Interactions, Adam Hilger, Bristol, 1982.

4. Grassie, A. D. C., The Superconducting State, Sussex University Press, London, 1975.

5. Weiss, R. A., "Thermodynamic Gauge Theory of Solids and Quantum Liquids with Internal Phase", Fifth Army Conference on Applied Mathematics and Computing, West Point, New York, ARO 88-1, June 15-18, 1987, p. 649.

6. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.

7. Hirschfelder, J. O., Curtiss, C. F. and Bird, R. B., Molecular Theory of Gases and Liquids, John Wiley, New York, 1954.

8. Beattie, J. A., "Thermodynamic Properties of Real Gases and Mixtures of Real Gases", article in Thermodynamics and Physics of Matter, edited by F. D. Rossini, Princeton University Press, 1955, p. 240.

9. Weiss, R. A., "Relativistic Wave Equations for Real Gases", Fourth Army Conference on Applied Mathematics and Computing, Cornell University, ARO 87-1, May 27-30, 1986, p. 341.

10. Sears, F. W., Thermodynamics, the Kinetic Theory of Gases, and Statistical Mechanics, Addison-Wesley, Reading, MA, 1953.

11. Planck, M., Theory of Heat, MacMillan, New York, 1949.

12. Pauli, W., Pauli Lectures on Physics: Volume 3. Thermodynamics and the Kinetic Theory of Gases, MIT Press, Cambridge, MA, 1973.

13. Sommerfeld, A., Thermodynamics and Statistical Mechanics, Academic Press, New York, 1955.

14. Michelson, P. F., Price, J. C. and Taber, R. C., "Resonant-Mass Detectors of Gravitational Radiation", Science, Vol. 237, 10 July 1987, p. 150.

15. Winfree, A. T., The Geometry of Biological Time, Springer-Verlag, New York, 1980.

16. Winfree, A. T., When Time Breaks Down, Princeton University Press, 1987.

GAUGE THEORY OF ATOMIC PROCESSES

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. Atomic particle processes that occur within bulk matter or the vacuum are expected to be influenced by the broken symmetries of the thermodynamic ground and excited states of these systems. Internal phase angles are associated with the space and time coordinates and kinematic and dynamic variables of particles and radiation in bulk matter or vacuum with broken internal symmetries. A broken symmetry photon gas in bulk matter or vacuum is considered, and the radiation pressure and energy density is calculated. The geometric angles between kinematic variables and between dynamic variables have internal phase angles. This affects the description of the photoelectric effect, Compton effect, and Coulomb scattering in bulk matter and the vacuum. Thomson, Compton, Rutherford, Mott, Bhabha, and Møller scattering processes in broken symmetry systems are investigated. The Schrödinger and Dirac equations are developed for a particle located in bulk matter or vacuum with broken internal symmetries. This work will have applications to nuclear explosions and the interaction of directed energy beams with matter.

1. INTRODUCTION. In the past decade, great advances were made in the theory of the elementary forces which bind the universe. These advances developed through the realization that gauge theory is the natural framework for describing the four basic interactions that occur in nature.[1-3] Gauge theory was first formulated many years ago by Hermann Weyl, but only recently has its real importance to physics been understood.[1] In some cases when gauge symmetry is broken spontaneously by some special set of forces, a coherent state of matter can be formed as in the case of superconductivity where the Cooper pairs of electrons break the ground state gauge symmetry through electron-phonon interactions.[4]

It has been suggested that vacuum interactions with bulk matter may produce a coherent ground state which is described by thermodynamic functions that possess internal phase angles.[5] This coherent broken symmetry ground state can possibly influence the microscopic processes that take place in bulk matter. The effect can occur in two ways: first, through the Euler equations by which fluid elements are expected to have space and time coordinates and kinematic and dynamic variables that have internal phase angles. Secondly, a microscopic gauge interaction between material particles can be induced by Minkowski spacetime, and this complex number gauge interaction will impart internal phases to the space and time coordinates and to the kinematic and dynamic variables of individual particles. Therefore individual particles in bulk matter require complex numbers for their kinematic and dynamic descriptions and for their coordinate locations in space and time. The same conclusions are valid for the vacuum with broken internal symmetries, because the vacuum can be considered as a special simplified case of bulk matter.

The coherent state of bulk matter is due to spacetime interactions, and these have been described by a bulk matter relativistic trace equation whose scalar form for symmetrical bulk matter is[6]

$$U_s + T\left(\frac{dU_s}{dT}\right)_{P_sV} - 3V\frac{d}{dV}(P_sV)_{U_s} = U^a + T\left(\frac{dU^a}{dT}\right)_{pa_V} \tag{1}$$

where U_s = renormalized internal energy for symmetric bulk matter, P_s = renormalized pressure for symmetric bulk matter, T = absolute temperature, V = volume of substance, and U^a and P^a = corresponding nonrelativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic (unrenormalized) calculations. The complex number form of the relativistic trace equation that describes the coherent broken symmetry state of bulk matter is given by[5]

$$\bar{U} + T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} - 3V\frac{d}{dV}(\bar{P}V)_{\bar{U}} = U^a + T\left(\frac{dU^a}{dT}\right)_{pa_V} \tag{2}$$

or equivalently as

$$(1 - \bar{b} + T\frac{\partial}{\partial T} - \bar{b}V\frac{\partial}{\partial V})\bar{E} - 3(1 + \bar{\gamma} + V\frac{\partial}{\partial V} - \bar{\gamma}T\frac{\partial}{\partial T})\bar{P} = \psi^a \tag{3}$$

where

$$\psi^a = (T\frac{\partial}{\partial T} - b^aV\frac{\partial}{\partial V} + 1 - b^a)E^a \tag{4}$$

and where $\bar{U}$, $\bar{E}$, $\bar{P}$, $\bar{\gamma}$, and $\bar{b}$ are complex number representations of the internal energy, energy density, pressure, and the gauge parameters.[5] With their right hand sides set equal to zero, equations (2) and (3) describe the broken symmetry thermodynamic ground state of the vacuum. Therefore the broken symmetry thermodynamic ground state of the vacuum is a simpler special case of the broken symmetry state of bulk matter.

Due to the spacetime interactions with bulk matter, the single particle energy must contain a gauge potential that produces the difference between U and U^a at the macroscopic level. Corresponding to equation (1) the noninteracting single particle energy is given by[7]

$$\varepsilon_{is}^{free} = \sqrt{c^2p_s^2 + m^2c^4} + V_g^s \tag{5}$$

where p_s = single particle momentum for a symmetrical system, c = light speed, m = proper mass, and V_g^s = scalar gauge potential for symmetrical matter. The gauge potential is zero when PV = αU where α = constant, and $U = U^a$.[6] When it has a non-zero value, the gauge potential breaks the Lorentz symmetry of the system.[6] The condition $V_g^s \neq 0$ is valid for the general case of a noninteracting

zero temperature Fermi gas (for which $PV \neq \alpha U$), except for the low density non-relativistic case and the high density ultra-relativistic case for which $PV = \alpha U$ and $V_g^s = 0$.[6] The single particle energy for the non-interacting case corresponding to equation (2) for broken symmetry matter is given by

$$\bar{\varepsilon}_i^{free} = \sqrt{c^2\bar{p}^2 + m^2c^4} + \bar{V}_g \tag{6}$$

$$\bar{V}_g = V_g e^{j\theta_{Vg}} \tag{7}$$

where $\bar{p}$ = complex number single particle momentum, and $\bar{V}_g$ = complex number gauge potential. In a similar fashion, $\bar{V}_g = 0$ when $\bar{P}\bar{V} = \alpha\bar{U}$. It is just the derivative terms in equations (1) and (2), required for gauge invariance, which produce the spacetime interaction gauge potentials that prevent the single particle energy and momentum from being four vectors in equations (5) and (6) when $PV \neq \alpha U$ and $\bar{P}\bar{V} \neq \alpha\bar{U}$ respectively. For the interacting case, the single particle energy is written as

$$\varepsilon_i^s = \sqrt{c^2p_s^2 + m^2c^4} + V_g^s + V_e^s \tag{8}$$

corresponding to equation (1) for a symmetrical system, and as

$$\bar{\varepsilon}_i = \sqrt{c^2\bar{p}^2 + m^2c^4} + \bar{V}_g + \bar{V}_e \tag{9}$$

corresponding to the relativistic trace equation (2) for a system with broken internal symmetry. For the broken symmetry case the external potential is written as

$$\bar{V}_e = V_e e^{j\theta_{Ve}} \tag{10}$$

In this paper the "s" refers to a symmetrical renormalized system.

The gauge potentials V_g or $\bar{V}_g$ are determined indirectly from the solution of the trace equations (1) or (2). Consider the trace equation (2). This equation is solved to determine the bulk matter internal energy $\bar{U}$ in terms of the unrenormalized internal energy U^a . The unrenormalized thermodynamic functions are determined from the unrenormalized partition function Z^a which is given by[8,9]

$$Z^a = \int \eta e^{-\beta H^a} dq_a\, dp_a \tag{11}$$

where η = degeneracy, $\beta = 1/(kT)$, q_a and p_a = conventional generalized coordinates and momenta respectively, and the unrenormalized Hamiltonian is given by

$$H^a(q_a,p_a) = \sqrt{c^2p_a^2 + m^2c^4} + V_e^a \tag{12}$$

where V_e^a = unrenormalized external potential, and[8,9]

$$U^a = -\left(\frac{\partial\ \ell n\ Z^a}{\partial\beta}\right)_V \tag{13}$$

$$P^a = \frac{1}{\beta}\left(\frac{\partial\ \ell n\ Z^a}{\partial V}\right)_\beta \tag{14}$$

The trace equation (2) is then used to determine the renormalized complex number internal energy and pressure $\bar{U}$ and $\bar{P}$ respectively. These values of $\bar{U}$ and $\bar{P}$ are then used to determine the complex number renormalized partition function $\bar{Z}$ from

$$\bar{U} = -\left(\frac{\partial\ \ell n\ \bar{Z}}{\partial\beta}\right)_V \tag{15}$$

$$\bar{P} = \frac{1}{\beta}\left(\frac{\partial\ \ell n\ \bar{Z}}{\partial V}\right)_\beta \tag{16}$$

where

$$\bar{Z} = \int \eta e^{-\beta\bar{H}} d\bar{q}\, d\bar{p} = Ze^{j\theta Z} \tag{17}$$

and $\bar{q}$ and $\bar{p}$ = complex number generalized coordinates and momenta respectively, and where the renormalized complex number Hamiltonian is given by

$$\bar{H}(\bar{q},\bar{p}) = \sqrt{c^2\bar{p}^2 + m^2c^4} + \bar{V}_g + \bar{V}_e \tag{18}$$

From equations (15), (16), and (17) it follows that

$$U\cos\theta_U = -\left(\frac{\partial\ \ell n\ Z}{\partial\beta}\right)_V \tag{19}$$

$$U\sin\theta_U = -\left(\frac{\partial\theta_Z}{\partial\beta}\right)_V \tag{20}$$

$$P\cos\theta_P = \frac{1}{\beta}\left(\frac{\partial\ \ell n\ Z}{\partial V}\right)_\beta \tag{21}$$

$$P\sin\theta_P = \frac{1}{\beta}\left(\frac{\partial\theta_Z}{\partial V}\right)_\beta \tag{22}$$

Therefore equations (19) through (22) can be used to determine Z and θ_Z for a relativistic system with broken internal symmetry. From a knowledge of $\bar{Z}$, equations (17) and (18) are then inverted to determine $\bar{V}_g$ and $\bar{V}_e$. In summary,

$$Z^a \rightarrow P^a, U^a \rightarrow \bar{P}, \bar{U} \rightarrow \bar{Z} \rightarrow \bar{V}_g, \bar{V}_e \tag{23}$$

Because $\bar{V}_g$ and $\bar{V}_e$ are complex numbers, it is expected that the coordinates of space and time must also be complex numbers. Note that it is the real parts of the complex number quantities such as coordinates, momentum, energy, pressure, frequency, angles and scattering cross sections that are the measured quantities.

Therefore, macroscopic local gauge invariance suggests the existence of a symmetry breaking microscopic gauge potential. Also, the macroscopic broken symmetry state required by equation (2) suggests that the space and time coordinates and the kinematic and dynamic quantities such as single particle velocity, acceleration, and force should be represented by complex numbers that include a description of internal phase angles. This paper indicates the effects of microscopic internal phase angles on the photon gas, and on such elementary atomic processes as the photoelectric effect, Compton effect, and Coulomb scattering. The forms of the Dirac and Schrödinger equations for particles in bulk matter or vacuum with broken symmetry are developed.

2. BROKEN SYMMETRY PHOTON GAS IN BULK MATTER. This section describes a photon gas with broken internal symmetry interacting with bulk matter that also has internal phase angles. The spectral energy density of a symmetrical photon gas is given by Planck's law as follows[10-12]

$$E_{\nu s} = \frac{A}{e^{h\nu/kT} - 1} = E^{(v)}_{\nu s} = E_{\nu a} = E^{(v)}_{\nu a} \tag{24}$$

where

$$A = \frac{8\pi h\nu^3}{c^3} \tag{25}$$

and where $E_{\nu s}$ = spectral energy density of radiation in symmetric matter, $E^{(v)}_{\nu s}$ = spectral energy density in the symmetric vacuum, h = Planck's constant, k = Boltzmann's constant, and T = absolute temperature. For this case the total energy density is given by the Stefan-Boltzmann law[10,11]

$$E_{rs} = \int_0^\infty E_{\nu s} d\nu = \sigma T^4 = E^{(v)}_{rs} = E_{ra} = E^{(v)}_{ra} \tag{26}$$

where E_{rs} and $E^{(v)}_{rs}$ = total energy density for a symmetrical photon gas in symmetric matter and the symmetric vacuum respectively, and σ = Stefan-Boltzmann constant. The pressure of the symmetrical photon gas is given for the symmetrical vacuum by

$$P_{\nu s}^{(v)} = \frac{1}{3} E_{\nu s}^{(v)} = P_{\nu a}^{(v)} = \frac{1}{3} E_{\nu a}^{(v)} \tag{27}$$

$$P_{rs}^{(v)} = \frac{1}{3} E_{rs}^{(v)} = \frac{1}{3} \sigma T^4 = \frac{1}{3} E_{ra}^{(v)} \tag{28}$$

where $P_{\nu s}^{(v)}$ and $P_{rs}^{(v)}$ = spectral and total radiation pressure respectively for the symmetrical photon gas in symmetric vacuum.

In order to write Planck's law for radiation in matter or the vacuum with broken internal symmetries, a complex number form of the radiation frequency is adopted and written as

$$\bar{\nu} = \nu e^{j\theta_\nu} = \nu(\cos \theta_\nu + j \sin \theta_\nu) \tag{29}$$

where $\bar{\nu}$ = complex number frequency, ν = magnitude of frequency, and θ_ν = frequency phase angle. For radiation in the asymmetrical vacuum, the radiation frequency is written as

$$\bar{\nu}^{(v)} = \nu e^{j\theta_\nu^{(v)}} \tag{30}$$

where $\theta_\nu^{(v)}$ = internal phase angle of the photon frequency in the asymmetric vacuum. Using equations (24) and (29), the complex number form of Planck's law is written as

$$\bar{E}_\nu = \frac{\bar{A}}{e^{h\bar{\nu}/kT} - 1} \tag{31}$$

where

$$\bar{A} = \frac{8\pi h \bar{\nu}^3}{c^3} \tag{32}$$

Placing equation (29) into equation (31) yields

$$\bar{E}_\nu = \frac{A}{D}(B + jC) = E_\nu e^{j\theta E\nu} \tag{33}$$

$$E_\nu = \frac{A}{D}\sqrt{B^2 + C^2} \tag{34}$$

$$\theta_{E\nu} = \tan^{-1}\left(\frac{C}{B}\right) \tag{35}$$

where A is given by equation (25) and where

$$D = e^{2x} - 2 \cos y\, e^{x} + 1 \tag{36}$$

$$B = \cos(3\theta_{\nu})\,(\cos y\, e^{x} - 1) + \sin(3\theta_{\nu}) \sin y\, e^{x} \tag{37}$$

$$C = \sin(3\theta_{\nu})\,(\cos y\, e^{x} - 1) - \cos(3\theta_{\nu}) \sin y\, e^{x} \tag{38}$$

$$x = \frac{h\nu}{kT} \cos\theta_{\nu} \tag{39}$$

$$y = \frac{h\nu}{kT} \sin\theta_{\nu} \tag{40}$$

The same expressions that are valid for radiation in matter with broken internal symmetry are also valid for radiation in the vacuum with broken internal symmetry if the substitution $\theta_{\nu} \rightarrow \theta_{\nu}^{(v)}$ is made in equations (36) through (40). This gives the photon energy density for radiation in the asymmetric vacuum as

$$\bar{E}_{\nu}^{(v)} = E_{\nu}^{(v)} e^{j\theta_{E\nu}^{(v)}} \tag{41}$$

where $E_{\nu}^{(v)}$ and $\theta_{E\nu}^{(v)}$ are obtained from equations (34) and (35) respectively.

Matter in general has a spectral index of refraction, but it does not enter into the photon spectral energy density as given by the Planck function.[14] This is because the Planck function is universal and does not include specific properties of matter.[14] The index of refraction does not enter the spectral energy calculation whether or not the radiation is symmetric or not. Later in this paper it will be shown that the index of refraction does enter the calculation of the total energy density of photons in matter due to an increased photon density in matter.

The spectral radiation pressure associated with the complex spectral radiation energy density given in equation (31) is obtained by a generalization of a radiation pressure formula given in the literature for mechanical radiation in matter as[15]

$$P_{\nu a} = \left(\frac{1}{3} + \frac{n}{W_{\nu}^{a}} \frac{dW_{\nu}^{a}}{dn}\right) E_{\nu a} = \left(\frac{1}{3} - \frac{n}{\mu_{\nu a}} \frac{d\mu_{\nu a}}{dn}\right) E_{\nu a} \tag{42}$$

where $P_{\nu a}$ = spectral radiation pressure, n = particle number density for matter, and W_ν^a = speed of waves of frequency ν. The generalization of equation (42) to the case of broken symmetry electromagnetic radiation is given by

$$\bar{P}_\nu = \left(\frac{1}{3} - \frac{n}{\bar{\mu}_\nu}\frac{d\bar{\mu}_\nu}{dn}\right)\bar{E}_\nu \tag{43}$$

where $\bar{P}_\nu$ = complex number spectral radiation pressure in matter, and $\bar{\mu}_\nu$ = complex number spectral index of refraction for electromagnetic waves in matter. For the vacuum with broken internal symmetry, equation (43) reduces to

$$\bar{P}_\nu^{(v)} = \frac{1}{3}\bar{E}_\nu^{(v)} \tag{44}$$

because $\bar{\mu}_\nu = 1$ for the vacuum. If the complex number spectral index of refraction is written as

$$\bar{\mu}_\nu = \mu_\nu e^{j\theta_{\mu\nu}} \tag{45}$$

then

$$\frac{n}{\bar{\mu}_\nu}\frac{d\bar{\mu}_\nu}{dn} = \frac{n}{\mu_\nu}\frac{d\mu_\nu}{dn} + jn\frac{d\theta_{\mu\nu}}{dn} = H_\nu e^{j\beta_{\mu\nu,n}}$$

and then equation (43) can be rewritten as

$$\bar{P}_\nu = E_\nu\left[\frac{1}{3}e^{j\theta_{E\nu}} - H_\nu e^{j(\theta_{E\nu}+\beta_{\mu\nu,n})}\right] \tag{47}$$

where

$$H_\nu = \sqrt{\left(\frac{n}{\mu_\nu}\frac{d\mu_\nu}{dn}\right)^2 + \left(n\frac{d\theta_{\mu\nu}}{dn}\right)^2} \tag{48}$$

and where

$$\tan\beta_{\mu\nu,n} = n\frac{d\theta_{\mu\nu}}{dn}\Big/\frac{n}{\mu_\nu}\frac{d\mu_\nu}{dn} \tag{49}$$

and finally where $\theta_{E\nu}$ is given by equation (35). For the vacuum $\mu_\nu = 1$ and $\theta_{\mu\nu} = 0$ so that

$$\bar{P}_{\nu}^{(v)} = \frac{1}{3} E_{\nu}^{(v)} e^{j\theta_{E\nu}^{(v)}} \tag{50}$$

where $\theta_{E\nu}^{(v)}$ is given by equation (35) with the substitution $\theta_{\nu} \rightarrow \theta_{\nu}^{(v)}$. For the symmetric vacuum $\theta_{\nu}^{(v)} = 0$ and $\theta_{E\nu}^{(v)} = 0$ so that

$$P_{\nu s}^{(v)} = \frac{1}{3} E_{\nu s}^{(v)} = \frac{1}{3} E_{\nu s} = P_{\nu a}^{(v)} \tag{51}$$

where $E_{\nu s} = E_{\nu s}^{(v)}$ for the symmetric vacuum or symmetric matter is given by equation (24). For symmetric radiation in symmetric matter equation (43) becomes

$$P_{\nu s} = \left(\frac{1}{3} - \frac{n}{\mu_{\nu s}} \frac{d\mu_{\nu s}}{dn} \right) E_{\nu s} \neq P_{\nu a} \tag{52}$$

where $P_{\nu s}$ = spectral radiation pressure for symmetric matter. Note that although $E_{\nu s}^{(v)} = E_{\nu s}$ one has $P_{\nu s} \neq P_{\nu s}^{(v)}$ on account of the spectral index of refraction that appears in equation (52).

Writing the complex spectral radiation pressure for asymmetric radiation as

$$\bar{P}_{\nu} = P_{\nu} e^{j\theta_{P\nu}} \tag{53}$$

and using equation (47) gives

$$P_{\nu} \cos \theta_{P\nu} = E_{\nu} \left[\frac{1}{3} \cos \theta_{E\nu} - H_{\nu} \cos (\theta_{E\nu} + \beta_{\mu\nu,n}) \right] \tag{54}$$

$$P_{\nu} \sin \theta_{P\nu} = E_{\nu} \left[\frac{1}{3} \sin \theta_{E\nu} - H_{\nu} \sin (\theta_{E\nu} + \beta_{\mu\nu,n}) \right] \tag{55}$$

Equations (54) and (55) give for the asymmetric vacuum

$$P_{\nu}^{(v)} \cos \theta_{P\nu}^{(v)} = \frac{1}{3} E_{\nu}^{(v)} \cos \theta_{E\nu}^{(v)} \tag{56}$$

$$P_{\nu}^{(v)} \sin \theta_{P\nu}^{(v)} = \frac{1}{3} E_{\nu}^{(v)} \sin \theta_{E\nu}^{(v)} \tag{57}$$

which can also be obtained directly from equation (50). From equations (56) and (57) it follows for the asymmetric vacuum that

$$P_{\nu}^{(v)} = \frac{1}{3} E_{\nu}^{(v)} \tag{58}$$

$$\theta_{P\nu}^{(v)} = \theta_{E\nu}^{(v)} \tag{59}$$

A comparison of equations (51) and (58) shows that this equation holds for both the symmetric and asymmetric vacuum. From equations (54) and (55) it follows that

$$\tan \theta_{P\nu} = \frac{\frac{1}{3} \sin \theta_{E\nu} - H_{\nu} \sin (\theta_{E\nu} + \beta_{\mu\nu,n})}{\frac{1}{3} \cos \theta_{E\nu} - H_{\nu} \cos (\theta_{E\nu} + \beta_{\mu\nu,n})} \tag{60}$$

and

$$P_{\nu} = E_{\nu} \sqrt{\frac{1}{9} + H_{\nu}^2 - \frac{2}{3} H_{\nu} \cos \beta_{\mu\nu,n}} \tag{61}$$

When $\beta_{\mu\nu,n} = 0$ and $\theta_{\mu\nu} = 0$ equation (61) reduces to the case of symmetric radiation in symmetric matter

$$P_{\nu s} = E_{\nu s}(\frac{1}{3} - H_{\nu s}) \tag{62}$$

where

$$H_{\nu s} = \frac{n}{\mu_{\nu s}} \frac{d\mu_{\nu s}}{dn} \tag{63}$$

for symmetric matter.

In order to determine the phase angles θ_{ν} and $\theta_{\nu}^{(v)}$ in terms of the pressure internal phase angle θ_P that is associated with the state equation of bulk matter, the mechanical equilibrium of matter and radiation with broken internal symmetries must be considered. Consider a piece of matter bathed in the surrounding radiation of a vacuum with broken internal symmetry. Were no radiation present, the matter would have a zero pressure $P = 0$ because the situation corresponds to a minimum value of the binding energy at the equilibrium density. When radiation is present inside the matter and outside in the vacuum, the density of matter shifts from its $P = 0$ equilibrium value to a new value for which $P \neq 0$. The new equilibrium density depends on the internal and external radiation densities (pressures) which in turn depends on the temperature and frequency for

monochromatic radiation, or solely on temperature for thermal radiation.[16] The relationship between the induced matter pressure and the radiation pressure of a monochromatic radiation field will now be developed for matter and radiation with broken internal symmetries.

For matter in mechanical equilibrium with internal and external monochromatic radiation fields, the equilibrium condition at the surface boundary is

$$\bar{P} = \bar{P}_{\nu}^{(v)} - \bar{P}_{\nu} \tag{64}$$

where $\bar{P}_{\nu}^{(v)}$ = spectral radiation pressure in vacuum, $\bar{P}_{\nu}$ = spectral radiation pressure in matter, and $\bar{P}$ = complex matter mechanical pressure induced by the radiation fields. In component form equation (64) can be rewritten as

$$P \cos \theta_P = P_{\nu}^{(v)} \cos \theta_{P\nu}^{(v)} - P_{\nu} \cos \theta_{P\nu} \tag{65}$$

$$P \sin \theta_P = P_{\nu}^{(v)} \sin \theta_{P\nu}^{(v)} - P_{\nu} \sin \theta_{P\nu} \tag{66}$$

Combining equations (65) and (66) with equations (54) through (57) gives

$$P \cos \theta_P = \frac{1}{3} E_{\nu}^{(v)} \cos \theta_{E\nu}^{(v)} - E_{\nu} \left[\frac{1}{3} \cos \theta_{E\nu} - H_{\nu} \cos (\theta_{E\nu} + \beta_{\mu\nu,n}) \right] \tag{67}$$

$$P \sin \theta_P = \frac{1}{3} E_{\nu}^{(v)} \sin \theta_{E\nu}^{(v)} - E_{\nu} \left[\frac{1}{3} \sin \theta_{E\nu} - H_{\nu} \sin (\theta_{E\nu} + \beta_{\mu\nu,n}) \right] \tag{68}$$

For the vacuum the following conditions have been used

$$\mu_{\nu}^{(v)} = 1 \qquad \theta_{\mu\nu}^{(v)} = 0 \qquad \beta_{\mu\nu,n}^{(v)} = 0 \qquad H_{\nu}^{(v)} = 0 \tag{69}$$

Note that for the asymmetric vacuum $\theta_{\nu}^{(v)} \neq 0$ just as for the case of radiation in matter one has $\theta_{\nu} \neq 0$. From equations (65) and (66) it follows that

$$\tan \theta_P = \frac{P_{\nu}^{(v)} \sin \theta_{P\nu}^{(v)} - P_{\nu} \sin \theta_{P\nu}}{P_{\nu}^{(v)} \cos \theta_{P\nu}^{(v)} - P_{\nu} \cos \theta_{P\nu}} \tag{70}$$

$$P^2 = [P_{\nu}^{(v)}]^2 + P_{\nu}^2 - 2P_{\nu}P_{\nu}^{(v)} \cos [\varphi_{P\nu}^{(v)} - \theta_{P\nu}] \tag{71}$$

For the case $\theta_{P\nu} = 0$ and $\theta_{P\nu}^{(v)} = 0$ equations (65) through (71) reduce to their

proper scalar forms, for instance equation (71) gives the induced pressure in symmetrical matter as

$$P_s = P_{\nu s}^{(v)} - P_{\nu s} = \frac{1}{3} E_{\nu s}^{(v)} - (\frac{1}{3} - H_{\nu s})E_{\nu s} \tag{72}$$

But for symmetrical matter and symmetrical vacuum $E_{\nu s}^{(v)} = E_{\nu s}$, so that equation (72) can be written as

$$P_s = H_{\nu s} E_{\nu s} \tag{73}$$

In general P and θ_P are functions of matter density, and equations (70) and (71) can be satisfied only if the equilibrium density of matter is altered by the radiation fields. Therefore equilibrium at a surface requires that the internal phase of matter and radiation are related by equations (70) and (71).

Now the total (integrated) energy density and associated pressure needs to be determined for radiation in matter and the vacuum with broken internal symmetries. The integrated radiation energy density is obtained from equations (31) through (40) by the following integral

$$\bar{E}_r = E_r e^{j\theta_{Er}} = \int \bar{E}_\nu \, d\bar{\nu} = \int_o^\infty E_\nu \sqrt{1 + (\nu \, d\theta_\nu/d\nu)^2} \; e^{j(\theta_{E\nu} + \theta_\nu + \beta_{\nu,\nu})} \, d\nu \tag{74}$$

where

$$\tan \beta_{\nu,\nu} = \nu \frac{d\theta_\nu}{d\nu} \tag{75}$$

and where the following result was used

$$d\bar{\nu} = e^{j\theta_\nu}(d\nu + j\nu d\theta_\nu) = e^{j(\theta_\nu + \beta_{\nu,\nu})} \sqrt{1 + (\nu \, d\theta_\nu/d\nu)^2} \, d\nu \tag{76}$$

Therefore the gauge rotated frequency must be used to evaluate the integrated radiation energy density. The radiation energy density has the following real and imaginary parts

$$E_r^R = E_r \cos \theta_{Er} = \int_o^\infty E_\nu \sqrt{1 + (\nu \, d\theta_\nu/d\nu)^2} \cos(\theta_{E\nu} + \theta_\nu + \beta_{\nu,\nu}) \, d\nu \tag{77}$$

$$E_r^I = E_r \sin \theta_{Er} = \int_o^\infty E_\nu \sqrt{1 + (\nu \, d\theta_\nu/d\nu)^2} \sin(\theta_{E\nu} + \theta_\nu + \beta_{\nu,\nu}) \, d\nu \tag{78}$$

The magnitude and phase angle of the radiation energy is given in terms of these integrals as follows

$$E_r = \sqrt{(E_r^R)^2 + (E_r^I)^2} \tag{79}$$

$$\tan \theta_{Er} = E_r^I / E_r^R \tag{80}$$

The evaluation of the integrals in equations (74), (77), and (78) is not simple due to the complicated form of the spectral energy density given by equations (34) and (35). It follows from equations (34), (35), (77), and (78) that the energy density for asymmetric radiation does not have the T^4 temperature behaviour that is valid for symmetric radiation according to equation (26). The radiation pressure is given by

$$\bar{P}_r = \frac{1}{3} \bar{E}_r + \Delta \bar{P}_r \tag{81}$$

where ΔP_r = small difference in radiation pressure due to internal phase angles, and considering only the complex number values of the Planck function. Material properties, such as the index of refraction, do not enter the Planck analysis.[17] Later in this paper it will be shown that the index of refraction enters the expressions for radiation energy density and pressure due to an increased photon density in matter. For the asymmetric vacuum the integrated radiation density is given by

$$\bar{E}_r^{(v)} = \int_0^\infty E_\nu^{(v)} \sqrt{1 + (\nu \, d\theta_\nu^{(v)}/d\nu)^2} \; e^{j[\theta_{E\nu}^{(v)} + \theta_\nu^{(v)} + \beta_{\nu,\nu}^{(v)}]} \, d\nu \tag{82}$$

$$= E_r^{(v)} e^{j\theta_{Er}^{(v)}} = E_r^{R(v)} + jE_r^{I(v)}$$

$$\tan \theta_{Er}^{(v)} = E_r^{I(v)} / E_r^{R(v)} \tag{83}$$

which is formally identical in structure to equations (74) through (80) for asymmetric radiation in matter. Asymmetric radiation in the vacuum also does not have a T^4 dependence. The radiation pressure for the asymmetric vacuum is given by

$$\bar{P}_r^{(v)} = \frac{1}{3} \bar{E}_r^{(v)} + \Delta \bar{P}_r^{(v)} \tag{84}$$

where $\Delta\bar{P}_r^{(v)}$ = small difference in vacuum radiation pressure due to internal phase angles.

Considering the effects of increased photon number density due to the index of refraction the energy density and pressure of symmetrical thermal radiation in symmetrical matter is given by[16]

$$E_{rMs} = \sigma\mu_s^3 T^4 \tag{85}$$

$$P_{rMs} = \sigma\mu_s^3 T^4\left(\frac{1}{3} - \frac{n}{\mu_s}\frac{d\mu_s}{dn}\right) \tag{86}$$

where E_{rMs} and P_{rMs} = measured radiation energy density and pressure for symmetrical matter and radiation, and where μ_s = density dependent index of refraction averaged over frequency. The term μ_s^3 arises from the general thermodynamic relations between pressure and energy denstiy.[16] A comparison of equations (26), (28), (85), and (86) shows that $E_{rMs} \neq E_{rs}$ and $P_{rMs} \neq P_{rs}$. The expressions in (26) and (28) are totally independent of any reference to material parameters (such as μ_s) and are the results of local thermodynamic equilibrium.[17] This is why the T^4 law is universal in the sense that it applies to all symmetric thermal radiation in symmetric matter or vacuum. The presence of the μ_s^3 term in equation (85) represents a diffusion effect where the photon number density is increased due to their slower speed in matter as compared to the vacuum.[16] The important point is that the μ_s^3 term (or any other dependence on material properties) does not originate from the Planck distribution. The subscript M (for measured value) is added to all expressions that include a μ_s^3 dependence.

In analogy to equations (85) and (86) for symmetric thermal radiation in symmetric matter, the measured radiation energy density and pressure for an asymmetric system is written as

$$\bar{E}_{rM} = \bar{W}(\bar{\mu})\bar{E}_r = E_{rM}e^{j\theta_{ErM}} \tag{87}$$

$$\bar{P}_{rM} = \bar{E}_{rM}\left(\frac{1}{3} - \frac{n}{\bar{\mu}}\frac{d\bar{\mu}}{dn}\right) = P_{rM}e^{j\theta_{PrM}} \tag{88}$$

where $\bar{E}_r$ = energy density for asymmetric radiation in matter as calculated from the complex number Planck function given in equation (74), $\bar{E}_{rM}$ = measured thermal radiation energy density in an asymmetric system, $\bar{\mu}$ = density dependent complex number index of refraction for asymmetric matter and averaged over frequency, and where $\bar{W}(\bar{\mu})$ = yet to be determined function of the complex number refraction index. The density dependent, frequency averaged, index of refraction for asymmetric matter is written as

$$\bar{\mu} = \mu e^{j\theta_\mu} \tag{89}$$

Note that equation (88) is already an approximation because the factor 1/3 holds only for symmetric radiation as shown in equation (81). For asymmetric matter and radiation the integrals in equations (77) and (78) have not been evaluated due to their complexity and so values of $\bar{E}_r$ in equation (87) have not been found, but it is clear from equation (74) that the leading term of $\bar{E}_r$ is the scalar σT^4 corresponding to symmetric radiation

$$\bar{E}_r = \sigma T^4 + \Delta\bar{E}_r \tag{90}$$

where $\Delta\bar{E}_r$ is small if θ_ν is small. From equations (81) and (90) it follows that

$$\bar{P}_r = \frac{1}{3}\sigma T^4 + \frac{\Delta\bar{E}_r}{3} + \Delta\bar{P}_r \tag{90A}$$

The determination of the exact value of $\bar{W}(\bar{\mu})$ is not possible since the value of $\bar{E}_r$ has not been determined. However, an approximate value of the factor $\bar{W}(\bar{\mu})$ can be determined by using the first order term in equation (90). This is done by first considering the Gibbs-Helmholtz relation (also called Maxwell's relation) applied to $\bar{E}_{rM}$ and $\bar{P}_{rM}$ as follows[16]

$$-n\,\frac{\partial\bar{E}_{rM}}{\partial n} + \bar{E}_{rM} = T\,\frac{\partial\bar{P}_{rM}}{\partial T} - \bar{P}_{rM} \tag{91}$$

Combining equations (87) and (88) with equation (91) yields

$$-n\,\frac{\partial}{\partial n}\,(\bar{W}\bar{E}_r) + \bar{W}\bar{E}_r = \left(\frac{1}{3} - \frac{n}{\bar{\mu}}\,\frac{d\bar{\mu}}{dn}\right)\left[T\,\frac{\partial}{\partial T}\,(\bar{W}\bar{E}_r) - \bar{W}\bar{E}_r\right] \tag{92}$$

Then assuming $\bar{W}$ is a function of density alone, and $\bar{E}_r$ is given by the scalar term in equation (90) so that $\bar{E}_r \sim \sigma T^4$, it follows from equation (92) that

$$-n\,\frac{d\bar{W}}{dn} + \bar{W} \sim 3\bar{W}\left(\frac{1}{3} - \frac{n}{\bar{\mu}}\,\frac{d\bar{\mu}}{dn}\right) \tag{93}$$

or

$$\frac{n}{\bar{W}}\,\frac{d\bar{W}}{dn} \sim 3\,\frac{n}{\bar{\mu}}\,\frac{d\bar{\mu}}{dn} \tag{94}$$

from which

$$\bar{W} \sim \bar{\mu}^3 \tag{95}$$

From equations (87), (88), and (95) the following approximations are obtained

$$\bar{E}_{rM} = \bar{\mu}^3 \bar{E}_r \tag{96}$$

$$\bar{P}_{rM} = \bar{\mu}^3 \bar{E}_r \left(\frac{1}{3} - \frac{n}{\bar{\mu}} \frac{d\bar{\mu}}{dn} \right) \tag{97}$$

From equations (87), (89), and (96) it follows that

$$E_{rM} = \mu^3 E_r \tag{98}$$

$$\theta_{ErM} = \theta_{Er} + 3\theta_{\mu} \tag{99}$$

where E_r and θ_{Er} are given by equations(79) and (80) respectively. All further analysis based on equations (96) and (97) is limited to the same approximations that went into the derivation of equations (96) and (97) namely, that all asymmetries are small.

The detailed calculation of the radiation pressure proceeds from equation (97). From equation (89) it follows that

$$\frac{n}{\bar{\mu}} \frac{d\bar{\mu}}{dn} = He^{j\beta_{\mu,n}} \tag{100}$$

where

$$H = \sqrt{\left(\frac{n}{\mu} \frac{d\mu}{dn} \right)^2 + \left(n \frac{d\theta_{\mu}}{dn} \right)^2} \tag{101}$$

$$\tan \beta_{\mu,n} = n \frac{d\theta_{\mu}}{dn} \Big/ \frac{n}{\mu} \frac{d\mu}{dn} \tag{102}$$

Then the measured thermal radiation pressure in asymmetric bulk matter is obtained from equations (97) through (99) as the following approximation

$$\bar{P}_{rM} = \mu^3 E_r \left[\frac{1}{3} e^{j\theta_{ErM}} - He^{j(\theta_{ErM} + \beta_{\mu,n})} \right] \tag{103}$$

The component forms of equation (103) are

$$P_{rM} \cos \theta_{PrM} = \mu^3 E_r \left[\frac{1}{3} \cos \theta_{ErM} - H \cos (\theta_{ErM} + \beta_{\mu,n}) \right] \tag{104}$$

$$P_{rM} \sin \theta_{PrM} = \mu^3 E_r \left[\frac{1}{3} \sin \theta_{ErM} - H \sin (\theta_{ErM} + \beta_{\mu,n}) \right] \tag{105}$$

From equations (104) and (105) it follows that

$$\tan \theta_{PrM} = \frac{\frac{1}{3} \sin \theta_{ErM} - H \sin (\theta_{ErM} + \beta_{\mu,n})}{\frac{1}{3} \cos \theta_{ErM} - H \cos (\theta_{ErM} + \beta_{\mu,n})} \tag{106}$$

$$P_{rM} = \mu^3 E_r \sqrt{\frac{1}{9} - \frac{2}{3} H \cos \beta_{\mu,n} + H^2} \tag{107}$$

For the case $\beta_{\mu,n} = 0$ and $\theta_{\mu} = 0$ for symmetric radiation, equation (107) becomes

$$P_{rMs} = \mu_s^3 E_{rs} (\frac{1}{3} - H_s) \tag{108}$$

which is just equation (86) because

$$H_s = \frac{n}{\mu_s} \frac{d\mu_s}{dn} \tag{109}$$

and E_{rs} is given by equation (26).

For the vacuum $\mu = 1$, $\theta_{\mu} = 0$, $\beta_{\mu,n} = 0$, and $H = 0$, so that from equations (98) and (99) it follows that

$$E_{rM}^{(v)} = E_r^{(v)} \tag{110}$$

$$\theta_{ErM}^{(v)} = \theta_{Er}^{(v)} \tag{111}$$

For radiation in the asymmetric vacuum it follows from equations (104) and (105) that the following approximate equations are valid

$$P_r^{(v)} \cos \theta_{Pr}^{(v)} = \frac{1}{3} E_r^{(v)} \cos \theta_{Er}^{(v)} \tag{112}$$

$$P_r^{(v)} \sin \theta_{Pr}^{(v)} = \frac{1}{3} E_r^{(v)} \sin \theta_{Er}^{(v)} \tag{113}$$

For the vacuum it follows from equations (112) and (113) that approximately

$$\theta_{Pr}^{(v)} = \theta_{Er}^{(v)} \tag{114}$$

$$P_r^{(v)} = \frac{1}{3} E_r^{(v)} \tag{115}$$

where $\theta_{Er}^{(v)}$ and $E_r^{(v)}$ are obtained from the evaluation of the integral in equation (82).

Consider now the equilibrium equations at the surface of asymmetric matter that is bathed in asymmetric thermal radiation of the vacuum. The condition for mechanical equilibrium at the surface of the body is that the induced mechanical pressure is

$$\bar{P} = \bar{P}_r^{(v)} - \bar{P}_{rM} \tag{116}$$

or equivalently

$$P \cos \theta_P = P_r^{(v)} \cos \theta_{Pr}^{(v)} - P_{rM} \cos \theta_{PrM} \tag{117}$$

$$P \sin \theta_P = P_r^{(v)} \sin \theta_{Pr}^{(v)} - P_{rM} \sin \theta_{PrM} \tag{118}$$

Combining equations (117), (118) and equations (104), (105), (112), and (113) gives the following approximations

$$P \cos \theta_P = \frac{1}{3} E_r^{(v)} \cos \theta_{Er}^{(v)} - \mu^3 \bar{E}_r \left[\frac{1}{3} \cos \theta_{ErM} - H \cos (\theta_{ErM} + \beta_{\mu,n}) \right] \tag{119}$$

$$P \sin \theta_P = \frac{1}{3} E_r^{(v)} \sin \theta_{Er}^{(v)} - \mu^3 \bar{E}_r \left[\frac{1}{3} \sin \theta_{ErM} - H \sin (\theta_{ErM} + \beta_{\mu,n}) \right] \tag{120}$$

From equations (117) and (118) it follows that

$$P^2 = [P_r^{(v)}]^2 + P_{rM}^2 - 2P_r^{(v)} P_{rM} \cos [\theta_{Pr}^{(v)} - \theta_{PrM}] \tag{121}$$

$$\tan \theta_P = \frac{P_r^{(v)} \sin \theta_{Pr}^{(v)} - P_{rM} \sin \theta_{PrM}}{P_r^{(v)} \cos \theta_{Pr}^{(v)} - P_{rM} \cos \theta_{PrM}} \tag{122}$$

For the case of symmetric matter and symmetric radiation in matter and the vacuum, equation (121) becomes[16]

$$P_s = P_{rs}^{(v)} - P_{rMs} \tag{123}$$

$$= P_{rs}^{(v)} - \mu_s^3 E_{rs} \left(\frac{1}{3} - H_s \right)$$

$$= P_{rs}^{(v)} [1 - 3\mu_s^3 \left(\frac{1}{3} - H_s \right)] = P_{rs}^{(v)} (1 - \mu_s^3 + 3\mu_s^3 H_s)$$

where $P_{rs}^{(v)}$ is given by equation (28). Equation (123) can also be obtained diectly from equation (119).

The functions θ_ν and E_ν have not been obtained explicitly, and therefore the radiation energy E_r is also unknown. For the purposes of the rest of this paper it is sufficient to understand that the frequency of photons in asymmetric bulk matter or vacuum has an internal phase angle that will manifest itself in the interactions of photons with other atomic particles.

3. BROKEN SYMMETRY OF ANGLES IN ASYMMETRIC BULK MATTER AND VACUUM. Within asymmetric bulk matter or the asymmetric vacuum, the internal phase angles of the coordinates produce a broken symmetry in various geometrical quantities such as, for example, angles. These broken symmetry angles enter the basic calculations of atomic processes that are treated in this paper. Consider first the fact that angles have internal phases, a result which can be deduced from the law of cosines which, for the complex number lengths that appear in asymmetric bulk matter or vacuum, can be written as

$$\cos \bar{\phi} = \frac{\bar{a}^2 + \bar{b}^2 - \bar{c}^2}{2\bar{a}\bar{b}} \tag{124}$$

where $\bar{a}$, $\bar{b}$, and $\bar{c}$ are the sides of a plane triangle and $\bar{\phi}$ is the angle opposite side $\bar{c}$. Therefore it is clear that $\bar{\phi}$ and $\cos \bar{\phi}$ are complex numbers and can be written as

$$\bar{\phi} = \phi e^{j\theta_\phi} \tag{125}$$

$$\cos \bar{\phi} = C_\phi e^{-j\theta_{c\phi}}$$

where C_ϕ = magnitude of $\cos \bar{\phi}$, and $\theta_{c\phi}$ = phase angle associated with $\cos \bar{\phi}$. It also follows that

$$\sin \bar{\phi} = S_\phi e^{j\theta_{s\phi}} \tag{127}$$

where S_ϕ = magnitude of $\sin \bar{\phi}$, and $\theta_{s\phi}$ = phase angle of $\sin \bar{\phi}$. From the following expression

$$\cos \bar{\phi} = \frac{1}{2} (e^{j\bar{\phi}} + e^{-j\bar{\phi}}) \tag{128}$$

it follows by elementary algebra that

$$\cos \bar{\phi} = \cos \phi_R \cosh \phi_I - j \sin \phi_R \sinh \phi_I \tag{129}$$

where

$$\bar{\phi} = \phi_R + j\phi_I = \phi(\cos\theta_\phi + j \sin\theta_\phi) \tag{130}$$

Then the combining equations (126), (129), and (130) gives

$$C_\phi = \sqrt{\cos^2(\phi \cos\theta_\phi) + \sinh^2(\phi \sin\theta_\phi)} \tag{131}$$

$$\tan\theta_{c\phi} = \tan(\phi \cos\theta_\phi) \tanh(\phi \sin\theta_\phi) \tag{132}$$

In a similar fashion from

$$\sin\bar{\phi} = \frac{1}{2j}(e^{j\bar{\phi}} - e^{-j\bar{\phi}}) \tag{133}$$

it follows that

$$\sin\bar{\phi} = \sin\phi_R \cosh\phi_I + j \cos\phi_R \sinh\phi_I \tag{134}$$

and combining equations (127), (130), and (134) gives

$$S_\phi = \sqrt{\sin^2(\phi \cos\theta_\phi) + \sinh^2(\phi \sin\theta_\phi)} \tag{135}$$

$$\tan\theta_{s\phi} = \cot(\phi \cos\theta_\phi) \tanh(\phi \sin\theta_\phi) \tag{136}$$

These results will be used in Sections 5 and 6 where particle scattering in asymmetric bulk matter or vacuum is considered. The measured angle is given by $\phi_m = \phi \cos\theta_\phi = \phi_a$ where ϕ_a = conventional angle between two lines.

4. PHOTOELECTRIC EFFECT IN ASYMMETRIC BULK MATTER OR VACUUM. A very simple atomic process is the photoelectric effect wherein a photon collides with an electron that is bound in matter. If the photon has sufficient energy it will overcome the binding energy of the electron, and the electron will leave its site in the matter lattice with an excess kinetic energy.[17,18] The description of the process that occurs in bulk matter or vacuum with broken symmetries is similar to that of the standard analysis for the case where the electrons and photons move in symmetrical bulk matter or vacuum, except that now the kinematic variables for the photons and electrons become complex numbers. This is related to the broken symmetry of space and time in bulk matter and the vacuum.

Within asymmetric bulk matter the binding energy of an electron is described by a complex number potential $\bar{W}$, so that the binding energy is $e\bar{W}$, where e = electric charge. The conservation of energy then requires[17]

$$\frac{1}{2} m\bar{v}^2 = h\bar{\nu} - e\bar{W} \tag{137}$$

where m = electron mass, $\bar{v}$ = complex number electron velocity, and as before $\bar{\nu}$ = complex number frequency of the photon. Within asymmetric bulk matter or vacuum the electron velocity has a broken internal symmetry and is written as

$$\bar{v} = ve^{j\theta_v} \tag{138}$$

where v and θ_v = magnitude and internal phase angle respectively of the electron velocity. The complex number binding potential is written as

$$\bar{W} = We^{j\theta_W} \tag{139}$$

where W and θ_W = magnitude and internal phase angle of the binding potential. As described in Section 2 the photon frequency is a complex number for a photon propagating in asymmetric bulk matter or vacuum, and is written as in equation (29). It is assumed that ν, θ_ν, W, and θ_W are known quantities and that equation (137) can be used to determine the unknown complex number speed of the ejected electron.

The two scalar equations equivalent to equation (137) are

$$\frac{1}{2} mv^2 \cos(2\theta_v) = h\nu \cos\theta_\nu - eW \cos\theta_W \tag{140}$$

$$\frac{1}{2} mv^2 \sin(2\theta_v) = h\nu \sin\theta_\nu - eW \sin\theta_W \tag{141}$$

These two equations can be used to determine the unknown quantities v and θ_v as follows

$$\tan(2\theta_v) = \frac{h\nu \sin\theta_\nu - eW \sin\theta_W}{h\nu \cos\theta_\nu - eW \cos\theta_W} \tag{142}$$

$$\frac{1}{4} m^2 v^4 = h^2\nu^2 + e^2W^2 - 2h\nu eW \cos(\theta_\nu - \theta_W) \tag{143}$$

A plot of the kinetic energy of the electron versus frequency is shown in Figure 1, while a plot of the internal phase angle of the electron kinetic energy $2\theta_v$ versus frequency is shown in Figure 2. These two figures show that there is a discontinuity in the kinetic energy magnitude and phase angle at a threshold frequency which is obtained from equation (140) by taking $2\theta_v = \pi/2$ or

$$h\nu_t \cos\theta_{\nu t} = eW \cos\theta_W \tag{144}$$

where ν_t = threshold frequency, and $\theta_{\nu t}$ = internal phase angle of the frequency at the threshold frequency. The electron kinetic energy at the threshold frequency is obtained from equation (141) to be

$$\frac{1}{2} mv_t^2 = h\nu_t \sin \theta_{\nu t} - eW \sin \theta_W \tag{145}$$

$$= eW (\cos \theta_W \tan \theta_{\nu t} - \sin \theta_W)$$

The threshold kinetic energy given by equation (145) is the minimum kinetic energy that the ejected electron can have in asymmetric bulk matter or vacuum. Below the threshold frequency the photoelectric process will not occur. If all phase angles are set equal to zero, the standard results are regained that the threshold frequency is given by $h\nu_t = eW$, and the minimum kinetic energy of the ejected electron is zero. Note that the measured electron kinetic energy is given by equation (140) which is linear in the photon frequency ν. The measured frequency is equal to $\nu \cos \theta_\nu$.

5. THOMSON SCATTERING AND THE COMPTON EFFECT IN ASYMMETRIC BULK MATTER AND VACUUM. The elastic scattering of photons by electrons is called Thomson scattering. For this case the photon energy is much smaller than the electron mass energy mc^2 and, if the electron is bound in an atom, the photon energy is larger than the binding energy so that the electrons can be considered to be free. For this case the differential cross section is given by[13]

$$I^a(\phi_a) = \frac{r_o^2}{2}(1 + \cos^2 \phi_a) \tag{146}$$

where r_o = classical electron radius, and where ϕ_a = conventional scattering angle. The corresponding differential cross section for Thomson scattering of photons by electrons in asymmetric bulk matter or vacuum is given by

$$\bar{I}(\bar{\phi}) = Ie^{j\theta_I} = \frac{r_o^2}{2}(1 + \cos^2 \bar{\phi}) \tag{147}$$

Combining equations (126) and (147) gives

$$I \cos \theta_I = \frac{r_o^2}{2}[1 + C_\phi^2 \cos (2\theta_{c\phi})] \tag{148}$$

$$I \sin \theta_I = -\frac{r_o^2}{2} C_\phi^2 \sin (2\theta_{c\phi}) \tag{149}$$

or

$$\tan \theta_I = -\frac{C_\phi^2 \sin (2\theta_{c\phi})}{1 + C_\phi^2 \cos (2\theta_{c\phi})} \tag{150}$$

$$I^2 = \frac{r_o^4}{4}[1 + C_\phi^4 + 2C_\phi^2 \cos (2\theta_{c\phi})] \tag{151}$$

The Compton effect is the name associated with the quantum scattering of photons by electrons with a transfer of momentum and energy from the photons to the electrons.[11,17] The description of this process using quanta of light was one of the early successes of quantum theory. This process is conventionally described by assuming that the photon and electron propagate in the symmetric vacuum, and applying the laws of conservation of energy and momentum to the colliding particles. When this process occurs within asymmetric bulk matter or vacuum, the same conservation laws are expected to be valid except now the kinematical parameters of the photon and the electron have broken symmetries and are represented by complex numbers.

Within bulk matter or vacuum with broken internal symmetries, a photon of initial frequency $\bar{\nu}$ collides with a stationary electron, and a new photon of frequency $\bar{\nu}'$ is emitted at an angle $\bar{\phi}$ with respect to the initial photon direction, and the electron recoils with speed $\bar{v}$ in a direction $\bar{\psi}$ with respect to the initial photon direction. Then the conservation of energy for the nonrelativistic case gives[11]

$$h\bar{\nu} = \frac{1}{2} m\bar{v}^2 + h\bar{\nu}' \tag{152}$$

while the conservation of momentum yields two equations[11]

$$\frac{h\bar{\nu}}{c} = \frac{h\bar{\nu}'}{c} \cos \bar{\phi} + m\bar{v} \cos \bar{\psi} \tag{153}$$

$$\frac{h\bar{\nu}'}{c} \sin \bar{\phi} = m\bar{v} \sin \bar{\psi} \tag{154}$$

These equations can be used to determine the three unknown complex number quantities $\bar{\nu}'$, $\bar{v}$, and $\bar{\phi}$ in terms of the known quantities $\bar{\nu}$ and $\bar{\psi}$. These three conservation equations are expressed in terms of complex numbers and are therefore equivalent to six scalar equations. The two components of the nonrelativistic energy conservation equation (152) are

$$h\nu \cos \theta_\nu = \frac{1}{2} mv^2 \cos (2\theta_v) + h\nu' \cos \theta'_\nu \tag{155}$$

$$h\nu \sin \theta_\nu = \frac{1}{2} mv^2 \sin (2\theta_v) + h\nu' \sin \theta'_\nu \tag{156}$$

The four momentum conservation equations obtained from equations (153) and (154) are respectively

$$\frac{h\nu}{c} \cos \theta_\nu = \frac{h\nu'}{c} C_\phi \cos (\theta'_\nu - \theta_{c\phi}) + mvC_\psi \cos (\theta_v - \theta_{c\psi}) \tag{157}$$

$$\frac{h\nu}{c} \sin \theta_\nu = \frac{h\nu'}{c} C_\phi \sin (\theta'_\nu - \theta_{c\phi}) + mvC_\psi \sin (\theta_v - \theta_{c\psi}) \tag{158}$$

$$\frac{h\nu'}{c} S_\phi = mvS_\psi \tag{159}$$

$$\theta'_\nu + \theta_{s\phi} = \theta_v + \theta_{s\psi} \tag{160}$$

where C_ϕ and S_ϕ are given by equations (131) and (135) respectively, and $\theta_{c\phi}$ and $\theta_{s\phi}$ are given by equations (132) and (136) respectively, and similarly

$$C_\psi = \sqrt{\cos^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)} \tag{161}$$

$$S_\psi = \sqrt{\sin^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)} \tag{162}$$

$$\tan \theta_{c\psi} = \tan(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \tag{163}$$

$$\tan \theta_{s\psi} = \cot(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \tag{164}$$

where

$$\bar{\psi} = \psi e^{j\theta\psi} = \psi_R + j\psi_I \tag{165}$$

$$\cos \bar{\psi} = C_\psi e^{-j\theta c\psi} \tag{166}$$

$$\sin \bar{\psi} = S_\psi e^{j\theta s\psi} \tag{167}$$

The six equations (155) through (160) can be solved simultaneously for the six unknowns ν', θ'_ν, v, θ_v, ϕ, and θ_ϕ in terms of the four known quantities ν, θ_ν, ψ, and θ_ψ.

For a bulk matter system or vacuum with broken internal symmetries, the relativistic analogs of the energy and momentum conservation equations (152) through (154) are[12,17]

$$h\bar{\nu} = mc^2(\bar{\gamma} - 1) + h\bar{\nu}' \tag{168}$$

$$\frac{h\bar{\nu}}{c} = \frac{h\bar{\nu}'}{c} \cos \bar{\phi} + m\bar{\gamma}\bar{v} \cos \bar{\psi} \tag{169}$$

$$\frac{h\bar{\nu}'}{c} \sin \bar{\phi} = m\bar{\gamma}\bar{v} \sin \bar{\psi} \tag{170}$$

where the complex number velocity factor for a particle with a velocity that has a broken symmetry is

$$\bar{\gamma} = \gamma e^{j\theta_\gamma} = (1 - \bar{v}^2/c^2)^{-1/2} \tag{171}$$

and where the magnitude and internal phase angle of the complex number boost is

$$\gamma = (f^2 + b^2)^{-1/4} \tag{172}$$

$$\tan(2\theta_\gamma) = b/f \tag{173}$$

where

$$b = v^2/c^2 \sin(2\theta_v) \tag{174}$$

$$f = 1 - v^2/c^2 \cos(2\theta_v) \tag{175}$$

The six scalar component equations corresponding to equations (168) through (170) are

$$h\nu \cos\theta_\nu = mc^2(\gamma \cos\theta_\gamma - 1) + h\nu' \cos\theta'_\nu \tag{176}$$

$$h\nu \sin\theta_\nu = mc^2\gamma \sin\theta_\gamma + h\nu' \sin\theta'_\nu \tag{177}$$

$$\frac{h\nu}{c} \cos\theta_\nu = \frac{h\nu'}{c} C_\phi \cos(\theta'_\nu - \theta_{c\phi}) + m\gamma v C_\psi \cos(\theta_\gamma + \theta_v - \theta_{c\psi}) \tag{178}$$

$$\frac{h\nu}{c} \sin\theta_\nu = \frac{h\nu'}{c} C_\phi \sin(\theta'_\nu - \theta_{c\phi}) + m\gamma v C_\psi \sin(\theta_\gamma + \theta_v - \theta_{c\psi}) \tag{179}$$

$$\frac{h\nu'}{c} S_\phi = m\gamma v S_\psi \tag{180}$$

$$\theta'_\nu + \theta_{s\phi} = \theta_\gamma + \theta_v + \theta_{s\psi} \tag{181}$$

In the limit $v/c \to 0$ equations (176) through (184) reduce to equations (155) through (160) by noting that

$$\gamma \rightarrow 1 + \frac{1}{2} v^2/c^2 \cos(2\theta_v) \rightarrow 1 \tag{182}$$

$$\theta_\gamma \rightarrow \frac{1}{2} v^2/c^2 \sin(2\theta_v) \rightarrow 0 \tag{183}$$

The six equations (155) through (160) or (176) through (181) can be solved numerically using Brown's algorithm for the solution of simultaneous nonlinear equations. This algorithm is a modification of Newton's method and requires no derivative evaluations.[19]

The solution of equations (168) through (170) can be obtained by direct analogy to the solution for the standard Compton effect as follows[17]

$$\bar{\lambda}' - \bar{\lambda} = \lambda_o(1 - \cos\bar{\phi}) \tag{184}$$

where

$$\bar{\lambda}' = \lambda' e^{j\theta'_\lambda} = c/\bar{\nu}' = c/\nu' e^{-j\theta'_\nu} \tag{185}$$

$$\bar{\lambda} = \lambda e^{j\theta_\lambda} = c/\bar{\nu} = c/\nu e^{-j\theta_\nu} \tag{186}$$

and where λ_o = Compton wavelength = $h/(mc)$. The scalar equivalents for equation (184) are

$$\lambda' \cos\theta'_\lambda = \lambda \cos\theta_\lambda + \lambda_o(1 - C_\phi \cos\theta_{c\phi}) \tag{187}$$

$$\lambda' \sin\theta'_\lambda = \lambda \sin\theta_\lambda + \lambda_o C_\phi \sin\theta_{c\phi} \tag{188}$$

From equations (187) and (188) it follows that

$$\tan\theta'_\lambda = \frac{\lambda \sin\theta_\lambda + \lambda_o C_\phi \sin\theta_{c\phi}}{\lambda \cos\theta_\lambda + \lambda_o(1 - C_\phi \cos\theta_{c\phi})} \tag{189}$$

$$(\lambda')^2 = \lambda^2 + 2\lambda\lambda_o[\cos\theta_\lambda - C_\phi \cos(\theta_\lambda + \theta_{c\phi})] + \lambda_o^2(1 - 2C_\phi \cos\theta_{c\phi} + C_\phi^2) \tag{190}$$

Equations (189) and (190) give the wavelength internal phase angle and wavelength magnitude respectively of the scattered photon in asymmetric bulk matter or vacuum. The corresponding frequency equations can be obtained from equations (189) and (190) by noting that $\lambda' = c/\nu'$, $\lambda = c/\nu$, $\theta'_\lambda = -\theta'_\nu$, and $\theta_\lambda = -\theta_\nu$. Note that equation (187) gives the change in measured wavelengths, and this wavelength difference is independent of the wavelength itself.

Consider now the differential cross section for Compton scattering in asymmetric bulk matter or vacuum. The standard Compton scattering differential cross section is given by the Klein-Nishina formula[13,20]

$$I^a(\phi_a) = \frac{r_o^2}{2}\left(\frac{\nu_a'}{\nu_a}\right)^2\left(\frac{\nu_a}{\nu_a'} + \frac{\nu_a'}{\nu_a} - \sin^2\phi_a\right) \tag{191}$$

where ν_a = conventionally determined initial photon frequency, and ν_a' = conventionally determined scattered photon frequency. The generalization to the differential scattering cross section for Compton scattering within bulk matter or the vacuum with broken internal symmetries follows from equation (191) as

$$\bar{I}(\bar{\phi}) = \frac{r_o^2}{2}\left(\frac{\bar{\nu}'}{\bar{\nu}}\right)^2\left(\frac{\bar{\nu}}{\bar{\nu}'} + \frac{\bar{\nu}'}{\bar{\nu}} - \sin^2\bar{\phi}\right) \tag{192}$$

or equivalently as

$$\bar{I}(\bar{\phi}) = \frac{r_o^2}{2}\left(\frac{\nu'}{\nu}\right)^2\left(\frac{\nu}{\nu'}e^{j\Gamma_1} + \frac{\nu'}{\nu}e^{j\Gamma_2} - S_\phi^2 e^{j\Gamma_3}\right) \tag{193}$$

where ν and ν' = magnitudes of the complex number initial and scattered photon frequencies respectively, and where

$$\Gamma_1 = \theta_\nu' - \theta_\nu \tag{194}$$

$$\Gamma_2 = 3(\theta_\nu' - \theta_\nu) \tag{195}$$

$$\Gamma_3 = 2(\theta_\nu' - \theta_\nu + \theta_{s\phi}) \tag{196}$$

Therefore from equation (193) it follows that

$$I\cos\theta_I = \frac{r_o^2}{2}\left(\frac{\nu'}{\nu}\right)^2\left(\frac{\nu}{\nu'}\cos\Gamma_1 + \frac{\nu'}{\nu}\cos\Gamma_2 - S_\phi^2\cos\Gamma_3\right) \tag{197}$$

$$I\sin\theta_I = \frac{r_o^2}{2}\left(\frac{\nu'}{\nu}\right)^2\left(\frac{\nu}{\nu'}\sin\Gamma_1 + \frac{\nu'}{\nu}\sin\Gamma_2 - S_\phi^2\sin\Gamma_3\right) \tag{198}$$

from which I and θ_I can easily be obtained. The measured differential cross section = $I\cos\theta_I$.

6. COULOMB SCATTERING IN BULK MATTER AND THE VACUUM WITH BROKEN INTERNAL SYMMETRIES. This section considers Rutherford, Mott, Bhabha, and Møller scat-

tering in asymmetric bulk matter and vacuum.

A. Rutherford Scattering

The α-particle scattering experiments of Rutherford are one of the cornerstones of knowledge about atomic structure. These experiments measured the scattering angles of α-particles interacting with atomic nuclei of charge Ze. The basic formulas for Rutherford scattering give the differential scattering cross section as[17]

$$I^a(\phi_a) = \frac{A}{v_a^4} \csc^4 \frac{\phi_a}{2} \tag{199}$$

where

$$A = \left(\frac{Z'Ze^2}{2m}\right)^2 \tag{200}$$

ϕ_a = measured scattering angle, v_a = conventionally calculated initial relative speed of the α-particle and the atomic nucleus, Z' = atomic number of incident particle (Z' = 2 for α-particle), Z = atomic number of the atomic nucleus, and m = reduced mass of the incident particle and the atomic nucleus. In addition to the differential cross section, the other quantity that is often calculated is the number of particles deviated through and angle between ϕ_a and $\phi_a + d\phi_a$ which is given by[17]

$$\frac{dN^a}{d\phi_a} = \frac{4\pi A}{v_a^4} \cot \frac{\phi_a}{2} \csc^2 \frac{\phi_a}{2} \tag{201}$$

These formulas were deduced by considering the scattering of an α-particle by an isolated atomic nucleus situated in a symmetrical vacuum.

For Rutherford scattering within asymmetric bulk matter or vacuum equations (199) and (201) need to be modified because the indicent α-particle speed $\bar{v}$ is now a complex number, and because the deflection angle $\bar{\phi}$ is also a complex number. Therefore equations (199) and (201) must now be written as

$$\bar{I}(\bar{\phi}) = \frac{A}{\bar{v}^4} \csc^4 \frac{\bar{\phi}}{2} \tag{202}$$

$$\frac{d\bar{N}}{d\bar{\phi}} = \frac{4\pi A}{\bar{v}^4} \cot \frac{\bar{\phi}}{2} \csc^2 \frac{\bar{\phi}}{2} \tag{203}$$

where $\bar{I}(\bar{\phi})$ = complex number differential scattering cross section, $\bar{\phi}$ = complex number deflection angle, $d\bar{N}/d\bar{\phi}$ = complex number of particles deviated through $\bar{\phi}$ and $\bar{\phi} + d\bar{\phi}$, and $\bar{v}$ = complex number initial α-particle speed. Because $\bar{v}$ and $\bar{\phi}$ are phase rotated, the number of α-particles scattered will also include a phase

rotated part, so that

$$\bar{N} = Ne^{j\theta_N} \tag{204}$$

where N and θ_N = magnitude and internal phase angle respectively of the number of scattered particles.

Using the following standard trigonometric formulas

$$\sin^2 \frac{\bar{\phi}}{2} = \frac{1}{2}(1 - \cos \bar{\phi}) \tag{205}$$

$$\tan \frac{\bar{\phi}}{2} = \frac{\sin \bar{\phi}}{1 + \cos \bar{\phi}} \tag{206}$$

$$\cos^2 \frac{\bar{\phi}}{2} = \frac{1}{2}(1 + \cos \bar{\phi}) \tag{207}$$

and combining them with equations (126) and (127) gives

$$\csc^2 \frac{\bar{\phi}}{2} = K_s e^{-jx_\phi} \tag{208}$$

$$\csc^4 \frac{\bar{\phi}}{2} = K_s^2 e^{-2jx_\phi} \tag{209}$$

$$\cot \frac{\bar{\phi}}{2} = K_t e^{-jz_\phi} \tag{211}$$

$$\sec^2 \frac{\bar{\phi}}{2} = K_{se} e^{jy_\phi} \tag{212}$$

where

$$K_s = 2(1 - 2C_\phi \cos \theta_{c\phi} + C_\phi^2)^{-1/2} \tag{213}$$

$$K_t = \frac{1}{S_\phi}(1 + 2C_\phi \cos \theta_{c\phi} + C_\phi^2)^{1/2} \tag{214}$$

$$K_{se} = 2(1 + 2C_\phi \cos \theta_{c\phi} + C_\phi^2)^{-1/2} \tag{215}$$

$$\tan x_\phi = \frac{C_\phi \sin \theta_{c\phi}}{1 - C_\phi \cos \theta_{c\phi}} \tag{216}$$

$$\tan z_\phi = \frac{\sin \theta_{s\phi} + C_\phi \sin (\theta_{c\phi} + \theta_{s\phi})}{\cos \theta_{s\phi} + C_\phi \cos (\theta_{c\phi} + \theta_{s\phi})} \tag{217}$$

$$\tan y_\phi = \frac{C_\phi \sin \theta_{c\phi}}{1 + C_\phi \cos \theta_{c\phi}} \tag{218}$$

where C_ϕ, S_ϕ, $\theta_{c\phi}$, and $\theta_{s\phi}$ are given by equations (131), (135), (132), and (136) respectively.

Combining equations (208) and (209) with equation (202) gives

$$\bar{I}(\bar{\phi}) = Ie^{j\theta_I} = \frac{AK_s^2}{v^4} e^{-j(4\theta_v + 2x_\phi)} \tag{219}$$

or

$$I = \frac{AK_s^2}{v^4} \tag{220}$$

$$\theta_I = -4\theta_v - 2x_\phi \tag{221}$$

which are the equations for the magnitude and internal phase of the complex number differential cross section for Rutherford scattering in asymmetric matter or vacuum. Combining equations (208) and (211) with equation (203) gives

$$\frac{d\bar{N}}{d\bar{\phi}} = \left|\frac{d\bar{N}}{d\bar{\phi}}\right| e^{j\theta_{N\phi}} = \frac{4\pi AK_sK_t}{v^4} e^{-j(4\theta_v + x_\phi + z_\phi)} \tag{222}$$

and therefore

$$\left|\frac{d\bar{N}}{d\bar{\phi}}\right| = \sqrt{\frac{\left(\frac{dN}{d\phi}\right)^2 + N^2\left(\frac{d\theta_N}{d\phi}\right)^2}{1 + \phi^2\left(\frac{d\theta_\phi}{d\phi}\right)^2}} = 4\pi AK_sK_t/v^4 \tag{223}$$

$$\theta_{N\phi} = \theta_N + \beta_{N,\phi} - \theta_\phi - \beta_{\phi,\phi} = -4\theta_v - x_\phi - z_\phi \tag{224}$$

where

$$\tan \beta_{N,\phi} = N \frac{d\theta_N/d\phi}{dN/d\phi} \tag{225}$$

$$\tan \beta_{\phi,\phi} = \phi \frac{d\theta_\phi}{d\phi} \tag{226}$$

which gives the magnitude and phase angle of the number of scattered particles. The measured scattering cross section is given by I cos θ_I .

B. Mott Scattering

Mott scattering describes the Coulomb scattering of two identical fermions such as, for example, two protons. The differential scattering cross section for two protons scattering in the symmetric vacuum is described in the center of mass coordinates by the following equation[21-26]

$$I^a(\phi_a) = \frac{A}{v_a^4}\left[\csc^4 \frac{\phi_a}{2} + \sec^4 \frac{\phi_a}{2} - \csc^2 \frac{\phi_a}{2} \sec^2 \frac{\phi_a}{2} \cos\left(2\xi_a \ell n \tan \frac{\phi_a}{2}\right)\right] \tag{227}$$

where

$$A = \left(\frac{e^2}{2m}\right)^2 \tag{228}$$

$$\xi_a = \frac{e^2}{\hbar v_a} \tag{229}$$

and m = reduced mass = $m_p/2$ where m_p = proton mass, and v_a = conventionally determined relative speed of the two protons. For the scattering of two protons within asymmetric bulk matter or vacuum, the differential scattering cross section is written as a complex number as follows

$$\bar{I}(\bar{\phi}) = \frac{A}{\bar{v}^4}\left[\csc^4 \frac{\bar{\phi}}{2} + \sec^4 \frac{\bar{\phi}}{2} - \csc^2 \frac{\bar{\phi}}{2} \sec^2 \frac{\bar{\phi}}{2} \cos\left(2\bar{\xi} \ell n \tan \frac{\bar{\phi}}{2}\right)\right] \tag{230}$$

where

$$\bar{\xi} = \frac{e^2}{\hbar \bar{v}} = \frac{e^2}{\hbar v} e^{-j\theta_v} = \xi e^{-j\theta_v} \tag{231}$$

Equation (230) can be rewritten as

$$\bar{I}(\bar{\phi}) = \frac{A}{v^4}\left(J_R e^{-j\theta JR} + J_E e^{-j\theta JE} + J_I e^{-j\theta JI}\right) \tag{232}$$

where the Rutherford term and the exchange term are written as

$$J_R = K_s^2 \tag{233}$$

$$J_E = K_{se}^2 \tag{234}$$

$$\theta_{JR} = 4\theta_v + 2x_\phi \tag{235}$$

$$\theta_{JE} = 4\theta_v - 2y_\phi \tag{236}$$

The interaction term is written as

$$\bar{J}_I = - K_s K_{se} \cos \bar{G}\, e^{-j(4\theta_v + x_\phi - y_\phi)} \tag{237}$$

where from equations (211) and (230) it follows that

$$\bar{G} = Ge^{j\theta_G} = 2\,\bar{\xi}\, \ell n \left(\frac{e^{jz_\phi}}{K_t}\right) \tag{238}$$

$$= 2\xi(\cos\theta_v - j \sin\theta_v)\,(jz_\phi - \ell n\, K_t)$$

$$= 2\xi[z_\phi \sin\theta_v - \ell n\, K_t \cos\theta_v + j(z_\phi \cos\theta_v + \ell n\, K_t \sin\theta_v)]$$

so that

$$G = 2\xi \sqrt{z_\phi^2 + (\ell n\, K_t)^2} \tag{239}$$

$$\tan\theta_G = \frac{z_\phi \cos\theta_v + \ell n\, K_t \sin\theta_v}{z_\phi \sin\theta_v - \ell n\, K_t \cos\theta_v} \tag{240}$$

The interaction term in equation (237) can be rewritten as

$$\bar{J}_I = J_I e^{-j\theta_{JI}} \tag{241}$$

where

$$J_I = - K_s K_{se} C_G \tag{242}$$

$$\theta_{JI} = 4\theta_v + x_\phi - y_\phi + \theta_{cG} \tag{243}$$

$$C_G = \sqrt{\cos^2 (G \cos \theta_G) + \sinh^2 (G \sin \theta_G)} \tag{244}$$

$$\tan \theta_{cG} = \tan (G \cos \theta_G) \tanh (G \sin \theta_G) \tag{245}$$

The two equations for determining I and θ_I are obtained from equations (232) through (245) as

$$I \cos \theta_I = \frac{A}{v^4} (J_R \cos \theta_{JR} + J_E \cos \theta_{JE} + J_I \cos \theta_{JI}) \tag{246}$$

$$I \sin \theta_I = - \frac{A}{v^4} (J_R \sin \theta_{JR} + J_E \sin \theta_{JE} + J_I \sin \theta_{JI}) \tag{247}$$

In this manner a theory of Mott scattering in an asymmetric medium is developed which is consistent with the gauge theory of the asymmetric background medium. The measured cross section is = I cos θ_I .

C. Bhabha Scattering

Bhabha scattering is electron-positron scattering $e^+ + e^- \rightarrow e^+ + e^-$ by photon exchange and pair annihilation. The differential scattering cross section in the center of mass system and in the high energy limit ($E_a >> m$) is given for the symmetrical vacuum by[27]

$$I^a = \frac{B}{\gamma_a^2 v_a^2} \left[\frac{1 + \cos^4 \frac{\phi_a}{2}}{\sin^4 \frac{\phi_a}{2}} + \frac{1}{2} (1 + \cos^2 \phi_a) - 2 \frac{\cos^4 \frac{\phi_a}{2}}{\sin^2 \frac{\phi_a}{2}} \right] \tag{248}$$

where

$$B = \frac{1}{2} \left(\frac{\alpha}{mc} \right)^2 \tag{249}$$

$$\gamma_a = (1 - v_a^2/c^2)^{-1/2} \tag{250}$$

α = fine structure constant, m = reduced mass of electron, and v_a = conventionally determined speed in center of mass system. The first term in equation (248) is the photon exchange term, the second term is the pair annihilation

contribution, while the third term represents the interference between the first two terms.[22]

The corresponding cross section for Bhabha scattering in asymmetric bulk matter or vacuum is written as

$$\bar{I}(\bar{\phi}) = \frac{B}{\bar{\gamma}^2\bar{v}^2}\left[\frac{1+\cos^4\frac{\bar{\phi}}{2}}{\sin^4\frac{\bar{\phi}}{2}} + \frac{1}{2}\left(1+\cos^2\bar{\phi}\right) - 2\frac{\cos^4\frac{\bar{\phi}}{2}}{\sin^2\frac{\bar{\phi}}{2}}\right] \qquad (251\text{-}260)$$

where $\bar{v}$ and $\bar{\gamma}$ are given by equations (138) and (171) respectively. Combining equations (208) through (215) with equation (260) gives

$$\bar{I}(\bar{\phi}) = \frac{B}{\gamma^2 v^2}\left(L_1 e^{-j\Phi_1} + L_2 e^{-j\Phi_2} + L_3 e^{-j\Phi_3} + L_4 e^{-j\Phi_4} + L_5 e^{-j\Phi_5}\right) \qquad (261)$$

where γ is now the magnitude of the boost for a broken symmetry system and is given by equation (172), and where

$$L_1 = K_s^2 \qquad L_4 = \frac{1}{2}C_\phi^2 \qquad (262)$$

$$L_2 = K_s^2/K_{se}^2 \qquad L_5 = -2K_s/K_{se}^2 \qquad (263)$$

$$L_3 = 1/2$$

$$\Phi_1 = 2(\theta_\gamma + \theta_v + x_\phi) \qquad \Phi_4 = 2(\theta_\gamma + \theta_v + \theta_{c\phi}) \qquad (264)$$

$$\Phi_2 = 2(\theta_\gamma + \theta_v + x_\phi + y_\phi) \qquad \Phi_5 = 2(\theta_\gamma + \theta_v + x_\phi/2 + y_\phi) \qquad (265)$$

$$\Phi_3 = 2(\theta_\gamma + \theta_v) \qquad (266)$$

where θ_γ is given by equation (173). From equation (261) it follows that

$$I\cos\theta_I = \frac{B}{\gamma^2 v^2}\left(L_1\cos\Phi_1 + L_2\cos\Phi_2 + L_3\cos\Phi_3 + L_4\cos\Phi_4 + L_5\cos\Phi_5\right) \qquad (267)$$

$$I\sin\theta_I = -\frac{B}{\gamma^2 v^2}\left(L_1\sin\Phi_1 + L_2\sin\Phi_2 + L_3\sin\Phi_3 + L_4\sin\Phi_4 + L_5\sin\Phi_5\right) \qquad (268)$$

from which I and θ_I can be easily determined. In equations (261), (267), and (268) the first two terms are due to photon exchange, terms three and four are due to pair annihilation, and term five is the interference term.

D. Møller Scattering

Møller scattering is electron-electron scattering $e^- + e^- \rightarrow e^- + e^-$ by photon exchange. The differential scattering cross section for this process in the center of mass system and for high energy ($E_a >> m$) is given for the symmetrical vacuum by[19,23,27]

$$I^a = \frac{B}{\gamma_a^2 v_a^2}\left[\frac{1+\cos^4\frac{\phi_a}{2}}{\sin^4\frac{\phi_a}{2}} + \frac{2}{\sin^2\frac{\phi_a}{2}\cos^2\frac{\phi_a}{2}} + \frac{1+\sin^4\frac{\phi_a}{2}}{\cos^4\frac{\phi_a}{2}}\right] \tag{269}$$

where ϕ_a = scattering angle, and γ_a = ordinary relativistic boost given by equation (250). The first term in equation (269) is due to direct scattering, the second is due to interference, and the third term is the result of exchange scattering.

The corresponding differential scattering cross section for Møller scattering in bulk matter or vacuum with broken internal symmetries is given by

$$\bar{I} = \frac{B}{\bar{\gamma}^2\bar{v}^2}\left[\frac{1+\cos^4\frac{\bar{\phi}}{2}}{\sin^4\frac{\bar{\phi}}{2}} + \frac{2}{\sin^2\frac{\bar{\phi}}{2}\cos^2\frac{\bar{\phi}}{2}} + \frac{1+\sin^4\frac{\bar{\phi}}{2}}{\cos^4\frac{\bar{\phi}}{2}}\right] \tag{270}$$

where the particle speed $\bar{v}$ and boost $\bar{\gamma}$ for a broken symmetry system are given by equations (138) and (171) respectively. Combining equation (270) with equations (208) through (215) gives

$$\bar{I} = \frac{B}{\gamma^2 v^2}\left(T_1 e^{-j\psi_1} + T_2 e^{-j\psi_2} + T_3 e^{-j\psi_3} + T_4 e^{-j\psi_4} + T_5 e^{-j\psi_5}\right) \tag{271}$$

where the boost γ for a broken symmetry system is given by equation (172), and where

$$T_1 = K_s^2 \qquad T_4 = K_{se}^2 \tag{272}$$

$$T_2 = K_s^2/K_{se}^2 \qquad T_5 = K_{se}^2/K_s^2 \tag{273}$$

$$T_3 = 2K_s K_{se} \tag{274}$$

where K_s and K_{se} are given by equations (213) and (215) respectively, and where

$$\psi_1 = 2(\theta_\gamma + \theta_v + x_\phi) \qquad \psi_4 = 2(\theta_\gamma + \theta_v - y_\phi) \tag{275}$$

$$\psi_2 = 2(\theta_\gamma + \theta_v + x_\phi + y_\phi) \qquad \psi_5 = 2(\theta_\gamma + \theta_v - x_\phi - y_\phi) \tag{276}$$

$$\psi_3 = 2(\theta_\gamma + \theta_v + x_\phi/2 - y_\phi/2) \tag{277}$$

where x_ϕ and y_ϕ are given by equations (216) and (218) respectively, and where θ_γ is expressed in terms of θ_v by equation (173). From equation (271) it follows that the magnitude and internal phase of the differential cross section for Møller scattering in a broken symmetry system is given by

$$I \cos\theta_I = \frac{B}{\gamma^2 v^2}(T_1 \cos\psi_1 + T_2 \cos\psi_2 + T_3 \cos\psi_3 + T_4 \cos\psi_4 + T_5 \cos\psi_5) \tag{278}$$

$$I \sin\theta_I = -\frac{B}{\gamma^2 v^2}(T_1 \sin\psi_1 + T_2 \sin\psi_2 + T_3 \sin\psi_3 + T_4 \sin\psi_4 + T_5 \sin\psi_5) \tag{279}$$

which can be solved for I and θ_I immediately.

7. DIRAC EQUATION FOR FERMIONS IN ASYMMETRIC BULK MATTER OR VACUUM. The Dirac equation determines the spectrum and eigenfunctions of half-integral spin particles moving in an external potential.[18] The eigenfunctions take the form of four-component spinors, and therefore the Dirac equation for a particle moving in the symmetric vacuum under the influence of an external potential must be equivalent to four equations. In fact the Dirac equation is a matrix equation involving 4 x 4 matrices and is written as[18,19,27-37]

$$(-i\gamma^\mu \frac{\partial}{\partial x_{\mu a}} + m + V_e^a)\psi^a = 0 \tag{280}$$

where $x_{\mu a} = t_a, x_a, y_a, z_a$, and where $\gamma_\mu = \gamma_o, \gamma_1, \gamma_2$, and γ_3 are the four Dirac matrices. Within asymmetric bulk matter or vacuum the Dirac equation is expected to be written as

$$(-i\gamma^\mu \frac{\partial}{\partial \bar{x}_\mu} + m + \bar{W})\bar{\psi} = 0 \tag{281}$$

where $\bar{\psi}$ = spinor with internal phase given by

$$\bar{\psi} = \psi e^{j\theta_\psi} = \psi_R + j\psi_I \tag{282}$$

$$\psi_R = \psi \cos \theta_\psi \tag{283}$$

$$\psi_I = \psi \sin \theta_\psi \tag{284}$$

and where the complex number potential is written as

$$\bar{W} = \bar{V}_e + \bar{V}_g \tag{285}$$

The gauge rotated time and space coordinates $\bar{x}_\mu = \bar{t}, \bar{x}, \bar{y}$, and $\bar{z}$ of a particle in bulk matter or vacuum with broken internal symmetries are written as

$$\bar{t} = te^{j\theta_t} \qquad \bar{x} = xe^{j\theta_x} \tag{286}$$

$$\bar{y} = ye^{j\theta_y} \qquad \bar{z} = ze^{j\theta_z} \tag{287}$$

The combined effects of gauge rotated coordinates, gauge rotated external potential (which is a function of the gauge rotated coordinates), and the gauge potential itself $\bar{V}_g$, will manifest themselves in the eigenvalues and eigenfunctions of the Dirac equation for a fermion located in an asymmetric system.

The space and time derivatives that appear in equation (281) are written as

$$\partial/\partial\bar{t} = e^{-j\theta_{dt}}\left[1 + \left(t\frac{\partial\theta_t}{\partial t}\right)^2\right]^{-1/2} \qquad \partial/\partial t = e^{-j\theta_{dt}} \cos\beta_{t,t}\, \partial/\partial t \tag{288}$$

$$\partial/\partial\bar{x} = e^{-j\theta_{dx}}\left[1 + \left(x\frac{\partial\theta_x}{\partial x}\right)^2\right]^{-1/2} \qquad \partial/\partial x = e^{-j\theta_{dx}} \cos\beta_{x,x}\, \partial/\partial x \tag{289}$$

$$\partial/\partial\bar{y} = e^{-j\theta_{dy}}\left[1 + \left(y\frac{\partial\theta_y}{\partial y}\right)^2\right]^{-1/2} \qquad \partial/\partial y = e^{-j\theta_{dy}} \cos\beta_{y,y}\, \partial/\partial y \tag{290}$$

$$\partial/\partial\bar{z} = e^{-j\theta_{dz}}\left[1 + \left(z\frac{\partial\theta_z}{\partial z}\right)^2\right]^{-1/2} \qquad \partial/\partial z = e^{-j\theta_{dz}} \cos\beta_{z,z}\, \partial/\partial z \tag{291}$$

where

$$\theta_{dt} = \theta_t + \beta_{t,t} = \theta_o + \beta_{o,o} = \theta_{do} \tag{292}$$

$$\theta_{dx} = \theta_x + \beta_{x,x} = \theta_1 + \beta_{1,1} = \theta_{d1} \tag{293}$$

$$\theta_{dy} = \theta_y + \beta_{y,y} = \theta_2 + \beta_{2,2} = \theta_{d2} \tag{294}$$

$$\theta_{dz} = \theta_z + \beta_{z,z} = \theta_3 + \beta_{3,3} = \theta_{d3} \tag{295}$$

and where

$$\tan \beta_{o,o} = \tan \beta_{t,t} = t\, \partial\theta_t/\partial t \tag{296}$$

$$\tan \beta_{1,1} = \tan \beta_{x,x} = x\, \partial\theta_x/\partial x \tag{297}$$

$$\tan \beta_{2,2} = \tan \beta_{y,y} = y\, \partial\theta_y/\partial y \tag{298}$$

$$\tan \beta_{3,3} = \tan \beta_{z,z} = z\, \partial\theta_z/\partial z \tag{299}$$

In this way the necessary space and time derivatives in Dirac's equation for broken symmetry systems are evaluated.

Equation (281) can then be rewritten as

$$(-ie^{-j\theta_{d\mu}} \cos \beta_{\mu,\mu}\, \gamma^\mu \frac{\partial}{\partial x_\mu} + m + \bar{W})\, \bar{\psi} = 0 \tag{300}$$

The two matrix equations corresponding to equation (300) are obtained by taking the real and imaginary parts in the internal space as follows

$$(-i \cos \theta_{d\mu} \cos \beta_{\mu,\mu}\, \gamma^\mu\, \partial/\partial x_\mu + m + W \cos \theta_W)\, \psi_R \tag{301}$$

$$- (i \sin \theta_{d\mu} \cos \beta_{\mu,\mu}\, \gamma^\mu\, \partial/\partial x_\mu + W \sin \theta_W)\, \psi_I = 0$$

$$(i \sin \theta_{d\mu} \cos \beta_{\mu,\mu}\, \gamma^\mu\, \partial/\partial x_\mu + W \sin \theta_W)\, \psi_R \tag{302}$$

$$+ (-i \cos \theta_{d\mu} \cos \beta_{\mu,\mu}\, \gamma^\mu\, \partial/\partial x_\mu + m + W \cos \theta_W)\, \psi_I = 0$$

where it is assumed that the mass is a real number that is not affected by the gauge rotations due to the asymmetric background. Note that from equation (285) it follows that

$$W \cos \theta_W = V_e \cos \theta_{Ve} + V_g \cos \theta_{Vg} \tag{303}$$

$$W \sin \theta_W = V_e \sin \theta_{Ve} + V_g \sin \theta_{Vg} \tag{304}$$

The equations (301) and (302) are equivalent to eight equations for the eight spinor components ψ_R^o , ψ_R^1 , ψ_R^2 , ψ_R^3 , ψ_I^o , ψ_I^1 , ψ_I^2 , and ψ_I^3 , or equivalently ψ^o , ψ^1 , ψ^2 , ψ^3 , $\theta_{\psi o}$, $\theta_{\psi 1}$, $\theta_{\psi 2}$, and $\theta_{\psi 3}$. Therefore Dirac's equation for a fermion located in a background with broken internal symmetry is equivalent to eight independent equations. An approximate solution ignores the imaginary wavefunction components, which gives

$$(- i \cos \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu \partial/\partial x_\mu + m + W \cos \theta_W) \psi_R = 0 \tag{301A}$$

$$(i \sin \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu \partial/\partial x_\mu + W \sin \theta_W) \psi_R = 0 \tag{302A}$$

as the Dirac equations with four spinor components.

Alternatively, equation (300) can be combined with equation (282) to give the following set of Dirac equations

$$[- i \cos \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu(\partial/\partial x_\mu + \partial\theta_\psi/\partial x_\mu) + m + W \cos \theta_W]\psi = 0 \tag{304A}$$

$$[i \sin \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu(\partial/\partial x_\mu + \partial\theta_\psi/\partial x_\mu) + W \sin \theta_W]\psi = 0 \tag{304B}$$

If the space and time derivatives of θ_ψ can be neglected these equations become

$$(- i \cos \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu \partial/\partial x_\mu + m + W \cos \theta_W)\psi = 0 \tag{304C}$$

$$(i \sin \theta_{d\mu} \cos \beta_{\mu,\mu} \gamma^\mu \partial/\partial x_\mu + W \sin \theta_W)\psi = 0 \tag{304D}$$

8. SCHRÖDINGER'S EQUATION FOR A PARTICLE WITHIN ASYMMETRIC BULK MATTER OR VACUUM. This section considers the effects of bulk matter and vacuum with broken internal symmetries on Schrödinger's equation for a particle moving in a potential field. The time dependent Schrödinger equation for a particle moving in a potential field in a symmetric vacuum is written as[38-47]

$$\left[\frac{1}{2m} \left(p_{xa}^2 + p_{ya}^2 + p_{za}^2 \right) + V_e^a \right] \psi^a = i\hbar \frac{\partial \psi^a}{\partial t_a} \tag{305}$$

where the single particle momentum and energy operators are given by

$$P_{\alpha a} = - i\hbar \partial/\partial \alpha_a \qquad E_a = i\hbar \partial/\partial t_a \tag{306}$$

with $\alpha = x, y, z$. Within asymmetric bulk matter or vacuum it is assumed that space, time, momentum operators, energy operator, potential, and wave functions exhibit broken symmetries and must be represented by complex numbers in internal space. For this case the time dependent Schrödinger equation is written as

$$\left[\frac{1}{2m} \left(\bar{p}_x^2 + \bar{p}_y^2 + \bar{p}_z^2 \right) + \bar{W} \right] \bar{\psi} = i\hbar \frac{\partial \bar{\psi}}{\partial \bar{t}} \tag{307}$$

where $\bar{W} = \bar{V}_e + \bar{V}_g$ and where

$$\bar{p}_\alpha = p_\alpha e^{j\theta p\alpha} = - i\hbar \partial/\partial \bar{\alpha} = - i\hbar \cos \beta_{\alpha,\alpha} \, e^{-j\theta d\alpha} \, \partial/\partial \alpha \tag{308}$$

$$\bar{E} = E e^{j\theta E} = i\hbar \partial/\partial \bar{t} = i\hbar \cos \beta_{t,t} \, e^{-j\theta dt} \, \partial/\partial t \tag{309}$$

where

$$\cos \beta_{\alpha,\alpha} = \left[1 + \left(\alpha \frac{\partial \theta_\alpha}{\partial \alpha} \right)^2 \right]^{-1/2} \tag{310}$$

$$\theta_{d\alpha} = \theta_\alpha + \beta_{\alpha,\alpha} \tag{311}$$

$$\cos \beta_{t,t} = \left[1 + \left(t \frac{\partial \theta_t}{\partial t} \right)^2 \right]^{-1/2} \tag{312}$$

$$\theta_{dt} = \theta_t + \beta_{t,t} \tag{313}$$

where $\beta_{t,t}$ and $\beta_{\alpha,\alpha}$ are given in equations (296) through (299). From equations (308) and (309) it follows that

$$p_\alpha = - i\hbar \cos \beta_{\alpha,\alpha} \, \partial/\partial \alpha \tag{314}$$

$$\theta_{p\alpha} = - \theta_{d\alpha} \tag{315}$$

$$E = i\hbar \cos \beta_{t,t} \, \partial/\partial t \tag{316}$$

$$\theta_E = - \theta_{dt} \tag{317}$$

It is easy to show from the Heisenberg uncertainty principle applied to $\bar{p}_\alpha$ and $\bar{\alpha}$ and to $\bar{E}$ and $\bar{t}$ that $\beta_{\alpha,\alpha} < 0$ and $\beta_{t,t} < 0$, so that from equations (296) through (299) it follows that θ_α is a decreasing function of α, and θ_t is a decreasing function of t.

The kinetic energy operator in equation (307) is written as

$$\sum_{\alpha=1}^{3} \frac{\bar{p}_\alpha^2}{2m} \bar{\psi} = - \frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos \beta_{\alpha,\alpha} \, e^{-j\theta_{d\alpha}} \frac{\partial}{\partial \alpha} \left(\cos \beta_{\alpha,\alpha} \, e^{-j\theta_{d\alpha}} \frac{\partial}{\partial \alpha} \right) \bar{\psi} \tag{318}$$

For simplicity it is assumed that β_α and θ_α are slowly varying functions of position so that equation (318) can be rewritten as

$$\sum_{\alpha=1}^{3} \frac{\bar{p}_\alpha^2}{2m} \bar{\psi} = - \frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos^2 \beta_{\alpha,\alpha} \, e^{-j2\theta_{d\alpha}} \frac{\partial^2 \bar{\psi}}{\partial \alpha^2} \tag{319}$$

Writing the wavefunction as $\bar{\psi} = \psi_R + j\psi_I$ allows equation (307) to be written as two component equations as follows

$$- \frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos^2 \beta_{\alpha,\alpha} \left[\cos (2\theta_{d\alpha}) \frac{\partial^2 \psi_R}{\partial \alpha^2} + \sin (2\theta_{d\alpha}) \frac{\partial^2 \psi_I}{\partial \alpha^2} \right] \tag{320}$$

$$+ W(\cos \theta_W \, \psi_R - \sin \theta_W \, \psi_I)$$

$$= i\hbar \cos \beta_{t,t} \left(\cos \theta_{dt} \frac{\partial \psi_R}{\partial t} + \sin \theta_{dt} \frac{\partial \psi_I}{\partial t} \right)$$

$$- \frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos^2 \beta_{\alpha,\alpha} \left[- \sin (2\theta_{d\alpha}) \frac{\partial^2 \psi_R}{\partial \alpha^2} + \cos (2\theta_{d\alpha}) \frac{\partial^2 \psi_I}{\partial \alpha^2} \right] \tag{321}$$

$$+ W(\sin \theta_W \, \psi_R + \cos \theta_W \, \psi_I)$$

$$= i\hbar \cos \beta_{t,t} \left(- \sin \theta_{dt} \frac{\partial \psi_R}{\partial t} + \cos \theta_{dt} \frac{\partial \psi_I}{\partial t} \right)$$

Equations (320) and (321) can be used to determine ψ_R and ψ_I. For the case of a stationary state the wave function components are written as

$$\psi_R = \phi_R e^{-i\varepsilon t/\hbar} \qquad \psi_I = \phi_I e^{-i\varepsilon t/\hbar} \tag{322}$$

and equations (320) and (321) become

$$-\frac{\hbar^2}{2m}\sum_{\alpha=1}^{3}\cos^2\beta_{\alpha,\alpha}\left[\cos(2\theta_{d\alpha})\frac{\partial^2\phi_R}{\partial\alpha^2}+\sin(2\theta_{d\alpha})\frac{\partial^2\phi_I}{\partial\alpha^2}\right] \tag{323}$$

$$+ W(\cos\theta_W\,\phi_R - \sin\theta_W\,\phi_I)$$

$$= \varepsilon\cos\beta_{t,t}\,(\cos\theta_{dt}\,\phi_R + \sin\theta_{dt}\,\phi_I)$$

$$-\frac{\hbar^2}{2m}\sum_{\alpha=1}^{3}\cos^2\beta_{\alpha,\alpha}\left[-\sin(2\theta_{d\alpha})\frac{\partial^2\phi_R}{\partial\alpha^2}+\cos(2\theta_{d\alpha})\frac{\partial^2\phi_I}{\partial\alpha^2}\right] \tag{324}$$

$$+ W(\sin\theta_W\,\phi_R + \cos\theta_W\,\phi_I)$$

$$= \varepsilon\cos\beta_{t,t}\,(-\sin\theta_{dt}\,\phi_R + \cos\theta_{dt}\,\phi_I)$$

It is generally quite difficult to determine ϕ_R and ϕ_I (and ε) from equations (323) and (324). The form and magnitude of the functions $\beta_{\alpha,\alpha}$, $\beta_{t,t}$, $\theta_{d\alpha}$, and θ_{dt} depend on the nature, density, and temperature of the asymmetric bulk matter or vacuum surrounding a particle.

Consider the case where the asymmetries are sufficiently small that the imaginary part of the wavefunction can be neglected in equation (323), so that this equation becomes

$$-\frac{\hbar^2}{2m}\sum_{\alpha=1}^{3}\cos^2\beta_{\alpha,\alpha}\cos(2\theta_{d\alpha})\frac{d^2\phi_R}{d\alpha^2} + W\cos\theta_W\,\phi_R = \varepsilon\cos\beta_{t,t}\cos\theta_{dt}\,\phi_R \tag{325}$$

For an isotropic system equation (325) becomes

At this point it is easy to treat the Klein-Gordon equation for a particle that is located in an asymmetric background. For a particle in a symmetric vacuum, the Klein-Gordon equation is written as[34-37]

$$\frac{\partial^2 \psi_a}{\partial t_a^2} = c^2 \nabla_a^2 \psi_a - \frac{m^2 c^4}{\hbar^2} \psi_a \tag{330}$$

Within asymmetric bulk matter or vacuum the spacetime interactions induce a broken symmetry in the wave function and in the space and time coordinates, so that the Klein-Gordon equation becomes

$$\frac{\partial^2 \bar{\psi}}{\partial \bar{t}^2} = c^2 \left(\frac{\partial^2 \bar{\psi}}{\partial \bar{x}^2} + \frac{\partial^2 \bar{\psi}}{\partial \bar{y}^2} + \frac{\partial^2 \bar{\psi}}{\partial \bar{z}^2} \right) - \frac{m^2 c^4}{\hbar^2} \bar{\psi} \tag{331}$$

Taking the real and imaginary components of equation (331) gives

$$R_{2t}(\psi_R, \psi_I) \sim c^2 [R_{2x}(\psi_R, \psi_I) + R_{2y}(\psi_R, \psi_I) + R_{2z}(\psi_R, \psi_I)] - \frac{m^2 c^4}{\hbar^2} \psi_R \tag{332}$$

$$I_{2t}(\psi_R, \psi_I) \sim c^2 [I_{2x}(\psi_R, \psi_I) + I_{2y}(\psi_R, \psi_I) + I_{2z}(\psi_R, \psi_I)] - \frac{m^2 c^4}{\hbar^2} \psi_I \tag{333}$$

where

$$R_{2\eta}(\psi_R, \psi_I) = \cos^2 \beta_{\eta,\eta} \left[\cos(2\theta_{d\eta}) \frac{\partial^2 \psi_R}{\partial \eta^2} + \sin(2\theta_{d\eta}) \frac{\partial^2 \psi_I}{\partial \eta^2} \right] \tag{334}$$

$$I_{2\eta}(\psi_R, \psi_I) = \cos^2 \beta_{\eta,\eta} \left[-\sin(2\theta_{d\eta}) \frac{\partial^2 \psi_R}{\partial \eta^2} + \cos(2\theta_{d\eta}) \frac{\partial^2 \psi_I}{\partial \eta^2} \right] \tag{335}$$

where $\eta = t, x, y, z$, and where

$$\theta_{d\eta} = \theta_\eta + \beta_{\eta,\eta} \tag{336}$$

9. CONCLUSION. On account of spacetime interactions with bulk matter and the vacuum, these systems exhibit broken internal symmetries. In the case of black body radiation in asymmetric bulk matter or vacuum, the photons have complex number frequencies which produce a radiation pressure and energy density that have broken internal symmetries. The space and time coordinates within a

$$\frac{d^2\phi_R}{dx^2} + \frac{2m}{3\hbar^2 \cos^2 \beta_{x,x} \cos(2\theta_{dx})} (\varepsilon \cos \beta_{t,t} \cos \theta_{dt} - W \cos \theta_W) \phi_R = 0 \qquad (326)$$

This can be rewritten as

$$\frac{d^2\phi_R}{dx^2} + \frac{2m^*}{3\hbar^2} (\varepsilon - W^*) \phi_R = 0 \qquad (327)$$

where m^* is an effective given by

$$m^* = \frac{m}{\cos^2 \beta_{x,x} \cos(2\theta_{dx})} \qquad (328)$$

and W^* is an energy dependent effective potential given by

$$W^* = \varepsilon(1 - \cos \beta_{t,t} \cos \theta_{dt}) + W \cos \theta_W \qquad (329)$$

Equations (323) and (324) can also be written in terms of the magnitude and phase angle of the wavefunction by writing $\phi_R = \phi \cos \theta_\phi$ and $\phi_I = \phi \sin \theta_\phi$. If the derivatives of the phase angle θ_ϕ are sufficiently small and can be neglected then equations (323) and (324) can be rewritten as

$$-\frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos^2 \beta_{\alpha,\alpha} \cos(2\theta_{d\alpha}) \frac{d^2\phi}{d\alpha^2} + W \cos \theta_W \phi \qquad (323A)$$

$$= \varepsilon \cos \beta_{t,t} \cos \theta_{dt} \phi$$

$$+\frac{\hbar^2}{2m} \sum_{\alpha=1}^{3} \cos^2 \beta_{\alpha,\alpha} \sin(2\theta_{d\alpha}) \frac{d^2\phi}{d\alpha^2} + W \sin \theta_W \phi \qquad (324A)$$

$$= -\varepsilon \cos \beta_{t,t} \sin \theta_{dt} \phi$$

For a one dimensional system the factor 3 that appears in equations (326) and (327) should be replaced by unity. Therefore in asymmetric bulk matter or vacuum the particle acquires an effective mass, due to spacetime interactions, which is larger than the bare mass. In addition an energy dependent effective potential arises whose value depends on the degree of asymmetry that exists in the background of the particle.

broken symmetry system are also gauge rotated and are described by internal phase angles. From this it follows that geometrical angles are described by complex numbers and have internal phase angles. The skewed nature of space and time affects the fundamental scattering processes of atomic particles. All atomic processes that occur in asymmetric bulk matter or vacuum should also have broken symmetries that are manifested in the measured differential cross sections. For broken symmetry quantum systems, the asymmetry produces an effective mass in the Schrödinger equation that is larger than the bare mass of a particle.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Chaichian, M. and Nelipa, N., Introduction to Gauge Field Theories, Springer-Verlag, New York, 1984.

2. Ryder, L. H., Quantum Field Theory, Cambridge University Press, New York, 1985.

3. Dodd, J., The Ideas of Particle Physics, Cambridge University Press, New York, 1984.

4. Aitchison, I. and Hey, A., Gauge Theories in Particle Physics, Adam Hilger, Bristol, 1982.

5. Weiss, R. A., "Thermodynamic Gauge Theory of Solids and Quantum Liquids with Internal Phase", Fifth Army Conference on Applied Mathematics and Computing, West Point, New York, ARO 88-1, June 15-18, 1987, p. 649.

6. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

7. Hagedorn, R., Relativistic Kinematics, Benjamin, New York, 1964.

8. Huang, K., Statistical Mechanics, John Wiley, New York, 1963.

9. Hill, T. L., An Introduction to Statistical Thermodynamics, Addison-Wesley, New York, 1960.

10. Planck, M., The Theory of Heat Radiation, Dover, New York, 1959.

11. Joos, G., Theoretical Physics, Hafner, New York, 1950.

12. Page, L., Introduction to Theoretical Physics, Van Nostrand, New York, 1952.

13. Rybicki, G. and Lightman, A., Radiative Processes in Astrophysics, John Wiley, New York, 1979.

14. Sommerfeld, A., Thermodynamics and Statistical Mechanics, Academic Press, New York, 1964, p. 148.

15. Brillouin, L., Tensors in Mechanics and Elasticity, Academic Press, New York, 1964.

16. Weiss, R. A., "Radiation Pressure of Light in a Refractive Medium", Journal of Applied Physics, Vol. 47, No. 1, Jan. 1976.

17. Born, M., Atomic Physics, Hafner, New York, 1953.

18. Bethe, H. and Salpeter, E., Quantum Mechanics of One- and Two-Electron Systems, Springer-Verlag, New York, 1957.

19. Brown, K. M., "Solutions of Simultaneous Non-Linear Equations", Communications of the ACM, Vol. 10, No. 11, Nov. 1967, p. 728.

20. DeBenedetti, S., Nuclear Interactions, John Wiley, New York, 1964.

21. Eder, G., Nuclear Forces, MIT Press, Cambridge, MA, 1968.

22. Kursunoglu, B., Modern Quantum Theory, Freeman, San Francisco, 1962.

23. Mott, N. and Massey, H., The Theory of Atomic Collisions, Oxford University Press, Oxford, 1965.

24. Sachs, R., Nuclear Theory, Addison-Wesley, New York, 1953.

25. Bethe, H. and Morrison, P., Elementary Nuclear Theory, John Wiley, New York, 1961.

26. Elton, L., Introductory Nuclear Theory, John Wiley-Interscience, New York, 1959.

27. Mandl, F. and Shaw, G., Quantum Field Theory, John Wiley, New York, 1984.

28. Bjorken, J. and Drell, S., Relativistic Quantum Mechanics, McGraw-Hill, New York, 1964.

29. Bogoliubov, N. and Shirkov, D., Introduction to the Theory of Quantized Fields, John Wiley-Interscience, New York, 1959.

30. Gasiorowicz, S., Elementary Particle Physics, John Wiley, New York, 1966.

31. Bethe, H. A. and Jackiw, R., Intermediate Quantum Mechanics, Benjamin, New York, 1968.

32. Heitler, W., The Quantum Theory of Radiation, Dover, New York, 1984.

33. Dirac, P., The Principles of Quantum Mechanics, Oxford University Press, New York, 1947.

34. Schweber, S., An Introduction to Relativistic Quantum Field Theory, Harper and Row, New York, 1962.

35. Akhiezer, A. and Berestetskii, V., Quantum Electrodynamics, John Wiley-Interscience, New York, 1965.

36. Schweber, S., Bethe, H. and de Hoffmann, F., Mesons and Fields, Vol. 1, Row-Peterson, Evanston, 1956.

37. Feynman, R., Quantum Electrodynamics, Benjamin, New York, 1962.

38. Rojansky, V., Introductory Quantum Mechanics, Prentice-Hall, New York, 1938.

39. Persico, E., Fundamentals of Quantum Mechanics, Prentice-Hall, New York, 1950.

40. Landau, L. and Lifshitz, E., Quantum Mechanics, Addison-Wesley, New York, 1958.

41. Pauling, L. and Wilson, E., Introduction to Quantum Mechanics, McGraw-Hill, New York, 1935.

42. Bohm, D., Quantum Theory, Prentice-Hall, New York, 1951.

43. Kemble, E., The Fundamental Principles of Quantum Mechanics, Dover, New York, 1958.

44. Powell, J. and Crasemann, B., Quantum Mechanics, Addison-Wesley, New York, 1961.

45. Messiah, A., Quantum Mechanics, John Wiley, New York, 1961.

46. Merzbacher, E., Quantum Mechanics, John Wiley, New York, 1961.

47. Mandl, F., Quantum Mechanics, Academic, New York, 1954.

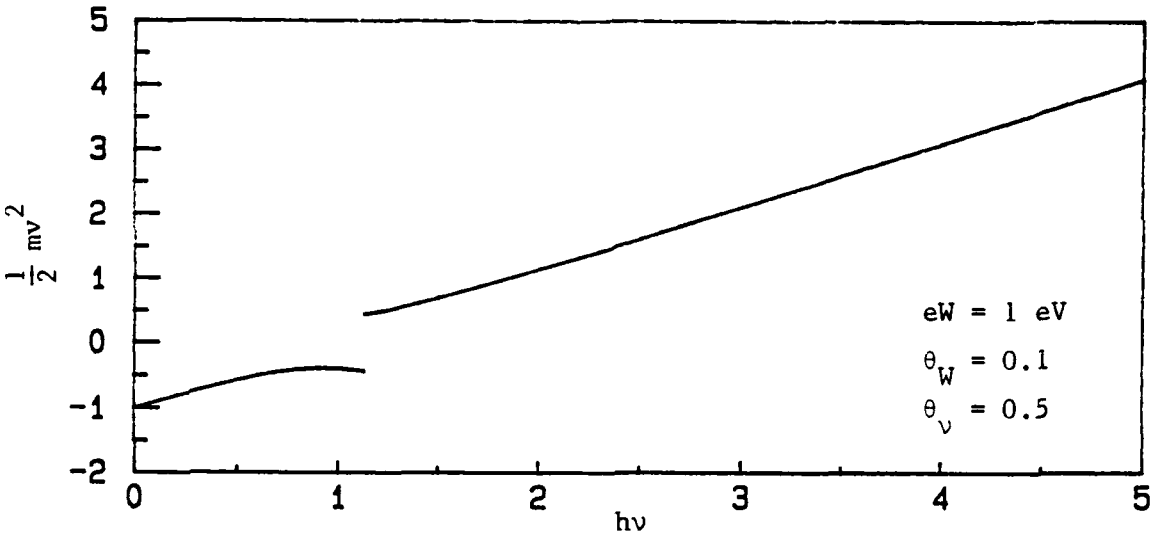


Figure 1. Magnitude of electron kinetic energy versus the magnitude of the photon energy.

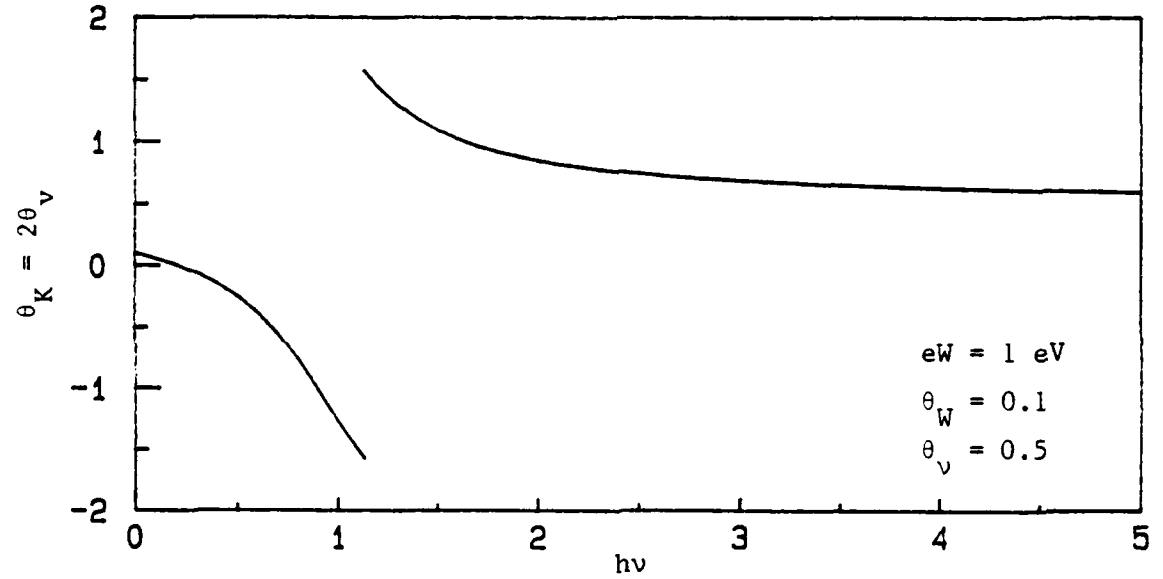


Figure 2. Phase angle of the electron kinetic energy versus the magnitude of the photon energy.

MAXWELL'S EQUATIONS WITH BROKEN INTERNAL SYMMETRIES

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. On account of the broken symmetries of the thermodynamic ground state and excited states of bulk matter and the vacuum, the electric and magnetic fields in bulk matter and the vacuum exhibit broken internal symmetries. Maxwell's equations are formulated for an electromagnetic field with broken internal symmetry. Lorentz covariance is expressed in terms of space and time coordinates that have broken symmetries represented by internal phase angles. Special relativity mechanics in bulk matter and the vacuum with broken symmetries is formulated for particles whose kinematic and dynamic variables exhibit internal phases. Electromagnetic wave equations for broken symmetry matter and vacuum are developed and the gauge conditions for the electromagnetic potential are developed. The vacuum state is shown to have properties that are essentially similar to those of a bulk matter system, and in particular both exhibit broken internal symmetry. The description of electromagnetic effects in matter and the vacuum must properly account for the broken symmetry of the fields and space and time coordinates, and the internal phase angles of the electromagnetic field vectors must be determined jointly with the internal phase angles of the space and time coordinates. A better knowledge of electromagnetic interactions in bulk matter will be useful for understanding electromagnetic wave propagation in the atmosphere and for comprehending the complex processes that occur when high energy microwave beams interact with matter.

1. INTRODUCTION. Electrodynamics is a theory that is based on the Lorentz covariant set of Maxwell's equations and on the symmetry of the gauge group $U(1)$.[1-3] This theory has charges and currents as the sources of the electromagnetic field. Maxwell's equations are a set of partial differential equations that determine the space and time variation of the electric and magnetic fields that are associated with the distribution of charges and currents. Classically, the charges and currents are situated in a passive space and time background (the vacuum) which is assumed to be inert and plays no active part in the determination of the fields. In quantum electrodynamics, the vacuum is taken to be a polarizable medium which can affect the energy levels of charged particle configurations. The active vacuum is one of the great discoveries of twentieth century physics, and has been experimentally verified in a number of ways including a measurement of the Lamb shift of energy levels.[4-6]

In this paper an additional vacuum effect on the electromagnetic field is suggested to manifest itself through the fact that space and time coordinates within asymmetric bulk matter or vacuum acquire internal phase angles (broken symmetries). The electric and magnetic field vectors also acquire broken symmetries. The internal phase angles of the space and time coordinates and of the electromagnetic field vectors are due to the interaction of Minkowski spacetime with bulk matter, the electromagnetic field, and the vacuum. The internal

phase angles of the electromagnetic field vectors must be determined jointly with the internal phase angles of the space and time coordinates, and it is the joint solution of Maxwell's equations and the equations of motion of charged particles for a system with broken internal symmetries that accomplishes this task.

The relativistic values of the electromagnetic field vectors in asymmetric bulk matter or vacuum must satisfy the relativistic trace equation for radiation.[7-8] This radiation trace equation relates the renormalized radiation pressure to the corresponding nonrelativistic radiation pressure. The trace equation for radiation is derived from the relativistic trace equation for the ground state of bulk matter which is written as[7,8]

$$\bar{U} + T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} - 3V\,\frac{d}{dV}(\bar{P}V)_{\bar{U}} = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{1}$$

or equivalently as

$$(1 - \bar{b} + T\,\frac{\partial}{\partial T} - \bar{b}V\,\frac{\partial}{\partial V})\bar{E} - 3(1 + \bar{\gamma} + V\,\frac{\partial}{\partial V} - \bar{\gamma}T\,\frac{\partial}{\partial T})\bar{P} = \psi^a \tag{2}$$

where

$$\psi^a = (T\,\frac{\partial}{\partial T} - b^aV\,\frac{\partial}{\partial V} + 1 - b^a)E^a \tag{3}$$

and where $\bar{U}$, $\bar{E}$, $\bar{P}$, $\bar{\gamma}$, and $\bar{b}$ are complex number representations of the internal energy, energy density, pressure, and the gauge parameters, T = absolute temperature, and V = volume of specified number of particles. The complex number Grüneisen parameter is defined as

$$\bar{\gamma} = \frac{V}{\bar{C}_V}\,\frac{\partial\bar{P}}{\partial T} = \frac{\partial\bar{P}/\partial T}{\partial\bar{E}/\partial T} = \gamma e^{\theta\gamma} \tag{4}$$

where γ and θ_γ = magnitude and phase of the Grüneisen parameter respectively. The corresponding equation for radiation in matter with internal phases is derived from equation (2) to be[8]

$$(1 - \bar{b} + T\,\frac{\partial}{\partial T} - \bar{b}V\,\frac{\partial}{\partial V})\bar{E}_r - \bar{\beta}_r(T\,\frac{\partial\bar{P}}{\partial T} - \bar{P}) \tag{5}$$

$$- 3[(1 + \bar{\gamma} + V\,\frac{\partial}{\partial V} - \bar{\gamma}T\,\frac{\partial}{\partial T})\bar{P}_r - \bar{\delta}_r(T\,\frac{\partial\bar{P}}{\partial T} - \bar{P})] = \psi^a_r$$

where

$$\gamma_r^a = (T\frac{\partial}{\partial T} - b^a V\frac{\partial}{\partial V} + 1 - b^a)E_r^a - \beta_r^a(T\frac{\partial P^a}{\partial T} - P^a) \quad (6)$$

and where $\bar{E}_r$, $\bar{P}_r$, $\bar{\beta}_r$, and $\bar{\delta}_r$ are the complex number radiation energy density, radiation pressure, and two radiation gauge functions respectively.[8] The radiation Grüneisen parameter is defined by

$$\gamma_r = \frac{\partial P_r}{\partial T} / \frac{\partial E_r}{\partial T} \quad (6A)$$

Throughout this paper the index "a" will refer to nonrelativistic (unrenormalized) calculations. Equation (1) with its right hand side set equal to zero represents the asymmetric ground state of the vacuum, while equation (5) with its right hand side equal to zero represents the excited (radiation) states of the asymmetric vacuum.

The relativistic trace equations for the ground and excited states of bulk matter and the vacuum imply that the ground state and excited state pressure fields have broken symmetries.[8] In turn, this implies that all of the descriptive variables of particles and fields located in asymmetric bulk matter or vacuum also exhibit broken symmetries. Therefore the space and time coordinates as well as the electric and magnetic field vectors will exhibit broken symmetries as manifested by internal phase angles. The space and time coordinates are written as

$$\bar{x} = xe^{j\theta_x} \quad (7)$$

$$\bar{y} = ye^{j\theta_y} \quad (8)$$

$$\bar{z} = ze^{j\theta_z} \quad (9)$$

$$\bar{t} = te^{j\theta_t} \quad (10)$$

and the derivatives with respect to the space and time coordinates are written as

$$\partial/\partial\bar{x} = e^{-j\theta_{dx}} \cos \beta_{x,x}\, \partial/\partial x \quad (11)$$

$$\partial/\partial\bar{t} = e^{-j\theta_{dt}} \cos \beta_{t,t}\, \partial/\partial t \quad (12)$$

where

$$\tan \beta_{x,x} = x\, \partial\theta_x/\partial x \quad (13)$$

$$\tan \beta_{t,t} = t\, \partial\theta_t/\partial t \quad (14)$$

$$\cos \beta_{x,x} = [1 + (x\, \partial\theta_x/\partial x)^2]^{-1/2} \tag{15}$$

$$\cos \beta_{t,t} = [1 + (t\, \partial\theta_t/\partial t)^2]^{-1/2} \tag{16}$$

$$\theta_{dx} = \theta_x + \beta_{x,x} \tag{17}$$

$$\theta_{dt} = \theta_t + \beta_{t,t} \tag{17A}$$

In a condensed notation the coordinates and derivatives are written with $\eta = t, x, y, z$ as follows

$$\bar{\eta} = \eta e^{j\theta_\eta} \tag{18}$$

$$\partial/\partial\bar{\eta} = e^{-j\theta_{d\eta}} \cos \beta_{\eta,\eta}\, \partial/\partial\eta \tag{19}$$

$$\theta_{d\eta} = \theta_\eta + \beta_{\eta,\eta} \tag{20}$$

$$\tan \beta_{\eta,\eta} = \eta\, \partial\theta_\eta/\partial\eta \tag{21}$$

$$\cos \beta_{\eta,\eta} = [1 + (\eta\, \partial\theta_\eta/\partial\eta)^2]^{-1/2} \tag{22}$$

Note that it is the real parts of the complex number quantities such as space and time coordinates, electric and magnetic field vectors, pressure and energy that are the measured quantities.

This paper develops Maxwell's equations for electromagnetic fields that have broken internal symmetries and for space and time coordinates that also have broken internal symmetries. Section 2 considers the fields and coordinates with broken symmetry, while Section 3 develops Maxwell's equations and the equations of motion of charged particles in an electromagnetic field in a broken symmetry vacuum or bulk matter system. Section 4 develops the consequences of assuming the validity of Lorentz covariance for coordinate systems with broken internal symmetry. In Section 5 the electromagnetic wave equations and their gauge conditions are written for systems with internal phase angles. Finally, Section 6 develops the equations of the relativistic vacuum from the corresponding bulk matter equations, and a broken symmetry condition for the vacuum state is suggested. Therefore all of the conclusions for the bulk matter state with broken internal symmetries are also valid for the broken symmetry vacuum state. The conventional coordinates t_a , x_a , y_a , z_a are related to the measured values t_m , x_m , y_m , z_m of the complex number coordinates by $t_a = t_m = t \cos \theta_t$, $x_a = x_m = x \cos \theta_x$, $y_a = y_m = y \cos \theta_y$ and $z_a = z_m = z \cos \theta_z$.

2. THE BROKEN SYMMETRY OF ELECTROMAGNETIC FIELDS. For electromagnetic waves within asymmetric bulk matter or vacuum, the electric and magnetic fields

are expected to acquire internal phase angles. This is due to the fact that the spacetime coordinates and the kinematical and dynamical variables of particles in asymmetric bulk matter or vacuum exhibit broken internal symmetries. In particular the particle velocity, and therefore the electric current for charged particles, has an internal phase angle. Therefore the cartesian components of the electric and magnetic field vectors in asymmetric bulk matter or vacuum can be written as

$$\bar{E}_\alpha = E_\alpha e^{j\theta_{E\alpha}} = E_{\alpha R} + jE_{\alpha I} \tag{23}$$

$$\bar{D}_\alpha = D_\alpha e^{j\theta_{D\alpha}} = D_{\alpha R} + jD_{\alpha I} \tag{24}$$

$$\bar{H}_\alpha = H_\alpha e^{j\theta_{H\alpha}} = H_{\alpha R} + jH_{\alpha I} \tag{25}$$

$$\bar{B}_\alpha = B_\alpha e^{j\theta_{B\alpha}} = B_{\alpha R} + jB_{\alpha I} \tag{26}$$

where α = x, y, and z. The phase angles $\theta_{E\alpha}$, $\theta_{D\alpha}$, $\theta_{H\alpha}$, and $\theta_{B\alpha}$ are in general functions of space and time of the general form

$$\theta_{E\alpha} = \theta_{E\alpha}(x,y,z,t,\theta_x,\theta_y,\theta_z,\theta_t) \tag{27}$$

The field vector amplitudes are also functions of space and time, as for example

$$E_\alpha = E_\alpha(x,y,z,t,\theta_x,\theta_y,\theta_z,\theta_t) \tag{28}$$

The imaginary number j will be used to refer to internal phase angles that are associated with broken symmetries, while the imaginary number i will refer to external phase angles. For plane waves the magnitudes of the field vectors in equations (23) through (26) may be written as

$$E_\alpha = A_{E\alpha} e^{i\xi} \tag{29}$$

$$D_\alpha = A_{D\alpha} e^{i\xi} \tag{30}$$

$$H_\alpha = A_{H\alpha} e^{i\xi} \tag{31}$$

$$B_\alpha = A_{B\alpha} e^{i\xi} \tag{32}$$

where

$$\xi = k_x x + k_y y + k_z z - \omega t \tag{32A}$$

where $A_{E\alpha}$, $A_{D\alpha}$, $A_{H\alpha}$, and $A_{B\alpha}$ are constants; k_x , k_y , k_z = wavenumber components, and ω = magnitude of the frequency.

In general the internal phase angles $\theta_{E\alpha}$, $\theta_{D\alpha}$, $\theta_{H\alpha}$, and $\theta_{B\alpha}$ are functions of space and time, and it follows from equations (18) through (22) and (23) through (26) that

$$\partial \bar{E}_\alpha / \partial \bar{\eta} = e^{j\Phi_{E\alpha,\eta}} \cos \beta_{\eta,\eta} \, W_{E\alpha,\eta} \tag{33}$$

$$\partial \bar{D}_\alpha / \partial \bar{\eta} = e^{j\Phi_{D\alpha,\eta}} \cos \beta_{\eta,\eta} \, W_{D\alpha,\eta} \tag{34}$$

$$\partial \bar{H}_\alpha / \partial \bar{\eta} = e^{j\Phi_{H\alpha,\eta}} \cos \beta_{\eta,\eta} \, W_{H\alpha,\eta} \tag{35}$$

$$\partial \bar{B}_\alpha / \partial \bar{\eta} = e^{j\Phi_{B\alpha,\eta}} \cos \beta_{\eta,\eta} \, W_{B\alpha,\eta} \tag{36}$$

where $\eta = t, x, y, z$ and where

$$W_{E\alpha,\eta} = \sqrt{(\partial E_\alpha / \partial \eta)^2 + (E_\alpha \, \partial \theta_{E\alpha} / \partial \eta)^2} \tag{37}$$

$$W_{D\alpha,\eta} = \sqrt{(\partial D_\alpha / \partial \eta)^2 + (D_\alpha \, \partial \theta_{D\alpha} / \partial \eta)^2} \tag{38}$$

$$W_{H\alpha,\eta} = \sqrt{(\partial H_\alpha / \partial \eta)^2 + (H_\alpha \, \partial \theta_{H\alpha} / \partial \eta)^2} \tag{39}$$

$$W_{B\alpha,\eta} = \sqrt{(\partial B_\alpha / \partial \eta)^2 + (B_\alpha \, \partial \theta_{B\alpha} / \partial \eta)^2} \tag{40}$$

$$\Phi_{E\alpha,\eta} = \theta_{E\alpha} + \beta_{E\alpha,\eta} - \theta_{d\eta} \tag{41}$$

$$\Phi_{D\alpha,\eta} = \theta_{D\alpha} + \beta_{D\alpha,\eta} - \theta_{d\eta} \tag{42}$$

$$\Phi_{H\alpha,\eta} = \theta_{H\alpha} + \beta_{H\alpha,\eta} - \theta_{d\eta} \tag{43}$$

$$\Phi_{B\alpha,\eta} = \theta_{B\alpha} + \beta_{B\alpha,\eta} - \theta_{d\eta} \tag{44}$$

$$\tan \beta_{E\alpha,\eta} = E_\alpha \frac{\partial\theta_{E\alpha}/\partial\eta}{\partial E_\alpha/\partial\eta} \tag{45}$$

$$\tan \beta_{D\alpha,\eta} = D_\alpha \frac{\partial\theta_{D\alpha}/\partial\eta}{\partial D_\alpha/\partial\eta} \tag{46}$$

$$\tan \beta_{H\alpha,\eta} = H_\alpha \frac{\partial\theta_{H\alpha}/\partial\eta}{\partial H_\alpha/\partial\eta} \tag{47}$$

$$\tan \beta_{B\alpha,\eta} = B_\alpha \frac{\partial\theta_{B\alpha}/\partial\eta}{\partial B_\alpha/\partial\eta} \tag{48}$$

and where

$$\theta_{d\eta} = \theta_\eta + \beta_{\eta,\eta} \tag{49}$$

If the electric and magnetic fields have an external time dependence given by equations (29) through (32) it follows that

$$\frac{\partial\bar{E}_\alpha}{\partial\bar{t}} = e^{j\theta_{E\alpha,t}} \cos \beta_{t,t} W_{E\alpha,t} \tag{50}$$

and

$$W_{E\alpha,t} = A_{E\alpha} e^{i\xi} \sqrt{-\omega^2 + \left(\frac{\partial\theta_{E\alpha}}{\partial t}\right)^2} = -iA_{E\alpha} e^{i\xi} \sqrt{\omega^2 - \left(\frac{\partial\theta_{E\alpha}}{\partial t}\right)^2} \tag{51}$$

where ξ is given by equation (32A), with similar expressions for the magnetic vector components. Similarly, the derivative with respect to the coordinate $\bar{x}$ is given by

$$\frac{\partial\bar{E}_\alpha}{\partial\bar{x}} = e^{j\theta_{E\alpha,x}} \cos \beta_{x,x} W_{E\alpha,x} \tag{52}$$

where

$$W_{E\alpha,x} = A_{E\alpha} e^{i\xi} \sqrt{-k_x^2 + \left(\frac{\partial\theta_{E\alpha}}{\partial x}\right)^2} = iA_{E\alpha} e^{i\xi} \sqrt{k_x^2 - \left(\frac{\partial\theta_{E\alpha}}{\partial x}\right)^2} \tag{53}$$

with similar expressions for the magnetic field vector components and for the derivatives with respect to $\bar{y}$ and $\bar{z}$.

It is sometimes conventient to write the derivatives in equations (33) and (36) in the following alternative forms

$$\frac{\partial \bar{E}_\alpha}{\partial \bar{\eta}} = \cos \beta_{\eta,\eta}\, e^{-j\theta_{d\eta}} \frac{\partial \bar{E}_\alpha}{\partial \eta} = R_{1\eta}(E_{\alpha R}, E_{\alpha I}) + jI_{1\eta}(E_{\alpha R}, E_{\alpha I}) \tag{54}$$

$$\frac{\partial \bar{B}_\alpha}{\partial \bar{\eta}} = \cos \beta_{\eta,\eta}\, e^{-j\theta_{d\eta}} \frac{\partial \bar{B}_\alpha}{\partial \eta} = R_{1\eta}(B_{\alpha R}, B_{\alpha I}) + jI_{1\eta}(B_{\alpha R}, B_{\alpha I}) \tag{55}$$

where

$$R_{1\eta}(E_{\alpha R}, E_{\alpha I}) = \cos \beta_{\eta,\eta}\, (\cos \theta_{d\eta}\, \partial E_{\alpha R}/\partial \eta + \sin \theta_{d\eta}\, \partial E_{\alpha I}/\partial \eta) \tag{56}$$

$$= \cos \beta_{\eta,\eta} \cos \Phi_{E\alpha,\eta}\, W_{E\alpha,\eta}$$

$$\sim \cos \beta_{\eta,\eta} \cos (\theta_{E\alpha} - \theta_{d\eta})\, \partial E_\alpha/\partial \eta$$

$$I_{1\eta}(E_{\alpha R}, E_{\alpha I}) = \cos \beta_{\eta,\eta}\, (-\sin \theta_{d\eta}\, \partial E_{\alpha R}/\partial \eta + \cos \theta_{d\eta}\, \partial E_{\alpha I}/\partial \eta) \tag{57}$$

$$= \cos \beta_{\eta,\eta} \sin \Phi_{E\alpha,\eta}\, W_{E\alpha,\eta}$$

$$\sim \cos \beta_{\eta,\eta} \sin (\theta_{E\alpha} - \theta_{d\eta})\, \partial E_\alpha/\partial \eta$$

$$R_{1\eta}(B_{\alpha R}, B_{\alpha I}) = \cos \beta_{\eta,\eta}\, (\cos \theta_{d\eta}\, \partial B_{\alpha R}/\partial \eta + \sin \theta_{d\eta}\, \partial B_{\alpha I}/\partial \eta) \tag{58}$$

$$= \cos \beta_{\eta,\eta} \cos \Phi_{B\alpha,\eta}\, W_{B\alpha,\eta}$$

$$\sim \cos \beta_{\eta,\eta} \cos (\theta_{B\alpha} - \theta_{d\eta})\, \partial B_\alpha/\partial \eta$$

$$I_{1\eta}(B_{\alpha R},B_{\alpha I}) = \cos\ \beta_{\eta,\eta}\ (-\sin\ \theta_{d\eta}\ \partial B_{\alpha R}/\partial\eta + \cos\ \theta_{d\eta}\ \partial B_{\alpha I}/\partial\eta) \tag{59}$$

$$= \cos\ \beta_{\eta,\eta}\ \sin\ \Phi_{B\alpha,\eta}\ W_{B\alpha,\eta}$$

$$\sim \cos\ \beta_{\eta,\eta}\ \sin\ (\theta_{B\alpha} - \theta_{d\eta})\ \partial B_{\alpha}/\partial\eta$$

and where

$$E_{\alpha R} = E_{\alpha}\ \cos\ \theta_{E\alpha} \tag{60}$$

$$E_{\alpha I} = E_{\alpha}\ \sin\ \theta_{E\alpha} \tag{61}$$

$$B_{\alpha R} = B_{\alpha}\ \cos\ \theta_{B\alpha} \tag{62}$$

$$B_{\alpha I} = B_{\alpha}\ \sin\ \theta_{B\alpha} \tag{63}$$

The second derivatives of the field vectors are obtained by consecutively applying the first derivative operators that appear in equation (54) as follows

$$\frac{\partial^2\bar{E}_{\alpha}}{\partial\bar{\eta}^2} = \cos\ \beta_{\eta,\eta}\ e^{-j\theta d\eta}\ \frac{\partial}{\partial\eta}\ (\cos\ \beta_{\eta,\eta}\ e^{-j\theta d\eta}\ \frac{\partial\bar{E}_{\alpha}}{\partial\eta}) \tag{64}$$

where $\alpha = x, y, z$ and $\eta = x, y, z, t$. For simplicity it will be assumed that $\beta_{\eta,\eta}$ and $\theta_{d\eta}$ are slowly varying functions of η. Within the limits of this approximation the second derivatives of the field vectors are written as

$$\frac{\partial^2\bar{E}_{\alpha}}{\partial\bar{\eta}^2} \sim \cos^2\ \beta_{\eta,\eta}\ e^{-2j\theta d\eta}\ \frac{\partial^2\bar{E}_{\alpha}}{\partial\eta^2} = R_{2\eta}(E_{\alpha R},E_{\alpha I}) + jI_{2\eta}(E_{\alpha R},E_{\alpha I}) \tag{65}$$

$$\sim \cos^2\ \beta_{\eta,\eta}\ e^{j\Psi E\alpha,\eta}\ T_{E\alpha,\eta}$$

$$\sim \cos^2\ \beta_{\eta,\eta}\ e^{j(\theta_{E\alpha} - 2\theta_{d\eta})}\ \frac{\partial^2 E_{\alpha}}{\partial\eta^2}$$

$$\frac{\partial^2 \bar{H}_\alpha}{\partial \bar{\eta}^2} \sim \cos^2 \beta_{\eta,\eta}\, e^{-2j\theta_{d\eta}} \frac{\partial^2 \bar{H}_\alpha}{\partial \eta^2} = R_{2\eta}(H_{\alpha R}, H_{\alpha I}) + jI_{2\eta}(H_{\alpha R}, H_{\alpha I}) \tag{66}$$

$$\sim \cos^2 \beta_{\eta,\eta}\, e^{j\Psi_{H\alpha,\eta}}\, T_{H\alpha,\eta}$$

$$\sim \cos^2 \beta_{\eta,\eta}\, e^{j(\theta_{H\alpha} - 2\theta_{d\eta})} \frac{\partial^2 H_\alpha}{\partial \eta^2}$$

where

$$T_{E\alpha,\eta} = \sqrt{\left(\frac{\partial^2 E_\alpha}{\partial \eta^2}\right)^2 + \left(E_\alpha \frac{\partial^2 \theta_{E\alpha}}{\partial \eta^2}\right)^2} \tag{66A}$$

$$T_{H\alpha,\eta} = \sqrt{\left(\frac{\partial^2 H_\alpha}{\partial \eta^2}\right)^2 + \left(H_\alpha \frac{\partial^2 \theta_{H\alpha}}{\partial \eta^2}\right)^2} \tag{66B}$$

$$\Psi_{E\alpha,\eta} = \theta_{E\alpha} + \delta_{E\alpha,\eta} - 2\theta_{d\eta} \tag{66C}$$

$$\Psi_{H\alpha,\eta} = \theta_{H\alpha} + \delta_{H\alpha,\eta} - 2\theta_{d\eta} \tag{66D}$$

and where

$$R_{2\eta}(E_{\alpha R}, E_{\alpha I}) = \cos^2 \beta_{\eta,\eta} \left[\cos(2\theta_{d\eta}) \frac{\partial^2 E_{\alpha R}}{\partial \eta^2} + \sin(2\theta_{d\eta}) \frac{\partial^2 E_{\alpha I}}{\partial \eta^2}\right] \tag{67}$$

$$\sim \cos^2 \beta_{\eta,\eta} \cos \Psi_{E\alpha,\eta}\, T_{E\alpha,\eta}$$

$$\sim \cos^2 \beta_{\eta,\eta} \cos(\theta_{E\alpha} - 2\theta_{d\eta}) \frac{\partial^2 E_\alpha}{\partial \eta^2}$$

$$I_{2\eta}(E_{\alpha R},E_{\alpha I}) = \cos^2 \beta_{\eta,\eta} \left[- \sin (2\theta_{d\eta}) \frac{\partial^2 E_{\alpha R}}{\partial \eta^2} + \cos (2\theta_{d\eta}) \frac{\partial^2 E_{\alpha I}}{\partial \eta^2} \right] \tag{68}$$

$$\sim \cos^2 \beta_{\eta,\eta} \sin \Psi_{E\alpha,\eta} T_{E\alpha,\eta}$$

$$\sim \cos^2 \beta_{\eta,\eta} \sin (\theta_{E\alpha} - 2\theta_{d\eta}) \frac{\partial^2 E_{\alpha}}{\partial \eta^2}$$

$$R_{2\eta}(H_{\alpha R},H_{\alpha I}) = \cos^2 \beta_{\eta,\eta} \left[\cos (2\theta_{d\eta}) \frac{\partial^2 H_{\alpha R}}{\partial \eta^2} + \sin (2\theta_{d\eta}) \frac{\partial^2 H_{\alpha I}}{\partial \eta^2} \right] \tag{69}$$

$$\sim \cos^2 \beta_{\eta,\eta} \cos \Psi_{H\alpha,\eta} T_{H\alpha,\eta}$$

$$\sim \cos^2 \beta_{\eta,\eta} \cos (\theta_{H\alpha} - 2\theta_{d\eta}) \frac{\partial^2 H_{\alpha}}{\partial \eta^2}$$

$$I_{2\eta}(H_{\alpha R},H_{\alpha I}) = \cos^2 \beta_{\eta,\eta} \left[- \sin (2\theta_{d\eta}) \frac{\partial^2 H_{\alpha R}}{\partial \eta^2} + \cos (2\theta_{d\eta}) \frac{\partial^2 H_{\alpha I}}{\partial \eta^2} \right] \tag{70}$$

$$\sim \cos^2 \beta_{\eta,\eta} \sin \Psi_{H\alpha,\eta} T_{H\alpha,\eta}$$

$$\sim \cos^2 \beta_{\eta,\eta} \sin (\theta_{H\alpha} - 2\theta_{d\eta}) \frac{\partial^2 H_{\alpha}}{\partial \eta^2}$$

and

$$\tan \delta_{E\alpha,\eta} = E_{\alpha} \frac{\partial^2 \theta_{E\alpha}}{\partial \eta^2} \Big/ \frac{\partial^2 E_{\alpha}}{\partial \eta^2} \tag{70A}$$

$$\tan \delta_{H\alpha,\eta} = H_{\alpha} \frac{\partial^2 \theta_{H\alpha}}{\partial \eta^2} \Big/ \frac{\partial^2 H_{\alpha}}{\partial \eta^2} \tag{70B}$$

Similar equations can be written for the D_α and B_α components. In this way the derivatives necessary for evaluating Maxwell's equations in asymmetric bulk matter or vacuum can be evaluated.

3. MAXWELL'S EQUATIONS WITH BROKEN INTERNAL SYMMETRIES. This section develops Maxwell's equations for electromagnetic fields whose field vectors have internal phase angles. The asymmetric bulk matter or vacuum in which the electromagnetic fields exist also have broken internal symmetries in the static pressure field, Grüneisen parameter, and in the space and time coordinates of each point in the system. The internal phase angles of the ambient medium such as θ_P, θ_γ, θ_x, θ_y, θ_z and θ_t must be calculated in conjunction with the internal phase angles of the electromagnetic field vectors. The quantities of θ_P and θ_γ are obtained from the ground state relativistic trace equation (1) and equation (4) that defines the relativistic Grüneisen function.

The unrenormalized Maxwell equations for charges and currents are written as[9-17]

$$\vec{\nabla}_a \cdot \vec{B}^a = 0 \tag{71}$$

$$\vec{\nabla}_a \cdot \vec{D}^a = \rho_q^a \tag{72}$$

$$\vec{\nabla}_a \times \vec{H}^a = \partial\vec{D}^a/\partial t_a + \vec{j}^a \tag{73}$$

$$\vec{\nabla}_a \times \vec{E}^a = -\partial\vec{B}^a/\partial t_a \tag{74}$$

where $\vec{B}^a$ = unrenormalized magnetic induction vector, $\vec{D}^a$ = unrenormalized electric displacement vector, ρ_q^a = unrenormalized charge density, $\vec{H}^a$ = unrenormalized magnetic field vector, $\vec{j}^a$ = unrenormalized current density vector, and $\vec{E}^a$ = unrenormalized electric field vector. Equations (71) through (74) represent eight equations, six of which are independent. The simplest constitutive equations are the following

$$\vec{B}^a = \mu^a\vec{H}^a \tag{75}$$

$$\vec{D}^a = \varepsilon^a\vec{E}^a \tag{76}$$

where μ^a = unrenormalized magnetic permeability, and ε^a = unrenormalized dielectric constant (permittivity). More complicated constitutive equations are often used.[18] In general

$$\mu^a = \mu^a(P^a,\gamma^a) \qquad \varepsilon^a = \varepsilon^a(P^a,\gamma^a) \tag{77}$$

where P^a and γ^a are functions of density and temperature.

Within asymmetric bulk matter or vacuum a similar set of Maxwell's equations must be valid except now the renormalized electric and magnetic field vectors must have internal phases, and the space and time coordinates must also have internal phase angles. Therefore equations (71) through (74) can be written for the electromagnetic field in bulk matter or vacuum with broken internal symmetries as follows

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{78}$$

$$\vec{\nabla} \cdot \vec{D} = \rho_q \tag{79}$$

$$\vec{\nabla} \times \vec{H} = \partial\vec{D}/\partial t + \vec{j} \tag{80}$$

$$\vec{\nabla} \times \vec{E} = -\partial\vec{B}/\partial t \tag{81}$$

where $\vec{B}$ = renormalized complex number magnetic induction vector, $\vec{D}$ = renormalized complex number electric displacement vector, ρ_q = renormalized charge density, $\vec{H}$ = renormalized complex number magnetic field vector, $\vec{j}$ = renormalized complex number current density vector, and $\vec{E}$ = renormalized complex number electric field vector. Equations (78) through (81) are complex number vector equations and represent a total of sixteen equations, twelve of which are independent. Note that $\rho_q \neq \rho_q^a$ as can be seen from equations (72) and (79) on account of $\vec{D} \neq \vec{D}^a$. Since $\rho_q = nq$ and $\rho_q^a = nq^a$, where q and q^a = renormalized and unrenormalized charge per particle, it follows that $q \neq q^a$. The renormalized current density is given by $\vec{j} = nq\vec{v}$, where $\vec{v}$ = vector particle velocity with internal phase. The internal phase angle of the particle velocity is a function of θ_x, θ_y, θ_z, θ_t and clearly $\vec{j} \neq \vec{j}^a$.

The simplest renormalized constitutive equations are written as

$$\vec{B} = \mu\vec{H} \tag{82}$$

$$\vec{D} = \varepsilon\vec{E} \tag{83}$$

where μ = renormalized magnetic permeability, and ε = renormalized permittivity. Taking account of the fact that equations (82) and (83) are vector equations with real and imaginary parts, it is clear that they represent twelve equations.

The state equations for μ and ε will be assumed to be given by

$$\mu = \mu^a(P,\gamma) \tag{84}$$

$$\varepsilon = \varepsilon^a(P,\gamma) \tag{85}$$

where the functions are evaluated at the values of the renormalized pressure and Grüneisen function as determined from a solution of equation (1).

The component form of the renormalized Maxwell equations (78) through (81) are written as

$$\partial\bar{B}_x/\partial\bar{x} + \partial\bar{B}_y/\partial\bar{y} + \partial\bar{B}_z/\partial\bar{z} = 0 \tag{86}$$

$$\partial\bar{D}_x/\partial\bar{x} + \partial\bar{D}_y/\partial\bar{y} + \partial\bar{D}_z/\partial\bar{z} = \rho_q \tag{87}$$

$$\partial\bar{H}_y/\partial\bar{x} - \partial\bar{H}_x/\partial\bar{y} = \partial\bar{D}_z/\partial\bar{t} + \bar{j}_z \tag{88}$$

$$\partial\bar{H}_z/\partial\bar{y} - \partial\bar{H}_y/\partial\bar{z} = \partial\bar{D}_x/\partial\bar{t} + \bar{j}_x \tag{89}$$

$$\partial\bar{H}_x/\partial\bar{z} - \partial\bar{H}_z/\partial\bar{x} = \partial\bar{D}_y/\partial\bar{t} + \bar{j}_y \tag{90}$$

$$\partial\bar{E}_y/\partial\bar{x} - \partial\bar{E}_x/\partial\bar{y} = - \partial\bar{B}_z/\partial\bar{t} \tag{91}$$

$$\partial\bar{E}_z/\partial\bar{y} - \partial\bar{E}_y/\partial\bar{z} = - \partial\bar{B}_x/\partial\bar{t} \tag{92}$$

$$\partial\bar{E}_x/\partial\bar{z} - \partial\bar{E}_z/\partial\bar{x} = - \partial\bar{B}_y/\partial\bar{t} \tag{93}$$

Using equations (33) through (49) to evaluate derivatives allows Maxwell's equations to be rewritten as

$$W_{Bx,x} \cos \Phi_{Bx,x} \cos \beta_{x,x} + W_{By,y} \cos \Phi_{By,y} \cos \beta_{y,y} \tag{94}$$

$$+ W_{Bz,z} \cos \Phi_{Bz,z} \cos \beta_{z,z} = 0$$

$$W_{Bx,x} \sin \Phi_{Bx,x} \cos \beta_{x,x} + W_{By,y} \sin \Phi_{By,y} \cos \beta_{y,y} \tag{95}$$

$$+ W_{Bz,z} \sin \Phi_{Bz,z} \cos \beta_{z,z} = 0$$

$$W_{Dx,x} \cos \Phi_{Dx,x} \cos \beta_{x,x} + W_{Dy,y} \cos \Phi_{Dy,y} \cos \beta_{y,y} \tag{96}$$

$$+ W_{Dz,z} \cos \Phi_{Dz,z} \cos \beta_{z,z} = \rho_q$$

$$W_{Dx,x} \sin \Phi_{Dx,x} \cos \beta_{x,x} + W_{Dy,y} \sin \Phi_{Dy,y} \cos \beta_{y,y} \tag{97}$$

$$+ W_{Dz,z} \sin \Phi_{Dz,z} \cos \beta_{z,z} = 0$$

$$W_{Hy,x} \cos \Phi_{Hy,x} \cos \beta_{x,x} - W_{Hx,y} \cos \Phi_{Hx,y} \cos \beta_{y,y} \tag{98}$$

$$= W_{Dz,t} \cos \Phi_{Dz,t} \cos \beta_{t,t} + j_z \cos \theta_{jz}$$

$$W_{Hy,x} \sin \Phi_{Hy,x} \cos \beta_{x,x} - W_{Hx,y} \sin \Phi_{Hx,y} \cos \beta_{y,y} \tag{99}$$

$$= W_{Dz,t} \sin \Phi_{Dz,t} \cos \beta_{t,t} + j_z \sin \theta_{jz}$$

$$W_{Hz,y} \cos \Phi_{Hz,y} \cos \beta_{y,y} - W_{Hy,z} \cos \Phi_{Hy,z} \cos \beta_{z,z} \tag{100}$$

$$= W_{Dx,t} \cos \Phi_{Dx,t} \cos \beta_{t,t} + j_x \cos \theta_{jx}$$

$$W_{Hz,y} \sin \Phi_{Hz,y} \cos \beta_{y,y} - W_{Hy,z} \sin \Phi_{Hy,z} \cos \beta_{z,z} \tag{101}$$

$$= W_{Dx,t} \sin \Phi_{Dx,t} \cos \beta_{t,t} + j_x \sin \theta_{jx}$$

$$W_{Hx,z} \cos \Phi_{Hx,z} \cos \beta_{z,z} - W_{Hz,x} \cos \Phi_{Hz,x} \cos \beta_{x,x} \tag{102}$$

$$= W_{Dy,t} \cos \Phi_{Dy,t} \cos \beta_{t,t} + j_y \cos \theta_{jy}$$

$$W_{Hx,z} \sin \Phi_{Hx,z} \cos \beta_{z,z} - W_{Hz,x} \sin \Phi_{Hz,x} \cos \beta_{x,x} \tag{103}$$

$$= W_{Dy,t} \sin \Phi_{Dy,t} \cos \beta_{t,t} + j_y \sin \theta_{jy}$$

$$W_{Ey,x} \cos \Phi_{Ey,x} \cos \beta_{x,x} - W_{Ex,y} \cos \Phi_{Ex,y} \cos \beta_{y,y} \tag{104}$$

$$= - W_{Bz,t} \cos \Phi_{Bz,t} \cos \beta_{t,t}$$

$$W_{Ey,x} \sin \Phi_{Ey,x} \cos \beta_{x,x} - W_{Ex,y} \sin \Phi_{Ex,y} \cos \beta_{y,y} \tag{105}$$

$$= - W_{Bz,t} \sin \Phi_{Bz,t} \cos \beta_{t,t}$$

$$W_{Ez,y} \cos \Phi_{Ez,y} \cos \beta_{y,y} - W_{Ey,z} \cos \Phi_{Ey,z} \cos \beta_{z,z} \tag{106}$$

$$= - W_{Bx,t} \cos \Phi_{Bx,t} \cos \beta_{t,t}$$

$$W_{Ez,y} \sin \Phi_{Ez,y} \cos \beta_{y,y} - W_{Ey,z} \sin \Phi_{Ey,z} \cos \beta_{z,z} \tag{107}$$

$$= - W_{Bx,t} \sin \Phi_{Bx,t} \cos \beta_{t,t}$$

$$W_{Ex,z} \cos \Phi_{Ex,z} \cos \beta_{z,z} - W_{Ez,x} \cos \Phi_{Ez,x} \cos \beta_{x,x} \tag{108}$$

$$= - W_{By,t} \cos \Phi_{By,t} \cos \beta_{t,t}$$

$$W_{Ex,z} \sin \phi_{Ex,z} \cos \beta_{z,z} - W_{Ez,x} \sin \phi_{Ez,x} \cos \beta_{x,x} \tag{109}$$

$$= - W_{By,t} \sin \phi_{By,t} \cos \beta_{t,t}$$

Maxwell's equations can be written in an alternative but equivalent form by using equations (54) through (59) as follows

$$R_{lx}(B_{xR},B_{xI}) + R_{ly}(B_{yR},B_{yI}) + R_{lz}(B_{zR},B_{zI}) = 0 \tag{110}$$

$$I_{lx}(B_{xR},B_{xI}) + I_{ly}(B_{yR},B_{yI}) + I_{lz}(B_{zR},B_{zI}) = 0 \tag{111}$$

$$R_{lx}(D_{xR},D_{xI}) + R_{ly}(D_{yR},D_{yI}) + R_{lz}(D_{zR},D_{zI}) = \rho_q \tag{112}$$

$$I_{lx}(D_{xR},D_{xI}) + I_{ly}(D_{yR},D_{yI}) + I_{lz}(D_{zR},D_{zI}) = 0 \tag{113}$$

$$R_{lx}(H_{yR},H_{yI}) - R_{ly}(H_{xR},H_{xI}) = R_{lt}(D_{zR},D_{zI}) + j_{zR} \tag{114}$$

$$I_{lx}(H_{yR},H_{yI}) - I_{ly}(H_{xR},H_{xI}) = I_{lt}(D_{zR},D_{zI}) + j_{zI} \tag{115}$$

$$R_{ly}(H_{zR},H_{zI}) - R_{lz}(H_{yR},H_{yI}) = R_{lt}(D_{xR},D_{xI}) + j_{xR} \tag{116}$$

$$I_{ly}(H_{zR},H_{zI}) - I_{lz}(H_{yR},H_{yI}) = I_{lt}(D_{xR},D_{xI}) + j_{xI} \tag{117}$$

$$R_{lz}(H_{xR},H_{xI}) - R_{lx}(H_{zR},H_{zI}) = R_{lt}(D_{yR},D_{yI}) + j_{yR} \tag{118}$$

$$I_{lz}(H_{xR},H_{xI}) - I_{lx}(H_{zR},H_{zI}) = I_{lt}(D_{yR},D_{yI}) + j_{yI} \tag{119}$$

$$R_{lx}(E_{yR},E_{yI}) - R_{ly}(E_{xR},E_{xI}) = - R_{lt}(B_{zR},B_{zI}) \tag{120}$$

$$I_{lx}(E_{yR},E_{yI}) - I_{ly}(E_{xR},E_{xI}) = - I_{lt}(B_{zR},B_{zI}) \tag{121}$$

$$R_{1y}(E_{zR},E_{zI}) - R_{1z}(E_{yR},E_{yI}) = - R_{1t}(B_{xR},B_{xI}) \tag{122}$$

$$I_{1y}(E_{zR},E_{zI}) - I_{1z}(E_{yR},E_{yI}) = - I_{1t}(B_{xR},B_{xI}) \tag{123}$$

$$R_{1z}(E_{xR},E_{xI}) - R_{1x}(E_{zR},E_{zI}) = - R_{1t}(B_{yR},B_{yI}) \tag{124}$$

$$I_{1z}(E_{xR},E_{xI}) - I_{1x}(E_{zR},E_{zI}) = - I_{1t}(B_{yR},B_{yI}) \tag{125}$$

where

$$j_{zR} = j_z \cos \theta_{jz} \qquad j_{zI} = j_z \sin \theta_{jz} \tag{126}$$

$$j_{xR} = j_x \cos \theta_{jx} \qquad j_{xI} = j_x \sin \theta_{jx} \tag{127}$$

$$j_{yR} = j_y \cos \theta_{jy} \qquad j_{yI} = j_y \sin \theta_{jy} \tag{128}$$

The radiation pressure is related to the radiation energy density for isotropic radiation with broken internal symmetry by the following approximate formula

$$\bar{P}_r = P_r e^{j\theta Pr} = \frac{1}{3} \bar{E}_r = \frac{1}{3} E_r e^{j\theta Er} \tag{129}$$

Equation (129) is exact only for symmetrical isotropic radiation. Equation (129) gives the following approximate equations

$$P_r = \frac{1}{3} E_r \tag{130}$$

$$\theta_{Pr} = \theta_{Er} \tag{131}$$

where $\bar{E}_r$ = radiation energy density with broken internal symmetry, that is related to the electromagnetic field vectors as follows

$$\bar{E}_r = \frac{\varepsilon}{2} (\bar{E}_x^2 + \bar{E}_y^2 + \bar{E}_z^2) + \frac{\mu}{2} (\bar{H}_x^2 + \bar{H}_y^2 + \bar{H}_z^2) \tag{132}$$

Equation (132) is equivalent to the following two equations

$$\mathcal{E}_r \cos \theta_{\mathcal{E}r} = \frac{\varepsilon}{2}\left[E_x^2 \cos (2\theta_{Ex}) + E_y^2 \cos (2\theta_{Ey}) + E_z^2 \cos (2\theta_{Ez})\right] \qquad (133)$$

$$+ \frac{\mu}{2}\left[H_x^2 \cos (2\theta_{Hx}) + H_y^2 \cos (2\theta_{Hy}) + H_z^2 \cos (2\theta_{Hz})\right]$$

$$\mathcal{E}_r \sin \theta_{\mathcal{E}r} = \frac{\varepsilon}{2}\left[E_x^2 \sin (2\theta_{Ex}) + E_y^2 \sin (2\theta_{Ey}) + E_z^2 \sin (2\theta_{Ez})\right] \qquad (134)$$

$$+ \frac{\mu}{2}\left[H_x^2 \sin (2\theta_{Hx}) + H_y^2 \sin (2\theta_{Hy}) + H_z^2 \sin (2\theta_{Hz})\right]$$

from which $\mathcal{E}_r$ and $\theta_{\mathcal{E}r}$ can be immediately obtained. The unrenormalized radiation density is given by equation (132) with the bars removed and with the superscript "a" inserted on all quantities.

In addition to Maxwell's equations several other equations are required to form a complete set of equations to determine the phase angles of the spacetime coordinates as well as the phase angles of the electromagnetic field vectors. Six of the additional equations required are the equations of motion for charged bulk matter (plasma). These six equations are given by the complex number vector nonrelativistic Euler equations combined with the Lorentz force as follows[19-22]

$$\rho\, d\bar{v}_\alpha/d\bar{t} = -\partial\bar{P}/\partial\bar{\alpha} - \partial\bar{P}_r/\partial\bar{\alpha} - \partial\bar{W}/\partial\bar{\alpha} + \rho_q(\vec{\bar{E}} + \vec{\bar{v}} \times \vec{\bar{B}})_\alpha \qquad (135)$$

where $\alpha = x, y, z$, ρ = mass density, $\bar{v}_\alpha$ = spatial components of particle velocity with internal phase, $\bar{P}$ = static pressure with internal phase, $\bar{P}_r$ = radiation pressure with internal phase, $\bar{W}$ = external potential (such as gravity) with internal phase, and $\vec{\bar{v}}$ = particle velocity vector with internal phase = $(\bar{v}_x, \bar{v}_y, \bar{v}_z)$. The static pressure and external potential are complex numbers with internal phase angles and are written as

$$\bar{P} = Pe^{j\theta P} \qquad (136)$$

$$\bar{W} = We^{j\theta W} \qquad (137)$$

The time derivatives of the velocity components in equation (135) are given by the following six equations

$$\bar{a}_\alpha = d\bar{v}_\alpha/d\bar{t} = \partial\bar{v}_\alpha/\partial\bar{t} + \sum_\sigma \bar{v}_\sigma\, \partial\bar{v}_\alpha/\partial\bar{\sigma} \qquad (138)$$

where the following six equations define the complex number velocity

$$\bar{v}_\alpha = d\bar{\alpha}/d\bar{t} = v_\alpha e^{j\theta v\alpha} \tag{139}$$

$$v_\alpha = \sqrt{\frac{(d\alpha/dt)^2 + (\alpha\, d\theta_\alpha/dt)^2}{1 + (t\, d\theta_t/dt)^2}} \tag{140}$$

$$\theta_{v\alpha} = \theta_\alpha + \beta_{\alpha,t} - \theta_t - \beta_{t,t} \tag{141}$$

where

$$\tan \beta_{\alpha,t} = \alpha \frac{d\theta_\alpha/dt}{d\alpha/dt} \tag{142}$$

$$\frac{d\theta_\alpha}{dt} = \frac{\partial\theta_\alpha}{\partial t} + \frac{\partial\theta_\alpha}{\partial x}\frac{dx}{dt} + \frac{\partial\theta_\alpha}{\partial y}\frac{dy}{dt} + \frac{\partial\theta_\alpha}{\partial z}\frac{dz}{dt} \tag{143}$$

In addition, the continuity equation

$$\partial\rho_q/\partial\bar{t} + \vec{\bar{\nabla}} \cdot (\rho_q \vec{\bar{v}}) = 0 \tag{144A}$$

is necessary to determine θ_t and $\rho_q(x,y,z,t)$. Equation (144A) has two components because it is a complex number scalar equation.

The Maxwell equations for broken symmetry matter, equations (94) through (109) or equivalently equations (110) through (125), are not sufficient by themselves to determine the internal phase angles of the space and time coordinates. This is because the twelve independent Maxwell equations (88) through (93), the twelve constitutive equations (82) and (83), the two components of the ground state trace equation (1), the two components of the ground state Grüneisen parameter equation (4), the two components of the excited state trace equation (5), the two components of the radiation Grüneisen parameter equation (6A), the two state equations (84) and (85) for the renormalized magnetic permeability and electric permittivity, and the two components of the continuity equation, represent thirty-six equations. However, thus far only thirty-five field and matter variables have been enumerated and these are: E_α, $\theta_{E\alpha}$; H_α, $\theta_{H\alpha}$; B_α, $\theta_{B\alpha}$; E, θ_E; γ, θ_γ; E_r, θ_{Er}; γ_r, $\theta_{\gamma r}$; ε, μ; and ρ_q. But these thirty-five quantities are related to nineteen kinematic and dynamic variables because of the space and time derivatives of the field vectors in Maxwell's equations (33) through (48) and because of the appearance of the current density (velocity) in Maxwell's equations. The nineteen kinematic and dynamic variables are: x, y, z, v_x, v_y, v_z, a_x, a_y, a_z and the corresponding phase angles θ_x, θ_y, θ_z, θ_{vx}, θ_{vy}, θ_{vz},

ϑ_{ax}, ϑ_{ay}, ϑ_{az} and by itself ϑ_t. In these calculations the magnitude of the time t is taken to be a free and independent variable. Therefore a total of fifty-four unknown quantities need to be calculated, and thus far only thirty-six equations have been enumerated. The additional necessary eighteen equations are: the six equations of motion (135), the six kinematic acceleration equations (138), and the six kinematic velocity equations (139). Thus there are fifty-four equations and fifty-four unknown variables to be determined. Note that the two components of the complex number scalar continuity equation introduces only one new unknown quantity ρ_q, and this leaves the second component equation to determine θ_t which stands by itself because t is taken to be an independent variable.

The relativistic trace equations (1) and (5) play an important part in the calculation of the renormalized electromagnetic fields in asymmetric bulk matter or vacuum. Starting with the unrenormalized ground state energy density and Grüneisen parameter, E^a and γ^a respectively, equation (1) is used to calculate the renormalized values of the ground state energy density E, ϑ_E and the ground state Grüneisen parameter γ, ϑ_γ. The renormalized values of magnetic permeability μ and dielectric constant ϵ are expressed in terms of P and γ through equations (84) and (85). In addition to Maxwell's equations, the radiation trace equation (5), in conjunction with equations (133) and (134) that relate the radiation energy density to the electromagnetic field vectors, determines the renormalized field vectors in terms of the corresponding unrenormalized values. The solution of the unrenormalized Maxwell equations (71) through (74) gives the unrenormalized field vectors in terms of the unrenormalized charge density ρ_q^a and current density j^a. The unrenormalized energy density is then calculated in terms of the unrenormalized field vectors using equations (133) and (134). Then equation (5) is applied again in conjuction with equations (133) and (134) and Maxwell's equations to obtain the renormalized field vectors. Finally the renormalized charge and current density are obtained from the renormalized field vectors by using equations (79) and (80).

For electromagnetic waves in the vacuum, the total density ρ and the charge density ρ_q that appear in equations (135) refer to test charges placed within the vacuum to measure the electromagnetic field strengths. Therefore the case of electromagnetic waves in the vacuum is formally equivalent to the case of electromagnetic waves in bulk matter. For the vacuum

$$
\begin{array}{lll}
\vartheta_x = \vartheta_x^{(v)} & \vartheta_y = \vartheta_y^{(v)} & \vartheta_z = \vartheta_z^{(v)} \\
\vartheta_{vx} = \vartheta_{vx}^{(v)} & \vartheta_{vy} = \vartheta_{vy}^{(v)} & \vartheta_{vz} = \vartheta_{vz}^{(v)} \\
\vartheta_{ax} = \vartheta_{ax}^{(v)} & \vartheta_{ay} = \vartheta_{ay}^{(v)} & \vartheta_{az} = \vartheta_{az}^{(v)} \\
\vartheta_t = \vartheta_t^{(v)} & \vartheta_P = \vartheta_P^{(v)} & \vartheta_U = \vartheta_U^{(v)}
\end{array}
\qquad (145)
$$

where (v) refers to the vacuum state (see Section 6).

The energy conservation equation is the first integral of equation (135) and in its simplest form is written as[19]

$$\frac{1}{2}\rho\bar{v}^2 + \bar{P} + \bar{P}_r + \bar{W} - \rho_q \int \vec{\bar{E}} \cdot \vec{\bar{v}}\, d\bar{t} = \text{constant} \tag{146}$$

where $\vec{\bar{v}}$ = complex number vector velocity whose complex number magnitude is given by

$$\bar{v}^2 = \bar{v}_x^2 + \bar{v}_y^2 + \bar{v}_z^2 \tag{147}$$

where

$$\bar{v} = ve^{j\theta_v} \tag{147A}$$

$$\bar{v}_\alpha = v_\alpha e^{j\theta_{v\alpha}} \tag{147B}$$

Note that

$$\vec{\bar{E}} \cdot \vec{\bar{v}} = \sum_{\alpha=1}^{3} \bar{E}_\alpha \bar{v}_\alpha \tag{148}$$

Were it possible to neglect the charge density term by making ρ_q vanishingly small, equation (146) becomes

$$\frac{1}{2}\rho\bar{v}^2 + \bar{P}_r + \bar{P} + \bar{W} = \text{constant} \tag{149}$$

where the mass density ρ refers to a test probe. Finally, the dynamical equations for relativistic bulk matter with broken internal symmetry are given by the following generalization of equation (135)[23]

$$[\rho + (\bar{P} + \bar{P}_r)/c^2]\bar{\gamma}^2 \frac{d\bar{v}_\alpha}{d\bar{t}} = -\left[\frac{\partial \bar{P}}{\partial \bar{\alpha}} + \frac{\partial \bar{P}_r}{\partial \bar{\alpha}} + \bar{v}_\alpha\left(\frac{\partial \bar{P}}{\partial \bar{t}} + \frac{\partial \bar{P}_r}{\partial \bar{t}}\right)\right] \tag{150}$$

$$- \frac{\partial \bar{W}}{\partial \bar{\alpha}} + \rho_q \bar{\gamma}\left(\vec{\bar{E}} + \vec{\bar{v}} \times \vec{\bar{B}}\right)_\alpha$$

where $\bar{\gamma}$ = complex velocity factor that is defined in Section 4, and where c = light speed in the vacuum.

4. LORENTZ INVARIANCE IN ASYMMETRIC BULK MATTER AND VACUUM. This section considers the Lorentz invariance of Maxwell's equations in bulk matter and vacuum with broken internal symmetries. Maxwell's equations for symmetric systems, equations (71) through (74), are invariant under the Lorentz transformation of coordinate systems[9,10,16,23-34]

$$x'_a = \gamma_a(x_a - v_a t_a) \tag{151}$$

$$t'_a = \gamma_a(t_a - v_a x_a/c^2) \tag{152}$$

where v_a = relative speed of coordinate systems, and the standard velocity factor is given by

$$\gamma_a = (1 - \beta_a^2)^{-1/2} \tag{153}$$

where $\beta_a = v_a/c$, where c = light speed in vacuum. The Lorentz transformation can be obtained by requiring the form invariance of the Minkowski metric as follows

$$x_a^2 - c^2 t_a^2 = x_a'^2 - c^2 t_a'^2 \tag{154}$$

General relativity, which is not considered in this paper, uses a Riemann metric.[23-25]

The form of Maxwell's equations for charges and currents in asymmetric bulk matter or vacuum, equations (78) through (81), is the same as that for symmetric bulk matter or vacuum, equations (71) through (74). The only difference is that in asymmetric systems the field vectors, current density, and spacetime coordinates are complex numbers. Therefore by the same analysis that shows the symmetric Maxwell equations (71) through (74) to be form covariant under the real number Lorentz transformation equations (151) through (153), it follows that the asymmetric Maxwell equations (78) through (81) are form covariant under the following complex number Lorentz transformations

$$\bar{x}' = \bar{\gamma}(\bar{x} - \bar{v}\bar{t}) \tag{155}$$

$$\bar{t}' = \bar{\gamma}(\bar{t} - \bar{v}\bar{x}/c^2) \tag{156}$$

where $\bar{v}$ = complex number relative speed of the two coordinate systems, and the complex number velocity factor for an asymmetric system is given by

$$\bar{\gamma} = (1 - \bar{\beta}^2)^{-1/2} \tag{157}$$

where $\bar{\beta} = \bar{v}/c$. Also, simple algebra shows that equations (155) through (157) satisfy

$$\bar{x}'^2 - c^2\bar{t}'^2 = \bar{x}^2 - c^2\bar{t}^2 \tag{158}$$

where $\bar{x}$, $\bar{t}$, $\bar{x}'$ and $\bar{t}'$ are space and time coordinates that exhibit broken internal symmetry, and in general

$$\bar{x} = xe^{j\theta_x} \qquad \bar{x}' = x'e^{j\theta'_x} \tag{159}$$

$$\bar{t} = te^{j\theta_t} \qquad \bar{t}' = t'e^{j\theta'_t} \tag{160}$$

and the relative speed of the coordinate systems is written as

$$\bar{v} = ve^{j\theta_v} \qquad \bar{\beta} = \beta e^{j\theta_v} \tag{161}$$

where θ_x , θ'_x , θ_t , θ'_t and θ_v are functions of P and θ_p of the ambient asymmetric bulk matter or vacuum.

Combining equations (157) and (161) gives

$$\bar{\gamma} = (f - jb)^{-1/2} = \sqrt{\frac{f + jb}{f^2 + b^2}} = \gamma e^{j\theta_\gamma} \tag{162}$$

where

$$f = 1 - \beta^2 \cos(2\theta_v) \tag{163}$$

$$b = \beta^2 \sin(2\theta_v) \tag{164}$$

From equation (162) it follows that for an asymmetric system the magnitude and internal phase angle of the velocity factor are given by

$$\gamma = (f^2 + b^2)^{-1/4} = [1 - 2\beta^2 \cos(2\theta_v) + \beta^4]^{-1/4} \tag{165}$$

$$\tan(2\theta_\gamma) = \frac{\beta^2 \sin(2\theta_v)}{1 - \beta^2 \cos(2\theta_v)} \tag{166}$$

Note that $\beta = v/c$, where now v = magnitude of the complex number velocity that appears in equation (161). Also, if $\theta_v = 0$ then equation (165) reduces to equation (153).

The Lorentz transformations in equations (155) and (156) can be written in the form of real and imaginary components as follows

$$x' \cos \theta'_x = \gamma[x \cos (\theta_x + \theta_\gamma) - vt \cos (\theta_v + \theta_t + \theta_\gamma)] \tag{167}$$

$$x' \sin \theta'_x = \gamma[x \sin (\theta_x + \theta_\gamma) - vt \sin (\theta_v + \theta_t + \theta_\gamma)] \tag{168}$$

$$t' \cos \theta'_t = \gamma[t \cos (\theta_t + \theta_\gamma) - vx/c^2 \cos (\theta_v + \theta_x + \theta_\gamma)] \tag{169}$$

$$t' \sin \theta'_t = \gamma[t \sin (\theta_t + \theta_\gamma) - vx/c^2 \sin (\theta_v + \theta_x + \theta_\gamma)] \tag{170}$$

where γ is given by equation (165). From equations (167) and (168) x' and θ'_x can be calculated as follows

$$x'^2 = \gamma^2[x^2 + v^2t^2 - 2vtx \cos (\theta_t + \theta_v - \theta_x)] \tag{171}$$

$$\tan \theta'_x = \frac{x \sin (\theta_x + \theta_\gamma) - vt \sin (\theta_v + \theta_t + \theta_\gamma)}{x \cos (\theta_x + \theta_\gamma) - vt \cos (\theta_v + \theta_t + \theta_\gamma)} \tag{172}$$

while from equations (169) and (170) t' and θ'_t can be calculated in the following manner

$$t'^2 = \gamma^2[t^2 + v^2x^2/c^4 - 2vxt/c^2 \cos (\theta_x + \theta_v - \theta_t)] \tag{173}$$

$$\tan \theta'_t = \frac{t \sin (\theta_t + \theta_\gamma) - vx/c^2 \sin (\theta_v + \theta_x + \theta_\gamma)}{t \cos (\theta_t + \theta_\gamma) - vx/c^2 \cos (\theta_v + \theta_x + \theta_\gamma)} \tag{174}$$

From equations (171) and (173) the Minkowski interval can be written as

$$x'^2 - c^2t'^2 = \gamma^2[(1 - \beta^2)(x^2 - c^2t^2) - 4vtx \sin \theta_v \sin (\theta_x - \theta_t)] \tag{175}$$

where γ is given by equation (165). If $\theta_v = 0$ equation (175) reduces to equation (154).

Consider now some properties of the velocity factor γ given by equation (165). The first thing to see is that γ is not singular for real values of β. In fact it is easy to show that the roots of the denominator in equation (165) are given by

$$\beta^4 - 2\beta^2 \cos (2\theta_v) + 1 = 0 \tag{176}$$

or

$$\beta = e^{\pm j\theta_v} \tag{177}$$

Only when $\theta_v = 0$ does equation (176) have a real root $\beta = 1$ which agrees with equation (153). By taking the derivative of γ given by equation (165) and setting the result equal to zero it is easy to show that γ has a maximum value (assuming θ_v to be independent of velocity) given by

$$[\gamma]_{max} = [\sin(2\theta_v)]^{-1/2} \tag{178}$$

and this maximum value of γ occurs at a value of β given by

$$[\beta]_{max\ \gamma} = [\cos(2\theta_v)]^{1/2} < 1 \tag{179}$$

Combining equations (166) and (179) gives the following value of θ_γ at the maximum point of γ

$$[\theta_\gamma]_{max\ \gamma} = \pi/4 - \theta_v \tag{180}$$

The values of γ and θ_γ for $\beta = 1$ are obtained from equations (165) and (166) respectively as

$$[\gamma]_{\beta=1} = [2 \sin \theta_v]^{-1/2} \tag{181}$$

$$[\theta_\gamma]_{\beta=1} = \pi/4 - \theta_v/2 \tag{182}$$

The functions γ and θ_γ appear in Figures 1 and 2 respectively. As shown by equations (179) the maximum value of γ occurs for $\beta < 1$, and if θ_v is small the maximum value of γ occurs close to $\beta = 1$. Within asymmetric bulk matter or vacuum γ is nonsingular. For $\beta \sim 0$ it follows from equations (165) and (166) that

$$\gamma \sim 1 + \frac{1}{2}\beta^2 \cos(2\theta_v) \tag{183}$$

$$\theta_\gamma \sim \frac{1}{2}\beta^2 \sin(2\theta_v) \tag{184}$$

For $\beta \to \infty$ it follows from equations (165) and (166) that

$$\gamma \to 1/\beta \to 0 \tag{185}$$

$$\theta_\gamma \to \pi/2 - \theta_v \tag{186}$$

where θ_v is assumed to be independent of particle velocity. It is assumed that θ_v depends on P and θ_p of the ambient medium.

The complex number de Broglie wavelength for a relativistic particle moving at velocity $\bar{v}$ is[35]

$$\bar{\lambda} = h/\bar{p} = h/(m\bar{\gamma}\bar{v}) = \lambda e^{j\theta\lambda} \tag{187}$$

where $\bar{p}$ = complex number momentum whose magnitude is given by $p = m\gamma v$, and h = Planck's constant. From equation (187) it follows that

$$\lambda = h/(m\gamma v) = \lambda_c/(\gamma\beta) \tag{188}$$

$$\theta_\lambda = -\theta_\gamma - \theta_v \tag{189}$$

where λ_c = Compton wavelength given by[30]

$$\lambda_c = h/(mc) \tag{190}$$

Three special cases can be considered.

$$\underline{\text{Case 1.}} \quad \beta = [\beta]_{\max\ \gamma} = [\cos(2\theta_v)]^{1/2} \tag{191}$$

It follows from equations (178), (179), (180), (188), and (189) that

$$\lambda = \lambda_c[\tan(2\theta_v)]^{1/2} \tag{192}$$

$$\theta_\lambda = -\pi/4 \tag{193}$$

$$\underline{\text{Case 2.}} \quad \beta = 1 \tag{194}$$

In this case it follows from equations (181), (182) and (188) that

$$\lambda = \lambda_c\,(2\sin\theta_v)^{1/2} \tag{195}$$

$$\theta_\lambda = -\pi/4 - \theta_v/2 \tag{196}$$

$$\underline{\text{Case 3.}} \quad \beta \to \infty \tag{197}$$

In the limit $\beta \to \infty$ equations (185), (186), (188) and (189) give

$$\lambda \rightarrow \lambda_c \tag{198}$$

$$\theta_\lambda \rightarrow -\pi/2 \tag{199}$$

Therefore as β increases without limit the de Broglie wavelength increases to a limiting value of λ_c. It is assumed that θ_v is independent of velocity.

The total energy of a particle located in asymmetric bulk matter or vacuum is given by the following generalization of the standard results of special relativity[24]

$$\bar{\varepsilon} = \varepsilon e^{j\theta_\varepsilon} = \bar{\gamma} mc^2 \tag{200}$$

where the complex number velocity factor is given by equation (157), and m = proper mass. Therefore the total energy of a particle has the same properties as $\bar{\gamma}$, so that

$$\varepsilon = \gamma mc^2 \tag{201}$$

$$\theta_\varepsilon = \theta_\gamma \tag{202}$$

where the magnitude and internal phase angle of the velocity factor is given by equations (165) and (166) respectively. The kinetic energy of a particle in asymmetric bulk matter or vacuum is given by[24]

$$\bar{\varepsilon}_K = \varepsilon_K e^{j\theta_K} = (\bar{\gamma} - 1)mc^2 \tag{203}$$

The component form of equation (203) is written as

$$\varepsilon_K \cos \theta_K = (\gamma \cos \theta_\gamma - 1)mc^2 \tag{204}$$

$$\varepsilon_K \sin \theta_K = \gamma \sin \theta_\gamma \, mc^2 \tag{205}$$

and therefore for a broken symmetry system

$$\tan \theta_K = \frac{\gamma \sin \theta_\gamma}{\gamma \cos \theta_\gamma - 1} \tag{206}$$

$$\varepsilon_K = mc^2(\gamma^2 - 2\gamma \cos \theta_\gamma + 1)^{1/2} \tag{207}$$

Placing equations (183) and (184) into equations (206) and (207) shows that for $\beta \sim 0$

$$\theta_K \sim 2\theta_v \tag{208}$$

$$\varepsilon_K \sim \frac{1}{2} mv^2 \tag{209}$$

which agrees with the nonrelativistic limit obtained directly from equations (157) and (203) namely

$$\bar{\varepsilon}_K \sim \frac{1}{2} m\bar{v}^2 \tag{210}$$

Figures 3 and 4 show ε_K and θ_K in terms of β.

Specific values of the total energy, kinetic energy, and momentum will now be considered for some characteristic values of β.

Case 1. $\beta \sim 0$

$$\gamma \sim 1 + \beta^2/2 \cos(2\theta_v) \tag{211}$$

$$\theta_\gamma \sim \beta^2/2 \sin(2\theta_v) \tag{212}$$

$$\varepsilon \sim mc^2 + \frac{1}{2} mv^2 \tag{212A}$$

$$\theta_\varepsilon \sim \beta^2/2 \sin(2\theta_v) \tag{212B}$$

$$\varepsilon_K \sim \frac{1}{2} mv^2 \tag{212C}$$

$$\theta_K \sim 2\theta_v \tag{212D}$$

$$p \sim mv \tag{212E}$$

$$\theta_p \sim \theta_v \tag{212F}$$

Case 2. $\beta = [\beta]_{\max \gamma} = [\cos(2\theta_v)]^{1/2}$

$$\gamma = [\sin(2\theta_v)]^{-1/2} \tag{213}$$

$$\theta_\gamma = \pi/4 - \theta_v \tag{214}$$

$$\varepsilon = mc^2[\sin(2\theta_v)]^{-1/2} \tag{215}$$

$$\theta_\varepsilon = \pi/4 - \theta_v \tag{216}$$

$$\varepsilon_K = mc^2\left[\frac{1}{\sin(2\theta_v)} - \frac{2\cos(\pi/4 - \theta_v)}{\sqrt{\sin(2\theta_v)}} + 1\right]^{1/2} \tag{217}$$

$$\varepsilon_K \sim \varepsilon \text{ for small } \theta_v \tag{218A}$$

$$\tan\theta_K = \frac{\sin(\pi/4 - \theta_v)}{\cos(\pi/4 - \theta_v) - \sqrt{\sin(2\theta_v)}} \tag{218B}$$

$$\theta_K \sim \theta_\gamma \text{ for small } \theta_v \tag{218C}$$

$$p/mc = \gamma\beta = [\cot(2\theta_v)]^{1/2} \tag{218D}$$

$$\theta_p = \pi/4 \tag{218E}$$

Case 3. $\beta = 1$

$$\gamma = (2\sin\theta_v)^{-1/2} \tag{219}$$

$$\theta_\gamma = \pi/4 - \theta_v/2 \tag{220}$$

$$\varepsilon = mc^2(2\sin\vartheta_v)^{-1/2} = cp \tag{221}$$

$$\vartheta_\varepsilon = \pi/4 - \theta_v/2 \tag{222}$$

$$\varepsilon_K = mc^2\left[\frac{1}{2\sin\theta_v} - \frac{2\cos(\pi/4 - \theta_v/2)}{\sqrt{2\sin\theta_v}} + 1\right]^{1/2} \tag{223}$$

$$\varepsilon_K \sim \varepsilon \text{ for small } \theta_v \tag{224}$$

$$\tan \theta_K = \frac{\sin (\pi/4 - \theta_v/2)}{\cos (\pi/4 - \theta_v/2) - \sqrt{2 \sin \theta_v}} \tag{224A}$$

$$\theta_K \sim \theta_\gamma \text{ for small } \theta_v \tag{224B}$$

$$p/mc = \gamma\beta = (2 \sin \theta_v)^{-1/2} \tag{224C}$$

$$\theta_p = \pi/4 + \theta_v/2 \tag{224D}$$

Case 4. $\beta \to \infty$

$$\gamma \to 1/\beta \to 0 \tag{225}$$

$$\theta_\gamma \to \pi/2 - \theta_v \tag{226}$$

$$\varepsilon \to mc^2/\beta \to 0 \tag{227}$$

$$\theta_\varepsilon \to \pi/2 - \theta_v \tag{228}$$

$$\varepsilon_K \to mc^2[1 - \frac{1}{\beta} \cos (\pi/2 - \theta_v)] \to mc^2 \tag{229}$$

$$\theta_K \to \pi \tag{230A}$$

$$p/mc = \gamma\beta \to 1 \tag{230B}$$

$$\theta_p \to \pi/2 \tag{230C}$$

In direct analogy to the standard expression for relativistic momentum, the momentum of a particle located in asymmetric bulk matter or vacuum is written as[24]

$$\bar{p} = pe^{j\theta_p} = m\bar{\gamma}\bar{v} \tag{231}$$

so that

$$p = m\gamma v \tag{232}$$

$$\theta_p = \theta_\gamma + \theta_v \tag{233}$$

where γ and θ_γ are given by equations (165) and (166) respectively. From equations (157), (204) and (231) it follows that the single particle energy is

$$\bar{\varepsilon}^2 = c^2\bar{p}^2 + m^2c^4 \tag{234}$$

which shows the four-vector status of $\bar{\varepsilon}$ and $\bar{p}$. Equation (234) has two component equations

$$\varepsilon^2 \cos(2\theta_\varepsilon) = c^2p^2 \cos(2\theta_p) + m^2c^4 \tag{235}$$

$$\varepsilon^2 \sin(2\theta_\varepsilon) = c^2p^2 \sin(2\theta_p) \tag{236}$$

Equations (231) through (236) are equivalent to equations (165), (166), (205) and (206). From equations (235) and (236) it follows that

$$\tan(2\theta_\varepsilon) = \frac{p^2 \sin(2\theta_p)}{p^2 \cos(2\theta_p) + m^2c^2} \tag{237}$$

$$\gamma = \frac{\varepsilon}{mc^2} = \left[\left(\frac{p}{mc}\right)^4 + 2\left(\frac{p}{mc}\right)^2 \cos(2\theta_p) + 1\right]^{1/4} \tag{238}$$

It should be remembered that for asymmetric matter or vacuum, an interaction potential and a gauge potential needs to be added to obtain the total single particle energy

$$\bar{\varepsilon}_i = \bar{\varepsilon} + \bar{V}_e + \bar{V}_g \tag{239}$$

Equation (232) can be written as

$$p/mc = \gamma\beta \tag{240A}$$

Combining equations (165) and (240A) and setting the derivative of the momentum equal to zero gives the following value of β for maximum momentum

$$[\beta]_{\max p} = [\cos(2\theta_v)]^{-1/2} = 1/[\beta]_{\max \gamma} > 1 \tag{240B}$$

for which

$$[\gamma]_{\max p} = [\cot(2\theta_v)]^{1/2} = [p/mc]_{\max \gamma} \tag{240C}$$

so that the maximum momentum is

$$[p/mc]_{\max} = [\sin(2\theta_v)]^{-1/2} = [\gamma]_{\max}$$

where equation (178) has been used. Figures 5 and 6 give p and θ_p in terms of β.

The following arguments show how numerical values of θ_v for the asymmetric vacuum can be obtained from the experimental results of the Michelson-Morley experiment.[10] The generalization of the standard relativistic velocity addition formula to a system with broken internal symmetry is[24]

$$\bar{w} = \frac{\bar{u} + \bar{v}}{1 + \bar{u}\bar{v}/c^2} \tag{241}$$

where $\bar{u}$ = particle velocity relative to a reference frame that itself is moving at a velocity $\bar{v}$, and $\bar{w}$ = particle velocity relative to a frame of reference from which the moving frame has a velocity $\bar{v}$. Writing the velocities as

$$\bar{w} = we^{j\theta_w} \qquad \bar{u} = ue^{j\theta_u} \qquad \bar{v} = ve^{j\theta_v} \tag{241A}$$

gives the following velocity addition formula for asymmetric bulk matter or vacuum

$$we^{j\theta_w} = \frac{A + jB}{C + jD} \tag{242}$$

$$w = \left[\frac{A^2 + B^2}{C^2 + D^2}\right]^{1/2} \tag{243}$$

$$\theta_w = \theta_N - \theta_D \tag{244}$$

$$\tan \theta_N = B/A \tag{245}$$

$$\tan \theta_D = D/C \tag{246}$$

$$A = u \cos \theta_u + v \cos \theta_v \tag{247}$$

$$B = u \sin \theta_u + v \sin \theta_v \tag{248}$$

$$C = 1 + uv/c^2 \cos (\theta_u + \theta_v) \tag{249}$$

$$D = uv/c^2 \sin (\theta_u + \theta_v) \tag{250}$$

It is easy to show that

$$A^2 + B^2 = u^2 + v^2 + 2uv \cos (\theta_u - \theta_v) \tag{251}$$

$$C^2 + D^2 = 1 + u^2v^2/c^4 + 2uv/c^2 \cos (\theta_u + \theta_v) \tag{252}$$

It will be assumed that the internal phase of the particle velocity is independent of the magnitude of the velocity so that $\theta_u = \theta_v = \theta$ and

$$A^2 + B^2 = (u + v)^2 \tag{253}$$

$$C^2 + D^2 = 1 + u^2v^2/c^4 + 2uv/c^2 \cos (2\theta) \tag{254}$$

Consider the case $u = c$ and $v = c$, then equations (243), (253) and (254) give

$$w = c/\cos \theta \tag{255}$$

For the case $\theta = 0$, the standard result $w = c$ is regained.

In order to determine θ, consider the case where $u = c$ and $v =$ speed of the earth in its orbit which is much less than c. From equations (253) and (254) it follows that for this case

$$A^2 + B^2 = (c + v)^2 \tag{256A}$$

$$C^2 + D^2 = 1 + v^2/c^2 + 2v/c \cos (2\theta) = (1 + v/c)^2 - 4v/c \sin^2 \theta \tag{256B}$$

From equation (243) it follows that

$$w = \frac{c + v}{\sqrt{(1 + \beta)^2 - 4\beta \sin^2 \theta}} \tag{257}$$

$$\sim c\left[1 + \frac{2\beta \sin^2 \theta}{(1 + \beta)^2} - \cdots\right] \tag{258}$$

$$\sim c(1 + 2\beta \sin^2 \theta - \cdots) \tag{259}$$

where $\beta = v/c << 1$. The Galilean result would be $w = c + v = c(1 + \beta)$ instead of equation (257), while the standard special relativistic result without broken symmetry would be $w = c$ which can be regained from equations (257) through (259) by taking $\theta = 0$.

The details of the Michelson-Morley experiment are described in many references, and only the briefest description will be given here.[10,24-34] Using the Galilean assumption $w = c + v$ and $w = c - v$ respectively for the speed of light propagating with and against the ether, the number of interference fringes to be expected in a Michelson interferometer whose arms are parallel and perpendicular to the earth's motion is given according to the Galilean assumption by[10]

$$N_G = -2L/\lambda \ \beta^2 \tag{260}$$

where λ = wavelength of light, and L = length of the arms of the interferometer. The experimental value N_E of the number of fringes has been getting smaller relative to N_G as more accurate experiments are performed, and following Reference 10, N_E is given by

$$N_E \sim \frac{1}{400} N_G \tag{261}$$

On the other hand for the broken symmetry vacuum case equation (259), the predicted number of interference fringes N_{BS} is given by

$$N_{BS} = -2L/\lambda \ (2\beta \sin^2 \theta)^2 = -8L/\lambda \ \beta^2 \sin^4 \theta \tag{262}$$

If it is assumed that $N_{BS} = N_E$ it follows from equations (260) through (262) that

$$\sin^4 \theta \sim \frac{1}{1600} \tag{263}$$

$$\theta \sim 9.1° \tag{264}$$

Since future Michelson-Morley experiments may find values of N_E lower than the one used in this paper one can conclude that $\theta = \theta_v^{(v)} < 9°$ for the broken sym-

metry angle of particle velocity in the vacuum. The Michelson-Morley experiment not only shows that the Galilean velocity addition formula $c \pm v$ for a light source is invalid, but gives a positive result in the form of an upper limit to the value of the velocity asymmetry angle θ_v .

Alternatively, measurements of the velocity factor γ from particle accelerator experiments may eventually produce a maximum value of γ which will immediately determine the value of θ_v . For this, particles with $\beta > 1$ would have to be observed. If no such particles are ever found, it would show that $\theta_v = 0$ for the vacuum, and that the vacuum is symmetric. Experiment can only resolve this issue. Experiments to determine θ_v for asymmetric bulk matter may be easier because θ_v for bulk matter is expected to be larger than $\theta_v^{(v)}$ for the vacuum. Note that astronomical objects with $\beta > 1$ have apparently already been observed, and their explanation in terms of conventional effects can be given only with much difficulty.[36]

Finally, the laws of motion of a relativistic particle in asymmetric bulk matter or vacuum are considered. Newton's law of motion is modified by special relativity to give the following dynamical equation of motion for a force in the direction of motion[24]

$$F^a = \frac{d}{dt}(m\gamma_a v_a) = m\gamma_a^3 \, dv_a/dt_a = m\gamma_a^3 a_a \tag{265}$$

where a_a = conventionally calculated acceleration, and γ_a is given by equation (153). The generalization of this equation to the case of particle motion in asymmetric bulk matter or vacuum is

$$\bar{F} = \frac{d}{d\bar{t}}(m\bar{\gamma}\bar{v}) = m\bar{\gamma}^3 \, d\bar{v}/d\bar{t} = m\bar{\gamma}^3 \bar{a} \tag{266}$$

where $\bar{\gamma}$ is given by equation (157) and where $\bar{t}$, $\bar{v}$, $\bar{a}$, and $\bar{F}$ are the gauge rotated time, velocity, acceleration and force respectively. Therefore

$$\bar{a} = ae^{j\theta_a} \tag{267}$$

$$\bar{F} = Fe^{j\theta_F} \tag{268}$$

Combining equation (162) with equations (266) through (268) gives the force in the direction of motion as

$$F = m\gamma^3 a \tag{269}$$

$$\theta_F = 3\theta_\gamma + \theta_a \tag{270}$$

where γ and θ_γ are given by equations (165) and (166) respectively.

5. ELECTROMAGNETIC WAVE EQUATIONS. A direct result of Maxwell's equations is a set of wave equations that describe the time and space dependence of the electric and magnetic field vectors in a material body or vacuum.[9,10] This section considers the construction of electromagnetic wave equations for matter and radiation with broken internal symmetries. The standard equation of telegraphy that determines the electric (or magnetic) field in a conducting medium is[9,10]

$$\nabla_a^2 E_\alpha^a = \varepsilon^a \mu^a \frac{\partial^2 E_\alpha^a}{\partial t_a^2} + \mu^a \sigma^a \frac{\partial E_\alpha^a}{\partial t_a} = 0 \tag{271}$$

where $\alpha = x$, y and z , and σ^a = unrenormalized conductivity. The Laplacian operator is defined as

$$\nabla_a^2 = \frac{\partial^2}{\partial x_a^2} + \frac{\partial^2}{\partial y_a^2} + \frac{\partial^2}{\partial z_a^2} \tag{272}$$

The prescription introduced in this paper to handle electromagnetism in matter or vacuum with broken internal symmetries is to use gauge rotated field vectors and gauge rotated space and time coordinates. Applying this prescription to equation (271) yields

$$\bar{\nabla}^2 \bar{E}_\alpha = \varepsilon\mu \frac{\partial^2 \bar{E}_\alpha}{\partial \bar{t}^2} + \mu\sigma \frac{\partial \bar{E}_\alpha}{\partial \bar{t}} \tag{273}$$

The complex number Laplacian is given by

$$\bar{\nabla}^2 = \frac{\partial^2}{\partial \bar{x}^2} + \frac{\partial^2}{\partial \bar{y}^2} + \frac{\partial^2}{\partial \bar{z}^2} \tag{274}$$

The first and second derivative terms in equation (273) have already been evaluated in Section 2. Using the notation developed in equations (54) through (59) and (65) through (70) allows equation (273) to be written as six real number relations as follows

$$\sum_\beta R_{2\beta}(E_{xR}, E_{xI}) \sim \varepsilon\mu R_{2t}(E_{xR}, E_{xI}) + \mu\sigma R_{1t}(E_{xR}, E_{xI}) \tag{275}$$

$$\sum_\beta R_{2\beta}(E_{yR}, E_{yI}) \sim \varepsilon\mu R_{2t}(E_{yR}, E_{yI}) + \mu\sigma R_{1t}(E_{yR}, E_{yI}) \tag{276}$$

$$\sum_\beta R_{2\beta}(E_{zR}, E_{zI}) \sim \varepsilon\mu R_{2t}(E_{zR}, E_{zI}) + \mu\sigma R_{1t}(E_{zR}, E_{zI}) \tag{277}$$

$$\sum_{\beta} I_{2\beta}(E_{xR},E_{yR}) \sim \varepsilon\mu I_{2t}(E_{xR},E_{xI}) + \mu\sigma I_{1t}(E_{xR},E_{xI}) \tag{278}$$

$$\sum_{\beta} I_{2\beta}(E_{yR},E_{yI}) \sim \varepsilon\mu I_{2t}(E_{yR},E_{yI}) + \mu\sigma I_{1t}(E_{yR},E_{yI}) \tag{279}$$

$$\sum_{\beta} I_{2\beta}(E_{zR},E_{zI}) \sim \varepsilon\mu I_{2t}(E_{zR},E_{zI}) + \mu\sigma I_{1t}(E_{zR},E_{zI}) \tag{280}$$

where the sum is over $\beta = x, y$ and z.

The standard equations that determine the electromagnetic potentials are written as[9,10]

$$\nabla_a^2\phi^a - \varepsilon^a\mu^a \frac{\partial^2\phi^a}{\partial t_a^2} = -\rho_q^a/\varepsilon^a \tag{281}$$

$$\nabla_a^2 A_\alpha^a - \varepsilon^a\mu^a \frac{\partial^2 A_\alpha^a}{\partial t_a^2} = -\mu^a j_\alpha^a \tag{282}$$

The generalization of these equations to electromagnetic fields in asymmetric bulk matter or vacuum is as follows

$$\bar{\nabla}^2\bar{\phi} - \varepsilon\mu \frac{\partial^2\bar{\phi}}{\partial\bar{t}^2} = -\rho_q/\varepsilon \tag{283}$$

$$\bar{\nabla}^2\bar{A}_\alpha - \varepsilon\mu \frac{\partial^2\bar{A}_\alpha}{\partial\bar{t}^2} = -\mu\bar{j}_\alpha \tag{284}$$

where the complex number electromagnetic potentials are written as

$$\bar{\phi} = \phi e^{j\theta\phi} = \phi_R + j\phi_I \tag{285}$$

$$\bar{A}_\alpha = A_\alpha e^{j\theta A\alpha} = A_{\alpha R} + jA_{\alpha I} \tag{286}$$

Using the notation of equations (65) and (66) allows equation (283) to be written as the following two relations

$$\sum_{\beta} R_{2\beta}(\phi_R,\phi_I) - \varepsilon\mu R_{2t}(\phi_R,\phi_I) \sim -\rho_q/\varepsilon \tag{287}$$

$$\sum_{\beta} I_{2\beta}(\phi_R,\phi_I) - \varepsilon\mu I_{2t}(\phi_R,\phi_I) \sim 0 \tag{288}$$

while equation (284) can be written as the following six approximations

$$\sum_{\beta} R_{2\beta}(A_{xR},A_{xI}) - \varepsilon\mu R_{2t}(A_{xR},A_{xI}) \sim - \mu j_x \cos \theta_{jx} \tag{289}$$

$$\sum_{\beta} R_{2\beta}(A_{yR},A_{yI}) - \varepsilon\mu R_{2t}(A_{yR},A_{yI}) \sim - \mu j_y \cos \theta_{jy} \tag{290}$$

$$\sum_{\beta} R_{2\beta}(A_{zR},A_{zI}) - \varepsilon\mu R_{2t}(A_{zR},A_{zI}) \sim - \mu j_z \cos \theta_{jz} \tag{291}$$

$$\sum_{\beta} I_{2\beta}(A_{xR},A_{xI}) - \varepsilon\mu I_{2t}(A_{xR},A_{xI}) \sim - \mu j_x \sin \theta_{jx} \tag{292}$$

$$\sum_{\beta} I_{2\beta}(A_{yR},A_{yI}) - \varepsilon\mu I_{2t}(A_{yR},A_{yI}) \sim - \mu j_y \sin \theta_{jy} \tag{293}$$

$$\sum_{\beta} I_{2\beta}(A_{zR},A_{zI}) - \varepsilon\mu I_{2t}(A_{zR},A_{zI}) \sim - \mu j_z \sin \theta_{jz} \tag{294}$$

Finally the gauge conditions for an electromagnetic field with broken internal symmetry is written as[9,10]

$$\vec{\nabla} \cdot \vec{A} + \varepsilon\mu \frac{\partial \bar{\phi}}{\partial \bar{t}} = 0 \tag{295}$$

which can be written in terms of real and imaginary components as

$$\sum_{\beta} R_{1\beta}(A_{\beta R},A_{\beta I}) + \varepsilon\mu R_{1t}(\phi_R,\phi_I) = 0 \tag{296}$$

$$\sum_{\beta} I_{1\beta}(A_{\beta R},A_{\beta I}) + \varepsilon\mu I_{1t}(\phi_R,\phi_I) = 0 \tag{297}$$

where the sum is over $\beta = x, y$ and z.

6. VACUUM WITH BROKEN INTERNAL SYMMETRIES. Of special importance to the propagation of electromagnetic waves are the properties of the vacuum state. The vacuum state may exhibit the same broken internal symmetries as does bulk

matter. Consider the vacuum state to have a zero temperature state coupled to a thermal state in such a way that the vacuum energy density and pressure for low temperatures are given by

$$\bar{E}^{(v)} = \bar{E}_o^{(v)} + \bar{E}_j^{(v)} T^j + \cdots = E^{(v)} e^{j\theta_E^{(v)}} \tag{298}$$

$$\bar{P}^{(v)} = \bar{P}_o^{(v)} + \bar{P}_j^{(v)} T^j + \cdots = P^{(v)} e^{j\theta_P^{(v)}} \tag{299}$$

where $\bar{E}^{(v)}$ and $\bar{P}^{(v)}$ = vacuum energy density and pressure respectively, $\bar{E}_o^{(v)}$ and $\bar{P}_o^{(v)}$ = zero temperature vacuum energy density and pressure respectively, and $\bar{E}_j^{(v)}$ and $\bar{P}_j^{(v)}$ = thermal coefficients for the vacuum energy density and pressure respectively. The vacuum Grüneisen parameter is defined by

$$\bar{\gamma}_o^{(v)} = \gamma_o^{(v)} e^{j\theta_{\gamma o}^{(v)}} \tag{300}$$

$$= \left[\frac{\partial \bar{P}^{(v)}/\partial T}{\partial \bar{E}^{(v)}/\partial T}\right]_{T=o} = \frac{\bar{P}_j^{(v)}}{\bar{E}_j^{(v)}} = \frac{1}{(j-1)} \frac{V}{\bar{U}_j^{(v)}} \frac{\partial \bar{U}_j^{(v)}}{\partial V}$$

where $\bar{U}_j^{(v)} = V\bar{E}_j^{(v)}$, and where j = index that describes the thermal properties of the vacuum.

The energy density $\bar{E}_o^{(v)}$ and Grüneisen parameter $\bar{\gamma}_o^{(v)}$ for the zero temperature vacuum are calculated from the simultaneous solution of two differential equations .

$$\bar{E}_o^{(v)} - 3\left\{[1 + \bar{\gamma}_o^{(v)}]\bar{P}_o^{(v)} - \bar{K}_o^{(v)}\right\} = 0 \tag{301}$$

$$1 + j + \frac{j\bar{\gamma}_o^{(v)}\bar{P}_o^{(v)}}{\bar{P}_o^{(v)} - \bar{K}_o^{(v)}} + 3n \frac{d\bar{\gamma}_o^{(v)}}{dn} = 0 \tag{302}$$

which are just equations (252) and (253) of Reference 8 with their right hand sides set equal to zero. A trivial solution of equations (252) and (253) of Reference 8 with their right hand sides equal to zero is just $\bar{E}_o^{(v)} = 0$ and $\bar{E}_j^{(v)} = 0$ which is equivalent to the unrenormalized vacuum $E_o^a = 0$ and $E_j^a = 0$. A non-trivial solution is obtained by simultaneously solving equations (301)

and (302). It is easy to show that equation (301) can be written as

$$3n^2 \frac{d^2\bar{E}_o^{(v)}}{dn^2} - 3[1 + \bar{\gamma}_o^{(v)}]n \frac{d\bar{E}_o^{(v)}}{dn} + [3\bar{\gamma}_o^{(v)} + 4]\bar{E}_o^{(v)} = 0 \qquad (303)$$

The vacuum radiation energy density and pressure are written as[8]

$$\bar{E}_r^{(v)} = \bar{E}_{or}^{(v)} + \bar{E}_{jr}^{(v)} T^j + \cdots \qquad (304)$$

$$\bar{P}_r^{(v)} = \bar{P}_{or}^{(v)} + \bar{P}_{jr}^{(v)} T^j + \cdots \qquad (305)$$

while the zero temperature radiation Gruneisen parameter for the vacuum is given by

$$\bar{\gamma}_{or}^{(v)} = \gamma_{or}^{(v)} e^{j\theta_{\gamma or}^{(v)}} \qquad (306)$$

$$= \left[\frac{\partial \bar{P}_r^{(v)}/\partial T}{\partial \bar{E}_r^{(v)}/\partial T}\right]_{T=o} = \frac{\bar{P}_{jr}^{(v)}}{\bar{E}_{jr}^{(v)}} = \frac{1}{(j-1)} \frac{V}{\bar{U}_{jr}^{(v)}} \frac{\partial \bar{U}_{jr}^{(v)}}{\partial V}$$

The vacuum radiation equations are then written as[8]

$$\bar{E}_{or}^{(v)} - 3\left\{[1 + \bar{\gamma}_o^{(v)}]\bar{P}_{or}^{(v)} - \bar{K}_{or}^{(v)}\right\} - 3 \frac{\bar{E}_{jr}^{(v)}}{\bar{E}_j^{(v)}} \bar{P}_o^{(v)}[\bar{\gamma}_{or}^{(v)} - \bar{\gamma}_o^{(v)}] = 0 \qquad (307)$$

$$j\bar{E}_j^{(v)}[\bar{\alpha}^{(v)}\bar{K}_{or}^{(v)} - \bar{\beta}^{(v)}\bar{P}_{or}^{(v)}] + \bar{E}_{jr}^{(v)}\bar{S}_{jr}^{(v)} = 0 \qquad (308)$$

which are just equations (287) and (288) of Reference 8 with their right hand sides set equal to zero and a superscript (v) added to indicate a vacuum solution.

Therefore in principle asymmetric vacuum state is formally identical to the asymmetric bulk matter state. In fact, the vacuum is simpler than the bulk matter state as can be seen by comparing equations (301), (302), (307) and (308) with equations (252), (253), (287) and (288) respectively of Reference 8. The vacuum is expected to exhibit a broken internal symmetry state that is described by $\theta_p^{(v)}$ and $\theta_\gamma^{(v)}$. The broken symmetry of the vacuum will impress bro-

ken symmetries on the kinematic and dynamic variable of particles moving in the vacuum. Similarly, electromagnetic waves in the vacuum are expected to possess electric and magnetic fields and a spacetime coordinate description that exhibit internal phase angles.

7. CONCLUSION. The effects of the broken symmetry of space and time on electromagnetism in matter and the vacuum is considered, and Maxwell's equations with broken internal symmetries are developed. The Lorentz covariance of these equations is assumed to be valid but must now be represented in the form of complex number Lorentz transformations. The results of the Michelson-Morley experiment can be used to place a limit on the magnitude of the internal phase angle of the velocity of a particle moving in the vacuum, but more accurate experiments are required. Experiments conducted in asymmetric bulk matter may be fruitful becuase the internal symmetries of spacetime are larger in this case than for the vacuum. The wave equations and gauge conditions for electromagnetic waves with broken internal symmetries are easily developed. Finally, the broken symmetry properties of the vacuum are obtained by solving a set of coupled differential equations which are similar in form to the corresponding equations for asymmetric bulk matter.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. O'Raifeartaigh, L., Group Structure of Gauge Theories, Cambridge University Press, Cambridge, 1986.

2. Taylor, J., Gauge Theories of Weak Interactions, Cambridge University Press, Cambridge, 1976.

3. Ramond, P., Field Theory: A Modern Primer, Benjamin, New York, 1981.

4. Mandl, F. and Shaw, G., Quantum Field Theory, John Wiley, New York, 1984.

5. Itzykson, C. and Zuber, J., Quantum Field Theory, McGraw-Hill, New York, 1980.

6. de Wit, B. and Smith, J., Field Theory in Particle Physics, North-Holland, New York, 1986.

7. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

8. Weiss, R. A., "Thermodynamic Gauge Theory of Solids and Quantum Liquids with Internal Phase", Fifth Army Conference on Applied Mathematics and Computing, West Point, New York, June 15-18, 1987, p. 649.

9. Jackson, J., Classical Electrodynamics, John Wiley, New York, 1975.

10. Panofsky, W. and Phillips, M., Classical Electricity and Magnetism, Addison-Wesley, Reading, MA, 1955.

11. Stratton, J., Electromagnetic Theory, McGraw-Hill, New York, 1941.

12. Smythe, W., Static and Dynamic Electricity, McGraw-Hill, New York, 1950.

13. Becker, R., Electromagnetic Fields and Interactions, Dover, New York, 1964.

14. Jeans, J., The Mathematical Theory of Electricity and Magnetism, Cambridge University Press, Cambridge, 1951.

15. Menzel, D., Mathematical Physics, Prentice-Hall, New York, 1953.

16. Whittaker, E., A History of the Theories of Aether and Electricity, Philosophical Library, New York, 1951.

17. Sommerfeld, A., Electrodynamics, Academic, New York, 1952.

18. Kong, J., Electromagnetic Wave Theory, John Wiley, New York, 1986.

19. Chandrasekhar, S., Plasma Physics, University of Chicago Press, Chicago, 1960.

20. Artsimovich, L. A., A Physicists ABC on Plasma, MIR Publishers, Moscow, 1978.

21. Spitzer, L., Physics of Fully Ionized Gases, John Wiley-Interscience, New York, 1964.

22. Thompson, W. B., An Introduction to Plasma Physics, Addison-Wesley, Reading, Massachusetts, 1962.

23. Weinberg, S., Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity, John Wiley, New York, 1972.

24. Pauli, W., Theory of Relativity, Pergamon, New York, 1958.

25. Møller, C., The Theory of Relativity, Oxford, New York, 1952.

26. Synge, J. L., Relativity: The Special Theory, North-Holland, Amsterdam, 1965.

27. Bergmann, P., Introduction to the Theory of Relativity, Prentice-Hall, New York, 1953.

28. Eddington, A., The Mathematical Theory of Relativity, Cambridge University Press, Cambridge, 1952.

29. Weyl, H., Space-Time-Matter, Dover, New York, 1922.

30. Robertson, H. and Noonan, T., Relativity and Cosmology, Saunders, Philadelphia, 1968.

31. Tolman, R., Relativity Thermodynamics and Cosmology, Oxford, New York, 1934.

32. Aharoni, J., The Special Theory of Relativity, Oxford, New York, 1959.

33. Misner, C., Thorne, K. and Wheeler, J., Gravitation, Freeman, San Francisco, 1973.

34. Zeldovich, Ya. and Novikov, I., Relativistic Astrophysics Volume 1 Stars and Relativity, University of Chicago Press, Chicago, 1971.

35. Born, M., Atomic Physics, Hafner, New York, 1953.

36. Zensus, J. A. and Pearson, T. J., Eds., Superluminal Radio Sources, Cambridge University Press, New York, 1987.

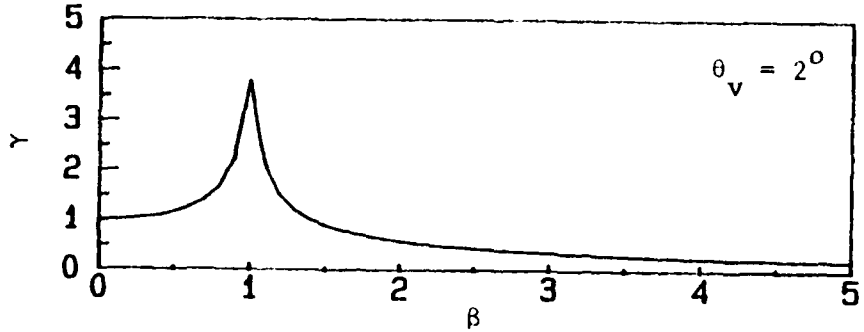


Figure 1. Magnitude of velocity factor versus magnitude of velocity ratio.

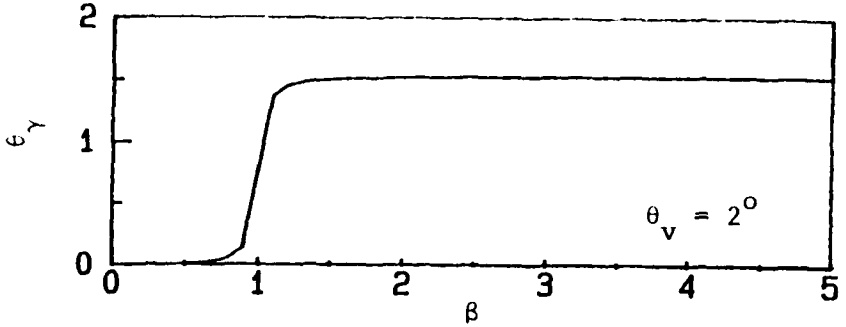


Figure 2. Phase angle of velocity factor versus magnitude of velocity ratio.

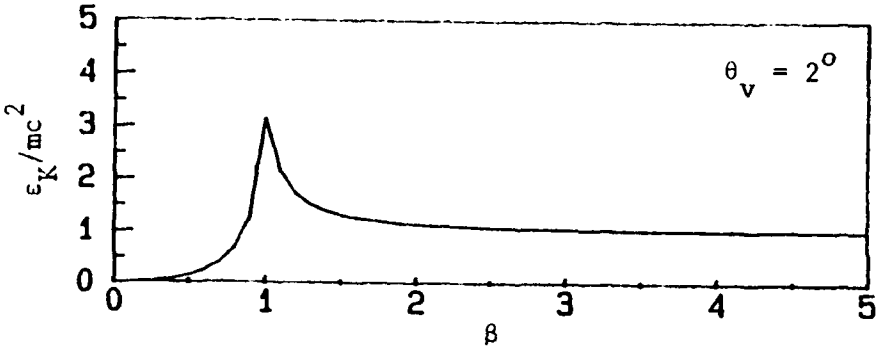


Figure 3. Normalized magnitude of kinetic energy versus magnitude of velocity ratio.

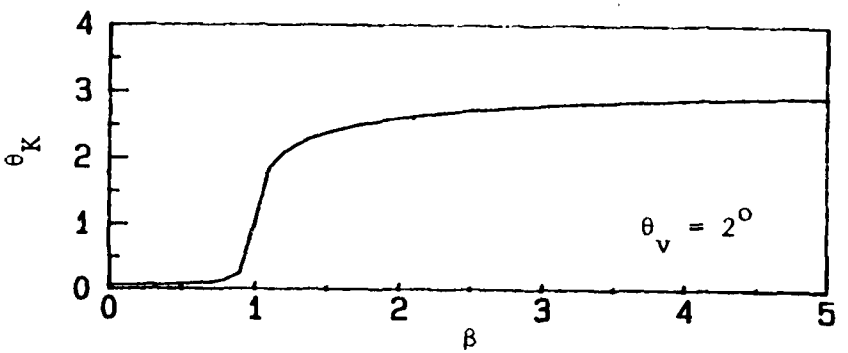


Figure 4. Phase angle of kinetic energy versus magnitude of velocity ratio.

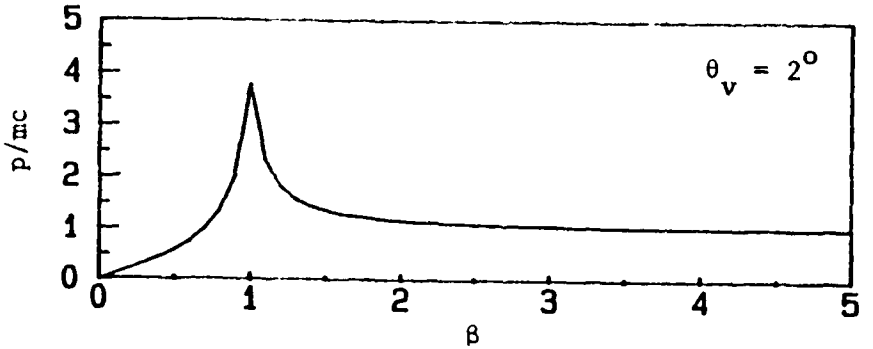


Figure 5. Normalized magnitude of momentum versus magnitude of velocity ratio.

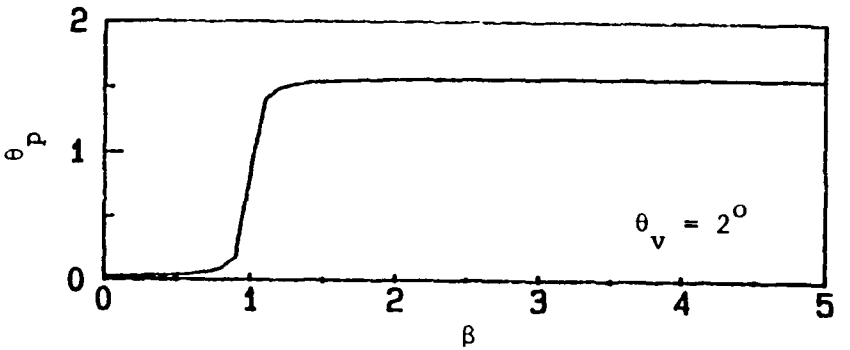


Figure 6. Phase angle of momentum versus magnitude of velocity ratio.

THE BROKEN SYMMETRY OF SPACE AND TIME IN BULK MATTER AND THE VACUUM

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. Because the pressure and internal energy of bulk matter and the vacuum are associated with internal phase angles, the space and time coordinates and the kinematic and dynamic variables of an interacting system of particles also exhibit broken internal symmetries. Specifically, in bulk matter or the vacuum with broken internal symmetries, the internal phase angles of the particle velocity, acceleration, and space and time coordinates are related to the internal phase angles of the pressure and internal energy. A procedure is developed for determining the internal phase angles of the kinematic and dynamic variables and of the space and time coordinates in terms of Euler's equations of motion. Continuum mechanics and elasticity solutions for bulk matter require the joint determination of phase angles for the space and time coordinates and the magnitude and internal phase angle of the pressure. Rotating matter with broken space and time symmetries is treated, and it is shown that the conservation of angular momentum is valid for such a system. The gravitational equilibrium configurations of stars and planets are treated for state equations that have broken internal symmetries, and equations are developed that relate the internal phase angles of the space and time coordinates to the internal phase angle of the pressure. Newtonian gravity in matter with broken internal symmetry is considered and applications to the earth's gravity field are suggested. These results will also affect the predicted trajectories of ballistic missiles.

1. INTRODUCTION. The fundamental interactions in nature are formulated as gauge theories. For instance, the theory of gravity is formulated as a gauge theory based on the Lorentz group $SO(3,1)$, while electromagnetism is based on the gauge group $U(1)$.[1] The nongravitational forces are thought to be described by the gauge group $SU(3) \times SU(2) \times U(1)$.[2,3] In fact the Lie group $U(1)$ and its real value analog $e^{\pm\phi}$ have been shown to be the gauge groups of relativistic thermodynamics.[4] The pressure and energy density of matter described by relativistic thermodynamics are associated with broken symmetries. This is related to the fact that the pressure and energy density can be gauge rotated in such a way as to leave the terms of the basic trace equation of relativistic thermodynamics gauge invariant.[4]

For an interacting bulk matter system the broken symmetry of the state equation is vacuum induced and results from the solution of a complex number trace equation that relates the renormalized (relativistic) state equation to the corresponding ordinary state equation. This trace equation is given by[5,6]

$$\bar{U} + T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} - 3V\,\frac{d}{dV}(\bar{P}V)_{\bar{U}} = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{1}$$

or equivalently as

$$\left(1 - \bar{b} + T\,\frac{\partial}{\partial T} - \bar{b}V\,\frac{\partial}{\partial V}\right)\bar{E} - 3\left(1 + \bar{\gamma} + V\,\frac{\partial}{\partial V} - \bar{\gamma}T\,\frac{\partial}{\partial T}\right)\bar{P} = \psi^a \tag{2}$$

where $\bar{U}$, $\bar{E}$, $\bar{P}$, $\bar{\gamma}$, and $\bar{b}$ are complex number representations of the renormalized internal energy, energy density, pressure, and the gauge parameters, and where

$$\bar{b} = \frac{T\,\frac{\partial \bar{P}}{\partial T}}{\bar{P} - \bar{K}_T} \tag{3}$$

$$\bar{K}_T = -V\left(\frac{\partial \bar{P}}{\partial V}\right)_T \tag{4}$$

$$\bar{\gamma} = \frac{V}{\bar{C}_V}\,\frac{\partial \bar{P}}{\partial T} = \frac{\partial \bar{P}/\partial T}{\partial \bar{E}/\partial T} \tag{5}$$

$$\psi^a = \left(T\,\frac{\partial}{\partial T} - b^a V\,\frac{\partial}{\partial V} + 1 - b^a\right)E^a \tag{6}$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \tag{7}$$

The quantities E^a, P^a, and K_T^a = unrenormalized values of the energy density, pressure, and bulk modulus respectively. Throughout this paper the index "a" will refer to nonrelativistic (unrenormalized) calculations. The complex number thermodynamic state functions that appear in equations (1) and (2) will be written in terms of their internal phase angles as follows

$$\bar{U} = Ue^{j\theta_U} \tag{8}$$

$$\bar{E} = \bar{U}/V = Ee^{j\theta_U} \tag{9}$$

$$\bar{P} = Pe^{j\theta_P} \tag{10}$$

$$\bar{\gamma} = \gamma e^{j\theta_\gamma} \tag{11}$$

$$\bar{b} = be^{j\theta_b} \tag{12}$$

where θ_U, θ_p, θ_γ, and θ_b = internal phase angles of the internal energy, pressure, Grüneisen parameter, and b gauge parameter respectively. The relativistic ground state of the vacuum is described by equation (1) or (2) with their right hand sides set equal to zero. The vacuum state also has a broken symmetry and in fact the bulk matter state is essentially mathematically equivalent to the vacuum state.

On account of the broken symmetry of the pressure and energy density of bulk matter or the vacuum, time may not unfold in a purely linear fashion but may also rotate in an internal space. Spatial coordinates in bulk matter or the vacuum may also have broken internal symmetries that are associated with internal phase angles. The broken symmetries of space and time in bulk matter or the vacuum are related to the broken symmetries of the state equations for these systems. Thermodynamic and continuum mechanics theories will require the joint determination of the internal phase angles of space and time coordinates along with the pressure and internal energy and their internal phase angles. The gauge rotated space and time coordinates have an effect on the equations of motion of a system of particles and will affect the equilibrium configurations of atomic nuclei, planets and the stars. Note that it is the real parts of the complex number quantities such as space and time coordinates, pressure, energy, velocity and acceleration that are the measured quantities.

The broken symmetry of space and time is related to the broken symmetry of the pressure and internal energy of bulk matter or the vacuum as determined from solutions of equation (1). The right hand side of equation (1) is equal to zero for the case of the vacuum. When matter is present the broken symmetry of space and time can be calculated in two ways: 1, at the macroscopic level through Euler's equations and the complex pressure field for interacting matter (Section 6), and 2, at the single particle level through the action of a complex gauge potential that is induced by vacuum effects. For the vacuum only the second method is possible because the matter density is zero, and the complex gauge potential for the vacuum must be determined.

The complex gauge potential is calculated from the relativistic internal energy and pressure that are obtained from equation (1). This is done by calculating the renormalized complex valued partition function which is defined as[7,8]

$$\bar{Z} = \int \eta e^{-\beta \bar{H}} d\bar{q}\, d\bar{p} \tag{13}$$

where η = degeneracy, $\beta = 1/(kT)$, and where the complex number Hamiltonian is given by

$$\bar{H} = \frac{\bar{p}^2}{2m} + \bar{W} \tag{14}$$

where $\bar{W} = V^a + \bar{V}_g$, where V^a = ordinary external potential, $\bar{V}_g$ = complex number gauge potential that is responsible for the difference between $\bar{U}$ and U^a given in equation (1). The connection between the internal energy and pressure and the partition function is given by[7,8]

$$\bar{U} = -\left(\frac{\partial\ \ell n\ \bar{Z}}{\partial\beta}\right)_V \qquad \bar{P} = \frac{1}{\beta}\left(\frac{\partial\ \ell n\ \bar{Z}}{\partial V}\right)_\beta \tag{15}$$

where $\bar{U}$ is given by equation (1), so that equations (13) through (15) can be used to determine the complex gauge potential $\bar{V}_g$ in terms of P and θ_P of the complex matter fields. These equations relate the macroscopic pressure field given by equation (1) to the microscopic gauge potential $\bar{V}_g$. For the broken symmetry vacuum the partition function is

$$\bar{Z}^{(v)} = \int \eta e^{-\beta \bar{V}_g}\ d\bar{q}^{(v)}\ d\bar{p}^{(v)} \tag{13A}$$

from which $\bar{U}^{(v)}$ and $\bar{P}^{(v)}$ can be obtained using equation (15). These values of $\bar{U}^{(v)}$ and $\bar{P}^{(v)}$ must agree with the vacuum solutions of equation (1), and this determines $\bar{V}_g$.

The broken symmetry of the state functions of interacting bulk matter and the vacuum impart a broken symmetry to the velocity, acceleration and space and time coordinates of particles located in bulk matter or the vacuum. Forces exerted in bulk matter or the vacuum will also exhibit broken internal symmetries. The aim of this paper is to relate the broken symmetries of space, time and the kinematic and dynamical variables, to the broken symmetry of the state equations for interacting bulk matter or the vacuum. The paper is organized as follows: Section 2. introduces gauge rotated coordinates, Section 3. treats the geometry of broken internal symmetry, Section 4. considers the kinematics and dynamics of broken symmetry particle systems, Section 5. studies rotating systems with broken internal symmetry, Section 6. introduces the Euler equations for bulk matter with broken symmetry, and Section 7. considers the equilibrium equations of stars and planets whose matter has internal phase.

2. GAUGE ROTATED SPACE AND TIME. In bulk matter or the vacuum the thermodynamic functions such as pressure and internal energy exhibit internal phases (broken symmetry).[6] This suggests that space and time coordinates in bulk matter or the vacuum may also possess broken symmetries. Accordingly the space and time coordinates of particles in bulk matter are written as

$$\bar{x} = xe^{j\theta_x} \tag{16}$$

$$\bar{y} = ye^{j\theta_y} \tag{17}$$

$$\bar{z} = ze^{j\theta_z} \tag{18}$$

$$\bar{t} = te^{j\theta_t} \tag{19}$$

where the phase angles θ_x, θ_y, θ_z, and θ_t manifest the broken symmetry. It will be assumed that in bulk matter the phase angles can be represented as

$$\theta_x = \theta_x(x,y,z,t) \tag{20}$$

$$\theta_y = \theta_y(x,y,z,t) \tag{21}$$

$$\theta_z = \theta_z(x,y,z,t) \tag{22}$$

$$\theta_t = \theta_t(x,y,z,t) \tag{23}$$

For the vacuum, coordinates will be written as

$$\bar{x}^{(v)} = x^{(v)} e^{j\theta_x^{(v)}} \tag{23A}$$

$$\bar{y}^{(v)} = y^{(v)} e^{j\theta_y^{(v)}} \tag{23B}$$

$$\bar{z}^{(v)} = z^{(v)} e^{j\theta_z^{(v)}} \tag{23C}$$

$$\bar{t}^{(v)} = t^{(v)} e^{j\theta_t^{(v)}} \tag{23D}$$

The differentials of the space and time coordinates can be written as

$$d\bar{x} = e^{j\theta_x}(dx + jxd\theta_x) = e^{j\theta_x}\left[\left(1 + jx\frac{\partial\theta_x}{\partial x}\right)dx + jx\frac{\partial\theta_x}{\partial y}dy + jx\frac{\partial\theta_x}{\partial z}dz + jx\frac{\partial\theta_x}{\partial t}dt\right] \tag{24}$$

$$d\bar{y} = e^{j\theta_y}(dy + jyd\theta_y) = e^{j\theta_y}\left[jy\frac{\partial\theta_y}{\partial x}dx + \left(1 + jy\frac{\partial\theta_y}{\partial y}\right)dy + jy\frac{\partial\theta_y}{\partial z}dz + jy\frac{\partial\theta_y}{\partial t}dt\right] \tag{25}$$

$$d\bar{z} = e^{j\theta_z}(dz + jzd\theta_z) = e^{j\theta_z}\left[jz\frac{\partial\theta_z}{\partial x}dx + jz\frac{\partial\theta_z}{\partial y}dy + \left(1 + jz\frac{\partial\theta_z}{\partial z}\right)dz + jz\frac{\partial\theta_z}{\partial t}dt\right] \tag{26}$$

$$d\bar{t} = e^{j\theta_t}(dt + jtd\theta_t) = e^{j\theta_t}\left[jt\frac{\partial\theta_t}{\partial x}dx + jt\frac{\partial\theta_t}{\partial y}dy + jt\frac{\partial\theta_t}{\partial z}dz + \left(1 + jt\frac{\partial\theta_t}{\partial t}\right)dt\right] \tag{27}$$

From equations (24) through (27) it follows that

$$\partial\bar{x}/\partial x = \sqrt{1 + x^2(\partial\theta_x/\partial x)^2}\; e^{j(\theta_x + \beta_{x,x})} \tag{28}$$

$$\partial\bar{x}/\partial y = x\partial\theta_x/\partial y\, e^{j(\theta_x + \pi/2)} \tag{29}$$

$$\partial\bar{x}/\partial z = x\partial\theta_x/\partial z\, e^{j(\theta_x + \pi/2)} \tag{30}$$

$$\partial\bar{x}/\partial t = x\partial\theta_x/\partial t\, e^{j(\theta_x + \pi/2)} \tag{30A}$$

$$\partial\bar{y}/\partial x = y\partial\theta_y/\partial x\, e^{j(\theta_y+\pi/2)} \tag{31}$$

$$\partial\bar{y}/\partial y = \sqrt{1 + y^2(\partial\theta_y/\partial y)^2}\; e^{j(\theta_y+\beta y,y)} \tag{32}$$

$$\partial\bar{y}/\partial z = y\partial\theta_y/\partial z\, e^{j(\theta_y+\pi/2)} \tag{33}$$

$$\partial\bar{y}/\partial t = y\partial\theta_y/\partial t\, e^{j(\theta_y+\pi/2)} \tag{33A}$$

$$\partial\bar{z}/\partial x = z\partial\theta_z/\partial x\, e^{j(\theta_z+\pi/2)} \tag{34}$$

$$\partial\bar{z}/\partial y = z\partial\theta_z/\partial y\, e^{j(\theta_z+\pi/2)} \tag{35}$$

$$\partial\bar{z}/\partial z = \sqrt{1 + z^2(\partial\theta_z/\partial z)^2}\; e^{j(\theta_z+\beta z,z)} \tag{36}$$

$$\partial\bar{z}/\partial t = z\partial\theta_z/\partial t\, e^{j(\theta_z+\pi/2)} \tag{36A}$$

$$\partial\bar{t}/\partial x = t\partial\theta_t/\partial x\, e^{j(\theta_t+\pi/2)} \tag{37A}$$

$$\partial\bar{t}/\partial y = t\partial\theta_t/\partial y\, e^{j(\theta_t+\pi/2)} \tag{37B}$$

$$\partial\bar{t}/\partial z = t\partial\theta_t/\partial z\, e^{j(\theta_t+\pi/2)} \tag{37C}$$

$$\partial\bar{t}/\partial t = \sqrt{1 + t^2(\partial\theta_t/\partial t)^2}\; e^{j(\theta_t+\beta t,t)} \tag{37D}$$

where in equations (28) through (37) the following notation is used

$$\tan \beta_{x,x} = x\,\frac{\partial\theta_x}{\partial x} \tag{38}$$

$$\tan \beta_{y,y} = y\,\frac{\partial\theta_y}{\partial y} \tag{39}$$

$$\tan \beta_{z,z} = z\,\frac{\partial\theta_z}{\partial z} \tag{40}$$

$$\tan \beta_{t,t} = t\,\frac{\partial\theta_t}{\partial t} \tag{41}$$

The following angles are also useful

$$\tan \beta_{x,y} = x \frac{\partial \theta_x}{\partial y} \qquad \tan \beta_{x,z} = x \frac{\partial \theta_x}{\partial z} \qquad \tan \beta_{x,t} = x \frac{\partial \theta_x}{\partial t} \tag{42}$$

$$\tan \beta_{y,x} = y \frac{\partial \theta_y}{\partial x} \qquad \tan \beta_{y,z} = y \frac{\partial \theta_y}{\partial z} \qquad \tan \beta_{y,t} = y \frac{\partial \theta_y}{\partial t} \tag{43}$$

$$\tan \beta_{z,x} = z \frac{\partial \theta_z}{\partial x} \qquad \tan \beta_{z,y} = z \frac{\partial \theta_z}{\partial y} \qquad \tan \beta_{z,t} = z \frac{\partial \theta_z}{\partial t} \tag{44}$$

$$\tan \beta_{t,x} = t \frac{\partial \theta_t}{\partial x} \qquad \tan \beta_{t,y} = t \frac{\partial \theta_t}{\partial y} \qquad \tan \beta_{t,z} = t \frac{\partial \theta_t}{\partial z} \tag{44A}$$

From equations (16) through (19) it also follows that

$$\frac{\partial}{\partial \bar{\eta}} = e^{-j\theta_{d\eta}} \cos \beta_{\eta,\eta} \frac{\partial}{\partial \eta} \tag{45}$$

where $\eta = x, y, z,$ and $t,$ and

$$\theta_{d\eta} = \theta_\eta + \beta_{\eta,\eta} \tag{46}$$

$$\tan \beta_{\eta,\eta} = \eta \frac{\partial \theta_\eta}{\partial \eta} \tag{47}$$

$$\cos \beta_{\eta,\eta} = \frac{1}{\sqrt{1 + (\eta \, \partial \theta_\eta / \partial \eta)^2}} \tag{48}$$

The result in equation (45) follows from the fact that if $\bar{y}$, $\bar{z}$, and $\bar{t}$ are constant, then their respective magnitudes y, z, and t are also constant. The measured space and time coordinates are $x_m = x \cos \theta_x$, $y_m = y \cos \theta_y$, $z_m = z \cos \theta_z$ and $t_m = t \cos \theta_t$ respectively. Space and time can be represented by helices whose spiral lengths are $L_x = x \sec \beta_{x,x}$; $L_y = y \sec \beta_{y,y}$; $L_z = z \sec \beta_{z,z}$ and $L_t = t \sec \beta_{t,t}$. The conventional coordinates t_a , x_a , y_a , z_a are related to the gauge rotated coordinates by $t_a = t_m$, $x_a = x_m$, $y_a = y_m$ and $z_a = z_m$.

The following relationships hold for spherical polar coordinates

$$\bar{r} = r e^{j\theta_r} \tag{49}$$

$$\bar{\psi} = \psi e^{j\theta_\psi} \tag{50}$$

$$\bar{\phi} = \phi e^{j\theta_\phi} \tag{51}$$

where ψ = zenith angle, ϕ = azimuth angle, and where $\theta_r = \theta_r(r,\psi,\phi,t)$, $\theta_\psi = \theta_\psi(r,\psi,\phi,t)$, and $\theta_\phi = \theta_\phi(r,\psi,\phi,t)$ which gives

$$d\bar{r} = e^{j\theta_r}(dr + jrd\theta_r) = e^{j\theta_r}[(1 + jr\frac{\partial\theta_r}{\partial r})dr + jr\frac{\partial\theta_r}{\partial\psi}d\psi + jr\frac{\partial\theta_r}{\partial\phi}d\phi + jr\frac{\partial\theta_r}{\partial t}dt] \quad (52)$$

$$d\bar{\psi} = e^{j\theta_\psi}(d\psi + j\psi d\theta_\psi) = e^{j\theta_\psi}[j\psi\frac{\partial\theta_\psi}{\partial r}dr + (1 + j\psi\frac{\partial\theta_\psi}{\partial\psi})d\psi + j\psi\frac{\partial\theta_\psi}{\partial\phi}d\phi + j\psi\frac{\partial\theta_\psi}{\partial t}dt] \quad (53)$$

$$d\bar{\phi} = e^{j\theta_\phi}(d\phi + j\phi d\theta_\phi) = e^{j\theta_\phi}[j\phi\frac{\partial\theta_\phi}{\partial r}dr + j\phi\frac{\partial\theta_\phi}{\partial\psi}d\psi + (1 + j\phi\frac{\partial\theta_\phi}{\partial\phi})d\phi + j\phi\frac{\partial\theta_\phi}{\partial t}dt] \quad (54)$$

$$d\bar{t} = e^{j\theta_t}(dt + jtd\theta_t) = e^{j\theta_t}[jt\frac{\partial\theta_t}{\partial r}dr + jt\frac{\partial\theta_t}{\partial\psi}d\psi + jt\frac{\partial\theta_t}{\partial\phi}d\phi + (1 + jt\frac{\partial\theta t}{\partial t})dt] \quad (54A)$$

and

$$\partial\bar{r}/\partial r = \sqrt{1 + (r\partial\theta_r/\partial r)^2}\, e^{j(\theta_r + \beta_{r,r})} \quad (55)$$

$$\partial\bar{r}/\partial\psi = r\partial\theta_r/\partial\psi\, e^{j(\theta_r + \pi/2)} \quad (56)$$

$$\partial\bar{r}/\partial\phi = r\partial\theta_r/\partial\phi\, e^{j(\theta_r + \pi/2)} \quad (57)$$

$$\partial\bar{r}/\partial t = r\partial\theta_r/\partial t\, e^{j(\theta_r + \pi/2)} \quad (57A)$$

$$\partial\bar{\psi}/\partial r = \psi\partial\theta_\psi/\partial r\, e^{j(\theta_\psi + \pi/2)} \quad (58)$$

$$\partial\bar{\psi}/\partial\psi = \sqrt{1 + (\psi\partial\theta_\psi/\partial\psi)^2}\, e^{j(\theta_\psi + \beta_{\psi,\psi})} \quad (59)$$

$$\partial\bar{\psi}/\partial\phi = \psi\partial\theta_\psi/\partial\phi\, e^{j(\theta_\psi + \pi/2)} \quad (60)$$

$$\partial\bar{\psi}/\partial t = \psi\partial\theta_\psi/\partial t\, e^{j(\theta_\psi + \pi/2)} \quad (60A)$$

$$\partial\bar{\phi}/\partial r = \phi\partial\theta_\phi/\partial r\, e^{j(\theta_\phi + \pi/2)} \quad (61)$$

$$\partial\bar{\phi}/\partial\psi = \phi\partial\theta_\phi/\partial\psi\, e^{j(\theta_\phi + \pi/2)} \quad (62)$$

$$\partial\bar{\phi}/\partial\phi = \sqrt{1 + (\phi\partial\theta_\phi/\partial\phi)^2}\, e^{j(\theta_\phi + \beta_{\phi,\phi})} \quad (63)$$

$$\partial\bar{\phi}/\partial t = \phi\partial\theta_\phi/\partial t\, e^{j(\theta_\phi + \pi/2)} \quad (63A)$$

$$\partial\bar{t}/\partial r = t\partial\theta_t/\partial r\, e^{j(\theta_t+\pi/2)} \tag{63B}$$

$$\partial\bar{t}/\partial\psi = t\partial\theta_t/\partial\psi\, e^{j(\theta_t+\pi/2)} \tag{63C}$$

$$\partial\bar{t}/\partial\phi = t\partial\theta_t/\partial\phi\, e^{j(\theta_t+\pi/2)} \tag{63D}$$

where

$$\tan\beta_{r,r} = r\,\frac{\partial\theta_r}{\partial r} \tag{64}$$

$$\tan\beta_{\psi,\psi} = \psi\,\frac{\partial\theta_\psi}{\partial\psi} \tag{65}$$

$$\tan\beta_{\phi,\phi} = \phi\,\frac{\partial\theta_\phi}{\partial\phi} \tag{66}$$

where the connection between r, θ_r, ψ, θ_ψ, ϕ, θ_ϕ and x, θ_x, y, θ_y and z, θ_z is given in Section 3. One can also define the following angles

$$\tan\beta_{r,\psi} = r\,\frac{\partial\theta_r}{\partial\psi} \qquad \tan\beta_{r,\phi} = r\,\frac{\partial\theta_r}{\partial\phi} \qquad \tan\beta_{r,t} = r\,\frac{\partial\theta_r}{\partial t} \tag{67}$$

$$\tan\beta_{\psi,r} = \psi\,\frac{\partial\theta_\psi}{\partial r} \qquad \tan\beta_{\psi,\phi} = \psi\,\frac{\partial\theta_\psi}{\partial\phi} \qquad \tan\beta_{\psi,t} = \psi\,\frac{\partial\theta_\psi}{\partial t} \tag{68}$$

$$\tan\beta_{\phi,r} = \phi\,\frac{\partial\theta_\phi}{\partial r} \qquad \tan\beta_{\phi,\psi} = \phi\,\frac{\partial\theta_\phi}{\partial\psi} \qquad \tan\beta_{\phi,t} = \phi\,\frac{\partial\theta_\phi}{\partial t} \tag{69}$$

$$\tan\beta_{t,r} = t\,\frac{\partial\theta_t}{\partial r} \qquad \tan\beta_{t,\psi} = t\,\frac{\partial\theta_t}{\partial\psi} \qquad \tan\beta_{t,\phi} = t\,\frac{\partial\theta_t}{\partial\phi} \tag{69A}$$

The derivatives with respect to the complex spherical polar coordinates are now written in the same form as in equation (45) where now $n = t, r, \psi, \phi$. The measured space and time coordinates are $r_m = r\cos\theta_r$, $\psi_m = \psi\cos\theta_\psi$, $\phi_m = \phi\cos\theta_\phi$ and $t_m = t\cos\theta_t$ respectively.

The effects of the different types of forces on the gauge rotation of space and time depend on the relative magnitude and ranges of the forces. Over small distances $< 10^{-18}$ cm the color force dominates, $< 10^{-13}$ cm the strong nuclear force between nucleons dominates, $< 10^{-8}$ cm the electric and magnetic forces of electrons and nuclei dominate.[9,10] For ranges $> 10^{-8}$ cm the long range gravitational force dominates. Therefore when equations (24) through (69) are written, the origin of the coordinates is associated with the origin of the forces involved. Thus for gravity the origin is taken to be the center of the planet or star in question, and the range of r is throughout the gravitating body and beyond because gravity has an infinite range. For nuclear forces the range of

r is $r < 10^{-13}$ cm, while for electric forces in an atom $r < 10^{-8}$ cm. The values of θ_r and θ_t depend on the scale at which the dominant forces act. It is the real part of a complex number coordinate that is the quantity measured when a space or time coordinate measurement is made.

3. GEOMETRY OF SPACE IN BULK MATTER AND THE VACUUM. The broken symmetry of coordinates of particles located in bulk matter or the vacuum will influence the calculation of the effects of the basic forces that operate in these media, such as for example pressure and gravity. This section considers the effects of the broken symmetry of coordinates on basic geometrical quantities such as angles, areas, and path lengths. For example, the simple law of cosines for a plane triangle located in a medium with broken symmetry is written as

$$\cos \bar{\phi} = \frac{\bar{a}^2 + \bar{b}^2 - \bar{c}^2}{2\bar{a}\bar{b}} \tag{70}$$

where $\bar{a}$, $\bar{b}$ and $\bar{c}$ are the complex number sides of a plane triangle, and $\bar{\phi}$ is the complex angle opposite side $\bar{c}$. The complex number sides of the triangle can be written as

$$\bar{a} = ae^{j\theta a} \tag{71}$$

$$\bar{b} = be^{j\theta b} \tag{72}$$

$$\bar{c} = ce^{j\theta c} \tag{73}$$

then

$$\cos \bar{\phi} = \frac{1}{2}\frac{a}{b}e^{j(\theta a - \theta b)} + \frac{1}{2}\frac{b}{a}e^{j(\theta b - \theta a)} - \frac{c^2}{2ab}e^{j(2\theta c - \theta a - \theta b)} \tag{74}$$

From equation (74) it is clear that $\bar{\phi}$ and $\cos \bar{\phi}$ are complex numbers so that

$$\bar{\phi} = \phi e^{j\theta\phi} \tag{75}$$

$$\cos \bar{\phi} = C_\phi e^{-j\theta c\phi} \tag{76}$$

where C_ϕ = magnitude of $\cos \bar{\phi}$, and $\theta_{c\phi}$ = phase angle associated with $\cos \bar{\phi}$. In the same manner it follows that

$$\sin \bar{\phi} = S_\phi e^{j\theta s\phi} \tag{77}$$

where S_ϕ = magnitude of $\sin \bar{\phi}$, and $\theta_{s\phi}$ = phase angle associated with $\sin \bar{\phi}$. From the well known relation

$$\cos \bar{\phi} = \frac{1}{2} \left[e^{j\bar{\phi}} + e^{-j\bar{\phi}} \right] \tag{78}$$

it follows that

$$\cos \bar{\phi} = \cos \phi_R \cosh \phi_I - j \sin \phi_R \sinh \phi_I \tag{79}$$

where from equation (38)

$$\bar{\phi} = \phi_R + j\phi_I = \phi(\cos \theta_\phi + j \sin \theta_\phi) \tag{80}$$

Combining equations (76), (79), and (80) gives

$$C_\phi = \sqrt{\cos^2 (\phi \cos \theta_\phi) + \sinh^2 (\phi \sin \theta_\phi)} \tag{81}$$

$$\tan \theta_{c\phi} = \tan (\phi \cos \theta_\phi) \tanh (\phi \sin \theta_\phi) \tag{82}$$

In a similar manner from

$$\sin \bar{\phi} = \frac{1}{2j} \left[e^{j\bar{\phi}} - e^{-j\bar{\phi}} \right] \tag{83}$$

it follows that

$$\sin \bar{\phi} = \sin \phi_R \cosh \phi_I + j \cos \phi_R \sinh \phi_I \tag{84}$$

and combining equations (77), (80) and (84) that

$$S_\phi = \sqrt{\sin^2 (\phi \cos \theta_\phi) + \sinh^2 (\phi \sin \theta_\phi)} \tag{85}$$

$$\tan \theta_{s\phi} = \cot (\phi \cos \theta_\phi) \tanh (\phi \sin \theta_\phi) \tag{86}$$

The law of sines for a plane triangle is given by

$$\frac{\bar{a}}{\sin \bar{A}} = \frac{\bar{b}}{\sin \bar{B}} = \frac{\bar{c}}{\sin \bar{C}} \tag{87}$$

where

$$\bar{A} = Ae^{j\theta A} \tag{88}$$

with similar expressions for $\bar{B}$ and $\bar{C}$. It follows from equation (87) that

$$\frac{a}{S_A} = \frac{b}{S_B} = \frac{c}{S_C} \tag{89}$$

and

$$\theta_a - \theta_{sA} = \theta_b - \theta_{sB} = \theta_c - \theta_{sC} \tag{90}$$

and where

$$S_A = \sqrt{\sin^2 (A \cos \theta_A) + \sinh^2 (A \sin \theta_A)} \tag{91}$$

$$\tan \theta_{sA} = \cot (A \cos \theta_A) \tanh (A \sin \theta_A) \tag{92}$$

with similar expressions for S_B, S_C, θ_{sB}, and θ_{sC} . It should be noted that for spherical triangles equations (89) and (90) become respectively

$$\frac{S_a}{S_A} = \frac{S_b}{S_B} = \frac{S_c}{S_C} \tag{93}$$

and

$$\theta_{sa} - \theta_{sA} = \theta_{sb} - \theta_{sB} = \theta_{sc} - \theta_{sC} \tag{94}$$

Consider now simple plane areas located within a medium with broken internal symmetry. For example, the area of a triangle of sides $\bar{a}$, $\bar{b}$ and $\bar{c}$ with $\bar{\phi}$ = angle between sides $\bar{a}$ and $\bar{b}$ is given by

$$\bar{A} = \frac{1}{2} \bar{a}\bar{b} \sin \bar{\phi} = Ae^{j\theta_A} \tag{95}$$

where A = magnitude of area, and θ_A = phase angle of area. Combining equations (71), (72), (77), and (88) gives

$$A = \frac{1}{2} abS_\phi \tag{96}$$

$$\theta_A = \theta_a + \theta_b + \theta_{s\phi} \tag{97}$$

where S_ϕ and $\theta_{s\phi}$ are given by equations (85) and (86) respectively. Now consider the area of a circular sector of angle $\bar{\phi}$ which is

$$\bar{A} = \frac{1}{2}\bar{r}^2(\bar{\phi} - \sin\bar{\phi}) \tag{98}$$

$$= \frac{1}{2} r^2[\phi e^{j(2\theta_r+\theta_\phi)} - S_\phi e^{j(2\theta_r+\theta_{s\phi})}]$$

then it follows that

$$A\cos\theta_A = \frac{1}{2} r^2[\phi\cos(2\theta_r + \theta_\phi) - S_\phi\cos(2\theta_r + \theta_{s\phi})] \tag{99}$$

$$A\sin\theta_A = \frac{1}{2} r^2[\phi\sin(2\theta_r + \theta_\phi) - S_\phi\sin(2\theta_r + \theta_{s\phi})] \tag{100}$$

From equations (99) and (100) it follows that

$$\tan\theta_A = \frac{\phi\sin(2\theta_r + \theta_\phi) - S_\phi\sin(2\theta_r + \theta_{s\phi})}{\phi\cos(2\theta_r + \theta_\phi) - S_\phi\cos(2\theta_r + \theta_{s\phi})} \tag{101}$$

$$A^2 = \frac{1}{4} r^4\{\phi^2 + S_\phi^2 - 2\phi S_\phi\cos(\theta_\phi - \theta_{s\phi})\} \tag{102}$$

For a full circle obviously

$$A = \pi r^2 \tag{103}$$

$$\theta_A = 2\theta_r \tag{104}$$

For a rectangle of sides $\bar{x}$ and $\bar{y}$ one has

$$A = xy \tag{105}$$

$$\theta_A = \theta_x + \theta_y \tag{106}$$

For these cases, measured area = $A\cos\theta_A$.

Now consider various coordinate systems located in bulk matter or vacuum with broken internal symmetries. For example, for plane polar coordinates

$$\bar{x} = \bar{r}\cos\bar{\phi} = xe^{j\theta x} \tag{107}$$

$$\bar{y} = \bar{r}\sin\bar{\phi} = ye^{j\theta y} \tag{108}$$

and

$$\bar{x}^2 + \bar{y}^2 = \bar{r}^2 = r^2e^{2j\theta r} \tag{109}$$

The scalar equivalents of equations (107) and (108) are

$$x = rC_{\phi} \tag{110}$$

$$y = rS_{\phi} \tag{111}$$

$$\theta_x = \theta_r - \theta_{c\phi} \tag{112}$$

$$\theta_y = \theta_r + \theta_{s\phi} \tag{113}$$

The scalar equivalents of equation (109) are

$$x^2 \cos(2\theta_x) + y^2 \cos(2\theta_y) = r^2 \cos(2\theta_r) \tag{114}$$

$$x^2 \sin(2\theta_x) + y^2 \sin(2\theta_y) = r^2 \sin(2\theta_r) \tag{115}$$

or equivalently

$$r^4 = x^4 + y^4 + 2x^2y^2 \cos[2(\theta_x - \theta_y)] \tag{116}$$

and

$$\tan(2\theta_r) = \frac{x^2 \sin(2\theta_x) + y^2 \sin(2\theta_y)}{x^2 \cos(2\theta_x) + y^2 \cos(2\theta_y)} \tag{117}$$

Finally, substituting equations (110) through (113) into equation (116) gives

$$1 = C_{\phi}^4 + S_{\phi}^4 + 2C_{\phi}^2 S_{\phi}^2 \cos[2(\theta_{c\phi} + \theta_{s\phi})] \tag{118}$$

Consider now spherical coordinates located within bulk matter. For this system

$$\bar{x} = \bar{r} \sin\bar{\psi} \cos\bar{\phi} \tag{119}$$

$$\bar{y} = \bar{r} \sin\bar{\psi} \sin\bar{\phi} \tag{120}$$

$$\bar{z} = \bar{r} \cos\bar{\psi} \tag{121}$$

$$\bar{x}^2 + \bar{y}^2 + \bar{z}^2 = \bar{r}^2 \tag{122}$$

The scalar equivalent equations for equations (119) through (121) are

$$x = rS_{\psi}C_{\phi} \tag{123}$$

$$y = rS_{\psi}S_{\phi} \tag{124}$$

$$z = rC_{\psi} \tag{125}$$

and

$$\theta_x = \theta_r + \theta_{s\psi} - \theta_{c\phi} \tag{126}$$

$$\theta_y = \theta_r + \theta_{s\psi} + \theta_{s\phi} \tag{127}$$

$$\theta_z = \theta_r - \theta_{c\psi} \tag{128}$$

where C_{ϕ} and S_{ψ} are defined in equations (81) and (85) respectively, and $\theta_{c\phi}$ and $\theta_{s\psi}$ be equations (82) and (86) respectively. From equation (122) it follows that

$$x^2 \cos(2\theta_x) + y^2 \cos(2\theta_y) + z^2 \cos(2\theta_z) = r^2 \cos(2\theta_r) \tag{129}$$

$$x^2 \sin(2\theta_x) + y^2 \sin(2\theta_y) + z^2 \sin(2\theta_z) = r^2 \sin(2\theta_r) \tag{130}$$

Equations (129) and (130) give

$$r^4 = x^4 + y^4 + z^4 + 2x^2y^2\cos[2(\theta_x - \theta_y)] \tag{131}$$

$$+ 2y^2z^2 \cos[2(\theta_y - \theta_z)] + 2x^2z^2 \cos[2(\theta_x - \theta_z)]$$

$$\tan(2\theta_r) = \frac{x^2 \sin(2\theta_x) + y^2 \sin(2\theta_y) + z^2 \sin(2\theta_z)}{x^2 \cos(2\theta_x) + y^2 \cos(2\theta_y) + z^2 \cos(2\theta_z)} \tag{132}$$

From equations (123) through (125) and equation (131) it follows that

$$1 = S_{\psi}^4C_{\phi}^4 + S_{\psi}^4S_{\phi}^4 + C_{\psi}^4 + 2S_{\psi}^4C_{\phi}^2S_{\phi}^2 \cos[2(\theta_{c\phi} + \theta_{s\phi})] \tag{133}$$

$$+ 2S_{\psi}^2S_{\phi}^2C_{\psi}^2 \cos[2(\theta_{s\psi} + \theta_{s\phi} + \theta_{c\psi})]$$

$$+ 2S_{\psi}^2C_{\phi}^2C_{\psi}^2 \cos[2(\theta_{s\psi} - \theta_{c\phi} + \theta_{c\psi})]$$

The last type of coordinate system that will be considered is the polar space coordinates which utilizes direction cosines as follows

$$\bar{x} = \bar{r} \cos \bar{\alpha} \tag{134}$$

$$\bar{y} = \bar{r} \cos \bar{\beta} \tag{135}$$

$$\bar{z} = \bar{r} \cos \bar{\gamma} \tag{136}$$

$$\bar{r}^2 = \bar{x}^2 + \bar{y}^2 + \bar{z}^2 \tag{137}$$

It follows from equations (134) through (136) that

$$x = rC_\alpha \tag{138}$$

$$y = rC_\beta \tag{139}$$

$$z = rC_\gamma \tag{140}$$

$$\theta_x = \theta_r - \theta_{c\alpha} \tag{141}$$

$$\theta_y = \theta_r - \theta_{c\beta} \tag{142}$$

$$\theta_z = \theta_r - \theta_{c\gamma} \tag{143}$$

where

$$C_\alpha = \sqrt{\cos^2 (\alpha \cos \theta_\alpha) + \sinh^2 (\alpha \sin \theta_\alpha)} \tag{144}$$

$$\tan \theta_{c\alpha} = \tan (\alpha \cos \theta_\alpha) \tanh (\alpha \sin \theta_\alpha) \tag{145}$$

with similar expressions for C_β, C_γ, $\theta_{c\beta}$, and $\theta_{c\gamma}$. Equations (129) through (132) also hold for polar space coordinates. The equivalent of equation (133) for polar space coordinates is

$$1 = C_\alpha^4 + C_\beta^4 + C_\gamma^4 + 2C_\alpha^2 C_\beta^2 \cos [2(\theta_{c\beta} - \theta_{c\alpha})] \tag{146}$$

$$+ 2C_\beta^2 C_\gamma^2 \cos [2(\theta_{c\gamma} - \theta_{c\beta})] + 2C_\alpha^2 C_\gamma^2 \cos [2(\theta_{c\gamma} - \theta_{c\alpha})]$$

Consider now the case of rotation of coordinates in a plane that is located

within bulk matter or vacuum with broken symmetry. The values of the coordinates in the cartesian system that is rotated through an angle $\bar{\phi}$ are

$$\bar{x}' = \bar{x} \cos \bar{\phi} + \bar{y} \sin \bar{\phi} \tag{147}$$

$$\bar{y}' = \bar{y} \cos \bar{\phi} - \bar{x} \sin \bar{\phi} \tag{148}$$

The component equations for equation (147) are

$$x' \cos \theta_x' = xC_\phi \cos (\theta_x - \theta_{c\phi}) + yS_\phi \cos (\theta_y + \theta_{s\phi}) \tag{149}$$

$$x' \sin \theta_x' = xC_\phi \sin (\theta_x - \theta_{c\phi}) + yS_\phi \sin (\theta_y + \theta_{s\phi}) \tag{150}$$

while the component equations for equation (148) are

$$y' \cos \theta_y' = yC_\phi \cos (\theta_y - \theta_{c\phi}) - xS_\phi \cos (\theta_x + \theta_{s\phi}) \tag{151}$$

$$y' \sin \theta_y' = yC_\phi \sin (\theta_y - \theta_{c\phi}) - xS_\phi \sin (\theta_x + \theta_{s\phi}) \tag{152}$$

From equations (149) through (152) it follows that

$$(x')^2 = x^2C_\phi^2 + y^2S_\phi^2 + 2xyC_\phi S_\phi \cos [\theta_x - \theta_y - \theta_{c\phi} - \theta_{s\phi}] \tag{153}$$

$$(y')^2 = y^2C_\phi^2 + x^2S_\phi^2 - 2xyC_\phi S_\phi \cos [\theta_x - \theta_y + \theta_{c\phi} + \theta_{s\phi}] \tag{154}$$

The coordinate internal phase angles in the rotated system are given by

$$\tan \theta_x' = \frac{xC_\phi \sin (\theta_x - \theta_{c\phi}) + yS_\phi \sin (\theta_y + \theta_{s\phi})}{xC_\phi \cos (\theta_x - \theta_{c\phi}) + yS_\phi \cos (\theta_y + \theta_{s\phi})} \tag{155}$$

$$\tan \theta_y' = \frac{yC_\phi \sin (\theta_y - \theta_{c\phi}) - xS_\phi \sin (\theta_x + \theta_{s\phi})}{yC_\phi \cos (\theta_y - \theta_{c\phi}) - xS_\phi \cos (\theta_x + \theta_{s\phi})} \tag{156}$$

From equations (153) and (154) it follows that

$$(x')^2 + (y')^2 = (x^2 + y^2)(C_\phi^2 + S_\phi^2) + 4xyC_\phi S_\phi \sin (\theta_x - \theta_y) \sin (\theta_{c\phi} + \theta_{s\phi}) \tag{156A}$$

which reduces to the standard cartesian result when the internal phase angles are set equal to zero. The Lorentz group of rotations in spacetime are considered in an accompanying paper where Maxwell's equations with broken internal symmetry are considered.

4. BROKEN SYMMETRY OF KINEMATICAL AND DYNAMICAL VARIABLES. This section considers the effects of gauge rotated space and time on kinematics and dynamics. The gauge rotated space and time coordinates that were introduced in Section 2 can be used to define gauge rotated velocity and acceleration of particles located within bulk matter or the vacuum. For instance the components of the velocity of a particle are given by

$$\bar{v}_x = \frac{d\bar{x}}{d\bar{t}} = v_x e^{j\theta_{vx}} \tag{157}$$

$$\bar{v}_y = \frac{d\bar{y}}{d\bar{t}} = v_y e^{j\theta_{vy}} \tag{158}$$

$$\bar{v}_z = \frac{d\bar{z}}{d\bar{t}} = v_z e^{j\theta_{vz}} \tag{159}$$

where

$$v_x = \sqrt{\frac{\left(\frac{dx}{dt}\right)^2 + x^2\omega_{\theta x}^2}{1 + t^2\omega_{\theta t}^2}} \tag{160}$$

$$v_y = \sqrt{\frac{\left(\frac{dy}{dt}\right)^2 + y^2\omega_{\theta y}^2}{1 + t^2\omega_{\theta t}^2}} \tag{161}$$

$$v_z = \sqrt{\frac{\left(\frac{dz}{dt}\right)^2 + z^2\omega_{\theta z}^2}{1 + t^2\omega_{\theta t}^2}} \tag{162}$$

$$\theta_{vx} = \theta_x - \theta_t + \beta_{x,t} - \beta_{t,t} \tag{163}$$

$$\theta_{vy} = \theta_y - \theta_t + \beta_{y,t} - \beta_{t,t} \tag{164}$$

$$\theta_{vz} = \theta_z - \theta_t + \beta_{z,t} - \beta_{t,t} \tag{165}$$

where the internal angular velocities are given by

$$\omega_{\theta t} = d\theta_t/dt \qquad \omega_{\theta x} = d\theta_x/dt \tag{166}$$

$$\omega_{\theta y} = d\theta_y/dt \qquad \omega_{\theta z} = d\theta_z/dt \tag{167}$$

and where

$$\tan \beta_{x,t} = x \frac{d\theta_x/dt}{dx/dt} \tag{168}$$

$$\tan \beta_{y,t} = y \frac{d\theta_y/dt}{dy/dt} \tag{169}$$

$$\tan \beta_{z,t} = z \frac{d\theta_z/dt}{dz/dt} \tag{170}$$

and where $\beta_{t,t}$ is given by equation (41). The internal angular velocities can also be written as

$$\omega_{\theta x} = \frac{d\theta_x}{dt} = \frac{\partial\theta_x}{\partial t} + \frac{\partial\theta_x}{\partial x}\frac{dx}{dt} + \frac{\partial\theta_x}{\partial y}\frac{dy}{dt} + \frac{\partial\theta_x}{\partial z}\frac{dz}{dt} \tag{171}$$

$$\omega_{\theta y} = \frac{d\theta_y}{dt} = \frac{\partial\theta_y}{\partial t} + \frac{\partial\theta_y}{\partial x}\frac{dx}{dt} + \frac{\partial\theta_y}{\partial y}\frac{dy}{dt} + \frac{\partial\theta_y}{\partial z}\frac{dz}{dt} \tag{172}$$

$$\omega_{\theta z} = \frac{d\theta_z}{dt} = \frac{\partial\theta_z}{\partial t} + \frac{\partial\theta_z}{\partial x}\frac{dx}{dt} + \frac{\partial\theta_z}{\partial y}\frac{dy}{dt} + \frac{\partial\theta_z}{\partial z}\frac{dz}{dt} \tag{173}$$

$$\omega_{\theta t} = \frac{d\theta_t}{dt} = \frac{\partial\theta_t}{\partial t} + \frac{\partial\theta_t}{\partial x}\frac{dx}{dt} + \frac{\partial\theta_t}{\partial y}\frac{dy}{dt} + \frac{\partial\theta_t}{\partial z}\frac{dz}{dt} \tag{173A}$$

The conventional special relativistic momentum of a particle moving with a velocity v_x^a is given for a conventional dynamical system by the following standard formula[11]

$$p_x^a = m\gamma_x^a v_x^a \tag{174}$$

where m = mass, $v_x^a = dx_a/dt_a = dx_m/dt_m$ = conventionally calculated velocity, and γ_x^a = ordinary velocity factor (boost) given by[11]

$$\gamma_x^a = [1 - (v_x^a/c)^2]^{-1/2} \tag{175}$$

where c = light speed in the vacuum. These standard formulas are developed by considering the particle to be attached to a coordinate system moving with velocity $v = v_x^a$.[11] In this paper the generalization to bulk matter or vacuum with broken internal symmetries is made by considering the particle to be attached to a coordinate system moving with complex velocity $\bar{v} = \bar{v}_x$, so that the single particle momentum is written as

$$\bar{p}_x = m\bar{\gamma}_x\bar{v}_x = m\gamma_x v_x e^{j(\theta_{vx}+\theta_{\gamma x})} \tag{176}$$

where

$$\bar{\gamma}_x = \gamma_x e^{j\theta_{\gamma x}} = (1 - \bar{v}_x^2/c^2)^{-1/2} \tag{177}$$

gives the complex number velocity factor. The magnitude and phase angle of the complex number velocity factor is given by

$$\gamma_x = [1 - 2(v_x/c)^2 \cos(2\theta_{vx}) + (v_x/c)^4]^{-1/4} \tag{178}$$

$$\tan(2\theta_{\gamma x}) = \frac{(v_x/c)^2 \sin(2\theta_{vx})}{1 - (v_x/c)^2 \cos(2\theta_{vx})} \tag{179}$$

The results in equations (178) and (179) are obtained as a simple generalization of standard special relativity results to the case where space and time have intrinsic broken symmetry, and reduce to the standard result in equation (175) if the internal phase angles are set equal to zero. Note that the measured velocity is $v_{xm} = v_x \cos\theta_{vx} \neq v_x^a$.

The magnitude of the particle velocity is obtained by noting that the complex number particle velocity is written as

$$\bar{v} = ve^{j\theta_v} \tag{180}$$

and from equations (157) through (159) and equation (180) it follows that

$$\bar{v}^2 = \bar{v}_x^2 + \bar{v}_y^2 + \bar{v}_z^2 \tag{181}$$

or

$$v^2 e^{2j\theta_v} = v_x^2 e^{2j\theta_{vx}} + v_y^2 e^{2j\theta_{vy}} + v_z^2 e^{2j\theta_{vz}} \tag{182}$$

The component equations corresponding to equation (182) are

$$v^2 \cos(2\theta_v) = v_x^2 \cos(2\theta_{vx}) + v_y^2 \cos(2\theta_{vy}) + v_z^2 \cos(2\theta_{vz}) \quad (183)$$

$$v^2 \sin(2\theta_v) = v_x^2 \sin(2\theta_{vx}) + v_y^2 \sin(2\theta_{vy}) + v_z^2 \sin(2\theta_{vz}) \quad (184)$$

From equations (183) and (184) it follows that

$$v^4 = v_x^4 + v_y^4 + v_z^4 + 2v_x^2 v_y^2 \cos[2(\theta_{vx} - \theta_{vy})] \quad (185)$$

$$+ 2v_x^2 v_z^2 \cos[2(\theta_{vx} - \theta_{vz})] + 2v_y^2 v_z^2 \cos[2(\theta_{vy} - \theta_{vz})]$$

and

$$\tan(2\theta_v) = \frac{v_x^2 \sin(2\theta_{vx}) + v_y^2 \sin(2\theta_{vy}) + v_z^2 \sin(2\theta_{vz})}{v_x^2 \cos(2\theta_{vx}) + v_y^2 \cos(2\theta_{vy}) + v_z^2 \cos(2\theta_{vz})} \quad (186)$$

where v_x, v_y, v_z, θ_{vx}, θ_{vy}, and θ_{vz} are given by equations (160) through (165) respectively. The measured velocity = $v \cos \theta_v$.

The acceleration components are written as

$$\bar{a}_x = \frac{d\bar{v}_x}{d\bar{t}} = a_x e^{j\theta ax} \quad (187)$$

$$\bar{a}_y = \frac{d\bar{v}_y}{d\bar{t}} = a_y e^{j\theta ay} \quad (188)$$

$$\bar{a}_z = \frac{d\bar{v}_z}{d\bar{t}} = a_z e^{j\theta az} \quad (189)$$

where using the Eulerian derivative gives

$$\bar{a}_x = \frac{d\bar{v}_x}{d\bar{t}} = \frac{\partial \bar{v}_x}{\partial \bar{t}} + \bar{v}_x \frac{\partial \bar{v}_x}{\partial \bar{x}} + \bar{v}_y \frac{\partial \bar{v}_x}{\partial \bar{y}} + \bar{v}_z \frac{\partial \bar{v}_x}{\partial \bar{z}} \quad (190)$$

$$= \bar{a}_x^{(o)} + \bar{a}_x^{(1)} + \bar{a}_x^{(2)} + \bar{a}_x^{(3)}$$

$$= a_x^{(o)} e^{j\psi xo} + a_x^{(1)} e^{j\psi x1} + a_x^{(2)} e^{j\psi x2} + a_x^{(3)} e^{j\psi x3}$$

$$\bar{a}_y = \frac{d\bar{v}_y}{d\bar{t}} = \frac{\partial \bar{v}_y}{\partial \bar{t}} + \bar{v}_x \frac{\partial \bar{v}_y}{\partial \bar{x}} + \bar{v}_y \frac{\partial \bar{v}_y}{\partial \bar{y}} + \bar{v}_z \frac{\partial \bar{v}_y}{\partial \bar{z}} \tag{191}$$

$$= \bar{a}_y^{(o)} + \bar{a}_y^{(1)} + \bar{a}_y^{(2)} + \bar{a}_y^{(3)}$$

$$= a_y^{(o)} e^{j\psi_{yo}} + a_y^{(1)} e^{j\psi_{y1}} + a_y^{(2)} e^{j\psi_{y2}} + a_y^{(3)} e^{j\psi_{y3}}$$

$$\bar{a}_z = \frac{d\bar{v}_z}{d\bar{t}} = \frac{\partial \bar{v}_z}{\partial \bar{t}} + \bar{v}_x \frac{\partial \bar{v}_z}{\partial \bar{x}} + \bar{v}_y \frac{\partial \bar{v}_z}{\partial \bar{y}} + \bar{v}_z \frac{\partial \bar{v}_z}{\partial \bar{z}} \tag{192}$$

$$= \bar{a}_z^{(o)} + \bar{a}_z^{(1)} + \bar{a}_z^{(2)} + \bar{a}_z^{(3)}$$

$$= a_z^{(o)} e^{j\psi_{zo}} + a_z^{(1)} e^{j\psi_{z1}} + a_z^{(2)} e^{j\psi_{z2}} + a_z^{(3)} e^{j\psi_{z3}}$$

where

$$a_x^{(o)} = \sqrt{\frac{\left(\frac{\partial v_x}{\partial t}\right)^2 + v_x^2\left(\frac{\partial \theta_{vx}}{\partial t}\right)^2}{1 + t^2\left(\frac{\partial \theta_t}{\partial t}\right)^2}} \qquad a_x^{(1)} = v_x \sqrt{\frac{\left(\frac{\partial v_x}{\partial x}\right)^2 + v_x^2\left(\frac{\partial \theta_{vx}}{\partial x}\right)^2}{1 + x^2\left(\frac{\partial \theta_x}{\partial x}\right)^2}} \tag{193}$$

$$a_x^{(2)} = v_y \sqrt{\frac{\left(\frac{\partial v_x}{\partial y}\right)^2 + v_x^2\left(\frac{\partial \theta_{vx}}{\partial y}\right)^2}{1 + y^2\left(\frac{\partial \theta_y}{\partial y}\right)^2}} \qquad a_x^{(3)} = v_z \sqrt{\frac{\left(\frac{\partial v_x}{\partial z}\right)^2 + v_x^2\left(\frac{\partial \theta_{vx}}{\partial z}\right)^2}{1 + z^2\left(\frac{\partial \theta_z}{\partial z}\right)^2}} \tag{194}$$

$$a_y^{(o)} = \sqrt{\frac{\left(\frac{\partial v_y}{\partial t}\right)^2 + v_y^2\left(\frac{\partial \theta_{vy}}{\partial t}\right)^2}{1 + t^2\left(\frac{\partial \theta_t}{\partial t}\right)^2}} \qquad a_y^{(1)} = v_x \sqrt{\frac{\left(\frac{\partial v_y}{\partial x}\right)^2 + v_y^2\left(\frac{\partial \theta_{vy}}{\partial x}\right)^2}{1 + x^2\left(\frac{\partial \theta_x}{\partial x}\right)^2}} \tag{195}$$

$$a_y^{(2)} = v_y \sqrt{\frac{\left(\frac{\partial v_y}{\partial y}\right)^2 + v_y^2\left(\frac{\partial \theta_{vy}}{\partial y}\right)^2}{1 + y^2\left(\frac{\partial \theta_y}{\partial y}\right)^2}} \qquad a_y^{(3)} = v_z \sqrt{\frac{\left(\frac{\partial v_y}{\partial z}\right)^2 + v_y^2\left(\frac{\partial \theta_{vy}}{\partial z}\right)^2}{1 + z^2\left(\frac{\partial \theta_z}{\partial z}\right)^2}} \tag{196}$$

$$a_z^{(o)} = \sqrt{\frac{\left(\frac{\partial v_z}{\partial t}\right)^2 + v_z^2\left(\frac{\partial \theta_{vz}}{\partial t}\right)^2}{1 + t^2\left(\frac{\partial \theta_t}{\partial t}\right)^2}} \qquad a_z^{(1)} = v_x\sqrt{\frac{\left(\frac{\partial v_z}{\partial x}\right)^2 + v_z^2\left(\frac{\partial \theta_{vz}}{\partial x}\right)^2}{1 + x^2\left(\frac{\partial \theta_x}{\partial x}\right)^2}} \tag{197}$$

$$a_z^{(2)} = v_y\sqrt{\frac{\left(\frac{\partial v_z}{\partial y}\right)^2 + v_z^2\left(\frac{\partial \theta_{vz}}{\partial y}\right)^2}{1 + y^2\left(\frac{\partial \theta_y}{\partial y}\right)^2}} \qquad a_z^{(3)} = v_z\sqrt{\frac{\left(\frac{\partial v_z}{\partial z}\right)^2 + v_z^2\left(\frac{\partial \theta_{vz}}{\partial z}\right)^2}{1 + z^2\left(\frac{\partial \theta_z}{\partial z}\right)^2}} \tag{198}$$

$$\psi_{xo} = \theta_{vx} - \theta_t + \beta_{vx,t} - \beta_{t,t} \tag{199}$$

$$\psi_{x1} = 2\theta_{vx} - \theta_x + \beta_{vx,x} - \beta_{x,x} \tag{200}$$

$$\psi_{x2} = \theta_{vy} + \theta_{vx} - \theta_y + \beta_{vx,y} - \beta_{y,y} \tag{201}$$

$$\psi_{x3} = \theta_{vz} + \theta_{vx} - \theta_z + \beta_{vx,z} - \beta_{z,z} \tag{202}$$

$$\psi_{yo} = \theta_{vy} - \theta_t + \beta_{vy,t} - \beta_{t,t} \tag{203}$$

$$\psi_{y1} = \theta_{vx} + \theta_{vy} - \theta_x + \beta_{vy,x} - \beta_{x,x} \tag{204}$$

$$\psi_{y2} = 2\theta_{vy} - \theta_y + \beta_{vy,y} - \beta_{y,y} \tag{205}$$

$$\psi_{y3} = \theta_{vz} + \theta_{vy} - \theta_z + \beta_{vy,z} - \beta_{z,z} \tag{206}$$

$$\psi_{zo} = \theta_{vz} - \theta_t + \beta_{vz,t} - \beta_{t,t} \tag{207}$$

$$\psi_{z1} = \theta_{vx} + \theta_{vz} - \theta_x + \beta_{vz,x} - \beta_{x,x} \tag{208}$$

$$\psi_{z2} = \theta_{vy} + \theta_{vz} - \theta_y + \beta_{vz,y} - \beta_{y,y} \tag{209}$$

$$\psi_{z3} = 2\theta_{vz} - \theta_z + \beta_{vz,z} - \beta_{z,z} \tag{210}$$

where

$$\tan \beta_{vx,t} = v_x \frac{\partial \theta_{vx}/\partial t}{\partial v_x/\partial t} \qquad \tan \beta_{vx,x} = v_x \frac{\partial \theta_{vx}/\partial x}{\partial v_x/\partial x} \tag{211}$$

$$\tan \beta_{vx,y} = v_x \frac{\partial \theta_{vx}/\partial y}{\partial v_x/\partial y} \qquad \tan \beta_{vx,z} = v_x \frac{\partial \theta_{vx}/\partial z}{\partial v_x/\partial z} \tag{212}$$

$$\tan \beta_{vy,t} = v_y \frac{\partial \theta_{vy}/\partial t}{\partial v_y/\partial t} \qquad \tan \beta_{vy,x} = v_y \frac{\partial \theta_{vy}/\partial x}{\partial v_y/\partial x} \tag{213}$$

$$\tan \beta_{vy,y} = v_y \frac{\partial \theta_{vy}/\partial y}{\partial v_y/\partial y} \qquad \tan \beta_{vy,z} = v_y \frac{\partial \theta_{vy}/\partial z}{\partial v_y/\partial z} \tag{214}$$

$$\tan \beta_{vz,t} = v_z \frac{\partial \theta_{vz}/\partial t}{\partial v_z/\partial t} \qquad \tan \beta_{vz,x} = v_z \frac{\partial \theta_{vz}/\partial x}{\partial v_z/\partial x} \tag{215}$$

$$\tan \beta_{vz,y} = v_z \frac{\partial \theta_{vz}/\partial y}{\partial v_z/\partial y} \qquad \tan \beta_{vz,z} = v_z \frac{\partial \theta_{vz}/\partial z}{\partial v_z/\partial z} \tag{216}$$

Equations (199) through (210) can be further reduced by using equations (163) through (165).

Combining equations (187) through (189) with (190) through (192) gives

$$a_x \cos \theta_{ax} = a_x^{(o)} \cos \psi_{xo} + a_x^{(1)} \cos \psi_{x1} + a_x^{(2)} \cos \psi_{x2} + a_x^{(3)} \cos \psi_{x3} \tag{217}$$

$$a_x \sin \theta_{ax} = a_x^{(o)} \sin \psi_{xo} + a_x^{(1)} \sin \psi_{x1} + a_x^{(2)} \sin \psi_{x2} + a_x^{(3)} \sin \psi_{x3} \tag{218}$$

$$a_y \cos \theta_{ay} = a_y^{(o)} \cos \psi_{yo} + a_y^{(1)} \cos \psi_{y1} + a_y^{(2)} \cos \psi_{y2} + a_y^{(3)} \cos \psi_{y3} \tag{219}$$

$$a_y \sin \theta_{ay} = a_y^{(o)} \sin \psi_{yo} + a_y^{(1)} \sin \psi_{y1} + a_y^{(2)} \sin \psi_{y2} + a_y^{(3)} \sin \psi_{y3} \tag{220}$$

$$a_z \cos \theta_{az} = a_z^{(o)} \cos \psi_{zo} + a_z^{(1)} \cos \psi_{z1} + a_z^{(2)} \cos \psi_{z2} + a_z^{(3)} \cos \psi_{z3} \quad (221)$$

$$a_z \sin \theta_{az} = a_z^{(o)} \sin \psi_{zo} + a_z^{(1)} \sin \psi_{z1} + a_z^{(2)} \sin \psi_{z2} + a_z^{(3)} \sin \psi_{z3} \quad (222)$$

These equations can be used to determine a_x, a_y, a_z, θ_{ax}, θ_{ay}, and θ_{az}. For the special case when there is no spatial variation of the velocity field it follows that

$$a_x = a_x^{(o)} \quad (223)$$

$$a_y = a_y^{(o)} \quad (224)$$

$$a_z = a_z^{(o)} \quad (225)$$

$$\theta_{ax} = \psi_{xo} = \theta_x - 2\theta_t - 2\beta_{t,t} + \beta_{vx,t} + \beta_{x,t} \quad (226)$$

$$\theta_{ay} = \psi_{yo} = \theta_y - 2\theta_t - 2\beta_{t,t} + \beta_{vy,t} + \beta_{y,t} \quad (227)$$

$$\theta_{az} = \psi_{zo} = \theta_z - 2\theta_t - 2\beta_{t,t} + \beta_{vz,t} + \beta_{z,t} \quad (228)$$

The complex magnitude of the particle acceleration is written as

$$\bar{a} = ae^{j\theta_a} \quad (229)$$

and from equations (187) through (189) it follows that

$$\bar{a}^2 = \bar{a}_x^2 + \bar{a}_y^2 + \bar{a}_z^2 \quad (230)$$

The component equations corresponding to equation (230) are

$$a^2 \cos (2\theta_a) = a_x^2 \cos (2\theta_{ax}) + a_y^2 \cos (2\theta_{ay}) + a_z^2 \cos (2\theta_{az}) \quad (231)$$

$$a^2 \sin (2\theta_a) = a_x^2 \sin (2\theta_{ax}) + a_y^2 \sin (2\theta_{ay}) + a_z^2 \sin (2\theta_{az}) \quad (232)$$

It follows from equations (231) and (232) that

$$a^4 = a_x^4 + a_y^4 + a_z^4 + 2a_x^2 a_y^2 \cos [2(\theta_{ax} - \theta_{ay})] \quad (233)$$

$$+ 2a_x^2 a_z^2 \cos [2(\theta_{ax} - \theta_{az})] + 2a_y^2 a_z^2 \cos [2(\theta_{ay} - \theta_{az})]$$

and

$$\tan (2\theta_a) = \frac{a_x^2 \sin (2\theta_{ax}) + a_y^2 \sin (2\theta_{ay}) + a_z^2 \sin (2\theta_{az})}{a_x^2 \cos (2\theta_{ax}) + a_y^2 \cos (2\theta_{ay}) + a_z^2 \cos (2\theta_{az})} \tag{234}$$

where a_x, a_y, a_z, and θ_{ax}, θ_{ay}, θ_{az} are given by equations (217) through (222). The measured acceleration = $a \cos \theta_a$.

For a particle moving in bulk matter or vacuum with broken symmetry and not acted upon by forces, the momentum is constant and equation (176) gives

$$m\gamma_x v_x = C_{vx} \tag{235}$$

$$\theta_{vx} + \theta_{\gamma x} = C'_{vx} \tag{236}$$

where C_{vx} and C'_{vx} are constants of the motion. Equations (160), (163), (235), and (236) give

$$\frac{m^2\gamma_x^2\left[\left(\frac{dx}{dt}\right)^2 + x^2\omega_{\theta x}^2\right]}{1 + t^2\omega_{\theta t}^2} = C_{vx}^2 \tag{237}$$

$$\theta_{\gamma x} + \theta_x - \theta_t + \beta_{x,t} - \beta_{t,t} = C'_{vx} \tag{238}$$

Equation (237) shows that there is a transfer of energy between the linear motion and the internal phase motion. Equation (238) shows that there is also a transfer between θ_x and θ_t because equation (238) can be rewritten as

$$\theta_{\gamma x} + \theta_x + \tan^{-1}\left(x \frac{d\theta_x/dt}{dx/dt}\right) - \theta_t - \tan^{-1}(t\, d\theta_t/dt) = C'_{vx} \tag{239}$$

The nonrelativistic equations of motion of a particle moving in a potential field W^a are given by[11]

$$m\ddot{x}^a = m\dot{v}_x^a = ma_x^a = -\partial W^a/\partial x^a \tag{240}$$

$$m\ddot{y}^a = m\dot{v}_y^a = ma_y^a = -\partial W^a/\partial y^a \tag{241}$$

$$m\ddot{z}^a = m\dot{v}_z^a = ma_z^a = -\partial W^a/\partial z^a \tag{242}$$

The corresponding relativistic equations of motion for particles in a medium not having broken internal symmetry are[12]

$$m(\gamma_x^a)^3 a_x^a = -\partial W^a/\partial x^a \tag{243}$$

$$m\gamma_x^a a_y^a = -\partial W^a/\partial y^a \tag{244}$$

$$m\gamma_x^a a_z^a = -\partial W^a/\partial z^a \tag{245}$$

where the standard velocity factor is given by equation (175). Consider now a conservative force acting on a particle located in bulk matter or vacuum with broken internal symmetries in the space and time coordinates. If the complex number potential is written as

$$\bar{W} = We^{j\theta_W} \tag{246}$$

then the nonrelativistic equations of motion are written as

$$m\bar{a}_x = ma_x e^{j\theta_{ax}} = -\partial\bar{W}/\partial\bar{x} \tag{247}$$

$$m\bar{a}_y = ma_y e^{j\theta_{ay}} = -\partial\bar{W}/\partial\bar{y} \tag{248}$$

$$m\bar{a}_z = ma_z e^{j\theta_{az}} = -\partial\bar{W}/\partial\bar{z} \tag{249}$$

where $\bar{a}_x$, $\bar{a}_y$, and $\bar{a}_z$ are given by equations (190) through (192); and a_x, a_y, a_z, θ_{ax}, θ_{ay}, and θ_{az} are obtained from equations (217) through (222). If the theory of special relativity is considered in conjunction with broken internal symmetry, equations (247) through (249) become

$$m\bar{\gamma}_x^3\bar{a}_x = m\gamma_x^3 a_x e^{j(\theta_{ax}+3\theta_{\gamma x})} = -\partial\bar{W}/\partial\bar{x} \tag{250}$$

$$m\bar{\gamma}_x\bar{a}_y = m\gamma_x a_y e^{j(\theta_{ay}+\theta_{\gamma x})} = -\partial\bar{W}/\partial\bar{y} \tag{251}$$

$$m\bar{\gamma}_x\bar{a}_z = m\gamma_x a_z e^{j(\theta_{az}+\theta_{\gamma x})} = -\partial\bar{W}/\partial\bar{z} \tag{252}$$

where $\bar{\gamma}_x$ is given by equation (177), γ_x and $\theta_{\gamma x}$ are given by equations (178) and (179) respectively, and where the particle is moving instantaneously along the x axis with velocity $\bar{v}_x$. Equations (250) through (252) are simple generalizations of the standard special relativistic inertia terms to the case of particle motion in media with broken internal symmetries.

The derivatives of the broken symmetry potential can be written as

$$-\frac{\partial \bar{W}}{\partial \bar{x}} = -\frac{(dW + jWd\theta_W)}{(dx + jxd\theta_x)} e^{j(\theta_W - \theta_x)} \tag{253}$$

$$= \sqrt{\frac{\left(\frac{\partial W}{\partial x}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial x}\right)^2}{1 + x^2\left(\frac{\partial \theta_x}{\partial x}\right)^2}} e^{j(\pi + \theta_W + \beta_{W,x} - \theta_x - \beta_{x,x})}$$

$$-\frac{\partial \bar{W}}{\partial \bar{y}} = \sqrt{\frac{\left(\frac{\partial W}{\partial y}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial y}\right)^2}{1 + y^2\left(\frac{\partial \theta_y}{\partial y}\right)^2}} e^{j(\pi + \theta_W + \beta_{W,y} - \theta_y - \beta_{y,y})} \tag{254}$$

$$-\frac{\partial \bar{W}}{\partial \bar{z}} = \sqrt{\frac{\left(\frac{\partial W}{\partial z}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial z}\right)^2}{1 + z^2\left(\frac{\partial \theta_z}{\partial z}\right)^2}} e^{j(\pi + \theta_W + \beta_{W,z} - \theta_z - \beta_{z,z})} \tag{255}$$

where

$$\tan \beta_{W,x} = W \frac{\partial \theta_W / \partial x}{\partial W / \partial x} \tag{256}$$

$$\tan \beta_{W,y} = W \frac{\partial \theta_W / \partial y}{\partial W / \partial y} \tag{257}$$

$$\tan \beta_{W,z} = W \frac{\partial \theta_W / \partial z}{\partial W / \partial z} \tag{258}$$

and where $\beta_{x,x}$; $\beta_{y,y}$ and $\beta_{z,z}$ are given by equations (38) through (40) respectively. Combining equations (187) through (198) with equations (250) through (255) gives the following relativistic equations of motion for a particle located in bulk matter or vacuum with broken internal symmetries.

$$m\gamma_x^3 a_x = \sqrt{\frac{\left(\frac{\partial W}{\partial x}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial x}\right)^2}{1 + x^2\left(\frac{\partial \theta_x}{\partial x}\right)^2}} \tag{259}$$

$$m\gamma_x a_y = \sqrt{\frac{\left(\frac{\partial W}{\partial y}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial y}\right)^2}{1 + y^2\left(\frac{\partial \theta_y}{\partial y}\right)^2}} \tag{260}$$

$$m\gamma_x a_z = \sqrt{\frac{\left(\frac{\partial W}{\partial z}\right)^2 + W^2\left(\frac{\partial \theta_W}{\partial z}\right)^2}{1 + z^2\left(\frac{\partial \theta_z}{\partial z}\right)^2}} \tag{261}$$

$$\theta_{ax} + 3\theta_{\gamma x} = \pi + \theta_W + \beta_{W,x} - \theta_x - \beta_{x,x} \tag{262}$$

$$\theta_{ay} + \theta_{\gamma x} = \pi + \theta_W + \beta_{W,y} - \theta_y - \beta_{y,y} \tag{263}$$

$$\theta_{az} + \theta_{\gamma x} = \pi + \theta_W + \beta_{W,z} - \theta_z - \beta_{z,z} \tag{264}$$

where γ_x and $\theta_{\gamma x}$ are given by equations (178) and (179) respectively, and where a_x, a_y, a_z and θ_{ax}, θ_{ay}, and θ_{az} are obtained from equations (217) through (222).

A useful form of the nonrelativistic equations of motion for a particle located in asymmetric matter is obtained from equation (247) as follows

$$m\frac{d^2\bar{x}}{d\bar{t}^2} = -\partial\bar{W}/\partial\bar{x} \tag{264A}$$

where $\bar{x}$ is a complex number given by equation (16) whose real and imaginary components are written as $x_R = x\cos\theta_x$ and $x_I = x\sin\theta_x$. Then it follows that

$$\frac{d\bar{x}}{d\bar{t}} = \cos\beta_{t,t}\, e^{-j\theta_{dt}}\frac{d\bar{x}}{dt} = R_{lt}(x_R, x_I) + jI_{lt}(x_R, x_I) \tag{264B}$$

where

$$\theta_{dt} = \theta_t + \beta_{t,t} \tag{264C}$$

$$R_{1t}(x_R,x_I) = \cos\beta_{t,t}\left(\cos\theta_{dt}\frac{dx_R}{dt} + \sin\theta_{dt}\frac{dx_I}{dt}\right) \tag{264D}$$

$$\sim \cos\beta_{t,t}\cos(\theta_x - \theta_{dt})\frac{dx}{dt}$$

$$I_{1t}(x_R,x_I) = \cos\beta_{t,t}\left(-\sin\theta_{dt}\frac{dx_R}{dt} + \cos\theta_{dt}\frac{dx_I}{dt}\right) \tag{264E}$$

$$\sim \cos\beta_{t,t}\sin(\theta_x - \theta_{dt})\frac{dx}{dt}$$

Then it follows that

$$\frac{d^2\bar{x}}{dt^2} = \cos\beta_{t,t}\, e^{-j\theta_{dt}}\frac{d}{dt}\left(\cos\beta_{t,t}\, e^{-j\theta_{dt}}\frac{d\bar{x}}{dt}\right) \tag{264F}$$

$$\sim \cos^2\beta_{t,t}\, e^{-2j\theta_{dt}}\frac{d^2\bar{x}}{dt^2} = R_{2t}(x_R,x_I) + jI_{2t}(x_R,x_I)$$

where

$$R_{2t}(x_R,x_I) = \cos^2\beta_{t,t}\left[\cos(2\theta_{dt})\frac{d^2x_R}{dt^2} + \sin(2\theta_{dt})\frac{d^2x_I}{dt^2}\right] \tag{264G}$$

$$\sim \cos^2\beta_{t,t}\cos(\theta_x - 2\theta_{dt})\frac{d^2x}{dt^2}$$

$$I_{2t}(x_R,x_I) = \cos^2\beta_{t,t}\left[-\sin(2\theta_{dt})\frac{d^2x_R}{dt^2} + \cos(2\theta_{dt})\frac{d^2x_I}{dt^2}\right] \tag{264H}$$

$$\sim \cos^2\beta_{t,t}\sin(\theta_x - 2\theta_{dt})\frac{d^2x}{dt^2}$$

The derivative of the complex potential can be written by noting that $W_R = W\cos\theta_W$ and $W_I = W\sin\theta_W$ so that

$$\frac{\partial \bar{W}}{\partial \bar{x}} = \cos \beta_{x,x} \, e^{-j\theta_{dx}} \frac{\partial \bar{W}}{\partial x} = R_{1x}(W_R, W_I) + jI_{1x}(W_R, W_I) \tag{264I}$$

where

$$\theta_{dx} = \theta_x + \beta_{x,x} \tag{264J}$$

$$R_{1x}(W_R, W_I) = \cos \beta_{x,x} \left(\cos \theta_{dx} \frac{\partial W_R}{\partial x} + \sin \theta_{dx} \frac{\partial W_I}{\partial x} \right) \tag{264K}$$

$$\sim \cos \beta_{x,x} \cos (\theta_W - \theta_{dx}) \frac{\partial W}{\partial x}$$

$$I_{1x}(W_R, W_I) = \cos \beta_{x,x} \left(- \sin \theta_{dx} \frac{\partial W_R}{\partial x} + \cos \theta_{dx} \frac{\partial W_I}{\partial x} \right) \tag{264L}$$

$$\sim \cos \beta_{x,x} \sin (\theta_W - \theta_{dx}) \frac{\partial W}{\partial x}$$

Newton's dynamical equation (264A) can now be written in the following *approximate forms*

$$mR_{2t}(x_R, x_I) \sim - R_{1x}(W_R, W_I) \tag{264M}$$

$$mI_{2t}(x_R, x_I) \sim - I_{1x}(W_R, W_I) \tag{264N}$$

where R_{2t} and I_{2t} are given by equations (264G) and (264H) respectively, and R_{1x} and I_{1x} are given by equations (264K) and (264L) respectively. A further approximation for relations (264M) and (264N) yields

$$m \cos^2 \beta_{t,t} \cos (\theta_x - 2\theta_{dt}) \frac{d^2x}{dt^2} \sim - \cos \beta_{x,x} \cos (\theta_W - \theta_{dx}) \frac{\partial W}{\partial x} \tag{264O}$$

$$m \cos^2 \beta_{t,t} \sin (\theta_x - 2\theta_{dt}) \frac{d^2x}{dt^2} \sim - \cos \beta_{x,x} \sin (\theta_W - \theta_{dx}) \frac{\partial W}{\partial x} \tag{264P}$$

which gives

$$\theta_x - 2\theta_{dt} \sim \theta_W - \theta_{dx} \tag{264Q}$$

$$m \cos^2 \beta_{t,t} \frac{d^2x}{dt^2} \sim - \cos \beta_{x,x} \frac{\partial W}{\partial x} \tag{264R}$$

which are the approximate equations of motion for a particle in a potential field that is located in asymmetric bulk matter or vacuum. For a nonrelativistic system the measured acceleration is given by

$$a_{xm} = a_x \cos \theta_{ax} \sim \cos^2 \beta_{t,t} \cos \theta_{ax} \frac{d^2x}{dt^2} \tag{264S}$$

while the conventionally calculated acceleration is given by

$$a_x^a = \frac{d^2x_a}{dt_a^2} = \frac{d^2x_m}{dt_m^2} \sim \frac{\cos \theta_x}{\cos^2 \theta_t} \frac{d^2x}{dt^2} \tag{264T}$$

and therefore $a_{xm} \neq a_x^a$. Relations (264Q) and (264R) can be applied to many specific dynamical systems that are located in an asymmetric medium. For instance the vibration of molecules and atoms located in matter are expected to be described by these equations.

There are twenty three unknown variables that are needed to describe a particle in bulk matter or vacuum with broken internal symmetries: x, θ_x, y, θ_y, z, θ_z, v_x, θ_{vx}, v_y, θ_{vy} , v_z, θ_{vz}, a_x, θ_{ax}, a_y, θ_{ay}, a_z, θ_{az}, P, θ_p, γ, θ_γ, and θ_t . The magnitude of the time t is taken to be a totally independent parameter. Twenty two equations have been derived thus far in an attempt to determine the twenty three unknowns: two ground state relativistic equations (1), two equations for the ground state Grüneisen parameter (5), the six kinematic velocity equations (160) through (165), the six kinematic acceleration equations (217) through (222), and the six dynamical equations (259) through (264). By means of the twenty two equations the kinematical and dynamical variables have been expressed in terms of the potential components W and θ_W . But the single particle potential parameters W and θ_W are related through a gauge potential to the complex macroscopic state equation variables P, θ_p, γ, and θ_γ that are determined from equations (1) and (5). This connection is made through a partition function as shown in equations (13) through (15) which determine the gauge potential. But through equation (1), P, θ_p, γ, and θ_γ are related to the unrenormalized pressure and Grüneisen function P^a and γ^a respectively. Therefore it should be possible to express all of the kinematical and dynamical variables in terms of t, P^a, γ^a and the unrenormalized potential V^a .

Clearly one additional equation is necessary in order to have a total of twenty three equations that are needed to determine the twenty three unknown

variables. The needed equation is given by the following complex number continuity equation for broken symmetry matter

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{v}) = 0 \tag{265}$$

where ρ = mass density. Equation (265) can be rewritten as two real number equations as follows

$$G_t \cos \phi_t + G_x \cos \phi_x + G_y \cos \phi_y + G_z \cos \phi_z = 0 \tag{266}$$

$$G_t \sin \phi_t + G_x \sin \phi_x + G_y \sin \phi_y + G_z \sin \phi_z = 0 \tag{267}$$

where

$$G_t = \frac{\partial \rho / \partial t}{\sqrt{1 + t^2 \left(\frac{\partial \theta_t}{\partial t} \right)^2}} \tag{268A}$$

$$G_x = \sqrt{\frac{\left[\frac{\partial}{\partial x} (\rho v_x) \right]^2 + \rho^2 v_x^2 \left(\frac{\partial \theta_{vx}}{\partial x} \right)^2}{1 + x^2 \left(\frac{\partial \theta_x}{\partial x} \right)^2}} \tag{268B}$$

$$G_y = \sqrt{\frac{\left[\frac{\partial}{\partial y} (\rho v_y) \right]^2 + \rho^2 v_y^2 \left(\frac{\partial \theta_{vy}}{\partial y} \right)^2}{1 + y^2 \left(\frac{\partial \theta_y}{\partial y} \right)^2}} \tag{268C}$$

$$G_z = \sqrt{\frac{\left[\frac{\partial}{\partial z} (\rho v_z) \right]^2 + \rho^2 v_z^2 \left(\frac{\partial \theta_{vz}}{\partial z} \right)^2}{1 + z^2 \left(\frac{\partial \theta_z}{\partial z} \right)^2}} \tag{268D}$$

$$\phi_t = -\theta_t - \beta_{t,t} = -\theta_{dt} \tag{268E}$$

$$\phi_x = \theta_{vx} + \beta_{\rho vx,x} - \theta_x - \beta_{x,x} \tag{268F}$$

$$\phi_y = \theta_{vy} + \beta_{\rho vy,y} - \theta_y - \beta_{y,y} \tag{268G}$$

$$\phi_z = \theta_{vz} + \beta_{\rho vz,z} - \theta_z - \beta_{z,z} \tag{268H}$$

where

$$\tan \beta_{\rho v\alpha,\alpha} = \frac{\rho v_\alpha \dfrac{\partial \theta_{v\alpha}}{\partial \alpha}}{\dfrac{\partial}{\partial \alpha}(\rho v_\alpha)} \tag{269}$$

where $\alpha = x, y, z$. Therefore the complex number equation (265) has two real components that can be used along with the previously elaborated twenty two equations to obtain ρ and θ_t . There are now a total of twenty four equations and twenty four unknown quantities which can be determined to give

$$x = x(t,P^a,\gamma^a,V^a) \qquad \theta_x = \theta_x(t,P^a,\gamma^a,V^a) \tag{270A}$$

$$y = y(t,P^a,\gamma^a,V^a) \qquad \theta_y = \theta_y(t,P^a,\gamma^a,V^a) \tag{270B}$$

$$z = z(t,P^a,\gamma^a,V^a) \qquad \theta_z = \theta_z(t,P^a,\gamma^a,V^a) \tag{270C}$$

$$v_x = v_x(t,P^a,\gamma^a,V^a) \qquad \theta_{vx} = \theta_{vx}(t,P^a,\gamma^a,V^a) \tag{271A}$$

$$v_y = v_y(t,P^a,\gamma^a,V^a) \qquad \theta_{vy} = \theta_{vy}(t,P^a,\gamma^a,V^a) \tag{271B}$$

$$v_z = v_z(t,P^a,\gamma^a,V^a) \qquad \theta_{vz} = \theta_{vz}(t,P^a,\gamma^a,V^a) \tag{271C}$$

$$a_x = a_x(t,P^a,\gamma^a,V^a) \qquad \theta_{ax} = \theta_{ax}(t,P^a,\gamma^a,V^a) \tag{272A}$$

$$a_y = a_y(t,P^a,\gamma^a,V^a) \qquad \theta_{ay} = \theta_{ay}(t,P^a,\gamma^a,V^a) \tag{272B}$$

$$a_z = a_z(t,P^a,\gamma^a,V^a) \qquad \theta_{az} = \theta_{az}(t,P^a,\gamma^a,V^a) \tag{272C}$$

$$P = P(P^a,\gamma^a,V^a) \qquad \theta_P = \theta_P(P^a,\gamma^a,V^a) \tag{273A}$$

$$\gamma = \gamma(P^a,\gamma^a,V^a) \qquad \theta_\gamma = \theta_\gamma(P^a,\gamma^a,V^a) \tag{273B}$$

$$\rho = \rho(t,P^a,\gamma^a,V^a) \tag{274A}$$

$$\theta_t = \theta_t(t,P^a,\gamma^a,V^a) \tag{274B}$$

where time is treated as an independent variable, and where V^a = unrenormalized potential.

The first integral of equations (247) through (249) is

$$\frac{1}{2}m\bar{v}^2 + \bar{W} = \bar{E} \tag{275}$$

where $\bar{E}$ = complex number total energy, and where $\bar{v}^2$ is given by equation (181). The two scalar component equations corresponding to equation (275) are

$$\frac{1}{2}mv^2 \cos(2\theta_v) + W \cos\theta_W = E \cos\theta_E \tag{276}$$

$$\frac{1}{2}mv^2 \sin(2\theta_v) + W \sin\theta_W = E \sin\theta_E \tag{277}$$

where v and θ_v are given by equations (185) and (186) respectively. The corresponding first integral of the relativistic equations of motion (259) through (261) is[12]

$$(\bar{\gamma}_x - 1)mc^2 + \bar{W} = \bar{E} \tag{278}$$

where the particle is instantaneously moving along the x axis and $\bar{\gamma}_x$ is given by equation (177) The component form of equation (178) is written as

$$(\gamma_x \cos\theta_{\gamma x} - 1)mc^2 + W \cos\theta_W = E \cos\theta_E \tag{279}$$

$$\gamma_x \sin\theta_{\gamma x} mc^2 + W \sin\theta_W = E \sin\theta_E \tag{280}$$

The energy equations determine the magnitudes of coordinates and velocities.

5. ROTATING SYSTEMS IN BULK MATTER AND THE VACUUM. In Section 3 it was shown that the angle between two lines located in bulk matter or the vacuum with broken internal symmetry is expected to have an internal phase angle. This suggests that angular velocity also has a broken internal symmetry. Accordingly the angular speed associated with a complex number geometrical angle given by

$$\bar{\phi} = \phi e^{j\theta\phi} \tag{281}$$

is written as

$$\bar{\omega} = \omega e^{j\theta_\omega} = \frac{d\bar{\phi}}{d\bar{t}} = e^{j(\theta_\phi - \theta_t)} \left(\frac{d\phi + j\phi d\theta_\phi}{dt + jtd\theta_t} \right) \tag{282}$$

$$= \sqrt{\frac{\left(\frac{d\phi}{dt}\right)^2 + \phi^2\left(\frac{d\theta_\phi}{dt}\right)^2}{1 + t^2\left(\frac{d\theta_t}{dt}\right)^2}} \; e^{j(\theta_\phi + \beta_{\phi,t} - \theta_t - \beta_{t,t})}$$

so that

$$\omega = \sqrt{\frac{\left(\frac{d\phi}{dt}\right)^2 + \phi^2\left(\frac{d\theta_\phi}{dt}\right)^2}{1 + t^2\left(\frac{d\theta_t}{dt}\right)^2}} = \sqrt{\frac{\omega_\phi^2 + \phi^2\omega_{\theta\phi}^2}{1 + t^2\omega_{\theta t}^2}} \tag{283}$$

and

$$\theta_\omega = \theta_\phi + \beta_{\phi,t} - \theta_t - \beta_{t,t} \tag{284}$$

where

$$\tan \beta_{\phi,t} = \phi \frac{d\theta_\phi/dt}{d\phi/dt} \tag{285}$$

The angular speed associated with the internal phase angle of the geometrical phase angle is written as

$$\omega_{\theta\phi} = \frac{d\theta_\phi}{dt} = \frac{\partial\theta_\phi}{\partial t} + \frac{\partial\theta_\phi}{\partial r}\frac{dr}{dt} + \frac{\partial\theta_\phi}{\partial\phi}\frac{d\phi}{dt} \tag{286}$$

and the angular speed of the internal phase angle of the time coordinate $\omega_{\theta t}$ is given in equation (173A) or equivalently by

$$\omega_{\theta t} = \frac{d\theta_t}{dt} = \frac{\partial\theta_t}{\partial t} + \frac{\partial\theta_t}{\partial r}\frac{dr}{dt} + \frac{\partial\theta_t}{\partial\phi}\frac{d\phi}{dt} \tag{286A}$$

and finally where

$$\omega_\phi = \frac{d\phi}{dt} \tag{287}$$

is the ordinary angular speed associated with the magnitude ϕ of the geometrical angle. Equation (283) is the general expression for angular speed within bulk matter or vacuum with broken internal symmetries. The measured angular speed is given by $\omega \cos \theta_\omega$.

For short periods of time equation (283) shows that

$$\omega \sim \omega_\phi \tag{288}$$

while for long periods of time

$$\omega \sim \langle \omega_\phi \rangle \frac{\omega_{\theta\phi}}{\omega_{\theta t}} \tag{289}$$

where

$$\langle \omega_\phi \rangle = \frac{1}{t} \int_0^t \omega_\phi dt = \phi/t \tag{290}$$

In fact equation (283) shows that for a small t

$$\omega = \omega_\phi \left[1 + \frac{1}{2} \left(\langle \omega_\phi \rangle^2 \frac{\omega_{\theta\phi}^2}{\omega_\phi^2} - \omega_{\theta t}^2 \right) t^2 - \cdots \right] \tag{291}$$

for $\phi^2 \omega_{\vartheta\phi}^2 / \omega_\phi^2 << 1$ and $t^2 \omega_{\vartheta t}^2 << 1$, while for large t it follows that

$$\omega = \frac{\omega_{\theta\phi}}{\omega_{\theta t}} \langle \omega_\phi \rangle \left[1 + \frac{1}{2} \left(\frac{\omega_\phi^2}{\langle \omega_\phi \rangle^2 \omega_{\vartheta\phi}^2} - \frac{1}{\omega_{\theta t}^2} \right) t^{-2} + \cdots \right] \tag{292}$$

Consider now the velocity of a particle in bulk matter or vacuum that has a radial and a transverse component. The radial component is given by

$$\bar{v}_r = v_r e^{j\theta_{vr}} = \frac{d\bar{r}}{d\bar{t}} = e^{j(\theta_r - \theta_t)} \left(\frac{dr + jrd\theta_r}{dt + jtd\theta_t} \right) \tag{293}$$

$$= \sqrt{\frac{\left(\frac{dr}{dt}\right)^2 + r^2 \left(\frac{d\theta_r}{dt}\right)^2}{1 + t^2 \left(\frac{d\theta_t}{dt}\right)^2}} \; e^{j(\theta_r + \beta_{r,t} - \theta_t - \beta_{t,t})}$$

so that

$$v_r = \sqrt{\frac{\left(\frac{dr}{dt}\right)^2 + r^2\omega_{\theta r}^2}{1 + t^2\omega_{\theta t}^2}} \tag{294}$$

$$\theta_{vr} = \theta_r + \beta_{r,t} - \theta_t - \beta_{t,t} \tag{295}$$

where

$$\tan \beta_{r,t} = r \frac{d\theta_r/dt}{dr/dt} \tag{296}$$

$$\omega_{\theta r} = d\theta_r/dt = \frac{\partial \theta_r}{\partial t} + \frac{\partial \theta_r}{\partial r}\frac{dr}{dt} + \frac{\partial \theta_r}{\partial \phi}\frac{d\phi}{dt} \tag{297}$$

The transverse component of velocity is given by

$$\bar{v}_\phi = v_\phi e^{j\theta v\phi} = \bar{r}\frac{d\bar{\phi}}{d\bar{t}} = \bar{r}\bar{\omega} \tag{298}$$

Combining equations (282) and (298) gives

$$v_\phi = r\omega \tag{299}$$

$$\theta_{v\phi} = \theta_r + \theta_\phi + \beta_{\phi,t} - \theta_t - \beta_{t,t} = \theta_r + \theta_\omega \tag{300}$$

where ω is given by equation (283). The magnitude of the vector sum of the radial and transverse velocities is given by

$$\bar{v}^2 = \bar{v}_r^2 + \bar{v}_\phi^2 = v^2 e^{2j\theta_v} \tag{301}$$

which has the following scalar components

$$v^2 \cos(2\theta_v) = v_r^2 \cos(2\theta_{vr}) + v_\phi^2 \cos(2\theta_{v\phi}) \tag{302}$$

$$v^2 \sin(2\theta_v) = v_r^2 \sin(2\theta_{vr}) + v_\phi^2 \sin(2\theta_{v\phi}) \tag{303}$$

From equation (302) and (303) it follows that

$$v^4 = v_r^4 + v_\phi^4 + 2v_r^2 v_\phi^2 \cos\left[2(\theta_{vr} - \theta_{v\phi})\right] \tag{304}$$

$$\tan(2\theta_v) = \frac{v_r^2 \sin(2\theta_{vr}) + v_\phi^2 \sin(2\theta_{v\phi})}{v_r^2 \cos(2\theta_{vr}) + v_\phi^2 \cos(2\theta_{v\phi})} \tag{305}$$

where

$$\theta_{vr} - \theta_{v\phi} = \beta_{r,t} - \theta_\phi - \beta_{\phi,t} \tag{306}$$

The measured speed = $v \cos \theta_v$.

For ordinary matter rotating about a center of force, the radial and transverse accelerations are written as[11]

$$a_r^a = \frac{d^2 r_a}{dt_a^2} - \omega_a^2 r_a \tag{307}$$

$$a_\phi^a = r_a \frac{d^2 \phi_a}{dt_a^2} + 2 \frac{dr_a}{dt_a} \omega_a \tag{308}$$

The acceleration of a particle that is orbiting about a center of force located within bulk matter or the vacuum with broken internal symmetry will have the following radial and transverse components

$$\bar{a}_r = \frac{d^2 \bar{r}}{d\bar{t}^2} - \bar{\omega}^2 \bar{r} = a_r e^{j\theta_{ar}} \tag{309}$$

$$\bar{a}_\phi = \bar{r} \frac{d^2 \bar{\phi}}{d\bar{t}^2} + 2 \frac{d\bar{r}}{d\bar{t}} \bar{\omega} = a_\phi e^{j\theta_{a\phi}} \tag{310}$$

Each of the four terms in equations (309) and (310) can be evaluated in terms of previously calculated quantities.

The linear radial acceleration term in equation (309) is given by

$$\frac{d^2 \bar{r}}{d\bar{t}^2} = \frac{d\bar{v}_r}{d\bar{t}} = \bar{a}_{rr} = a_{rr} e^{j\theta_{arr}} \tag{311}$$

where the time derivative of the radial velocity is given by the Eulerian derivative as follows

$$\bar{a}_{rr} = \frac{d\bar{v}_r}{d\bar{t}} = \frac{\partial \bar{v}_r}{\partial \bar{t}} + \bar{v}_r \frac{\partial \bar{v}_r}{\partial \bar{r}} + \frac{\bar{v}_\phi}{\bar{r}} \frac{\partial \bar{v}_r}{\partial \bar{\phi}} \tag{312}$$

$$= \bar{a}_{rr}^{(o)} + \bar{a}_{rr}^{(1)} + \bar{a}_{rr}^{(2)}$$

$$= a_{rr}^{(o)} e^{j\psi_{ro}} + a_{rr}^{(1)} e^{j\psi_{r1}} + a_{rr}^{(2)} e^{j\psi_{r2}}$$

where

$$a_{rr}^{(o)} = \sqrt{\frac{\left(\frac{\partial v_r}{\partial t}\right)^2 + v_r^2\left(\frac{\partial \theta_{vr}}{\partial t}\right)^2}{1 + t^2\left(\frac{\partial \theta_t}{\partial t}\right)^2}} \tag{313A}$$

$$a_{rr}^{(1)} = v_r \sqrt{\frac{\left(\frac{\partial v_r}{\partial r}\right)^2 + v_r^2\left(\frac{\partial \theta_{vr}}{\partial r}\right)^2}{1 + r^2\left(\frac{\partial \theta_r}{\partial r}\right)^2}} \tag{313B}$$

$$a_{rr}^{(2)} = \frac{v_\phi}{r} \sqrt{\frac{\left(\frac{\partial v_r}{\partial \phi}\right)^2 + v_r^2\left(\frac{\partial \theta_{vr}}{\partial \phi}\right)^2}{1 + \phi^2\left(\frac{\partial \theta_\phi}{\partial \phi}\right)^2}} \tag{313C}$$

$$\psi_{ro} = \theta_{vr} + \beta_{vr,t} - \theta_t - \beta_{t,t} \tag{313D}$$

$$\psi_{r1} = 2\theta_{vr} - \theta_r + \beta_{vr,r} - \beta_{r,r} \tag{313E}$$

$$\psi_{r2} = \theta_{v\phi} + \theta_{vr} - \theta_r - \theta_\phi + \beta_{vr,\phi} - \beta_{\phi,\phi} \tag{313F}$$

where $\beta_{t,t}$, $\beta_{r,r}$, and $\beta_{\phi,\phi}$ are given by equations (41), (64), and (66) respectively, and v_r and θ_{vr} are given by equations (294) and (295) respectively, and v_ϕ and $\theta_{v\phi}$ are given by equations (299) and (300) respectively, and where

$$\tan \beta_{vr,t} = v_r \frac{\partial \theta_{vr}/\partial t}{\partial v_r/\partial t} \tag{314A}$$

$$\tan \beta_{vr,r} = v_r \frac{\partial \theta_{vr}/\partial r}{\partial v_r/\partial r} \tag{314B}$$

$$\tan \beta_{vr,\phi} = v_r \frac{\partial \theta_{vr}/\partial \phi}{\partial v_r/\partial \phi} \tag{314C}$$

From equations (331) and (312) it follows that

$$a_{rr} \cos \theta_{arr} = a_{rr}^{(o)} \cos \psi_{ro} + a_{rr}^{(1)} \cos \psi_{r1} + a_{rr}^{(2)} \cos \psi_{r2} \tag{315}$$

$$a_{rr} \sin \theta_{arr} = a_{rr}^{(o)} \sin \psi_{ro} + a_{rr}^{(1)} \sin \psi_{r1} + a_{rr}^{(2)} \sin \psi_{r2} \tag{316}$$

from which a_{rr} and θ_{arr} can be obtained immediately. For the special case where there is no spatial variation of the velocity field the acceleration equation become

$$a_{rr} = a_{rr}^{(o)} \tag{317}$$

$$\theta_{arr} = \psi_{ro} \tag{318}$$

The centrifugal radial acceleration term in equation (309) is written as

$$\bar{\bar{r}}\omega^2 = r\omega^2 e^{j(\theta_r + 2\theta_\omega)} = a_{cen} e^{j\theta_{acen}} \tag{319}$$

so that

$$a_{cen} = r\omega^2 \tag{320}$$

$$\theta_{acen} = \theta_r + 2\theta_\omega = \theta_r + 2(\theta_\phi + \beta_{\phi,t} - \theta_t - \beta_{t,t}) \tag{321}$$

where ω and θ_ω are given by equations (283) and (284) respectively.

The first term in the angular acceleration given by equation (310) is written as

$$\bar{r}\,\frac{d^2\bar{\phi}}{d\bar{t}^2} = \bar{r}\,\frac{d\bar{\omega}}{d\bar{t}} = \bar{a}_{\phi\phi} = a_{\phi\phi}e^{j\theta}a_{\phi\phi} \tag{322}$$

where $\bar{\omega}$ is given by equation (282). The time derivative is taken to be an Eulerian derivative (which accounts for differential rotation) as follows

$$\bar{a}_{\phi\phi} = \bar{r}\,\frac{d\bar{\omega}}{d\bar{t}} = \bar{r}\left(\frac{\partial\bar{\omega}}{\partial\bar{t}} + \bar{v}_r\,\frac{\partial\bar{\omega}}{\partial\bar{r}} + \frac{\bar{v}_\phi}{\bar{r}}\,\frac{\partial\bar{\omega}}{\partial\bar{\phi}}\right) \tag{323}$$

$$= \bar{a}^{(o)}_{\phi\phi} = \bar{a}^{(1)}_{\phi\phi} + \bar{a}^{(2)}_{\phi\phi}$$

$$= a^{(o)}_{\phi\phi}e^{j\psi_{\phi o}} + a^{(1)}_{\phi\phi}e^{j\psi_{\phi 1}} + a^{(2)}_{\phi\phi}e^{j\psi_{\phi 2}}$$

where

$$a^{(o)}_{\phi\phi} = r\sqrt{\frac{\left(\frac{\partial\omega}{\partial t}\right)^2 + \omega^2\left(\frac{\partial\theta_\omega}{\partial t}\right)^2}{1 + t^2\left(\frac{\partial\theta_t}{\partial t}\right)^2}} \tag{324A}$$

$$a^{(1)}_{\phi\phi} = rv_r\sqrt{\frac{\left(\frac{\partial\omega}{\partial r}\right)^2 + \omega^2\left(\frac{\partial\theta_\omega}{\partial r}\right)^2}{1 + r^2\left(\frac{\partial\theta_r}{\partial r}\right)^2}} \tag{324B}$$

$$a^{(2)}_{\phi\phi} = v_\phi\sqrt{\frac{\left(\frac{\partial\omega}{\partial\phi}\right)^2 + \omega^2\left(\frac{\partial\theta_\omega}{\partial\phi}\right)^2}{1 + \phi^2\left(\frac{\partial\theta_\phi}{\partial\phi}\right)^2}} \tag{324C}$$

$$\psi_{\phi o} = \theta_r + \theta_\omega + \beta_{\omega,t} - \theta_t - \beta_{t,t} \tag{324D}$$

$$\psi_{\phi 1} = \theta_{vr} + \theta_\omega + \beta_{\omega,r} - \beta_{r,r} \tag{324E}$$

$$\psi_{\phi 2} = \theta_{v\phi} + \theta_\omega + \beta_{\omega,\phi} - \theta_\phi - \beta_{\phi,\phi} \tag{324F}$$

where

$$\tan \beta_{\omega,t} = \omega \frac{\partial\theta_\omega/\partial t}{\partial\omega/\partial t} \tag{325A}$$

$$\tan \beta_{\omega,r} = \omega \frac{\partial\theta_\omega/\partial r}{\partial\omega/\partial r} \tag{325B}$$

$$\tan \beta_{\omega,\phi} = \omega \frac{\partial\theta_\omega/\partial \phi}{\partial\omega/\partial \phi} \tag{325C}$$

From equations (322) and (323) it follows that

$$a_{\phi\phi} \cos \theta_{a\phi\phi} = a_{\phi\phi}^{(o)} \cos \psi_{\phi o} + a_{\phi\phi}^{(1)} \cos \psi_{\phi 1} + a_{\phi\phi}^{(2)} \cos \psi_{\phi 2} \tag{326}$$

$$a_{\phi\phi} \sin \theta_{a\phi\phi} = a_{\phi\phi}^{(o)} \sin \psi_{\phi o} + a_{\phi\phi}^{(1)} \sin \psi_{\phi 1} + a_{\phi\phi}^{(2)} \sin \phi_{\phi 2} \tag{327}$$

from which $a_{\phi\phi}$ and $\theta_{a\phi\phi}$ can be immediately obtained. For the special case where there is no spatial variation of the angular velocity (uniform rotation) it follows that

$$a_{\phi\phi} = a_{\phi\phi}^{(o)} \tag{328}$$

$$\theta_{a\phi\phi} = \psi_{\phi o} \tag{329}$$

Finally the Coriolis term in equation (310) is written as

$$2\bar{\omega}\bar{v}_r = a_c e^{j\theta_{ac}} = 2\omega v_r e^{j(\theta_\omega + \theta_{vr})} \tag{330}$$

and therefore for the Coriolis acceleration

$$a_c = 2\omega v_r \tag{331}$$

$$\theta_{ac} = \theta_\omega + \theta_{vr} = \theta_r + \beta_{r,t} + \theta_\phi + \beta_{\phi,t} - 2(\theta_t + \beta_{t,t}) \tag{332}$$

where v_r is given by equation (294), and θ_ω and θ_{vr} by equations (284) and (295) respectively. Combining equations (284), (324D), (329), and (332) shows that for uniform rotation

$$\theta_{a\phi\phi} = \theta_{ac} - \beta_{r,t} + \beta_{\omega,t} \tag{333}$$

All of the terms in equations (309) and (310) have been evaluated and these equations can be written as

$$\bar{a}_r = a_r e^{j\theta_{ar}} = a_{rr} e^{j\theta_{arr}} - a_{cen} e^{j\theta_{acen}} \tag{334}$$

$$\bar{a}_\phi = a_\phi e^{j\theta_{a\phi}} = a_{\phi\phi} e^{j\theta_{a\phi\phi}} + a_c e^{j\theta_{ac}} \tag{335}$$

The magnitudes and internal phase angles of the radial and transverse components of the acceleration a_r, θ_{ar}, a_ϕ, and $\theta_{a\phi}$ have yet to be calculated. This is done using equations (334) and (335). From equation (334) it follows that

$$a_r \cos \theta_{ar} = a_{rr} \cos \theta_{arr} - a_{cen} \cos \theta_{acen} \tag{336}$$

$$a_r \sin \theta_{ar} = a_{rr} \sin \theta_{arr} - a_{cen} \sin \theta_{acen} \tag{337}$$

and

$$\tan \theta_{ar} = \frac{a_{rr} \sin \theta_{arr} - a_{cen} \sin \theta_{acen}}{a_{rr} \cos \theta_{arr} - a_{cen} \cos \theta_{acen}} \tag{338}$$

$$a_r^2 = a_{rr}^2 + a_{cen}^2 - 2a_{rr}a_{cen} \cos (\theta_{arr} - \theta_{acen}) \tag{239}$$

From equation (335) it follows that

$$a_\phi \cos \theta_{a\phi} = a_{\phi\phi} \cos \theta_{a\phi\phi} + a_c \cos \theta_{ac} \tag{340}$$

$$a_\phi \sin \theta_{a\phi} = a_{\phi\phi} \sin \theta_{a\phi\phi} + a_c \sin \theta_{ac} \tag{341}$$

and

$$\tan \theta_{a\phi} = \frac{a_{\phi\phi} \sin \theta_{a\phi\phi} + a_c \sin \theta_{ac}}{a_{\phi\phi} \cos \theta_{a\phi\phi} + a_c \cos \theta_{ac}} \tag{342}$$

$$a_\phi^2 = a_{\phi\phi}^2 + a_c^2 + 2a_c a_{\phi\phi} \cos (\theta_{a\phi\phi} - \theta_{ac}) \tag{343}$$

In order to complete the calculation of the acceleration, the magnitude and phase angle of the vector sum of the radial and transverse components of acceleration need to be calculated. The complex number magnitude of the vector sum will be written as

$$\bar{a} = ae^{j\theta_a} \tag{344}$$

so that

$$\bar{a}^2 = \bar{a}_\phi^2 + \bar{a}_r^2 \tag{345}$$

from which it follows that

$$a^2 \cos(2\theta_a) = a_\phi^2 \cos(2\theta_{a\phi}) + a_r^2 \cos(2\theta_{ar}) \tag{346}$$

$$a^2 \sin(2\theta_a) = a_\phi^2 \sin(2\theta_{a\phi}) + a_r^2 \sin(2\theta_{ar}) \tag{347}$$

where a_ϕ and $\theta_{a\phi}$ are given by equations (343) and (342) respectively, and a_r and θ_{ar} are given by equations (339) and (338) respectively. From equations (346) and (347) it follows that

$$a^4 = a_\phi^4 + a_r^4 + 2a_r^2 a_\phi^2 \cos[2(\theta_{a\phi} - \theta_{ar})] \tag{348}$$

$$\tan(2\theta_a) = \frac{a_\phi^2 \sin(2\theta_{a\phi}) + a_r^2 \sin(2\theta_{ar})}{a_\phi^2 \cos(2\theta_{a\phi}) + a_r^2 \cos(2\theta_{ar})} \tag{349}$$

The measured acceleration is equal to $a \cos \theta_a$.

The relativistic force equations for a particle moving in bulk matter or vacuum with broken internal symmetry are best written in terms of normal and tangential components. The equations of motion of a particle under the action of normal and tangential forces are written as[11]

$$\bar{F}_N = m\bar{\gamma}_T \bar{a}_N = m\gamma_T a_N e^{j(\theta_{\gamma T} + \theta_{aN})} \tag{350A}$$

$$\bar{F}_T = m\bar{\gamma}_T^3 \bar{a}_T = m\gamma_T^3 a_T e^{j(3\theta_{\gamma T} + \theta_{aT})} \tag{350B}$$

where $\bar{F}_N$ amd $\bar{F}_T$ = normal and transverse complex number forces, $\bar{a}_N$ and $\bar{a}_T$ = complex number normal and transverse accelerations written as

$$\bar{a}_N = a_N e^{j\theta_{aN}} \tag{351A}$$

$$\bar{a}_T = a_T e^{j\theta_{aT}} \tag{351B}$$

and where the transverse velocity boost is written as

$$\bar{\gamma}_T = \gamma_T e^{j\theta\gamma T} = (1 - \bar{v}_T^2/c^2)^{-1/2} \tag{352A}$$

$$\bar{v}_T = v_T e^{j\theta vT} \tag{352B}$$

with

$$\gamma_T = \left[1 - 2(v_T/c)^2 \cos(2\theta_{vT}) + (v_T/c)^4\right]^{-1/4} \tag{353}$$

$$\tan(2\theta_{\gamma T}) = \frac{(v_T/c)^2 \sin(2\theta_{vT})}{1 - (v_T/c)^2 \cos(2\theta_{vT})} \tag{354}$$

Consider now the question of the conservation of angular momentum of a body under the action of a radial force field in uniformly rotating bulk matter or vacuum with broken internal symmetry. For a radial force field in a broken symmetry system, equations (340) and (341) become

$$a_{\phi\phi} \cos\theta_{a\phi\phi} + a_c \cos\theta_{ac} = 0 \tag{355}$$

$$a_{\phi\phi} \sin\theta_{a\phi\phi} + a_c \sin\theta_{ac} = 0 \tag{356}$$

In order for equations (355) and (356) to be satisfied, remembering that $a_{\phi\phi} > 0$ and $a_c > 0$, the following conditions must hold

$$a_{\phi\phi} - a_c = 0 \tag{357}$$

$$\tan\theta_{a\phi\phi} = \tan\theta_{ac} \tag{358}$$

or

$$\theta_{a\phi\phi} = \theta_{ac} - \pi \tag{359}$$

For uniform rotation and a radial force, the combination of equations (323), (324A), (331), (294), and (357) gives the following equation

$$r\sqrt{\left(\frac{d\omega}{dt}\right)^2 + \omega^2\left(\frac{d\theta_\omega}{dt}\right)^2} - 2\omega\sqrt{\left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta_r}{dt}\right)^2} = 0 \tag{360}$$

Because $d\omega/dr < 0$ equation (360) can be written as

$$r \frac{d\omega}{dr} \sqrt{1 + \left(\omega \frac{d\theta_\omega}{d\omega}\right)^2} + 2\omega \sqrt{1 + \left(r \frac{d\theta_r}{dr}\right)^2} = 0 \tag{361}$$

Combining equations (333) and (359) gives for uniform rotation and a central force

$$\beta_{\omega,t} = \beta_{r,t} - \pi \tag{362}$$

From equation (362) it follows that for a radial force and uniform rotation

$$\tan \beta_{\omega,t} = \tan \beta_{r,t} \tag{363}$$

Combining equations (296) and (325A) with equation (363) gives

$$r \frac{d\theta_r}{dr} = \omega \frac{d\theta_\omega}{d\omega} \tag{364}$$

Substituting equation (364) into equation (361) gives

$$r \frac{d\omega}{dr} + 2\omega = 0 \tag{365}$$

a differential equation whose solution is

$$\omega r^2 = \text{constant} \tag{366}$$

where ω is given by equation (283). Dividing equations (364) and (365) gives also

$$2\theta_r + \theta_\omega = 2\theta_r + \theta_\phi + \beta_{\phi,t} - \theta_t - \beta_{t,t} = \text{constant} \tag{367}$$

so that in fact combining equations (366) and (367) gives

$$\bar{\omega}\bar{r}^2 = \text{constant} \tag{368}$$

which is the expression for the conservation of angular momentum for a particle of unit mass uniformly rotating in a central force field that is located in bulk matter or vacuum wherein the space and time coordinates exhibit a broken symmetry.

Equations (283) and (366) show that

$$r^2 \sqrt{\frac{\omega_\phi^2 + \phi^2 \omega_{\theta\phi}^2}{1 + t^2 \omega_{\theta t}^2}} = \text{constant} \tag{369}$$

Equation (369) allows a connection to be made between the $t = 0$ and $t = \infty$ rotational states of a central force system located in bulk matter or vacuum with broken internal symmetries namely

$$r_o^2 \omega_\phi(0) = r_\infty^2 \langle \omega_\phi(\infty) \rangle \frac{\omega_{\theta\phi}(\infty)}{\omega_{\theta t}(\infty)} \tag{370}$$

where

$$\langle \omega_\phi(\infty) \rangle = \lim_{t \to \infty} \frac{\phi}{t} \tag{371}$$

In a similar way equation (367) allows a connection to be made between the $t = 0$ and $t = \infty$ values of the internal phase angles of the coordinates of a particle in a central force system located in bulk matter or vacuum

$$2\theta_r(0) + \theta_\phi(0) + \beta_{\phi,t}(0) - \theta_t(0) - \beta_{t,t}(0) \tag{372}$$

$$= 2\theta_r(\infty) + \theta_\phi(\infty) + \beta_{\phi,t}(\infty) - \theta_t(\infty) - \beta_{t,t}(\infty)$$

Equation (369) shows that rotational motion is shared between external and internal angular motions, and this equation may perhaps be of value for describing the rotation of galaxies, neutron stars, molecules, atoms, and atomic nuclei where internal angular motions may exist.

A special case of interest, expecially for gravitationally bound systems such as stars or planets, is the situation where $\theta_r \neq 0$ but $d\theta_r/dt = 0$, $\theta_t = 0$ and $\theta_\phi = 0$. This gives the following results

$$v_r = \frac{dr}{dt} \qquad v_\phi = r \frac{d\phi}{dt} \qquad \omega = \omega_\phi = \frac{d\phi}{dt} \tag{373}$$

$$a_{rr} = \frac{dv_r}{dt} = \frac{d^2 r}{dt^2} \qquad a_r = a_{rr} - r\omega_\phi^2 \tag{374}$$

$$a_{\phi\phi} = r \frac{d\omega_\phi}{dt} = r \frac{d^2\phi}{dt^2} \qquad a_c = 2\omega_\phi \frac{dr}{dr} \qquad a_\phi = a_{\phi\phi} + a_c \tag{375}$$

$$a^2 = a_r^2 + a_\phi^2 \qquad v^2 = v_r^2 + v_\phi^2 \tag{376}$$

$$\theta_{vr} = \theta_r \qquad \theta_{v\phi} = \theta_r \qquad \theta_{\omega} = 0 \tag{377}$$

$$\theta_{arr} = \theta_r \qquad \theta_{ar} = \theta_r \qquad \theta_{a\phi\phi} = \theta_r - \pi \tag{378}$$

$$\theta_{ac} = \theta_r \qquad \theta_{acen} = \theta_r \qquad \theta_{a\phi} = \theta_r \tag{379}$$

$$\theta_a = \theta_r \qquad \beta_{r,t} = 0 \qquad \beta_{\omega,t} = -\pi \tag{380}$$

Therefore the case of a time independent θ_r combined with $\theta_t = 0$ and $\theta_\phi = 0$ gives the standard kinematic and dynamic equations (373) through (376). Thus the effects of a time dependent θ_r with $\theta_t \neq 0$ and $\theta_\phi \neq 0$ can be discerned from anomalies in the rotational motion of stars, molecules, atoms, and atomic nuclei. However, the effects of a time independent θ_r with $\theta_t = 0$ and $\theta_\phi = 0$ can be discovered in non-rotating systems through its effect on the gravity and pressure of non-rotating (or slowly rotating) stars and planets. Section 7 shows the effects of θ_r on the equilibrium configurations of stars and planets.

6. EULER EQUATIONS FOR BROKEN SYMMETRY MATTER. This section considers Euler's equations of motion for a broken symmetry fluid, and is a prelude to the study of stellar and planetary equilibrium which is considered in Section 7. The standard special relativistic Euler equations for the radial and transverse directions are written as[12,13]

$$(\rho + P^a/c^a)\gamma_a^2 a_r^a = -\left(\frac{\partial P^a}{\partial r^a} + v_r^a \frac{\partial P^a}{\partial t^a}\right) - \frac{\partial W^a}{\partial r^a} \tag{381}$$

$$(\rho + P^a/c^2)\gamma_a^2 a_\phi^a = -\left(\frac{1}{r^a}\frac{\partial P^a}{\partial \phi^a} + v_\phi^a \frac{\partial P^a}{\partial t^a}\right) - \frac{1}{r^a}\frac{\partial W^a}{\partial \phi^a} \tag{382}$$

where a_r^a and a_ϕ^a are the conventional radial and transverse components of acceleration, $\rho \triangleq$ proper mass density, P^a = pressure, W^a = macroscopic external force potential, and where

$$\gamma_a = (1 - \beta_a^2)^{-1/2} \qquad \beta_a = v_a/c \qquad v_a^2 = v_{ra}^2 + v_{\phi a}^2 \tag{383}$$

In section 7. W will be taken to be the gravitational potential.

It has been shown that in bulk matter the pressure has an internal phase angle as represented by equation (10), and that the coordinates within bulk matter also have internal phases such as, for example, is represented by equation (49) for the radial coordinate. Therefore the generalization of the special relativistic Euler equations to the case of bulk matter with broken internal symmetries is written as

$$\left(\rho + \frac{\bar{P}}{c^2}\right)\bar{\gamma}^2\bar{a}_r = \gamma^2 a_r\left(\rho + \frac{P}{c^2}\, e^{j\theta_P}\right)e^{j(\theta_{ar}+2\theta_\gamma)} \tag{384}$$

$$= -\left(\frac{\partial\bar{P}}{\partial\bar{r}} + \bar{v}_r\,\frac{\partial\bar{P}}{\partial\bar{t}}\right) - \frac{\partial\bar{W}}{\partial\bar{r}}$$

$$\left(\rho + \frac{\bar{P}}{c^2}\right)\bar{\gamma}^2\bar{a}_\phi = \gamma^2 a_\phi\left(\rho + \frac{P}{c^2}\, e^{j\theta_P}\right)e^{j(\theta_{a\phi}+2\theta_\gamma)} \tag{385}$$

$$= -\left(\frac{1}{\bar{r}}\,\frac{\partial\bar{P}}{\partial\bar{\phi}} + \bar{v}_\phi\,\frac{\partial\bar{P}}{\partial\bar{t}}\right) - \frac{1}{\bar{r}}\,\frac{\partial\bar{W}}{\partial\bar{\phi}}$$

where $\bar{a}_r$ and $\bar{a}_\phi$ are given by equations (334) and (335) respectively, and $\bar{W}$ can be written as in equation (246). The complex boost is written as

$$\bar{\gamma} = (1 - \bar{\beta}^2)^{-1/2} = \gamma e^{j\theta_\gamma} \tag{386}$$

where

$$\bar{\beta} = \bar{v}/c \qquad \bar{v} = ve^{j\theta_v} \qquad \bar{v}^2 = \bar{v}_r^2 + \bar{v}_\phi^2 \tag{387}$$

and where the boost magnitude and internal phase angle are given as

$$\gamma = [1 - 2\beta^2 \cos(2\theta_v) + \beta^4]^{-1/4} \tag{388}$$

$$\tan(2\theta_\gamma) = \frac{\beta^2 \sin(2\theta_v)}{1 - \beta^2\cos(2\theta_v)} \tag{388A}$$

The generalization of the relativistic Euler equations for bulk matter with broken internal symmetries can also be written for the $\bar{x}$, $\bar{y}$, and $\bar{z}$ coordinates as[12]

$$(\rho + \bar{P}/c^2)\bar{\gamma}^2\bar{a}_x = -\left(\frac{\partial\bar{P}}{\partial\bar{x}} + \bar{v}_x\,\frac{\partial\bar{P}}{\partial\bar{t}}\right) - \frac{\partial\bar{W}}{\partial\bar{x}} \tag{389}$$

$$(\rho + \bar{P}/c^2)\bar{\gamma}^2\bar{a}_y = -\left(\frac{\partial\bar{P}}{\partial\bar{y}} + \bar{v}_y\,\frac{\partial\bar{P}}{\partial\bar{t}}\right) - \frac{\partial\bar{W}}{\partial\bar{y}} \tag{390}$$

$$(\rho + \bar{P}/c^2)\bar{\gamma}^2\bar{a}_z = -\left(\frac{\partial\bar{P}}{\partial\bar{z}} + \bar{v}_z\,\frac{\partial\bar{P}}{\partial\bar{t}}\right) - \frac{\partial\bar{W}}{\partial\bar{z}} \tag{391}$$

where $\bar{a}_x$, $\bar{a}_y$, and $\bar{a}_z$ are given by equations (190) through (192), $\bar{\gamma}$ is given in terms of $\bar{v}/c$ by equation (386), and where

$$\bar{v}^2 = \bar{v}_x^2 + \bar{v}_y^2 + \bar{v}_x^2 \tag{391A}$$

Equations (384) and (385) or equations (389) through (391) are simple generalizations of the standard special relativistic Euler equations to the case of bulk matter with broken internal symmetry.

Euler's equations will be used to relate the internal phase angles of the coordinates to the internal phase angle of the pressure. From the radial acceleration equation (384) it follows for $\partial\bar{P}/\partial\bar{t} = 0$ that

$$\gamma^2 a_r\left(\rho + \frac{P}{c^2} e^{j\theta_P}\right)e^{j(\theta_{ar}+2\theta_\gamma)} = D_P e^{j(\Phi_P+\pi)} + D_W e^{j(\Phi_W+\pi)} \tag{392}$$

where

$$\frac{\partial\bar{P}}{\partial\bar{r}} = D_P e^{j\Phi_P} \qquad \frac{\partial\bar{W}}{\partial\bar{r}} = D_W e^{j\Phi_W} \tag{393}$$

with

$$D_P = \sqrt{\frac{\left(\frac{\partial P}{\partial r}\right)^2 + P^2\left(\frac{\partial\theta_P}{\partial r}\right)^2}{1 + r^2\left(\frac{\partial\theta_r}{\partial r}\right)^2}} \tag{394}$$

$$D_W = \sqrt{\frac{\left(\frac{\partial W}{\partial r}\right)^2 + W^2\left(\frac{\partial\theta_W}{\partial r}\right)^2}{1 + r^2\left(\frac{\partial\theta_r}{\partial r}\right)^2}} \tag{395}$$

$$\Phi_P = \theta_P + \beta_{P,r} - \theta_r - \beta_{r,r} \tag{396}$$

$$\Phi_W = \theta_W + \beta_{W,r} - \theta_r - \beta_{r,r} \tag{397}$$

and where

$$\tan \beta_{P,r} = P \frac{\partial \theta_P / \partial r}{\partial P / \partial r} \tag{398}$$

$$\tan \beta_{W,r} = W \frac{\partial \theta_W / \partial r}{\partial W / \partial r} \tag{399}$$

and where $\beta_{r,r}$ is given by equation (64). For the case of an external potential it follows from the radial equation of motion (392) that

$$\gamma^2 a_r [\rho \cos (\theta_{ar} + 2\theta_\gamma) + \frac{P}{c^2} \cos (\theta_{ar} + 2\theta_\gamma + \theta_P)] \tag{400}$$

$$= D_P \cos (\Phi_P + \pi) + D_W \cos (\Phi_W + \pi)$$

$$= - D_P \cos \Phi_P - D_W \cos \Phi_W$$

$$\gamma^2 a_r [\rho \sin (\theta_{ar} + 2\theta_\gamma) + \frac{P}{c^2} \sin (\theta_{ar} + 2\theta_\gamma + \theta_P)] \tag{401}$$

$$= D_P \sin (\Phi_P + \pi) + D_W \sin (\Phi_W + \pi)$$

$$= - D_P \sin \Phi_P - D_W \sin \Phi_W$$

From equation (400) and (401) it follows that

$$\gamma^4 a_r^2 (\rho^2 + \frac{P^2}{c^4} + 2\rho \frac{P}{c^2} \cos \theta_P) = D_P^2 + D_W^2 + 2 D_P D_W \cos (\Phi_W - \Phi_P) \tag{402}$$

Equations (400) and (401) determine a_r and θ_{ar} . Note that a_r and θ_{ar} are related to the component acceleration terms through equations (338) and (339). Expressions similar to equations (400) and (401) can be derived for the transverse acceleration from equation (385).

Consider now the case of static equilibrium. In this case the acceleration terms in equations (400) and (401) are equal to zero, with result

$$D_P = D_W \tag{403}$$

$$\tan (\Phi_P + \pi) = \tan (\Phi_W + \pi) \qquad \text{or} \qquad \tan \Phi_P = \tan \Phi_W \tag{404}$$

$$\Phi_P = \Phi_W + \pi \tag{405}$$

Because $D_P > 0$ and $D_W > 0$, the only way equations (400) and (401) can have their left hand sides equal to zero is to have $D_P = D_W$ and

$$\cos \Phi_P = - \cos \Phi_W \tag{406}$$

$$\sin \Phi_P = - \sin \Phi_W \tag{407}$$

which requires equation (405) to be valid while at the same time satisfying equation (404). Combining equations (396), (397) and (405) gives

$$\theta_P + \beta_{P,r} = \theta_W + \beta_{W,r} + \pi \tag{408}$$

Equations (403) and (408) are the equations for static equilibrium for the Euler equations describing bulk matter with broken internal symmetries under the action of an external potential (which also has a broken symmetry). The phase angle θ_P is determined by the relativistic state equation as shown in Reference 6 for solids and quantum liquids, and in an accompanying paper for the real gases. Therefore since θ_W and $\beta_{W,r}$ are related to the coordinates r and θ_r, it is equations (403) and (408) that relate the phase angle θ_r of the radial coordinate to P and θ_P of the equation of state. This will be made explicitly clear in Section 7 where gravitational equilibrium in stars and planets is considered.

Strictly speaking, only for a bulk matter system in which an external potential acts can one define a variation of θ_r with spatial coordinates, because only in this case can a physical choice or origin of coordinates be made (such as the center of a star or planet) from which to measure the coordinate r and thereby evaluate the denominators in equations (394) and (395). Only then is there a fixed reference point from which to calculate the variation of the phase angles such as θ_r, θ_v, and θ_a over macroscopic distances. However, θ_r, θ_v, and θ_a are determined by the broken symmetry of the local pressure θ_P and the broken symmetry of the local potential θ_W through equations (384) and (385).

7. EQUILIBRIUM OF STARS AND PLANETS. The equilibrium of stars and planets that are composed of matter with broken internal symmetries can be obtained from the complex number form of Euler's equation (384) or the equivalent equations (403) and (408). The gravitational potential energy that includes the effects of the broken symmetry of the space coordinates is written as

$$\bar{W} = We^{j\theta_W} = - \frac{GM\rho}{\bar{r}} = \frac{GM\rho}{r} e^{j(\pi - \theta_r)} \tag{409}$$

corresponding to a gravitation force

$$\bar{F} = Fe^{j\theta_F} = -\frac{GM\rho}{\bar{r}^2} = \frac{GM\rho}{r^2} e^{j(\pi-2\theta_r)} \tag{410}$$

so that

$$W = \frac{GM}{r} \tag{411}$$

$$\theta_W = \pi - \theta_r \tag{412}$$

where $M = M(r)$ = mass at radius r. Newtonian gravity is assumed to be valid in this paper, so that the force is dependent only on $\bar{r}$ (through $\bar{r}^{-2}$). No explicit dependence on the angular coordinates $\bar{\psi}$ or $\bar{\phi}$ is assumed. However, the radial coordinate phase angle θ_r can depend on angles, $\theta_r = \theta_r(r,\psi,\phi)$.

The first equilibrium condition that is derived from the Euler equation is given by equation (403). Substituting equations (411) and (412) into equation (395) gives

$$D_W = \frac{GM\rho}{r^2} \tag{413}$$

and therefore substituting equations (394) and (413) into equation (403) gives the first equilibrium equation for a gravitating star as

$$\sqrt{\left(\frac{\partial P}{\partial r}\right)^2 + P^2\left(\frac{\partial \theta_P}{\partial r}\right)^2} = \frac{GM\rho}{r^2}\sqrt{1 + \left(r\frac{\partial \theta_r}{\partial r}\right)^2} \tag{414}$$

Considering the fact that in a gravitating star or planet $\partial P/\partial r < 0$, equation (414) can be rewritten as

$$\frac{\partial P}{\partial r}\sqrt{1 + P^2\left(\frac{\partial \theta_P/\partial r}{\partial P/\partial r}\right)^2} = -\frac{GM\rho}{r^2}\sqrt{1 + \left(r\frac{\partial \theta_r}{\partial r}\right)^2} \tag{415}$$

which reduces to the standard stellar equilibrium equation for $\theta_P = 0$ and $\theta_r = 0$, namely[14]

$$\frac{\partial P}{\partial r} = -\frac{GM\rho}{r^2} \tag{416}$$

where the mass is related to the density and radial coordinates by

$$\cos \theta_{r,r}\frac{\partial M}{\partial r} = 4\pi r^2\rho \tag{417}$$

Note that equation (415) can also be rewritten as

$$\frac{\cos \beta_{r,r}}{\cos \beta_{P,r}} \frac{\partial P}{\partial r} = - \frac{GM\rho}{r^2} \tag{418}$$

If the terms involving the internal phase angles in equation (415) are assumed to be small it follows from this equation by expanding the radicals and solving a quadratic equation for $\partial P/\partial r$ that to a first approximation

$$\frac{\partial P}{\partial r} = - \frac{GM\rho}{r^2} \Psi \tag{419}$$

where

$$\Psi = 1 + \frac{1}{2} \left(r \frac{\partial \theta_r}{\partial r} \right)^2 - \frac{1}{2} \left(\frac{rP}{GM\rho} \right)^2 \left(r \frac{\partial \theta_P}{\partial r} \right)^2 \tag{420}$$

Therefore to first order the pressure gradient in equation (419) for stellar and planetary interiors with broken internal symmetry differs from the conventional result given in equation (416) by two opposing terms that are related to θ_r and θ_P respectively. Solving for the mass M from equation (414) and placing the expression in equation (417) gives the following combined equilibrium equation

$$\cos \beta_{r,r} \frac{1}{r^2} \frac{\partial}{\partial r} \left[\frac{r^2}{\rho} \sqrt{\frac{\left(\frac{\partial P}{\partial r} \right)^2 + P^2 \left(\frac{\partial \theta_P}{\partial r} \right)^2}{1 + r^2 \left(\frac{\partial \theta_r}{\partial r} \right)^2}} \right] = 4\pi G\rho \tag{421}$$

or equivalently as

$$\cos \beta_{r,r} \frac{1}{r^2} \frac{\partial}{\partial r} \left[\frac{r^2}{\rho} \frac{\partial P}{\partial r} \cos \beta_{r,r} \sqrt{1 + P^2 \left(\frac{\partial \theta_P / \partial r}{\partial P / \partial r} \right)^2} \right] = - 4\pi G\rho \tag{422}$$

Similarly, using equation (419) for this purpose gives

$$\cos \beta_{r,r} \frac{1}{r^2} \frac{\partial}{\partial r} \left(\frac{r^2}{\rho \Psi} \frac{\partial P}{\partial r} \right) = - 4\pi G\rho \tag{423}$$

where Ψ is given by equation (420).

The second gravitational equilibrium equation can be obtained by noting that equations (399), (412) and (64) yield

$$\beta_{W,r} = \beta_{r,r} - \pi \tag{424}$$

where $\beta_{r,r}$ is given by equation (64), so that it follows from equations (408), (412), and (424) that the second gravitational equilibrium equation is

$$\theta_P + \beta_{P,r} = \beta_{r,r} - \theta_r + \pi \tag{425}$$

where $\beta_{P,r}$ is given by equation (398). Equation (425) can be used to solve θ_r in terms of θ_P because this equation can be written as

$$\theta_P + \tan^{-1}\left(P \frac{\partial\theta_P/\partial r}{\partial P/\partial r}\right) = \tan^{-1}\left(r \frac{\partial\theta_r}{\partial r}\right) - \theta_r + \pi \tag{426}$$

Equation (427) can be simplified by writing

$$\beta_{P,r} = \pi + \beta'_{P,r} \tag{427}$$

where $\beta'_{P,r}$ is a small quantity which can be positive or negative. Combining equations (425) and (427) gives the second gravitational equilibrium condition as

$$\theta_P + \beta'_{P,r} = \beta_{r,r} - \theta_r \tag{428}$$

From equation (398) it follows that the case of $\theta_P > 0$ and $\partial\theta_P/\partial r < 0$ (corresponding to planets and degenerate stars such as neutron stars and white dwarfs) gives $\beta_{P,r} > \pi$ or $\beta'_{P,r} > 0$, and from equations (428) and (64) it follows that $\theta_r < 0$ and $\beta_{r,r} > 0$. For gaseous stars it may be possible to have $\theta_P > 0$ or $\theta_P < 0$ because of a degeneracy in the state equation of the relativistic real gas (see accompanying paper on real gases). For gaseous stars with $\theta_P < 0$ and $\partial\theta_P/\partial r > 0$ it follows from equation (398) that $\beta_{P,r} < \pi$ or $\beta'_{P,r} < 0$ and therefore from equation (428) it follows that $\theta_r > 0$ and $\beta_{r,r} < 0$. This analysis assumes that $\partial P/\partial r < 0$ for all stars and planets. Combining equations (396) and (397) with equations (412), (424), and (425) gives

$$\phi_P = \pi - 2\theta_r \tag{429}$$

$$\phi_W = -2\theta_r \tag{430}$$

Equation (429) follows from the fact that

$$\frac{\partial\bar{P}}{\partial\bar{r}} = -\frac{GM_0}{\bar{r}^2} \tag{431}$$

Equation (428) is the second equilibrium equation derived from the general Euler equilibrium equation (408).

Equation (422), or the approximation equation (423), along with the equilibrium equation for the internal phases given in equation (428) are the two equilibrium equations for a gravitationally bound star or planet. These equations involve P, ρ, θ_P, and θ_r, so that clearly two additional equations are required for a complete solution of the equilibrium configuration (actually an energy generation equation is also required). The two additional equations that are required are the state equations which specify

$$P = P(\rho,T) \tag{432}$$

$$\theta_P = \theta_P(\rho,T) \tag{433}$$

the magnitude and internal phase angle of the complex number pressure. Equations (432) and (433) can be used to develop the following relationships

$$\frac{\partial P}{\partial r} = \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial r} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial r} \tag{434A}$$

$$\frac{\partial P}{\partial \psi} = \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial \psi} \tag{434B}$$

$$\frac{\partial P}{\partial \phi} = \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial \phi} \tag{434C}$$

$$\frac{\partial \theta_P}{\partial r} = \frac{\partial \theta_P}{\partial \rho}\frac{\partial \rho}{\partial r} + \frac{\partial \theta_P}{\partial T}\frac{\partial T}{\partial r} \tag{435A}$$

$$\frac{\partial \theta_P}{\partial \psi} = \frac{\partial \theta_P}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \frac{\partial \theta_P}{\partial T}\frac{\partial T}{\partial \psi} \tag{435B}$$

$$\frac{\partial \theta_P}{\partial \phi} = \frac{\partial \theta_P}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \frac{\partial \theta_P}{\partial T}\frac{\partial T}{\partial \phi} \tag{435C}$$

where r, ψ, and ϕ are the spherical coordinates whose origin is at the center of the star. Defining the following quantities

$$\tan \beta_{P,\rho} = P\,\frac{\partial \theta_P/\partial \rho}{\partial P/\partial \rho} \tag{436}$$

$$\tan \beta_{P,T} = P\,\frac{\partial \theta_P/\partial T}{\partial P/\partial T} \tag{437}$$

allows equations (435A) through (435C) to be written as

$$\frac{\partial \theta_P}{\partial r} = \frac{1}{P}\left(\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial r} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial r}\right) \tag{438A}$$

$$\frac{\partial \theta_P}{\partial \psi} = \frac{1}{P}\left(\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial \psi}\right) \tag{438B}$$

$$\frac{\partial \theta_P}{\partial \phi} = \frac{1}{P}\left(\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial \phi}\right) \tag{438C}$$

also

$$\tan \beta_{P,r} = \frac{\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial r} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial r}}{\frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial r} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial r}} \tag{439A}$$

$$\tan \beta_{P,\psi} = \frac{\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial \psi}}{\frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial \psi}} \tag{439B}$$

$$\tan \beta_{P,\phi} = \frac{\tan \beta_{P,\rho} \frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \tan \beta_{P,T} \frac{\partial P}{\partial T}\frac{\partial T}{\partial \phi}}{\frac{\partial P}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \frac{\partial P}{\partial T}\frac{\partial T}{\partial \phi}} \tag{439C}$$

Similarly for the internal phase angle of the radial coordinate

$$\frac{\partial \theta_r}{\partial r} = \frac{\partial \theta_r}{\partial \rho}\frac{\partial \rho}{\partial r} + \frac{\partial \theta_r}{\partial T}\frac{\partial T}{\partial r} \tag{440A}$$

$$\frac{\partial \theta_r}{\partial \psi} = \frac{\partial \theta_r}{\partial \rho}\frac{\partial \rho}{\partial \psi} + \frac{\partial \theta_r}{\partial T}\frac{\partial T}{\partial \psi} \tag{440B}$$

$$\frac{\partial \theta_r}{\partial \phi} = \frac{\partial \theta_r}{\partial \rho}\frac{\partial \rho}{\partial \phi} + \frac{\partial \theta_r}{\partial T}\frac{\partial T}{\partial \phi} \tag{440C}$$

which can be used to evaluate equations (64) and (67). Equations similar to equations (440) hold for θ_ψ and θ_ϕ, but these internal phase angles are taken to be zero in the simplest theory of gravitational equilibrium. In any case,

it is clear that the determination of the equilibrium configuration of stars and planets require the determination of $\theta_r(r)$ and $\theta_p(r)$ as part of the solution. Both of these phase angles must approach their vacuum values, $\theta_r^{(v)}$ and $\theta_p^{(v)}$, at the surface of the star or planet.

The magnitude $P(\rho,T)$ and internal phase angle $\theta_p(\rho,T)$ of the relativistic pressure are obtained from a solution of the relativistic trace equation (1) along with the magnitudes and internal phase angles of the other thermodynamic functions.[6] A T = 0 degenerate neutron gas state equation with a pressure described by $P^o(\rho)$ and $\theta_p^o(\rho)$, which is obtained from the solution of the T = 0 form of the relativistic trace equation (1), can serve as an adequate description of a neutron star.[6] The radial variation of the internal phase angle of the radial coordinates of a neutron star can be determined from $P^o(\rho)$ and $\theta_p^o(\rho)$ using equations (422) and (428). For the interacting classical or quantum gases that occur in ordinary stars, the internal phase angle $\theta_p(\rho,T)$ can be evaluated from the relativistic third and higher virial coefficients of a real classical or quantum gas at high temperatures. The relativistic third and higher virial coefficients are obtained from a solution of the relativistic trace equation (1) for the real gases. Therefore the relativistic third and higher virial coefficients of the state equation of real gases will play an important role in the determination of the equilibrium conditions of ordinary gaseous stars.

The equilibrium of gravitating planets is treated in a slightly different manner than for stars, but the two basic equilibrium equations (422) and (428) are also valid for gravitating planets. Equation (422) will be written in a slightly different form for planets. As in the case for stars, the complex number equilibrium equation is written as

$$\frac{\partial \bar{P}}{\partial \bar{r}} = - \frac{G\rho M}{\bar{r}^2} \tag{441}$$

or

$$D_p e^{j\Phi_p} = \frac{G\rho M}{r^2} e^{j(\pi - 2\theta_r)} \tag{442}$$

where D_p is given by equation (394) and Φ_p is given by equation (396). Equation (441) can be rewritten in terms of a density derivative by introducing the bulk modulus at constant entropy $\bar{K}_S$. In order to determine $\bar{K}_S$, the bulk modulus at constant temperature $\bar{K}_T$ must first be introduced. The constant temperature bulk modulus is given by[6]

$$\bar{K}_T = \rho \frac{\partial \bar{P}}{\partial \rho} = K_T e^{j(\theta_p + \beta_{P,\rho})} \tag{443}$$

where

$$K_T = \sqrt{\left(\rho \frac{\partial P}{\partial \rho}\right)^2 + P^2\left(\rho \frac{\partial \theta_P}{\partial \rho}\right)^2} \tag{444}$$

and where $\beta_{P,\rho}$ is given by equation (436). The bulk modulus at constant entropy is easily found to be given by

$$\bar{K}_S = \bar{K}_T + \bar{\gamma}\left(T \frac{\partial \bar{P}}{\partial T}\right) = K_S e^{j\theta KS} \tag{445}$$

where the complex number Grüneisen function $\bar{\gamma}$ is given in equation (5). Equation (445) can be written in component form as

$$K_S \cos \theta_{KS} = K_T \cos (\theta_P + \beta_{P,\rho}) + \gamma N \cos (\theta_\gamma + \theta_P + \beta_{P,T}) \tag{446}$$

$$K_S \sin \theta_{KS} = K_T \sin (\theta_P + \beta_{P,\rho}) + \gamma N \sin (\theta_\gamma + \theta_P + \beta_{P,T}) \tag{447}$$

where expressions for the magnitude γ and internal phase θ_γ of the Grüneisen function are given in Reference 6, $\beta_{P,\rho}$ and $\beta_{P,T}$ are given by equations (436) and (437) respectively, and where[6]

$$N = \sqrt{\left(T \frac{\partial P}{\partial T}\right)^2 + P^2\left(T \frac{\partial \theta_P}{\partial T}\right)^2} \tag{448}$$

Equations (446) and (447) give immediately

$$\tan \theta_{KS} = \frac{K_T \sin (\theta_P + \beta_{P,\rho}) + \gamma N \sin (\theta_\gamma + \theta_P + \beta_{P,T})}{K_T \cos (\theta_P + \beta_{P,\rho}) + \gamma N \cos (\theta_\gamma + \theta_P + \beta_{P,T})} \tag{449}$$

$$K_S^2 = K_T^2 + \gamma^2 N^2 + 2\gamma N K_T \cos (\theta_\gamma + \beta_{P,T} - \beta_{P,\rho}) \tag{450}$$

which allow the calculation of the phase angle and magnitude of the bulk modulus at constant entropy.

Combining equations (434A), (441) and (445) gives the following approximation for a planet with broken symmetry matter[17]

$$\frac{\partial \rho}{\partial \bar{r}} = -\frac{G\rho^2 M}{\bar{K}_S \bar{r}^2} = -\frac{G\rho M}{\bar{v}_S^2 \bar{r}^2} \tag{451}$$

where the adiabatic velocity of elastic waves in a material with broken internal symmetry is given as

$$\bar{v}_S^2 = v_S^2 e^{2j\theta_{vS}} = \frac{\bar{K}_S}{\rho} \tag{452}$$

or

$$v_S^2 = \frac{K_S}{\rho} \tag{453}$$

$$2\theta_{vS} = \theta_{KS} \tag{454}$$

Equation (451) can also be rewritten as

$$\cos \beta_{r,r} \frac{\partial \rho}{\partial r} = - \frac{G\rho M}{r^2 v_S^2} \tag{455}$$

$$2\theta_{vS} = \beta_{r,r} - \theta_r = \tan^{-1}\left(r \frac{\partial \theta_r}{\partial r} \right) - \theta_r \tag{456}$$

where

$$\cos \beta_{r,r} = [1 + (r\, \partial\theta_r/\partial r)^2]^{-1/2} \tag{457}$$

Substituting the expression for the mass in equation (455) into equation (417) gives

$$\cos \beta_{r,r} \frac{1}{r^2} \frac{\partial}{\partial r} \left(\frac{r^2}{\rho} \frac{\partial \rho}{\partial r} v_S^2 \cos \beta_{r,r} \right) = - 4\pi G\rho \tag{458}$$

where v_S is obtained from equations (450) and (453). Equation (458) is the first equilibrium equation for gravitating planets and is the analog of equation (422) for stars, while equation (456) is the second equilibrium equation for gravitating planets with broken internal symmetry and is the analog ot equation (428) for stars with broken internal symmetry. Finally it should be pointed out that for matter with broken internal symmetries the adiabatic wave velocity is given by a simple formula, analogous to the conventional formula for symmetric matter, as follows[17]

$$\bar{v}_S^2 = \bar{\alpha}^2 - \frac{4}{3} \bar{\beta}^2 \tag{459}$$

where $\bar{\alpha}$ and $\bar{\beta}$ = compression and shear wave velocities respectively for matter with broken internal symmetries. Writing

$$\bar{\alpha} = \alpha e^{j\theta_\alpha} \qquad \bar{\beta} = \beta e^{j\theta_\beta} \tag{460}$$

gives

$$v_S^2 \cos(2\theta_{vS}) = \alpha^2 \cos(2\theta_\alpha) - \frac{4}{3}\beta^2 \cos(2\theta_\beta) \tag{461}$$

$$v_S^2 \sin(2\theta_{vS}) = \alpha^2 \sin(2\theta_\alpha) - \frac{4}{3}\beta^2 \sin(2\theta_\beta) \tag{462}$$

which are equivalent to the following equations

$$\tan(2\theta_{vS}) = \frac{\alpha^2 \sin(2\theta_\alpha) - \frac{4}{3}\beta^2 \sin(2\theta_\beta)}{\alpha^2 \cos(2\theta_\alpha) - \frac{4}{3}\beta^2 \cos(2\theta_\beta)} \tag{463}$$

$$v_S^2 = \alpha^4 + \frac{16}{9}\beta^4 - \frac{8}{3}\alpha^2\beta^2 \cos[2(\theta_\alpha - \theta_\beta)] \tag{464}$$

The measured adiabatic wave velocity = $v_S \cos\theta_{vS}$, while the measured compression and shear wave velocities = $\alpha \cos\theta_\alpha$ and $\beta \cos\theta_\beta$.

A knowledge of P and θ_P as a function of density and temperature can be obtained experimentally from high pressure measurements on earth materials such as olivine and gabbro. Alternatively P and θ_P can be obtained from the solution of the relativistic trace equation (1) if the unrenormalized pressure P^a and Grüneisen function γ^a can be estimated from atomic structure.[6] The seismic wave velocity v_S and its internal phase angle θ_{vS} can then be obtained from equations (453) and (454) respectively. Finally, equations (463) and (464) can be inverted to find the relativistic values of the compression wave velocity α and the shear wave velocity β . It may be possible to reverse the arguments and measure α and β which gives v_S and θ_{vS} by equations (463) and (464) and then obtain P and θ_P from equations (449), (450), (453), and (454). Equations (456) and (458) are the equilibrium equations for a planet whose solution gives $\rho(r)$ and $\theta_r(r)$ in terms of P and θ_P . As in the case of the equilibrium calculation for stars, two auxiliary state equations of the form given in equations (432) and (433) are required. In any case, it is clear that P, θ_P, and θ_r are required for an understanding of the equilibrium configuration and seismic properties of a planet.

From the previous analysis it is clear that the Newtonian force of gravity acting on a unit mass at a distance r from the center of a spherical body of mass M(r) with broken internal symmetry is written as

$$\bar{F} = -\frac{GM}{\bar{r}^2} = -\frac{GM}{r^2} e^{j2\theta_r} \tag{465}$$

$$F_R = -\frac{GM}{r^2}\cos(2\theta_r) = -GM/r_m^2 \cos^2\theta_r \cos(2\theta_r) \tag{466}$$

where $r_m = r \cos \theta_r$ = measured value of the radial distance between two points, $\bar{F}$ = complex Newtonian gravity force with internal phase, and F_R = real part of the gravity force in the radial direction which is the measured gravity force. The force F_R must be compared to the force F_a = conventional Newtonian gravity force for asymmetric matter which is given by

$$F_a = -\frac{GM}{r_m^2} = -\frac{GM}{r^2 \cos^2 \theta_r} \tag{467}$$

The difference $F_R - F_a$ is given by

$$F_D = F_R - F_a \tag{468}$$

$$= \frac{GM}{r^2}\left[\cos^{-2} \theta_r - \cos(2\theta_r)\right]$$

$$= \frac{GM}{r_m^2}\left[1 - \cos^2 \theta_r \cos(2\theta_r)\right]$$

$$\sim + 3\theta_r^2 GM/r^2$$

$$\sim + 3\theta_r^2 GM/r_m^2$$

where the last two approximations are valid for small θ_r , and where $\theta_r = \theta_r(r,\psi,\phi)$ is a function of the spherical polar coordinates of the unit mass. Therefore the effect of broken symmetry matter on Newtonian gravity is to imply that there is a new additional repulsive gravity force F_D in operation which does not have a strictly r^{-2} dependence on radial coordinates. But in fact gravity in the planets is Newtonian in form (neglecting general relativity effects) and has a $\bar{r}^{-2}$ dependence as given in equation (465) for broken symmetry matter. The apparent deviation from Newtonian gravity is due to the internal phase angle $\theta_r(r,\psi,\phi)$ of the radial coordinate which can have a complicated coordinate dependence because of the inhomogeneous nature of the earth's core, mantle and crust. Equation (466) shows that F_R does not have an r^{-2} (or r_m^{-2}) dependence on coordinates.

The rate of change of the force of gravity with radial distance is obtained for broken symmetry matter from equation (466) to be

$$\frac{\partial F_R}{\partial r} = \frac{2GM}{r^3}\left[\cos(2\theta_r) + r\frac{\partial \theta_r}{\partial r}\sin(2\theta_r)\right] - 4\pi G\rho \cos(2\theta_r) \tag{469}$$

and for radial variations only

$$\frac{\partial F_R}{\partial r_m} = \frac{\partial F_R}{\partial r}\frac{\partial r}{\partial r_m} = \frac{\partial F_R}{\partial r} / (\cos\theta_r - \sin\theta_r\, r\frac{\partial \theta_r}{\partial r}) \tag{470}$$

The corresponding variation for the conventionally calculated Newtonian gravity force given by equation (467) is

$$\frac{\partial F_a}{\partial r} = \frac{2GM}{r^3\cos^2\theta_r}(1 - \tan\theta_r\, r\frac{\partial \theta_r}{\partial r}) - \frac{4\pi G\rho}{\cos^2\theta_r} \tag{471}$$

$$\frac{\partial F_a}{\partial r_m} = \frac{\partial F_a}{\partial r} / (\cos\theta_r - \sin\theta_r\, r\frac{\partial \theta_r}{\partial r}) \tag{472}$$

Introduce the parameter

$$D = \frac{\partial F_R/\partial r_m - \partial F_a/\partial r_m}{\partial F_a/\partial r_m} = \frac{\partial F_R/\partial r - \partial F_a/\partial r}{\partial F_a/\partial r} \tag{473}$$

then a simple calculation shows that to second order in θ_r (there are no first order terms)

$$D = -3\theta_r^2(1 - \eta) \tag{474}$$

where

$$\eta = \frac{\frac{r}{\theta_r}\frac{\partial \theta_r}{\partial r}}{1 - \frac{2\pi\rho r^3}{M}} \tag{475}$$

For planets $\theta_r < 0$ and $\partial\theta_r/\partial r > 0$, so that in general η should be small and negative.

Therefore experimental measurements of the variation of the force of gravity with height should indicate $D < 0$, while measurements of the gravity force itself should yield $F_D > 0$. The net result of the internal phase of the radial coordinate is that the measured gravity force given by equation (466) should be slightly weaker than that predicted by the conventional Newtonian force given by equation (467). A weaker than Newtonian gravity force has been experimentally observed in geophysical measurements and in new Eötvös experiments.[18-23] These results have been interpreted to be due to a new finite range repulsive force associated with gravity.[19,20] Reference 21 contains many citations to the literature in this field. However the results of the present paper show

that in fact the weaker attractive force that is observed may be due to ordinary Newtonian gravity operating in matter with broken internal symmetries as in equation (465). A complete understanding of the earth's gravity field will require a detailed knowledge of the internal phase of the radial coordinate $\theta_r(r,\psi,\phi)$ and its variation with location. The orbits of satellites and ballistic missiles will be affected by the internal phase function $\theta_r(r,\psi,\phi)$, and perturbations in these orbits that are not explained totally by shape and density variations in the earth may lead to techniques for determining local values of $\theta_r(r,\psi,\phi)$.

8. CONCLUSION. By means of a relativistic trace equation, the Minkowski metric of spacetime impresses a broken symmetry on the matter and vacuum that are embedded in spacetime. The broken internal symmetries of matter and the vacuum are manifested at the microscopic level through the internal phase angles that are associated with the coordinates and the kinematic and dynamic variables for single particles. At the macroscopic level the broken symmetries appear in the thermodynamic functions such as pressure and internal energy of interacting systems. Within bulk matter and the vacuum, space and time exhibit broken symmetries that are manifested by internal phase angles that produce the broken symmetries of the kinematic and dynamic parameters and the broken symmetry of geometrical constructs such as angles, lengths and areas. The physical rotation of matter must also be associated with the rotation of the internal phase angles of the space and time coordinates. The internal phase angles of the kinematic and dynamic parameters of bulk matter fluid elements are determined by the Euler equations for broken symmetry matter. The calculation of the equilibrium configurations of stars, planets and other gravitationally bound systems such as galaxies must include the determination of the spatial dependence of the internal phase angle of the radial coordinate along with the spatial variation of the pressure and density. This can only be done if the state equation of broken symmetry matter is known from solutions of the basic complex number relativistic trace equation. The fact that time and space are gauge rotated quantities should affect the basic calculations of astrophysics, geophysics, and the engineering disciplines.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. O'Raifeartaigh, L. O., Group Structure of Gauge Theories, Cambridge University Press, London, 1986.

2. Becher, P., Böhm, M. and Joos, H., Gauge Theories, John Wiley, New York, 1984.

3. Konopleva, N. and Popov, V., Gauge Fields, Harwood, New York, 1981.

4. Weiss, R. A., "Scale Invariant Equations for Relativistic Waves", Fourth Army Conference on Applied Mathematics and Computing, Cornell University, ARO 87-1, May 27-30, 1986, p. 307.

5. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

6. Weiss, R. A., "Thermodynamic Gauge Theory of Solids and Quantum Liquids with Internal Phase", Fifth Army Conference on Applied Mathematics and Computing, West Point, ARO 88-1, June 15-18, 1987, p. 649.

7. Huang, K., Statistical Mechanics, John Wiley, New York, 1963.

8. Yourgrau, W., van der Merwe, A., and Raw, G., Treatise on Irreversible and Statistical Thermophysics, Dover Publications, Inc., New York, 1982.

9. Zee, A., Unity of Forces in the Universe, Vols. 1 and 2, World Scientific, Singapore, 1982.

10. Sakurai, J., Invariance Principles and Elementary Particles, Princeton University Press, Princeton, 1964.

11. Goldstein, H., Classical Mechanics, Addison-Wesley, New York, 1980.

12. Pauli, W., Theory of Relativity, Pergamon, New York, 1958.

13. Weinberg, S., Gravitation and Cosmology, John Wiley, New York, 1972.

14. Chandrasekhar, S., An Introduction to the Study of Stellar Structure, Dover, New York, 1957.

15. Jeffreys, H., The Earth, Cambridge University Press, New York, 1962.

16. Gutenberg, B., Physics of the Earth's Interior, Academic Press, New York, 1959.

17. Magnitskiy, V. A., The Internal Structure and Physics of the Earth, NASA Technical Translation TT F-395, NASA, Wash., D. C., April 1967.

18. Stacey, F. D. in Science Underground, edited by Nieto, M. M., Haxton, W.C., Hoffman, C. M., Kolb, E. W., Sandberg, V. D., and Toevs, J. W., AIP Conf. Proc. No. 96, AIP, New York, 1983.

19. Fischbach, E., Sudarsky, D., Szafer, A., and Talmadge, C., "Reanalysis of the Eötvös Experiment," Phys. Rev. Lett., Vol. 56, Jan. 6, 1986, p. 3.

20. Nieto, M. M., Goldman, T. and Hughes, R. J., "Phenomenological Aspects of New Gravitational Forces, I. Rapidly Rotating Compact Objects, II. Static Planetary Potentials, III. Slowly Rotating Astronomical Bodies" (with Macrae, K. I.), Phys. Rev. D, Vol. 36, No. 12, 15 Dec. 1987, p. 3684.

21. Stacey, F. D., Tuck, G. J., Moore, G. I., Holding, S. C., Goodwin, B. D., and Zhou, R., "Geophysics and the Law of Gravity", Rev. Mod. Phys., Vol. 59, No. 1, Jan 1987, p. 157.

22. Stacey, F. D., Tuck, G. J. and Moore, G. I., Phys. Rev. D. 36, 2374, 1987.

23. Fitch, V. L., Isaila, M. V. and Palmer, M. A., "Limits on the Existence of a Material-Dependent Intermediate-Range Force", Phys. Rev. Lett., Vol. 60, May 2, 1988, p. 1801.

NONLINEAR PROBLEMS IN THE STUDY OF WATER MOVEMENT IN FROZEN SOILS

Yoshisuke Nakano

U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, NH 03755

ABSTRACT

Water in unsaturated frozen soils generally exists in three phases: vapor, unfrozen (liquid) water and ice. Recent experimental data indicate that the flux of water f in frozen soils may be written in a general form:

$$f = - \rho D_1(w,T) \frac{\partial w}{\partial x} - \rho D_2(w,T) \frac{\partial T}{\partial x}$$

where ρ is dry density, and D_1 and D_2 are the properties of a given soil that generally depend on the content of total water in three phases w and the temperature T. Since D_1 and D_2 may vanish depending on w and T, the equation of mass balance becomes a quasilinear, degenerate equation of parabolic type. Our presentation is focused on a couple of special cases of this quasilinear problem which we encountered during our search for accurate experimental methods to determine D_1 and D_2.

INTRODUCTION

Water in unsaturated frozen soils generally exists in three phases: vapor, unfrozen (liquid) water and ice. We will denote the content of water in three phases by w. Reported experimental data[1-8,12] indicate that a gradient of w and a gradient of temperature T are two major driving forces of water in unsaturated frozen soils. Hence, the unidirectional flux of water f is given as

$$f = f_1 + f_2 \tag{1}$$

$$f_1 = - \rho D_1(w,T) \frac{\partial w}{\partial x} \tag{2a}$$

$$f_2 = - \rho D_2(w,T) \frac{\partial T}{\partial x} \tag{2b}$$

where ρ is the dry density of the soil that is a given positive number, x is a coordinate, and t is time. Nonnegative functions D_1 and D_2 are the properties of a given soil that must be determined experimentally. We will describe experimental methods for measuring D_1 and D_2 and discuss mathematical problems associated with these experiments below.

FUNCTION $D_1(w,T)$

The experiment[1-8] consisted of connecting two long columns of soil that were of the same size and the same dry density under an isothermal condition. One of them was uniformly dry with a negligibly small water

content, while the other was uniformly wet. At time t = 0 we connected the two columns to make a single column from which no water escaped. While we maintained the column at a specified temperature, water was transported from the wet part to the dry part across the contact surface between the wet and the dry columns. After a specific time passed, the soil column was quickly sectioned into many equal segments. The water content of each segment was determined gravimetrically.

The experiment is described by the following initial value problem

$$\frac{\partial w}{\partial t} = \frac{\partial}{\partial x}\left[D(w,T)\,\frac{\partial w}{\partial x}\right] \qquad -\infty < x < \infty \tag{3}$$

$$\begin{aligned} w(x,0) &= w_o && -\infty < x < 0 \\ &= 0 && 0 \le x < +\infty \end{aligned} \tag{4}$$

where T is a given temperature and w_o is a given positive number. When we seek a similarity solution $u(\eta) = w(x,t)$ with $\eta = x\,t^{-1/2}$, Eqs. 3 and 4 are reduced to

$$[D_1(u,T)\,u']' + \frac{1}{2}\,\eta\,u' = 0 \qquad -\infty < \eta < +\infty \tag{5}$$

$$u(-\infty) = w_o \quad , \quad u(+\infty) = 0 \tag{6}$$

where primes denote differentiation with respect to η.

Since D_1 vanishes at $w = 0$, Eq. 3 degenerates at this point. It is known[9] that the problem of Eqs. 3 and 4 has a unique weak solution. It is also known[10,11] that the problem of Eqs 5 and 6 has a unique weak solution that is the asymptotic solution of the problem of Eqs 3 and 4. Integrating Eq 5, we obtain

$$D_1[u(\eta)] = \frac{1}{2}\left[\int_{\infty}^{\eta} u\,d\eta - u\eta\right]/u' \tag{7}$$

Using measured probes $w(x)$ in the place of u in Eq. 7 we determined the value of $D_1(w,T)$.

The measured $D_1(w,T)$ of Morin clay is presented in Figure 1 as a function of w with T being a parameter. It is known that the unfrozen water content w* in a frozen soil depends mainly on the temperature T and that w* decreases with decreasing T. Since the mobility of water in a frozen soil is mainly due to the unfrozen water, the function $D_1(w,T)$ decreases with decreasing T. A common feature found in Figure 1 is that $D_1(w,T)$ for each given T has two peaks. One of them is around a point where w = 1.0% and the other is not far from a point where w is equal to the maximum unfrozen water content w* at T. For instance, the value of w* at T = -1.0°C is about 12.7%. The content of water in the solid phase (ice) increases as w increases beyond w*. Since the presence of ice tends to decrease the mobility of unfrozen water, D_1 decreases with increasing ice content.

FUNCTION $D_2(w,t)$

We will consider a problem in which a closed soil column initially with given uniform ρ and w is subjected on one end $x = 0$ to a temperature T_w and on the other end $x = x_o$ to a temperature $T_c < T_w$ at time $t > 0$. We assume that the temperature distribution T(x) is strictly linear, namely

$$\frac{\partial T}{\partial x} = - a = (T_c - T_w)/x_o \tag{8}$$

where a is a positive number. We will describe the problem in mathematical terms as follows.

The equation of mass balance for water is given as

$$\rho \frac{\partial}{\partial t} w(x,t) = - \frac{\partial}{\partial x} f \qquad \text{for} \qquad x_o > x > 0 , \quad t > 0 \tag{9}$$

The initial condition is given as

$$w(x, 0) = w_o \tag{10}$$

where w_o is a positive number. The boundary condition is given as

$$f(0, t) = f(x_o, t) = 0 \tag{11}$$

It follows from Eq. 8 that there is a one-to-one correspondence between x and T. Substituting T by x in Eqs. 2a and 2b and using Eq. 8, we rewrite f as

$$f = - \rho[D_1(w,x) \frac{\partial w}{\partial x} - a D_2(w, x)] \tag{12}$$

We will introduce a new function $\phi(w,x)$ defined as

$$\phi(w,x) = \int_0^w D_1(w,x)dw - a \int_0^x D_2(w,x)dx . \tag{13}$$

Using ϕ, we reduce the problem of Eqs. 9, 10 and 11 to a commonly used form given as

$$\frac{\partial w}{\partial t} = \frac{\partial^2}{\partial x^2} \phi(w,x) \qquad \text{for} \qquad x_o > x > 0 , \quad t > 0 \tag{14}$$

$$w(x,0) = w_o \tag{15}$$

$$\frac{\partial}{\partial x} \phi(w,0) = \frac{\partial}{\partial x} \phi(w,x_o) = 0 \tag{16}$$

When we seek a stationary (time-independent) solution $w^+(x)$ to the problem of Eqs. 14, 15 and 16, w (x) must satisfy the following equation if it exists:

$$D_1(w^+,x) \frac{dw^+}{dx} = a D_2(w^+,x) \tag{17}$$

It follows from Eq. 17 that the stationary solution $w^+(x)$ is a nondecreasing function in a part where D_1 is positive if D_2 are nonnegative.

The solution w^+ corresponds to a stationary state of the soil column in which the net flux of water f vanishes everywhere in the column while the dry density is kept at the initial value. It should be mentioned that this stationary state may not necessarily be a state of equilibrium so that a local circulation of water may occur. When the soil column initially with uniform ρ and w is subjected to the temperature gradient given by Eq. 8, the transport of water is expected to occur in the positive direction of x because of Eq. 2b. As water moves toward the cold end, the initial uniformity of w breaks down and a driving force of water toward the warmer end starts to build up because of Eq. 2a. Sooner or later two driving forces of water, one due to a temperature gradient and the other to a gradient of water content, tend to balance each other while the profile of water content w(x,t) asymptotically approaches the stationary profile with increasing time. If we are able to measure $w^+(x)$ experimentally, then D_2 can be determined by Eq. 17 for a given soil with known D_1.

It is not certain that the anticipated event described above actually takes place. For instance, the time required for the column to reach a stationary state may turn out to be too long for the method to be practical. These problems must be examined experimentally.

(1) Experimental Results

A typical evolution of the water content profile w(x,t) with time[12,13] is presented in Figure 2 under conditions that w_o = 15%, T_w = 1.40°C, T_c = -4.95°C and a = 0.310 °C/cm. It is clear from Figure 2 that water moves in the direction of negative temperature gradient and that w tends to converge to a stationary profile as time increases. The profiles at t = 22 days and at t = 34 days differ little. This implies that the flux of water f(x) almost diminishes everywhere after t = 22 days. An interesting feature of these profiles is the appearance of a maximum.

The effect of w_o on the stationary profile is shown in Figure 3. These experiments were conducted under the same thermal conditions as those in Figure 2. Among the four profiles in Figure 3, the measured stationary profile for the case of w_o = 5% is monotonically increasing, and the maximum appears in the profiles for three other cases. We conducted many experiments similar to those presented in Figure 3 under various thermal conditions. From these experiments we found that the stationary profile $w^+(x)$ generally consists of three parts, R_1, R_2 and R_3, depending upon the value of the first derivative $\dot{w}^+_x(x)$, when w_o is greater than 10% or so. These three parts are characterized as

$$\dot{w}^+_x > 0 \quad \text{in} \quad R_1(0 < x < x_m) \tag{18a}$$

$$< 0 \quad \text{in} \quad R_2(x_m < x < x_n) \tag{18b}$$

$$= 0 \ (w = w_o) \quad \text{in} \quad R_3(x_n < x < x_o) \tag{18c}$$

where x_m is the point where w^+ attains its maximum. We did not assign the value of $\dot{w}^+_x$ at $x = x_m$ because we are not certain about the continuity of $\dot{w}^+_x$ at x_m in view of our measured profiles that often had a maximum resembling a sharp peak.

As the first step we calculated the value of D_2 from Eq. 17 by using part of w^+ in R_1 to find the properties of D_2 as a function of w and T. The calculated values of D_2 are presented at four temperatures, -1.00, -0.50, -0.25 and -0.10°C, in Figure 4 where curves are drawn to show the general trend of data points. A common feature found in Figure 4 is that $D_2(w,T)$ for each given T has one peak. D_2 increases as w increases up to some point w and then decreases as w increases beyond this point. This decrease of D_2 is caused by increasing ice content.

(2) Mathematical Problem

Let us assume that D_1 and D_2 are smooth functions of w and T given by Figures 1 and 4, respectively. Under such an assumption it is easy to explain the appearance of a maximum at an interior point x_m as described by Eqs. 18a, 18b and 18c from a physical point of view because the lack of water movement in R_3 causes the accumulation of water in some part where $x < x_n$ when water moves in the direction of negative tmeperature gradient. An important question arises whether the solution to the problem of Eqs. 14, 15 and 16 under the above assumption actually behaves like the measured profiles. We do not have the answer to this problem. However, we will show below that Eq. 14 degenerates at the point x_m where w attains its maximum if the profile characterized by Eqs. 18a, 18b and 18c is a solution to the problem of Eqs. 14, 15 and 16.

We will consider the profile w(x,t) in the earlier stage of an experiment such as Exp. 1 in Figure 2. From such a profile we find.

$$\dot{w}_x > 0, \; f > 0 \qquad \text{in} \quad R_1(0 < x < x_m) \tag{19a}$$

$$< 0, \quad > 0 \qquad \text{in} \quad R_2(x_m < x < x_n) \tag{19b}$$

$$= 0, \quad = 0 \qquad \text{in} \quad R_3(x_n \leq x < x_o) \tag{19c}$$

We will evaluate $\dot{\phi}_w$ of the profiles given by Eqs. 19a, 19b and 19c.

Differentiating Eq. 13 with respect to w once, we obtain

$$\dot{\phi}_w = D_1(w,x) - a \frac{\partial}{\partial w} \int_0^x D_2(w,x)dx \tag{20}$$

It follows from Eqs. 19a, 19b and 19c that a one-to-one correspondence generally does not exist between w and x for a given time t. However, we can find such a correspondence in each of R_1 and R_2 separately. Hence, we reduce Eq. 20 to

$$\dot{\phi}_w = D_1(w,x) - a\, D_2(w,x)/\dot{w}_x \qquad x < x_n \quad \text{and} \quad x \neq x_m \tag{21}$$

Using Eq. 12, we reduce Eq. 21 to

$$\dot{\phi}_w = -\, f/(\rho\, \dot{w}_x) \qquad x < x_n \quad \text{and} \quad x \neq x_m \tag{22}$$

Since $f > 0$ in R_1 and R_2 from Eq. 22 we obtain

$$\dot{\phi}_w < 0 \qquad \text{in} \quad R_1 \tag{23a}$$

> 0 in R_2 (23b)

It follows from Eqs. 23a and 23b that $\dot{\phi}_w$ must vanish at x_m under the assumption and that Eq. 14 degenerates at this point.

Oleinik et al.[9] showed the existence of a unique weak solution to the problem of Eqs. 14, 15 and 16 with the condition that $\dot{\phi}_x = 0$ at $w = 0$ and $\dot{\phi}_x > 0$ for $w > 0$. In their problem Eq. 14 degenerates at $w = 0$. They showed that $\dot{w}_x$ may not be continuous at a point of transition between a part $w > 0$ and a part $w = 0$ and that a point of degeneracy propagates with a finite speed. A boundary where a degeneracy occurs is often referred to as a free (or moving) boundary. In our problem Eq. 14 degenerates not only at a point where $w = 0$ but also at a point where w attains its maximum ($\dot{\phi}_w$ changes its sign). Equations such as Eq. 14 in which the coefficient of the highest derivative changes its sign have been intensively investigated lately[14]. The results of such investigation are needed to understand the mechanism of water transport in frozen soils.

REFERENCES

1. Nakano, Y., Tice, A.R., Oliphant, J.L. and Jenkins, T.F. Transport of water in frozen soil: I. Experimental determination of soil-water diffusivity under isothermal conditions. Advances in Water Resources, 1982, 5, 221.

2. Nakano, Y., Tice, A.R., Oliphant, J.L. and Jenkins, T.F. Transport of water in frozen soil: II. Effects of ice on the transport of water under isothermal conditions. Advances in Water Resources, 1983a, 6, 15.

3. Nakano, Y., Tice, A.R., Oliphant, J.L. and Jenkins, T.F. Soil-water diffusivity of unsaturated soil at subzero temperatures. Proc. 4th Permafrost Conference, 1983b, 889.

4. Oliphant, J.L., Tice, A.R. and Nakano, Y. Water migration due to a temperature gradient in frozen soil. Proc. 4th Permafrost Conference, 1983c, 951.

5. Nakano, Y., Tice, A.R. and Oliphant, J.L. Transport of water in frozen soil: III. Experiments on the effects of ice content. Advances in Water Resources, 1984a, 7, 28.

6. Nakano, Y., Tice, A.R. and Oliphant, J.L. Transport of water in frozen soil: IV. Analysis of experimental results on the effects of ice content. Advances in Water Resources, 1984b, 7, 58.

7. Nakano, Y., Tice, A.R. and Jenkins, T.F. Transport of water in frozen soil: V. Method for measuring the vapor diffusivity when ice is absent. Advances in Water Resources, 1984c, 7, 172.

8. Nakano, Y. and Tice, A.R. Transport of water in frozen soil: VI. Effects of Temperature. Advances in Water Resources, 1987, 10, 44.

9. Oleinik, O.A., Kalashnikov, A.S. and Yui-Lin, Chzhu. The Cauchy and boundary problems for equations of the type of non-stationary infiltration. Izv. Akad. Nauk. USSR Ser. Mat., 2, 1958, 667.

10. Atkinson, F.V. and Peletier, L.A. A similarity solutions of the nonlinear diffusion equation. Arch. Rational Mech. Analysis, 54, 1974, 373.

11. Van Duyn, C.J. and Peletier, L.A. A class of similarity solutions of the nonlinear diffusion equation. Nonlinear Analysis, Theory, Methods and Applications, 1, 1977, 223.

12. Nakano, Y. and Tice, A.R. A method for measuring the rate of water transport due to temperature gradients in unsaturated frozen soils. To appear in Proc. 5th Permafrost Conference, 1988.

13. Nakano, Y. and Tice, A.R. Transport of water due to a temperature gradient in unsaturated frozen clay. To appear in Advances in Water Resources, 1988.

14. Lavent'ev Jr. M.M. Solvability of boundary-value problems for certain parabolic equations with degeneracies, Sibirskii Mat. Zh. 28, 1987, 79.

FIGURES

1. Function $D_1(w,T)$ vs. the water content w%.

2. Typical evolution of $w(x,t)$ with time.

3. Effect of w_o on the stationary profile $w^+(x)$.

4. Function $D_2(w,T)$ vs. the water content w%.

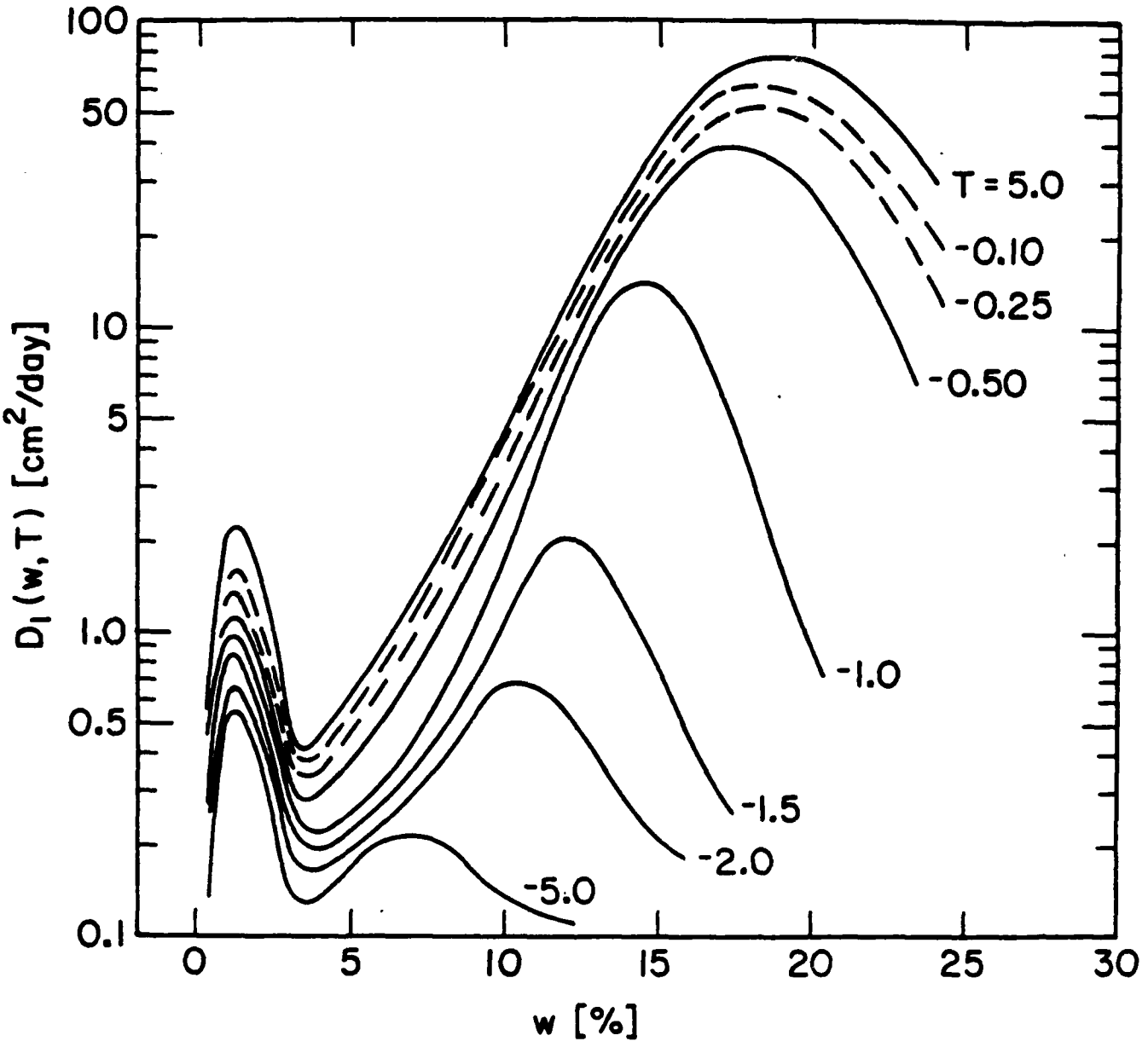


FIGURE 1

Function $D_1(w,T)$ vs. the water content w%.

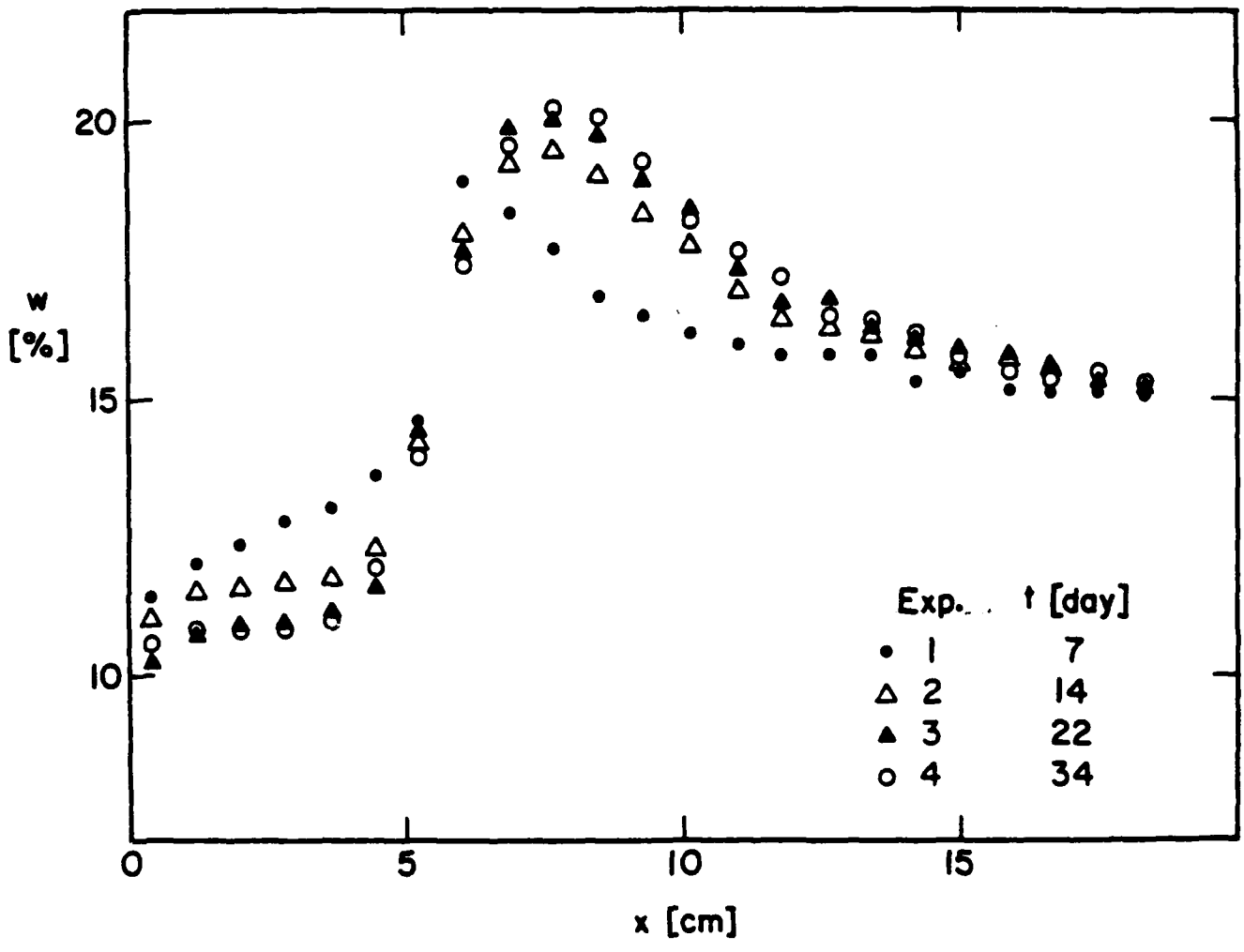


FIGURE 2

Typical evolution of w(x,t) with time.

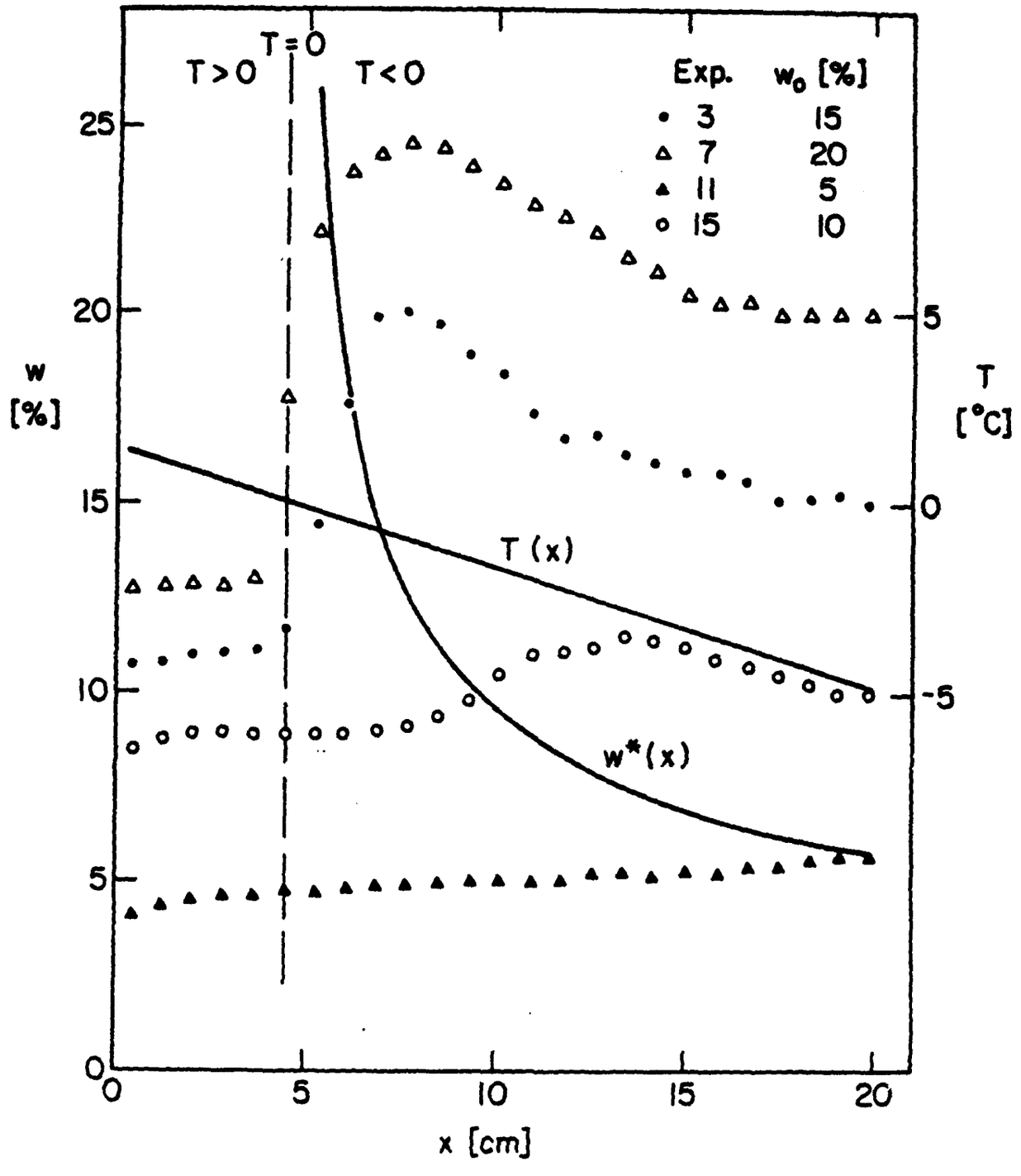


FIGURE 3

Effect of w_o on the stationary profile $w^T(x)$.

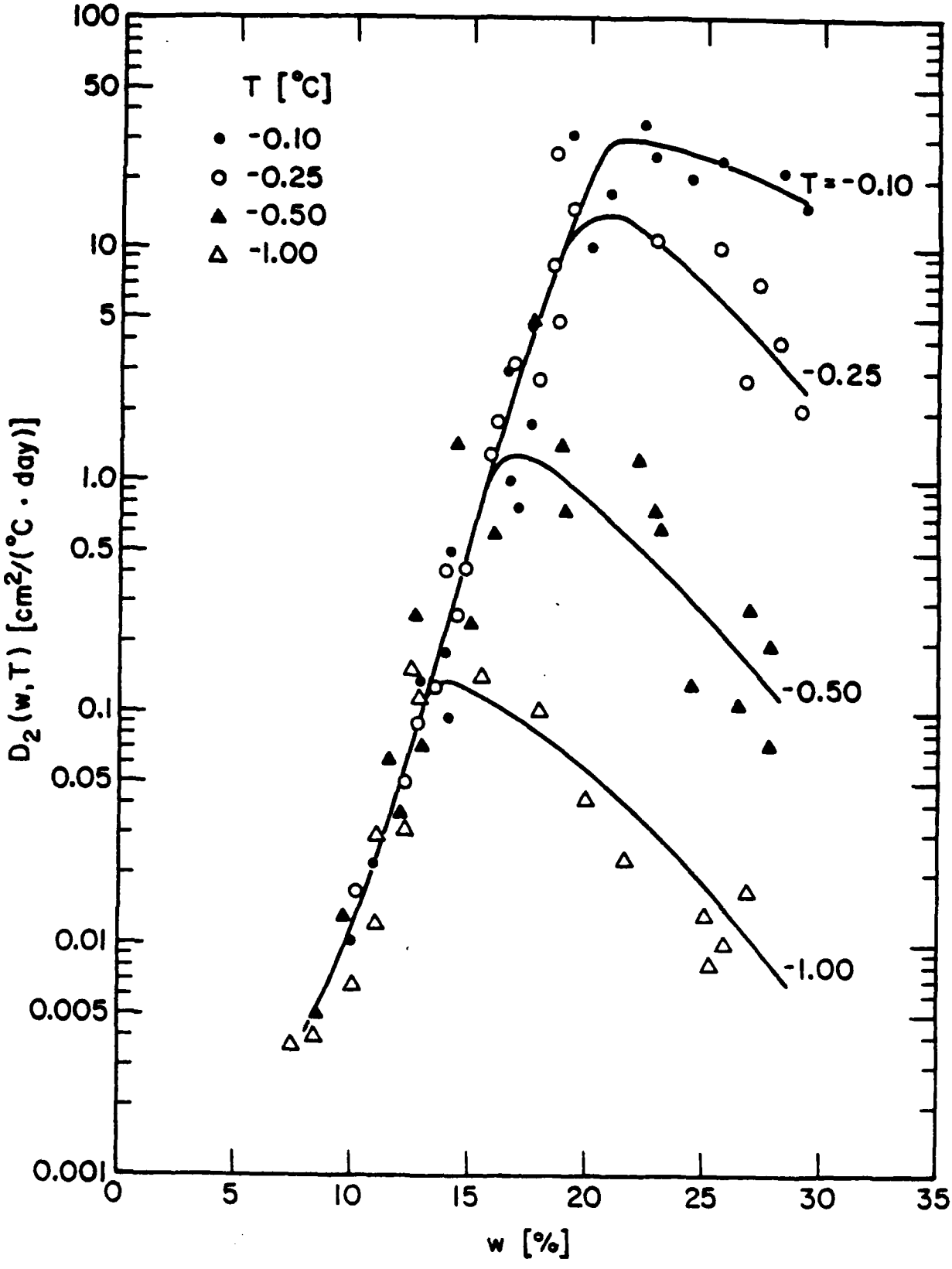


FIGURE 4

Function $D_2(w,T)$ vs. the water content w%.

INCOMPRESSIBLE RUBBER ELASTICITY FINITE ELEMENT ANALYSIS USING AN ELIMINATION METHOD

A. R. Johnson and C. J. Quigley

Mechanics and Structures Branch
U. S. Army Materials Technology Laboratory
Watertown, MA 02172-0001

SUMMARY

Finite element analysis of rubberlike materials requires the enforcement of an incompressibility condition. Penalty, Lagrange multiplier and mixed methods are typically used to enforce the constraint of incompressibility. These methods can lead to poorly conditioned tangent matricies or add a large number of variables increasing the size of the tangent matrix. In this effort the use of the implicit variable elimination method is investigated for enforcing incompressibility in rubber elasticity finite element analysis. No penalty parameters or Lagrange multipliers are used but it is difficult to generalize the method for two- and three-dimensional analyses. The one-dimensional inflation of a thick rubber cylinder is formulated and solved to demonstrate the method.

INTRODUCTION

The formulation of algorithms for the finite element analysis of large deformations of incompressible materials has involved many efforts since the mid 1960's. There are several methods in use at the present time. They include penalty, Lagrange multiplier and mixed methods solved with either updated or total Lagrangian algorithms. The basic problem is; compute the minimum of some energy functional such that a system of constraint equations is simultaneously satisfied. This is a constrained optimization problem and can be solved using methods from nonlinear constrained optimization theory. Additional techniques include successive quadratic programming and implicit variable elimination methods. Below we briefly mention references to the current methods in use and then describe a one-dimensional implicit variable elimination algorithm for the thick rubber cylinder. The review is intentionally brief.

BACKGROUND

The hydrostatic pressure was modeled with a Lagrange multiplier variable and used to attach the incompressibility constraint to the potential energy for rubber by Levinson [1]. He was then able to generate, solve and investigate the stability of solutions to the equilibrium equations for the internally pressurized Neo-Hookean sphere. Finite element methods using the Lagrange multiplier method were formulated and reviewed by Oden [2]. In Oden's formulations nodal displacements and pressure variables are related through the nonlinear equations which represent stationary points of the energy functional. He then solves illustrative examples including large deformations of rubber membranes and solids of revolution.

Tielking and Feng [3] considered problems for which the incompressibility constraint can be satisfied directly. That is, problems for which computation of the force of constraint (hydrostatic pressure) is not an issue. Instead of using displacements as variables they demonstrated the advantage of using configurations as variables. The Ritz method was then applied globally to obtain solutions to membrane problems. The configuration variable approach could then be used to construct a finite element algorithm.

While studying plasticity problems Nagtegaal, Parks and Rice [4] made an important contribution to finite element theory when they recognized the problem of dependent or redundant constraint equations. Too many or too few constraint equations in the finite element model cause numerical difficulties; either poor convergence rates or locking. Efforts were then concentrated on how many pressure variables were best for a given element formulation.

An extensive review of work done by Argyris, et al [5] included a special "fluid filled" finite element. These elements develop an internal energy when their volume (area) changes. The energy is minimized during the solution process making it a penalty like formulation. These penalty and mixed methods were under review at the same time by Hughes and Malkus[6,7]. Equivalence of the penalty method in the limit to the Lagrange multiplier method was proven. Also, it was noticed that quadratic convergence of the Newton - Raphson method is lost when large penalty

parameters are used (near incompressibility). The use of configuration variables mentioned above and a penalty enforcement of incompressiblity was investigated by Fried, Johnson, and Quigley [8,9,10]. These formulations allow for efficient computation of gradient and tangent matricies but are still subject to poor convergence when large penalty parameters are used.

A completely different approach to enforcing incompressibility was investigated by Needleman and Shih [11]. They used an implicit variable elimination method to enforce the divergence equation (incompressibility constraint) for small strain plasticity problems. The number of displacement variables are reduced by this method and the hydrostatic pressures are determined after the displacement solution is obtained. Because of the element to element interdependence of the incompressibility constraint superelements must be constructed during the variable elimination process.

The enforcement of contact constraints for large deformantion minimization problems involving configuration variables has been investigated by Johnson and Quigley [12,13,14]. Penalty, successive quadratic programming and implicit variable elimination methods have been used successfully.

In this effort we investigate implicit variable elimination for large strain rubberlike deformations. A formulation and results are presented for the one - dimensional axisymmetric deformations of an internally pressurized thick rubber cylinder.

INFINITE CYLINDER MODEL

In this section we construct unconstrained gradient and tangent matricies for the one - dimensional expansion of a thick rubber cylinder, see Figure 1. We let (α,r) represent the (undeformed,deformed) configurations. Then, the principle stretch ratios become:

$$\lambda_1 = 1.0$$

$$\lambda_2 = \frac{dr}{d\alpha} \qquad r = r(\alpha) \tag{1}$$

$$\lambda_3 = \frac{r}{\alpha}$$

Using linear two - node elements we have

$$\begin{aligned} \alpha(\varepsilon) &= (1 - \varepsilon)\alpha_1 + \varepsilon\alpha_2 = \phi^T\boldsymbol{\alpha} \quad ; \quad \boldsymbol{\alpha}^T = (\alpha_1, \alpha_2) \\ r(\varepsilon) &= (1 - \varepsilon)r_1 + \varepsilon r_2 = \phi^T\mathbf{r} \quad ; \quad \mathbf{r}^T = (r_1, r_2) \end{aligned} \tag{2}$$

where vectors are displayed in boldface. Then, we write

$$\lambda_2 = \frac{\phi_\varepsilon{}^T\mathbf{r}}{\phi_\varepsilon{}^T\boldsymbol{\alpha}}$$

$$\lambda_3 = \frac{\phi^T\mathbf{r}}{\phi^T\boldsymbol{\alpha}} \tag{3}$$

$$\lambda_2\lambda_3 = \frac{1}{2}\left(\frac{\mathbf{r}^T(\phi_\varepsilon\phi^T + \phi\phi_\varepsilon{}^T)\mathbf{r}}{\boldsymbol{\alpha}^T\phi\phi_\varepsilon{}^T\boldsymbol{\alpha}}\right)$$

We select the unconstrained [10] energy density function

$$w = C_1(I_1 - 3I_3{}^{1/3}) + C_2(I_2 - 3I_3{}^{2/3}) \tag{4}$$

where $I_1 = 1 + \lambda_2{}^2 + \lambda_3{}^2$

$$I_2 = \lambda_2{}^2 + \lambda_3{}^2 + \lambda_2{}^2\lambda_3{}^2$$

$$I_3 = \lambda_2{}^2\lambda_3{}^2$$

Assuming an element height h = 1 (Figure 1) we have the internal energy for an element on (α_1, α_2) as

$$U = 2\pi \int_{\alpha_1}^{\alpha_2} [C_1(I_1 - 3I_3{}^{1/3}) + C_2(I_2 - 3I_3{}^{2/3})]\ \alpha\ d\alpha \tag{5}$$

The element gradient and tangent matricies then become

$$\mathbf{g} = 2\pi \int_{\alpha_1}^{\alpha_2} (f_2\lambda_{2\mathbf{r}} + f_3\lambda_{3\mathbf{r}})\ \alpha\ d\alpha \quad ; \quad \lambda_{i\mathbf{r}} = \frac{d\lambda_i}{d\lambda_{\mathbf{r}}} \tag{6}$$

$$\mathbf{k} = 2\pi \int_{\alpha_1}^{\alpha_2} (g_3(\lambda_{3r}\lambda_{2r}^T + \lambda_{2r}\lambda_{3r}^T) + g_2\lambda_{2r}\lambda_{2r}^T + g_{3T}\lambda_{3r}\lambda_{3r}^T)\,\alpha\,d\alpha$$

where

$$f_2 = 2(C_1 + C_2)\lambda_2 + 2C_2\lambda_2\lambda_3^2 - 2C_1\lambda_3(\lambda_2\lambda_3)^{-1/3} - 4C_2\lambda_3(\lambda_2\lambda_3)^{1/3}$$

$$f_3 = 2(C_1 + C_2)\lambda_3 + 2C_2\lambda_2^2\lambda_3 - 2C_1\lambda_2(\lambda_2\lambda_3)^{-1/3} - 4C_2\lambda_2(\lambda_2\lambda_3)^{1/3}$$

$$g_2 = 2(C_1 + C_2) + 2C_2\lambda_3^2 + 2/3\ C_1\lambda_2^{-4/3}\lambda_3^{2/3} - 4/3\ C_2\lambda_2^{-2/3}\lambda_3^{4/3}$$

$$g_3 = 4C_2\lambda_2\lambda_3 - 4/3\ C_1\lambda_2^{-1/3}\lambda_3^{-1/3} - 16/3\ C_2\lambda_2^{1/3}\lambda_3^{1/3}$$

$$g_{3T} = 2(C_1 + C_2) + 2C_2\lambda_2^2\ \ 2/3\ C_1\lambda_2^{2/3}\lambda_3^{-4/3} - 4/3\ C_2\lambda_2^{4/3}\lambda_3^{-2/3}$$

After changing variables $(\alpha_1,\alpha_2) \longrightarrow (0,1)$ and using one point integration we have:

$$\lambda_2 = \frac{r_2 - r_1}{\alpha_2 - \alpha_1}$$

$$\lambda_3 = \frac{r_1 + r_2}{\alpha_2 + \alpha_1}$$

$$\mathbf{g} = 2\pi \left(\frac{f_2}{\alpha_2 - \alpha_1} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \frac{f_3}{\alpha_1 + \alpha_2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \frac{\alpha_2^2 - \alpha_1^2}{2} \qquad (7)$$

$$\mathbf{k} = 2\pi \left(\frac{g_3}{\alpha_2^2 - \alpha_1^2} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} + \frac{g_2}{(\alpha_2 - \alpha_1)^2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} + \frac{g_{3T}}{(\alpha_1 + \alpha_2)^2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right) \frac{\alpha_2^2 - \alpha_1^2}{2}$$

We can now quickly assemble global gradient and tangent matricies using equations (7).

REDUCED GRADIENT AND TANGENT / UPDATE

Given the unconstrained internal energy function above and internal pressure, F, the minimization problem defining deformed configurations is

$$\min \Pi = \sum_e U_e - \pi F(r_1^2 - \alpha_1^2)$$

$$\text{such that} \quad r_2^2 - r_1^2 = \alpha_2^2 - \alpha_1^2 \tag{8}$$

$$r_3^2 - r_2^2 = \alpha_3^2 - \alpha_2^2$$

$$\text{etc.}$$

We now linearize the constraint equations. That is, use

$$v_1 = v_{1o} + \frac{\delta v_1}{\delta r_1}\bigg|_o \delta r_1 + \frac{\delta v_1}{\delta r_2}\bigg|_o \delta r_2 + \ldots$$

$$v_2 = v_{2o} + \frac{\delta v_2}{\delta r_2}\bigg|_o \delta r_2 + \frac{\delta v_2}{\delta r_3}\bigg|_o \delta r_3 + \ldots \tag{9}$$

$$\text{etc.}$$

In this one - dimensional problem there will be one free variable - all others defined by the constraint equation (9) and we find

$$\begin{bmatrix} r_2/r_1 & 0 & 0 & 0 \\ -r_2 & r_3 & 0 & 0 \\ 0 & -r_3 & r_4 & 0 \\ 0 & 0 & -r_4 & r_5 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \cdot \begin{bmatrix} \delta r_2 \\ \delta r_3 \\ \delta r_4 \\ \delta r_5 \\ \cdot \end{bmatrix} = \begin{bmatrix} \delta r_1 \\ 0 \\ 0 \\ 0 \\ \cdot \end{bmatrix} \tag{10}$$

Solving (10) we find

$$\begin{bmatrix} \delta r_1 \\ \delta r_2 \\ \vdots \\ \delta r_n \end{bmatrix} = \begin{bmatrix} 1 \\ r_1/r_2 \\ \vdots \\ r_1/r_n \end{bmatrix} \begin{bmatrix} \delta r_1 \end{bmatrix} \tag{11}$$

Equation (11) can be used to compute the reduced global gradient and tangent consistent with the linearized constraint equation (9). We have

$$g_r = \sum_{i=1}^{n} a_i g_i \qquad k_r = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k_{ij} \tag{12}$$

where $a_i = r_1/r_i$ and g_i, k_{ij} are from the global maricies.
Equation (12) is used to update r_1 using the Newton - Raphson method. The remaining variables are updated by sequentially solving the constraint equations.

RESULTS AND DISCUSSION

We solved the infinite cylinder problem discussed by Oden [2]. In particular the cylinder inner and outer radii were 7.0 inches and 18.625 inches respectively. The Mooney - Rivlin constants C_1 and C_2 (eq(5)) were taken as 80.0 and 20.0 psi and ten elements were used. The implicit variable elimination method reduced the eleven variable unconstrained problem to ONE variable. A Lagrange multiplier method would require twenty - one variables when the ten element hydrostatic pressure variables are added. Figure 2 shows the convergence of the inner radius with respect to the Newton - Raphson steps. The initial configuration was a poor guess but after two steps the log of the reduced gradient converged linearly, see Figure 3. The converged solution is the correct solution and is shown in Figure 4. In addition we solved another one dimensional problem involving the stretching of a rubber rod. Again, the reduced gradient converged linearly.

It is important to note the difficulties involved in extending this method to two - dimensional problems (our original intent). The superlements suggested by Needleman and Shih [11] are apparently unavoidable. The implicit variable elimination method would be very attractive if the reduced gradient and tangent matricies could be computed at an element level. That is, if internal element displacement variables could be eliminated using the linearization of the constraint equations. Then, there would be no bandwidth change, the constrained gradient and tangent would be computed almost as quickly as the compressible case. This

method of eliminating internal variables fails because the eliminated variables cannot be updated so that the element volumes return to their original values. This is due to the interelement dependence of the constraint equations. One can carefully identify a set of global variables which can be eliminated and updated, etc. but a system of nonlinear equations must be solved at each Newton - Raphson iteration to exactly enforce incompressibility.

REFERENCES

1. M. Levinson, 'The application of the principle of stationary potential energy to some problems in finite elasticity', J. Appl. Mech., ASME, **87**, 656-660 (1965).

2. J. T. Oden, Finite elements of nonlinear continua, McGraw-Hill, New York, 1972.

3. J. T. Tielking and W. W. Feng, 'The application of the minimum potential energy principle to nonlinear axisymmetric membrane problems', J. Appl. Mech., 491-496 (1974).

4. J. C. Nagtegaal, D. M. Parks and J. R. Rice, 'On numerically accurate finite element solutions in the fully plastic range', Comp. Meths. Appl. Mech. Eng., 4, 153-177 (1974).

5. J. H. Argyris, H. Balmer, J. St. Doltsinis, P. C. Dunne, M. Haase, M. Kleiber, G. A. Malejannakis, H. P. Mlejnek, M. Muller and D. W. Scharpf, 'Finite element method - the natural approach', Comp. Meths. Appl. Mech. Eng., **17/18**, 1-106 (1979).

6. D. S. Malkus and T. J. R. Hughes, 'Mixed finite element methods - reduced and selective integration techniques: a unification of

concepts', Comp. Meths. Appl. Mech. Eng., **15**, 63-81 (1978).

7. D. S. Malkus, 'Finite elements with penalties in nonlinear elasticity', Int. J. Num. Meth. Eng., **16**, 121-136 (1980).

8. A. R. Johnson, C. J. Quigley and I. Fried, 'Large deformations of elastomer cylinders subjected to end thrust and probe penetration', Trans. of the Third Army Conf. on Applied Mathematics and Computing, Army Research Office, Report No. 86-1, 1986.

9. I. Fried and A. R. Johnson, 'Nonlinear computation of axisymmetric solid rubber deformation', Comp. Meths. Appl. Mech. Eng., **67**, 241-253 (1988).

10. I. Fried and A. R. Johnson, 'A note on elastic energy density functions for largely deformed compressible rubber solids', Comp. Meths. Appl. Mech. Eng., **69**, 53-64 (1988).

11. A. Needleman and C. F. Shih, 'A finite element method for plane strain deformations of incompressible solids', Comp. Meths. Appl. Mech. Eng., **15**, 223-240 (1978).

12. A. R. Johnson and C. J. Quigley, 'Buckled elastica in contact - finite element solutions', Trans. Second Army Conf. on Applied Mathematics and Computing, Army Research Office (AD-P004904), 1985.

13. A. R. Johnson and C. J. Quigley, 'Element level elimination of nonlinear constraints in total Lagrangian finite element formulations', Trans. Fifth Army Conf. on Applied Mathematics and Computing, Army Research Office, 1988.

14. A. R. Johnson and C. J. Quigley, 'Frictionless geometrically nonlinear contact using quadratic programming', In. J. Num. Meths. Eng., **27**, (1989).

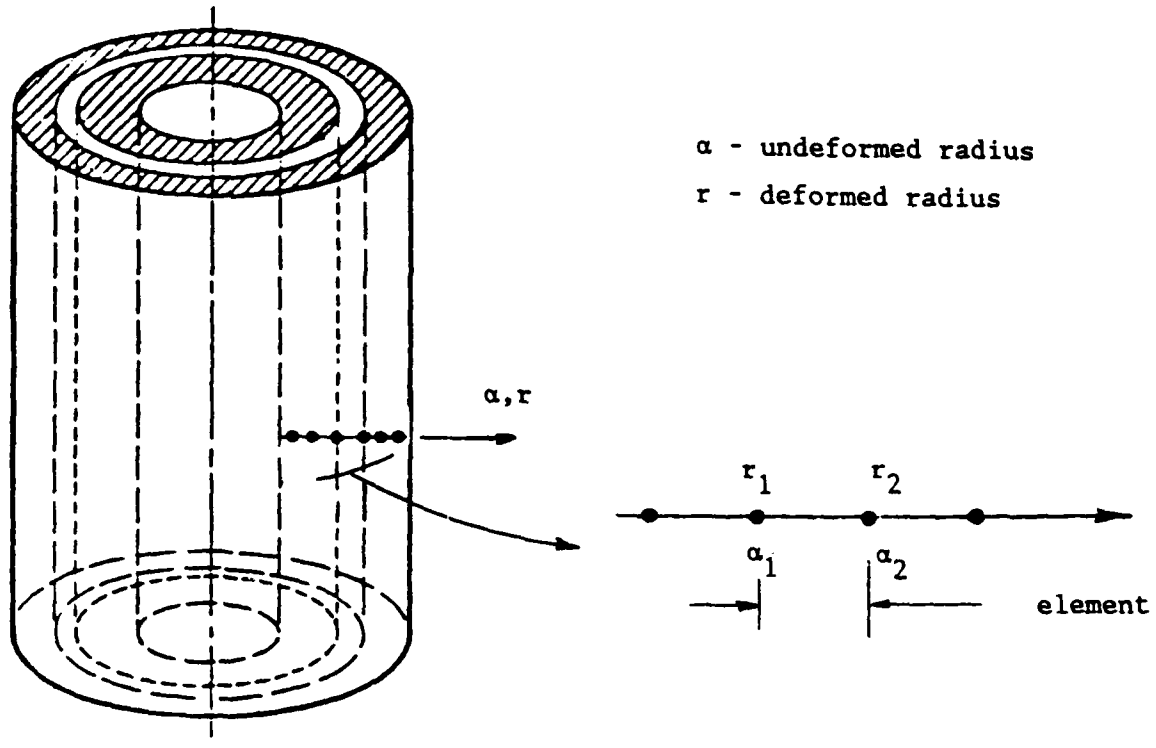


Figure 1. Infinite cylinder and finite element mesh.

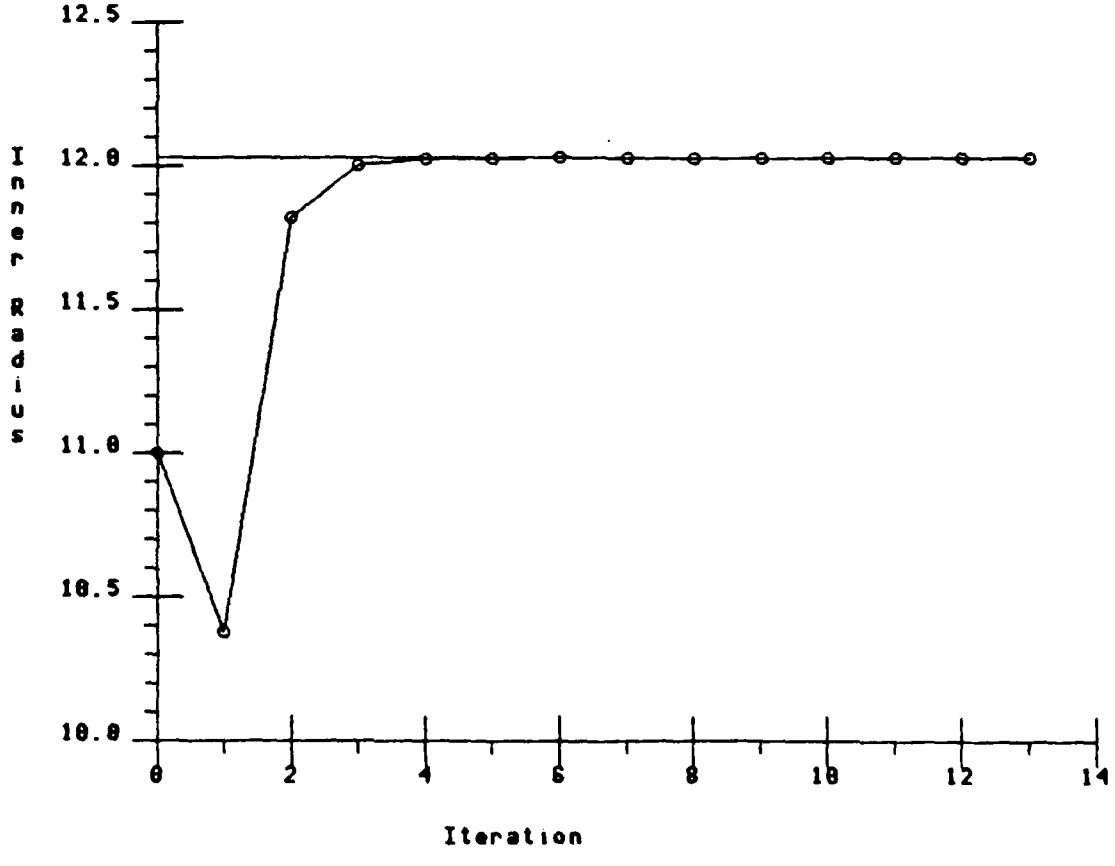


Figure 2. Inner radius vs Newton - Raphson iteration.

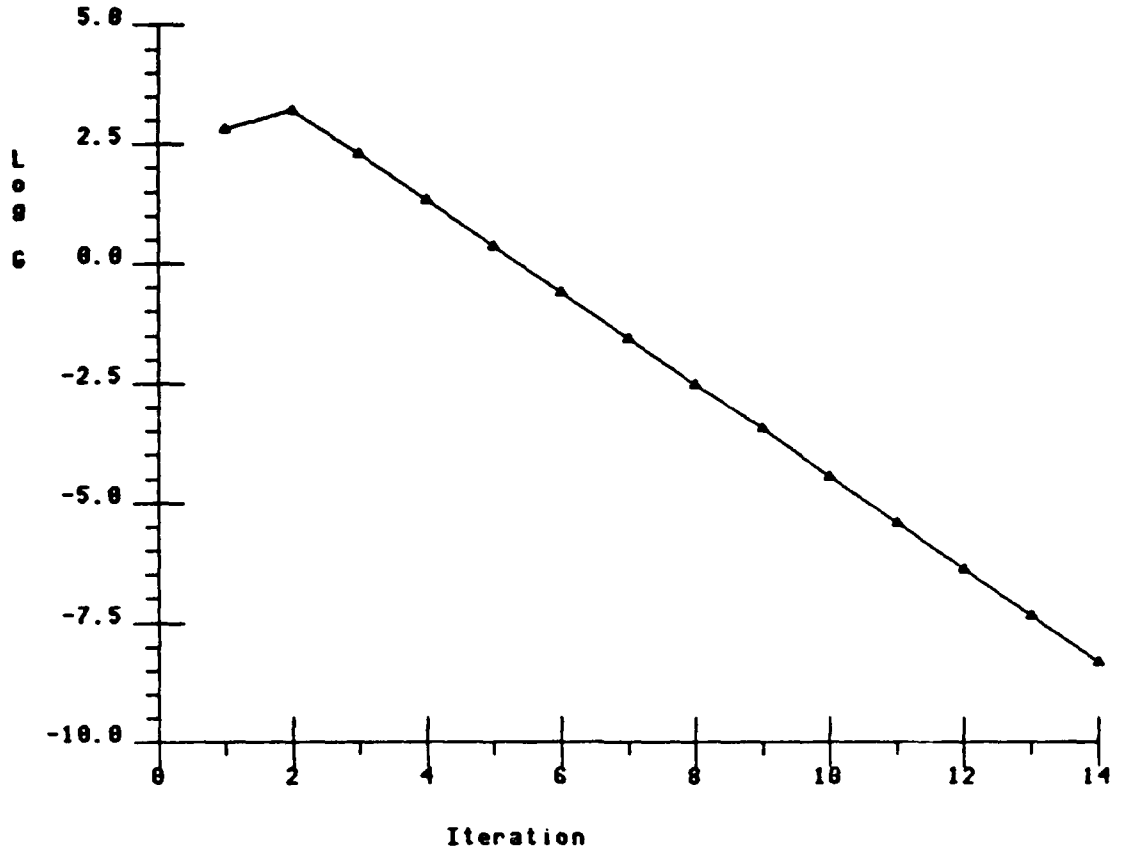


Figure 3. Log (reduced gradient) vs Newton - Raphson iteration.

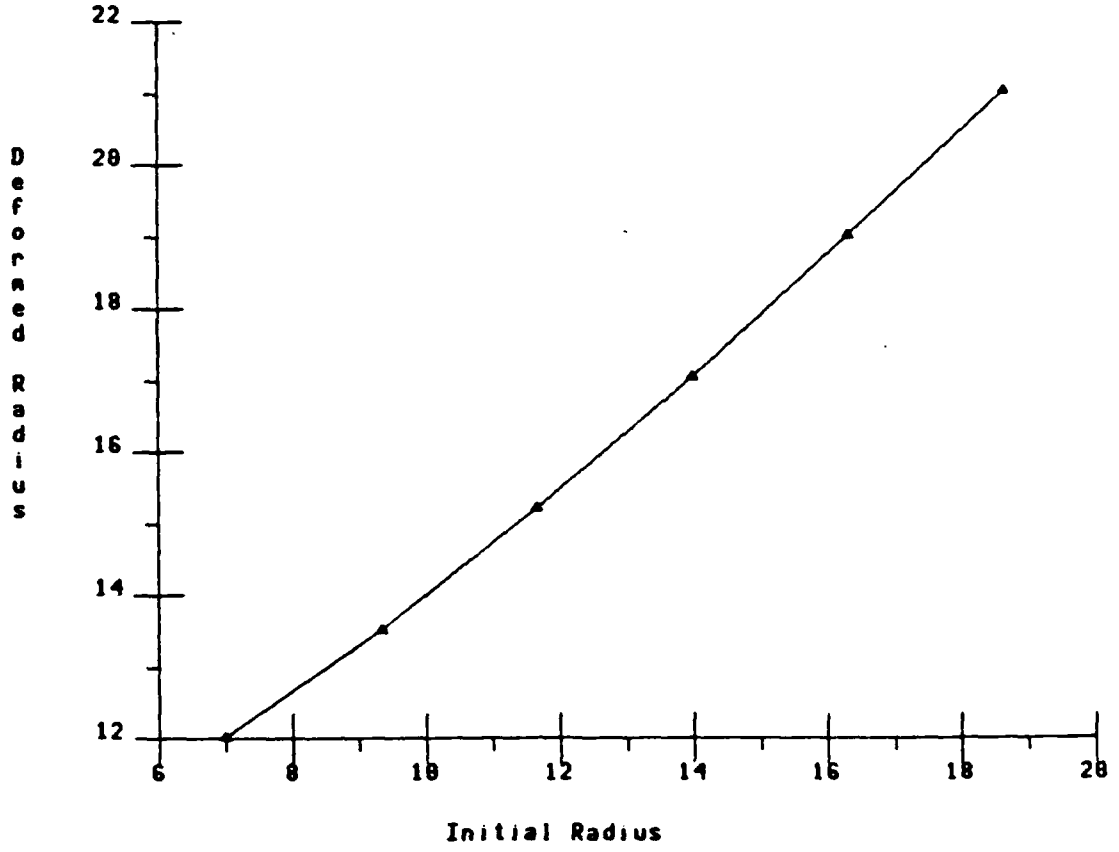


Figure 4. Deformed radius vs undeformed radius.

STRESS DISTRIBUTIONS NEAR MICROSTRUCTURAL INHOMOGENEITIES

Dennis M. Tracey and Paul J. Perrone

Mechanics and Structures Branch
U.S. Army Laboratory Command
Materials Technology Laboratory
Watertown, Massachusetts 02172-0001

ABSTRACT

Three-dimensional analyses have been conducted of elastic-plastic continua which contain pairs of spherical particles and voids. Response to shear loading was investigated with the intention of characterizing stress states at the microstructural level which result in void nucleation and softening, leading to shear strain localization in ultra-high strength steels.

INTRODUCTION. There is a great deal of evidence that ductile fracture of metallic alloys stems from the nucleation of voids at second phase microstructural particles. Nucleation occurs when either critical conditions at the interface are achieved, or when the strength of the particle is reached, causing a fracture of the particle. Either event produces local crack damage which deforms into a void as the plastic deformation of the sample proceeds. Plasticity theory has been applied to the case of void deformation in the presence of triaxial tension, and results have demonstrated that the void surface can experience strain levels far in excess of nominal values when the mean stress is above yield stress levels, Rice and Tracey (1). Consequently, the voids grow and the material progressively weakens as neighboring voids coalesce by impingement in such stress environments. Gurson (2) has developed a plasticity constitutive theory, including yield criterion and flow rule, to represent materials which dilate from the void growth mechanism. This theory is most properly applied to cases involving significant regions of high triaxial tension.

When the mean stress is low compared to the yield stress, the "void sheet" mechanism of internal damage is commonly observed in planes of maximum shear, Rogers (3). This may involve nucleation from different size scale populations of particles. For instance, pairs of relatively large voids nucleated from grain refinement particles might elevate the stress and strain fields locally to nucleate a number of smaller voids from strengthening particles between pairs. Coalescence would occur by cracking of ligaments after a critical spacing is achieved.

In this report three-dimensional elastic-plastic results are given for the stress and strain fields that develop near void and particle pairs. The matrix material has been modeled as a non-hardening elastic-plastic metal, while the particles are considered to be elastic with a modulus twice that of steel. The results vividly demonstrate how nominally uniform shear conditions are perturbed near interacting inhomogeneities. Comparisons with plane

strain solutions are made and these demonstrate the importance of including three-dimensional effects into micromechanical computer simulations.

The analyses modeled a sample of metal, nominally under uniform shear loading, containing one inhomogeneity pair (either a pair of voids or a pair of particles) buried within the sample far from its boundaries. Spheres placed at a distance of three diameters is the pair configuration we have limited our discussion to in this paper. Two separate orientations of the pair with respect to the direction of applied shear were considered, as illustrated in Figure 1. As shown in the top quarter section, one orientation has the applied shear directed parallel to the pair centerline. The bottom quarter section illustrates the other orientation which has the applied shear directed perpendicular to the centerline.

NUMERICAL FORMULATION. A finite element formulation was employed in the study to ascertain fully plastic solutions within the small strain theory of non-hardening plasticity. These solutions can be used to approximate the conditions that would prevail near interacting voids and particles at the point of incipient flow localization on the macroscale. Not considered here are solutions representing conditions of large deformation which develop after localization has initiated.

Specifically, an incremental elastic-plastic finite element formulation was used. The fully plastic solution which provides the local flow field of interest is achieved numerically by incrementally tracing the loading parameters (boundary displacement here, as described below) from the initial unstressed state. The approach consists of approximating the undetermined displacement rate field with standard piecewise defined finite element interpolations. The primary discrete variables are nodal displacement rates (increments) which are determined at each step of loading.

To achieve the desired uniform remote strain state, boundary nodes were constrained to displace according to the specified state. These constraints were imposed at each increment and magnitudes were maintained in fixed proportions. If these specified displacement increments are denoted u_1 through u_m, the matrix equation for the vector of undetermined values $\underline{u}$ is given by

$$\underline{K}\,\underline{u} = -\underline{K}_1 u_1 - \ldots - \underline{K}_m u_m$$

In this equation, $\underline{K}$ is the constrained stiffness matrix and $\underline{K}_i$ are stiffness columns corresponding to the specified components. The stiffness terms vary according to the position of the elastic-plastic boundary and stress state (flow rule). An implicit scheme is used at each step to average the flow rule at each position within the plastic zone. The load history is discretized through an adaptive incrementation procedure discussed by Tracey and Freese (4, 5). The planar and three-dimensional versions of this formulation are embodied in the MTL FORTRAN code EPFE which was utilized in this study.

The treatment of a pair of inhomogeneities can be contrasted with formulations which have considered periodic arrangements. Figure 2 illustrates typical idealizations used in plane strain and axisymmetric analyses. The

pair model would appear to allow a more realistic assessment of interaction in actual microstructures. If loadings are restricted to tractions perpendicular to the model (unit cell) boundaries, discretization requirements are essentially the same for the periodic arrangements and the pair configuration. Results presented below suggest limited usefulness of the cylindrical geometry and that the spherical geometry should be modeled instead. The axisymmetric formulation displayed in the bottom of Figure 2 treats spheres but suffers from the requirement that the centerline of the spheres must be a principal stress direction. The three-dimensional pair model employed in this work allows treatment of any applied loading, including the cases of interest which have the centerline in the plane of maximum shear.

In Figure 1 if the coordinate axes are centered between the spheres, the planes x=y=z=0 then serve to define planes of reflective symmetry of the model. Geometrically the total region can be viewed as an assembly of eight identical subregions, each containing a single quarter sphere. The regions displayed in Figure 1 are unions of two of these elemental subregions. Actually, only an interior subregion is displayed. The total region had dimensions 13x10x10 relative to the sphere diameter D. By noting conditions of skew anti-symmetry in simple shear, it was possible to perform the analysis by discretizing a single subregion (octant) of the total model.

If the entire region were to be modeled, the simple shear state would be enforced in the top problem of Figure 1 in the following way. The two yz boundary faces would have the x displacement varying linearly with y, and on these faces the y component of displacement would be zero. The xz faces would have a constant value for the x displacement and a zero value of y displacement. The z component of traction would be zero on these four faces, corresponding to zero valued xz and yz shear stresses. Finally, the xy faces would be completely traction free.

When the skew anti-symmetry conditions are invoked on the planes of geometric symmetry, the following boundary conditions produce the state of nominal simple shear. In the top problem, on x=0 the y component of displacement as well as the x and z components of traction are zero. On y=0, the x component of displacement and the y and z traction components are zero. Finally, on z=0 the z displacement and x and y tractions are zero. Similar conditions can be applied to the faces of the elemental octant in the bottom problem where the applied shear is directed perpendicular to the centerline.

The finite element mesh used over the octant consisted of constant strain tetrahedra. The mesh was generated by first developing a field of eight-node brick elements which were individually subdivided into five tetrahedra. The mesh refinement was different in the analyses of the two void problems. The case of parallel shear had a mesh consisting of 4500 elements and 1200 nodes, each with three degrees of freedom. The perpendicular shear analysis was more refined in that there were 7100 elements and 1800 nodes in the mesh. The analysis of the pair of particles was conducted using the refined mesh for both load orientations. The additional complexity in the particle analysis involved discretization of the particles themselves. The quarter particle appearing in the octant was represented by 1300 elements to give a total mesh of 8400 elements and 2000 nodes.

VOID PAIR INTERACTIONS. The elastic solution for an isolated spherical void in simple shear has been described by Love (6). Referring to Figure 1, the maximum stress occurs at the two points on the void surface on the xz plane with tangent in the direction of applied shear. For a Poisson's ratio of 0.3, the stress concentration factor at these locations is 1.91, suggesting that void surface yielding should commence when the nominal shear level equals 1/1.91=0.52 times the material's yield strain in shear.

Four stages of the elastic-plastic solution are illustrated in Figure 3 for a pair of spherical voids spaced at a distance of three diameters under a remote shear directed parallel to the centerline. Plastic zones are represented in a quadrant by regions consisting of tetrahedron elements which have met the yield condition at the load level indicated. As anticipated from the classical elasticity solution, yielding first occurs in this quadrant at the void surfaces 90^0 from the pair centerline in the xz plane. As load is increased, plasticity spreads from these locations. In the top left, corresponding to a remote strain 0.80 times the yield strain, most of the void surfaces have yielded, but there is no plasticity between voids. Significant yielding between the voids has occurred at 94% of yield, as demonstrated in the bottom left. At 96% of yield, bottom right, the separate plastic zones have merged, leading the way for a mechanism of extensive plastic straining between voids.

The strain intensification that occurs along the centerline of the void pair is summarized in Figure 4. Data are plotted for the two spherical void pair problems and also for the cylindrical void pair problem that has been discussed by Tracey, Freese, and Perrone (7). These problems are individually considered in the two top and the bottom left plots of Figure 4. The results of the three problems are contrasted in the bottom right plot which has peak local strain plotted against nominal strain level.

The component of strain that is plotted for each case corresponds to the nominal simple shear state, e.g. yz component for the top left problem. The data are presented relative to the material's yield strain in shear. The distributions along the centerline are plotted for x/D values from 0.5 to 2.5, which corresponds to the distance between void surfaces.

When the applied shear is directed perpendicular to the pair centerline (top left), the centerline strain maxima occur on the void surfaces. The results for incipient yield (nominal strain = 0.49 times yield strain) demonstrate the extremely localized effects of inhomogeneities in an elastic field. As can be seen, the elastic solution has the strain elevated over the nominal value only within a distance of one void radius from the void surfaces. Hence, there is effectively no interaction in the elastic pair problem with a center spacing of three diameters. Consistent with Love's (6) isolated void result, the strain maxima are approximately twice the nominal value before plastic yielding intervenes. At general yield, the strain maxima have increased to about three times the nominal value and interaction is evident with mid-centerline strain magnitudes significantly exceeding the nominal value.

The analysis of the spherical void pair with shear parallel to the centerline was conducted using a mesh that was too coarse to adequately capture the shear free condition which holds at x=0.5 D and 2.5 D.

Nonetheless, the character of the elastic-plastic solution is thought to be reasonably represented in the top right plot. As in the other case, the elastic solution shows strain variations only within one radius of the void surfaces, with the nominal strain value realized over the middle half of the span between the voids. At general yield, the strain exceeds the nominal value over the entire ligament. The plot shows a modest peak at roughly 3/4 of a radius from the surfaces and strain levels roughly 30% over the nominal strain.

The cylindrical void pair analysis shows strain amplification levels greatly exceeding those found in the spherical void analyses. For this plane strain case, the mesh refinement was adequate to capture the shear free conditions at x=0.5 D and 2.5 D. The elastic solution shows a strong gradient out to a distance roughly 1/2 of a radius from the surfaces, otherwise reaching a near uniform state between the voids. Interaction is evident in this problem even in the elastic regime with the strain between voids approximately 50% over the nominal value before yielding occurs. Plastic zones develop at the void surfaces and separately in the center of the ligament for this problem. When these zones link, distinct strain maxima develop at positions roughly 3/4 of a void radius from the surfaces. The strain intensification is seen to increase in severity as general yield conditions are approached.

The three solutions are compared in the bottom right plot of Figure 4. Curves show the variation of local peak strain for each case as a function of nominal strain level. Of the two spherical pair cases, it can be seen that the orientation perpendicular to the load induces the highest local strains. Nonetheless, a comparison of the top left and right plots shows that the strain level attained in the middle of the centerline is essentially independent of orientation. The strain magnitudes found in the cylindrical void pair case are intermediate to the spherical pair results when plastic zone size is small. As can be seen, at approximately 75% of general yield corresponding to extensive local yielding, the peak strain values begin to take on values exceeding those found in the spherical pair cases. At general yield the local strain and strain rates for this case greatly exceed the values found for the spherical void problems.

It is the strain rate field that is most useful in assessing the local intensification of the nominal state once general yield conditions are achieved. Before general yield this field continually changes, as the plastic zones change, but thereafter, within the small deformation and nonhardening assumptions, the field remains constant relative to the nominal value. Figure 5 illustrates contours of shear strain rate (normalized by nominal strain rate) for the perpendicular loading. Results are given for nominal strain levels before (.94) and after (1.03) general yield. The contours are drawn over a quadrant of the xz midplane of the model. It is apparent that the maximum strain rate occurs in each case at the point of strain concentration at the void surface on the centerline.

At the lower load level, the maximum strain rate is approximately 4.3 times the nominal value. The gradient is steep, with the middle of the centerline experiencing a modest value of approximately 1.5. Little interaction is apparent at this nominal strain, as the 4.3 value holds on the opposite side of the void surface as well as at the surface-centerline

intersection point. At the higher load level, interaction is suggested by a maximum of 10.0 and the somewhat lower value 8.7 across the surface. In this case, the mid-centerline strain rate is 4.5 times the nominal value, indicating significant elevated strain rates along the entire centerline after general yield is achieved.

PARTICLE PAIR INTERACTION. The void pair analyses obviously have neglected the presence of nucleating particles, and thus are applicable to the study of post nucleation effects resulting from the creation of interior traction free surfaces. The field near perfectly bonded elastic spherical particles was studied by performing an elastic-plastic finite element analysis which modeled particles as elastic with infinite yield strength and a modulus twice that of the elastic modulus of the elastic-plastic matrix in which they reside. As in the void pair analyses, a particle pair with a three diameter spacing was considered.

The strain intensification is plotted in Figure 6 for the two loading orientations considered above. Curves display the strain distribution through the particles and along the centerline between them. In these problems incipient yield was found to occur at a nominal shear strain equal to 77% of the shear yield strain. Eshelby's (8) analysis of isolated ellipsoidal particles in elastic fields demonstrated a uniform strain state within particles. The finite element results displayed in Figure 6 agree with this result and have a near uniform state within the particles even after extensive matrix yielding has occurred. At incipient yield the shear strain of the particles is approximately 50% of the shear yield strain of the matrix, consistent with the difference in elastic moduli.

In the left plot of Figure 6, for the case of orientation perpendicular to the shear load, it can be seen that the distribution is continuous across the particle/matrix interface. The nominal value of shear strain is reached at the middle of the centerline. For this orientation, there is essentially no strain intensification over the nominal value along the void pair centerline.

The right plot of Figure 5 displays the shear strain intensification for the case of particles oriented in the direction of the applied load. At incipient yield the magnitude of strain in the particles is slightly higher than 50% of the shear yield strain of the matrix and at a load slightly greater than general yield, the magnitude is approximately 70% of the shear yield strain. Across the interface the shear strain is discontinuous and jumps from a subnominal value in the particle to the maximum value found in the matrix. At incipient yield, this maximum value of strain is equal to the shear yield strain of the matrix. The severe gradient shows a decrease to the nominal value of shear strain within one half of a particle radius into the matrix. As loading progresses, the shear strain rate intensifies on the matrix side of the interface corresponding to the occurrence of extensive plastic deformation.

SUMMARY. Results have been presented for the three-dimensional aspects of interaction of pairs of voids and particles in shear. While the work has been motivated by metallurgical needs, particularly the need to develop

microstructures for the delay of void nucleation, clearly, much remains to be done to guide alloying from a mechanics basis. Future work on pair interaction in shear must address void nucleation, the spacing issue and a more complete assessment of orientation effects. Ultimately, the goal is to consolidate the simulation features, so that the necessary data and methodology will be available to allow microstructural design for ultra-high strength and toughness.

REFERENCES

1. J. R. Rice and D. M. Tracey, "On The Ductile Enlargement of Voids in Triaxial Stress Fields", Journal of the Mechanics and Physics of Solids, Vol. 17, 1969, pp. 201-217.

2. A. L. Gurson, "Continuum Theory of Ductile Rupture Void Nucleation and Growth: Part I Yield Criteria and Flow Rules for Porous Ductile Materials", Journal of Engineering Materials and Technology, Vol. 99, 1977, pp. 2-15.

3. H. C. Rogers, "The Tensile Fracture of Ductile Metals", Transaction of The Society of AIME, Vol. 218, 1960, pp. 498-506.

4. D. M. Tracey and C. E. Freese, "Adaptive Load Incrementation in Elastic-Plastic Finite Element Analysis", Computers and Structures, Vol. 13, 1981, pp. 45-53.

5. D. M. Tracey and C. E. Freese, "A Variable Load Stop Solution Approach for Incremental Tangent Modulus Finite Element Analysis", AMMRC TR 79-47, 1979.

6. A. E. H. Love, A Treatise on the Mathematical Theory of Elasticity, 4th ed., Dover Publications, New York, 1944.

7. D. M. Tracey, C. E. Freese, and P. J. Perrone, "Micromechanics of Shear Banding in High Strength Steel", Trans. Fourth Army Conference on Applied Mathematics and Computing, ARO Report 87-1, pp. 103-116, 1987.

8. J. D. Eshelby, "The Determination of the Elastic Field of an Ellipsoidal Inclusion, and Related Problems", Proceedings of the Royal Society, A241, 1957, p. 376.

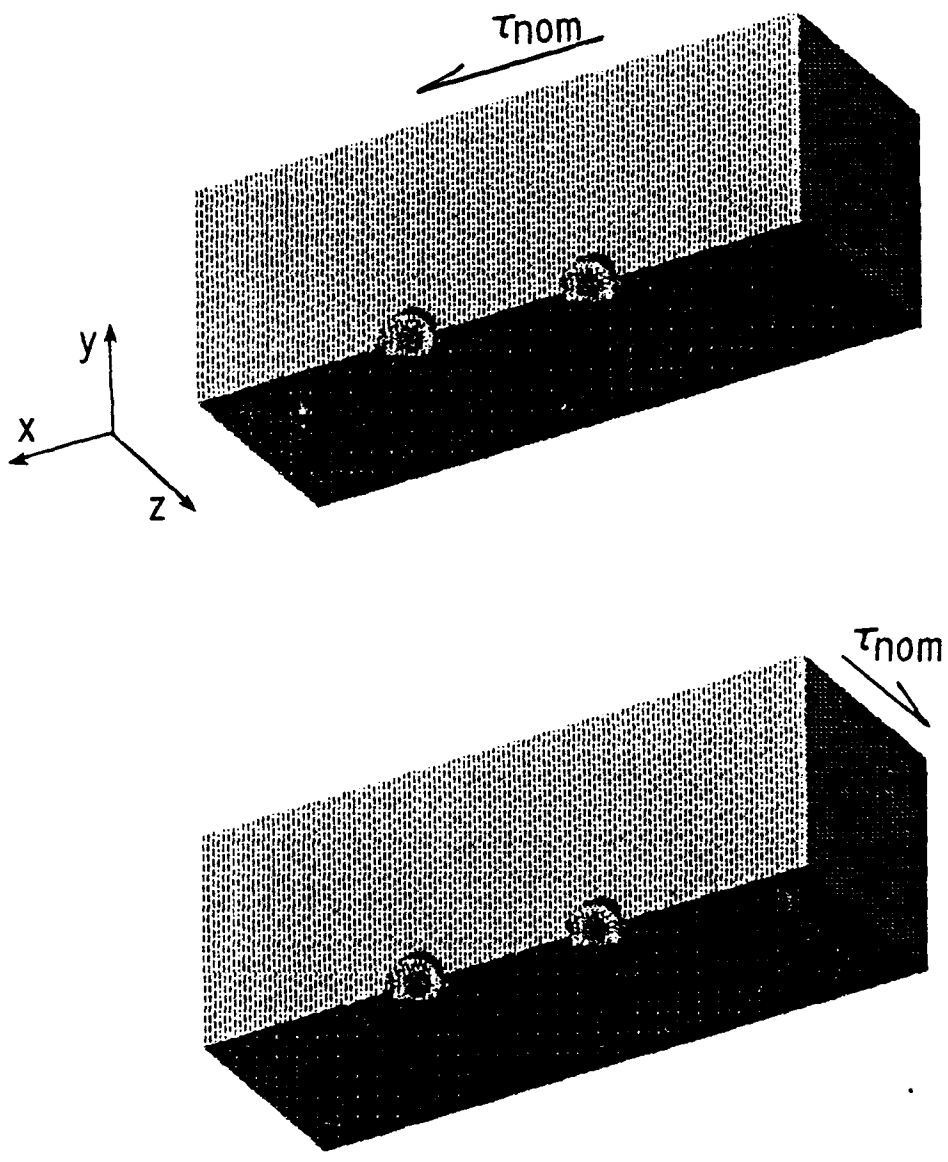


Figure 1. Quarter section of inner region containing void/particle pair under far field simple shear loading. In top drawing, shear load is directed parallel to pair centerline. Bottom drawing has shear load perpendicular to centerline.

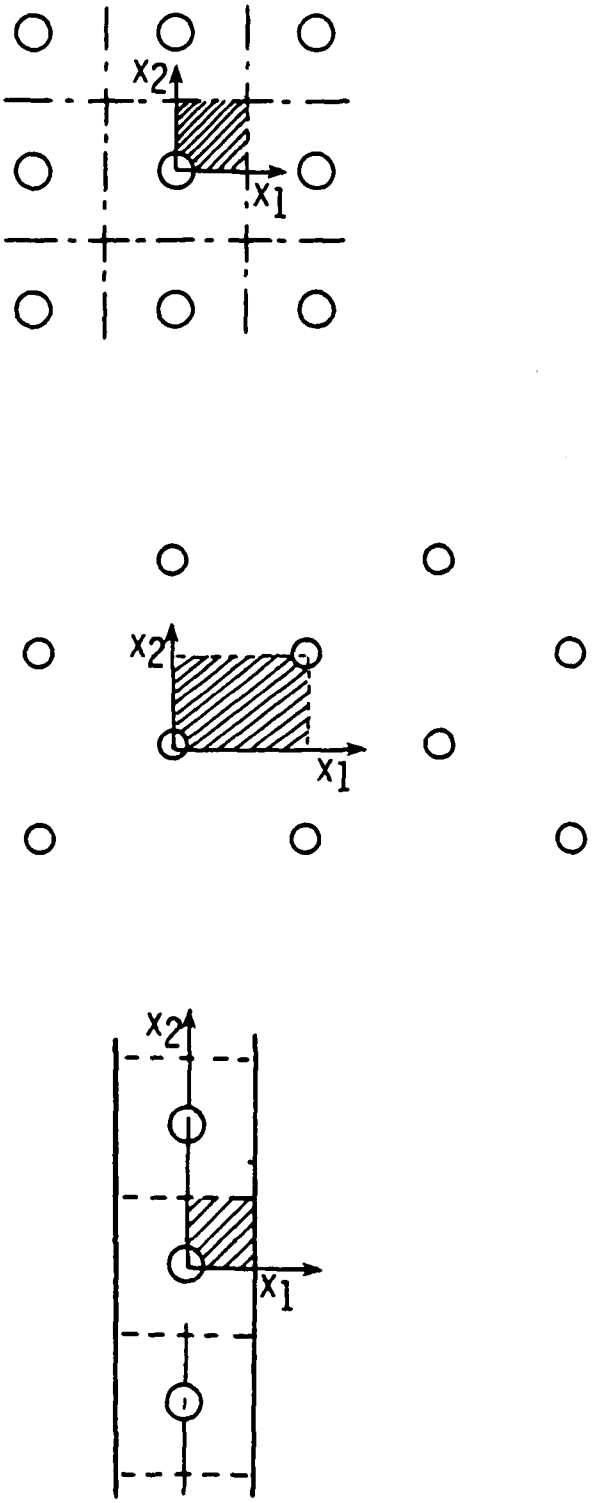


Figure 2. Unit cells (shaded) for two-dimensional analysis of interaction in periodic void arrangements and biaxial loading. Top and middle: circular cylindrical voids in plane strain, bottom: spherical voids under axisymmetric loading.

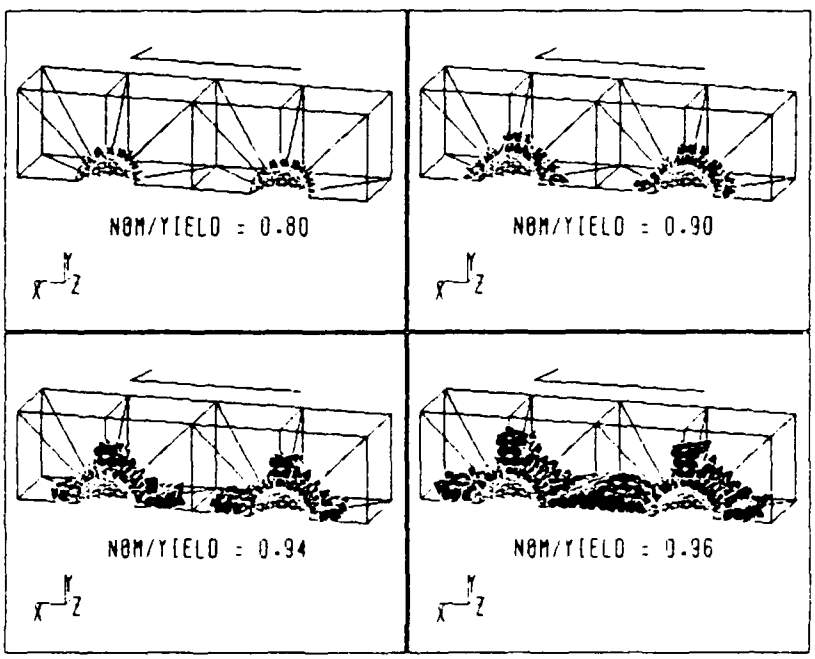


FIGURE 3. Plastic zone growth near spherical void pair before general yield conditions are achieved for shear parallel to centerline of voids.

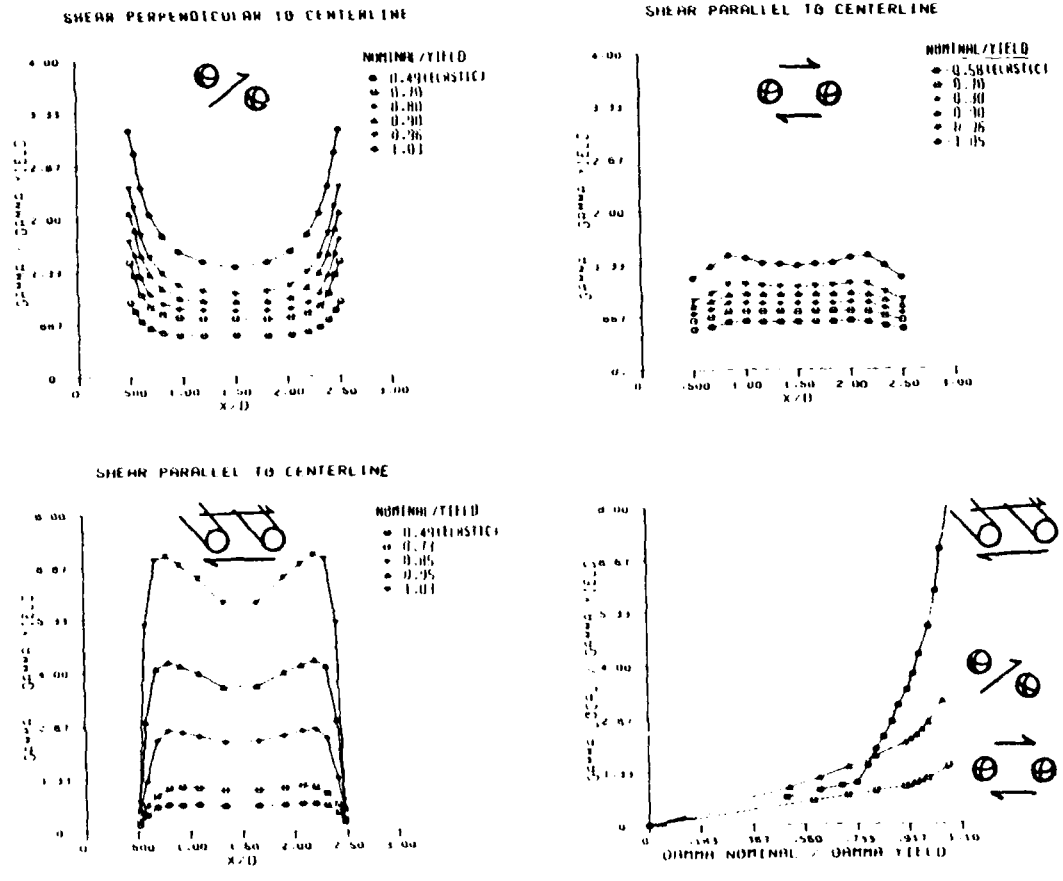


Figure 4. Strain variation along centerline for the spherical void problems (top) and cylindrical void problem (bottom left). In bottom right, peak local strain is plotted versus nominal strain.

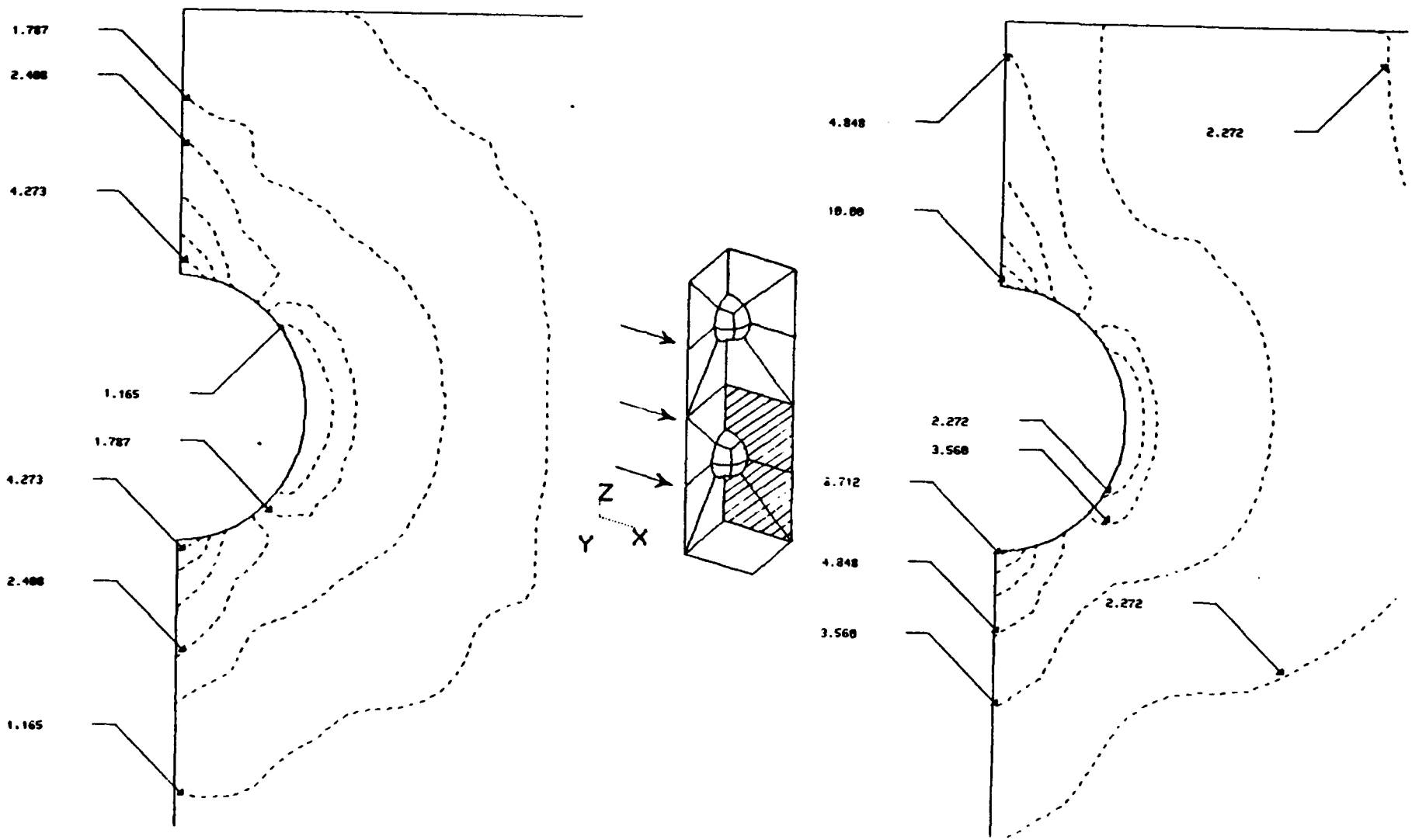


Figure 5. Contour plots of shear strain rate on midplane quadrant shaded in sketch, for voids oriented perpendicular to shear loading. Results are at nominal strain levels of 0.94 (left) and 1.03 (right) times yield strain in shear.

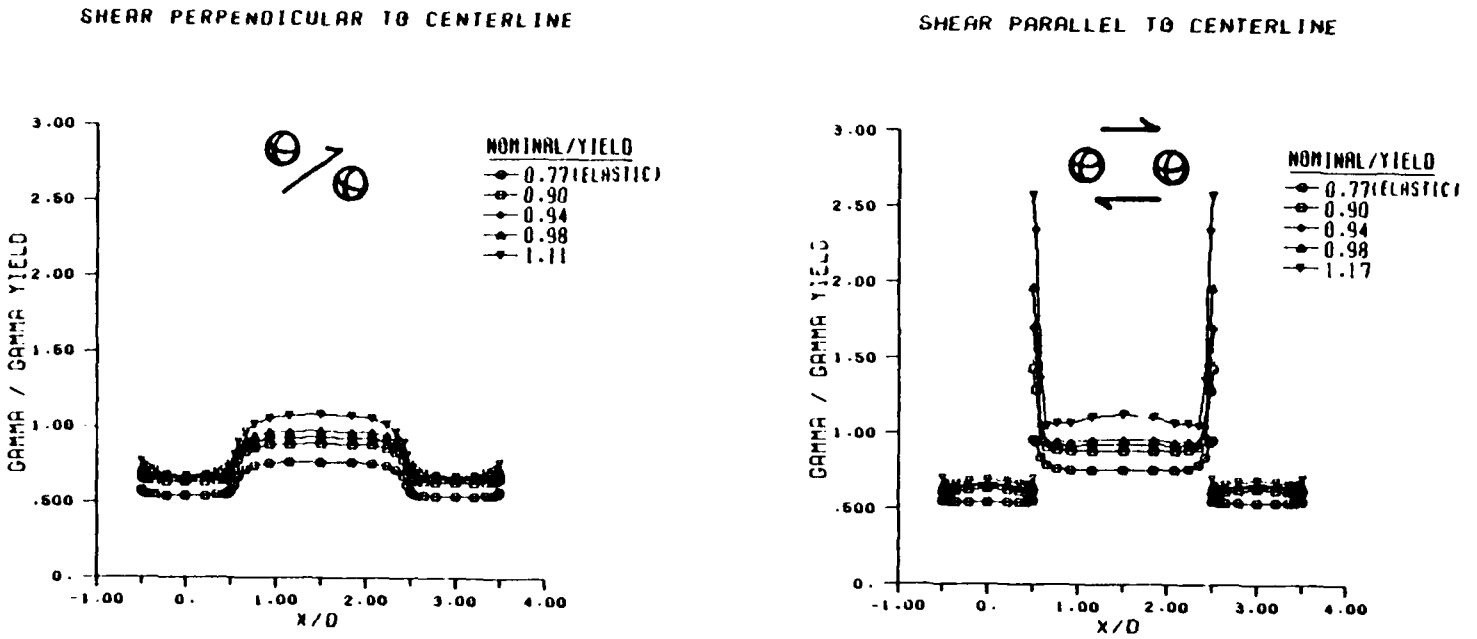


Figure 6. Strain distributions within elastic spherical particles and elastic-plastic matrix for shear loadings perpendicular and parallel to pair centerline. Distributions are plotted for nominal strain levels from incipient to general yield.

FINITE ELEMENT ANALYSIS OF SWAGE AUTOFRETTAGE PROCESS

Peter C. T. Chen
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

ABSTRACT. Swage autofrettage process is often used to produce favorable residual stresses in the tube. In this paper a finite element analysis of the swage autofrettage process is reported. The nonlinear finite element program (ABAQUS) is used to obtain numerical results for the displacements, strains, and stresses in the tube during and after autofrettage. Approximate solutions are obtained for one- and two-dimensional tubes pressed by rigid or elastic mandrel. The longitudinal effect and the elasticity of the mandrel on the permanent bore enlargement and the residual stresses are discussed.

INTRODUCTION. To increase the maximum elastic carrying capacity and to enhance the fatigue life, residual stresses are often produced in tubes through autofrettage [1]. Many solutions are reported for the hydraulic autofrettage process. The thick-walled cylinders are subjected to uniform internal pressure of sufficient magnitude to cause plastic deformation and then the pressure is removed.

A more economical way of producing residual stresses in thick-walled cylinders is the swage autofrettage process. This process is carried out by a mandrel, the diameter of which is greater than the inner diameter of the tube. The mandrel is driven through the tube from one end to the other. A rigorous analysis of this process is difficult. Recently a simple analysis of the swage autofrettage process was reported [2]. The model used was a one-dimensional plane-strain problem of mandrel-tube assembly. The steel tube was assumed to be elastically-ideally plastic, obeying Tresca's yield criterion and the associated flow theory, but the tungsten carbide mandrel was elastic. The deformation and stress distribution during swaging were obtained by solving the shrink-fit problem beyond the elastic limit. After swaging, the permanent bore enlargement and residual stresses were calculated by an unloading analysis [2], taking into account the Bauschinger effect and the strain-hardening during unloading [3].

The solution reported in [2] is in closed-form and the numerical results indicate that the agreement between the calculated and experimental data is excellent in zones with larger wall ratios but not so good in zones with wall ratios less than two. The differences in thinner sections may be due to the longitudinal bending effect since the simplified analytical analysis is one-dimensional and bending is neglected. In order to determine the longitudinal effect, a two-dimensional analysis based on the finite element method is made. In this paper, the finite element solutions are presented for both one- and two-dimensional models and a comparison of the results is given.

METHOD OF ANALYSIS. Since the total length of the tube is about sixty times the diameter of the mandrel, a complete finite element analysis of the swage autofrettage process is very difficult. As the mandrel is driven through the tube from one to the other, the simulation requires the study of elastic-plastic moving contact and separation history between two deforming bodies. In addition, a considerable amount of computer storage and run time would be required. In the present study, however, approximate finite element models are chosen to represent swaging in only a part of the tube (zone 3). We consider the process as quasi-static and neglect the effect of sliding and friction between the mandrel and tube. We want to obtain the information about the deformations and stresses for a section at only two particular stages, i.e., when the mandrel is at or far away from the position of interest. To achieve this purpose, we can simplify the simulation by studying two related problems, i.e., shrink-fit and complete unloading. When the mandrel is at the position of interest, we consider a shrink-fit problem of the mandrel-tube assembly to obtain the maximum deformation and stresses during swaging. When the mandrel is driven far away from the section, we study it as a complete unloading problem of the mandrel-tube assembly to obtain the information about the permanent bore enlargement and residual stresses after swaging. Figure 1 shows a one-dimensional interference-fit problem of the mandrel-tube assembly. Initially, the inner and outer radii of the mandrel is c. Given the interference I = c-a, we can determine the interference pressure p and the deformation and stresses in the mandrel and tube. In general, this problem can be solved only by an iterative approach. If the mandrel were rigid, then the direct approach using displacement constraints could be applied. The results based on this approach were obtained so we could discuss the effect of elasticity in the mandrel. The actual strength ratio of tungsten carbide to steel is about three. For the problem considered here, it is reasonable to assume that the steel tube is elastic-plastic, obeying Mises' yield criterion and the associated flow theory, but the tungsten carbide mandrel remains elastic. The finite element analysis is carried out by using the nonlinear program, ABAQUS [4]. Two types of elements used are shown in Figure 2. The axisymmetric solid elements (CAX4) are used to model the tube and mandrel. The interface elements (INTER2A) are used to model the separation or interference fit between the mandrel and tube. Truss elements (CID2) can also be used to model the mandrel because the displacement U_1 is directly related to the external pressure p by

$$U_1/c = -(1-\nu_1-2\nu_1^2)p/E$$

where E_1, ν_1 are elastic constants of the mandrel.

FINITE ELEMENT MODEL. Figure 3 shows a two-dimensional finite element model (E3) chosen to represent the swaging process in zone 3. The model is considered symmetric with respect to z = 0 so that only half of the model is shown. We have used 133 and 21 elements of type CAX4 to represent the tube and mandrel, respectively, with a = 1, b = 1.431, c = 1.007415. There are eight interface elements of type INTER2A to represent the interaction between the tube and mandrel. Figure 3a shows an interference-fit problem of the mandrel-tube assembly. This model is used to determine the maximum deformation and stresses during swaging. Figure 3b shows a complete unloading problem when the two parts

are separated. This problem is used to determine the permanent deformation and residual stresses after swaging. In order to determine the longitudinal bending effect, we would like to compare the two-dimensional analysis with the one-dimensional analysis. The one-dimensional model (E1) consists of ten elements (of type CAX4) each for the tube and mandrel with one interface element (of type INTER2A). Another one-dimensional model (E2) is to replace the mandrel by one or two Truss elements of type CID2. The material constants used are E = 200, σ_o = 1, ν = 0.3 for the high strength steel and E_1 = 590, σ_1 = 3.33, ν_1 = 0.258 for the tungsten carbide. The materials exhibit no strain-hardening. In the modeling and computation we have used the dimensionless quantities with the inner radius (2.283 inches) as the unit length and the initial yield stress (150 Ksi) as the unit stress. The actual quantities can be obtained easily if needed.

In the above three models (E1, E2, E3), the tube is elastic-plastic, but the mandrel remains elastic. If the strength ratio of the mandrel material to tube material is very large, then the mandrel can be regarded as rigid. In order to determine the effect of elasticity in the mandrel, we have chosen three finite element models (R1, R2, R3). Models R1 and R2 represent one-dimensional plane-strain and plane-stress cases, respectively. We have used ten elements of type CAX4 to represent the tube. The model R3 is the same as the model E3 shown in Figure 3 except that the mandrel is replaced by a rigid block.

Following the instructions given in Reference [4], we have prepared the input data for each of the six finite element models. For each model we ran the problem in two steps, i.e., loading and unloading. The input deck for the finite element analysis of model E1 is shown in Table 1.

RESULTS AND DISCUSSIONS. For each of the six models (R1, R2, R3, E1, E2, E3) discussed in the preceding section, we have run the finite element program successfully. The numerical results for the displacements, strains, and stresses in the tube during and after swaging have been obtained. Only the results for the stresses along the radial direction near z = 0 and the displacements along the bore are presented graphically.

When the mandrel is assumed to be rigid, the displacement at the bore is equal to the given interference. The results for the stresses based on models (R1, R2, R3) are presented in Figures 4 through 6. When the interference is only half of the maximum, the state of stresses remains elastic as shown in Figures 4 and 5. When the maximum interference (I = 0.007415) is reached, the state of stresses is elastic-plastic. The effect of interference on the distributions of hoop and axial stresses can be seen in Figures 4 and 5, respectively. By comparing the results for model R1 (one-dimensional, plane-strain case) and model R3 (two-dimensional case), we can also see the influence of the longitudinal effect on the hoop and axial stresses. The influence on the maximum axial stresses is very significant as shown in Figure 5. Unloading after the maximum interference is reached, we have obtained the residual stresses as shown in Figures 5 and 6 for the axial and hoop stresses. A comparison of these residual stresses indicates that the differences between one- and two-dimensional models (R1 and R3) are very minor. Models R1 and R2 represent plane-strain and plane-stress cases, respectively, and both models are one-dimensional.

TABLE 1. THE FINITE ELEMENT INPUT DECK FOR MODEL E1

```
*HEADING
TUBE-MANDREL ASSEMBLY AND SEPARATION
*NODE
1,.
11,1.007415
21,1.0
31,1.431
101,.        , 0.05
111,1.007415, 0,05
121,1.0      , 0.05
131,1.431    , 0.05
*NGEN,NSET=SIDE1
1,11
101,111
*NGEN,NSET=SIDE2
21,31
121,131
*NSET,NSET=BORE
1,101
*ELEMENT,TYPE=CAX4
1,1,2,102,101
11,21,22,122,121
*ELGEN,ELSET=MANDREL
11,10
*ELGEN,ELSET=TUBE
11,10
*SOLID SECTION,ELSET=MANDREL,MATERIAL=CARBIDE
*MATERIAL,NAME=CARBIDE
*ELASTIC
590., .258
*PLASTIC
3.33
*SOLID SECTION,ELSET=TUBE,MATERIAL=STEEL
*MATERIAL,NAME=STEEL
*ELASTIC
2.E2, .3
*PLASTIC
1.
*ELEMENT,TYPE=INTER2A,ELSET=SFIT
101,111,11,121,21
*INTERFACE,ELSET=SFIT
*FRICTION
 .0
*BOUNDARY
SIDE1,2
SIDE2,2
*STEP, NLGEOM,CYCLE=10
*STATIC,PTOL=1.E-4,DIRECT
1. , 1.
*END STEP
*STEP,NLGEOM
*STATIC,PTOL=1.E-4 ,DIRECT
1.,1.
*MODEL CHANGE, REMOVE
MANDREL,SFIT
*END STEP
```

When the mandrel is considered as elastic, the interference-fit assembly is solved iteratively. The same results for the one-dimensional models (E1 and E2) have been obtained. A comparison of two models (E1 and R1) for the hoop stresses during and after swaging is shown in Figure 7. The elasticity in the mandrel reduces the amount of overstrain from 70 to 60 percent. The numerical results for the two-dimensional model (E3) are presented in Figures 8 through 11. Figure 8 shows the distributions of hoop stresses during and after swaging. Figure 9 shows the corresponding distributions of maximum and residual axial stresses. Also shown in Figures 8 and 9 are the one-dimensional results based on model E1. A comparison of the results based on models E1 and E3 can determine the two-dimensional effect on these stresses. In Figure 10 we show the results for the radial stresses based on four models (E1, E3, R1, R3). Finally, the results based on several models for the radial displacement along the bore are presented in Figure 11. The displacements during and after swaging are represented by U and U", respectively. Also shown in the figure is the measured permanent bore enlargement. By comparing the results based on models E1 and R1, the elasticity effect gives a smaller value for U". If we include the two-dimensional effect with model E3, we get a value for U" even smaller than that based on the one-dimensional model.

REFERENCES

1. Davidson, T. E. and Kendall, D. P., "The Design of Pressure Vessels for Very High Pressure Operation," Mechanical Behavior of Materials Under Pressure, (H.L.P. Pugh, ed.), Elsevier Co., 1970.

2. Chen, P. C. T., "A Simple Analysis of Swage Autofrettage Process," Transactions of the Fifth Army Conference on Applied Mathematics and Computing, in press; also Technical Report ARCCB-TR-88030, Benet Laboratories, Watervliet, NY, July 1988.

3. Milligan, R. V., Koo, W. H., and Davidson, T. E., "The Bauschinger Effect in a High Strength Steel," Journal of Basic Engineering, Vol. 88, 1966, pp. 480-488.

4. "ABAQUS Users' Manual," Version 4.6, Hibbit, Karlsson, and Sorensen, Inc., 1987.

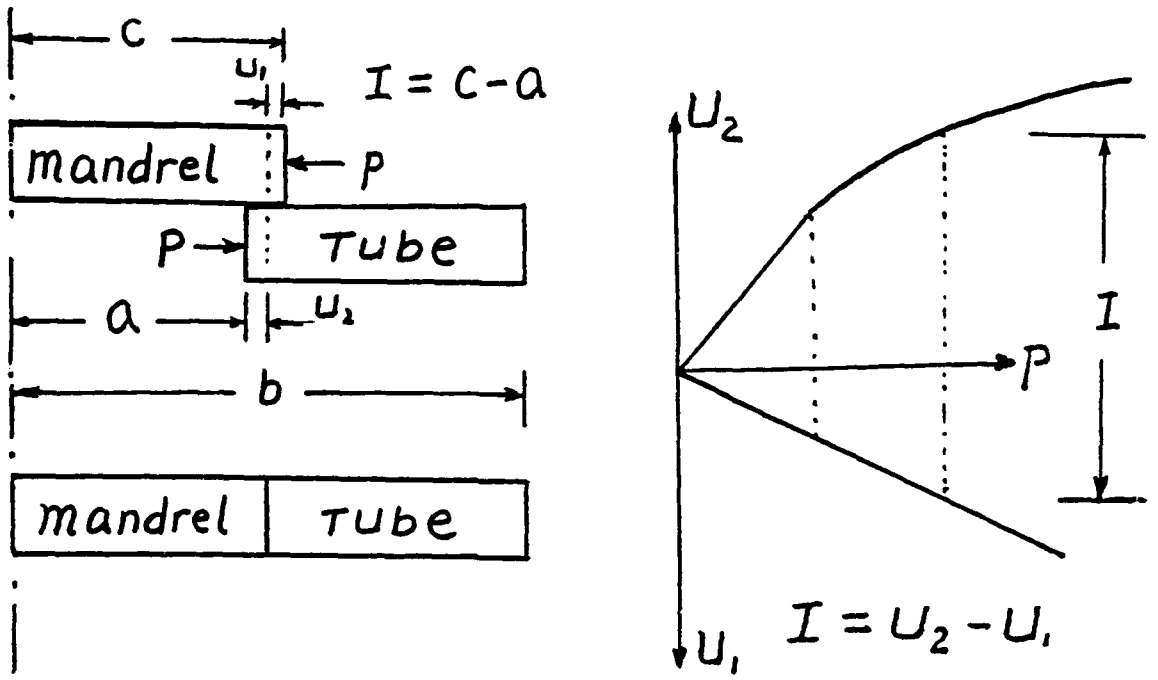
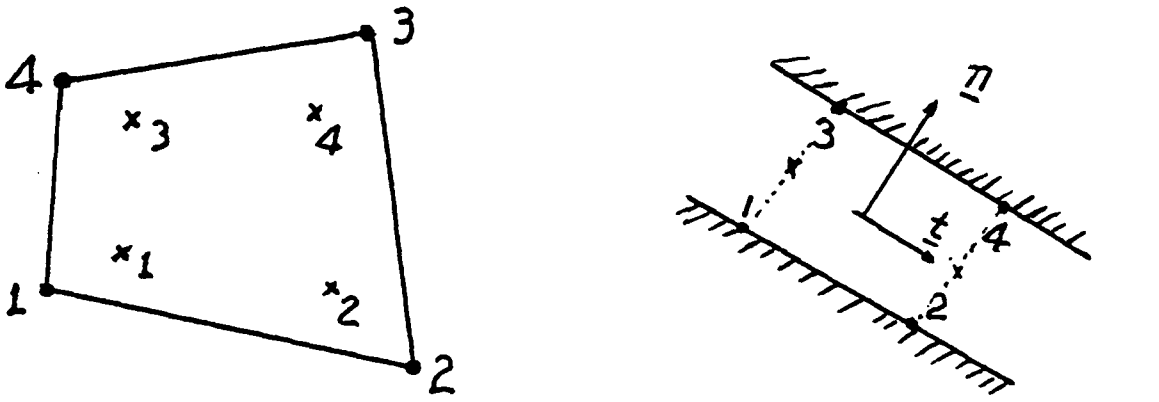


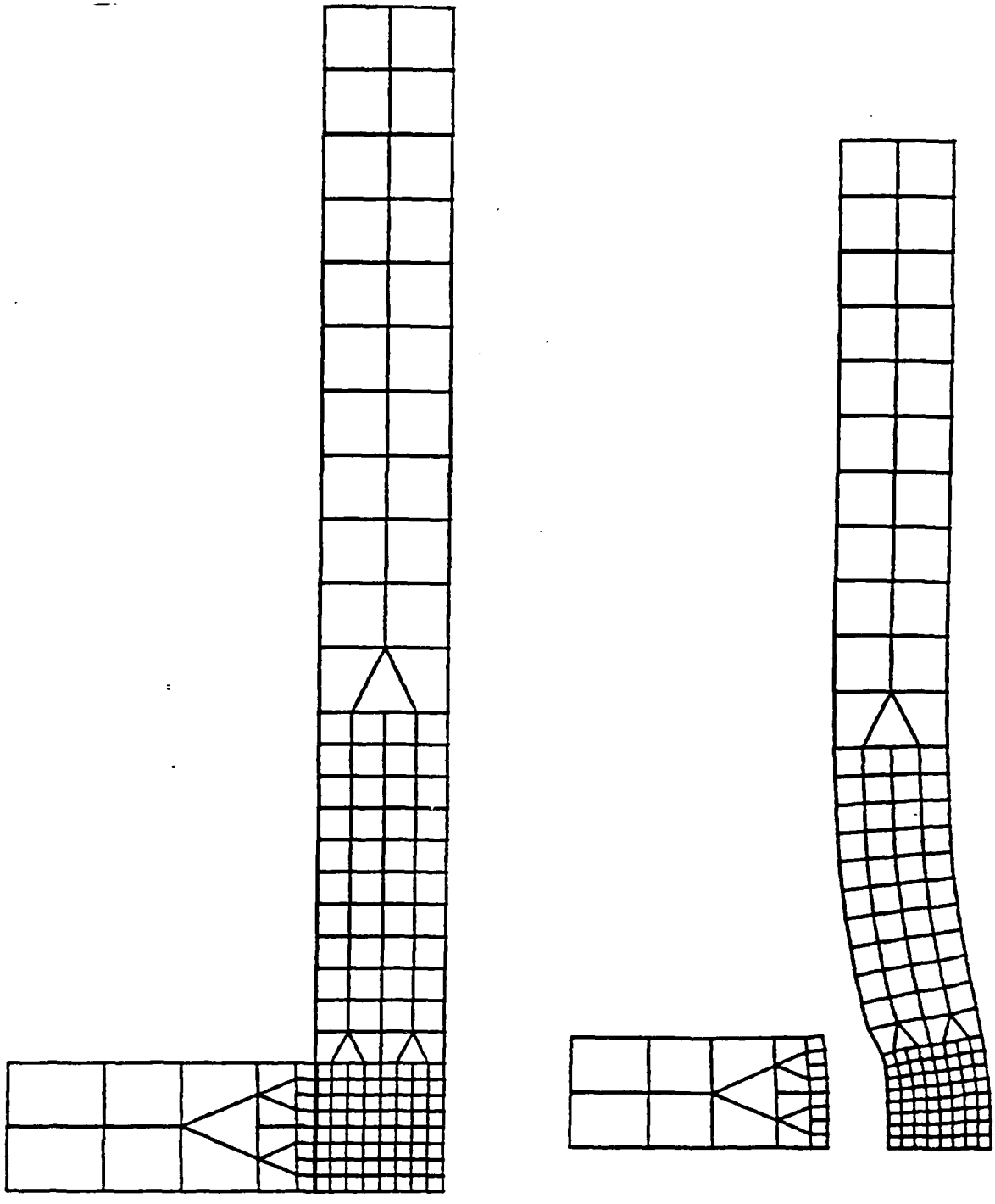
Figure 1. One-dimensional interference-fit assembly.



(a) solid element (CAX4)

(b) interface element (INTER2A)

Figure 2. Axisymmetric solid and interface elements.



(a) interference-fit (b) separation

Figure 3. A two-dimensional finite element model.

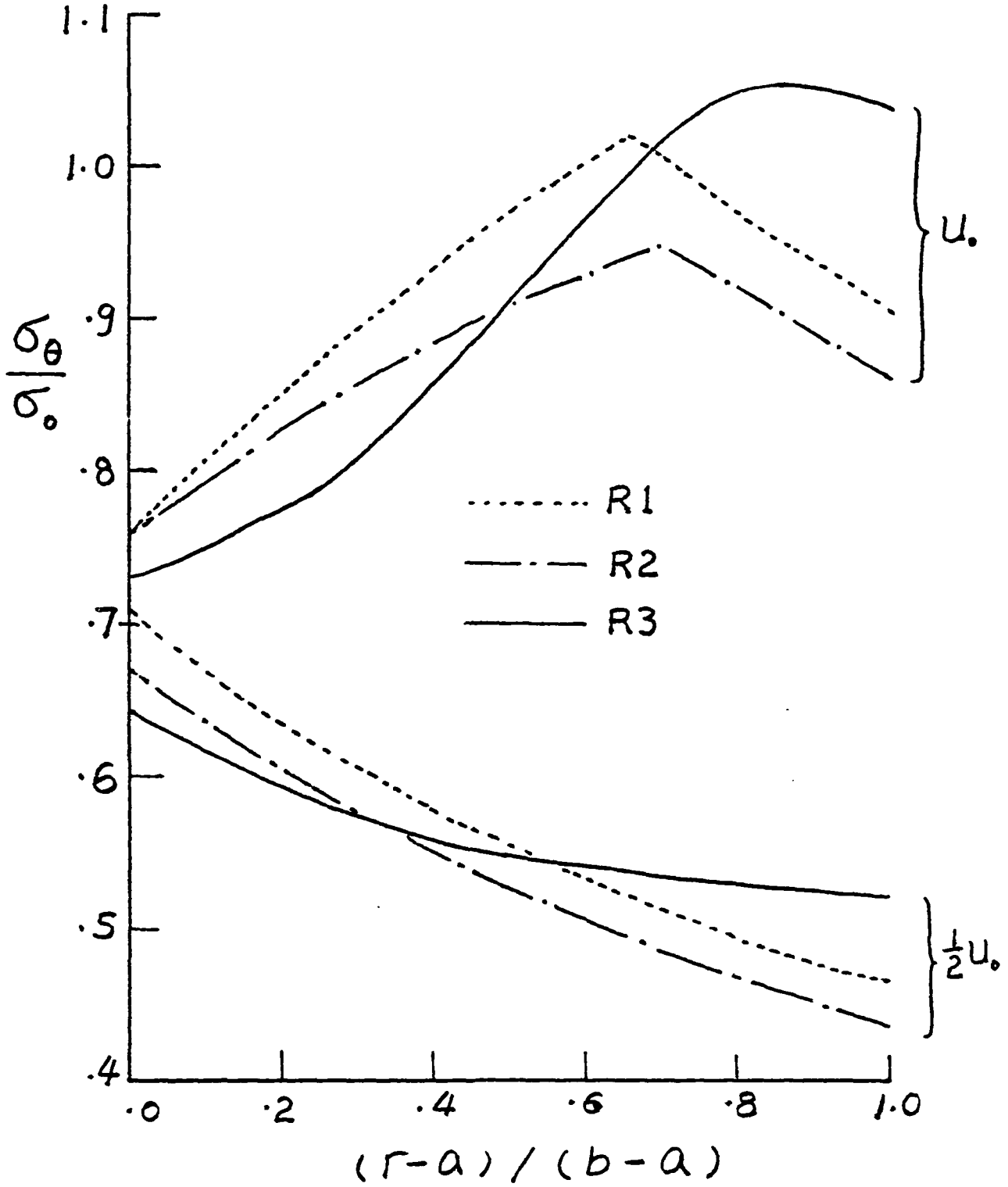


Figure 4. The effect of interference on the hoop stresses using rigid mandrels.

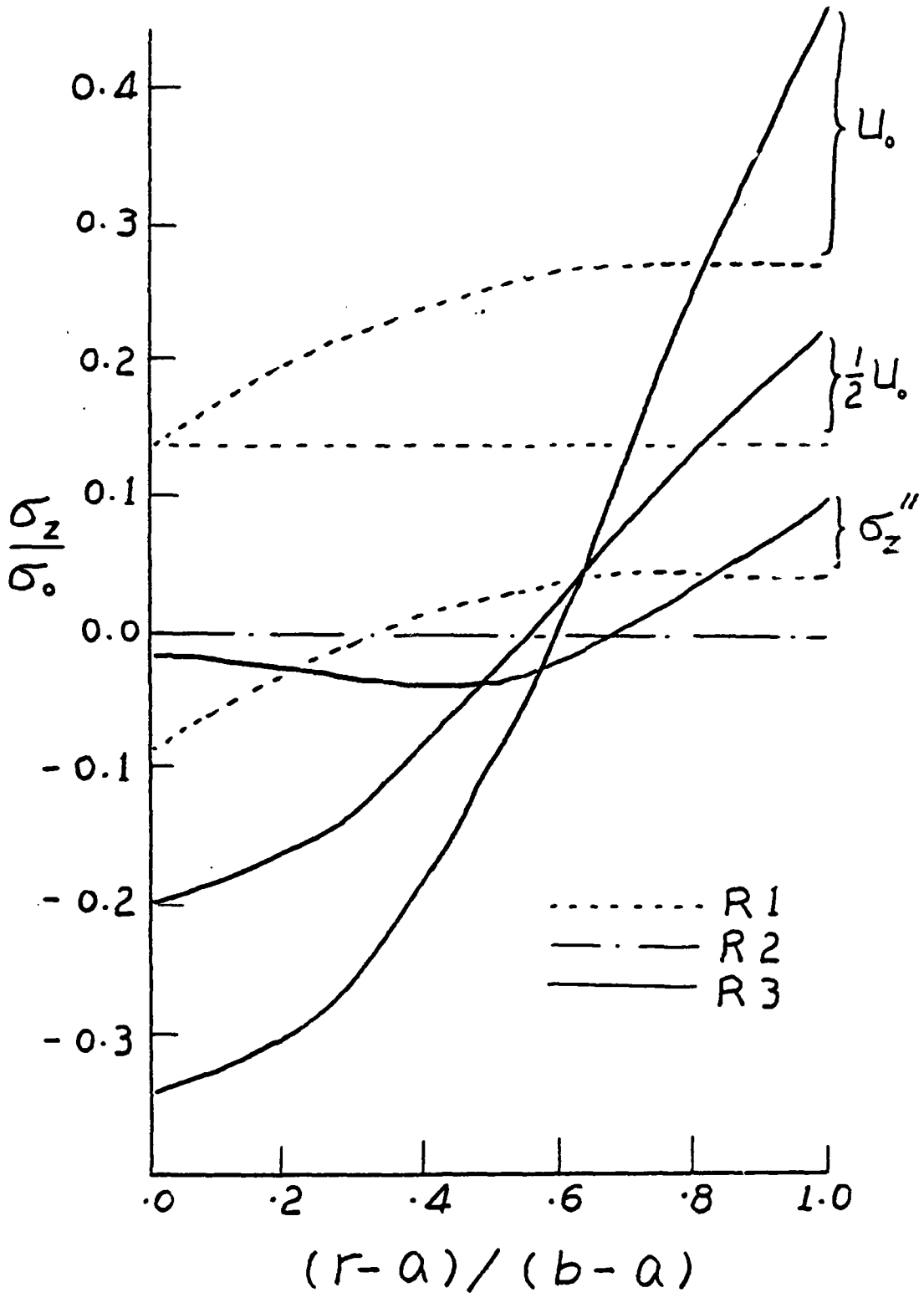


Figure 5. The effect of interference on the axial stresses using rigid mandrels.

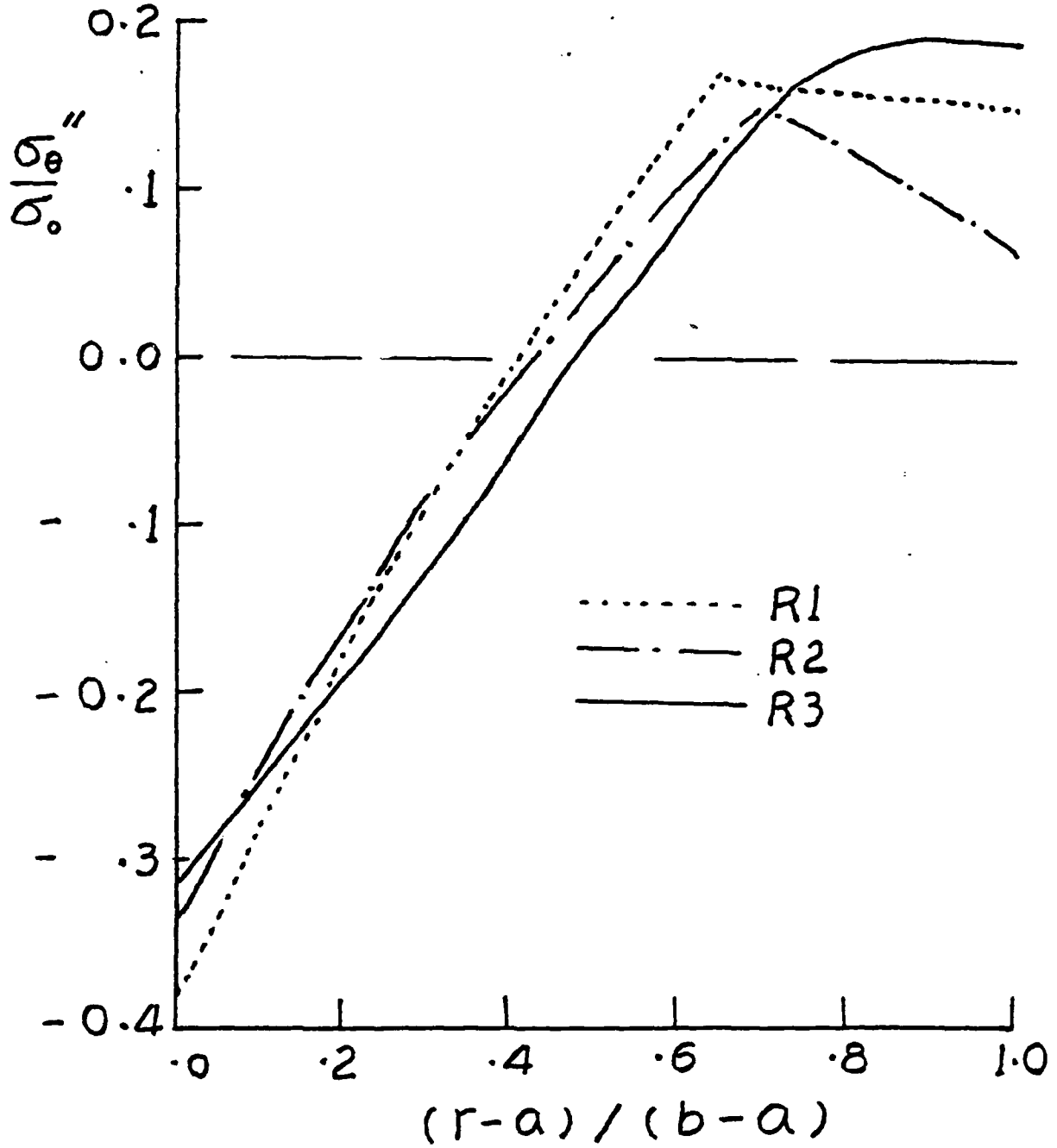


Figure 6. The effect of three rigid mandrels on the residual hoop stresses.

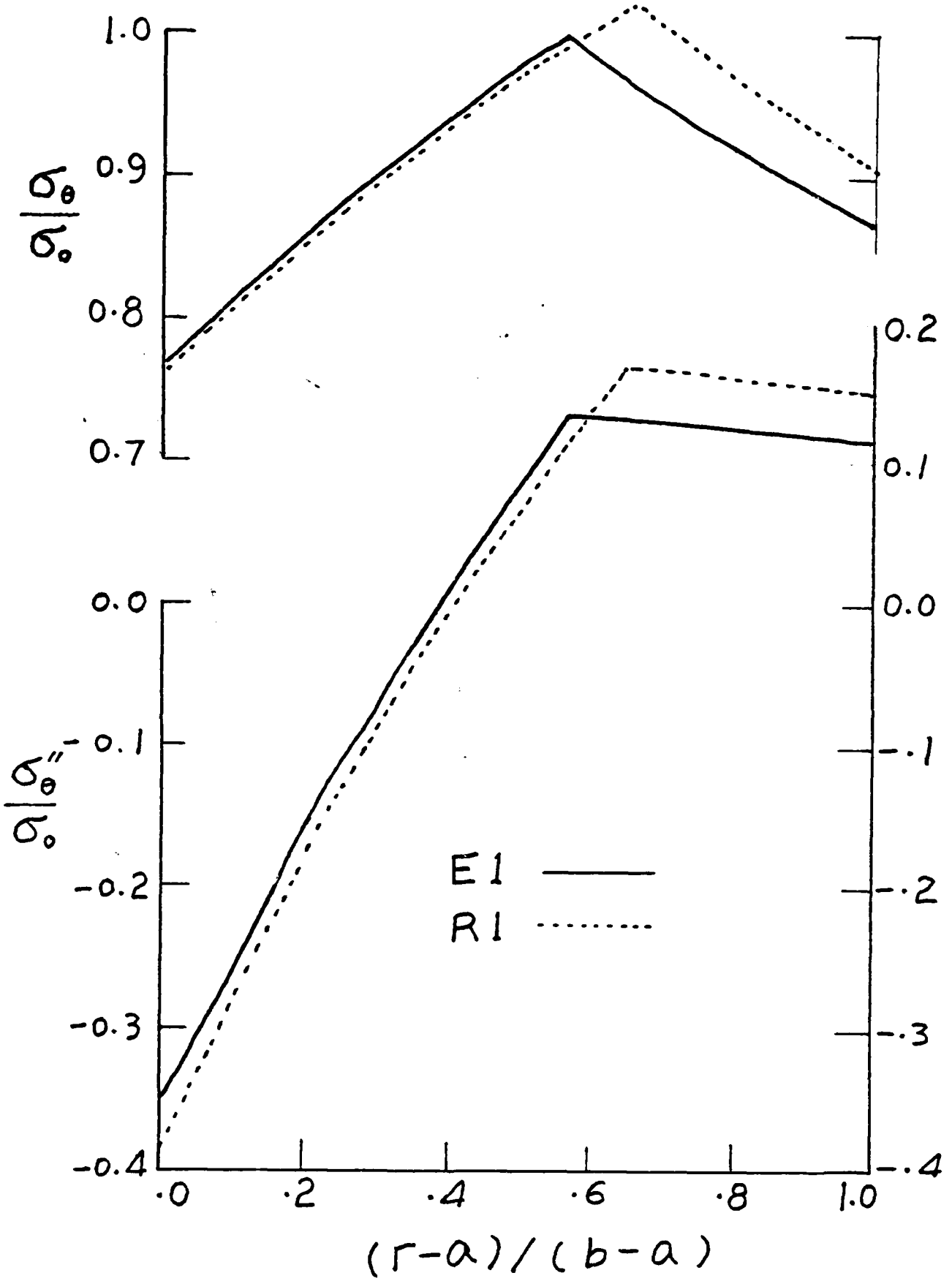


Figure 7. The effect of elasticity in the mandrel on the residual hoop stresses.

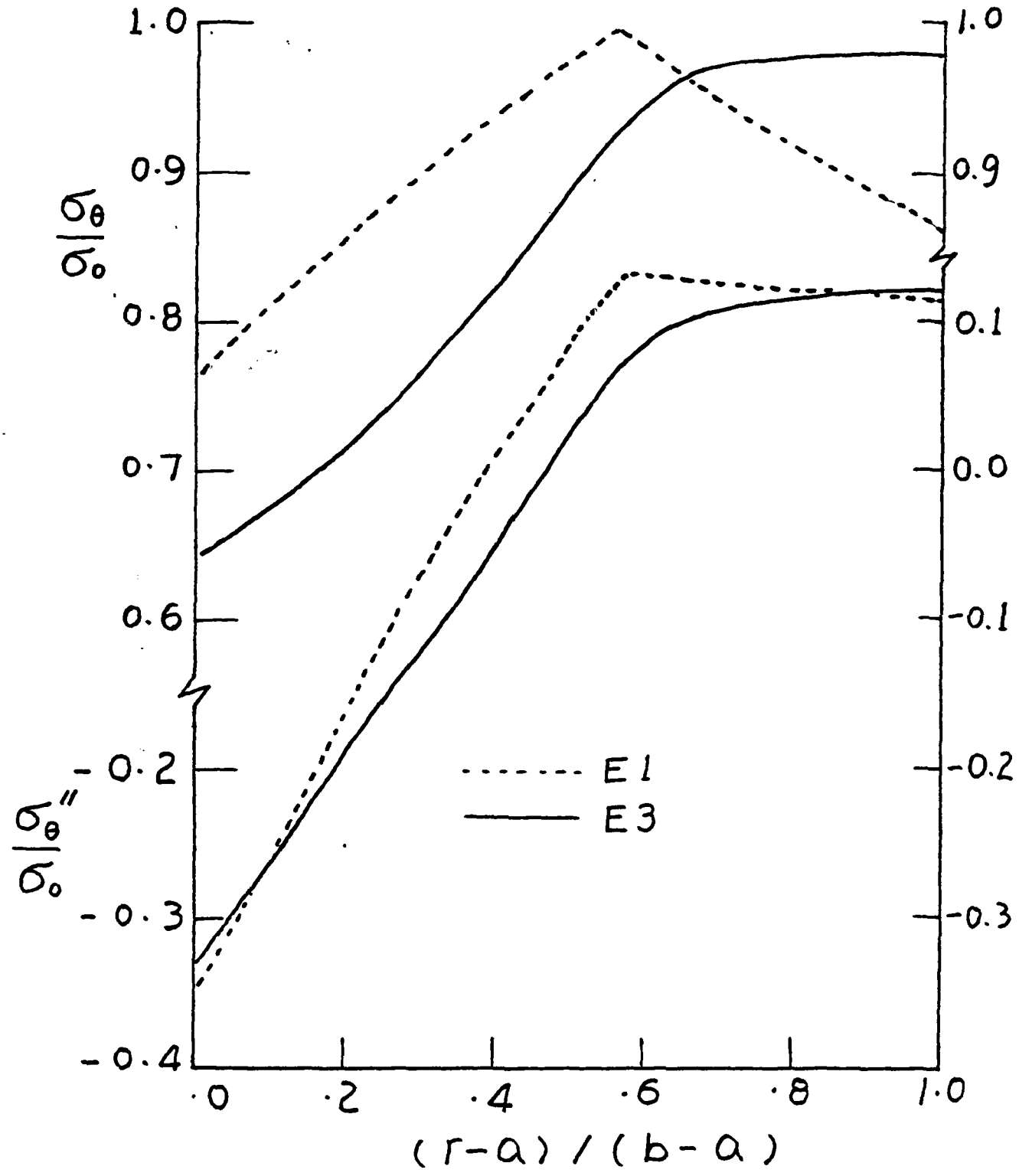


Figure 8. The two-dimensional effect of the mandrel on the residual hoop stresses.

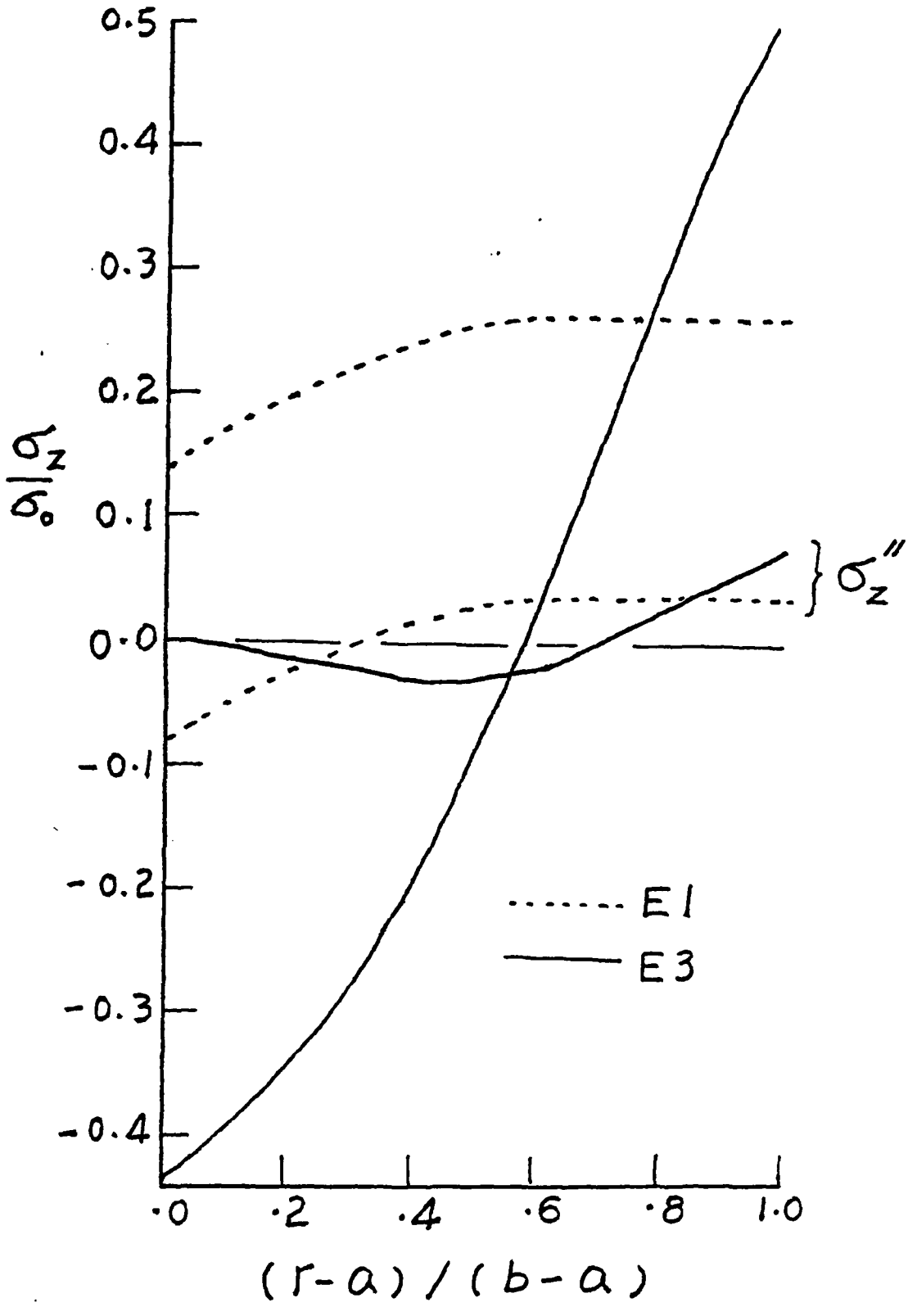


Figure 9. The two-dimensional effect of the mandrel on the axial stresses.

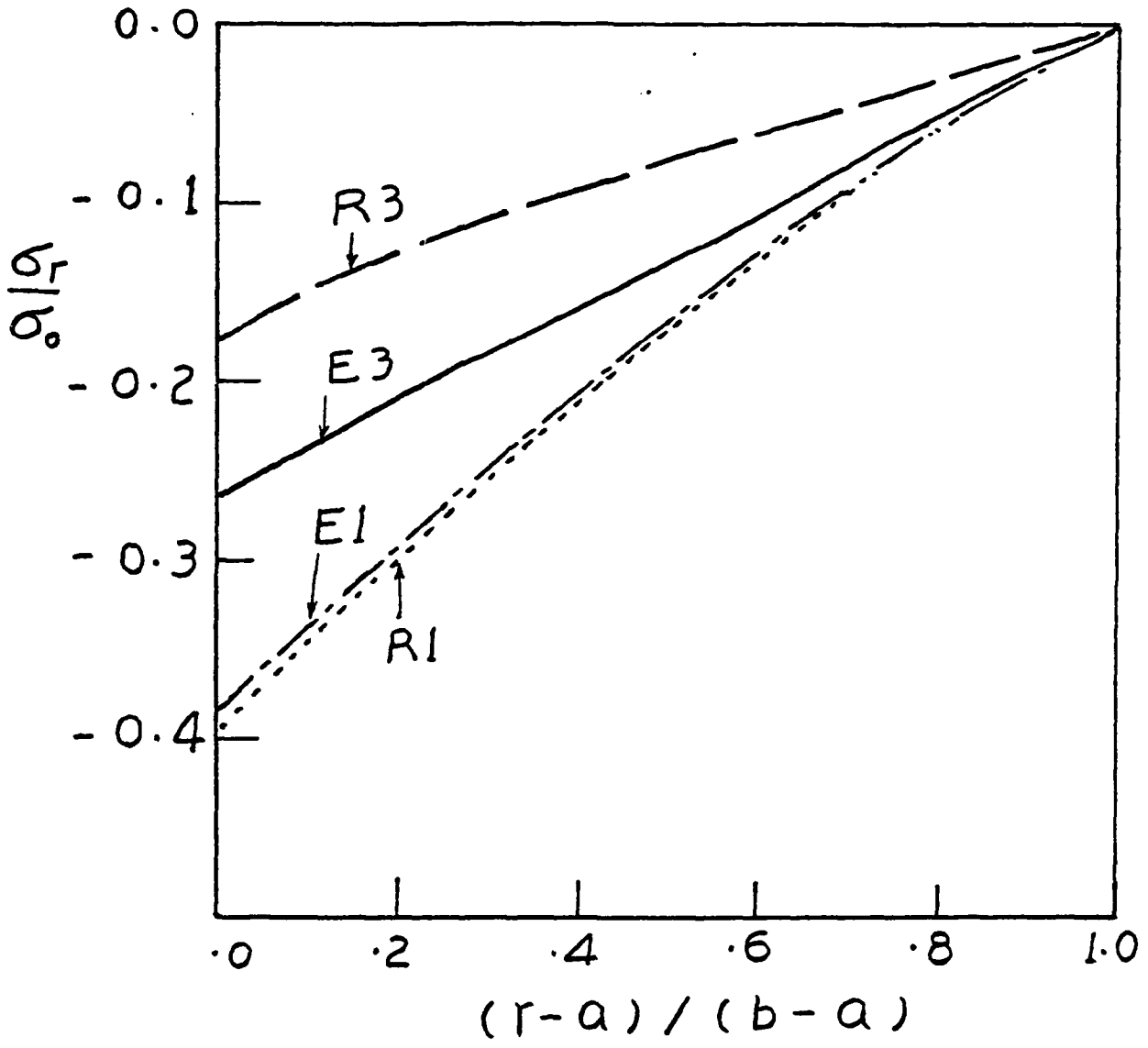


Figure 10. A comparison of radial stresses for four models.

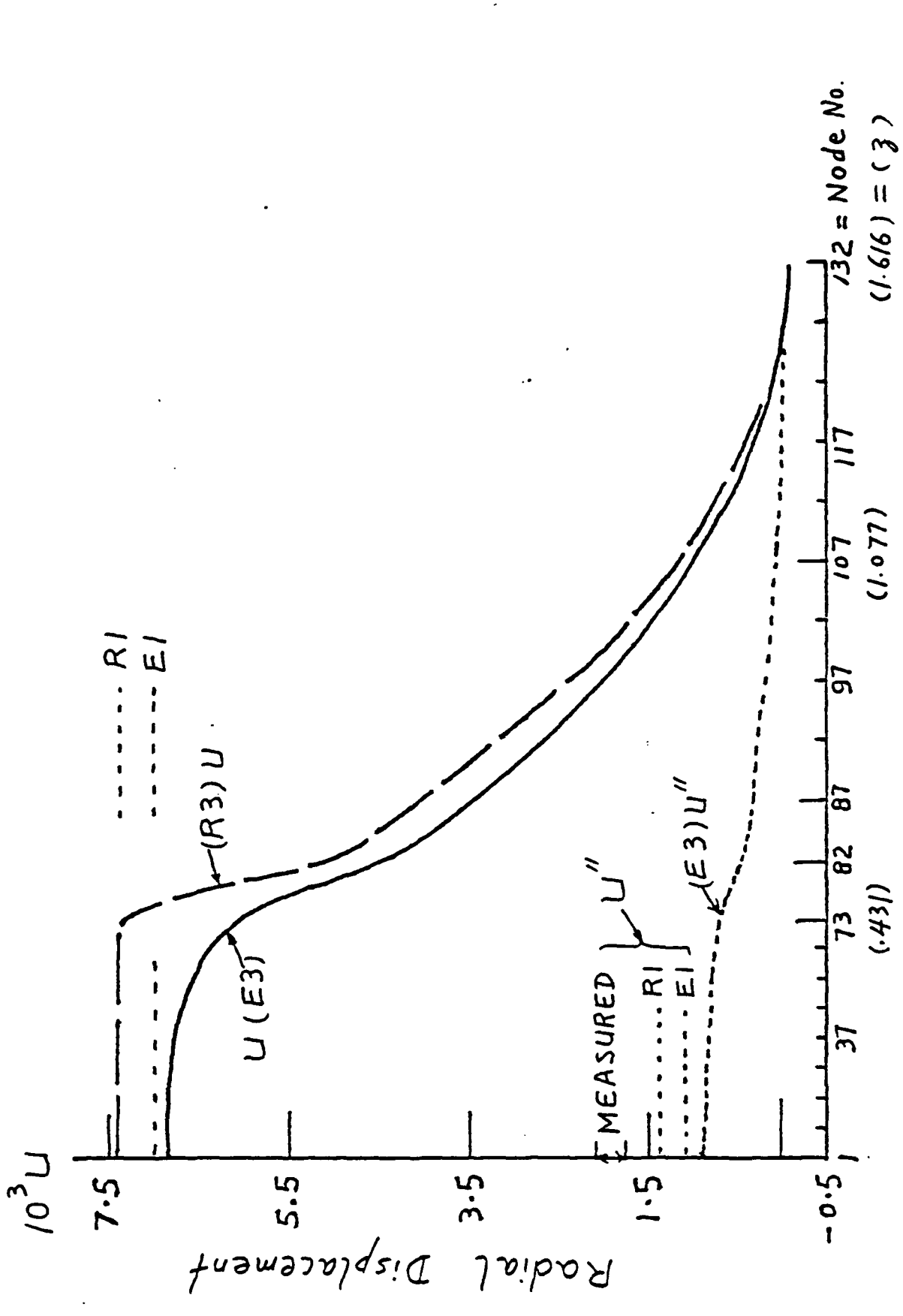


Figure 11. Radial displacements along the bore.

NONMONOTONIC STRESS-STRAIN LAWS: BIZARRE BEHAVIOR AND ITS REPERCUSSIONS ON NUMERICAL SOLUTIONS*

Ted Belytschko and David Lasry
Department of Mechanical Engineering
Northwestern University
Evanston, Illinois 60208

ABSTRACT

The properties of solutions with nonmonotonic stress-strain laws are described. Some particular properties are: severe mesh dependence, an apparent lack of convergence, and chaotic results for converging waves. These results are partially explained by examining a closed form solution for a simple problem. It shows that in a nonmononic continuum, the unstable dynamic response localizes to a set of measure zero.

To remedy this difficulty, localization limiters have been introduced to provide solutions where the deformation concentrates in regions of finite size. Several formulations of such limiters are discussed, with particular reference to stability and computational issues. Various applications are presented.

1. INTRODUCTION

Solutions of problems involving a strain-softening material law are fraught with serious difficulties, both from mathematical and numerical points of view. In dynamics, these difficulties were illustrated in Ref. [1] using a simple one-dimensional wave propagation model: an elastic wave propagating in a bar travels with a velocity proportional to $\sqrt{E_T}$, where E_T is the tangent modulus. When E_T becomes negative, as is the case for a strain-softening material, the wave cannot propagate anymore, giving rise to what Freund [2] calls a deformation-trapping phenomenon: the deformation is trapped in a certain zone of the body and no information can be transmitted to the rest of the material. Bazant and Belytschko [1] have shown by a closed form solution that strain-softening in transient problems is characterized by the appearance of infinite strains on a set of measure zero. This is reflected in numerical simulations by a strong dependency of the results upon the refinement of the mesh [3]. When the equations are discretized, by finite elements for example, the deformation will localize in the smallest discrete cell of material capable of representing that set of measure zero, namely one element in constant-strain elements in one-dimension or a one element-wide band in two dimensions. The problem of mesh dependency is not an intrinsically numerical one, but rather stems from the more fundamental loss of strict hyperbolicity of the equations of motion [4] upon attaining the strain-softening regime.

---

*Supported by the U.S. Army Research Office (Contract DAAL03-87-K-0035)

In statics, localization has been associated with the loss of ellipticity of the incremental equilibrium equations [5], and the existence of a bifurcation from a homogeneous state of deformation into a nonhomogeneous ones and the appearance of multiple equilibrium paths. This approach provides the orientation of the localization band and the critical load for which localization may be triggered but does not provide any length parameter for the subsequent behavior; in this respect, it is somewhat different from the dynamic case, where a localization zone (albeit reduced to a single point or line) appears in the closed form solutions [3]. This difference is reflected in the numerical simulations as well. If a solid with no imperfections is submitted to a homogeneous state of deformation, the numerical solution for a static problem will follow that homogeneous deformation path even when it becomes unstable beyond the bifurcation point, provided the machine precision is sufficient to prevent round-off error from triggering an inhomogeneous mode.

In order to circumvent these difficulties, the concept of localization limiters has been proposed in [3,4]. The essential idea of these limiters is to change the character of the equations so that the region of localization does not degenerate to a set of measure zero. The limiters proposed in [3] and [4] were respectively of two distinct types: integral limiters based on nonlocal constitutive equations and differential limiters based on higher order derivatives of the strain.

One purpose of this paper is to present analyses of the governing equations with and without limiters in one dimension and in the case of antiplane motion in two dimensions. It is shown that without limiters, the static equations lose ellipticity for strain softening materials and nonassociated plastic laws, while the dynamic equations lose strict hyperbolicity. With the gradient-type localization limiter, the dynamic equations change from hyperbolic to parabolic, which introduces a length scale.

It is also shown in this paper that the two types of localization limiters, differential and integral, possess very similar characteristics. Both limiters (1) exhibit a stable response to short wavelength input and and unstable response to long wavelength, and (2) limit the localization to a width dependent strictly on the length parameter. It is noted that even with the limiter the discrete tangent stiffness does not maintain positive definiteness and the numerical difficulties associated with strain-softening in local materials also appear.

The paper is organized as follows: Section 2 deals with the relationship between localization and change of type in the governing equations. Section 3 classifies the different localization limiters. In section 4, solutions are given for simple problems.

2. CONDITION FOR LOCALIZATION AND CHANGE OF TYPE

In order to better understand the difficulties associated with the localization phenomenon and the role of the gradient localization limiter, the relation between the onset of localization and a change of type of the governing equations is investigated here.

2.1 Change of type in statics and dynamics

The purpose of this section is to derive for a simple problem the condition for the onset of localization in statics and dynamics and relate it to the type of the system of PDE's governing the problem.

For the general three-dimensional case, the equations of motion are written:

$$\sigma_{ij,i} + b_j = \rho\, v_{j,t} \tag{2.1}$$

where $\boldsymbol{\sigma}$ is the Cauchy stress tensor, $\mathbf{v}$ the velocity vector, ρ the mass density and $\mathbf{b}$ the vector of body forces. Subscript indices preceded by comma denote partial derivatives. The body forces intervene in the governing equation only as a forcing term, so we can omit them in the study of the character of the equations.

The constitutive law relating the stress and strain rates is written:

$$\dot{\sigma}_{ij} = C_{ijkl}\, \dot{\varepsilon}_{kl} \tag{2.2}$$

where the tensor C_{ijkl} has minor symmetries $C_{ijkl} = C_{jikl} = C_{ijlk}$. $\boldsymbol{\varepsilon}$ is the strain tensor defined as:

$$\varepsilon_{kl} = \frac{1}{2}(u_{k,l} + u_{l,k}) \tag{2.3}$$

The equation will be analyzed for a simple *antiplane shear problem*, but the main results remain valid for the general three-dimensional case. The antiplane shear problem has been studied for a class of incompressible hyperelastic materials by Knowles [7] in statics and by Freund et al.[8] and Toulios [9] in dynamics. For this problem, the displacement and stress fields are as follows:

$$u_1 = u_2 = 0\ ,\ u_3 = u_3(x_1,x_2)\ ,\ \sigma_{13} = \sigma_{13}\,(x_1,x_2)\ ,\ \sigma_{23} = \sigma_{23}\,(x_1,x_2) \tag{2.4}$$

In statics, the equilibrium equations (2.1) then reduce to:

$$\sigma_{13,1} + \sigma_{23,2} = 0 \tag{2.5}$$

The constitutive law reads:

$$\begin{aligned} \dot{\sigma}_{13} &= 2C_{1313}\, \dot{\varepsilon}_{13} + 2C_{1323}\, \dot{\varepsilon}_{23} \\ \dot{\sigma}_{23} &= 2C_{2313}\, \dot{\varepsilon}_{13} + 2C_{2323}\, \dot{\varepsilon}_{23} \end{aligned} \tag{2.6}$$

We make the simplifying assumption that stress is a single-valued function of strain. This holds for elastic-plastic laws as long as there is no unloading at any point. We can write:

$$\sigma_{13,1} = \frac{\partial \sigma_{13}}{\partial \varepsilon_{13}} \frac{\partial \varepsilon_{13}}{\partial x_1} + \frac{\partial \sigma_{13}}{\partial \varepsilon_{23}} \frac{\partial \varepsilon_{23}}{\partial x_1} = 2C_{1313}\,\varepsilon_{13,1} + 2C_{1323}\,\varepsilon_{23,1} \tag{2.7}$$

and a similar relation for $\sigma_{23,2}$. We look for solutions that have discontinuities in $\varepsilon_{\alpha 3,\beta}$ along a line Γ defined by its local normal $\mathbf{n}$ $(n_1,n_2,0)$. The tangent vector to Γ at the current point is $\mathbf{s}$ $(s_1,s_2)=(-n_2,n_1,0)$.

The governing equations along Γ (equilibrium, compatibility, directional derivatives) can be cast in a matrix form:

$$\begin{bmatrix} C_{1313} & C_{1323} & C_{2313} & C_{2323} \\ 0 & 1 & -1 & 0 \\ s_1 & s_2 & 0 & 0 \\ 0 & 0 & s_1 & s_2 \end{bmatrix} \begin{bmatrix} \varepsilon_{13,1} \\ \varepsilon_{13,2} \\ \varepsilon_{23,1} \\ \varepsilon_{23,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \varepsilon_{13,s} \\ \varepsilon_{23,s} \end{bmatrix} \tag{2.8}$$

This relation is of the form $\mathbf{A}\,\boldsymbol{\varepsilon}^D = \mathbf{c}$; in order for $\boldsymbol{\varepsilon}^D$ not to be unique, we must require

$$\det \mathbf{A} = 0 \tag{2.9}$$

which yields in this case:

$$-s_1^2 C_{2323} + s_1 s_2 (C_{1323} + C_{2313}) - s_2^2 C_{1313} = 0 \tag{2.10}$$

or in terms of the normal vector $\mathbf{n}$:

$$n_2^2 C_{2323} + n_1 n_2 (C_{1323} + C_{2313}) + n_1^2 C_{1313} = 0 \tag{2.11}$$

The above can be written

$$\det (n_i\, C_{ijkl}\, n_l) = \det (\mathbf{n}\, \mathbf{C}\, \mathbf{n}) = 0 \tag{2.12}$$

which is the classical localization condition [5,10]. The loss of uniqueness corresponds to the loss of ellipticity of the governing equations, or in other words, to the appearance of real characteristics which are associated with equations of a hyperbolic type.

We focus now on the **dynamic** case. The equation of motion for the antiplane problem is:

$$\sigma_{13,1} + \sigma_{23,2} = \rho\, v_{3,t} \tag{2.13}$$

The cross-derivative relations:

$$\varepsilon_{13,t} = v_{3,1} \quad , \quad \varepsilon_{23,t} = v_{3,2} \tag{2.14}$$

are combined with the equation of motion to yield a system of first order PDE's:

$$\begin{bmatrix} 0 & C_{1313} & C_{1323} \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_3 \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}_{,1} + \begin{bmatrix} 0 & C_{2313} & C_{2323} \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_3 \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}_{,2} + \begin{bmatrix} -\rho & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_3 \\ \varepsilon_{13} \\ \varepsilon_{23} \end{bmatrix}_{,t} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{2.15}$$

This system is of the form:

$$\mathbf{A}^1 \mathbf{U}_{,1} + \mathbf{A}^2 \mathbf{U}_{,2} + \mathbf{A}^t \mathbf{U}_{,t} = \mathbf{0} \tag{2.16}$$

where $\mathbf{A}^t$ is nonsingular. The condition for $\Phi(x_1,x_2,t) = 0$ to be a characteristic surface of (2.15) is [11]:

$$\det(\mathbf{A}) = 0 \tag{2.17a}$$

where

$$\mathbf{A} = \mathbf{A}^1 \Phi_{,1} + \mathbf{A}^2 \Phi_{,2} + \mathbf{A}^t \Phi_{,t} \tag{2.17b}$$

which yields here:

$$\Phi_{,t}(-\rho\Phi_{,t}^2 + (C_{1323} + C_{2313})\,\Phi_{,1}\Phi_{,2} + C_{1313}\Phi_{,1}^2 + C_{2323}\Phi_{,2}^2) = 0 \tag{2.18}$$

The extra factor $\Phi_{,t}$ in (2.18) corresponds to a characteristic surface with zero velocity and is a result of introducing an additional dependent variable by choosing strains and velocity as the dependent variables [12]. In order to better understand the meaning of (2.18), we define the constitutive matrix **D** such that:

$$D_{kl} = C_{k3l3} \tag{2.19}$$

and select a new coordinate system ($\tilde{1}$, $\tilde{2}$) defined by the principal directions of **D**, so that in the new coordinate system:

$$\mathbf{D} = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \tag{2.20}$$

Equation (2.18) can thus be written as:

$$-\rho\Phi_{,t}^2 + D_1\Phi_{,1}^2 + D_2\Phi_{,2}^2 = 0 \tag{2.21}$$

Characteristic surfaces are cones of elliptic section, as illustrated in figure 1; the equation of the cone passing through a point ($\tilde{x}_o$, $\tilde{y}_o$, t_o) is:

$$(t - t_o)^2 = \frac{1}{c_1^2}(\bar{x} - \bar{x}_o)^2 + \frac{1}{c_2^2}(\bar{y} - \bar{y}_o)^2 \tag{2.22}$$

where

$$c_1 = \sqrt{\frac{D_1}{\rho}} \quad \text{and} \quad c_2 = \sqrt{\frac{D_2}{\rho}} \tag{2.23}$$

As **D** loses positive definiteness, say for example D_1 remains strictly positive and D_2 approaches zero, the cone collapses to a plane surface. Considered as a function of the variables ($\bar{y}$, t), the system loses strict hyperbolicity, or equivalently real waves no longer propagate in every direction (in our case they stop propagating in the $\bar{y}$ direction). It is therefore seen that here the condition of strict hyperbolicity of the system of governing equations and the condition of strong ellipticity are equivalent.

It should be pointed out that when a **viscoplastic** constitutive law is used, the equations of motion do not lose hyperbolicity. This is readily seen by observing that for viscoplastic models the rate constitutive relation is written as:

$$\dot{\sigma}_{ij} = C^e_{ijkl}\,\dot{\varepsilon}_{kl} - R_{ij}(\boldsymbol{\sigma}) \tag{2.24}$$

where C^e_{ijkl} is the elastic tensor and the inelastic part is embedded in the term R_{ij}. The type of the system of governing equations is determined by C^e_{ijkl}, so that it remains strictly hyperbolic, the inelastic effects appearing only as a forcing term.

2.2 Relation between strain softening and localization for elasto-plastic materials

The rate constitutive relations for an elasto-plastic material are written in tensor form as:

$$\dot{\boldsymbol{\sigma}} = \mathbf{C} : \dot{\boldsymbol{\varepsilon}} = \mathbf{C}^e : \dot{\boldsymbol{\varepsilon}} - \frac{\mathbf{C}^e : \mathbf{P}}{h + \mathbf{Q} : \mathbf{C}^e : \mathbf{P}}\ \mathbf{Q} : \mathbf{C}^e : \dot{\boldsymbol{\varepsilon}} \tag{2.25}$$

where **P** and **Q** are symmetric first order tensors giving respectively the direction of the plastic deformation and the outer normal to the yield surface, h is the rate of hardening, and $\mathbf{C}^e$ the elasticity tensor:

$$C^e_{ijkl} = \lambda\,\delta_{ij}\,\delta_{kl} + G\,(\,\delta_{ik}\delta_{jl} + \delta_{il}\,\delta_{jk}\,) \tag{2.26}$$

and λ and G are Lamé's constants. The case $\mathbf{P} = \mathbf{Q}$ corresponds to plastic normality, $\mathbf{P} \neq \mathbf{Q}$ corresponds to a non-associative flow rule.

We consider again the antiplane shear problem and focus on the relation between the localization condition and the strain-hardening modulus in the case of an elasto-plastic material model. For localization to occur on a plane of normal **n**, the condition (2.12) has to be met. We can write that condition in a set of cartesian axes $\mathbf{n}$, $\mathbf{e}_3 \times \mathbf{n}$ and $\mathbf{e}_3$ ($\mathbf{e}_3$ is the

unit vector in the "3" direction). With subscripts denoting components in that set of axes, and for the constitutive law (2.25)-(2.26), the localization condition reduces to:

$$C_{1313} = 2\left(G - \frac{4G^2 P_{13}}{h + 4 P_{13} 2G Q_{13}} Q_{13}\right) = 0 \tag{2.27}$$

or equivalently

$$\frac{h}{G} = -2 (2 P_{13} Q_{13}) \tag{2.28}$$

This expression shows that if plastic normality holds (i.e. $\mathbf{P} = \mathbf{Q}$), then localization can only occur with negative h, that is in a strain-softening regime, whereas if normality does not apply, it is possible for localization to be triggered with a positive h.

This result, obtained for the particular case of the antiplane shear problem, is in fact general as shown by Rudnicki and Rice [13] and Rice [5]. In the three-dimensional case, (2.28) can be generalized to (with α, β denoting components on cartesian axes in the plane of localization):

$$\frac{h}{G} = -2 P_{\alpha\beta} Q_{\alpha\beta} - \frac{2\lambda}{\lambda+2G} P_{\alpha\alpha} Q_{\beta\beta} \tag{2.29}$$

and the conclusions derived previously remain valid. An example of a material model where localization occurs for positive h can be found in [13].

3. LOCALIZATION LIMITERS

Localization limiters can be classified as follows:

1. nonlocal or integral limiters where the strain measure includes an integral of the deformation over a finite domain [3].
2. differential limiters where the strain or stress measures include derivatives of order higher than one [4,14-17].
3. rate limiters, where a time dependence is built into the equations[18].

The rationale underlying the nonlocal limiters is that a classical local theory does not take into account the influence of the length scale associated with a rapidly varying strain field on the stress distribution, an essential part of the localization phenomena.

In the case of a one-dimensional rod with strain-softening, a nonlocal limiter is obtained by defining the stress field $\sigma(x)$ as a function of a nonlocal strain $\bar{\varepsilon}(x)$ [3]:

$$\sigma(x) = \sigma(\bar{\varepsilon}(x)) \tag{3.1}$$

with

$$\bar{\varepsilon}(x) = \frac{1}{\ell} \int_{-\ell/2}^{\ell/2} \varepsilon(x+s) w(s)\, ds \tag{3.2}$$

where $\left[x - \frac{\ell}{2}, x + \frac{\ell}{2} \right]$ is a domain around x, and w(s) a weighting function. For the sake of simplicity, we will assume in what follows a uniform weighting function $w(s) = \frac{1}{\ell}$.

The gradient-type limiter in a one dimensional context is given by [4]:

$$\hat{\varepsilon}(x) = \varepsilon(x) + \alpha \varepsilon_{,xx}(x) \tag{3.3}$$

These two limiters are related through a Taylor expansion [4] and actually differ by a function of order $o(\ell^2)$, provided that:

$$\alpha = \frac{\ell^2}{24} \tag{3.4}$$

In dynamic problems, the effect of the differential and rate limiters from a mathematical point of view is that the governing equations no longer become elliptic with the onset of strain-softening. This can be seen in a one-dimensional context for a path independent material, by combining the equation of motion and the compatibility condition into a system of first order partial differential equations:

$$\begin{bmatrix} v \\ \varepsilon \end{bmatrix}_{,t} + \begin{bmatrix} 0 & -\frac{\sigma'(\varepsilon)}{\rho} \\ -1 & 0 \end{bmatrix} \begin{bmatrix} v \\ \varepsilon \end{bmatrix}_{,x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{3.5}$$

where v, $\sigma(\varepsilon)$, ε and ρ are respectively the particle velocity, stress, strain and density, and subscript comma denotes a partial derivative. This system is of the type

$$\mathbf{A}\,\mathbf{U}_{,t} + \mathbf{B}\,\mathbf{U}_{,x} = \mathbf{c}(\mathbf{U}) \tag{3.6}$$

where one of the matrices **A** or **B**, e.g. **A**, is nonsingular and **c** is a forcing vector. The nature of (3.6) is determined by the roots of the characteristic determinant, $\det(\mathbf{B} - \lambda\mathbf{A})$ (or $\det(\mathbf{A} - \lambda\mathbf{B})$ if **A** is singular).In the case of Eq. (3.5)

$$\det(\mathbf{B} - \lambda\mathbf{A}) = \det(\mathbf{B} - \lambda\mathbf{I}) = \lambda^2 - \frac{\sigma'(\varepsilon)}{\rho} \tag{3.7}$$

so that the system becomes elliptic when $\sigma'(\varepsilon)<0$ (which corresponds to strain-softening) because the determinant (3.7) does not possess real roots anymore.

When the differential limiter defined in (3.3) is included in the formulation, a modified system of P.D.E's is obtained:

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \varepsilon \\ w_1 \\ w_2 \\ v \end{bmatrix}_{,t} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ \frac{\sigma'(\varepsilon)}{\rho} & 0 & \frac{\alpha\sigma'(\varepsilon)}{\rho} & 0 \end{bmatrix} \begin{bmatrix} \varepsilon \\ w_1 \\ w_2 \\ v \end{bmatrix}_{,x} = \begin{bmatrix} w_1 \\ 0 \\ w_2 \\ 0 \end{bmatrix} \tag{3.8}$$

where $w_1 = \varepsilon_{,x}$ and $w_2 = \varepsilon_{,xx}$. The characteristic determinant

$$\det (\mathbf{A} - \lambda \mathbf{B}) = -\frac{\alpha\sigma'(\varepsilon)}{\rho}\lambda^4 \tag{3.9}$$

possesses four real roots, all equal, irrespective of the sign of $\sigma'(\varepsilon)$, so that the system is parabolic. It should be pointed out that, when a rate-type limiter is used via a viscoplastic material model, the governing equations remain hyperbolic[18].

In order to better understand the behavior of the integral and differential limiters, a Fourier analysis by the method of frozen coefficients is useful. In this analysis a displacement disturbance δu is applied to the body, and the material is considered to be in a strain-softening state over an interval $[x_1,x_2]$:

$$\delta\sigma(x) = -|E_t| \ \delta\varepsilon(x) \qquad \text{for } x \text{ in } [x_1,x_2] \tag{3.10}$$

where the tangent modulus $E_t < 0$ is assumed constant. We then look for possible wave solutions of the form

$$\delta u(x,t)=A\, e^{ik(x-vt)} \tag{3.11}$$

for the equation of motion for δu:

$$\delta u_{,tt} + \frac{|E_t|}{\rho} \ \delta\varepsilon^{\text{nonloc.}}(x)_{,x} = 0 \tag{3.12}$$

The following dispersion relation was obtained in [4] with the gradient-type limiter $\hat{\varepsilon}$ defined by Eqn. (3.3):

$$kv = i \left\{ \frac{|E_t|}{\rho}(1- \alpha k^2) \right\}^{1/2} k = i\, \hat{\gamma}(k) \tag{3.13}$$

A similar analysis can be done for the integral limiter $\varepsilon^{\text{nonloc.}}(x) = \bar{\varepsilon}(x)$ defined in Eqn. (3.2). Looking for wave solutions of the equation of motion

$$\delta u_{,tt} + \frac{|E_t|}{\rho}\,\frac{1}{\ell}\left(\int_{-\ell/2}^{\ell/2} \delta\varepsilon(x+s)ds\right)_{,x} = 0 \qquad (3.14)$$

one obtains the following relation:

$$kv = i\left\{\frac{|E_t|}{\rho}\,\frac{2k}{\ell}\,\sin\frac{k\ell}{2}\right\}^{1/2} k = i\,\bar{\gamma}(k) \qquad (3.15)$$

The two functions $\hat{\gamma}(x)$ and $\bar{\gamma}(k)$ are plotted in fig. 2, for values of α and ℓ related through eqn. (3.4). In [4] the plot of $\bar{\gamma}(k)$ was interpreted to mean that the growth in short wavelength inputs susceptible to develop in the narrow localization zones is bounded when the limiter is present.

It is interesting to notice that for small values of α, the expression of $\bar{\gamma}(k)$ can be expanded, and using (3.4):

$$\bar{\gamma}(k) \approx \left\{\frac{|E_t|}{\rho}\,\frac{2k}{\ell}\left[\frac{k\ell}{2} - \frac{1}{3!}\left(\frac{k\ell}{2}\right)^3\right]\right\}^{1/2} \approx \left\{\frac{|E_t|}{\rho}\,\frac{k}{\sqrt{6\alpha}}\,k\sqrt{6\alpha}\,(1-\alpha k^2)\right\}^{1/2} = \hat{\gamma}(k) \qquad (3.16)$$

A perturbation analysis [4] reveals that, when the differential limiter defined in (3.3) is used, the width of the zone in the strain-softening regime varies with the square root of the parameter α. Numerical simulations confirmed this type of dependence, and for the integral limiter they yield a zone size proportional to the averaging length ℓ [3] as can be expected from relation (3.4). Thus both localization limiters prevent the growth of waves of the scale of the localization bands which are generated by the presence of strain-softening.

As far as static problems are concerned, the only attempts to derive closed form solutions using limiters known to the author are due to Aifantis and co-workers [16,17] Coleman and Hodgdon [15] and Schreyer and Chen [14]. In the former approach, higher order terms are included in the evolution equation of the flow stress, and in that sense it is quite similar to the work of Schreyer and Chen. In Coleman and Hodgdon [15], a second order strain gradient is added directly into the constitutive equation without modifying the yield function. The common denominator to all of these approaches is that they make the stress field dependent in some way on the spatial derivatives of the strain field. We follow the approach of Coleman and Hodgdon [15], but do not limit our formulation to rigid plastic materials. The expression for the stress is given by:

$$\sigma = \phi(\varepsilon) - \alpha\,\nabla^2\varepsilon \qquad (3.17)$$

where $\phi(\varepsilon)$ is the usual elastoplastic constitutive law(stress-strain relationship) and $\alpha > 0$ is a coefficient having the dimensions of a force. The stiffness matrix corresponding to the finite element formulation of (3.17) is developed in [21].

4. NUMERICAL EXAMPLES

In this section numerical solutions are presented with and without limiters for simple problems. Dynamic situations, where no stiffness matrix has to be constructed if explicit time integration is used are first considered; static problems are considered in subsection 4.3.

4.1 Wave propagation in a rod

This problem was considered in [3], see fig. 3a. Equal and opposite velocities v_0 are applied to the two ends of a rod of length 2L made of a strain-softening material, so that tensile waves are generated at the ends. The magnitude of the strain is slightly less than the strain corresponding to the onset of strain-softening. These tensile waves propagate elastically to the center; when they meet at the center, the stress would double if the behavior remained elastic, so that strain-softening starts at this midpoint.

The analytical solution for this problem was proposed in [1]: localization occurs at the midpoint where the strain becomes infinite. The solution, symmetric about the midpoint x=L, is expressed for the left half as:

$$\varepsilon = \frac{v_0}{c_0} \left[H(t - \frac{x}{c_0}) - H(t - \frac{2L-x}{c_0}) + 4 \langle c_0 t - L \rangle \delta(x-L) \right] \tag{4.1}$$

where H is the Heaviside step function, $\langle A \rangle = A$ if $A > 0$, $A = 0$ otherwise, δ is the Dirac-delta function, c_0 the elastic wave speed in the material. Numerical studies of this problem based on nonlocal approaches were conducted in [3]. Here we will use the localization limiter

$$\hat{\varepsilon} = \varepsilon + \alpha \varepsilon_{,xx} \tag{4.2}$$

The development of a finite element formulation corresponding to (4.2) can be found in [4]. Particular provisions are made to avoid zero energy modes, and stiffness proportional damping is added in order to prevent oscillations ahead of the wave front, see[4].

The stress-strain law considered for the calculations is illustrated in the enclosed box in fig. 4. Other parameters used in the calculations were: density $\rho=1.$, end velocity $v_0=0.6$, and for the stress-strain relation, $E = 1.$, yield stress $\sigma_p = 1.$, $E_t = -.25$, $\varepsilon_f = 5.$, nearly horizontal tail of slope $E_f = .001$ beyond ε_f.

It was first checked (see fig. 4) that, without introducing the localization limiter (that is, for $\alpha=0$), the strain profiles are severely dependent on the mesh refinement, and the localization zone shrinks to one element, irrespective of its size. Furthermore, the total energy dissipated in the mesh tends to a zero value as the mesh is refined, as seen in fig. 6a. Convergence studies were then performed with the localization limiter defined in (4.2), for a value $\alpha = .1667$, for different meshes with increasing number of elements (fig. 5). They exhibit a localization limited to a finite size zone, the length of that zone and the strain profiles being independent of the mesh refinement. Moreover, the total energy dissipated in the rod is independent of the mesh size, all other parameters remaining equal, as illustrated in fig. 6a.

Calculations were also conducted at fixed mesh size for different values of α, see fig. 6b. showing that the length of the localization zone is *linearly dependent on* $\sqrt{\alpha}$. This is consistent with the results of [3], that found a linear dependence in ℓ (averaging length), since a Taylor expansion yielded the linear relation (3.4) between α and ℓ^2.

4.2 Spherically symmetric problem

This problem (see fig. 3b) was considered with strain-softening materials in [20]. A sphere made of a strain-softening material is loaded with a uniform traction on its exterior surface. To better appreciate the complexity of this problem, consider the load to be a ramp function in time. Before the onset of strain-softening at an interior surface S, a portion of the stress will have passed through S. Due to the spherical geometry, the stresses in this wave are amplified as they pass to the center and trigger the formation of additional strain-softening surfaces. As conjectured in [20], it seems that an infinite number of localization surfaces will appear, although no analytical solution has been proposed so far.

The localization limiter defined previously in (4.2) was used to solve numerically this problem. We considered a sudden application of a uniform normal traction $\sigma_r = p_0 H(t)$ at the exterior surface R_2=100; the interior surface is R_1=10. The applied surface pressure was chosen as p_0=.708; for this boundary conditions, the wave propagating from the outer surface remains elastic until the wavefront reaches $0.7R_2$. The same material constants as in section 4.1 were considered.

It was first noticed (see fig. 7) that without the localization limiter (that is for $\alpha = 0$), as the number of elements is increased, several points of localization develop, and these points change arbitrarily with mesh refinement, even in the presence of damping. These points of localization can be appreciated both in the volumetric strain plots, with the presence of spikes, and in the radial displacements plots, where sharp discontinuities indicate separation along a surface.

The next group of solutions (figs. 8) examines the effect of the localization limiter. These solutions converge well with mesh refinement, and furthermore, they are very similar to those found with the imbricate elements approach [20].

4.3 Static problems

When conducting the numerical simulations for static problems, where a stiffness matrix has to be developped, it was noticed that the introduction of the localization limiter did not remove completely all the unpleasant features present in calculations involving strain-softening materials. More precisely, when the strain-softening regime is incipient, the Newton-Raphson procedure often results in iterations that oscillate between two or more states and fail to converge to one equilibrium state. From a numerical point of view, this is linked to the tangent stiffness $\mathbf{K}_T$ does not remain positive definite. In [19], a remedy for this difficulty was proposed; it consists of posing the problem as the minimization of the length of the residual vector:

$$\text{Minimize: } \mathcal{F} = \mathbf{r}^T(\mathbf{d})\,\mathbf{r}(\mathbf{d}) \qquad (4.3)$$

where

$$\mathbf{r}(\mathbf{d}) = \mathbf{f}^{ext} - \mathbf{f}^{int} \tag{4.4}$$

and require $\mathcal{F} = 0$ at the minimum.

This provides a more well-behaved problem for the line-search procedure and the rate of convergence of the Newton method is improved substantially. The method was also adapted in [19] so as to combine it with arc-length procedures.

To test the effectiveness of the localization limiter and the solution strategy, we consider the problem of a one-dimensional rod, subjected to equal and opposite loading at its two ends, as illustrated in figure 9. One node in the mesh is held fixed, so as to prevent rigid body translations.To trigger the appearance of a non-homogeneous strain-distribution,a small imperfection is introduced. In the present example, this was accomplished by making the cross-section of the center element 1% smaller than the cross-section of all other elements.

Numerical studies of this problem were conducted based on the localization limiter defined previously in Eqn. (3.17), which in one-dimension reduces to:

$$\sigma = \phi(\varepsilon) - \alpha\, \varepsilon_{,xx} \tag{4.5}$$

The elasto-plastic strain-stress law considered is also illustrated in fig. 9. It consists of a linear elastic part, and an exponential branch including a strain hardening portion followed by a softening one. At any point the unloading is elastic with Young's modulus E. The physical parameters used for the calculation were: Young's modulus E=200, yield strain $\varepsilon_1 = .05$, $\varepsilon_m = 0.3$, exponential branch: $\phi(\varepsilon) = E\,\varepsilon_1 \frac{(\varepsilon+\delta_0)}{(\varepsilon_1+\delta_0)}\, e^{g(\varepsilon)-g(\varepsilon_1)}$ where $g(\varepsilon) = \left(1 - \frac{\varepsilon+\delta_0}{\varepsilon_m+\delta_0}\right)$, parameter controlling the convexity $\delta_0 = 0.11$.

To solve this problem, the line-search technique combined with the arc-length method with a linearized constraint equation described in [19] was used. It was first checked that without introducing the localization limiter, that is for $\alpha=0$, the deformation localizes in the element with imperfection, irrespective of its size, while all other elements unload elastically. In a load-displacement curve, a sharp decrease is observed once strain-softening is attained, and even a snap-back behavior can be observed, which could not be captured with a pure displacement control strategy.

Calculations were then conducted with the localization limiter defined in Eqn. (4.5), for several values of the parameter α. The strain distribution along the rod for various load levels is given in fig. 10a. These strain-profiles are very close in shape to the ones obtained by Coleman and Hodgdon[15] in their study of the effects of the localization limiter (4.5) on strain-localization for a rigid plastic material with a parabolic law. Essentially, a finite localization zone emerges, practically constant in size, where the strain increases but remains bounded, while in the rest of the rod, the material unloads elastically. In the finite element calculation, that localized zone spans a few elements of the mesh. The size of the zone is directly related to the value of α.

Load-displacement curves for various values of α are reported in fig. 10b. They exhibit a milder negative slope with increasing α. It should be pointed out that, without the use of the line search procedure described above, the Newton-Raphson procedure fails to converge near the critical point.

5. CONCLUSIONS

Localization limiters can be classified as nonlocal, differential and rate limiters. A Fourier analysis of the wave-propagation problem shows that the introduction of nonlocal or differential limiters leads to governing equations where short waves, which are likely to develop with the onset of strain-softening, have a bounded growth. In dynamic problems, strain-softening causes the governing equations to lose strict hyperbolicity; it was shown for example that they become elliptic in at least one direction for the antiplane problem. With the gradient-type localization limiter, the dynamic equations change from hyperbolic to parabolic for the one-dimensional case. The character of the amplification spectrum of the integral and differential limiters is similar and they become identical in the limit as the magnitude of the parameter governing the limiter goes to zero.

The differential localization limiter proposed by Coleman and Hodgdon [15] based on the introduction of the second derivative of the strain in the stress expression was implemented in the context of static problems. Numerical studies showed that it allows for the development of a localized strain zone spanning over several elements of the mesh. However, the addition of the limiter does not guarantee positive definiteness of the tangent matrix.

REFERENCES

[1] Z.P. Bazant, T.B. Belytschko, "Wave Propagation in Strain-Softening Bar: Exact Solution," J. Egng Mech. ASCE, **111**, 381-389 (1985).

[2] F.H. Wu and L.B. Freund, "Deformation Trapping due to Thermoplastic Instability in One-Dimensional Wave Propagation," J. Mech. Phys. Solids, **32**, 119-132 (1984).

[3] Z.P. Bazant, T.B. Belytschko and T.P. Chang, "Continuum Theory for Strain-Softening," J. Egng Mech. ASCE, **110**, 1666-1692 (1984).

[4] D. Lasry and T. Belytschko, "Localization Limiters in Transient Problems," Int. J. Solids and Structures, **24**, 581-597(1988).

[5] J.R. Rice, "The Localization of Plastic Deformation," In Theoretical and Applied Mechanics, Proc. 14th Int. Congr. Theoret. Appl. Mech. (ed. W.T. Koiter, North-Holland, Amsterdam), 207-220 (1977).

[6] T. Belytschko, J. Fish, B.E. Engelmann, "A Finite Element with Embedded Localization Zones," Comp. Meth. Appl. Mech. Engng., in press (1988).

[7] J. K. Knowles, "The Finite Antiplane Shear Field Near the Tip of a Crack for a Class of Incompressible Elastic Solids," Int. J. of Fracture, **13**, 611-639 (1977).

[8] L.B. Freund, F.H. Wu and M Toulios, "Initiation and Propagation of Shear Band in Antiplane Shear Deformation," Proceedings of the Intl. Symposium on Plastic Instability, Considère Memorial, Paris Sept. 1985 (Publ. by Presses de l'Ecole Nationale des Ponts et Chaussées, Paris), 125-134 (1985).

[9] M Toulios, "The wavefront induced in a homogeneously shearing solid by a localized material imperfection," Quarterly of Applied Math., **43**, 225-235 (1985).

[10] R. Hill, "Acceleration Waves in Solids," J.Mech. Phys. Solids, **10**, 1-16 (1962).

[11] R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. II, Interscience, New York (1962).

[12] R. J. Clifton, "A Difference Method for Plane Problems in Dynamic Elasticity," Quarterly of Applied Math., **25**, 97-116 (1967).

[13] J.W. Rudnicki and J.R. Rice, "Conditions for the Localization of Deformation in Pressure Sensitive Dilatant Materials," J. Mech. Phys. Solids, **23**, 371-394 (1975).

[14] H.L. Schreyer and Z. Chen, "The Effect of Localization on the Softening Behavior of Structural Members," In Constitutive Equations: Macro and Computational Aspects, ed. K.J. William, ASME, New York, 193-203 (1984).

[15] B.D. Coleman and M.L. Hodgdon, "On Shear Bands in Ductile Materials," Arch. Rational Mech. Anal., **90**, 219-247 (1985).

[16] N. Triantafyllidis and E.C. Aifantis, "A Gradient Approach to Localization of Deformation. I. Hyperelastic Materials," J. of Elasticity, **16**, 225-237 (1986).

[17] H.M. Zbib and E.C. Aifantis, "On the Post-localization of Plastic Deformation. I. The Evolution of Shear Bands in Plastic Materials," Michigan Institute of Technology, MM Report 14, (1987).

[18] A. Needleman, "Material Rate Dependence and Mesh Sensitivity in Localization Problems," Comp. Meth. Appl. Mech. Engng., **67**, 61-85 (1988).

[19] T. Belytschko, D. Lasry and T. Lodygowski, "A Modified Newton Procedure For Post-Bifurcation Analysis, as in Strain-Softening," submitted for publication in Int. J. Num. Meth. Engg.(1988).

[20] T.B. Belytschko, Z.P. Bazant, Y.W. Hyun and T.P. Chang, "Strain-Softening Materials and Finite-Element Solutions," Computers & Structures, **23**(2), 163-180 (1986).

[21] T. Belytschko and D. Lasry, "A Study of Localization Limiters for Strain-Softening in Statics and Dynamics," submitted for publication in Computers and Structures (1988).

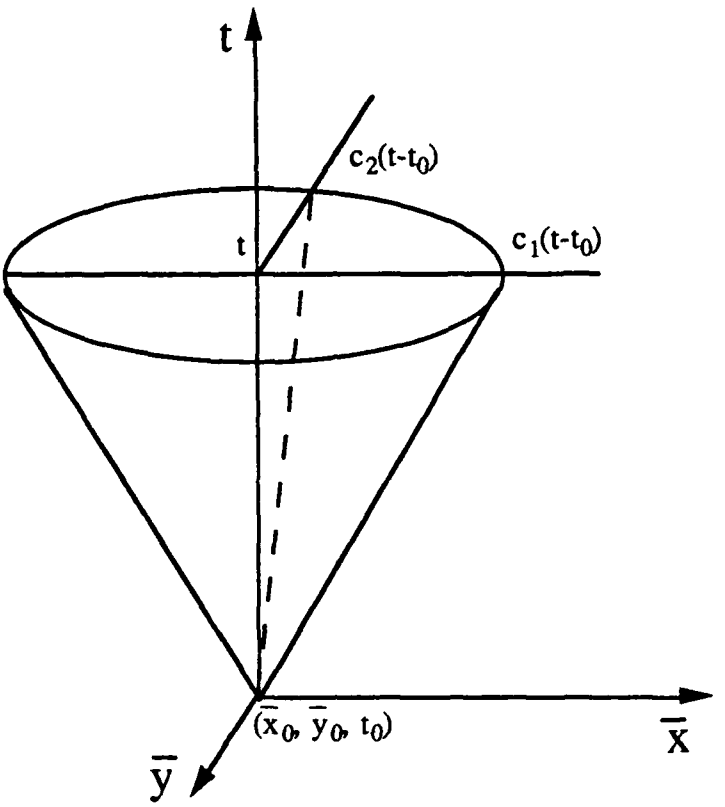


Figure 1. Charateristic cone for the dynamic equations.

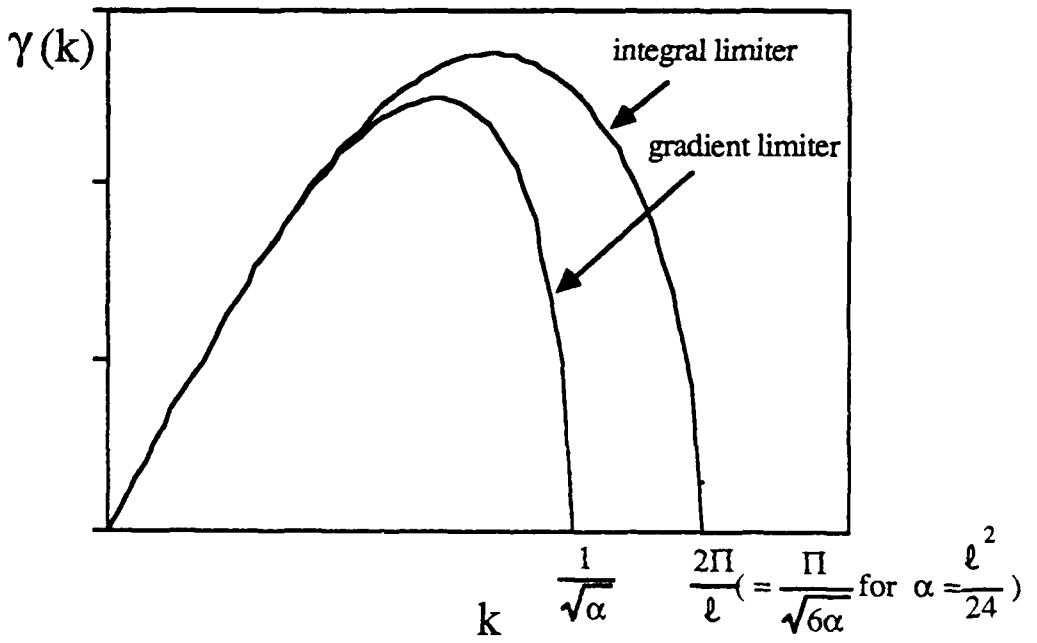


Figure 2. Linearized analysis: $\hat{\gamma}(x)$ (gradient-type limiter) and $\bar{\gamma}(k)$ (integral-type limiter) versus the wavelength k.

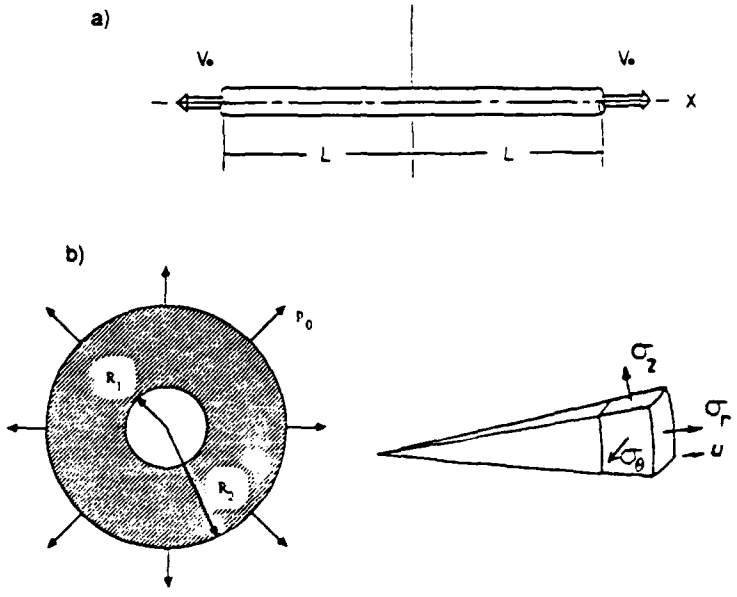


Fig. 3. Problem descriptions. a) 1D-rod problem, 2L=40; b) Spherically symmetric problem, interior radius R_1=10., exterior radius R_2=100 .

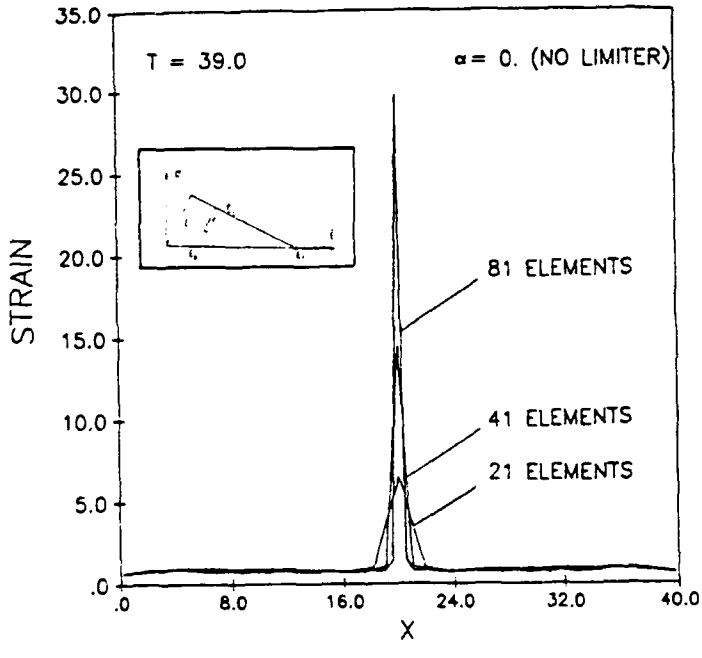


Fig. 4. Rod problem, strain plots at time T=39.0 for different meshes, no higher order term limiter (α=0). In enclosed box, stress-strain curve: E is Young's modulus; for the spherically symmetric problem, E is replaced by K (bulk modulus), E_T by K_T, ε by ε_v (volumetric strain).

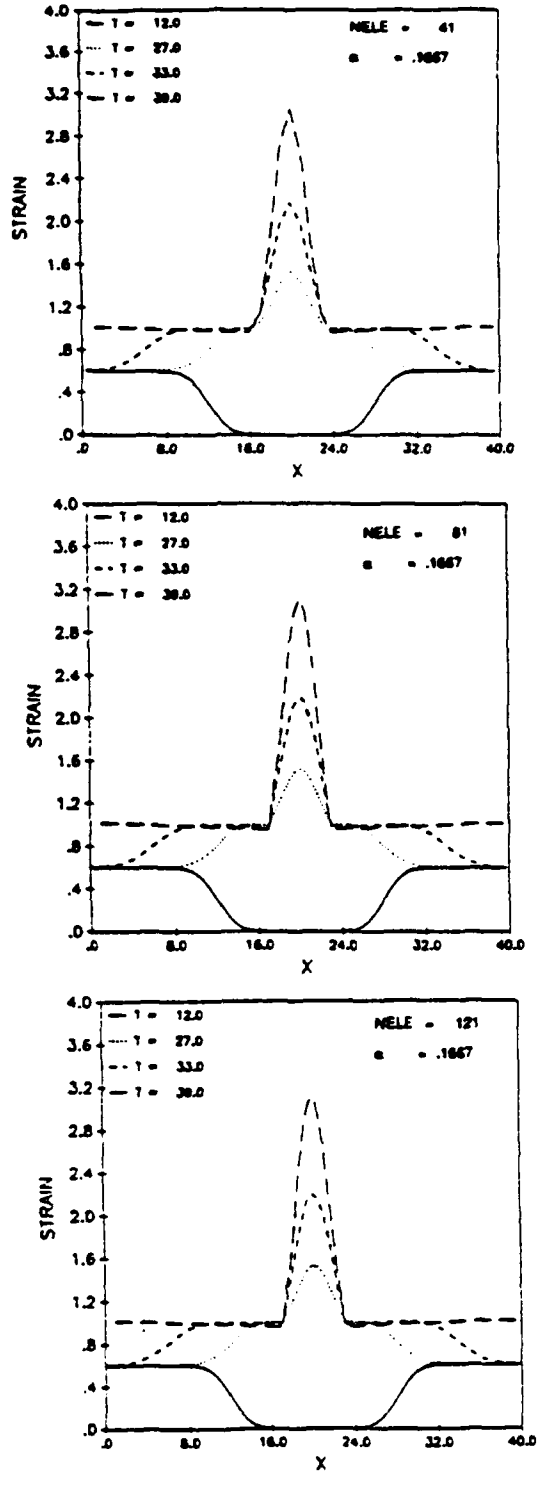


Fig. 5. Rod problem, strain plots obtained with localization limiter (α=.1667), meshes of 41, 81 and 121 elements.

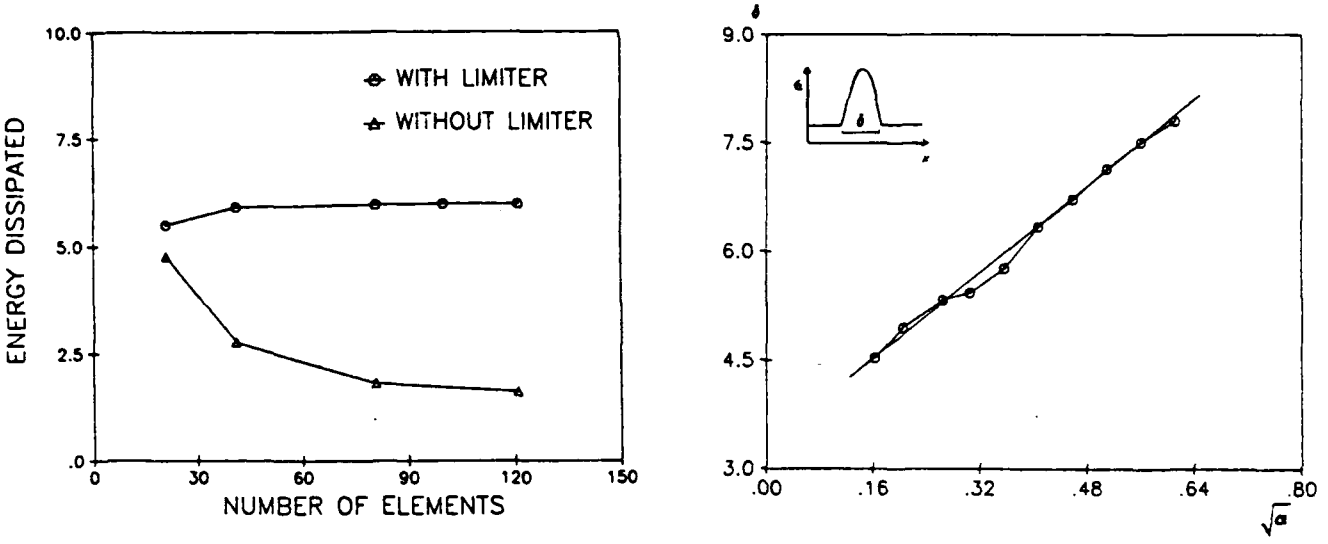


Fig. 6. Rod problem: a) Energy dissipated vs number of elements, with and without localization limiter. b)Size δ of the localization zone as a function of $\sqrt{\alpha}$.

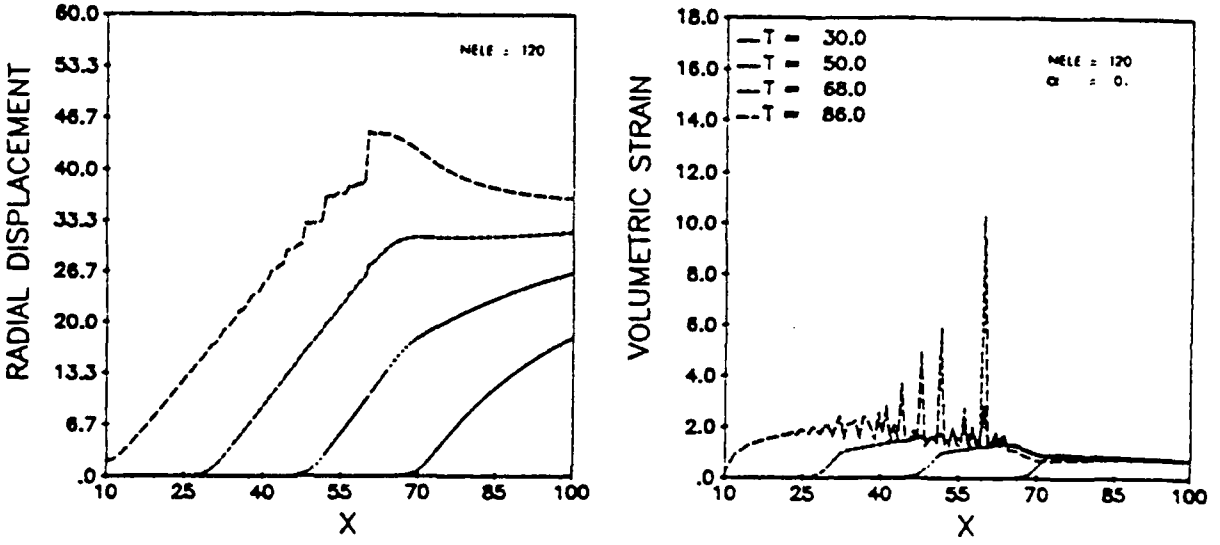


Fig. 7. Spherically symmetric problem, radial displacement and volumetric strain plots, without using the higher order term limiter (α=0.).

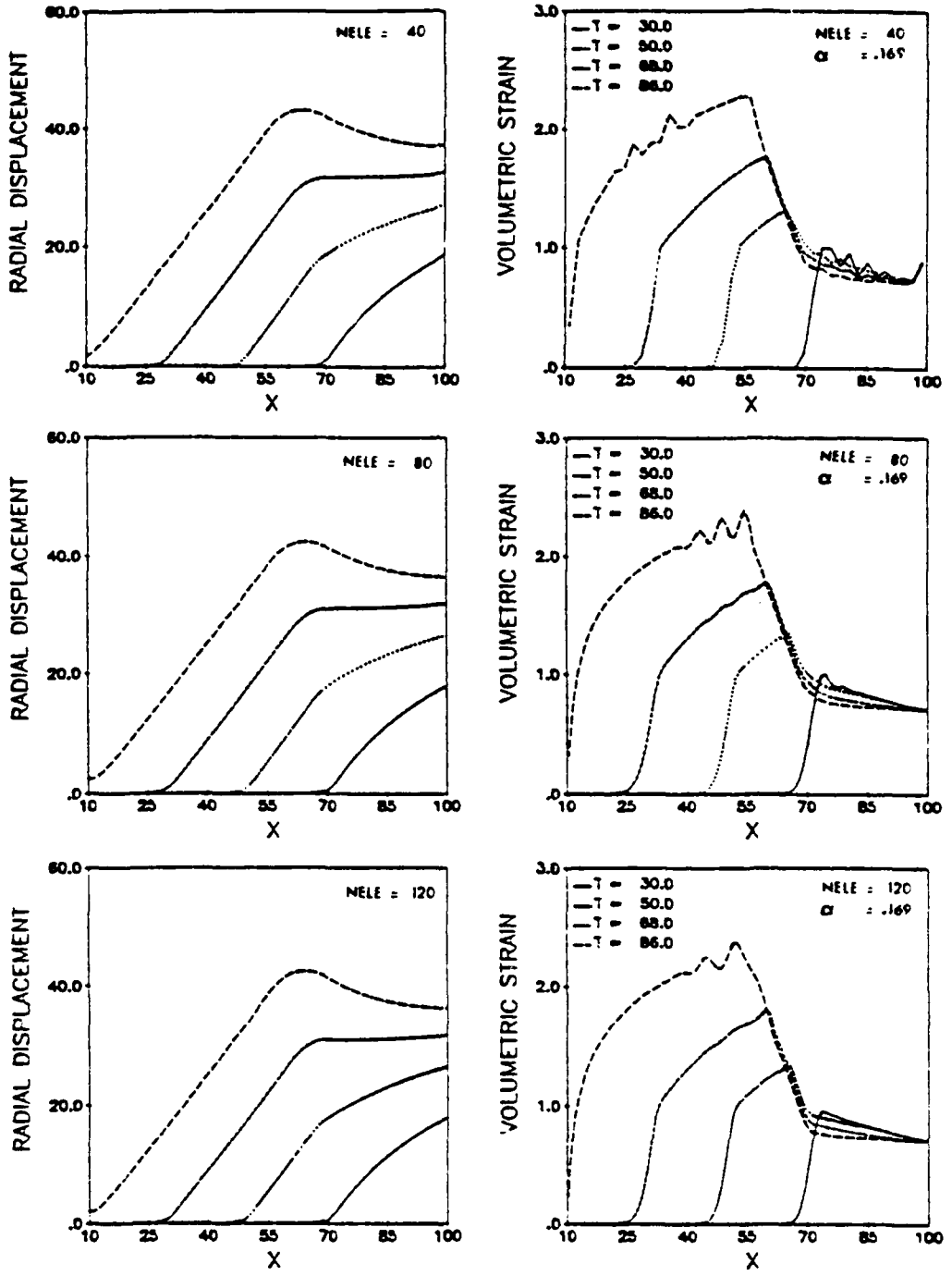


Fig. 8. Spherically symmetric problem, radial displacement and volumetric strain plots, with limiter (α=.169), with meshes of 41, 81 and 121 elements.

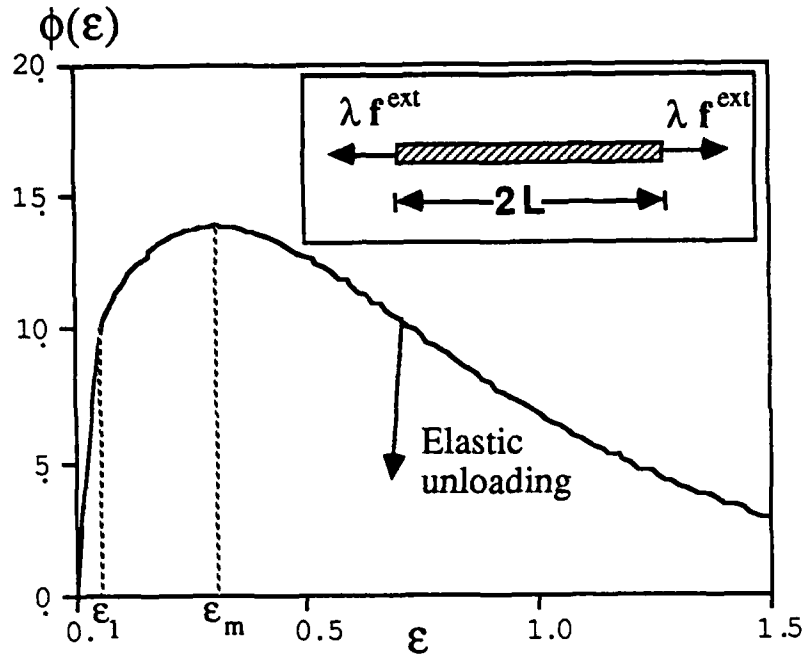


Figure 9. Traction curve ε-ϕ(ε). In enclosed box, problem description: rod length 2L=10, 40 nodal points.

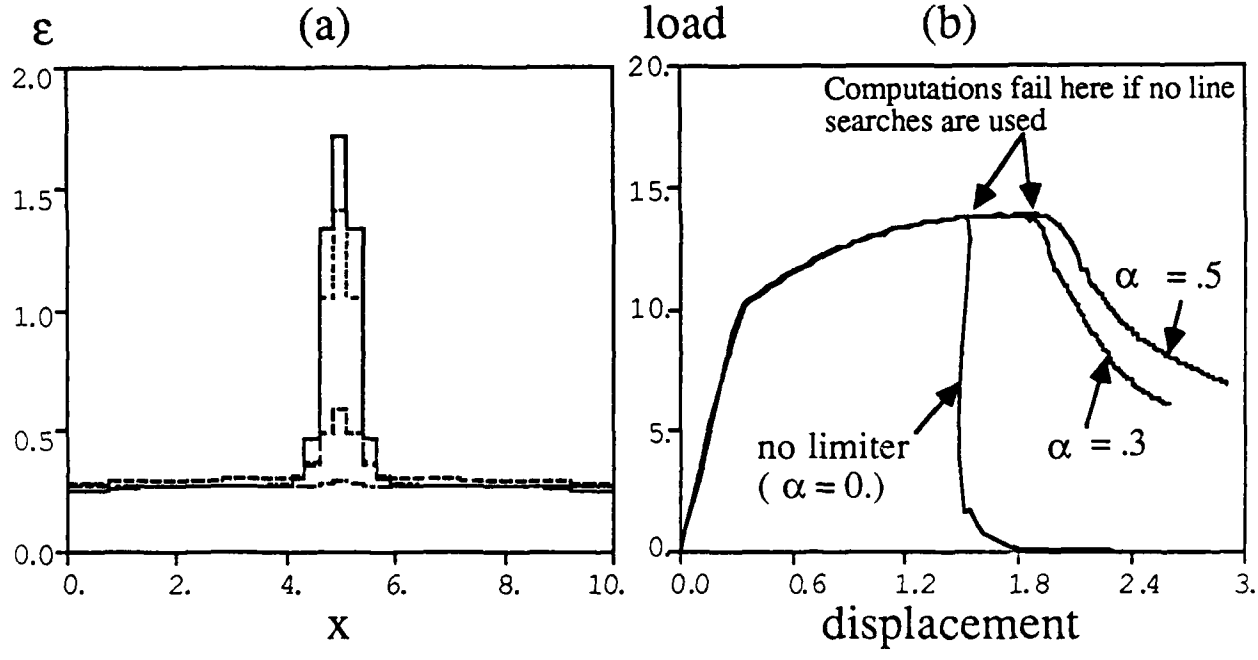


Figure 10. a) Strain vs spatial coordinate profiles at four different load steps with higher order term limiter (α=.5). b) Load-displacement curves for different values of α.

MODELING TWO-DIMENSIONAL DETONATIONS WITH DETONATION SHOCK DYNAMICS*

J. B. Bdzil
Los Alamos National Laboratory, Los Alamos, New Mexico 87545

D. S. Stewart
University of Illinois, Urbana, Illinois, 61801

Abstract

In any explosive device, the chemical reaction of the explosive takes place in a thin zone just behind the shock front. The finite size of the reaction zone is responsible for: the pressure generated by the explosive being less near the boundaries, for the detonation velocity being lower near a boundary than away from it, and for the detonation velocity being lower for a divergent wave than for a plane wave.

In computer models that are used for engineering design calculations, the simplest treatment of the explosive reaction zone is to ignore it completely. Most explosive modeling is still done this way. The neglected effects are small when the reaction zone is very much smaller than the explosive's physical dimensions. When the ratio of the explosive's detonation reaction-zone length to a representative system dimension is of the order of 1/100, neglecting the reaction zone is not adequate.

An obvious solution is to model the reaction zone in full detail. At present, there is not sufficient computer power to do so economically. Recently we have developed an alternative to this standard approach. By transforming the governing equations to the proper intrinsic-coordinate frame, we have simplified the analysis of the two-dimensional reaction-zone problem. When the radius of curvature of the detonation shock is large compared to the reaction-zone length, the calculation of the two-dimensional reaction zone can be reduced to a sequence of one-dimensional problems.

I. Introduction

Describing the propagation of detonation in complex multi-dimensional explosive geometries is an important and ongoing problem in the design process for explosively driven devices. In order for the design of the explosive system to be successful, two requirements need to be met. First, the detonation of the explosive system must be robust, that is relatively insensitive to variations in the initial conditions, such as changes in temperature and variations in the initiation system. At the same time, the explosive system must be safe from accidental initiation of detonation. The parameter $\mathcal{P}$, which is the ratio of the explosive's detonation reaction-zone length to a representative system dimension, is the parameter that controls these properties. When $\mathcal{P}$ is small (relatively fast reaction) the system is robust, but prone to accidental initiation. When $\mathcal{P}$ is large the explosive is near its failure limit making it harder to set off accidentally but also more sensitive to variations in the initial condition. A value of $\mathcal{P}$ of about .01 is a good compromise. Problems of accidental initiation are minimized, yet at the same time the detonation is relatively insensitive to initial conditions.

For most explosive geometries, this ratio is small enough so that the integrated momentum

*Supported by the U. S. Department of Energy.

through the reaction zone is small in comparison to that in the broad region where the reaction products expand and do work on their surroundings. Thus the reaction zone has little direct influence on the process of driving inert materials that are in contact with it. However, the indirect influences of the reaction zone on the calculation can be much more important. When $\rho = .01$ a significant fraction of the explosive charge experiences such things as reduced detonation pressure and velocity near boundaries, as well as a slower detonation velocity everywhere for a divergent detonation than for a plane one. These, in turn, lead to large errors in zeroth-order effects such as the time of detonation arrival and the two-dimensional detonation wave shape. From the point of view of the designer, this is a difficult computational regime. Not only does he need to resolve the broad region where the reaction products expand and do work on their surroundings, but he must also resolve the thin reaction zone.

Because of the disparate lengths of the reaction zone and the products expansion wave, most of the explosive design codes in use today employ some variant of the constant-detonation-velocity "Huygens" construction to propagate the detonation wave. This method for propagating the detonation only works well for explosives for which the reaction zone can be ignored (i.e., ρ is less than 1/1000). Ad hoc "fixes" of this simple model have been tried to model systems with larger values of ρ. For example, the detonation velocity may arbitrarily be set to some lower value near the edge. These have met with only limited success.

With all of its shortcomings, the simple "Huygens" method has one real advantage, computational speed. Since the reaction zone does not need to be modeled, design calculations are fast enough to allow many design iterations to be tried. This is an important feature that design codes need to have.

In order to improve on this simple method, the reaction zone must be modeled. This of course requires knowledge of the equation of state (eos) of the partially reacted explosive and of the reaction rate. When explicit information is available, one can in principle follow the standard approach and do multi-dimensional simulations that resolve both the reaction zone and the explosive products region. Typically we have only limited constitutive information: the shock Hugoniot of the "unreacted" explosive, an equation of state of the explosive products, and a compatible energy-release rate callibrated to one-dimensional experiments.

To be useful, a numerical simulation of the reaction zone must be able to resolve all of the important features of the flow. Fickett[1] has shown that when the standard one-dimensional (1D) Lagrangian-mesh artificial-viscosity methods are used, roughly 15 computational cells are needed in the reaction zone to get 10% accuracy. This translates into many tens of thousands of computational cells for a typical two-dimensional (2D) numerical calculation done with a uniform grid method. Even with today's supercomputers, such calculations take many hours of computation time; they are not practical for routine use. When one reduces the number of cells in the calculation in order to get sensible computational times, the accuracy of the calculations suffers.

In large measure, the inordinately large computation time is a result of the lack of sophistication of the standard uniform grid methods. The mesh size that is needed to achieve reasonable resolution in the reaction zone is excessively fine for the broad products expansion region. Today researchers are developing a variety of improved methods that include such features as: (1) multi-grid techniques that employ moving fine zoning near shocks,[2] (2) schemes based on the method of characteristics such as CIR and Godunov,[2,3] and (3) shock-tracking methods.[4] To date, however, none of these methods has reached the point of maturity where they could replace the standard method for routine detonation calculations.

The central issue in improved 2D calculations of detonation is a high-accuracy calculation of the reaction-zone structure, plus a relatively coarse-grid calculation of the following products release wave. One way of getting a high-accuracy calculation of the reaction-zone structure is to do it analytically. This alternative brings with it the direct computational benefit plus the advantage of a theoretical understanding of the 2D detonation process. With such an understanding, we could make a fast high-resolution wave-tracking code that solves the reaction-zone flow analytically and the broad products region with a coarse-grid numerical simulation. This increased knowledge also brings with it the insights that lead to the improvements that are necessary if some of the more sophisticated computational methods mentioned above are to become practical tools.

An analytical solution of the general 2D time-dependent detonation problem is not within reach. However, in many applications of explosives, one observes that the radius of curvature of the detonation shock is large in comparison with the reaction-zone length. Recently we have developed an alternative to the standard numerical approach that is based on the large radius of curvature limit. By transforming the governing equations to the proper intrinsic-coordinate frame, we have simplified the analysis of the 2D reaction-zone problem, and reduced it to a sequence of one-dimensional problems. The coordinate frame of choice is one in which the spatial coordinate axes are everywhere locally parallel and perpendicular to the shock. The governing equations consist of a kinematic equation that describes the progress of disturbances moving along the shock, and equations for the reaction-zone dynamics that describe the quasi-steady flow normal to the shock (i.e., through the reaction zone). We call this method DETONATION SHOCK DYNAMICS (DSD).

This paper gives a brief review of DSD. We have divided it into four sections. In Section II, we give an overview of the theoretical model. This section is divided into three subsections. In Shock Kinematics, we briefly describe our coordinate system and the kinematics of the detonation shock. The subsection entitled Boundary Conditions is devoted to a discussion of the boundary conditions that are applied at the edges of the explosive. In Reaction-Zone Dynamics, the Euler equations are transformed to the intrinsic-coordinate frame, and the analysis that leads to the quasi-steady description is briefly reviewed. In Section III, we demonstrate how our theory can be used to study a representative explosive design problem. In Section IV, we summarize our results.

II. Overview of the Theory

The thrust behind our theory is the concept that the response of the detonation shock is local, and is governed by its current local configuration. Philosophically, it is an extension of Whitham's geometrical shock dynamics to detonation.[5] Our theory is a uniform perturbation theory, which is based on the notion that the radius of curvature of the shock is large compared to the reaction-zone length. It is a nonlinear theory that can be used to describe arbitrarily large departures of the detonation shock shape from the plane one-dimensional state. From the results of our theoretical calculations, the following picture has emerged. In many situations, the dynamics of the detonation reaction zone is decoupled from the evolution of the large following reaction products expansion wave, and is controlled by the flow near the shock. As a result, we find that the important waves in the reaction zone, either rarefactions or compressions, are transverse waves. Our theory describes how waves on the shock are generated (e.g., near an explosive edge) and move along the shock (see Figure 1).

There are three components to the theory: (1) a kinematic condition for the shock surface, (2) conditions to be satisfied at the boundaries of the explosive, and (3) the flow dynamics in the direction normal to the shock. We will briefly describe each of these.

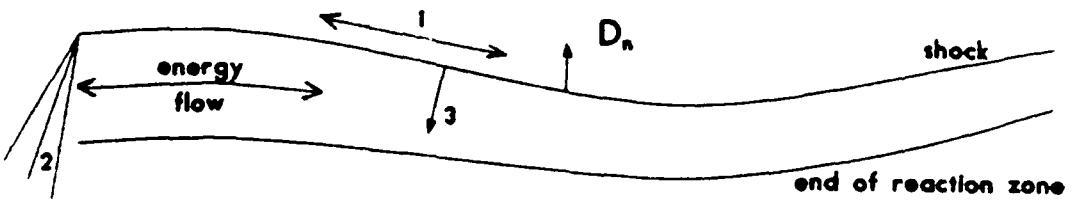


Figure 1. A schematic diagram that shows how chemical/mechanical energy are transported laterally through the reaction zone. The kinematic condition is applied along (1), boundary conditions are applied at (2) and the reaction-zone dynamics describes the flow along (3). To leading order, the reaction zone is insulated from rarefactions from the rear.

Our theory is based on the time-dependent, two-dimensional, reactive Euler equations. As a consequence, the detonation shock (shock) is a surface of discontinuity. Since we wish to treat detonation-wave evolution in complicated two-dimensional geometries, we have developed our theory in a problem-determined intrinsic-coordinate system (see Figure 2). It is a shock-centered frame that moves with the local normal detonation-shock velocity (D_n). The space variables are

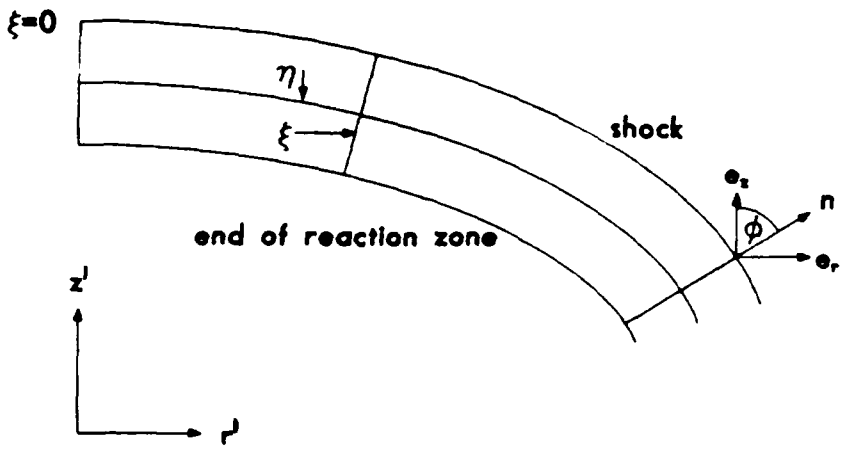


Figure 2. The intrinsic-coordinate system that was used in the calculation. The shock curvature is $\kappa = \phi_{,\xi}$ and $z^\ell = z_o^\ell - \eta\cos\phi \quad , r^\ell = r_o^\ell - \eta\sin\phi \quad .$

the distances ξ and η locally parallel and perpendicular to the shock.

a. *Shock Kinematics*

The principal object of the theory is to calculate the shock shape as a function of time. The intrinsic representation of a curve, such as the shock, is in terms of its curvature (κ) as a function of arc length along the shock (ξ) and time (t). In this coordinate system, the shock shape is described by the shock angle (ϕ) as a function of ξ and t. In terms of these variables, the shock curvature is $\kappa \equiv \phi_{,\xi}$, where the $,\xi$ indicates a partial derivative with respect to arc length. The laboratory coordinates for the shock are returned by

$$z_o^\ell = z_e^\ell - \int_o^\xi \sin(\phi)d\xi \quad , r_o^\ell = r_e^\ell + \int_o^\xi \cos(\phi)d\xi \quad , \tag{1}$$

where z_e^ℓ and r_e^ℓ are the coordinates of the edge. Typically we are most interested in describing the changes in the shock shape that are the result of the interaction that occurs between the shock and an explosive edge. For such problems, having the zero of arc length coincide with the edge is the most convenient origin to use for ξ. Figure 3 shows a schematic representation of the shock including the independent variable (ξ) and the definition of the dependent variables D_n and ϕ. The cartesian unit vectors are $\hat{e}_z$ and $\hat{e}_r$.

The geometric compatibility conditions for a moving two-dimensional surface are given in Whitham[5]

$$\phi_{,\alpha} = -\frac{1}{A} D_{n,\beta} \tag{2}$$

and

$$\phi_{,\beta} = \frac{1}{D_n} A_{,\alpha} \quad . \tag{3}$$

The variable α is equivalent to time, and labels a particular shock surface. The constant-β rays are orthogonal to the shock and are its propagators. The streamtube area is A, where at fixed α

$$d\xi = A d\beta \tag{4}$$

(i.e., the shock area between two adjacent constant-β rays or streamlines).

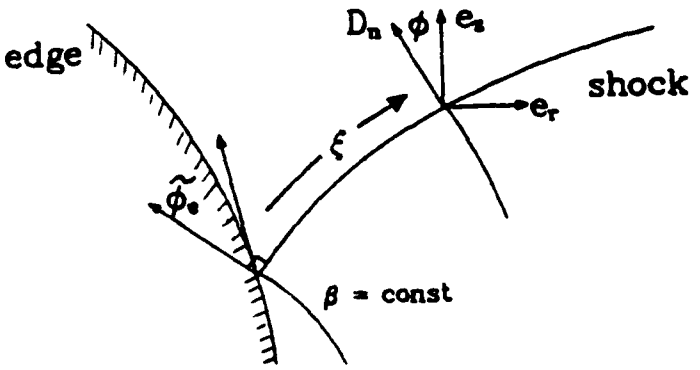


Figure 3. Intrinsic coordinates and shock kinematics. The independent variables are arc length (ξ) and time (t), while the dependent variables are the normal shock velocity (D_n) and the shock normal angle (ϕ). The curves β = constant are normal to the shock, and $\tilde{\phi}_e$ is the angle between the tangent to the edge and normal to the shock.

For the problems of interest in condensed-phase detonation, the shock is seldom normal to the explosive boundary. As a result, the coordinate β is not a convenient independent variable since boundary conditions must be applied at the edge. Changing independent variables from (α, β) to (t, ξ), we have

$$d\xi = A d\beta + B d\alpha \tag{5}$$

and

$$dt = d\alpha \quad , \tag{6}$$

where the coefficient B describes the change in arc length with time along a constant β ray. Under this transformation, the surface kinematics [i.e., Eq. (2)] takes the form of a one-dimensional Burgers equation along the shock: B is the wave velocity and $D_{n,\xi}$ is the transport term

$$\phi_{,t} + B\phi_{,\xi} = -D_{n,\xi} \quad . \tag{7}$$

The coefficient B is obtained by requiring that the transformation [Eqs. (5) and (6)] be solvable, from which it follows that

$$A_{,\alpha} = B_{,\beta} \quad . \tag{8}$$

From Eqs. (3) and (8) it follows that

$$B = \int_o^{\xi} \phi_{,\xi} D_n d\xi + B_o(t) \quad . \tag{9}$$

The function $B_o(t)$ is the rate at which shock arc length crosses the β = constant ray that intercepts the edge. It is given by

$$B_o(t) = D_{ne} \tan(\tilde{\phi}_e) \quad . \tag{10}$$

This intrinsic form of the shock-surface kinematics is fundamental to any shock-tracking method that seeks to describe the evolution of shocks of arbitrary shape in a uniform manner. Clearly, Eqs. (7) and (9) simply yield a constraint between D_n and $\kappa = \phi_{,\xi}$. However, if a second algebraic relation between D_n and κ can be obtained, then this contraint can be converted into a one-dimensional partial-differential equation for the shock surface. Further, if we then prescribed the initial shape (ϕ) of the surface, as well as some boundary condition at the intersection of the shock and the explosive boundary, then Eq. (7) could be solved to get the 2D shock locus at any subsequent time.

b. *Boundary Conditions*

For the problems we consider here, we do not need to study the complex flow or the detailed boundary conditions that apply in the vicinity of the explosive boundary. It will be sufficient to consider only the condition, if any, that must be applied at the locus generated by the intersection of the shock and the edge. We consider only an explosive/vacuum interface.

At such an interface, the flow experiences a singularity. In the explosive, the pressure just behind the detonation shock is near the Chapman-Jouguet (cj) pressure; just outside the explosive, the pressure is at or near zero. In order for the flow to execute such a transition, a singularity of Prandtl-Meyer (PM)—type must be embedded in the flow at the intersection of the shock and the edge. Since locally the flow at this point is quasi-steady, it can only be either a sonic or a supersonic flow (as seen by an observer riding along the edge/shock intersection locus). We will discuss the consequences that result from having flows of these two types.

Along the edge/shock locus, the sonic parameter is a function of the normal detonation velocity along the edge, D_{n_e}, and the shock interface angle, $\tilde{\phi}_e$. For a polytropic eos, with γ the polytropic exponent, the expression is

$$C^2 - |U|^2 = D_{ne}^2 \left\{ \frac{\gamma}{\gamma+1} \sqrt{1 - \frac{D_{n*}^2}{D_{ne}^2}} - \frac{1}{\gamma+1} \left(1 - \frac{D_{n*}^2}{D_{ne}^2} \right) - \tan^2(\tilde{\phi}_e) \right\} \quad , \tag{11}$$

where C is the sound speed, $|U|$ is the magnitude of the particle velocity in the edge/shock locus frame and D_{n*} is the minimum value of D_n for a one-dimensional detonation.

If the flow is supersonic along the locus, then disturbances from the edge can not propagate into the detonation reaction zone. The interface moves faster laterally than do acoustic waves. For this case, no boundary condition is applied, and the interface does not affect the detonation. As the flow turns subsonic, then D_{ne} and $\tilde{\phi}_e$ must be adjusted so that the sonic condition, $C^2 - |U|^2 = 0$, is maintained. This condition serves as a boundary condition for the flow.

The following rule summarizes the the edge/shock locus boundary condition: monitor the sonic parameter on the locus. If $C^2 - |U|^2 < 0$, the flow is supersonic and no condition is applied. When the flow is either sonic or subsonic, then D_{ne} and $\tilde{\phi}_e$ must be adjusted to satisfy the condition $C^2 - |U|^2 = 0$.

c. *Reaction-Zone Dynamics*

As noted above, Eq. (7) is a one-dimensional partial-differential relation that D_n and ϕ must satisfy if they are to describe a two-dimensional shock. If a second relation between D_n and ϕ can be found, we can convert this relation to a partial-differential equation (pde), and in the process reduce the two-dimensional shock tracking problem to a one-dimensional one. For a number of cases, we have found such a second relation between D_n and $\kappa = \phi_{,\xi}$. When it exists, this relation contains all the necessary reaction-zone dynamics; the consequences of the interaction of the chemical-heat release with the flow. To find it, we must solve the time-dependent two-dimensional Euler equations. In order to solve these equations for complex explosive geometries, we must express them in terms of a natural system of coordinates that simplifies their form. In the limit that the radius of curvature of the shock is large compared to the reaction-zone length, the coordinates shown in Figure 2 are particularly convenient. Bertrand curves that are everywhere parallel to the shock are the constant-η coordinates; the lines perpendicular to these curves are the constant-ξ coordinates. These coordinates are related to the laboratory cartesian frame, by

$$z^\ell = z_o^\ell - \eta \cos\phi \tag{12}$$

and

$$r^\ell = r_o^\ell - \eta \sin\phi \quad , \tag{13}$$

where z_o^ℓ and r_o^ℓ are given by Eq. (1). Expressed in these coordinates, the Euler equations are

$$\textit{mass} \qquad \mathcal{L}\rho + \rho(\kappa U_\eta - U_{\eta,\eta} + U_{\xi,\xi}) + \ldots = 0 \quad , \tag{14}$$

$$\eta - \textit{momentum} \qquad \mathcal{L}U_\eta - \frac{1}{\rho}P_{,\eta} + \ldots = 0 \quad , \tag{15}$$

$$\xi - \textit{momentum} \qquad \mathcal{L}U_\xi + \frac{1}{\rho}P_{,\xi} - D_{n,\xi}U_\eta + \ldots = 0 \quad , \tag{16}$$

and

$$\textit{energy} \qquad \mathcal{L}E - \frac{P}{\rho^2}\mathcal{L}\rho + \ldots = 0 \quad . \tag{17}$$

The chemical rate law is

$$\textit{rate} \qquad \mathcal{L}\lambda + \ldots = \mathcal{R} \quad . \tag{18}$$

We have displayed only those terms that are necessary to do the leading order theory in the small κ-limit. In the above, the operator $\mathcal{L}$ is

$$\mathcal{L} \equiv \frac{\partial}{\partial t} + (D_n - U_\eta)\frac{\partial}{\partial \eta} + B\frac{\partial}{\partial \xi} \quad , \tag{19}$$

ρ is the density, U_η is the η-component of the particle velocity (at leading order $U_\eta > 0$ and $U_{\eta,\eta} < 0$), U_ξ is the ξ-particle velocity ($U_\xi = 0$ at the shock), P is the pressure, λ is the degree of reaction ($\lambda = 0$ at the shock), $\mathcal{R}$ is the chemical rate and E is the specific internal energy. The above equations, the standard one-dimensional shock conditions, the kinematics [Eq. (7)] and appropriate initial/boundary conditions completely define the 2D problem that must be solved. Even in the small-κ limit, this is a formidable task.

What we have shown recently is that for certain rate-law forms (i.e., expressions for $\mathcal{R}$), the important large-scale dynamics is quasi steady.[6] We considered relatively long-scale disturbances to the shock

$$\epsilon^2 = 0(\kappa) \ll 1 \tag{20}$$

$$D_n = D_{cj} + 0(\epsilon^2) \quad , \tag{21}$$

and two spatio-temporal regimes:

$$\begin{aligned} &(1)\ \text{"fast" dynamics} && \left\{t_1 = \epsilon t \quad , \xi_{1/2} = \epsilon^{1/2}\xi\right\} \\ &\quad \text{shock deflection} && \phi = 0(\epsilon^{3/2}) \end{aligned} \tag{22}$$

and

$$\begin{aligned} &(2)\ \text{quasi-steady dynamics} && \left\{t_2 = \epsilon^2 t,\ \ \xi_1 = \epsilon\xi\right\} \\ &\quad \text{shock deflection} && \phi = 0(\epsilon),\ \text{or larger} \quad . \end{aligned} \tag{23}$$

The "fast" scale problem was necessary to treat the early influence of the two-dimensional initial/boundary data, and to describe the hydrodynamic wavehead that separates the reaction zone into parts that are either influenced or uninfluenced by the edge. As the flow evolves, the "fast" scale perturbations become smaller, and the disturbances to the one-dimensional state

became larger and quasi-steady. This quasi-steady regime is particularly simple; the Euler equations reduce to the steady nozzle equations [a steady cylindrically-symmetric system of ordinary-differential equations (ode)]

$$\left[(D_n - U_\eta)\rho\right]_{,\eta} + \rho\kappa U_\eta = 0 \quad , \tag{24}$$

etc.

The only parameters in these equations, besides the fixed constitutive parameters, are D_n and κ. That is, the initial/boundary data did not appear in the large-change reaction-zone dynamics. In some sense then, the dynamics is local and universal. The resulting one-dimensional problem is simply the detonation "eigenvalue" problem considered by Wood & Kirkwood.[7] In this limit, detonation shock propagation problem decouples from the product expansion region. Therefore for detonation, no *ad hoc* approximations are necessary to get a theory for the shock evolution that is local. At least this is the case for diverging detonation.

The quasi-steady problem defines $D_n(\kappa)$. With κ specified, D_n is determined by solving an eigenvalue problem. In addition to yielding $D_n(\kappa)$, this solution also gives the reaction zone end state as a function of κ. Thus for an important class of problems, the reaction-zone dynamics is given by $D_n(\kappa)$, and the two-dimensional shock-evolution problem is reduced to a one-dimensional problem.

Two points are worth noting. First, the $D_n(\kappa)$ relation only contains limited constitutive information about the explosive. The constants in this relation are integrals through the reaction zone of this information. Secondly, $D_n(\kappa)$ is independent of initial/boundary data. Therefore, when detailed constitutive information about the reaction zone is not known (the typical situation for condensed phase explosives), $D_n(\kappa)$ can be measured directly via simple steady-state two-dimensional hydrodynamic experiments. Thus we have a way of using simple experiments to calibrate the reaction-zone dynamics. In turn, the calibrated $D_n(\kappa)$ relation can be used to predict the shock evolution in complex explosive geometries.

Direct calculations of $D_n(\kappa)$ performed with the simple polytropic eos, show that $D_n(\kappa)$ is sensitive to the form of the rate law.[8] Calculations were done for two state-independent rates with different depletion forms; square-root depletion

$$\mathcal{R} = \sqrt{1-\lambda} \tag{25}$$

and simple depletion

$$\mathcal{R} = (1-\lambda) \quad . \tag{26}$$

The $D_n(\kappa)$ rule for Eq. (25) is

$$D_n = 1 - \alpha\kappa \quad , \tag{27}$$

while for Eq. (26) we have

$$D_n = 1 + \beta\kappa\ell n(\kappa) - \alpha\kappa \quad . \tag{28}$$

The constants α and β are not be confused with Whitham's curvilinear coordinates. Compacted into these two constants is everything that we need to know about the constitutive laws. D_{cj} is set to one. In the next section, we give a brief tutorial that describes how this theory can be applied to explosive engineering design problems.

III. Applications

a. *Chapman-Jouguet Wave*

The simplest time-dependent problem that can be done is the constant-velocity detonation or "Huygens" construction for a diverging detonation. For convenience we take $D_n = 1$. Equation (7) then becomes the simple nonlinear-wave equation for the shock angle (see Figure 3)

$$\phi_{,t} + (\phi - \phi_e)\phi_{,\xi} = 0 \quad , \tag{29}$$

where ϕ_e is the value of ϕ at the edge (i.e., at $\xi = 0$). Equation (29) states that ϕ = constant along the characteristic lines $\xi - (\phi - \phi_e)t$ = constant, that is

$$\phi = \phi_o \qquad \text{along} \qquad \xi - (\phi_o - \phi_e)t = \xi_o \quad . \tag{30}$$

If we consider a flow where the two-dimensional shock is convergent initially, then the initial angle, ϕ_o, is a decreasing function of the initial arc length, ξ_o. Such a flow looks compressive, in the sense that the characteristic lines are convergent. After a finite time, some of the characteristics cross one another and the solution becomes multi valued. Physically, the rule $D_n = 1$ does not apply to a convergent detonation, so we will not consider this case further.

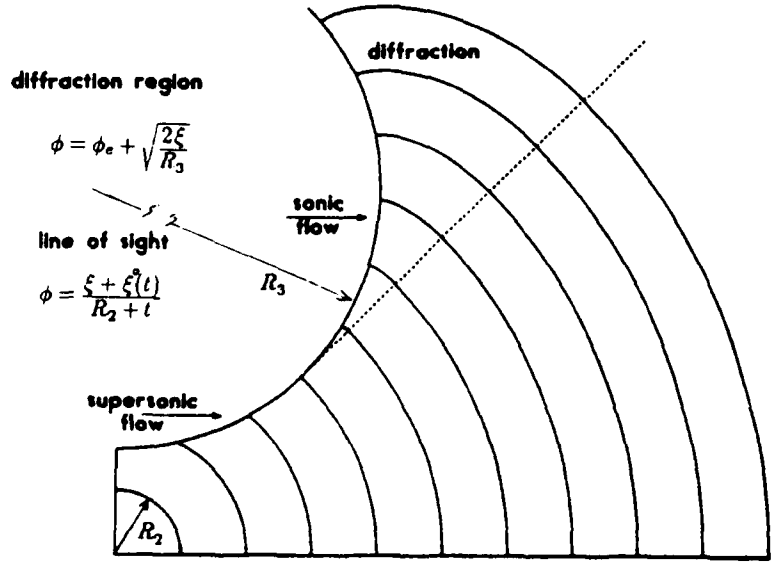


Figure 4. A prototypical diverging detonation problem. The wave is propagated with $D_n = 1$, a "Huygens" construction. Below the dashed line, the wave is free of boundary effects and expands as a circle. Above the dashed line, the wave shape is determined by applying the sonic condition along the radius R_3 circular edge.

When the two-dimensional shock is initially divergent, the initial angle is an increasing function of arc length, and the characteristic lines are rarefaction like. An example of a divergent-wave

problem that is often encountered in designs is shown in Figure 4. It is a prototypical example of a diverging detonation that features the diffraction of the detonation (i.e., the "shadow zone" problem). The left-most vertical line is a symmetry plane; the lower horizontal line and the upper circular arc are the edges of the explosive. The wave is initially circular with a radius R_2. Since the wave is perpendicular to the horizontal edge, the flow along that edge/shock locus is sonic, and the edge does not influence the shock evolution. When the expanding wave first reaches the circular boundary, the flow along the upper edge/shock locus is supersonic. It remains supersonic until the detonation reaches the point where the dashed line is tangent to the arc. The region above the dashed line is not in direct line of sight of the initial data; it is a "shadow zone." Diffraction is the process that allows the wave to spread into this region. The solution in this region is determined by the boundary data supplied along the circular edge.

In both regions of the problem, the solution takes a simple form. The great advantage of our formulation over older methods is this simplicity of representation. The calculations shown in Figure 4 are free of reaction-zone effects. We conclude this section by showing how detonation shock dynamics can be used to include the important finite size reaction-zone effects for this example.

b. *DSD Wave*

We assume that the reaction-zone dynamics is given by Eq. (27)

$$D_n = 1 - \alpha\kappa$$

and introduce the change of variable

$$\phi = \phi_e + \tilde{\phi} \quad , \tag{31}$$

where ϕ_e is the angle that the tangent to the edge makes with the reference direction $\hat{e}_x$. Substituting these into the kinematic equation [i.e., Eq. (27)], yields a "Burgers" equation

$$\tilde{\phi}_{,t} - \frac{R_2 D_{ne}}{R_3 \cos(\tilde{\phi}_e)} + B\tilde{\phi}_{,x} = \frac{\alpha}{R_2}\tilde{\phi}_{,xx} \quad , \tag{32}$$

as the propagator for the shock. The independent variables in Eq. (32) are scaled time (t) and scaled arc length (x). The finite length reaction-zone effects enter this equation as the transport term on the right-hand side. This is similar to the structure of wave-hierarchy problems that arise in one-dimensional wave propagation problems in reactive materials.[9] The second term on the left-hand side represents the diffraction effect. Equation (32) is a one-dimensional parabolic pde. Thus in the quasi-steady limit, the reactivity acts to smooth the shock locus.

Equation (32) was solved numerically for the design problem shown in Figure 4. A mesh was used with one thousand points along the shock. The computation time was one minute on the Cray-1 supercomputer. The results of the wave tracking calculation for a set of parameter values that highlight the finite-length reaction-zone effects are shown in Figure 5.

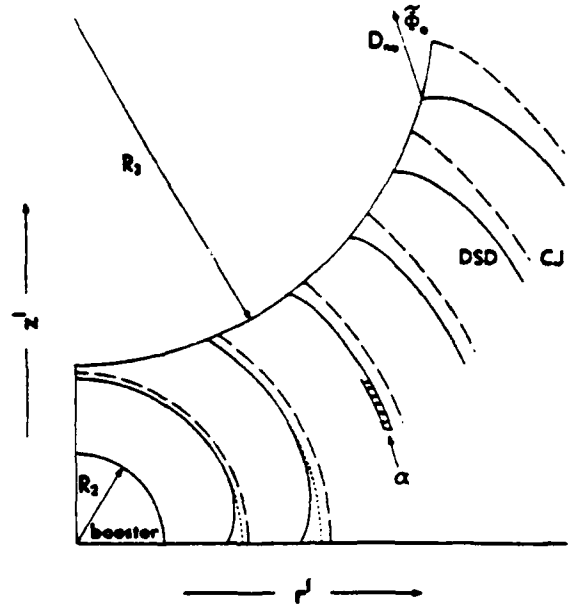


Figure 5. The DSD calculation of the example considered in Figure 4. The reaction-zone dynamics rule was $D_n = 1 - \alpha\kappa$, where the magnitude of α is shown. Three calculations are displayed: (— - — - — - —) $D_n = 1$ "Huygens." (. . . . .) $D_n = 1 - \alpha\kappa$ circularly expanding wave and (————) the full DSD calculation.

The important parameters in this calculation are (α/R_2), the ratio of the reaction-zone length parameter to the radius of the booster, and (R_2/R_3) the ratio of the booster to the edge radius. The dashed contours correspond to the standard "Huygens" construction studied in Figure 4. The dotted contours show the cylindrically expanding finite-length reaction-zone wave without any edge effect. The solid contours show the complete DSD calculation, including the edge effects. Although the results shown in the figure speak well for themselves, a few comments are in order. Even in regions of the flow that are not influenced by the edge, the finite-length reaction-zone effects cause the detonation to lag behind the "Huygens" wave. Near the lower edge, the complete DSD calculation is strongly curled back. Along this edge, the phase velocity of the detonation wave is initially low, but as time passes it builds back to that for a cylindrically expanding wave. Along the upper surface, no edge effect is observed until the detonation wave passes into the "shadow zone." After this occurs, the detonation wave is continually undergoing wave diffraction. Since the phase velocity at the edge quickly reaches a steady value that is well below D_{cj}, the curl back is more pronounced in this region than at the lower edge. The value of this velocity is a function of the radius of the upper explosive/vacuum interface.

IV. Summary

We have developed a theory for propagating two-dimensional detonation shocks in complex explosive assemblies. The three components of our method are:

(1) shock kinematics [Eq. (7)],

(2) boundary conditions [Eq. (11)], and

(3) reaction-zone dynamics [e.g., Eq. (27)].

In spirit it is the detonation analog of Whitham's inert shock propagation theory, geometrical shock dynamics. It is a rationally derived theory that applies when the radius of curvature of the detonation shock is large compared to the reaction-zone length. A fully nonlinear theory, it describes the large amplitude changes in the two-dimensional detonation shock that occur over long times.

The DSD method that we have developed is a powerful tool that can be used to efficiently model reaction-zone effects in numerical simulations of detonation. Using this method, a model explosive design calculations was performed with about one minute of supercomputer time. This is to be compared to the many hours that are required for modest resolution full numerical simulations of explosive assemblies. In addition to the direct computational benefit, this theory also increases our understanding of time-dependent two-dimensional detonation. For example, this theory defines the relationship between the detonation wave phase velocity and the radius of the explosive edge in the "shadow zone."

Acknowledgments

We would like to thank W. C. Davis and Wildon Fickett for many useful and stimulating discussions concerning this work.

References

1. Fickett, Wildon, "*Accuracy of the Conventional Lagrangian Scheme for One-Dimensional Hydrodynamics*," Los Alamos National Laboratory report LA-6454 (1976).

2. Woodward, P. and Colella, P., "*The Numerical Simulation of Two-Dimensional Fluid Flow with Strong Shocks*," J. Comput. Phys. **54**, 115–73 (1984).

3. Addessio, F., "*CAVEAT: A Computer Code for Fluid Dynamics Problems with Large Distortion and Internal Slip*," Los Alamos National Laboratory report LA-10613-MS (1986).

4. Glimm, J., Isacson, E., Marchesin, D., and McBryan, O., "*Front Tracking for Hyperbolic Systems*," Adv. Appl. Math **2**, 91–119 (1981).

5. Whitham, G., *Linear and Nonlinear Waves*, (Wiley-Interscience, New York, 1974), pp. 281–4.

6. Bdzil, J. B. and Stewart, D. S., "*Time-Dependent Two-Dimensional Detonation: The Interaction of Edge Rarefactions with Finite Length Reaction Zones*," J. Fluid Mech. **171**, 1–26.

7. Fickett, Wildon and Davis, W. C., *Detonation*, Univ. of California Press, Berkeley (1979), pp. 199–219.

8. Stewart, D. S. and Bdzil, J. B., "*The Shock Dynamics of Stable Multi-Dimensional Detonation*," Comb. Flame **72**, 311–23 (1988).

9. Whitman, G., *Linear and Nonlinear Waves*, (Wiley-Interscience, New York, 1974), pp. 339–59.

Reactive-Euler Induction Models

J. Bebernes and D. Kassoy
University of Colorado
Boulder, CO 80309

ABSTRACT. A unified formulation for the induction period for all thermal reaction problems is presented using high activation energy asymptotics. The important parameters in the nondimensional equations are the ratios of characteristic reaction, acoustic, and conduction times in the thermally disturbed parcel of a reactive gas of dimension L. In larger systems transport effects are negligible and the induction period is controlled by reactive gasdynamics equations. Two of these models are analyzed.

1. INTRODUCTION. The evolution of thermal explosions in gaseous systems depends on the interaction between chemical heat release, conductive thermal losses, and the effects of compressibility. The latter factor can accelerate reaction rates in constant volume systems where compression heating plays a role [1], [2], [7]. In unconfined systems however, the conversion of some thermal energy to kinetic energy may retard the appearance of thermal runaway. Systems in which conductive losses are unimportant will inevitably explode, perhaps faster than diffusive systems. In this sense it is important to be able to predict which physical processes control the evolution of an exothermic reaction in a specific gaseous system. In this paper, we present the results of our recent studies which provide a rational basis for deciding the correct induction model for the given physical system and analyze these models mathematically.

Consider a reactive viscous heat conducting compressible gas in an equilibrium state defined by the dimensional quantities $p_0 = p(x,0)$, $\rho_0 = \rho(x,0)$, $T_0 = T(x,0)$, $y_0 = y(x,0)$, and $u_0 = u(x,0)$ which represent pressure, density, temperature, concentration, and velocity, respectively.

At time $t = 0$, assume a small initial disturbance is created on a length scale L. Define $\bar{x} = x/L$ as the new position vector. Let $\bar{t} = t/t_R$ be the new time scale where t_R is a reference time to be determined later. Nondimensionalize the system variables, letting $\bar{p} = p/p_0$, $\bar{\rho} = \rho/\rho_0$, $\bar{T} = T/T_0$, $\bar{y} = y/y_0$, and $\bar{u} = u/(L/t_R)$. Assume a single one-step irreversible reaction which has a rate law described by Arrhenius kinetics. The complete combustion system can then be written in nondimensional form, where the bar notation has been dropped, as:

$$
(1)\qquad \begin{cases} \rho_t + \nabla\cdot(\rho u) = 0 \\ \rho(u_t + u\cdot\nabla u) = -\frac{1}{\gamma}(\frac{t_R}{t_A})^2\nabla p + \mathrm{Pr}(\frac{t_R}{t_c})\mu[\Delta u + \frac{1}{3}\nabla(\nabla\cdot u)] \\ \rho c_v(T_t + u\cdot\nabla T) = \gamma(\frac{t_R}{t_c})\nabla\cdot(k\nabla T) - (\gamma-1)p(\nabla\cdot u) \\ \qquad\qquad +2\mu\left[\frac{\gamma(\gamma-1)\mathrm{Pr}t_A^2}{t_R t_c}\right]\left[D:\nabla\otimes u - \frac{1}{3}(\nabla\cdot u)^2\right] \\ \qquad\qquad +t_R B\rho y\exp[-\frac{1}{\epsilon T}] \\ \rho(y_t + u\cdot\nabla y) = \mathrm{Le}(\frac{t_R}{t_c})\nabla\cdot(\rho D\nabla y) - t_R B\rho y\exp(-\frac{1}{\epsilon T}) \\ p = \rho T \end{cases}
$$

where $\bar{\mu} = \mu/\mu_0$, $\bar{D} = D/D_0$, $\bar{c}_p = c_p/c_{p_0}$, $\bar{c}_v = c_v/c_{v_0}$, $\bar{k} = k/k_0$, and $\bar{K} = K/K_0$ where $K = k/\rho c_p$ is the thermal diffusivity, c_p and c_v are the specific heats. Also, $\gamma = c_{p_0}/c_{v_0}$ is the gas

parameter, $\epsilon = RT_0/E$ is the nondimensional inverse of the activation energy, $\Pr = c_{p_0}\mu_0/k_0$, the Prandtl number, $\text{Le} = D_0/K_0$ the Lewis number, $c_0 = (\gamma RT_0)^{1/2}$ the initial sound speed, $t_A = L/c_0$ the acoustic time scale, $t_c = L^2/K_0$ the conduction time scale and $\bar{h} = hy_0/c_{v_0}T_0$ is the nondimensional heat of reaction.

2. INDUCTION PERIOD MODELS. As in [8], assume that $\Pr = O(1)$, $\text{Le} = O(1)$, $h = O(1)$ and that $\epsilon \ll 1$. Using the method of activation energy asymptotics, we seek simpler models of the combustion process. In (1c,d), the reaction terms contain an expression of the form $\exp(-\frac{1}{\epsilon T})$. For $\epsilon \ll 1$, an induction period theory can be described in terms of the perturbed variables

$$\text{(2)} \qquad \begin{cases} \rho = 1 + \epsilon m \quad p = 1 + \epsilon P \quad T = 1 + \epsilon\theta \\ u = \epsilon v \qquad y = 1 - \epsilon c \end{cases}$$

where we assume that the initial temperature disturbance is $O(\epsilon)$. If $O(\epsilon)$ terms are ignored, we obtain the induction model for a gaseous system from (1) using (2):

$$\text{(3)} \qquad \begin{cases} m_t + \nabla \cdot v = 0 \\ v_t = -\frac{1}{\gamma}\left(\frac{t_R}{t_A}\right)^2 \nabla P + \Pr\left(\frac{t_R}{t_c}\right)\mu[\Delta v + \frac{1}{3}\nabla(\nabla \cdot v)] \\ \theta_t = t_R B h \epsilon^{-1} e^{1/\epsilon} e^{\theta} + \gamma\left(\frac{t_R}{t_c}\right)\Delta\theta - (\gamma - 1)\nabla \cdot v \\ \qquad +2\gamma(\gamma-1)\frac{t_A^2}{t_R \cdot t_c}\epsilon \Pr\mu\left[-\frac{1}{3}(\nabla \cdot v)^2 + \{\nabla \otimes v + (\nabla \otimes v)^T\} : \nabla \otimes v\right] \\ c_t = t_R B \epsilon^{-1} e^{-1/\epsilon} e^{\theta} + \text{Le}\left(\frac{t_R}{t_c}\right)\Delta c \\ P = m + \theta \end{cases}$$

The induction model (3) contains three time scales t_R, t_A, and t_c which depend on the particular thermochemical system with the reference time t_R yet to be specified. The character of the induction models depends intimately on the ratios formed from these three time scales. We will consider initial temperature disturbances on a macroscopic length scale so that $t_A/t_c \ll 1$. If we assume that the perturbation temperature θ and the concentration c variations are caused by the chemical reaction process, then for ϵ small there should be a balance of the accumulation terms θ_t and c_t in (3) with the reaction terms involving e^{θ}. It follows that the reference time can be defined by

$$t_R = \frac{\epsilon e^{1/\epsilon}}{B} \tag{4}$$

which represents the chemical time for a reaction initiated at T_0 multiplied by ϵ. The three time scales are now completely defined and the reduced induction models depend on their ratios.

The first case to be considered is that for

$$\text{(5)} \qquad \frac{t_R}{t_A} \equiv a = O(1)$$

then the induction momentum, energy, and species equations of (3) can be written as

$$\text{(6)} \qquad \begin{cases} \left(\frac{t_A}{t_c}\right)^2 v_t = -\frac{a^2}{\gamma}\nabla P + a\left(\frac{t_R}{t_c}\right)^2 \Pr\mu\left[\Delta v + \frac{1}{3}\nabla(\nabla \cdot v)\right] \\ \theta_t = he^{\theta} + a\gamma\Delta\theta - (\gamma - 1)\nabla \cdot v \\ c_t = e^{\theta} + \text{Le}a\Delta c \end{cases}$$

Since we are assuming that the initial disturbances are spatially macroscopic so that $t_A/t_c \ll 1$, we have from the inductive momentum equation (6a) that $P = P(t)$ to a first approximation. Combining the mass equation (3a) and the energy equation (6b),

$$\theta_t = \frac{h}{\gamma}e^\theta + a\Delta\theta + \frac{\gamma-1}{\gamma}P'(t) \tag{7}$$

For a bounded container Ω since the total mass must be conserved, $\int_\Omega \rho(x,t)dx = \mathrm{vol}\,(\Omega)$ which implies $\int_\Omega m(x,t)dx = 0$ and hence

$$P(t) = \frac{1}{\mathrm{vol}\,\Omega}\int_\Omega \theta(x,t)dx.$$

We can thus rewrite (7) as

$$\theta_t - a\Delta\theta = \delta e^\theta + \frac{\gamma-1}{\gamma}\frac{1}{\mathrm{vol}\,\Omega}\int_\Omega \theta_t(x,t)dx \tag{8}$$

and impose initial-boundary conditions of the type

$$\begin{cases} \theta(x,0) = \theta_0(x), & x \in \Omega \\ \theta(x,t) = 0, & (x,t) \in \partial\Omega \times (0,\infty) \end{cases} \tag{9}$$

This model (8)–(9) with the last term representing the effects of spatially homogeneous gas compression was originally derived by Kassoy and Poland [7] and was analyzed in [1].

If the ratio $t_R/t_c \equiv a \ll O(1)$ so that the reaction time is much shorter than the conduction time, then (3b,c,d) can be written

$$\begin{aligned} v_t &= -\frac{1}{\gamma}\frac{(t_R/t_c)^2}{(t_A/t_c)^2}\nabla P + \mathrm{Pr}a\mu\left[\Delta v + \frac{1}{3}\nabla(\nabla\cdot v)\right] \\ \theta_t &= he^\theta - (\gamma-1)\nabla\cdot v + a\gamma\Delta\theta + 2\gamma(\gamma-1)\mathrm{Pr}\epsilon\mu\frac{(t_A/t_c)^2}{a}\cdot \\ &\qquad \left[-\frac{1}{3}(\nabla\cdot v)^2 + \{\nabla\otimes v + (\nabla\otimes v)^T\} : \nabla\otimes v\right] \\ c_t &= e^\theta + \mathrm{Le}\cdot a\cdot\Delta c. \end{aligned} \tag{10}$$

Because $a = o(1)$, viscous, conductive, and diffusive effects are weak. Three subcases are of interest, all of which lead to *reactive-Euler explosions.*

I) For $t_A \ll t_R \ll t_c$, then from (10a) $P = P(t)$ to a first approximation and the energy equation becomes

$$\begin{aligned} \theta_t &= \frac{h}{\gamma}e^\theta + \frac{\gamma-1}{\gamma}P'(t) \\ &= \frac{h}{\gamma}e^\theta + \frac{\gamma-1}{\gamma}\frac{1}{\mathrm{vol}\,\Omega}\int_\Omega \theta_t(x,t)dx \end{aligned} \tag{11}$$

where Ω is a bounded container.

II) For $O(t_R) = t_A \ll t_c$, to first order the momemtum equation (10a) becomes

$$v_t = -\frac{1}{\gamma}\frac{a^2}{(t_A/t_c)^2}\nabla P \tag{12}$$

and (3) reduces to

$$\theta_t - \frac{\gamma - 1}{\gamma}P_t = \frac{h}{\gamma}e^{\theta}$$

$$v_t + \frac{1}{\gamma}\frac{a^2}{(t_A/t_c)^2}\nabla P = 0 \tag{13}$$

$$\nabla \cdot v + \frac{1}{\gamma}P_t = \frac{h}{\gamma}e^{\theta}$$

III) For $t_R \ll t_A \ll t_c$, (10a) reduces to $v_t = 0$ or $v = v(x)$. This implies th inertial confinement of the heated gas is dominant. Aspects of short time inertial confinement have been discussed by Clarke et al. [3], Dold [4], and Jackson et al. [5], [6].

3. The First Reactive-Euler Model. For an arbitrary bounded container $\Omega \subset I\!R^N$, the reactive-Euler model (11) can be written as

$$\phi_t = \delta e^{\phi} + \frac{\gamma - 1}{\gamma}\frac{1}{\mathrm{vol}\,\Omega}\int_{\Omega}\phi_t(x,t)dx \tag{14}$$

with

$$\phi(x,0) = \phi_0(x) \tag{15}$$

assuming $\phi_0(x)$ is continuous and bounded on Ω. By integrating (14) over Ω, we see that (14) is equivalent to

$$\phi_t = \delta e^{\phi} + \beta\int_{\Omega} e^{\phi}dx \tag{16}$$

where $\beta = \frac{(\gamma-1)\delta}{\mathrm{vol}\,\Omega}$.

The IBVP (16)–(15) has a unique nonextendable solution $\phi(x,t)$ on $\Omega \times [0,\sigma)$ where $\sigma = +\infty$ or $\sigma < \infty$ with $\lim_{t\to\sigma^-}\sup\{\phi(x,t) : x \in \Omega\} = \infty$.

The inital value problem

$$a' = \delta e^{a}, \qquad (x,t) \in \Omega \times (0,T) \tag{17}$$

$$a(x,0) = \phi_0(x), \qquad x \in \bar{\Omega} \tag{18}$$

has the explicit solution

$$a(x,t) = -ln[e^{-\phi_0(x)} - \delta t] \tag{19}$$

which blows up in finite time $T = \delta^{-1}\exp(-\phi_0(x_m))$ where x_m is any point in Ω at which $\phi_0(x)$ attains its absolute maximum. Since $a(x,t)$ is a lower solution for (16)–(15), the solution $\phi(x,t)$ satisfies

$$\phi(x,t) \geq -ln[e^{-\phi_0(x)} - \delta t]$$

and hence $\phi(x,t)$ blows up in finite time σ with $\sigma \leq T$.

To get more information about $\phi(x,t)$, consider the implicit representation

$$\phi(x,t) = a(x,\tau(t)) + B(\tau(t)) \tag{20}$$

where $a(x,\tau)$ is the solution of (17)–(18) and $\tau(t), B(\tau)$ are scalar functions to be determined. As given in (20), $\phi(x,t)$ is a solution of (16)–(15) if and only if

$$\tau' = e^{B(\tau)}, \qquad \tau(0) = 0 \tag{21}$$

$$B' = \beta\int_\Omega e^{a(x,\tau)}dx = \beta\int_\Omega [e^{-\phi_0(x)} - \delta\tau]^{-1}dx, \quad B(0) = 0 \tag{22}$$

By integrating (22), (21) can be solved by quadrature to get

$$B(\tau) = \frac{\beta}{\delta}\int_\Omega [a(x,\tau) - \phi_0(x))]dx = \frac{\beta}{\delta}\int_\Omega ln\left[\frac{e^{-\phi_0(x)}}{e^{-\phi_0(x)} - \delta\tau}\right]dx \tag{23}$$

and τ satisfies

$$\tau' = \exp\left[\frac{\beta}{\delta}\int_\Omega ln\left[\frac{e^{-\phi_0(x)}}{e^{-\phi_0(x)} - \delta\tau}\right]dx\right], \quad \tau(0) = 0 \tag{24}$$

which can be solved by quadrature.

From (20), we thus have

Theorem 1. The number σ is the blowup time for the solution $\phi(x,t)$ of (14)–(15) if and only if $\tau(\sigma) = T$ is the blowup time for the solution $a(x,\tau)$ of (17)–(18), and thus $\sigma = \tau^{-1}\left(\frac{1}{\delta}e^{-\phi(x_m)}\right)$ where x_m is any point in Ω at which ϕ_0 has an absolute maximum.

By considering (20) and (23), we can observe that $\phi(x,t)$ blows up at those points x_m at which $\phi_0(x)$ has its absolute maximum provided that $B(\tau(\sigma)) < \infty$. This is true if and only if $\int_\Omega a(x,\tau(\sigma))dx < \infty$ which in turn is true provided that $\int_\Omega ln[e^{-\phi_0(x)} - e^{-\rho_0(x_m)}]dx > -\infty$. Similarly, $\phi(x,t)$ blows up everywhere in Ω at σ if and only if $B(\tau(\sigma)) = \infty$. Thus,

Theorem 2 (a) The solution $\phi(x,t)$ of (14)–(15) blows up only at those points x_m of Ω at which $\phi_0(x)$ has its absolute maximum if and only if

$$\int_\Omega ln[e^{-\phi_0(x)} - e^{-\phi_0(x_m)}]dx > -\infty. \tag{25}$$

(b) The solution $\phi(x,t)$ blows up everywhere in Ω at σ if and only if

$$\int_\Omega ln[e^{-\phi_0(x)} - e^{-\phi_0(x_m)}]dx = -\infty. \tag{26}$$

The integral in (25) is finite if there is at most a finite number of critical points $x_m \in \Omega$ at which ϕ_0 has an absolute maximum and if at each x_m $\phi_0(x)$ is strictly concave down and analytic in a neighborhood of x_m. In this case, blowup occurs only at those x_m at which ϕ_0 has an absolute maximum. If on the other hand ϕ_0 is too flat in a neighborhood of an x_m, then blowup occurs everywhere in Ω.

A second method for representing the solution $\phi(x,t)$ of (14)–(15) is to set

$$\Phi(x,t) = \phi(x,t) - \beta \int_\Omega \phi(x,t)dx \tag{27}$$

Then Φ satisfies

$$\Phi_t = \alpha F_t e^{\Phi} \tag{28}$$

with

$$\Phi(x,0) = \phi_0(x) - \beta \int_\Omega \phi_0(x)dx \tag{29}$$

where

$$F_t = e^{\beta \int_\Omega \phi(y,t)dy}, \qquad F(0) = 0. \tag{30}$$

By integrating (28) and using (30), we find that $\phi(x,t)$ can be expressed as

$$\phi(x,t) = \frac{\gamma - 1}{\text{vol}\,\Omega} \int_\Omega ln\frac{1}{G}dy + ln\frac{1}{G} \tag{31}$$

where $G(x,t) = ke^{-\phi_0(x)} - \alpha F(t)$, $k = e^{\beta \int \phi_0(y)dy}$. Note then that the blowup time σ for ϕ is given from (31) by $F(\sigma) = \frac{ke^{-\phi_0(x_m)}}{\delta}$.

Since $P_0(t) = \frac{1}{\text{vol}\,\Omega} \int_\Omega (\phi(x,t) - \phi_0(x))dx$, we have from (31)

$$P_0(t) = \frac{\gamma}{\text{vol}\,\Omega} \int_\Omega ln\frac{1}{\sigma(y,t)}dy - \frac{1}{\text{vol}\,\Omega} \int_\Omega \phi_0(y)dy \tag{32}$$

From (31) and (32), we have

$$\phi(x,t) = \phi_0(x) + \frac{\gamma - 1}{\gamma} P_0(t) - ln(1 - \frac{\delta}{k} e^{-\phi_0(x)} F(t)) \tag{33}$$

from which we can conclude that the temperature evolves from the initial value $\phi_0(x)$ through a purely time dependent term related to the homogeneous pressure increase and a logarithmic evolution term with spatial dependence which has a shape-preserving property.

4. The Second Reactive-Euler Model. In one spatial dimension, the reactive-Euler model (13) can be written as

$$\begin{aligned} \phi_t - \frac{\gamma-1}{\gamma} P_t &= \delta e^{\phi} \\ v_t' + \frac{1}{\gamma}\left(\frac{a'}{t_A/t_c}\right)^2 P_x &= 0, \quad (x,t) \in \mathbb{R} \times (0,\infty) \\ v_x' + \frac{1}{\gamma} P_t &= \delta e^{\phi} \end{aligned} \tag{34}$$

with

$$\phi(x,0) = \phi_0(x), \quad P(x,0) = P_0(x), \quad v'(x,0) = v_0(x) \tag{35}$$

continuous bounded functions on $\mathbb{R}$. Setting $a = \frac{\gamma-1}{\gamma}$, $b = \delta$, $c = \frac{1}{\gamma}\left(\frac{a'}{t_A/t_c}\right)^2$, $d = \frac{1}{\gamma}$, then, for $w = \phi - aP$, (34) becomes

$$\begin{aligned} w_t &= be^{w+aP} \\ v_t' + cP_x &= 0 \\ P_t + \frac{1}{d} v_x' &= \frac{b}{d} e^{w+aP} \end{aligned} \tag{36}$$

with

$$w(x,0) = \phi_0(x) - aP_0(x), \quad v'(x,0) = v_0(x), \quad P(x,0) = P_0(x) \tag{37}$$

Using the change of coordinate matrix

$$T = \begin{pmatrix} 1 & 1 \\ -(cd)^{-1/2} & (cd)^{1/2} \end{pmatrix} \quad \text{and setting} \quad \begin{pmatrix} \bar{v} \\ \bar{P} \end{pmatrix} = T^{-1} \begin{pmatrix} v' \\ P \end{pmatrix}$$

we have

$$\begin{aligned} w_t &= be^{w+\mu(\bar{P}-\bar{v})} \\ \bar{v}_t - \lambda \bar{v}_x &= -\frac{b\lambda}{2} e^{w+\mu(\bar{P}-\bar{v})} \\ \bar{P}_t + \lambda \bar{P}_x &= \frac{b\lambda}{2} e^{w+\mu(\bar{P}-\bar{v})} \end{aligned} \tag{38}$$

where $\mu = \frac{a}{(cd)^{1/2}}$ and $\lambda = \left(\frac{c}{d}\right)^{1/2}$. Set $u = \mu\bar{P}$, $v = -\mu\bar{v}$, $A = b = \delta$, $B = \frac{\gamma-1}{2}\delta$, then

$$\begin{cases} w_t & = Ae^{w+u+v} \\ u_t + \lambda u_x & = Be^{w+u+v} \\ v_t - \lambda v_x & = Be^{w+u+v} \end{cases} \tag{39}$$

with

$$\begin{cases} w(x,0) & = \phi_0(x) - a'P_0(x) \equiv \bar{w}(x) \\ u(x,0) & = \frac{a}{2}[(cd)^{-1/2}v_0(x) + P_0(x)] \equiv \bar{u}(x) \\ v(x,0) & = -\frac{a}{2}[(cd)^{-1/2}v_0(x) = P_0(x)] \equiv \bar{v}(x). \end{cases} \tag{40}$$

Thus, we have shown that the reactive-Euler induction problem (34)–(35) is equivalent to the more symmetric problem (39)–(40). Problem (39)–(40) is closely related to the low frequency—mean field equations considered by Majda and Rosales in [10] and the disturbance equations considered by Jackson, Kapila, and Stewart in [6].

Let $c^+ = \max[A,B]$, $c^- = \min[A,B]$, $m^+ = \max\{\bar{w}(x),\bar{u}(x),\bar{v}(x)\}$, $m^- = \min\{\bar{w}(x),\bar{u}(x),\bar{v}(x)\}$ and consider

$$\begin{cases} z' &= c^{\pm}e^{3z} \\ z(0) &= m^{\pm}. \end{cases} \tag{41}$$

By comparison with (39)–(40)

$$ln[e^{-3\bar{m}} - 3c^-t]^{-1/3} \le \left\{ \begin{array}{c} w(x,t) \\ u(x,t) \\ v(x,t) \end{array} \right\} \le ln[e^{-3m^+} - 3c^+t]^{-1/3} \tag{42}$$

Hence, every solution (w,u,v) of (39)–(40) blows up in finite time with

$$\frac{1}{3c^+e^{3m^+}} \le T \le \frac{1}{3c^-e^{3m^-}} \tag{43}$$

Note that $\phi(x,t) = w(x,t)+u(x,t)+v(x,t)$. Assume henceforth that $A+2B=1$ and that $\phi(x,t)$ blows up at $x_m \in \mathbb{R}$ at time T. We would like to describe how the blowup singularity evolves at (x_m,T). Make the backward similarity change of variables

$$\tau = -ln(T-t), \quad \eta = \frac{x-x_m}{(T-t)^{1/2}} \tag{44}$$

with

$$\begin{aligned} W &= w + Aln(T-t) \\ U &= u + Bln(T-t) \\ V &= v + Bln(T-t) \\ S &= W+U+V = \phi + ln(T-t), \end{aligned} \tag{45}$$

then

$$\begin{cases} W_\tau + \frac{\eta}{2}W_\eta = A(e^S-1) \\ U_\tau + \frac{\eta}{2}U_\eta + \lambda e^{-\tau/2}U_\eta = B(e^S-1) \\ V_\tau + \frac{\eta}{2}V_\eta - \lambda e^{-\tau/2}V_\eta = B(e^S-1) \\ S_\tau + \frac{\eta}{2}S_\eta + \lambda e^{-\tau/2}(U_\eta - V_\eta) = e^S - 1 \end{cases} \tag{46}$$

To describe how the blowup singularity evolves would require analyzing the behavior of solutions of (46) as τ becomes infinite. To get an idea of what to expect or hope for, consider the easier problem when there is no drift, i.e., $\lambda = 0$. The temperature ϕ blows up at

$$T_\phi = e^{-\phi_0(x_m)}$$

where x_m is an absolute maximum point for ϕ_0. Then we know exactly when and where blowup occurs. We can also describe precisely how the singularity evolves. Let $z = \phi + ln(T-t)$, then z is the solution of

$$(47) \qquad \begin{cases} z_\tau + \frac{\eta}{2} z_\eta & = e^z - 1 \\ z(\eta, -lnT) & = z_0(\eta) = \phi_0(\eta T^{\nu_2} + x_m) + lnT \end{cases}$$

which can be explicitly solved to give

$$(48) \qquad z(\eta, \tau) = -ln[1 - e^\tau(1 - e^{-z_0(\eta e^{-\tau/2})})]$$

Thus,

$$lim_{\tau \to \infty} z(\eta, \tau) = -ln\left[1 - z_0''(0)e^{-z_0(0)}\frac{\eta^2}{2}\right]$$
$$= -ln\left[1 - \frac{\Omega}{2}\eta^2\right] \equiv \bar{z}(\eta)$$

From this, we conclude that for $\lambda = 0$

$$\phi(x,t) + ln[(T-t) - \Omega(x - x_m)^2] \to 0$$

uniformly for $(x - x_m)^2 \leq \eta(T-t)$ as $t \to T^-$ which gives a description of how the blowup evolves. We expect a similar type behavior for (39)–(40). This has been confirmed formally by [6] and [4].

References

1. Bebernes, J. and A. Bressan, Thermal Behavior for a Confined Reactive Gas, *J. Differential Equations*, 44 (1982), 118-133.

2. Bebernes, J. and D. Kassoy, Characterizing self-similar blowup, *Mathematical Modeling in Combustion and Related Topics*, NATO ASI Series, M. Nijhoff, Pub. (1988), 383-392.

3. Clarke, J. F. and R. S. Cant, Nonsteady gasdynamic effects in the induction domain behind a strong shock wave, *Dynamics of Flames and Reactive Systems, Progress in Aeronautics and Astronautics*, 95 (1985), 142-163.

4. Dold, J. W., Dynamic transition of a self-igniting region, *Mathematical Modeling in Combustion and Related Topics*, C-M. Brauner and C. Schmidt-Lainé, Eds., NATO ASI Series, M. Nijhoff, Pub. (1988), 461-470.

5. Jackson, T. L. and A. K. Kapila, Shock-induced thermal runaway, *SIAM J. Appl. Math.* 45 (1985), 130-137.

6. Jackson, T., A. K. Kapila, and D. S. Stewart, Evolution of a reaction center in an explosive material, *Univ. of Illinois T & A.M.*, Report No. 484, February 1987.

7. Kassoy, D. and J. Poland, The Induction Period of a Thermal Explosion in a Gas Between Infinite Parallel Plates, *Combustion and Flame*, 50 (1983), 259-274.

8. D. R. Kassoy, A. K. Kapila, and D. S. Stewart, A unified formulation for diffusive and nondiffusive thermal explosion theory, *Comb. Sci. Tech.*, to appear.

9. Kassoy, D., N. Riley, J. Bebernes, and A. Bressan, The confined nondiffusive thermal explosion with spatially homogeneous pressure variation, *Comb. Sci. Tech.*, to appear.

10. Majda, A. and R. Rosales, Wave interactions in the induction-zone, *SIAM J. Appl. Math.* 47 (1987), 1017-1039.

NONBLOWUPS, PERIODICITIES, VORTEX SHREDDINGS IN COMBUSTION AND HYDRODYNAMIC FLOWS: A CONFERENCE REPORT

Karl Gustafson
Ed Ash[1]
Brian Eaton[2]
Kadosa Halasi[3]
Robert Leben[4]

Department of Mathematics, University of Colorado
Boulder, Colorado 80309-0426

ABSTRACT

We report here on some of my recent work with my Ph.D. students (the co-listed authors) on current bifurcation problems in Combustion Theory, Fluid Flow, and Aerodynamics. Our approach has been both computational and analytical. Although much work remains to be done, the results presented here are new and the sharpest to date.

1. Counting the Number of Solutions in Reactive Flow Problems.

The nonlinear elliptic partial differential equation

$$-\Delta u = \lambda e^{\frac{u}{1+\epsilon u}} \tag{1.1}$$

has been of considerable interest in Combustion Theory. In it, u represents a temperature in a self-heating body Ω near explosion, λ represents the lump exothermicity of the substance under consideration, and ϵ^{-1} is the activation energy. Equation (1.1) may possess anywhere from zero

[1]Current address: Dept. of Mathematics, Univ. of Colorado, Boulder, CO 80309-0426.

[2]Current address: NCAR, P.O. Box 3000, Boulder, CO 80307.

[3]Current address: Dept. of Mathematics, Kansas State Univ., Manhattan, KS 66506.

[4]Current address: CCAR, Univ. of Colorado, Boulder, CO 80309-0431.

to an infinite number of solutions, depending on the values of the parameters λ and ϵ under consideration. The determination of the exact number of solutions is of importance to the associated reactive flow problems.

Although physically a number of boundary conditions are relevant, here we shall restrict attention to the (usual) case of homogeneous Dirichlet boundary conditions $u = 0$ on the boundary $\partial\Omega$. Also we will consider here only the so-called type A geometries (spherical). The particular case of $n = 3$ dimensions is physically the most important, and the results given here will be for that case. The physical interest is in the case when all of u, λ, and ϵ are nonnegative.

The results to be presented here for $\epsilon > 0$ will be published in more detail in Ash, Eaton, and Gustafson [3]. For $\epsilon = 0$ the equation (1.1) has a long, varied, and distinguished history, found in the literature under the names Liouville, Poincaré, Bratu, Frank-Kamenetskii, Gelfand, Chandrasekhar, among others. See Gustafson [2] for a full historical account, including an exposition of Bratu's original work on the equation. In [2] many references to other recent work on this problem may be found, and we will not repeat them here. For our initial numerical work for the calculation of critical bifurcation points for equation (1.1) for $\epsilon > 0$, see Eaton and Gustafson, [1].

1.1 No Blow Up.

Qualitatively, the case $\epsilon = 0$ (which we call the Bratu approximation) and the case $\epsilon > 0$ (which we call the full Arrhenius equation) are fundamentally different. For $\epsilon = 0$ there exists a critical λ_c beyond which nonsingular solutions do not exist. On the other hand, for $\epsilon > 0$, solutions always exist for all positive λ. One way to view this situation is that the act of approximation (taking $\epsilon = 0$, i.e., taking activation energy $\alpha = \epsilon^{-1}$ to be infinite) introduces

a singularity into the problem. From this view, the singularity is artificial and should not be confused with solution explosion.

On the other hand, our recent numerical and analytical results [3] show, for very small ϵ, a pronounced tendency to a δ-function like nearly singular solution profile centered on a point in Ω (the center, for spherical geometries) at which very high temperature is concentrated. Remarkably, the λ_δ at which this occurs is near the next to the last turning point of the bifurcation diagram, rather than near the first turning point as one may have imagined from an $\epsilon = 0$ analysis.

1.2 The Last Turning Point.

Precise calculation of the last turning points of the bifurcation diagrams for equation (1.1) is difficult, both numerically and analytically, for small $\epsilon > 0$.

Figure 1 here, taken from [3], shows the exact bifurcation diagram for (1.1) for $\epsilon = 0.04$ (i.e., activation energy, $\alpha = 25$). The 6 turning points are so labelled on the curve. To compute this critical branching curve it is more convenient, following our approach of [1], to plot the bifurcation parameter vertically, rather than horizontally, as is usually done.

The solution u was found to be closest to a δ-function profile near the 5^{th} turning point. At the 6^{th} (and last) turning point, which is in the "noise level" along the horizontal axis in Figure 1, the solution profile snaps back to one very close to that of solutions along the leftmost first (stable) branch. Thereafter, although it cannot be seen from Figure 1, the final curve (stable) branch slowly rises and eventually increases to provide solutions for all λ.

The numerical scheme [1,2,3] that provides these results is called [2] HOC (Higher Order Calculus) inasmuch as it involves further implicit differentiation of the equations. This scheme provides an enlarged system, in some ways resembling the so-called inflation methods. In [3] we

also employ a rescaling trick which greatly increases the efficiency (shooting with only one iteration) over that of the original scheme in [1].

Our latest computations [3] have resolved the $\epsilon = 0.01$ case (unresolved in [2]). In this case there are 34 turning points, the last occurring at $\lambda \sim 10^{-38}$. This means that up to 35 solutions may occur for certain λ. See [3].

1.3 A Comparison Theorem.

Analytical lower bounds for the last turning point have been derived [3] using comparison techniques. Their proofs depend on and are motivated by the numerical procedures of the HOC scheme of [1,2,3]. One of them is the following:

$$\frac{d\lambda}{d\,(\text{Norm u})} \neq 0 \quad \text{whenever} \quad \pi^2 > \mu_2\,(\text{Norm U})\,,$$

where

$$\mu_2\,(\text{Norm u}) = \lambda\,(\text{Norm U}) \cdot 4\epsilon^2 e^{\frac{1-2\epsilon}{\epsilon}}\,.$$

For the case $\epsilon = 0.04$ this analytical bound estimates a lower bound for the last turning point to be: $\lambda_b = 1.57 \times 10^{-7}$. The numerically computed value of λ at the last turning point (see Table 2) was $\lambda = 2.44 \times 10^{-7}$. This is a very favorable comparison, and indicates the general comparison method we have used is a good one.

It would be very interesting and valuable to further investigate the general use of these numerical and analytical HOC methods on other reactive flow problems and in particular to study the implications of these results to Combustion Theory. For example, the presence of two (low and high temperature) stable branches for λ greater than a (very small) last turning point indicates an interpretation of explosion as a solution jump rather than singularity. Also there are interesting basic stiffness questions arising in the computations that need more under-

standing.

2. Vortex Dynamics in Cavity Flows

In Gustafson and Halasi [4] and [5] an in-depth study of lid driven cavity flow was carried out. The emphasis was on following the full dynamics of the unsteady (time-dependent) flow from an impulsive start. The full (viscous, incompressible) two-dimensional Navier-Stokes equations

$$u_t - \frac{1}{Re}\,\Delta u + (u\cdot\nabla)u = -\nabla p$$

$$\nabla\cdot u = 0 \tag{2.1}$$

were simulated under a MAC (marker and cell) primitive variable (velocity v and pressure p) discretization in which considerable care was given to maintaining correct incompressibility and pressure conditions near the boundary $\partial\Omega$ of the cavity Ω.

See [4] for a full accounting of previous work on this basic fluids problem, a fundamental geometry for the study of the effect of domain closure and corners on evolving fluid dynamics. In [5] the tolerance to varying grid size at $Re = 2000$ was determined and then a single long run of 360,000 time steps was carried out in a depth $A = 2$ cavity for the relatively high Reynolds number $Re = 10{,}000$.

From [5] it appeared that we had obtained a periodic solution, and hence had gone past a Hopf bifurcation at some Reynolds number between $Re = 2000$, where the solution became steady, and $Re = 10{,}000$, where it did not.

However, in recently writing the review chapter, Gustafson [6], I looked more closely at the results of the long run of [5] and came away with a different conclusion: tentatively, I will

assert that I found Feigenbaum's constant $\delta = 4.669201$ in the final oscillations. More details will be given in [6] but I will explain the finding briefly, in **2.3** below.

2.1 Computational Reliability

It is rather astounding, to those of us who started on a Royal McBee LGP-30 Machine (drum memory, 4096 words, electronic tubes failing all the time - but when used as an excuse for a coding error, tube failure was seldom the case!), that we may now routinely expect to do a Poisson Solver on a 40 by 80 rectangular grid 360,000 times without an interruption or logical error in the computation. Such is, however, the case these days.

Given this electronic reliability, we chose to use an extremely stable method (MAC) in the natural variables p and v. Our goal was to avoid numerical speed up tricks or stabilizing devices (no upwinding, etc.) to best follow what would be a representation of the physical flow. Physical experiments, by the way, to date cannot very well track secondary vortices lower in the cavity because the intensities fall off too quickly, e.g., by 10^{-4} in a vortex cascade in a corner.

2.2 Periodicities

After 180,000 time steps the flow at $Re = 10{,}000$ had settled into an oscillating pattern which clearly was not going to converge to a steady final solution. See the flow histories of [5] and Figure 3 here. The latter figure, from [6], shows final patterns of the flow after 330,000 time steps. We produced flow portraits only at every 1000 time steps, i.e., at each dimensionless time $t = 1$ second, having used $\Delta t = 0.001$. In these flow portraits, the velocity values have been normalized, namely, divided by their magnitude. Thus the portraits are qualitative: quantitative magnitudes are too small to show as more than points.

The "period" of the flow, e.g., see Figure 7 of [5], had appeared to be somewhere between 4 and 5 seconds.

2.3 Feigenbaum Frequency

As mentioned above, looking more closely at the final oscillation of the run of [5], I found [6] that the "period" of this oscillation is extremely close to Feigenbaum's universal constant $\delta \cong 4.669201$. To conclude this I took the portraits at 14 second intervals, as shown in Figure 3 here, and noted that $3\delta \cong 14.007$, already knowing the oscillation pattern to be repeating itself at a frequency somewhere between 4 and 5 seconds.

I have been mentioning this result at conferences since February 1988. The feedback has been interesting. It is of course objectionable that δ occurs here (it (approximately) definitely occurs, coincidence or not) as a "period", whereas one expects it to occur in a parameter ratio of increasing Reynolds number differences, for example.

Let me note however that there is a steady local Reynolds number buildup in the region of the left wall oscillation. Moreover I have found vestiges of at least one earlier period doubling in that critical flow region. And time here is really a dimensionless iteration parameter of a highly coupled quadratic dynamical system, as in the period doubling theory of Feigenbaum.

Finally, I have linked the Feigenbaum frequency to the actual shedding of vortices in the high shear interface region. This shedding, of alternatively signed tertiary vortices, is shown in Figure 4, taken from [6]. This shedding started earlier in the flow ($t \sim 92$) but could not maintain itself until later (after period doubling). Looking carefully at the quantitative velocity output shows a small chaotic fluctuation of trajectories about the (normalized) qualitative flow portraits, e.g., in an attractor like fashion.

It would be very interesting and valuable to have the resources to run this flow under different grid sizes and at aspect ratios and Reynolds numbers deviating slightly from $A = 2$ and $Re = 10{,}000$. A full parametric study would be of great value as the bifurcation diagram for cavity flow is not at all known.

The computational determination of valid flow conclusions for unsteady flows (e.g., how does one really conclude periodicity of a flow) will be a new chapter in numerical analysis.

3. Vortex Interactions in Aerodynamic Flows

We have begun a program to better understand the physically visualized vortex dynamics of flows over airfoils, and to investigate new numerical methods for their simulation. Initial results have been published in Gustafson and Leben [7,8,9].

3.1 Robust Multigrid Vortex Resolution

In [7,8] we have developed a numerical scheme which has successfully resolved up to 25 of the vortex cascade descending into a corner. This goes beyond the physics (the 25^{th} small corner vortex has intensity 10^{-111}) and is based upon a linear steady (Stokes) fluid model. No one really knows how many corner subvortices really persist in a nonlinear Navier-Stokes corner flow. But our method has proven its robustness.

3.2 Orthogonal Grid Generation

In [8,9] we have implemented a multigrid method to efficiently generate orthogonal grids around an airfoil, in body fitted coordinates. The equations describing the mapping are

$$(fx_\xi)_\xi + (f^{-1}x_\eta)_\eta = 0$$

$$(fy_\xi)_\xi + (f^{-1}y_\eta)_\eta = 0 \tag{3.1}$$

namely, two covariant Dirichlet problems are iteratively solved until a sufficient degree of orthogonality is obtained. The function f is a distortion function which must be interpolated into the domain. For details see [8,9].

Our method of grid generation in principle extends to 3 dimensions and it would be very interesting to examine its analytical and computational properties in that case, as well as its implementation to flow problems.

3.3 Vortex Shreddings

We have successfully simulated the full Navier-Stokes flow over an airfoil, in agreement with physical experiment. See [8,9] and Figure 5 here.

At moderate Reynolds numbers and constant acceleration we have been able to give the first demonstration the enhancement of lift by vortex *shreddings*. These simulations also agree with physical visualizations.

An example of vortex shredding is given in Figure 6 taken from [9]. Splitting of the primary positive lift vortex by the trailing edge vortex takes place in frame 20, causing a decrease in lift. Then shredding of the forward secondary negative lift vortex by the fragment of the primary vortex returning to the wing restores lift, thereby preventing stall under acceleration.

More details may be found in [9].

It would be very interesting and valuable to have the resources to do a full unsteady flow analysis using locally refined grids and the multigrid FAS feature to better understand the fundamentals of these vortex phenomena as they occur in aerodynamics.

References

1. B. Eaton and K. Gustafson, "Calculation of Critical Branching Points in Two-Parameter Bifurcation Problems," *J. Comp. Phys.* 50 (1983), 171-177.
2. K. Gustafson, "Combustion and Explosion Equations and Their Calculation," Chapter 7 of *Computational Techniques in Heat Transfer*, Eds: R. Lewis, K. Morgan, J. Johnson, W. Smith, Pineridge Press, U.K. (1985), 161-195.
3. E. Ash, B. Eaton, K. Gustafson, "Counting the Number of Solutions in Combustion and Reactive Flow Problems," to appear.
4. K. Gustafson and K. Halasi, "Vortex Dynamics of Cavity Flows," *J. Comp. Phys.* 64 (1986), 279-319.
5. K. Gustafson and K. Halasi, "Cavity Flow Dynamics at Higher Reynolds Number and Higher Aspect Ratio," *J. Comp. Phys.* 70 (1987), 271-283.
6. K. Gustafson, Chapter 5 in *The Mathematics of Vortex Methods and Vortex Motion*, SIAM Frontiers in Mathematics Series, Eds: K. Gustafson, J. Sethian, to appear.
7. K. Gustafson and R. Leben, "Multigrid Calculation of Subvortices," *Applied Math. and Comp.* 19 (1986), 89-102.
8. K. Gustafson and R. Leben, "Vortex Subdomains," *1st International Symposium on Domain Decomposition Methods for Partial Differential Equations*, Eds: R. Glowinski, G. Golub, G. Meurant, J. Periaux, SIAM (1988), 370-380.
9. K. Gustafson and R. Leben, *Proc. First National Fluid Dynamics Congress*, AIAA/ASME/ASCE/SIAM/APS, Cincinnati, July (1988).

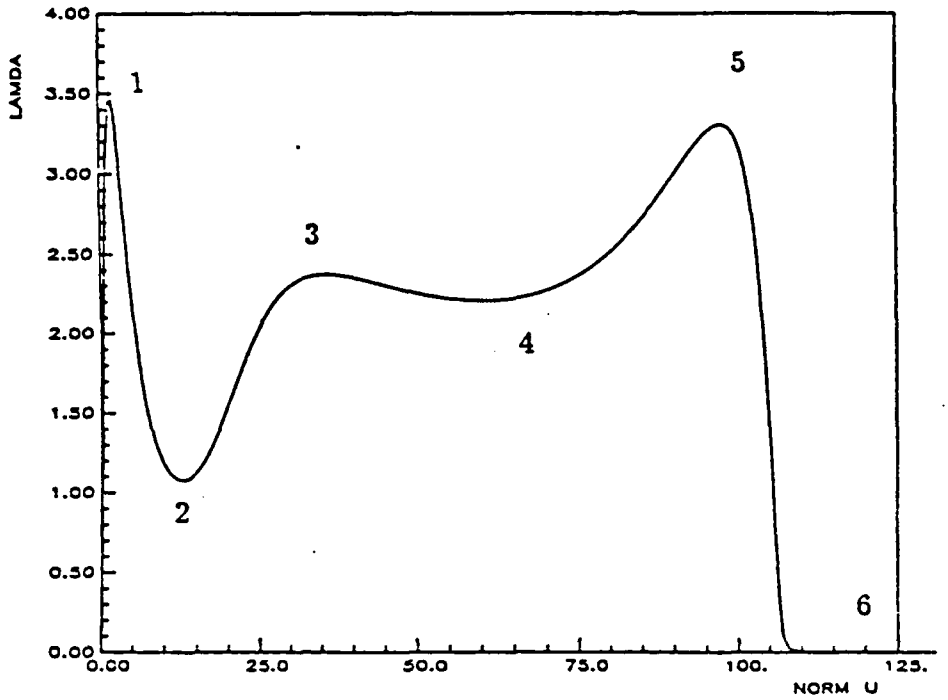
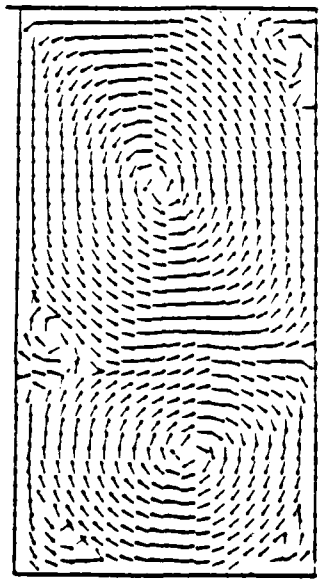


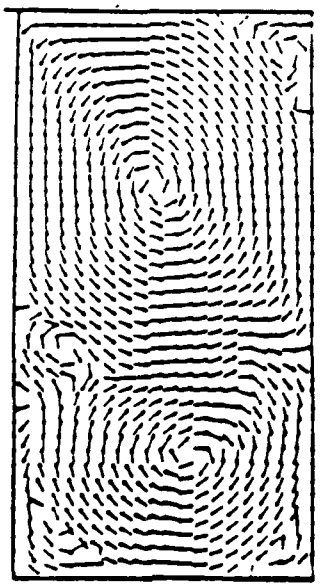
Figure 1. Bifurcation diagram and turning points for the full Arrhenius equation in 3 dimensions, for $\epsilon = 0.04$.

Number of Unstable Modes	Turning Points	Solution n = max u	Exothermicity λ
0	1	1.77223940	3.4828675
1	2	12.776963	1.0772377
2	3	35.563049	2.3736164
3	4	59.911631	2.2038626
2	5	97.240239	3.3059654
1	6	1054.2805	2.4383822×10^{-7}

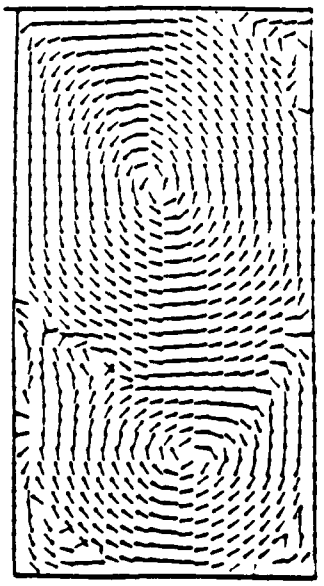
Table 2. Exact values of the turning points for $\epsilon = 0.04$.



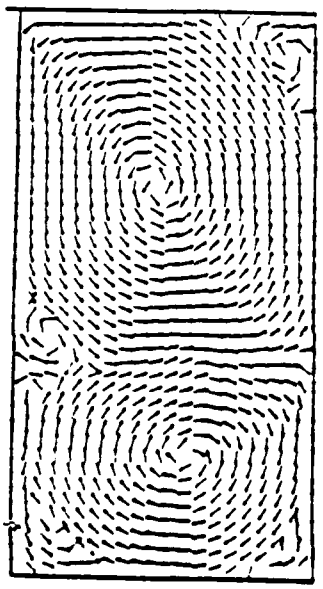
330 sec.



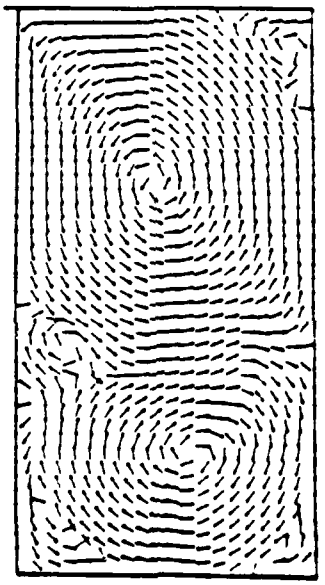
331 sec.



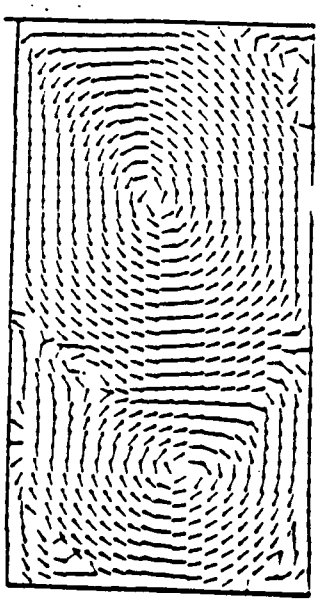
332 sec.



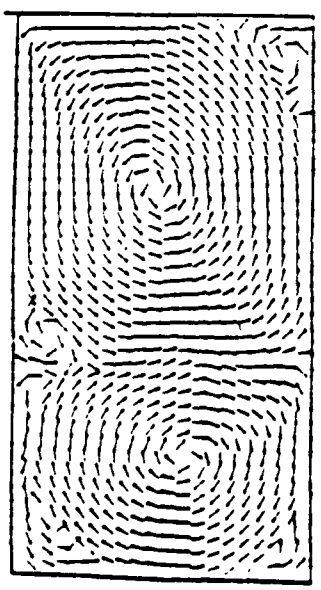
344 sec.



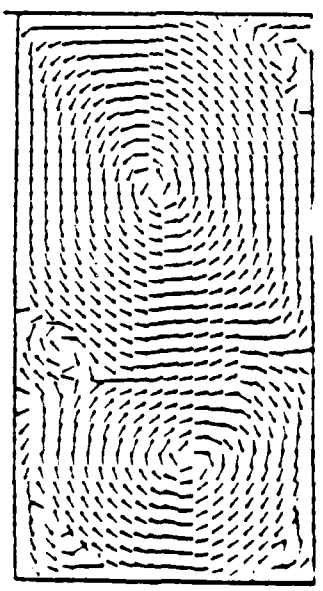
345 sec



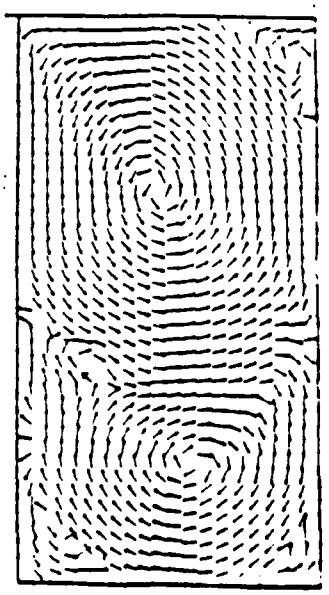
346 sec.



358 sec.



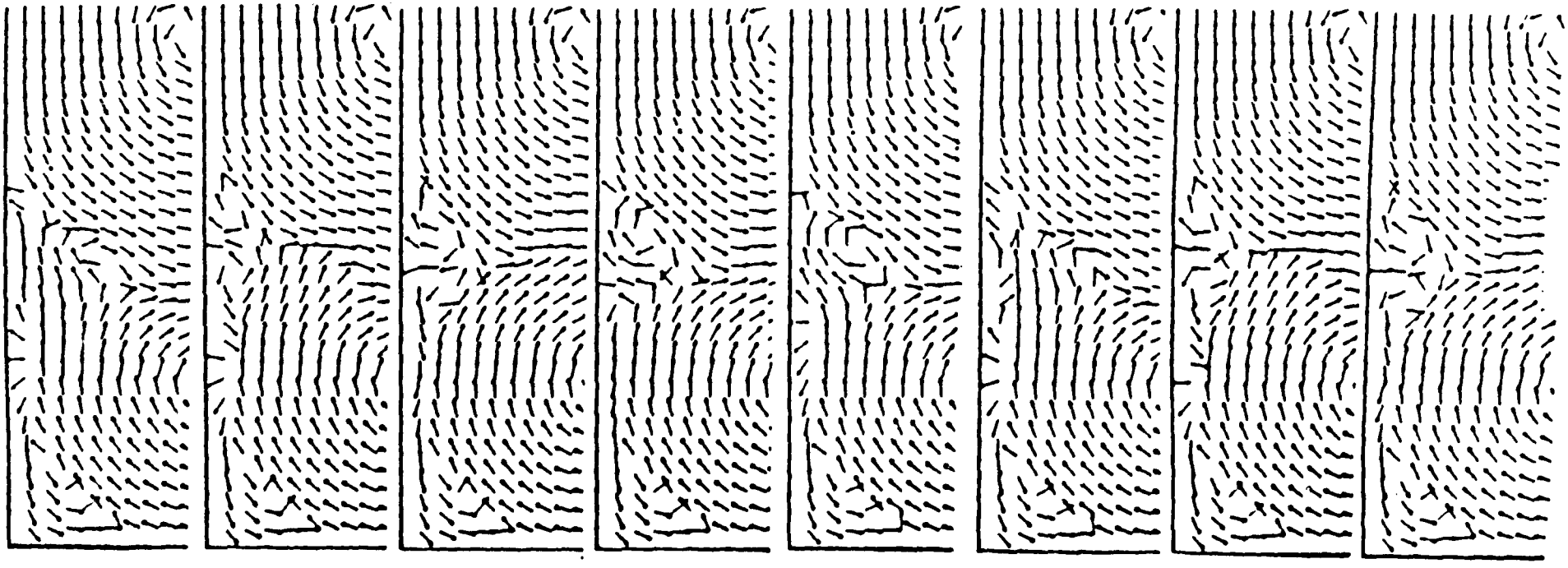
359 sec.



360 sec.

Figure 3. Final oscillation $\approx \delta = 4.669201$.

Figure 4. Vortex shedding manifestation of the period $\tau \approx \delta$. A small (−) tertiary vortex has just been split off from the midleft wall (−) secondary vortex, see t = 344, 345, 346 seconds in Fig. 3. Then a small thin (+) tertiary vortex is split off from the lower left corner (+) secondary vortex (t = 347 s), pulled into the major shear interface stress (t = 348 s), and shed into the cavity (t = 349 s). This is followed by another (−) shed from the midwall vortex (t = 350, 351 s), then the small thin (+) shed again (t = 352, 353 s), ····.



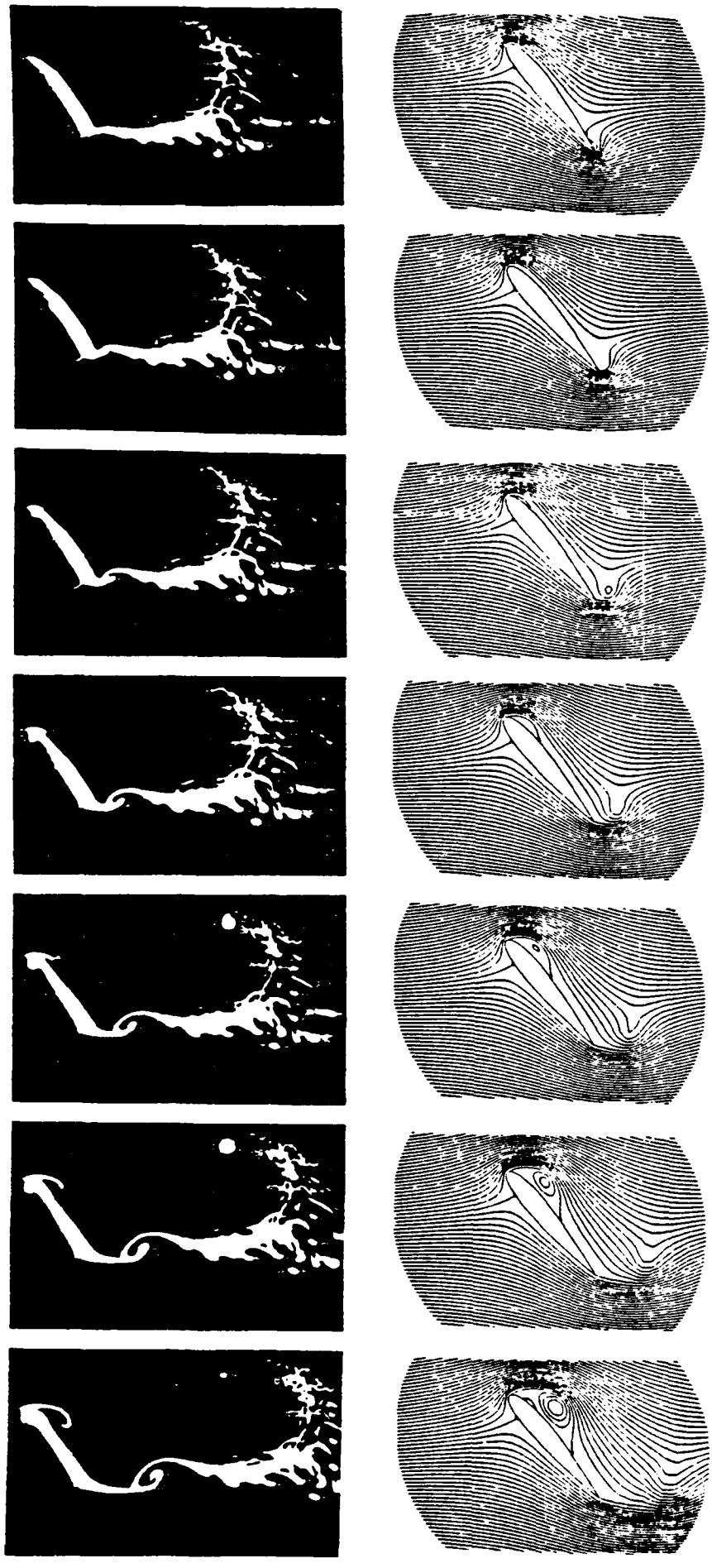


Figure 5(a). Constantly Accelerating Flow from Rest
$R_{acc} = 835,\ \alpha = 50°$

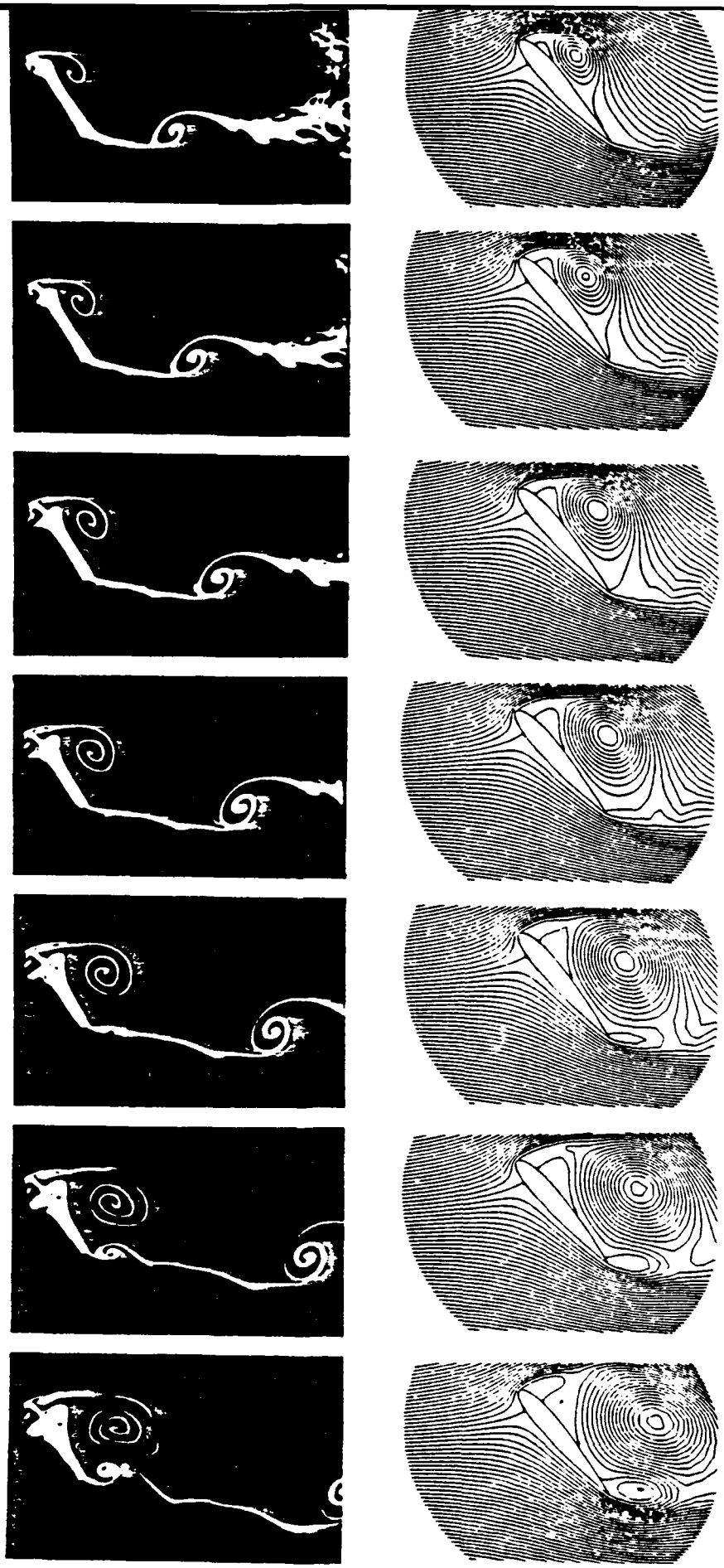


Figure 5(b). Constantly Accelerating Flow from Rest
$R_{acc} = 835,\ \alpha = 50°$

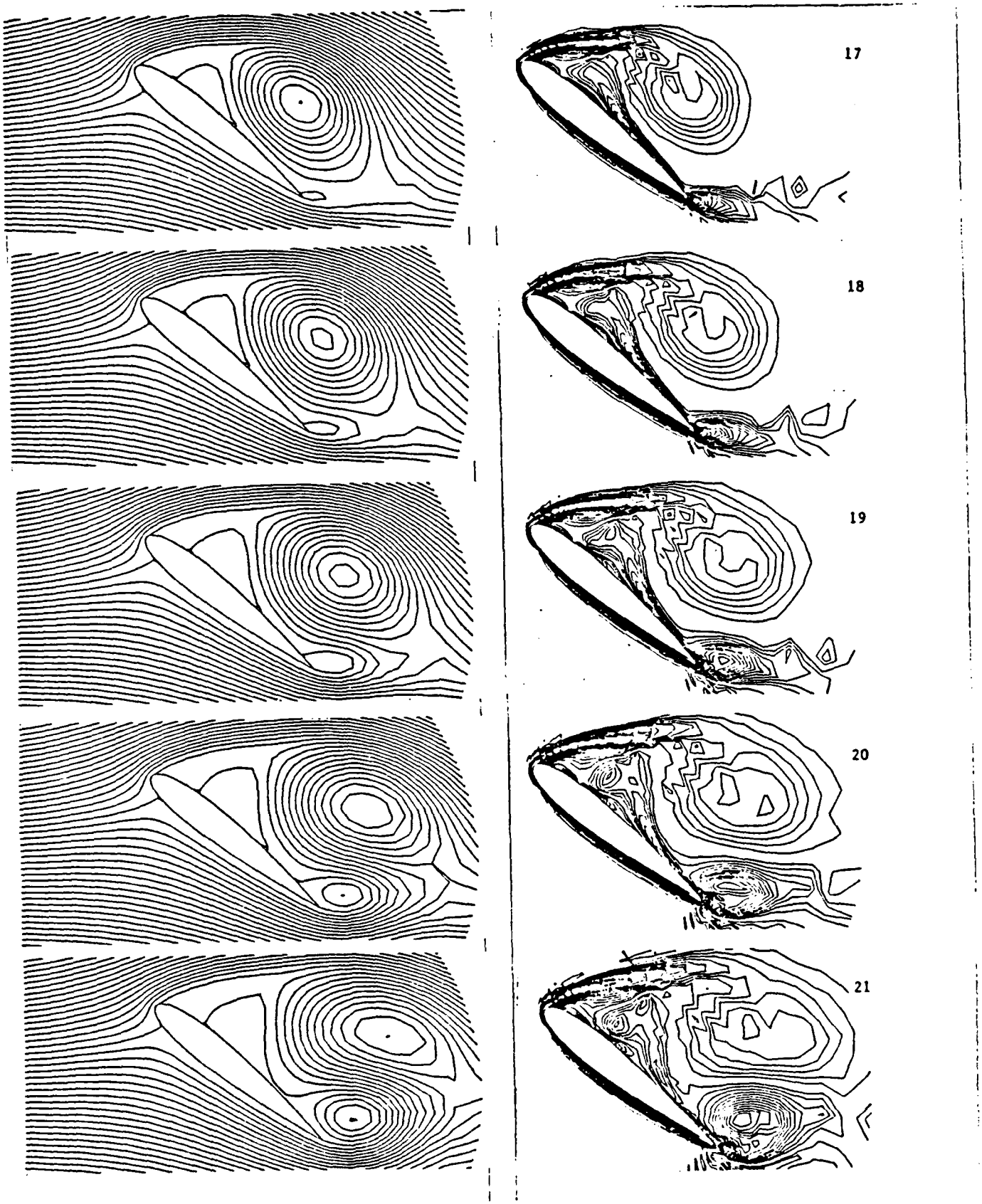


Figure 6(a). Constantly Accelerating Flow from Rest
$R_{acc} = 500,\ \alpha = 40°$

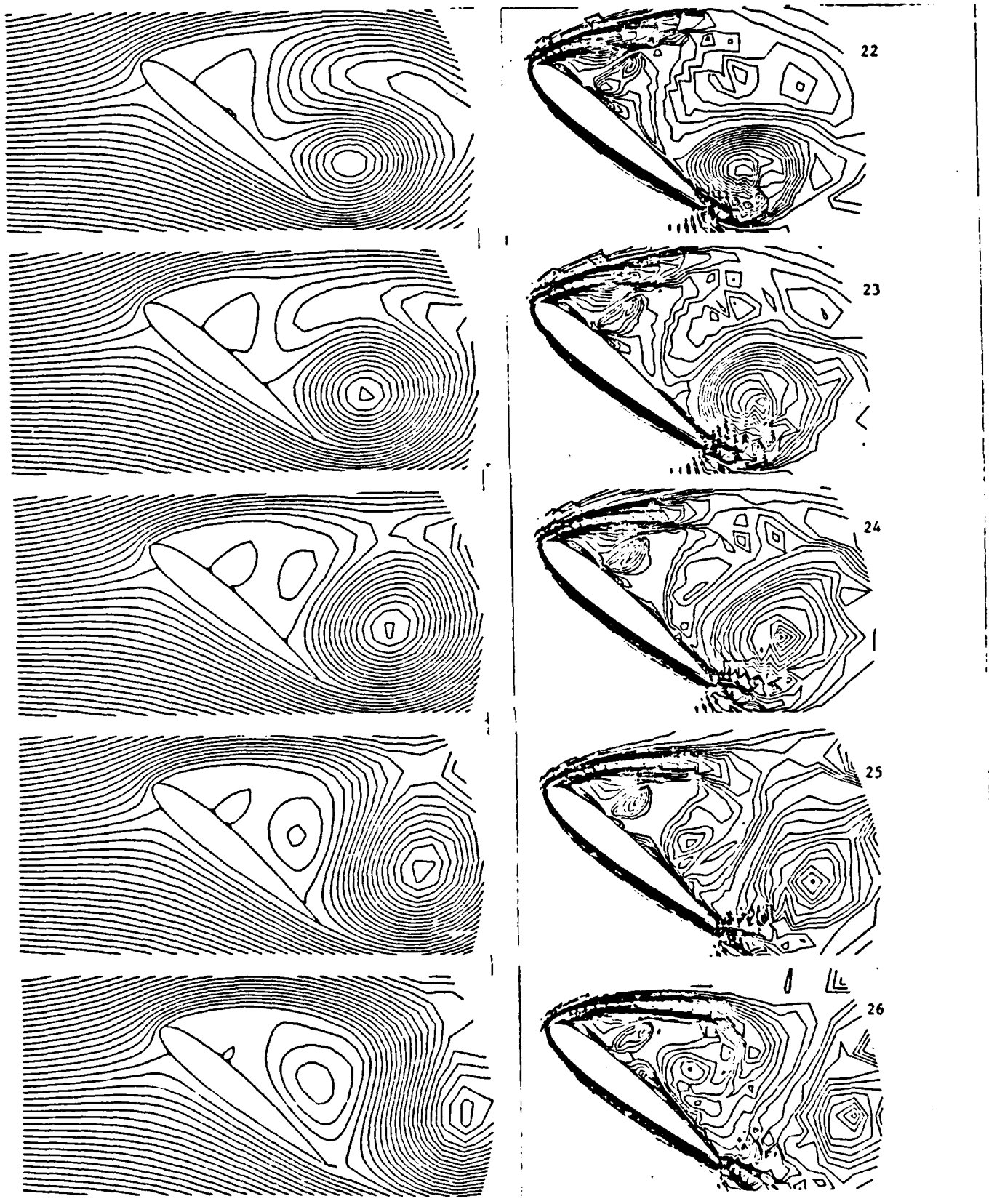


Figure 6(b). Constantly Accelerating Flow from Rest
$R_{acc} = 500,\ \alpha = 40°$

An Integrated Approach for Scientific Computing

An Extended Abstract

*PAUL S. WANG**
Department of Mathematical Sciences
Kent State University
Kent, Ohio 44242

1. Introduction. Modern workstations make it feasible to investigate such integrated computing environments. A workstation-based integrated scientific system should be the tool of choice for contemporary scientists and engineers. It is relatively simple to bring numeric, symbolic and graphics computing capabilities to a single computing system. What is more difficult is to have a truly integrated system where these techniques work together with very little barrier between them. More importantly, these three techniques should reinforce one another so that the whole is bigger than the sum of the parts. We briefly present some recent developments in this direction:

1. Symbolic derivation of numerical code for finite element analysis
2. Automatic numeric code generation based on derived formulas
3. Generating programs for parallel computers
4. Interactive graphing of curves and surfaces for mathematical formulas
5. Graphical user interface for mathematical systems
6. Software packages developed

This extended abstract is partially based on an earlier paper which appeared in the Proceedings of Compcon88, the 33rd IEEE Computer Society International Conference, Cathedral Hill Hotel, San Francisco, California, Feb. 29 – Mar. 4, 1988.

2. Symbolic Derivation for Finite Element Code. We have implemented a prototype software system to automate the derivation of formulas in finite element analysis and the generation of programs for the numerical calculation of these formulas. The generated code can be used with existing numerical packages. This is a general approach with good potential for many other scientific and engineering problems.

2.1. FINGER and GENTRAN.

From input provided by the user, either interactively or in a file, FINGER [8] will derive finite element characteristic arrays and generate FORTRAN code based on the derived formulas. The initial system handles the isoparametric element family. Element types include 2-D, 3-D, and shell elements in linear and nonlinear cases. The system allows easy extension to other finite element formulations.

*Work reported herein has been supported in part by the National Science Foundation under Grant CCR-8714836

3. GENTRAN. Actual generation of FORTRAN code from symbolic expressions or constructs is performed by the GENTRAN package [2,7] that we developed. It is a general purpose FORTRAN code generator/translator. It has the capability of generating control-flow constructs and complete subroutines and functions. Large expressions can be segmented into subexpressions of manageable size. Code formatting routines enable reasonable output formatting of the generated code. Routines are provided to facilitate the interleaving of code generation and other computations. Therefore, bits and pieces of code can be generated at different times and combined to form larger pieces. At the present time, work is going on to construct a LEX/YACC based code translator which will be much faster than GENTRAN and which can also produce vectorized f77 code.

4. Techniques for Generating Efficient Code. Our experiences in automatic code derivation and generation indicate that code generated naively will be voluminous and inefficient. We have used several techniques to generate better FORTRAN code.

(a) Automatic intermediate expression labeling.

(b) Using symmetry for generating functions and calls.

(c) Common subexpression identification.

(d) Using generated subroutines.

5. Generating Code for Parallel Processors. Carrying the automatic code derivation and generation idea one step further, current research at Kent State University addresses the derivation and generation of code for advanced parallel computers. As mentioned before, automatic generation of parallel code not only reduces manual mathematical manipulations but also helps engineers and scientists who are not computer experts take advantage of advanced parallel computers.

We have access to the Carnegie Mellon University (CMU) Warp systolic array computer [1] through dialout lines. We are able to make substantial progress experimentally with the Warp computer because Warp provides a good programming environment.

W2 is a simple Pascal-like high-level programming language [3] for the Warp array. W2 hides the low-level details of the Warp computer and provides a high-level abstraction for the Warp programmer. Using W2, a programmer can specify programs for each Warp cell and define inter-cell communications. It is the programmer's responsibility to devise an algorithm and map that algorithm to cell programs which can be executed in parallel efficiently. This is not a trivial task and is often central to finding a Warp solution to a problem. W2 is a convenient tool to program that solution.

6. P-FINGER AND GENW2. To generate key parts of the finite element computation into parallel code we have constructed the P-FINGER system [4]. P-FINGER runs under VAXIMA and is an enhanced version of FINGER to derive parallel code. Along with P-FINGER, a code generator package, GENW2 [6], has been developed by Trevor Tan at Kent State University. GENW2 is a parallel code generator written in Franz LISP and runs under the VAXIMA symbolic computation system. Given high-level algorithm specifications and expressions in symbolic representations, GENW2 outputs W2 code for the Warp systolic array computer. GENW2 can be used from VAXIMA top-level or invoked directly from the lisp

level. Generated routines may involve declarations, I/O statements, flow control, data distribution, subroutines, functions and macros. A code template can be specified by the user to render the output code in a designated format. The GENW2 package frees us from the syntax details of the target parallel language, W2, so we can concentrate on devising the parallel algorithms that will map important parts of finite element analysis on the Warp. The GENW2 package can also be used independently.

7. Graphics Display for Scientific Computation. Graphics display will play an important role in an integrated scientific computing environment. In such an environment graphics display should be an integral part of the user interface. A graphics package [9] for this purpose has been implemented to run under MACSYMA. This package features a highly interactive environment, a multiple window format and extensive help facilities. The capabilities include full color graphics, efficient hidden line removal, solid shading and cubic spline and least square curve fitting.

The package can display curves and surfaces given in either implicit or parametric form. The equations can be results of prior symbolic derivations. For plots involving many points, Fortran code is automatically generated to compute the function values more efficiently. The user has control over color, viewpoint, rotation, hidden line treatment etc. of plots. The control is provided alternatively through interactive menus or commands typed on the key-board. Plots can be superimposed using different colors.

The curve fitting capability allows the user to enter data points which are plotted as discreet points on the graphics display. A least square interpolation functions can then be calculated and the curve defined overlays the points. The equation for the fitted curve can be returned for further manipulation.

8. The GI/S Graphical User Interface. The user interface for a scientific computing system which combines numeric, symbolic and graphical capabilities should also be of advanced design which not only provides functionalities to control computations but is easy to learn and use. Recent studies in this direction resulted in the MathScribe [5] and the GI/S [10] user interface systems for REDUCE and MACSYMA respectively. These represent the initial steps in an investigation into suitable user interface designs for complicated scientific computing systems.

The trend is to take full advantage of the capabilities of a modern workstation. Multiple windows are provided to allow concurrent control of multiple activities. In GI/S, a mouse is used as a pointing device to select windows and expressions, to pop up menus and to issue commands. High resolution graphics is used for mathematical symbols, fonts and interactive plotting of points, curves and surfaces. An *emacs* style editor is active whenever and wherever user input is typed. Mouse-assisted "cut and paste " allows the user to rearrange text and graphics between windows. Mathematical expressions are displayed in a textbook-like two dimensional format. Using the mouse, subexpressions of mathematical formulas can be selected interactively. User specified operations can be applied to selected subexpressions.

8.1. GI/S windows

In the GI/S user interface system, two standard windows are displayed on the screen when the system begins. These are the input and display windows. The input window provides a command-line editor and a history mechanism to recall past commands. Results of computations are displayed in two dimensional form in

the display window. Other windows may be opened by the user as needed. There are several different types of windows:

1. Display window
2. Scratch window
3. Graphics window, and
4. Help window.

Windows are named. Each can be relocated and re-sized interactively by the user. A corner of each output window contains status information on how the computation controlled by the window is progressing. The mouse buttons are used for selection and for appropriate pop-up menus.

8.2 Mouse apply

One way to exploit the capability of the mouse in a scientific system is to use it to enhance mathematical operations. One such operation is singling out a part of a large expression and apply a user specified function to it with the result of the function replacing the original part *in place*. Let us call this operation "mouse apply".

Studies of user interface design of complicated scientific systems have just begun. Standards, protocols and conventions are still largely lacking. However, one can be sure that advances will be made and users will benefit much from the next generation interface systems.

9. Conclusions. Modern workstations offer a practical way to integrate numeric, symbolic and graphics computing systems into one comprehensive scientific computing environment. Operations such as symbolic formula derivation, automatic numerical program generation, derivation of parallel code, graphics display of data points and mathematical equations, and advanced user interfaces can work together and offer many desirable features and capabilities that are otherwise unavailable. Evolution of such integrated environment will one day provide a powerful tool for scientists and engineers for substantially increased productivity.

REFERENCES

1. Annaratone, M., Arnould, E., Gross, T., Kung, H. T., Lam, M. S., Menzilcioglu, O., and Webb, J. A., "The Warp Machine: Architecture, Implementation and Performance", IEEE Trans. on Computers, Dec. 1987, vol. C-36, no. 12, pp. 1523-1538.

2. GATES, B. L., "GENTRAN: An Automatic Code Generation Facility for REDUCE",

3. PAVELLE, R. AND WANG, P. S., "MACSYMA from F to G", Journal of Symbolic Computation, vol. 1, 1985, pp. 69-100, Academic Press.

4. SHARMA, N. AND WANG, P. S., "Symbolic Derivation and Automatic Generation of Parallel Routines for Finite Element Analysis", to appear in Proceedings, International Symposium on Symbolic and Algebraic Computation (ISSAC-88), Roma, Italy, July 4-8, 1988.

5. SMITH, C. J., SOIFFER, N., "MathScribe: A User Interface for Computer Algebra Systems," Proceedings, the 1986 Symposium on Symbolic and Algebraic Computation, 1986, pp. 7-12.

6. TAN T. AND WANG, P. S., "Automatic Generation of Parallel Code for the Warp Computer," to appear in Proceedings, International Workshop on Computer Algebra and Parallelism, Grenoble, France, June 29 – July 1, 1988.

7. WANG, P. S. AND GATES B., "A LISP-based RATFOR Code Generator", Proceedings, the Third MACSYMA Users Conference, August, 1984, pp. 319-329.

8. WANG, P. S., "FINGER: A Symbolic System for Automatic Generation of Numerical Programs in Finite Element Analysis", Journal of Symbolic Computation, vol. 2, 1986, pp. 305-316, Academic Press.

9. YOUNG D. A. AND WANG, P. S., "An Improved Plotting Package for VAXIMA", abstract, presented at ACM EUROCAL'85 Conference, April 1-3 1985, Linz Austria, Lecture Notes in Computer Science No. 204 (1985), Springer-Verlag, pp. 431-432.

10. YOUNG D. A. AND WANG, P. S., "GI/S: A Graphical User Interface For Symbolic Computation Systems", Journal of Symbolic Computation, Academic Press, Jan. 1988, pp. 365-380.

A Study of Symbolic Processing and Computational Aspects in Helicopter Dynamics*

S Ravichandran and G. Gaonkar
Florida Atlantic University
Boca Raton, Florida

J. Nagabhushanam
Indian Institue of Science
Bangalore, India

T.S.R. Reddy
University of Toledo
Toledo, Ohio

ABSTRACT

Even research models of helicopter dynamics often lead to a large number of equations of motions with periodic coefficients; and Floquet theory is a widely used mathematical tool for dynamic analysis. Presently, three types of approaches are used in generating the equations of motions. These are: 1) General purpose symbolic processors such as REDUCE and MACSYMA, 2) a special purpose symbolic processor DEHIM ---Dynamic Equations for Helicopter Interpretive Models---, and 3) completely numerical schemes. Comparative aspects of the first two purely algebraic approaches are studied by applying REDUCE and DEHIM to the same set of problems. These problems range from a linear model with one degree of freedom to a mildly nonlinear multi-bladed rotor model with several degrees of freedom. Further, computational issues in applying Floquet theory are also studied, which refer to: 1) the equilibrium solution for periodic forced response, 2) the transition matrix for perturbations about that response and 3) a small number of eigenvalues and eigenvectors of the unsymmetric transition matrix. That study shows the following: 1) Compared to REDUCE, DEHIM is far more portable and economical, but it is also less user-friendly, particularly during learning phases. 2) The problems of finding the periodic response and eigenvalues are well conditioned.

1. INTRODUCTION

Symbolic processing or computer algebra is a highly desirable adjunct of rotorcraft dynamics research [1-7]. For illustration, we select one research area ... aeroelastic stability in forward flight. Here, the complexity and extent of the process of deriving the equations of motions merit special mention. We broadly mention a few stages of that process, by passing details such as model description, ordering scheme, perturbation about a periodic orbit etc. For example, these stages include the following: 1) partial differential equations of inplane or lead-lag bending, out-of-plane or flap bending and elastic torsion, 2) rotor-support system or fuselage equations, 3) flow-field equations such as of downwash dynamics, 4) Galerkin-type discretization to generate ordinary differential equations with periodic coefficients, and 5) transformation of a complete set of equations in rotating or non-rotating coordinates, and state variable representation. Generally blade elasticity, blade-to-blade coupling and coupling between the rotor and the rotor-support system introduce a large number of state variables. In fact, use of nearly 50 state variables has become rather common even in simplified models of basic research (interpretive models). The corresponding picture in a stochastic environment e.g. rotorcraft in turbulence, is far more demanding. If we apply the second moment stability criterion, we need to generate "state equations" of order N(N+1)/2, [8-10]. In other words, a 40 th-order system requires 820 state equations.

*Supported by the U.S. Army Research Office.

Experience both with manual algebra and with computer algebra shows that computer algebra is the viable alternative to manual algebra. This viability is expected. After all, computer algebra is as much intrinsic to computers as is numerical computation, and the required expertise is comparable to that required in generating numerical results. In fact, once the user is used to a particular approach or system, computer algebra becomes rather routine, much more than numerical computation. Presently, three types of approaches are used: 1) general purpose or catholic symbolic processors such as MACSYMA and REDUCE [4,7], 2) a special purpose symbolic processor DEHIM ---Dynamic Equations for Helicopter Interpretive Models [1-3] and 3) completely numerical schemes [5,6], such as AGEM---Automatic Generation of Equations of Motions [6]. This study is restricted to the first two purely algebraic approaches.

With this as background, we now come to the two main objectives of this paper. The first one is to compare DEHIM with a general purpose processor, say REDUCE. The comparison is based on our experience in solving the same set of helicopter dynamics problems by the two approaches under reasonably identical conditions. Still a note of caution is in order. Such a comparison involves umpteen variables many of which defy quantification and it is subjective to a degree, and it may well be a boundless exercise. Moreover a multipurpose processor is virtually a finished product, provides numerous services and is less amenable to evolution. But a special purpose processor provides services restricted to a specialized area, it has modular structure and is constantly evolving. In spite of many gaps and constraints, the comparison of DEHIM with REDUCE should promote further research on the role and viability of special purpose processors in specialized areas, a research area in which only the barest beginnings have been made. Further, that comparison should contribute to finding better and improved means of comparing one approach with the other, including a completely numerical approach. The second objective is to broadly outline the computational aspects of the Floquet theory, particularly for high order (N>100) systems. We begin with this second objective.

2. APPLICATIONS OF FLOQUET THEORY

Rotorcraft models lead to mildly nonlinear ordinary differential equations, often with a large number of dominant periodic coefficients. The term "mildly nonlinear" implies that nonlinearity is important, but it does not dominate the solution. Thus, a perturbed linear solution about a periodic orbit is justified. Application of Floquet theory involves computation of three items [10]: 1) the periodic forced response, 2) the transition matrix for perturbations about that response and 3) a small number of eigenvalues and eigenvectors of the Floquet transition matrix, which is the state transition matrix at the end of one period. However, for many problems, we have to simultaneously and iteratively compute control settings along with response to obtain a periodic and desired system response, what is referred to as vehicle trim. In this paper the role of control settings is not studied. For completeness, we present a brief background of these three items, and then present a set of numerical coordinates, which provide a means of objectively describing the computational issues on the application of Floquet theory. We conclude this section with a discussion of numerical results pertaining to those coordinates.

2.1 Equilibrium State

The transient and forced responses are connected in a direct way in that the transient dynamics (about a periodic equilibrium) depend on that equilibrium solution. The Floquet transition matrix provides this connection. To elaborate, we introduce the Nx1 state vector x(t) and the T-periodic NxN state matrix A(t). For the Nx1 input vector G(t), the linear forced-response system can be expressed as

$$\{\dot{x}(t)\} = [A(t)]\ \{x(t)\} + \{G(t)\} \tag{1}$$

Now, the NxN state transition matrix $\Phi(t)$ is given by

$$[\dot{\Phi}(t)] = [A(t)]\ [\Phi(t)],\ \Phi(0) = I,\ 0 \le t \le T \tag{2}$$

To compute the initial state to give periodicity of the steady state, $\{x(0)\} = \{x(T)\}$, we first compute $x_E(T)$ which is the nonperiodic solution of the complete equation (1) at t = T for the zero initial state. Then we have

$$\{x(0)\} = [I - \Phi(T)]^{-1}\ \{x_E(T)\} \tag{3a}$$

Thus, the partial derivative of the errors, (x(T)-x(0)), with respect to the initial state x(0), is [I - FTM]. For the nonlinear case when G(t) in equation (1) is replaced by $G(x, \dot{x}, t)$, we iterate with an iterative adaptation of equation (3a):

$$[I - \Phi(T)]_{k+1}\ \{x_E(T) - x_E(0)\}_{k+1} = \{x_E(T) - x_E(0)\}_k \tag{3b}$$

where $x_E(0)$ is some Nx1 assumed initial state vector to start the iteration (k = 0). For details see references 11 and 12 which also include algorithmic aspects of sequentially perturbing each of the N elements of $x_E(0)$ by a small amount. Henceforth we will represent the Floquet transition matrix $\Phi(T)$ by FTM.

Concerning a solution strategy which couples Floquet theory to the response analysis, there is considerable similarity among the several trimming methods

[11,13,14] and for illustration we choose the method of periodic shooting [11,14]. In that method, we iterate on the initial conditions in order to find those that lead to a periodic solution of the nonlinear equations. The Floquet connection referred to earlier occurs in the iterative scheme, equation (3b), through the matrix [I-FTM]. Thus, the condition number of [I-FTM] quantifies how well-conditioned (or equivalently ill-conditioned) is the problem typified by equations (3); details in section 2.4 which introduces the concept and in section 2.5 which illustrates on the basis of numerical results.

2.2 Floquet Transition Matrix (FTM)

The FTM is part of the trimming analysis. For an nth-order system, the calculation of the FTM is equivalent to the solution of n, nth order initial-value problems or to one $n^2 x 1$ initial-value problem, what are referred to as n-pass and single-pass computations [15]. To effect this solution, several methods have been exercised, methods such as rectangular ripple [16], numerical perturbation [17], and recently finite elements [18-20]. Of these, time-marching in single-pass is by far the most popular. However, much promise exists for the finite element technique in the space-time domain [18-20]. A comparison of well tested IVP codes with the emerging finite element approaches to generate the FTMs and an exposition of the differences among the various finite element formulations present fruitful areas of research.

2.3 Eigenanalysis of the FTM

For large systems, the crux of the Floquet analysis is the eigenanalysis, which becomes more and more demanding with increasing order of the FTM. Due to algorithmic robustness and availability of well-documented computer codes, the generic QR-method (e.g., EISPACK version for a general matrix) is almost exclusively used for the eigenanalysis [21]. However, for high-order systems and for stochastic stability problems, such usage presents a computational barrier.

For a general matrix, the QR-method is the recommended method for a complete eigenalysis, as seen from its algorithmic structure (e.g., QR decomposition). While the operation counts and the machine time requirements grow cubically with the order of the FTM, the storage requirement grows as the square of the system order. Further, the Floquet analysis for stability requires only the dominant characteristic multiplier [10,12]. (In practice, we need a small subset of the dominant/sub-dominant eigenvalues, as well as the correspondent eigenvectors, due to frequency ambivalence and due to the necessity of identifying stability margins of critical modes). Thus, in summary, these restrictions show that the QR method is not practical for large systems and for the stochastic second moment stability of even a relatively small order system which requires an eigenanalysis of order N(N+1)/2. Two promising alternatives to the QR-method are: 1) the simultaneous iteration method [22-26] and 2) the generalized block-Lanczos method [22,27,28]. However further research is required to ascertain their viability for the Floquet eigenanalysis owing to nonsparsity of the FTMs.

2.4 Computational Reliability

The trimming analysis which includes the computation of the FTM, and the eigenanalysis of the FTM are subject to numerical perturbations which are involved and interdependent. For example, the characteristic multipliers

(eigenvalues of the FTM) are subject to numerical perturbations due to already existing perturbations in the FTM. It is necessary to know that we are not dealing with an ill-conditioned problem. That is, the perturbations or small changes in the data do not introduce large changes in the computed result or at least we have some means of ascertaining the goodness of the computations. The problem is ill-conditioned if the condition number is large, say, larger than 100, the ideal value being 1. To this end, following Ortega [29], we introduce the following computational reliability coordinates:

1. The matrix condition number of [I-FTM].

2. Condition number of characteristic multipliers and the vector of residual errors of eigenpairs (eigenvalue and the correspondent eigenvector).

The first coordinate concerns the periodic orbit analysis and it is a priori. The second set of two coordinates concerns the eigenanalysis of the FTM and is a posteriori. Though the condition number concept has a rigorous analytical basis [29], the corresponding condition number analysis for eigenvectors is too delicate to be practical [30]. Therefore here we use a combination of the eigenvalue condition number and the residual error of the correspondent eigenpair.

In the sequel we give a very brief account of these numerical coordinates with respect to a generic nonsymmetric real matrix A, right eigenvector x, left eigenvector y and eigenvalue λ. We use the 2-norm for the vector and the spectral norm for the matrix, that is,

$$\| x \|_2 = \sqrt{x^H x} \text{ and } \| A \|_2 = \sqrt{(\text{max. eigenvalue of } A^T A)} \tag{4}$$

where x^H is the Hermitian or complex conjugate transpose of x and A^T, the transpose of A. We mention in passing that $x^H y$ represents the inner product of x and y. The vectors are normalized such that

$$\| x \|_2 = \| x \| = \| x^H x \| = 1 = \| y \| = \| y^H y \| \tag{5}$$

The condition number of A or Cond.(A) is given by

$$\text{Cond.}(A) = [\text{maximum eigenvalue of } A^T A]^{1/2} / [\text{minimum eigenvalue of } A^T A]^{1/2} \tag{6}$$

and it satisfies the following inequality:

$$1 < \text{cond.}(A) < \infty \tag{7}$$

We assume that λ_j is a simple eigenvalue. Then the condition number with respect to each λ_j is given by

$$\text{Cond.}(\lambda_j) = \left| y_j^T x_j \right|^{-1} \tag{8}$$

where y_j and x_j are the left and right eigenvectors such that

$$Ax_j = \lambda_j x_j \text{ and } A^T y_j = \lambda_j y_j \tag{9}$$

It is good to emphasize that it is y_j^T that is used in equation (8) and not the Hermitian transpose y_j^H. Referring to the trimming analysis typified by equation (3b) we consider the following symbolic representation:

$$[A + \delta A]\{x + \delta x\} = \{b + \delta b\} \tag{10}$$

For example, A + δA represents I-FTM, x and b respectively representing $\{x_E(T) - x_E(o)\}_{k+1}$ and $\{x_E(T) - x_E(o)\}_k$. Under fairly general conditions it can be shown that [29]

$$\frac{\| \delta x \|}{\| x \|} < \text{cond.}(A) \left\{ \frac{\| \delta A \|}{\| A \|} + \frac{\| \delta b \|}{\| b \|} \right\} \tag{11}$$

Thus Cond.(A) represents the maximum magnification of the total relative errors in A and b. That is, the higher the value of cond.(A), the greater is the sensitivity of equation (3b) to computational perturbations, and consequently the less well conditioned is the problem of finding the periodic initial state. From equation (9), the relative residual error ε follows:

$$\epsilon = \frac{\| Ax_j - \lambda_j x_j \|}{\| \lambda_j x_j \|} = \frac{\| r \|}{\| \lambda_j \|}, \tag{12}$$

where r is the residual error. In the following section we present the numerical results on Cond.(A), Cond.(λ) and ε, which are respectively given by equations (6), (8) and (12).

2.5 Discussion of Results

In table 1, we present the condition numbers of the FTM and [I-FTM] together with the maximum eigenvalue condition number $\lambda_{j,max}$ and the

corresponding residual error of the eigenpairs. The physics of the problem refers to a multibladed rotor system with 3, 4 and 5 rigid blades. As sketched in figure 1 each blade has two degrees of freedom, flapping or out-of-plane motion and inplane or lead-lag motion. For the 3 and 4 bladed models the feedback system from the assumed unsteady aerodynamics or dynamic inflow model introduces 3 additional state variables, and for the 5 bladed model, it introduces 5 state variables. Thus we have 15 (3x4+3), 19 (4x4+3) and 25 (5x4+5) state variables. The first column contains the dimensionless velocity parameter μ, the higher the μ the more the dominance of periodic coefficients (and nonlinearity). While the second and third columns contain the condition numbers of the FTM and I-FTM, the fourth column contains the maximum eigenvalue condition number, that is the maximum value of cond.(λ_j) with respect to all the simple eigenvalues. (For the data in Table 2, all the N eigenvalues were simple or of multiplicity one.) The last column contains the residual error for the eigenpair corresponding to $\lambda_{j,max}$. The results are extremely interesting. The FTM is seriously ill-conditioned and this undesirable feature increases with increasing μ. But the crucial ingredient, [I-FTM], as seen through equations (3), is extremely well conditioned, the ideal value being one. This means that the problem of finding the periodic orbit as typified by equation (3) is well conditioned. These data show that though the FTM is ill-conditioned (with regard to its inverse), all the eigenvalues of the FTM are well conditioned. This feature is well corroborated by the corresponding residual error vector in the fifth or last column.

3. SPECIAL PURPOSE PROCESSOR DEHIM

The literature on the multipurpose processors such as MACSYMA, REDUCE and MAPLE is extensive. For example, the book by Davenport, Siret and Tournier [31] is encyclopedic. It contains an extensive bibliography and provides an excellent introduction to the general algorithmic basis of computer algebra and also in particular to the use of MACSYMA and REDUCE. By comparison, DEHIM, as is the case with special purpose processors, is restricted to a highly specialized area and merits some introduction. Details intended both of the learning and use of DEHIM are given in the Users' Manual [2] and in references 1 and 3. The introduction in the sequel is overly condensed. Nevertheless it should facilitate an appreciation of the view point that a special purpose processor can be developed as a natural predecessor to programming for numerical computations, and that the development and use of such processors are no more involved than programming for numerical results in such specialized areas.

3.1 Description of DEHIM

The four main aspects of DEHIM are the following: 1)Algebraic manipulations capabilities, 2) Commands, 3) Input-output details, and 4) Special features.

3.1.1 Algebraic Manipulations Capabilities

The manipulations consist of combining expressions, replacing variables in an expression by designated expressions or relations, and substituting numerical or logical values and tables into expressions. They also include the expansion of composite functions and expressions according to stipulated ordering schemes and the collection of coefficients of a specified variable in an expression. The algebraic manipulations of partial differentiation and integration, and matrix operations are carried out from the user supplied rules.

3.1.2 Commands

Several commands such as input commands form an important feature of the processor, and a brief account with illustration in parenthesis is given in appendix 1. Essentially, commands are constructed to perform various symbolic manipulations and they are oriented to the algebraic manipulations typical in helicopter dynamics as in deriving equations to ordering schemes and transforming into multiblade (non-rotating) coordinates.

3.1.3 Input-Output Details

The input to the program comprise the command names and their parameters which are in Alpha Numeric Format. Further, the processor gives two sets of outputs. The first set contains the resulting expressions of algebraic manipulation commands, perturbed linear equations and equations involving multiblade (nonrotating) coordinates. The expressions are printed term by term and one below the other for easy perusal by the user. The second set contains outputs which are coded FORTRAN statements of the equations as required in the subsequent numerical analysis. A typical input block diagram is sketched in Fig. 2. Appendix 2 gives a few samples of intermediate (optional) outputs. For example RYD there represents the y-component of the total time derivative (in a fixed frame of reference) of a dimensionless position vector r, as detailed in Figure 1.

3.1.4 Special Features

These features primarily refer to modular construction and portability. The modular structure permits the introduction of new commands or modifications of the old commands to consider major modifications in the formulation. Thus, the same program can be utilised to consider a variety of modifications or extensions of the original analytical model. Usually the implementation of symbolic manipulation systems on another computer requires a major effort in that it must take advantage of the specific features of the hardware and operating system of the host computer. The present program originally written in FORTRAN IV and now in 77 can be implemented with minimal assistance from the host computer, i.e. by utilising its Fortran compiler. As such, it is highly portable. A reset counter is also incorporated which erases all previous equations and saves core space for the next equation. If other language facilities such as LISP are available, required adaptation is routine. Other features include format free input and execution of several derivation steps through a single command. We conclude this section by mentioning that it is a routine exercise to incorporate ordering schemes and tables of formulae of trigonometric tables, perturbation scheme tables and multiblade coordinate transformation tables [2].

4. APPLICATIONS OF REDUCE AND DEHIM

We begin with the core space requirements to install these packages. As expected DEHIM takes far less core space; that is, as shown in table 2, 83 versus 1573 blocks. However, this comparison should be tempered by the fact that REDUCE provides numerous services, as is typical of a multipurpose processor. By comparison DEHIM provides services that are restricted to deriving equations of motions of rotorcraft dynamics models. In table 2 four cases are presented ---one-bladed and three-bladed rotors in combination with rigid flap and rigid

flap-lag blades. While in hover, the system has constant coefficients, in forward flight, the system has periodic coefficients. The treatment includes derivation of nonlinear equations, perturbed linear equations for stipulated trim conditions (no trim analysis) and transformation into multiblade or nonrotating coordinates for the three-bladed case. It is clearly seen that DEHIM is far more economical and this saving in machine time increases rapidly with N. Our experience with a wide range of problems also shows that DEHIM is remarkably portable. However, the feedback from users shows that during inital stages DEHIM is far less user-friendly compared to REDUCE. This is probably due to two reasons. First, the present USERS' manual does not take the user in small gradual steps and merits further elaboration on the basis of highly simplified graded examples. Second, all the users had used REDUCE earlier. The exercises of table 2 were treated as another set of problems to which REDUCE was applied once again, whereas with DEHIM those exercises were entirely a new experience.

5. CONCLUSIONS AND FUTURE WORK

The feasibility of programming with special purpose processor DEHIM for generating the equations of motions of helicopter dynamics models with a priori ordering schemes is demonstrated. Some examples treated range from a four bladed rotor model that has flap bending, lag bending and torsion degrees of freedom to a coupled rotor-body system with 3,4 and 5 rigid lag-flap blades with hinge offset and dynamic inflow under forward flight conditions [1-3]. The viability has been tested in including nonlinear airfoil characteristics and dynamic stall characteristics according to user supplied tabulated airfoil data and dynamic stall models. The program generates perturbed linear equations from the nonlinear ordinary (for rigid blades) or partial (for elastic blades) differential equations [1-3].

Compared to multipurpose processor REDUCE and with respect to a restricted class of helicopter dynamics problems, DEHIM is far more portable and economical, though it is found to be less user-friendly during learning phases.

The modular structure of the program allows the programmer to alter the existing modules and to add new subroutines. This program is oriented towards flexibility of application and user modification. Its application-oriented commands make user inputs minimal since many of the formulation steps are built into commands. The intermediate expression swell is significantly minimised since formulation procedures are carried out at term level rather than at expression level. DEHIM offers considerable promise in demonstrating that symbolic manipulation can be significantly exploited in deriving equations of motions of helicopter systems.

Concerning the computational reliability, the problem of finding the initial state that guarantees periodic forced response is found to be well-conditioned. That is, the condition number of [I-FTM] as typified by equations (3), is of the order of one, see table 1. This is remarkable in that the condition number of FTM, compared to that of [I-FTM] is extremely high and it generally increases with increasing nondimensional flight speed μ. The present study does not include the impact of control inputs. Therefore, how well-conditioned is the complete trim problem of finding the augmented vector of initial state for response periodicity and control inputs for desired response characteristics merits further research.

The problem of finding the eigenvalues in the Floquet analysis is well conditioned in that the eigenvalue condition numbers are of the order of one. This finding is further corroborated by the computed residual errors of the correspondent eigenpairs, as typified by equation (12).

Presently the QR method is almost exclusively used in the Floquet eigenanalysis for which the machine time grows cubically with the system dimension N. This fact practically precludes the use of the QR method for large systems (N>100) and for the stochastic second moment stability of even relatively small order systems (N≈25), since the latter case requires an eigenalysis of order N(N+1)/2. Floquet eigenanalysis in practice requires only a small subset of eigenvalues and eigenvectors. Therefore, though the FTM is generally not sparse, the feasibility of using simultaneous iteration and generalized Lanczos method for the unsymmetric eigenanalysis offers considerable promise.

6. ACKNOWLEDGEMENT

This work is sponsored by the U. S. Army Research Office under grant No. DAAL03. The view, opinions and/or findings contained in this report are those of the authors, and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

7. REFERENCES

1. Nagabhushanam, J., Gaonkar, G.H., and Reddy, T.S.R., "Automatic Generation of Equations for Rotor-Body Systems with Dynamic Inflow for a Priori Ordering Schemes," Seventh European Rotorcraft and Powered Lift Aircraft forum, Garmisch-Partenkirchen, Federal Republic of Germany, Sept. 8-11, 1981, paper no. 37.

2. Nagabhushanam, J., Gaonkar, G.H., Srinivasan, P., and Reddy, T.S.R., Users Manual for Automatic Generation of Equations of Motion and Damping Levels for Some Problems of Rotorcraft Flight Dynamics, R.&D Report, HAL-IISc Helicopter Training Program, Indian Institute of Science, Bangalore, October 1984. (Also see NASA TM-86750, 1985.)

3. Reddy, T.S.R., and Warmbrodt, William, "Forward Flight Aeroelastic Stability from Symbolically Generated Equations", Journal of he American Helicopter Society, July 1986, pp. 35-44.

4. Crespo da Silva, M.R.M., and Hodges, D.H., "The Role of Computerized Symbolic Manipulation in Rotorcraft Dynamic Analysis", Computers and Mathematics with Applications, Vol. 12A, no. 1, 1986, pp. 161-172.

5. Lytwyn, R.T., "Aeroelastic Stability Analysis of Hingeless Rotor Helicopters in Forward Flight Using Blade and Airframe Modes", 36th Annual Forum of the American Helicopter Society, Washington, D.C., May 1980, Preprint No. 80-24.

6. Done, G.T.S., Juggins, P.T.W., and Patel, M.H., "Further Experience with a New Approach to Helicopter Aeroelasticity," Thirteenth European Rotorcraft Forum, Arles, France, Sept. 8-11, 1987, Paper No. G·11.

7. Xin, Zhao and Curtiss, H.C., "A Linearized Model of Helicopter Dynamics including Correlation with Flight Test," Proceedings of the 2nd International Conference on Rotorcraft Basic Research, University of Maryland, February 16-18, 1988, pp. 2HAQ·1 - Zhao·14.

8. Prussing, J.E., Lin, Y.K., and Shian, T.N., "Rotor Blade Flap-Lag Stability and Response in Forward Flight in Turbulent Flows", Journal of the American Helicopter Society, Vol. 29, No. 4, Octdober 1984. pp. 81-82.

9. Lin, Y.K. and Gaonkar, G.H., "Motions of Rotor Blades in Turbulent Flow", The Second Technical Workshop on Dynamics and Aeroelastic Stability Modeling of Rotorcraft Systems, the U.S. Army Research Office and Florida Atlantic University, Boca Raton, Florida, November 18-20, 1987.

10. Gaonkar, G.H. and Peters, D.A., "Review of Floquet Theory in Stability and Response Analysis of Dynamic Systems with Periodic Coefficients," Recent Trends in Aeroelasticity, Structures and Structural Dynamics, Bisplinghoff Memorial Symposium Volume, University of Florida, University Presses of Florida, Edited by P. Hajela, pp. 102-120.

11. Peters, D.A. and Izadpanah, "Helicopter Trim by Periodic Shooting with Newton-Raphson Iteration", Proceedings of the 37th Annual National Forum of the American Helicopter Society, New Orleans, May 1981, paper No. 81-23.

12. Dugundji, J. and Wendell, J.H., "Some Analysis Methods for Rotating Systems with Periodic Coefficients", AIAA Journal, Vol. 21, No. 6, June 1983, pp. 390-897.

13. Johnson, W., A Comprehensive Analytical Model of Rotorcraft Aerodynamics and Dynamics, Part I: Analysis Development, USAAVRADCOM TR 90-A-5, June 1980.

14. O'Malley, James A., Izadpanah, Amir P., and Peters, D.A., "Comparison of Three Numerical Trim Methods for Rotor AirLoads", Ninth European Rotorcraft Forum, Stresa, Italy, September 13-15, 1983, Paper No. 58.

15. Gaonkar, G.H., Prasad, D.S., and Sastry, D., "On Computing Floquet Transition Matrices for Rotorcraft", Journal of the American Helicopter Society, Vol. 26, No. 3, July 1981, pp. 56-61.

16. Lowis, O.J., "The Stability of Rotor Blade Flapping Motion at High Tip Speed Ratios", Reports and Memoranda, No. 3544 (U.K.) January 1963.

17. Peters, D.A., "Flap-Lag Stability of Helicopter Rotor Blades in Forward Flight", Journal of the American Helicopter Society, Vol. 20, No. 4, October 1925,pp. 2-13.

18. Borri, M., "Helicopter Rotor Dynamics by Finite Element Time Approximation", Computers and Mathematics with Applications, Vol. 12A, 1986, No. 1, pp. 149-160.

19. Izadpanah, Amir, p-version Finite Elements for the Space-Time Domain with Application to Floquet Theory, Ph.D. Thesis, Georgia Institute of Technology, August 1986.

20. Panda Brahmananda, and Chopra, Inderjit, "Dynamics of Composite Rotor Blades in Forward Flight", Vertica, Vol. 11, No. 1/2, pp. 187-209, 1987.

21. Garbow, B.S. et al., Matrix Eigensystem Routines - EISPACK Guide Extension, Lecture Notes in Computer Science, No. 57, Springer-Verlag, New York, 1977.

22. Chatelin, Francoise, Spectral Approximation of Linear Operators, Academic Press, 1983, Chapter 1.

23. Steward, G.W., "Simultaneous Iteration for Computing Invariant Subspaces of Non-Hermitian Matrices", Numerical Mathematics, Vol. 25, 1976, pp. 123-136.

24. Stewart, G.W., SRRIT - A FORTRAN Subroutine to Calculate the Dominant Invariant Subspaces of a Real Matrix, Technical Report TR-514, Computer Science Department, University of Maryland, 1978.

25. Jennings, A. and Stewart, W.J., "Simultaneous Iteration for Partial Eigensolution of Real Matrices", J. Inst. Math. Appl., Vol. 15, 1975, pp. 351-361.

26. Stewart, W.J., and Jennings, Alan, "A Simultaneous Iteration Algorithm for Real Matrices", ACM Trans. Math. Software, Vol. 7, No. 2, 1981, pp. 184-198.

27. Cullum, Jane, and Willoughby, R.A., "A Lanczos Procedure for the Modal Analysis of Very Large Nonsymmetric Matrices", Proceedings of the 23rd IEEE Conference on Decision and Control, Dec. 12-14, 1984, Las Vegas, Nevada, pp. 1758-1761.

28. Cullum, Jane and Willoughby, R.A., "A Practical Procedure for Computing Eigenvalues of Large Sparse Nonsymmetric Matrices", Large Scale Eigenvalue Problems, North Hallandale Mathematics Studies, Vol. 127, 1986, pp. 193-240.

29. Ortega, J.M., Numerical Analysis: A Second Course, Academic Press, New York, 1972, Chapters 2 and 3.

30. Gourlay, A.R. and Watson, G.A., Computational Methods for Matrix Eigenproblems, John Wiley and Sons, New York, 1973, p. 24.

31. Davenport, J.H., Siret, Y. and Tournier, E., Computer Algebra, Academic Press, New York, 1988.

Table 1: Computational Reliability Coordinates for N = 15, 19 and 25.

N = 15

μ	COND.(FTM)	COND.(I-FTM)	MAX. COND(λ)	RESIDUAL ERROR
0.0	2.79E02	1.91	1.51	0.109E-14
0.1	9.61E01	1.95	2.89	0.222E-13
0.2	6.42E02	1.89	2.32	0.165E-12
0.3	7.69E03	1.86	2.19	0.264E-11
0.4	4.50E04	1.84	2.22	0.123E-10
0.5	1.27E06	1.83	2.07	0.252E-09

N = 19

0.0	1.10E02	2.05	1.49	0.770E-15
0.1	5.59E01	2.02	3.77	0.137E-13
0.2	6.08E02	1.92	2.32	0.136E-12
0.3	7.79E03	1.88	2.13	0.146E-11
0.4	9.77E04	1.87	2.10	0.146E-10
0.5	1.36E06	1.87	2.09	0.252E-09

N = 25

0.0	3.72E02	2.05	3.40	0.272E-14
0.1	1.57E02	1.86	5.37	0.890E-14
0.2	5.94E02	1.91	3.38	0.189E-14
0.3	7.65E03	1.89	3.57	0.889E-12
0.4	9.07E04	1.87	3.39	0.273E-14
0.5	1.09E06	1.87	3.40	0.130E-14

Table 2: Applications of DEHIM and REDUCE (Vax 8800)

Approach		REDUCE	DEHIM
Core Space (in blocks)		1573	83
CPU time (in secs.)			
NONLINEAR EQUATIONS (Single-Bladed Rotor)	Rigid Flap (N=2)		
	Hover	6.06	1.74
	Forward Flight	7.14	2.11
	Rigid Flap-lag(N=4)		
	Hover	9.02	5.25
	Forward Flight	12.00	8.26
LINEAR EQUATIONS (Single-Bladed Rotor)	Rigid FLAP (N=2)		
	Hover	6.59	2.81
	Forward Flight	8.16	3.25
	Rigid Flap-lag(N=4)		
	Hover	15.82	6.43
	Forward Flight	34.20	7.91
MULTIBLADE EQUATIONS (Three-bladed rotor)	Rigid Flap(N=6)		
	Hover	11.17	3.15
	Forward Flight	19.18	3.54
	Rigid Flap-lag(N=12)		
	Hover	139.00	9.06
	Forward Flight	362.00	14.39

Appendix 1: COMMANDS

- To input an expression

  :%FCT=5.X**4*Y**5+A*SIN(BT)*COS(BT)$

  (function $f_{ct}=5x^4y^5+a\ \sin\beta\cos\beta$)

- To input a matrix

  :?TR(3,3)=COS(P);0.;-SIN(P);0;1;0;SIN(P);0;COS(P)$

  (matrix of size 3x3),$[T_R]=\begin{bmatrix} COS(P) & 0 & SIN(P) \\ 0 & 1 & 0 \\ -SIN(P) & 0 & COS(P) \end{bmatrix}$

- To input a relation table

  :REL.TAB:#RTB1:SIN(ZE)=ZE+DZ;SIN(BT)=BB+DB-.5*BB**2*DB$

  (a table of relations named RTB1 containing $\sin(\zeta)=\bar{\zeta}+\delta\zeta$ and $\sin(\beta)=\bar{\beta}+\delta\beta-.5\bar{\beta}^2\delta\beta$)

- To assign order of magnitude to the variables

  :ORD.MAG:(BB,1,2),(DB,2,1)$

  (the variable β belongs to group 1 with order of magnitude ϵ^2 (ϵ,measure of magnitude of group 1 variables) and $\delta\beta$ belongs to group 2 with order of magnitude δ (δ,measure of magnitude of group 2 variables))

- Scheme for term retention

  :TER.RET:#TSCHM=(1,2),(2,1)$

  (term retention scheme TSCHM defines that during expansions retain terms whose magnitude is limited to ϵ^2 for group 1 variables and to δ for group 2 variables)

• Algebraic manipulation of matrices

:A1[DIFF,TAU,BT;SUBS,#RTB1]=?A2(TRAN)*?A3(INTG,RB,0.,1)

[SUBS,#RTB2;TRSH,#TSCHM]+?B1*?B3(DIFF,BTD)$

(matrix [A1])=[$\frac{\partial^2}{\partial\tau\,\partial\beta}\{[A_2]^T(\int_0^1 [A_3]dr)]$ with substitution

of table of relations RTB2 and application of retention scheme TSCHM } +

$[B1](\partial/\partial\beta[B_3]$ with substitution of relations of table RTB1]

•To input variables whose coefficients are to be collected

:VAR.COL.COE:#CVAR=DB,DBD$

(define a string of variables by CVAR which contains the names of the variables DB and DBD).

. To collect coefficients of an expression

:COL.COE:%A1(#CVAR,FORT,PTEQ)$

(collect the coefficients of the variables defined in the string of variables CVAR of the function A1 and transform the coefficient expressions into FORTRAN statements and store the details with index PTEQ)

. Multiblade Coordinate Transformation

:MUL.TRA:?MUEQ[#RMUB,4]=?PE$

(Transforms the expression PE which is written in rotating coordinates into expression with non-rotating coordinates (multiblade coordinates for blades) using relations defined in relation table RMUB of a 4 bladed rotor system)

Appendix 2: INTERMEDIATE OUTPUT

Details of expression RYD

```
 1.000000*COS(CY)*HEPS
+1.000000*COS(BT)*COS(ZE)*COS(CY)*RB
-1.000000*BTD*SIN(BT)*COS(ZE)*SIN(CY)*RB
-1.000000*COS(BT)*ZED*SIN(ZE)*SIN(CY)*RB
-1.000000*COS(BT)*SIN(ZE)*SIN(CY)*RB
-1.000000*BTD*SIN(BT)*SIN(ZE)*COS(CY)*RB
+1.000000*COS(BT)*ZED*COS(ZE)*COS(CY)*RB
```

(Output of the details of expression $\dot{R}_y = h_e \cos\psi + \bar{r} \cos\beta \cos\zeta \cos\psi - \bar{r}\dot{\beta} \sin\beta \cos\zeta \sin\psi - \bar{r}\dot{\zeta} \cos\beta \sin\zeta \sin\psi - \bar{r} \cos\beta \sin\zeta \sin\psi - \bar{r}\dot{\beta} \sin\beta \sin\zeta \cos\psi + \bar{r}\dot{\zeta} \cos\beta \cos\zeta \cos\psi$;

where R_y is the y-component of the time derivative of the position vector.)

Details of Matrix AA (3x1)

Terms of element (1,1)

```
1.*SIN(BT)*COS(CY)+5.*SZE
```

Terms of element (3,1)

```
10.5*SIN(CY)*LOG(X)
```

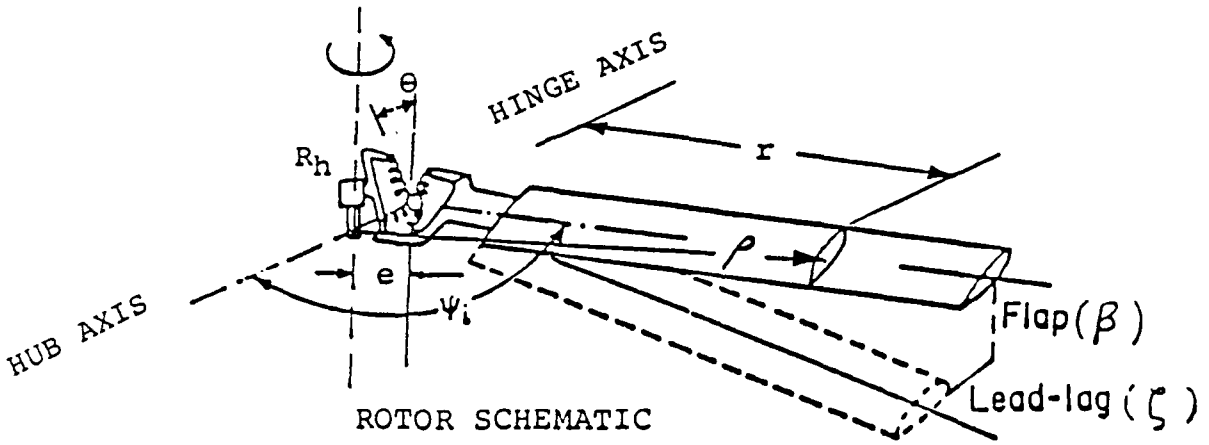
(output of matrix AA(size 3x1) which corresponds to $\begin{bmatrix} \sin\beta\cos\psi + 5\sin\zeta \\ 0 \\ 10.5\sin\psi\log x \end{bmatrix}$)

FORTRAN Statements

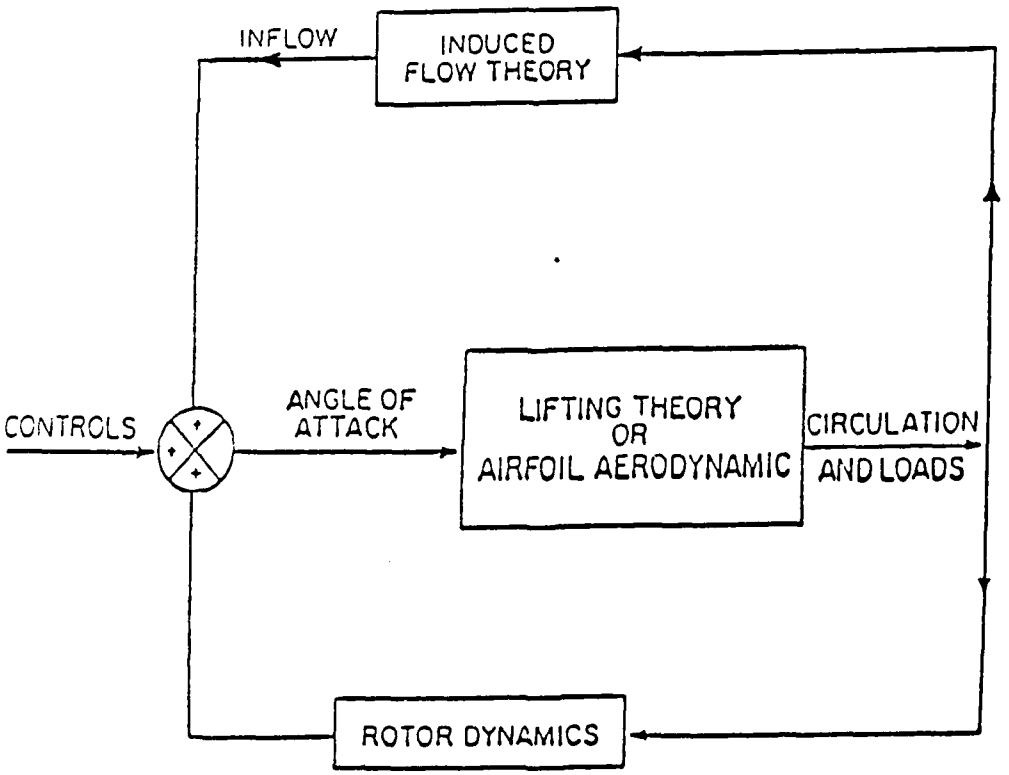
```
PEQN(5,1)=1.125*GAH2+1.5*MU*SIN(CY)**2
PEQN(6,3)=5.*BETA*COS(BETA)
```

(Fortran statements of equations P_{51} and P_{63} which correspond to $1.125\ \gamma_2 + 1.5\mu \sin^2(\psi)$ and $.5\beta \cos\beta$, respectively, where γ_2 is an aerodynamic force integral and μ, a dimensionless speed parameter).

VARIABLES	DESCRIPTION	REPRESENTATION
ρ	Position Vector from hub axis	R
$\bar{e}$	Hinge off-set/Rotor Radius	HEPS
$\bar{r}$	Location of the blade element (Dimensionless: r/(rotor radius - hinge offset)) from hinge axis	RB
R_h	Hub elasticity parameter	
θ	Blade pitch setting	



ROTOR SCHEMATIC



DYNAMIC INFLOW BLOCK DIAGRAM

Fig.1 ROTOR SYSTEM WITH DYNAMIC INFLOW FEEDBACK

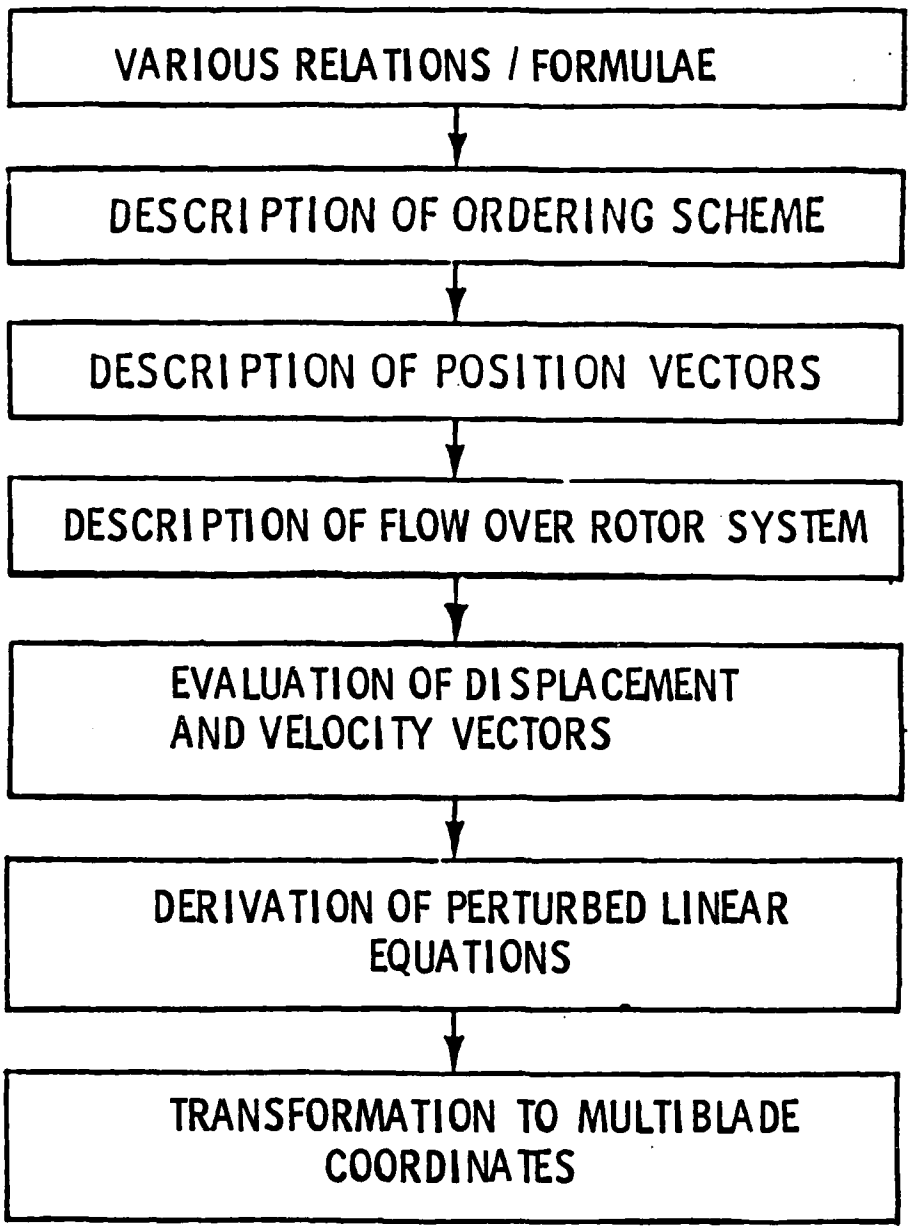


FIG. 2 INPUT BLOCK DIAGRAM

HYPERBOLIC WAVES AND NONLINEAR GEOMETRICAL ACOUSTICS

John K. Hunter

Colorado State University

ABSTRACT This paper reviews asymptotic methods for weakly nonlinear hyperbolic waves. When applied to compressible fluid flows, these methods give a theory of nonlinear geometrical acoustics.

1 INTRODUCTION

Nonlinear wave propagation is a unified scientific field largely because the basic phenomena are described by a relatively small number of canonical equations. These equations can be derived systematically from the primitive equations modelling the wave motion by means of formal asymptotic expansions. The aim of this paper is to summarise the canonical equations for weakly nonlinear hyperbolic waves, with or without the inclusion of small dissipative effects. We apply these results to the equations of motion of a compressible fluid, which gives a theory of nonlinear geometrical acoustics (NGA).

Exact solutions of the canonical equations are of particular interest. As well as providing quantitative asymptotic solutions, they often give considerable insight into the physical processes of the wave motion. We therefore note what is currently known about exact solutions of the canonical equations presented here. Unfortunately, in comparison with the situation for dispersive waves, few exact solutions are known. For example, the cylindrical KdV equation, the KP equation,

and the three wave resonant interaction equations are all solvable by the inverse scattering transform [1], [11]. None of the corresponding asymptotic equations for dissipative waves can be solved completely.

In section 2, we derive asymptotic equations for a single wave. The result is the kinematic wave equation (2.4) for inviscid waves, and the generalized Burgers' equation (2.19) for viscous waves. In section 3, we include diffraction. This gives the unsteady transonic small disturbance equation (3.5), and the Kuznetsov equation (3.7). In section 4, we consider wave–wave and wave–mean field interactions, leading to the integro–differential equations (4.4) and (4.13). Finally, in section 5, these equations are specialized to the case of sound waves in a fluid.

Our references are biased towards review articles and books, where they are available. These may be consulted for references to the original papers. For other reviews of asymptotic theories for hyperbolic waves, see Nayfeh [33] and Majda [28].

2 SINGLE WAVE EQUATIONS

2.1 The kinematic wave equation

Let us begin by considering a hyperbolic system of conservation laws in one space dimension,

$$U_t + f(U)_x = 0, \tag{2.1}$$

where $U(x,t) \in \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^m$. The weakly nonlinear expansion of solutions of (2.1) is

$$U = \epsilon a(\frac{x-\lambda t}{\epsilon},x,t)r + O(\epsilon^2), \quad \epsilon \to 0+ \quad \text{with } t = O(1). \tag{2.2}$$

Here, λ is an eigenvalue of

$$A = \nabla_U f(0),$$

and $\mathbf{r}$ is a corresponding eigenvector. We denote a left eigenvector by $\boldsymbol{\ell}$, and normalize it so that $\boldsymbol{\ell}\cdot\mathbf{r} = 1$. We assume throughout that λ is a simple eigenvalue (see [5], [17] for multiple characteristics). The wave amplitude $a(\theta,x,t)$ is determined by means of the method of multiple scales [21]. In this method, (θ,x,t) are treated as independent variables, and the final asymptotic solution is obtained by evaluating θ at $\epsilon^{-1}(x-\lambda t)$. The equation for a is found to be

$$a_t + \lambda a_x + Maa_\theta = 0. \tag{2.3}$$

In (2.3), the coefficient of the nonlinear term is

$$M = \nabla_U \lambda(U)\cdot\mathbf{r}(U)\big|_{U=0} = \boldsymbol{\ell}\cdot\nabla_U{}^2 f(0)(\mathbf{r},\mathbf{r}) = \sum_{i,j,k=1}^{m} \ell_i \frac{\partial^2 f_i}{\partial U_j \partial U_k} r_j r_k.$$

Thus, $M \neq 0$ if the wave is genuinely nonlinear [23]. If $M = 0$, nonlinear effects are negligable to leading order in ϵ and over times for which the asymptotic solution (2.2) is valid. Nonlinear effects may be retained by rescaling the amplitude in (2.2). In this paper we always assume that $M \neq 0$. For a discussion of other cases see [24], [40].

The amplitude a in (2.2) depends on a "fast" phase $\epsilon^{-1}\phi$, where $\phi = x-\lambda t$, and on "slow" variables (x,t). In this form, (2.2) is a weakly nonlinear extension of the high–frequency geometrical optics expansion. Therefore, one name for this method is weakly nonlinear geometrical optics (WNGO) [17]. Alternatively, note that (2.1) is invariant under the change of variables

$$\bar{x} = \epsilon^{-1}x, \quad \bar{t} = \epsilon^{-1}t.$$

In these variables, (2.2) is

$$U = \epsilon a(\bar{x}-\lambda\bar{t},\epsilon\bar{x},\epsilon\bar{t})\mathbf{r} + O(\epsilon^2), \qquad \epsilon \to 0+ \quad \bar{t} = O(\epsilon^{-1}),$$

which corresponds to a long–time/far–field expansion. It describes a wave which

changes slowly in a frame of reference moving with the linearized wave velocity λ. This point of view is adopted in the perturbation–reduction method (Taniuti et al [44]). Thus, provided that there are no lower order source terms in (2.1), the high–frequency and far–field expansions are equivalent.

If $a = a(\theta,t)$ is independent of x, then (2.2) describes a wave which is uniform in space and changing slowly in time. This form is appropriate for initial value problems. If $a = a(\theta,x)$ is independent of t, then (2.2) describes a wave which is modulated in space, and distance x is a "time–like" variable in (2.3). This form is appropriate for boundary value problems.

The wave–form of the wave in (2.2) is described by the dependence of a on θ. There are two main cases: oscillatory waves, and wavefronts. For oscillatory waves, a is a periodic (or almost periodic) function of θ, and (2.2) is valid in the limit

$$\epsilon \to 0+ \quad \text{with} \quad x,t = O(1).$$

The solution represents a rapidly oscillating wave field, with frequency and wavenumber of the order ϵ, which is modulated over distances and times of the order one.

A typical example of a wavefront expansion is when

$$a(\theta,x,t) \to a_+(x,t) \text{ as } \theta \to +\infty,$$

$$a(\theta,x,t) \to a_-(x,t) \text{ as } \theta \to -\infty,$$

as in a viscous shock profile. In this case, (2.2) is valid near the wavefront $x = \lambda t$, but not necessarily elsewhere. Formally, the limit is

$$\epsilon \to 0+ \quad \text{with} \quad t = O(1) \text{ and } x-\lambda t = O(\epsilon).$$

To put (2.3) in a standard form, we define

$$u(\theta,t,x-\lambda t) = Ma(\theta,x,t).$$

This implies that $u(x,t,\phi)$ satisfies

(2.4) $$u_t + uu_x = 0.$$

Here, we have renamed the independent variables – x in (2.4) is the phase variable, and not the original space variable – and ϕ occurs as a parameter. Equation (2.4) is called the *kinematic wave equation*, or the *inviscid Burgers' equation.* It is the canonical equation for a weakly nonlinear, genuinely nonlinear, hyperbolic wave.

Weak solutions of (2.4), in the conservative form

$$u_t + \frac{\partial}{\partial x}(\frac{1}{2}u^2) = 0,$$

continue to provide formally valid asymptotic solutions of (2.1) after shocks form [3].

Equation (2.4) can be solved exactly, in principle, by using the method of characteristics and shock fitting [46].

2.2 Nonplanar waves

The method of the previous section generalizes to nonplanar waves in several space dimensions and to nonuniform media which vary slowly over a wavelength. Consider a hyperbolic system of conservation laws,

(2.5) $$\sum_{i=0}^{n} \frac{\partial}{\partial x_i} \mathbf{f}_i(\mathbf{x},U) + \mathbf{g}(\mathbf{x},U) = 0,$$

where $\mathbf{x} = (x_0,\ldots,x_n) \in \mathbb{R}^{n+1}$, $U(\mathbf{x}) \in \mathbb{R}^m$, and $\mathbf{f}_i,\mathbf{g} : \mathbb{R}^{n+1}\times\mathbb{R}^m \to \mathbb{R}^m$. We assume that $U = 0$ is a solution of (2.5). Given a smooth nonzero solution of (2.5), this can always be arranged by a shift of dependent variables.

The weakly nonlinear solution of (2.5) is

(2.6) $$U = \epsilon a(\frac{\phi(\mathbf{x})}{\epsilon},\mathbf{x})\mathbf{r}(\mathbf{x}) + O(\epsilon^2), \quad \text{as } \epsilon \to 0+ \text{ with } \mathbf{x} = O(1).$$

This solution represents a small amplitude, locally plane wave. To show the latter fact, we expand U for $\mathbf{x}$ close to a fixed value $\mathbf{y}$. From (2.6) we find that

(2.7) $$U = \epsilon a[\frac{\boldsymbol{\kappa}\cdot(\mathbf{x}-\mathbf{y})}{\epsilon} + \frac{\phi(\mathbf{y})}{\epsilon},\mathbf{y}]\mathbf{r}(\mathbf{y}) + o(\epsilon), \quad \text{as } \epsilon \to 0+ \text{ with } \mathbf{x}-\mathbf{y} = o(\epsilon^{1/2}),$$

where

$$\kappa = D\phi(\mathbf{y}) = (\phi_{x_0},\ldots,\phi_{x_n})|_{\mathbf{x}=\mathbf{y}}.$$

For fixed $\mathbf{y}$, (2.7) is a plane wave with frequency–wavenumber vector $\epsilon^{-1}\kappa$.

The choice of scaling in (2.7), namely

dimensionless wave amplitude = $O(\epsilon)$,

relative change in frequency per period $\sim \dfrac{|\epsilon^{-1}D\kappa|}{|\epsilon^{-1}\kappa|}\,\epsilon = O(\epsilon)$,

leads to a balance between nonlinear and nonplanar effects. For amplitude $<< \epsilon$, one obtains linear geometrical optics in a first approximation; for $\epsilon <<$ amplitude $<< 1$ one obtains the weakly nonlinear plane wave solution described in section 2.1.

Equation (2.6) is an asymptotic solution of (2.5) if:

(a) The phase $\phi(\mathbf{x})$ solves the *eikonal equation*

$$\det[\sum_{i=0}^{n} \phi_{x_i} A_i] = 0, \tag{2.8}$$

where

$$A_i(\mathbf{x}) = \nabla_U f_i(\mathbf{x},0). \tag{2.9}$$

We denote right and left null–vectors of the matrix in (2.8) by $\mathbf{r}(\mathbf{x})$ and $\ell(\mathbf{x})$.

(b) The amplitude $a(\theta,\mathbf{x})$ solves the *transport equation*

$$a_s + Maa_\theta + Na = 0, \tag{2.10}$$

where

$$\frac{\partial}{\partial s} = \sum_{i=0}^{n} \ell\cdot A_i \mathbf{r}\frac{\partial}{\partial x_i},$$

$$M(\mathbf{x}) = \sum_{i=0}^{n} \phi_{x_i} \ell\cdot \nabla_U{}^2 f_i(\mathbf{x},U_1\mathbf{r},\mathbf{r}), \tag{2.11}$$

$$N(\mathbf{x}) = \sum_{i=0}^{n} \ell\cdot\frac{\partial}{\partial x_i}(A_i\mathbf{r}) + \ell\cdot\nabla_U g(\mathbf{x},0)\mathbf{r}.$$

To decribe the structure of these equations, we introduce the rays associated

with ϕ. The rays are curves in space–time $\mathbb{R}^{n+1}$ with parametric equation $\mathbf{x} = \mathbf{x}(s)$ defined by

$$\frac{d}{ds}x_i = \ell \cdot A_i r.$$

There is an n–parameter family of rays, which we label by $\beta(x) \in \mathbb{R}^n$. The above equations are valid in regions of space–time where the transformation between ray coordinates (s,β) and $\mathbf{x}$ is smooth. We shall write functions of $\mathbf{x}$ as functions of (s,β) whenever it is convenient to do so.

The eikonal equation (2.8) can be written as a system of ODE's along the rays associated with ϕ [17]. Equation (2.10) may be regarded as an n–parameter family of PDE's in one "space" variable, with one PDE for each ray. The time–like variable, s, is a one–to–one function of arclength along a ray and the space–like variable θ is the fast phase. The coefficients in (2.10) are functions of s (and the ray parameters β), but are independent of θ.

The eikonal equation (2.8) states that the local frequency–wavenumber vector $\epsilon^{-1}\kappa$ satisfies the local, linearized, high–frequency dispersion relation of (2.5). The transport equation is an energy balance equation for the wave.

The velocity of the rays (with $x_0 = t$ = time),

$$\frac{d}{dt}x_i = (\ell \cdot A_0 r)^{-1} \ell \cdot A_i r, \qquad i = 1,\ldots,n,$$

is called the group velocity of the wave. The phase velocity is the normal velocity of the wavefronts ϕ = constant. It is given by

$$-\phi_t |\nabla\phi|^2 \nabla\phi,$$

where

$$\nabla\phi = (\phi_{x_1},\ldots,\phi_{x_n}).$$

The phase and group velocities are not the same, in general. Equation (2.10) shows that wave energy propagates at the group velocity.

Because of the importance of rays, these theories are often called ray methods. The geometry of rays was first used to study the propagation of light, hence the name geometrical optics.

Equation (2.8) defines the same phase, and therefore the same rays, in the linear and the weakly nonlinear cases. Ostrovsky [37] therefore calls the weakly nonlinear theory the "linear ray approximation". An alternative point of view is obtained by writing a solution of (2.10) in the form

$$a(\theta,\mathbf{x}) = \alpha(s,\beta)F(\xi,\beta). \tag{2.12}$$

In (2.12), F is arbitrary, while α and $\xi(\theta,\mathbf{x})$ satisfy

$$\alpha_s + N\alpha = 0, \tag{2.13}$$

$$\theta = \xi + F(\xi,\beta) \int_0^s M(s',\beta)\alpha(s',\beta)ds'. \tag{2.14}$$

According to (2.12) – (2.14), the solution is of the same form as the linearized solution, but it depends on a perturbed phase $\xi(\epsilon^{-1}\phi,\mathbf{x})$, instead of on the linear phase $\epsilon^{-1}\phi$. This point of view is used in the analytical method of charactistics [22].

To put (2.10) in standard form, we define the new variables [26],

$$\begin{aligned} a(\theta,\mathbf{x}) &= E(s,\beta)u(\theta,s,\beta), \\ \sigma(s,\beta) &= \int_0^s M(s',\beta)E(s,\beta)\, ds', \\ E(s,\beta) &= \exp -\int_0^s N(s',\beta)\, ds'. \end{aligned} \tag{2.15}$$

Using (2.15) in (2.10), and renaming the independent variables, implies that $u(x,t,\beta)$ satisfies the kinematic wave equation (2.4). Thus, (2.10) can be solved by using the method of characteristics and shock fitting.

2.3 The generalized Burgers' equation

Now consider (2.5) with a "viscous" term,

$$\sum_{i=0}^{n} \frac{\partial}{\partial x_i} f_i(\mathbf{x},U) + g(\mathbf{x},U) = \mu \sum_{i,j=0}^{n} \frac{\partial}{\partial x_i}[D_{ij}(\mathbf{x},U)U_{x_j}]. \tag{2.16}$$

In (2.16), the D_{ij} are m×m matrices, and μ is a scalar (the "viscosity"). A balance between small nonlinear, nonplanar, and viscous effects is obtained in the limit (2.6), with

$$\mu = \epsilon^2\hat{\mu}, \quad \hat{\mu} = O(1) \quad \text{as} \quad \epsilon \to 0+.$$

The phase satisfies (2.8), as before, and the transport equation is

$$a_S + Maa_\theta + Na = Pa_{\theta\theta}, \tag{2.17}$$

where M and N are given in (2.11) and

$$P = \hat{\mu} \sum_{i,j=0}^{n} \phi_{x_i}\phi_{x_j}\boldsymbol{\ell}\cdot D_{ij}(\mathbf{x},0)\mathbf{r}. \tag{2.18}$$

The change of variables (2.15) puts (2.17) in the form

$$u_t + uu_x = \nu(t)u_{xx}, \tag{2.19}$$

where

$$\nu(s,\beta) = P(s,\beta)E^{-1}(s,\beta)M^{-1}(s,\beta), \tag{2.20}$$

and we write $\nu(t,\beta)$ as $\nu(t)$ in (2.19).

When ν is constant, (2.19) is called *Burgers' equation*, and when ν depends on t, it is called the *generalized Burgers' equation* (*GBE*). (It is customary to use "generalized" to denote variable coefficients, and "modified" to denote altered nonlinear terms.)

The initial value problem for Burgers' equation can be solved exactly by the Cole–Hopf transform [46]. The Cole–Hopf transform is a Backlund transform [39] between Burgers' equation and the linear heat equation. However, Nimmo and Crighton [34] show that ν = constant is the only case in which there is a Backlund transform connecting (2.19) with another parabolic PDE. Apart from some similarity solutions [41], [42], it seems necessary to use asymptotic or numerical methods to solve the GBE [35], [42].

Viscous effects on waves modelled by (2.16) are weak when $\epsilon >> \mu$, where the wavelength is of the order ϵ. Thus, this expansion uses a long wave approximation. Other long wave equations (e.g. the KdV equation for dispersive waves) are reviewed by Rosales [40].

3 DIFFRACTION

The equations described in the previous section are all based on a locally one–dimensional approximation to the wave. In this section, we describe some asymptotic equations, involving two space variables, which incorporate diffractive effects.

3.1 Weak transverse diffraction

First, let us consider a hyperbolic system of conservation laws in two space dimensions,

$$U_t + f(U)_x + g(U)_y = 0. \tag{3.1}$$

The linearized phase velocity of a wave propagating in the x–direction is $(\lambda,0)^T$, where λ is an eigenvalue of

$$A = \nabla_U f(0).$$

We assume that $\lambda \neq 0$. We denote associated right and left eigenvectors by $\mathbf{r}$ and $\boldsymbol{\ell}$ and normalize $\boldsymbol{\ell}$ so that $\boldsymbol{\ell}\cdot\mathbf{r} = 1$. The group velocity of the wave is $(\lambda,\mu)^T$ where

$$\mu = \boldsymbol{\ell}\cdot B\mathbf{r}, \quad B = \nabla_U g(0).$$

For anisotropic waves, μ is generally nonzero. The equations of the spatial projections of the rays are

(3.2) $\quad y - \lambda^{-1}\mu x = \text{constant}.$

The following ansatz desribes a weakly nonlinear wave propagating in the x–direction and diffracting slowly in a direction orthogonal to the ray projections (3.2):

(3.3) $\quad U = \epsilon a(\frac{x-\lambda t}{\epsilon}, \frac{y-\lambda^{-1}\mu x}{\epsilon^{1/2}}, x, y, t)\mathbf{r} + O(\epsilon^{3/2}), \quad \epsilon \to 0+ \quad \text{with} \quad t = O(1).$

The wave amplitude $a(\theta,\eta,x,y,t)$ satisfies

(3.4) $\quad \frac{\partial}{\partial\theta}(a_t + \lambda a_x + \mu a_y + Maa_\theta) + Qa_{\eta\eta} = 0.$

In (3.4),

$$M = \boldsymbol{\ell}\cdot\nabla_U^2 f(0)(\mathbf{r},\mathbf{r}),$$

$$Q = \boldsymbol{\ell}\cdot(B-\mu I)\mathbf{s},$$

where $\mathbf{s}$ is a solution of

$$(A-\lambda I)\mathbf{s} + (B-\mu I)\mathbf{r} = 0.$$

Assuming that M and Q are nonzero, and rescaling variables in (3.4),

$$a(\theta,\eta,x,y,t) = (MQ)^{-1}u(Q\theta,\eta,t,x-\lambda t,y-\mu t),$$

implies that $u(x,y,t,\phi,\psi)$ solves

(3.5) $\quad \frac{\partial}{\partial x}(u_t + uu_x) + u_{yy} = 0.$

Equation (3.5) was first derived in the context of transonic flows, and it is called the *unsteady transonic small disturbance equation* (*UTSDE*) [6]. The equation is a weakly nonlinear extension of the parabolic approximation to the wave equation [2].

For the general system (2.16), the asymptotic solution is [20]

$$U = \epsilon a(\frac{\phi(\mathbf{x})}{\epsilon}, \frac{\psi(\mathbf{x})}{\epsilon^{1/2}}, \mathbf{x})\mathbf{r}(\mathbf{x}) + O(\epsilon^{3/2}),$$

where ϕ is a solution of the eikonal equation (2.8), $\mathbf{r}$ and $\boldsymbol{\ell}$ are associated right and left eigenvectors, and ψ is constant along the rays asociated with ϕ i.e.

$$\sum_{i=0}^{n} \boldsymbol{\ell}\cdot A_i\mathbf{r}\ \psi_{x_i} = 0.$$

The amplitude $a(\theta,\eta,\mathbf{x})$ satisfies

$$\frac{\partial}{\partial\theta}(a_s + Maa_\theta + Na - Pa_{\theta\theta}) + Qa_{\eta\eta} = 0.$$

Here, ∂_s, M, N, and P are defined in (2.11) and (2.18), and

$$Q = \boldsymbol{\ell}\cdot Ks,$$

where s is a solution of

$$Hs + Kr = 0,$$

and

$$H = \sum_{i=0}^{n} \phi_{x_i} A_i, \quad K = \sum_{i=0}^{n} \psi_{x_i} A_i.$$

We make the change of variables

$$a(\theta,\eta,\mathbf{x}) = E(s,\beta)u(\theta,\eta,\sigma,\beta), \tag{3.6}$$

where E and σ are defined in (2.15). Then $u(x,y,t,\beta)$ satisfies

$$\frac{\partial}{\partial x}[u_t + uu_x - \nu(t)u_{xx}] + \delta(t)u_{yy} = 0, \tag{3.7a}$$

In (3.7a), ν is defined in (2.20) and

$$\delta = QE^{-1}M^{-1}.$$

This equation can be written in system form,

$$u_t + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2 - \nu u_x) + \frac{\partial}{\partial y}(\delta v) = 0, \tag{3.7b}$$

$$u_y - v_x = 0,$$

and in potential form for ϕ, where $u = \phi_x$ and

$$\phi_{xt} + \phi_x\phi_{xx} - \nu\phi_{xxx} + \delta\phi_{yy} = 0. \tag{3.7c}$$

If $\nu = 0$, solutions of (3.7) may contain shocks. The jump conditions for (3.7b) imply that, if the equation of the shock position is $x = s(y,t)$, then

$$s_t = < u > + \delta s_y^{\,2}.$$

Here,

$$< u > = \tfrac{1}{2}\{\lim_{x\to s+} u(x,y,t) + \lim_{x\to s-} u(x,y,t)\}$$

is the average of u ahead of and behind the shock.

For plane waves, ν and ρ are constants, and the change of variables

$$t \to (\nu\rho^2)^{1/3}t, \qquad x \to (\nu^{-1}\rho)^{1/3}x$$

transforms (3.7a) to

$$\frac{\partial}{\partial x}(u_t + uu_x - u_{xx}) + u_{yy} = 0. \tag{3.8}$$

Equation (3.8) is known by several names: the *2-D Burgers' equation*; the *Kuznetsov equation*; the *Zabolotskaya-Khokhlov (ZK) equation*; or the *Kuznetsov-Zabolotskaya-Khokhlov (ZKZ) equation.* Equation (3.7) is the *generalized Kuznetsov equation (GKE).*

A GKE may be transformed into another GKE with different coefficients by the change of variables

$$\begin{aligned}
\bar{t} &= K\int_0^t |\rho|^{-1/2}dt, \\
\bar{x} &= x + (4\rho)^{-1}y^2, \\
\bar{y} &= L\rho^{-1}y, \\
\bar{u} &= K^{-1}|\rho|^{1/2}u,
\end{aligned}$$

where $\delta(t) = \rho'(t)$ and K and L are constants. Then $\bar{u}$ satisfies

$$\frac{\partial}{\partial \bar{x}}[\bar{u}_{\bar{t}} + \overline{uu}_{\bar{x}} - \bar{\nu}(\bar{t})\bar{u}_{\overline{xx}}] + \bar{\delta}(\bar{t})\bar{u}_{\overline{yy}} = 0,$$

with

$$\bar{\nu} = K^{-1}|\rho|^{1/2}\nu, \qquad \bar{\delta} = L^2|\rho|^{-3/2}\delta.$$

If $\nu = t^p$ and $\delta = t^q$, then $\bar{\nu} = \bar{t}^{\bar{p}}$ and $\bar{\delta} = \bar{t}^{\bar{q}}$, where

$$\bar{p} = \frac{2p + q + 1}{1 - q}, \qquad \bar{q} = -\frac{3 + q}{2}.$$

For the cylindrical GKE (see section 5.3.1) p = 1 and q = −3. This transformation therefore reduces it to the planar Kuznetsov equation.

Unfortunately, not many exact solutions of these equations are known. Travelling wave solutions of the UTSDE (3.5),

$$u = c + U(x-ct,y),$$

satisfy the *steady transonic small disturbance equation* (*STSDE*), or *Karman–Guderley equation*,

(3.9a) $$\frac{\partial}{\partial x}(UU_x) + U_{yy} = 0.$$

This is often written in terms of a potential ϕ, where $U = \phi_x$ and

(3.9b) $$\phi_x\phi_{xx} + \phi_{yy} = 0.$$

Equation (3.9) is nonlinear and of mixed type. It is hyperbolic if $U < 0$ and elliptic if $U > 0$. Solutions can be found by using: (a) group invariance properties (similarity solutions); (b) the hodograph transformation [6]. The hodograph transformation linearizes (3.9), but it is difficult to use if there are shocks.

Another reduction of the UTSDE's to a system with two independent variables is obtained for scale–invariant solutions depending on $t^{-1}x$ and $t^{-1}y$. It is convenient to transform (3.5) to "polar" coordinates

$$r = x + \frac{y^2}{4t}, \quad \theta = \frac{y}{t},$$
$$w(r,\theta,t) = u(x,y,t),$$

which gives the cylindrical UTSDE

$$\frac{\partial}{\partial r}(w_t + ww_r + \frac{1}{2t}w) + \frac{1}{t^2}w_{\theta\theta} = 0.$$

The similarity solutions

$$w = w(\rho,\theta), \quad \rho = \frac{r}{t} = \frac{x}{t} + \frac{1}{4}(\frac{y}{t})^2,$$

satisfy

(3.10a) $$\frac{\partial}{\partial\rho}[(w-\rho)w_\rho + \frac{1}{2}w] + w_{\theta\theta} = 0.$$

The potential form is $w = \phi_\rho$ and

(3.10b) $$(\phi_\rho-\rho)\phi_{\rho\rho} + \phi_{\theta\theta} + \frac{1}{2}\phi_\rho = 0.$$

Equation (3.10) cannot be linearized by the hodograph transformation (because of the lower order term), but Zahalak and Myers [47] found some particular solutions in the

hodograph plane.

Gibbons and Kodama [12] give a generalization of the hodograph transformation which applies directly to the UTSDE (3.5). They use it to construct a family of solutions which are polynomials in appropriate dependent variables.

There seem to be no known exact solutions of the Kuznetsov equation (3.8) *that involve nontrivial nonlinearity, dissipation, and diffraction. Treatments thus far* have used numerical or approximate methods [41].

The UTSDE describes weakly nonlinear waves at singul ar rays [20]; in particular (3.10) decribes a shock at a singular ray [47], [15], [20]. The UTSDE is also a nonlinear caustic equation (see the next section); and it should describe the transition from regular to Mach reflection for weak shocks. The Kuznetsov equation has been used extensively to model acoustic beams, especially in connection with parametric acoustic arrays [14].

3.2 Caustics

The straightforward ray method expansions described in section 2 break down near caustics. A caustic is an envelope of rays. Straighforward ray methods predict that the wave amplitude is infinite at a caustic. In fact, the wave amplitude is limited by diffraction, and remains finite. However, the ratio of the amplitude at the caustic to the amplitude away from the caustic is unbounded as the wavelength $\epsilon \to 0+$. Therefore a modified asymptotic expansion is needed to describe a wave near a caustic.

The simplest case is a smooth convex caustic. On one side of the caustic (the "illuminated" region) there are two waves, an incident and a reflected wave. On the other side (the "shadow" region), the straightforward ray method expansion

predicts that there is no wave field; in fact, according to linearized theory, the wave amplitude decays exponentially into the shadow region. The illuminated region is doubly covered by rays, while no rays reach the shadow region.

Suppose that a wave forms a smooth convex caustic at the surface $\psi(\mathbf{x}) = 0$, and let the phases of the waves in the illuminated region $\psi > 0$ be

$$\phi \pm \frac{2}{3}\psi^{3/2}.$$

The associated null vectors are of the form

$$\mathbf{r} \pm \psi^{1/2}\mathbf{s}.$$

According to the linearized caustic theory for (2.5) [27], ϕ, ψ, $\mathbf{r}$, and $\mathbf{s}$ satisfy:

$$\det\begin{bmatrix} H & \psi K \\ K & H \end{bmatrix} = 0,$$
$$\begin{bmatrix} H & \psi K \\ K & H \end{bmatrix}\begin{bmatrix} \mathbf{r} \\ \mathbf{s} \end{bmatrix} = 0,$$

where

$$(3.11) \qquad H = \sum_{i=0}^{n} \phi_{x_i} A_i, \qquad K = \sum_{i=0}^{n} \psi_{x_i} A_i.$$

Here, A_i is defined in (2.9). We also use the left null vectors $\boldsymbol{\ell}$, $\mathbf{m}$ defined by

$$(\mathbf{m}, \boldsymbol{\ell})\cdot\begin{bmatrix} H & \psi K \\ K & H \end{bmatrix} = 0$$

The weakly nonlinear caustic expansion of solutions of (2.6) is

$$U = \epsilon^{2/3} a\left(\frac{\phi(\mathbf{x})}{\epsilon}, \frac{\psi(\mathbf{x})}{\epsilon^{2/3}}, \mathbf{x}\right)\mathbf{r}(\mathbf{x}) + O(\epsilon),$$
$$\text{as } \epsilon \to 0+ \text{ with } \mathbf{x} = O(1),\ \psi = O(\epsilon^{2/3}).$$

It uses the same phases as the linearized theory, but the amplitude $a(\theta,\eta,\mathbf{x})$ solves a nonlinear equation,

$$(3.12) \qquad \frac{\partial}{\partial\theta}[(Ma-\eta)a_\theta] + a_{\eta\eta} = 0.$$

In (3.12),

$$M = \boldsymbol{\ell}\cdot\sum_{i=0}^{n} \phi_{x_i} \nabla_U^2 f_i(0)(\mathbf{r},\mathbf{r}).$$

Equation (3.12) was derived by Giraud [13] and Hayes [16] for gas dynamics, and by

Hunter and Keller [19] for general systems.

To put (3.12) in a standard form, we define

$u(\theta,\eta) = Ma(\theta,\eta)$,

and we do not show the **x**–dependence explicitly, since **x** occurs in (3.12) as a parameter. Then u(x,y) satisfies

(3.13) $$\frac{\partial}{\partial x}[(u-y)u_x] + u_{yy} = 0.$$

The change of variables $u \to y+u$, reduces (3.13) to the STSDE (3.9).

When a smooth, weakly nonlinear wave (2.6) forms a caustic, its amplitude near the caustic is of the order $\epsilon^{5/6} << \epsilon^{2/3}$ [19]. Thus, the wave is described by the linearized version of (3.13) i.e. the Tricomi equation. However, if the incident wave contains a shock, the linearized theory is inconsistent, because it predicts a logarithmically infinite singularity in the reflected wave. Seebass [43] uses the nonlinear equation (3.13) to analyze a weak shock at a smooth caustic (although a complete formal justification of this procedure has not beeen given). He reduces (3.13) to the STSDE, and uses the hodograph transformation, but it does not seem possible to satisfy the shock conditions exactly.

A cusped caustic (or arete) is described by three functions, $\phi(\mathbf{x})$, $\psi(\mathbf{x})$, and $\chi(\mathbf{x})$. The caustic is at

$$(\frac{\psi}{2})^2 - (\frac{\chi}{3})^3 = 0,$$

and the cusp is at

$$\psi = \chi = 0.$$

To define the phases, let

$$\phi(\mathbf{x},\xi) = \phi(\mathbf{x}) + \psi(\mathbf{x})\xi - \frac{1}{2}\chi(\mathbf{x})\xi^2 + \frac{1}{4}\xi^4,$$

and denote the solutions of

$$\phi_\xi = \xi^3 - \chi\xi + \psi = 0$$

by $\xi = \xi_j(\mathbf{x})$. Then the phases are

$$\phi_j(\mathbf{x}) = \phi(\mathbf{x}, \xi_j(\mathbf{x})).$$

There are three phases inside the caustic and one (real) phase outside the caustic.

It follows from the linearized theory [27] that these functions, and their associated null–vectors, satisfy

$$\det L = 0,$$

$$L\begin{bmatrix} r \\ s \\ t \end{bmatrix} = 0, \quad (n, m, \ell)\cdot L = 0,$$

where

$$L = \begin{bmatrix} H & \frac{1}{2}\psi K & -\psi J \\ J & H - \frac{1}{2}\chi K & \psi J + \frac{1}{2}\psi K \\ \frac{1}{2}K & J & H - \frac{1}{2}\chi K \end{bmatrix}.$$

Here, H, K are defined in (3.11), and

$$J = \sum_{i=0}^{n} \chi_{x_i} A_i.$$

The weakly nonlinear cusped caustic expansion of (2.6) is

$$U = \epsilon^{1/2} a\left(\frac{\phi(\mathbf{x})}{\epsilon}, \frac{\psi(\mathbf{x})}{\epsilon^{3/4}}, \frac{\chi(\mathbf{x})}{\epsilon^{1/2}}, \mathbf{x}\right) r(\mathbf{x}) + O(\epsilon^{3/4}),$$

$$\text{as } \epsilon \to 0+ \quad \text{with } \mathbf{x} = O(1), \ \psi = O(\epsilon^{3/4}), \ \chi = O(\epsilon^{1/2}).$$

The equation for the amplitude $a(\theta, \eta, \zeta, \mathbf{x})$ is the UTSDE again,

$$\frac{\partial}{\partial\theta}(2a_\zeta + Maa_\theta) + a_{\eta\eta} = 0,$$

where

$$M = \ell \cdot \sum_{i=0}^{n} \phi_{x_i} \nabla_U^2 f_i(r, r).$$

The UTSDE was derived by Cramer and Seebass [8] for slowly focusing, weak shocks, and studied further by Obermeir [36]. The above derivation suggests that it should also describe strongly focusing weak shocks.

4 INTERACTING WAVES

So far, we have described asymptotic equations for a single wave. When several waves are present, they may interact and generate new waves. A wave may also generate a mean–field. The asymptotic equations for such processes are integro–differential equations.

4.1 Wave–wave interactions

For simplicity, we consider interactions between collinear waves, when the problem involves one space dimension. Interactions between oblique plane waves are described by similar equations, but the details are more complicated [18]. The asymptotic analysis for nonplanar wave interactions usually leads to a passage–through–resonance problem which has not been solved.

We denote the eigenvalues of $A = \nabla_U f(0)$ in (2.1) by

$$\lambda_1 < \lambda_2 < \ldots < \lambda_m,$$

and right and left eigenvectors associated with λ_j are denoted by $\mathbf{r}_j$ and $\boldsymbol{\ell}_j$. We normalize $\boldsymbol{\ell}_j$ so that $\boldsymbol{\ell}_j \cdot \mathbf{r}_j = 1$.

The asymptotic solution for m interacting, weakly nonlinear waves is

$$U = \epsilon \sum_{j=1}^{m} a_j\left(\frac{k_j x - \omega_j t}{\epsilon}, x, t\right)\mathbf{r}_j + O(\epsilon^2), \quad \epsilon \to 0+ \text{ with } x,t = O(1). \tag{4.1}$$

In (4.1), the wave amplitudes $a_j(\theta,x,t)$, and their derivatives, are zero–mean, almost periodic functions of θ. It is convenient to introduce explicitly wavenumbers k_j and frequencies ω_j, which satisfy the dispersion relation

$$\omega_j = \lambda_j k_j, \qquad j = 1,\ldots,m.$$

We define

$$\mu_{jpq} = \frac{k_j(\lambda_p - \lambda_j)}{k_p(\lambda_p - \lambda_q)},$$

for distinct j,p,q, so that

$$\omega_j = \mu_{jqp}\omega_p + \mu_{jpq}\omega_q,$$
$$k_j = \mu_{jqp}k_p + \mu_{jpq}k_q.$$

The wave amplitudes satisfy the following system of resonant interaction equations [29],

$$a_{jt}(\theta) + \lambda_j a_{jx}(\theta) + \frac{\partial}{\partial\theta}[\tfrac{1}{2}M_j a_j^2(\theta) + \sum_{p<q}^{j} \Gamma_{jpq} \lim_{T\to\infty} \frac{1}{T}\int_0^T a_p(\xi)a_q(\mu_{qpj}\theta + \mu_{qjp}\xi)d\xi] = 0. \tag{4.2}$$

In (4.2), we show only the θ–dependence and

$$\sum_{p<q}^{j} = \text{sum over all pairs (p,q) with } 1 \le p < q \le m,\ p\neq j \text{ and } q\neq j.$$

The coefficients in (4.2) are

$$\Gamma_{jpq} = \ell_j \cdot \nabla_U{}^2 f(0)(r_p, r_q),$$
$$M_j = \Gamma_{jjj}.$$

The simplest case is when: (a) there are three interacting waves e.g. (2.1) is a 3×3 system; (b) the waves are periodic in θ – we normalize the period to 2π; (c) there are no spatial modulations i.e. a_j is independent of x; (d) the frequencies and wave numbers satisfy the resonance condition

$$\omega_1 + \omega_2 + \omega_3 = 0,$$
$$k_1 + k_2 + k_3 = 0.$$

Then (4.2) becomes

$$a_{jt}(\theta) + \frac{\partial}{\partial\theta}[\tfrac{1}{2}M_j a_j^2(\theta) + \Gamma_{jpq}\frac{1}{2\pi}\int_0^{2\pi} a_p(\xi)a_q(-\theta-\xi)d\xi] = 0, \tag{4.3}$$

where (j,p,q) is a cyclic permutation of (1,2,3).

Equations (4.2) and (4.3) are in conservation form. They are valid in the weak sense after shocks form [3].

Rescaling variables in (4.3),

$$u_j(\theta,t) = \tilde{M}_j a_j(\theta,t),$$
$$\tilde{M}_j = \begin{cases} M_j & \text{if } M_j \neq 0 \\ 0 & \text{if } M_j = 0 \end{cases},$$

implies that $\{u_j(x,t)\}$ satisfies

$$\begin{aligned} &u_{1t}(x,t) + \frac{\partial}{\partial x}\{\tfrac{1}{2}\sigma_1 u_1^{\,2}(x,t) + \Gamma_1 \tfrac{1}{2\pi}\int_0^{2\pi} u_2(y,t)u_3(-x-y,t)\,dy\} = 0, \\ (4.4)\qquad &u_{2t}(x,t) + \frac{\partial}{\partial x}\{\tfrac{1}{2}\sigma_2 u_2^{\,2}(x,t) + \Gamma_2 \tfrac{1}{2\pi}\int_0^{2\pi} u_3(y,t)u_1(-x-y,t)\,dy\} = 0, \\ &u_{3t}(x,t) + \frac{\partial}{\partial x}\{\tfrac{1}{2}\sigma_3 u_3^{\,2}(x,t) + \Gamma_3 \tfrac{1}{2\pi}\int_0^{2\pi} u_1(y,t)u_2(-x-y,t)\,dy\} = 0. \end{aligned}$$

In (4.4),

$$\sigma_j = \begin{cases} 1 & \text{if } M_j \neq 0 \\ 0 & \text{if } M_j = 0 \end{cases},$$

$$\Gamma_j = \frac{\tilde{M}_j \Gamma_{jpq}}{M_p M_q}, \quad (j,p,q) = (1,2,3),\ (2,3,1),\ (3,1,2).$$

Smooth solutions of (4.4) satisfy the following conservation laws for $1 \le p < q \le 3$:

$$(4.5)\qquad \Gamma_p \int_0^{2\pi} u_q^{\,2}(x,t)dx - \Gamma_q \int_0^{2\pi} u_p^{\,2}(x,t)dx = \text{constant}.$$

Equation (4.5) is a generalization of the Manley–Rowe relations for dispersive waves [7]. Once shocks form, (4.5) must be modified to allow for the decrease in entropy across a shock (see [31] for the appropriate modification in the case of gas dynamics). If the interaction coeffecients $\{\Gamma_j\}$ are not all of the same sign, then (4.5) implies that solutions of (4.4) are bounded in the L^2–norm. If the interaction coeficients are all of the same sign, then (4.4) has "explosively" unstable solutions, which blow up in finite time. The weakly nonlinear approximation is inconsistent after the blow–up time.

Equation (4.4) simplifies for some special types of solutions, namely: (a) sawtooth waves; (b) travelling waves; (c) separable solutions.

(a) Sawtooth waves

We define the sawtooth function S(x) by

$$S(x) = x, \quad |x| < \pi,$$

$$S(x+2\pi) = S(x).$$

Equation (4.4) has solutions

$$u_j(x,t) = \alpha_j(t)S(x-\zeta_j), \tag{4.6}$$

where

$$\zeta_1 + \zeta_2 + \zeta_3 \equiv \pi \pmod{2\pi},$$

and $\{\alpha_j(t)\}$ satisies the ODE's,

$$\alpha_{jt} + \sigma_j\alpha_j^2 = \Gamma_j\alpha_p\alpha_q. \tag{4.7}$$

The solution (4.6) is admissible if

$$\sigma_j\alpha_j \geq 0.$$

Unless the Γ_j's are all positive, solutions of (4.7) typically become inadmissible after a finite time.

A simple, but interesting, special case of (4.4) which shows what can happen when a sawtooth wave solution becomes inadmissible is

$$M_2 = M_3 = \Gamma_2 = \Gamma_3 = 0,$$

$$M_1 = 1, \quad \Gamma_1 = -1,$$

$$u_2 = u_3 = S(x).$$

The equation for u_1 is

$$u_t + uu_x + S(x) = 0.$$

This has the sawtooth wave solution,

$$u(x,t) = -\tan(t)S(x),$$

which is admissible for $-\frac{\pi}{2} < t < 0$. Majda, Rosales, and Schonbeck [31] show that when $t > 0$, the shocks become "cusped rarefaction waves", containing a square root

singularity. Their solution is: for $0 < t < \frac{\pi}{2}$,

$$u = \begin{cases} -(\pi^2-x^2)^{1/2} & \pi\cos t \le x \le \pi, \\ -x\tan t & -\pi\cos t \le x \le \pi\cos t, \\ (\pi^2-x^2)^{1/2} & -\pi \le x \le -\pi\cos t; \end{cases}$$

and for $t > \frac{\pi}{2}$,

$$u = \begin{cases} -(\pi^2-x^2)^{1/2} & 0 < x \le \pi, \\ (\pi^2-x^2)^{1/2} & -\pi \le x < 0. \end{cases}$$

A particular solution of (4.6) and (4.7) is

$$u_j = t^{-1}K_j, \tag{4.8}$$

where the constants K_j satisfy

$$K_j + \sigma_j K_j^2 = \Gamma_j K_p K_q.$$

If $\sigma_j K_j \ge 0$, $j = 1,2,3$, then (4.8) is admissible for $t > 0$, and decays to zero as $t \to +\infty$. If $\sigma_j K_j \le 0$, then (4.8) is admissible for $t < 0$, and blows up as $t \to 0-$. Otherwise, (4.8) is inadmissible for all t.

(b) Travelling waves

Nonlinearity steepens the wave profile of any genuinely nonlinear hyperbolic wave and periodic travelling waves do not exist. Wave–wave interactions can balance nonlinearity, so that interacting travelling waves are possible. They are described by solutions of (4.4) of the form

$$u_j = U_j(x - c_j t - \zeta_j),$$

where the wave velocities c_j and the phase shifts ζ_j satisfy

$$c_1 + c_2 + c_3 = 0,$$

$$\zeta_1 + \zeta_2 + \zeta_3 \equiv 0 \pmod{2\pi},$$

and $\{U_j\}$ solves a nonlinear system of integral equations,

$$\tfrac{1}{2}\sigma_j U_j^2(x) - c_j U_j(x) + \Gamma_j \tfrac{1}{2\pi}\int_0^{2\pi} U_p(y)U_q(-x-y)dy = K_j. \tag{4.9}$$

In (4.9), K_1, K_2, K_3 are constants.

Pego [38] gives an exact solution in the special case of gas dynamics (see section 5.4). If $\sigma_j = 0$, and $c_j\Gamma_j > 0$ – which implies that the Γ_j are of mixed signs (the nonexplosive case) – (4.9) has the solution,

$$U_j(x) = 2\left(\frac{c_p c_q}{\Gamma_p \Gamma_q}\right)^{1/2} \cos(x).$$

Solutions of (4.9) with $\sigma_j \neq 0$, can be found in the limit $\Gamma_j >> 1$ by perturbing off this solution.

(c) Separable solutions

The separable solutions of (4.4) are

$$u_j(x,t) = t^{-1}X(x), \text{ where}$$

$$\frac{d}{dx}\left[\frac{1}{2}\sigma_j X_j^2(x) + \Gamma_j \frac{1}{2\pi}\int_0^{2\pi} X_p(y)X_q(-x-y)dy\right] = X_j(x). \tag{4.10}$$

A particular solution of (4.10), when the Γ_j have the same signs, is the sawtooth wave solution (4.8). Also, for $\sigma_j = 0$, (4.10) has the solution

$$X_j(x) = 2\gamma|\Gamma_p\Gamma_q|^{-1/2}\cos(x-\zeta_j),$$

where

$$\zeta_1 + \zeta_2 + \zeta_3 \equiv -\frac{\pi}{2} \pmod{2\pi},$$

$$\text{sgn}\Gamma_1 = \text{sgn}\Gamma_2 = \text{sgn}\Gamma_3 = \gamma.$$

Small amplitude solutions of (4.10) with $\sigma_j \neq 0$ may be found by perturbing off this solution in the limit $\Gamma_j >> 1$.

The resonant interaction equations can be generalized in a straightforward way to include other effects, such as weak viscosity [29] or dispersion. For example, consider (3.1), and suppose that the wave motion is isotropic ($\ell_j \cdot Br_j = 0$). Then the

following asymptotic solution describes interacting waves with diffraction:

$$U = \epsilon \sum_{j=1}^{m} a_j(\frac{k_j x-\omega_j t}{\epsilon}, \frac{y}{\epsilon^{1/2}}, x, y, t)r_j + O(\epsilon^2), \quad \epsilon \to 0+ \text{ with } x,t = O(1).$$

The amplitudes $a_j(\theta,\eta,x,y,t)$ satisfy

$$\frac{\partial}{\partial\theta}\{a_{jt}(\theta) + \lambda_j a_{jx}(\theta) + \frac{\partial}{\partial\theta}[\frac{1}{2}M_j a_j^2(\theta) + \sum_{p<q}^{j} \Gamma_{jpq} \lim_{T\to\infty} \frac{1}{T}\int_0^T a_p(\xi)a_q(\mu_{qpj}\theta+\mu_{qjp}\xi)d\xi]\} + Q_j a_{\eta\eta} = 0.$$

where

$$Q_j = \ell_j \cdot Bs_j,$$
$$(A - \lambda_j I)s_j + Br_j = 0.$$

4.2 Wave–mean field interactions

Averaging (2.1) with respect to x, shows that the mean of a bounded solution is constant. Therefore, waves cannot generate a mean field. However, suppose that there is a rapidly varying source term, so that

(4.11) $$U_t + f(U)_x + g(\frac{U}{\epsilon}) = 0.$$

(We consider one space dimension for simplicity – the analysis extends easily to several dimensions.) Then (4.11) has the following asymptotic solution,

(4.12) $$U = \epsilon a(\frac{x-\lambda t}{\epsilon}, x, t)r + \epsilon\bar{U}(x,t) + O(\epsilon^2), \quad \epsilon \to 0+ \text{ with } x,t = O(1).$$

In (4.12), λ is an eigenvalue of $A = \nabla_U f(0)$, and r and ℓ are eigenvectors, with $\ell\cdot r = 1$. We assume that a and its derivatives are periodic (or almost periodic), functions of θ with zero mean. It is also straightforward to include several waves, as in section 4.1. The mean field $\bar{U}$ satisfies the semi-linear equations,

(4.13a) $$\bar{U}_t + A\bar{U}_x + \langle g \rangle = 0,$$

where

$$\langle g \rangle(x,t) = \lim_{T\to+\infty} \frac{1}{T}\int_0^T g[\bar{U}(x,t) + a(\theta,x,t)r]d\theta.$$

The wave amplitude a satisfies

(4.13b) $\quad a_t + \lambda a_x + Maa_\theta + \ell \cdot [g(\bar{U} + a\mathbf{r}) - < g >] = 0,$

where M is given in (2.12). The wave drives the mean field, and the mean field modulates the wave.

A special case is when (4.11) is semi-linear, meaning that

$$f(U) = AU.$$

Then $V = \epsilon^{-1}U$ satisfies

$$V_t + AV + g(V) = 0.$$

The asymptotic solution is [32], [45]

$$V = a(\frac{x-\lambda t}{\epsilon}, x, t)\mathbf{r} + \bar{U}(x,t) + O(\epsilon), \quad \epsilon \to 0+ \quad \text{with} \quad x,t = O(1),$$

where a and $\bar{U}$ satisfy (4.13) with M = 0.

5 NONLINEAR GEOMETRICAL ACOUSTICS

Nonlinear acoustics is a well-developed subject with applications to sonic boom research, remote sensing in the ocean, and ultrasonic technology. Here, we summarize the basic asymptotic equations of nonlinear acoustics. They are obained as a special case of the general theory described in the previous sections. For other reviews, see Crighton [9], [10] and Hamilton [14].

5.1 Equations of motion of a compressible fluid

The equations of motion of a (one species) compressible fluid are

$$\rho_t + \operatorname{div}(\rho \mathbf{u}) = 0,$$

$$(\rho \mathbf{u})_t + \operatorname{div}(\rho \mathbf{u} \times \mathbf{u} - T) = \rho \mathbf{F}, \tag{5.1}$$

$$[\rho(\tfrac{1}{2}\mathbf{u}\cdot\mathbf{u}+e)]_t + \operatorname{div}[(\tfrac{1}{2}\mathbf{u}\cdot\mathbf{u}+e)\rho\mathbf{u} - T\mathbf{u} + \mathbf{q}] = 0.$$

In (5.1), ρ is the mass density of the fluid, $\mathbf{u} \in \mathbb{R}^n$ is the fluid velocity, T is the Cauchy stress tensor, e is the specific internal energy, $\mathbf{q}$ is the heat flux vector, and $\mathbf{F}$ is the body force per unit mass. We neglect any heat sources. The constitutive equations for T, $\mathbf{q}$ and e are

$$T = [-p + \mu' \operatorname{div}\mathbf{u}]I + 2\mu E, \quad E = \tfrac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^T),$$

$$\mathbf{q} = \kappa \nabla T.$$

Here, p is the pressure, μ is the shear viscosity, $\mu' = \mu_B - \frac{2}{3}\mu$ is the dilatational viscosity, with μ_B the bulk viscosity, κ is the thermal conductivity, and T is the temperature. The internal energy, temperature, and pressure are functions of ρ and S, and they satisfy the thermodynamic relation,

$$TdS = de + pd(\tfrac{1}{\rho}).$$

The conductivity and viscosities κ, μ, and μ' are also functions of ρ and S. We define the sound speed $c(\rho,S)$ by

$$c^2 = \left.\frac{\partial p}{\partial \rho}\right|_S.$$

For simplicity, we shall consider a polytropic gas, for which

$$e = c_v T,$$

$$p = R\rho T = K\exp(\tfrac{S}{c_v})\rho^\gamma,$$

$$c^2 = \gamma\frac{p}{\rho},$$

where c_v is the specific heat at constant volume, and

$$\gamma = \frac{c_p}{c_v}$$

is the ratio of specific heats. One obtains the similar asymptotic equations for

general equations of state. Only the values of the coefficients are affected.

The non–conservative form of (5.1) is

$$
\begin{aligned}
&\frac{D\rho}{Dt} + \rho \mathrm{div}\mathbf{u} = 0, \\
&\rho\frac{D\mathbf{u}}{Dt} + \nabla p = \nabla(\mu'\mathrm{div}\mathbf{u}) + \mathrm{div}(2\mu E) + \rho\mathbf{F}, \\
&\rho T\frac{DS}{Dt} = 2\mu E{:}E + \mu'(\mathrm{div}\mathbf{u})^2 + \mathrm{div}(\kappa\nabla T),
\end{aligned}
\tag{5.2}
$$

where

$$\frac{D}{Dt} = \partial_t + \mathbf{u}\cdot\nabla,$$

is the material derivative.

Linearizing the nondissipative version of (5.2), with $\mathbf{F} = 0$, about a constant solution, $\rho = \rho_0$, $\mathbf{u} = \mathbf{u}_0$, $S = S_0$, and $c = c_0$, we obtain the acoustics equations,

$$
\begin{aligned}
&\rho_T + \rho_0 \mathrm{div}\mathbf{u} = 0, \\
&\mathbf{u}_T + c_0^{\,2}\rho_0^{\,-1}\nabla\rho + \rho_0^{\,-1}p_{S0}\nabla S = 0, \\
&S_T = 0,
\end{aligned}
\tag{5.3}
$$

where $\partial_T = \partial_t + \mathbf{u}_0\cdot\nabla$. The plane wave solutions of (5.3) are

$$\rho = \hat{\rho}\exp i(\mathbf{k}\cdot\mathbf{x}-\omega t),\quad S = \hat{S}\exp i(\mathbf{k}\cdot\mathbf{x}-\omega t),\quad \mathbf{u} = \hat{\mathbf{u}}\exp i(\mathbf{k}\cdot\mathbf{x}-\omega t),$$

where

$$(\omega - \mathbf{u}_0\cdot\mathbf{k})^n[(\omega - \mathbf{u}_0\cdot\mathbf{k})^2 - c_0^{\,2}k^2] = 0.$$

The root

$$(\omega - \mathbf{u}_0\cdot\mathbf{k})^2 = c_0^{\,2}k^2,$$

is the dispersion relation of sound waves. The associated null vector is

$$\begin{bmatrix} \hat{\rho} \\ \hat{\mathbf{u}} \\ \hat{S} \end{bmatrix} = \begin{bmatrix} \rho_0 \\ c_0\hat{\mathbf{k}} \\ 0 \end{bmatrix}, \quad \hat{\mathbf{k}} = k^{-1}\mathbf{k}.$$

Sound waves are compressive, longitudinal, and isentropic. The root

$$\omega = \mathbf{u}_0\cdot\mathbf{k},$$

is a multiple eigenvalue in more than one space dimension. Such waves are

convected by the background flow. The associated null space is spanned by vectors of the form

$$\begin{bmatrix} \hat{\rho} \\ \hat{\mathbf{u}} \\ \hat{S} \end{bmatrix} = \begin{bmatrix} -p_{S0} \\ 0 \\ c_0^{\,2} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 \\ \mathbf{k}^{\perp} \\ 0 \end{bmatrix}, \quad \text{where} \quad \mathbf{k}^{\perp}\cdot\mathbf{k} = 0.$$

These waves carry either entropy at constant pressure, or vorticity.

5.2 A single sound wave

We begin with two examples. The first is a sound wave propagating through a stratified fluid. The equations of motion in one space dimension – including viscosity, heat conduction, and a gravitational body force – are:

$$\begin{bmatrix} \rho \\ u \\ S \end{bmatrix}_t + \begin{bmatrix} u & \rho & 0 \\ c^2/\rho & u & p_S/\rho \\ 0 & 0 & u \end{bmatrix} \begin{bmatrix} \rho \\ u \\ S \end{bmatrix}_x = \begin{bmatrix} 0 \\ -g \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \partial_x(\bar{\mu}u_x)/\rho \\ \bar{\mu}u_x^{\,2} + \partial_x(\kappa T_x)/\rho T \end{bmatrix}. \tag{5.3}$$

Here, $\bar{\mu} = \frac{4}{3}\mu + \mu_B$. We suppose that the unperturbed state is one of hydrostatic equilibrium,

$$\rho = \rho_0(x),\ u = 0,\ S = S_0(x),$$
$$p = p_0(x) = \exp(S_0/C_v)\rho_0^{\,\gamma},$$
$$c^2 = c_0^{\,2}(x) = \gamma p_0/\rho_0,$$

where

$$p_0' = -g\rho_0, \qquad ' = \frac{d}{dx}.$$

The scale height of the stratification is a typical value of

$$\rho_0/\rho_0'.$$

A special case is isothermal equilibrium, when the fluid is exponentially stratified,

$$\rho_0 = \rho_* \exp(-x/H), \quad p_0 = p_* \exp(-x/H), \quad c_0 = (\gamma g H)^{1/2}.$$

Here, H is the scale height, ρ_* and p_* are the density and pressure at $x = 0$, and

the sound speed is constant. In the atmosphere, the sound speed is not constant – it varies from about 330 ms^{-1} at sea level to 300 ms^{-1} at 10 km.

To explain when a sound wave may be described using the weakly nonlinear theory, we introduce the acoustic Mach number

$$M_a = \frac{u_{max}}{c_0}.$$

Here, u_{max} is the maximum particle velocity in the sound wave. We denote a typical wavenumber by k i.e.

$$\frac{|\nabla u|}{u_{max}} = O(k).$$

The wave is small amplitude if $M_a << 1$. The cumulative effect of nonlinearity is important over propagation distances $\ell_N = O(M_a k)^{-1}$. A more precise value is

$$\ell_N = (M_a k L)^{-1},$$

where L is the parameter of nonlinearity of the fluid, defined below (5.4). Then ℓ_N is the shock formation distance for a plane wave with maximum slope k. Nondimensionalizing lengths by ℓ_N, the weakly nonlinear theory describes the propagation of waves of amplitude order $\epsilon << 1$, and wavenumber of order ϵ^{-1} over distances of the order one.

A typical width of the N–wave in a sonic boom is 100m ($k = \frac{1}{50}$ m^{-1}). The the strength of a strong boom is $M_a = 10^{-3}$. This gives $\ell_N \simeq 40$ km. For a 20 kHz ultrasonic wave in water, at standard conditions, with strength $M_a = 10^{-4}$ (which corresponds to maximum overpressures of two atmospheres), one finds that $\ell_N \simeq 35$ m.

The importance of shear viscosity is measured by the acoustic Reynolds number.

$$Re = \frac{\rho_0 c_0}{\mu k}.$$

Viscous effects are important over propagation distances of the order

$$\ell_D = \mathrm{Re}\ k^{-1}.$$

For the sonic boom in air with $k = \frac{1}{50}\ m^{-1}$, this gives $\ell_D \simeq 10^6$ km. Thus shear viscosity has negligable influence over most of the N–wave. However, it may have an important effect in spreading out the shock waves. Also, dissipation due to relaxation effects, which we shall not consider here, is usually more important than that due to shear viscosity. For a 20 kHz ultrasonic wave in water $\ell_D \simeq 10^3$ km is also much longer than the nonlinear lengthscale. However, at higher frequencies of about 100 MHz, one finds that $\ell_N \simeq \ell_D \simeq 10^{-2}$ m, so that nonlinear and viscous effects are about the same magnitude.

A balance between weakly nonlinear, viscous, and nonuniform effects occurs in the limit $M_a \to 0+$, with $\ell_N \sim \ell_D \sim H$. This gives

$M_a = O(\epsilon)$, small amplitude,

$kH = O(\epsilon^{-1})$ high frequency,

$\mathrm{Re} = O(\epsilon^{-1})$ large acoustic Reynolds number.

These effects are significant over propagation distances of the order $(\epsilon k)^{-1}$. Special cases (e.g. an inviscid fluid) may be obtained from this expansion by neglecting the appropriate terms.

We nondimensionalize (5.3) by the scale height H, and the density ρ_* and sound speed c_* at $x = 0$. With the above scaling assumptions, the nondimensional viscosity and thermal conductivity are of the order ϵ^2,

$$\bar{\mu} = \epsilon^2 \hat{\mu}, \qquad \kappa = \epsilon^2 \hat{\kappa}.$$

The weakly nonlinear expansion for a sound wave propagating in the positive x–direction is

$$\begin{bmatrix} \rho \\ u \\ S \end{bmatrix} = \begin{bmatrix} \rho_0(x) \\ 0 \\ S_0(x) \end{bmatrix} + \epsilon a[\epsilon^{-1}\phi(x,t),x,t] \begin{bmatrix} \rho_0(x) \\ c_0(x) \\ 0 \end{bmatrix} + O(\epsilon^2),$$

where ϕ is the retarded time,

$$\phi = t - \int_0^x c_0(x)^{-1}dx.$$

The transport equation for the wave amplitude is,

$$a_t + c_0 a_x - \frac{\gamma+1}{2} a a_\theta + (2\rho_0)^{-1}(c_0\rho_0' + 3c_0'\rho_0)a = \frac{1}{2}\delta c_0^{-2} a_{\theta\theta} \tag{5.4}$$

where

$$\delta = \frac{\mu}{\rho_0} + \frac{(\gamma-1)\kappa}{\rho_0 c_p}.$$

The quantity $\epsilon^2\delta$ is the "diffusivity of sound" [26],

$$\epsilon^2\delta = \frac{\mu}{\rho}\left[\frac{4}{3} + \frac{\mu_B}{\mu} + \frac{\gamma-1}{Pr}\right], \qquad Pr = \frac{\gamma c_v \mu}{\kappa} = \text{Prandtl number.}$$

The coefficient of the nonlinear term,

$$L = \frac{\gamma+1}{2} = 1 + \frac{\gamma-1}{2},$$

is called the parameter of nonlinearity of the fluid. The "1" in L is due to convection of the wave by its own velocity field. The remaining part is due to the variation of sound speed with density. L = 1.2 for air, and L $\simeq$ 3.5 for water. The nonlinearity of sound waves in air is mainly due to convection, but in water it is mainly due to the dependence of sound speed on density.

The inviscid form of (5.4) may be written as,

$$\frac{\partial}{\partial t}(\rho_0 c_0^2 a^2) + \frac{\partial}{\partial x}(\rho_0 c_0^3 a^2) - \frac{\partial}{\partial\theta}\left(\frac{\gamma+1}{3}\rho_0 c_0^2 a^3\right) = 0, \tag{5.5}$$

which states that the average linearized wave energy density, $\rho_0 c_0^2 a^2$, is conserved. For a uniform fluid, in which ρ_0 and c_0 are equal to one, (5.4) reduces to Burgers' equation,

$$a_t + a_x - \frac{\gamma+1}{2} a a_\theta = \frac{1}{2}\delta a_{\theta\theta} \tag{5.6}$$

For an exponentially stratified fluid, (5.4) gives

$$a_t + a_x - \frac{\gamma+1}{2} a a_\theta - \frac{1}{2}a = \frac{1}{2}\delta_0 e^x a_{\theta\theta} \tag{5.7}$$

Here, δ_0 is the diffusivity of sound at x = 0. (We assume that the viscosities and

thermal conductivity depend only on temperature.) If nonlinearity and dissipation are neglected, the wave amplitude grows along a ray like $\exp(\frac{x}{2})$. This follows from conservation of energy (5.5), since the density decays like e^{-x}, and the sound speed is constant.

For our second example, we consider spherical sound waves. Geometrical acoustics applies at distances r which are much greater than a wavelength. The NGA solution for an outgoing spherical wave propagating through a uniform fluid is

$$\begin{bmatrix} \rho \\ u \\ S \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \epsilon a[\tfrac{r-t}{\epsilon}, r, t] \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + O(\epsilon^2),$$

where u is the radial velocity component. The transport equation is

$$a_t + a_r + \tfrac{\gamma+1}{2} a a_\theta + \tfrac{n-1}{2r} a = \tfrac{1}{2} \delta a_{\theta\theta}, \tag{5.9}$$

where n is the number of space dimensions (n = 2 for cylindrical waves and n = 3 for spherical waves). If $\delta = 0$, this has the conservative form

$$\frac{\partial}{\partial t}(r^{n-1} a^2) + \frac{\partial}{\partial r}(r^{n-1} a^2) + \frac{\partial}{\partial \theta}(\tfrac{\gamma+1}{3} r^{n-1} a^3) = 0.$$

The general solution of the linearized equation is therefore

$$a = a_0(r-t) r^{-(n-1)/2}.$$

The cross–sectional area of a cone of rays at a distance r, is proportional to r^{n-1}. Thus this formula states that (amplitude^2 × ray tube area) is constant along a ray, which also follows from conservation of wave energy.

For a sound wave propagating through a nonuniform, moving fluid the NGA solution of (5.1) is

$$\begin{bmatrix} \rho \\ \mathbf{u} \\ S \end{bmatrix} = \begin{bmatrix} \rho_0 \\ u_0 \\ S_0 \end{bmatrix} + \epsilon a[\epsilon^{-1} \phi(\mathbf{x},t), \mathbf{x}, t] \begin{bmatrix} \rho_0 \Omega \\ c_0^2 \mathbf{k} \\ 0 \end{bmatrix} + O(\epsilon^2). \tag{5.10}$$

In (5.10), $\rho_0(\mathbf{x},t)$, $\mathbf{u}_0(\mathbf{x}.t)$, $S_0(\mathbf{x},t)$ are smooth solutions of the nondissipative form of (5.1), and $c_0(\mathbf{x}.t)$ is the corresponding sound speed. The local frequency and

wavenumber are

$$\omega = -\phi_t, \quad \mathbf{k} = \nabla\phi,$$

and Ω is the frequency in a reference frame moving with the fluid,

$$\Omega = \omega - \mathbf{u}_0\cdot\mathbf{k} = -\phi_T,$$

where

$$\partial_T = \partial_t + \mathbf{u}_0\cdot\nabla.$$

The eikonal equation for ϕ is

$$\phi_T^2 = c_0^2|\nabla\phi|^2.$$

The transport equation for a is

$$a_T + \frac{\Omega}{k^2}\mathbf{k}\cdot\nabla a + \frac{\gamma+1}{2}\Omega^2 a a_\theta + \{\frac{\Omega_T + c_0^2 \mathrm{div}\mathbf{k}}{2\Omega} + \frac{\partial_T(\rho_0 c_0^2)}{2\rho_0 c_0^2} + \frac{1}{2}\mathrm{div}\mathbf{u}_0 + \frac{\mathbf{k}\cdot\nabla(\rho_0 c_0^4)}{2\rho_0 c_0^2\Omega}\}a = \frac{1}{2}\delta k^2 a_{\theta\theta} \tag{5.11}$$

If $\delta = 0$, (5.11) can be put in the conservative form,

$$A_t + \mathrm{div}[(\mathbf{u}_0 + \frac{\Omega}{k^2}\mathbf{k})A] = 0,$$

where

$$A = \rho_0 c_0^2 \Omega a^2,$$

is the wave action.

Chin et al [4] have used this expansion for (5.1) with heat sources, to study the nonlinear development of acoustic instabilities.

5.3 Diffraction

5.3.1 Transverse diffraction

Transverse version of the GBE's in section 5.1 are most easily obtained by the following heuristic argument. We consider plane waves in two space dimensions for

simplicity. In a stationary fluid, with sound speed one, the linearized dispersion relation is

$$\omega^2 = k^2 + \ell^2 + m^2. \tag{5.12}$$

In (5.2), k, ℓ, and m are the x, y and z components of the wavenumber vector. For waves propagating in the positive x–direction, with slow variation in the y and z–directions,

$$\frac{\ell}{k}, \frac{m}{k} << 1.$$

Expanding (5.12) gives

$$\omega \sim k + (2k)^{-1}(\ell^2+m^2). \tag{5.13}$$

The transverse version of (5.2) and (5.6), whose linearized dispersion relation agrees with (5.13) is

$$\begin{bmatrix} \rho \\ u \\ S \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \epsilon a(\frac{t-x}{\epsilon}, \frac{y}{\epsilon^{1/2}}, \frac{z}{\epsilon^{1/2}}, x, y, z, t) \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + O(\epsilon^2),$$

where $a(\theta, \eta, \zeta, x, y, z, t)$ satisfies

$$\partial_\theta(a_t + a_x - \tfrac{\gamma+1}{2} a a_\theta - \tfrac{1}{2}\delta a_{\theta\theta}) - \tfrac{1}{2}(a_{\eta\eta} + a_{\zeta\zeta}) = 0.$$

For an axially symmetric beam, this gives

$$\partial_\theta(a_t + a_x - \tfrac{\gamma+1}{2} a a_\theta - \tfrac{1}{2}\delta a_{\theta\psi}) - \tfrac{1}{2}(a_{\rho\rho} + \rho^{-1} a_\rho) = 0, \tag{5.14}$$

where

$$\rho = (\eta^2 + \zeta^2)^{1/2} = \epsilon^{-1/2} r(y,z), \quad r = (y^2 + z^2)^{1/2}.$$

This equation is not obtained directly by using $\psi = r$, which gives (5.14) without the term proportional to $\rho^{-1} a_\rho$. This is because r is not smooth at r = 0. The two expansions agree when $r >> \epsilon^{1/2}$.

For nonplanar waves (in a uniform, stationary fluid for simplicity), $\phi(\mathbf{x},t)$ solves the eikonal equation,

$$\omega^2 = k^2, \quad \omega = -\phi_t, \quad \mathbf{k} = \nabla\phi,$$

and $\psi(\mathbf{x})$ must satisfy

$$\nabla\phi\cdot\nabla\psi = 0.$$

The transport equation is

$$\frac{\partial}{\partial\theta}[a_t + \omega^{-1}\mathbf{k}\cdot\nabla a + \frac{\gamma+1}{2}\omega^2 a a_\theta + \frac{\omega_t+\mathrm{div}\mathbf{k}}{2\omega} - \frac{1}{2}k^2\delta a_{\theta\theta}] + c_0{}^2\frac{|\nabla\psi|^2}{2\omega}a_{\eta\eta} = 0.$$

For the case of a sound wave propagating vertically upwards through an exponentially stratified fluid and diffracting horizontally, the transverse equation corresponding to (5.7) is

$$\partial_\theta(a_t + a_x - \frac{\gamma+1}{2}aa_\theta - \frac{1}{2}a - \frac{1}{2}\delta_0 e^x a_{\theta\theta}) - \frac{1}{2}a_{\eta\eta} = 0,$$

where $\eta = \epsilon^{-1/2}y$. The change of variables (3.6), with an additional rescaling to remove constant factors, puts this equation in the form

$$\frac{\partial}{\partial x}[u_t + uu_x - tu_{xx}] + t^{-1}u_{yy} = 0, \quad t > 0.$$

For a outgoing cylindrical wave diffracting in the angular direction, the transverse equation corresponding to (5.9), with $n = 2$, is

$$\partial_\theta(a_t + a_r + \frac{\gamma+1}{2}aa_\theta + \frac{1}{2r}a - \frac{1}{2}\delta a_{\theta\theta}) + \frac{1}{2}r^{-2}a_{\eta\eta} = 0,$$

where $\eta = \epsilon^{-1/2}\tan^{-1}(\frac{y}{x})$. The change of variables (3.6) leads to

$$\frac{\partial}{\partial x}[u_t + uu_x - tu_{xx}] + t^{-3}u_{yy} = 0.$$

As shown in section 3.1, this equation is equivalent to the Kuznetsov equation (3.8).

5.3.2 Caustics

Suppose that a nonplanar sound wave propagates through a uniform fluid at rest and forms a smooth, convex caustic at $\psi(\mathbf{x}) = 0$. The weakly nonlinear solution is

$$\begin{bmatrix} \rho \\ \mathbf{u} \\ S \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \epsilon^{2/3}a[\epsilon^{-1}\varphi(\mathbf{x},t),\epsilon^{-2/3}\psi(\mathbf{x}),\mathbf{x},t]\begin{bmatrix} 1 \\ \mathbf{k}(\mathbf{x}) \\ 0 \end{bmatrix} + O(\epsilon).$$

Here,

$$\phi(\mathbf{x},t) = \varphi(\mathbf{x}) - t,$$

and the sound speed and the density of the unperturbed fluid are one. The functions φ and ψ are determined from

$$|\nabla\varphi|^2 + \psi|\nabla\psi|^2 = 1, \quad |\nabla\psi|\cdot|\nabla\varphi| = 0,$$

and $a(\theta,\eta,\mathbf{x},t)$ satisfies

$$\frac{\partial}{\partial\theta}[(Ma-\eta)a_\theta] + a_{\eta\eta} = 0,$$

where

$$M = |\nabla\psi|^{-2}(\gamma+1).$$

For example, consider a circular caustic at $r = R$, where (r,σ) are polar coorninates in the plane. Then

$$\varphi = R\sigma, \quad \frac{2}{3}\psi^{3/2} = (r^2-R^2)^{1/2} - R\cos^{-1}(\frac{R}{r}).$$

On $r = R$, this implies that

$$M = 2^{4/3}R^{2/3}\ (\gamma+1).$$

5.4 Interacting waves

5.4.1 Reflection of a sound wave off an entropy wave

The weakly nonlinear expansion for interacting wave solutions of (5.3) has the following form

$$\begin{bmatrix} \rho \\ u \\ S \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \epsilon a_1[\frac{k_1(x+t)}{\epsilon},x,t]\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + \epsilon a_2[\frac{k_2 x}{\epsilon},x,t]\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} + \epsilon a_3[\frac{k_3(x-t)}{\epsilon},x,t]\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + O(\epsilon^2).$$

Here, the entropy variations are nondimensionalized by c_p. This solution is a sum of a sound wave moving to the left (with amplitude a_1), an entropy wave (with

amplitude a_2), and a sound wave moving to the right (with amplitude a_3).

The resonant interaction equations for the wave amplitudes are

$$
\begin{aligned}
&a_{1t}(\theta) - a_{1x}(\theta) - k_1\tfrac{\gamma+1}{2}a_1(\theta)a_{1\theta}(\theta) \\
&\qquad - \tfrac{1}{4}k_2 \lim_{T\to\infty} \tfrac{1}{T}\int_0^T a_2'\left[\frac{k_2\theta}{2k_1} + \frac{k_2\xi}{2k_3}\right]a_3(\xi)\, d\xi = \tfrac{1}{2}k_1{}^2\delta a_{1\theta\theta}(\theta), \\
&a_{2t} = k_2{}^2\hat{\kappa}\, a_{2\theta\theta}, \qquad (5.15) \\
&a_{3t}(\theta) + a_{3x}(\theta) + k_3\tfrac{\gamma+1}{2}a_3(\theta)a_{3\theta}(\theta) \\
&\qquad + \tfrac{1}{4}k_2 \lim_{T\to\infty} \tfrac{1}{T}\int_0^T a_2'\left[\frac{k_2\theta}{2k_3} + \frac{k_2\xi}{2k_1}\right]a_1(\xi)\, d\xi = \tfrac{1}{2}k_3{}^2\delta\, a_{3\theta\theta}.
\end{aligned}
$$

In (5.15), we have not shown the dependence of a_j on (x,t) explicitly, and $a_2'(\theta) = \partial_\theta a_2(\theta)$. These equations consist of a pair of Burgers equations for the sound waves coupled by a correlation with the entropy wave. The entropy wave is determined independently of the sound waves from the diffusion equation (5.15b). Thus, (5.15) describes the reflection of sound waves off an entropy perturbation.

We obtain a simpler version of (5.15) if we neglect thermo–viscous effects ($\delta = \hat{\kappa} = 0$), and assume that there are no spatial modulations (a_j independent of x). Then from (5.15b)

$$a_2 = a_2(\theta)$$

is an arbitrary function of θ, which we assume to be 2π–periodic in θ. We shall look for 2π–periodic solutions for $\{a_1,a_2\}$. The correlations in (5.15) are only 2π–periodic if

$$\frac{k_2}{2k_1} = n, \quad \frac{k_2}{2k_3} = m,$$

where n and m are integers. Then, nondimensionalizing lengths so that $k_2 = 2$, the wavenumbers are

$$k_1 = \frac{1}{n}, \quad k_2 = 2, \quad k_3 = \frac{1}{m},$$

and the corresponding frequencies are

$$\omega_1 = -k_1, \quad \omega_2 = 0, \quad \omega_3 = k_3.$$

They satisfy the resonance condition,

$$k_2 = nk_1 + mk_3,$$
$$\omega_2 = n\omega_1 + m\omega_3.$$

Defining new variables

$$u(\theta,t) = k_3\frac{\gamma+1}{2}a_3(\theta,t),$$
$$v(\theta,t) = -k_1\frac{\gamma+1}{2}a_1(\theta,t),$$
$$K(\theta) = \frac{1}{2}a_2'(\theta),$$

in (5.15), gives the following pair of integro–differential equations for u(x,t) and v(x,t),

$$u_t + uu_x - \frac{n}{m}\frac{1}{2\pi}\int_0^{2\pi} K(mx + ny)v(y,t)\,dy = 0,$$
(5.16)
$$v_t + vv_x + \frac{m}{n}\frac{1}{2\pi}\int_0^{2\pi} K(nx + my)u(y,t)\,dy = 0.$$

These equations (with $|m| = |n| = 1$) have been studied analytically and numerically by Majda, Rosales, and Schonbeck [31]. For a sinusoidal entropy distribution, and $m = -n = 1$, (5.16) becomes

$$u_t + uu_x + \frac{1}{2\pi}\int_0^{2\pi} \sin(x - y)v(y,t)\,dy = 0,$$
(5.17)
$$v_t + vv_x + \frac{1}{2\pi}\int_0^{2\pi} \sin(x - y)u(y,t)\,dy = 0.$$

Pego [38] found an exact travelling wave solution of (5.17), namely

$$u = c + \beta[1 + a\cos(x-ct)]^{1/2},$$
(5.18)
$$v = c + \beta[1 + \sigma a\cos(x-ct)]^{1/2},$$

where $\sigma \in \{-1,+1\}$, $0 \le a \le 1$, and

$$\beta(a) = \frac{\sigma}{\pi a}\int_0^{2\pi} \cos y(1 + a\cos y)^{1/2}dy,$$
$$c(a) = -\beta(a)\frac{1}{2\pi}\int_0^{2\pi} (1 + a\cos y)^{1/2}$$

An interesting feature of this solution is that waves exist only up to a maximum amplitude, correponding to a =1. The limiting wave has a cusp at its crest or trough.

5.4.2 Wave induced combustion

The combustion equations contain source terms which are rapidly varying when the activation energy is large. A combination of weakly nonlinear–high frequency and high activation energy asymptotics leads to equations of the type described in section 4.2 [30]. The mean field equations are

$$\bar{a}_{1t} - \bar{a}_{1x} = (2\gamma)^{-1} e^{\bar{T}} < \exp(\gamma-1)a >,$$

$$\bar{a}_{2t} = (\gamma^2-\gamma)^{-1} e^{\bar{T}} < \exp(\gamma-1)a >,$$

$$\bar{a}_{3t} + \bar{a}_{3x} = (2\gamma)^{-1} e^{\bar{T}} < \exp(\gamma-1)a >,$$

and the equation for the wave is

$$a_t - a_x - \tfrac{1}{2}[(\gamma+1)\bar{a}_1 + (\gamma-1)\bar{a}_2 + (\gamma-3)\bar{a}_3]a_\theta - \tfrac{\gamma+1}{2} a a_\theta$$

$$= (2\gamma)^{-1} e^{\bar{T}} [\exp(\gamma-1)a - < \exp(\gamma-1)a >].$$

Here, $\epsilon\bar{a}_1(x,t)$, $\epsilon\bar{a}_2(x,t)$, $\epsilon\bar{a}_3(x,t)$ are the mean field perturbations in the left–moving sound wave, the entropy wave, and the right–moving sound wave, and

$$\bar{T} = (\gamma-1)(\bar{a}_1 + \bar{a}_2 + \bar{a}_3).$$

The amplitude of the left–moving high frequency sound wave is $\epsilon a(\frac{x+t}{\epsilon},x,t)$, where $a(\theta,x,t)$ is a zero–mean almost periodic function of θ, and

$$< \exp(\gamma-1)a > (x,t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \exp[(\gamma-1)a(\theta,x,t)]\, d\theta.$$

The solution for the mean field blows up in finite time. The blow–up time is shorter when $a \neq 0$ than when $a = 0$. This describes one way that a sound wave could enhance the detonatation of a reacting gas.

REFERENCES

1. M. J. Ablowitz and H. Segur, Solitons and the Inverse Scattering Transform, SIAM, Philadelphia (1981).

2. A. Bamberger, B. Engquist, L. Halpern and P. Joly, "Parabolic wave equations and approximations in heterogeneous media", SIAM J. Appl. Math 48, 99–128 (1988).

3. P. Cehelesky and R. R. Rosales, "Resonantly interacting weakly nonlinear waves in the presence of shocks: a single space variable in a homogeneous time independent medium", Stud. Appl. Math. 74, 117–138 (1986).

4. R. C. Y. Chin, J. C. Garrison, C. D. Levermore, and J. Wong, "Weakly nonlinear acoustic instabilities in inviscid fluids", Wave Motion 8, 537–559 (1986).

5. Y. Choquet–Bruhat, "Ondes asymptotique at approchees pour systemes d'equations aux derivees partielles nonlineaires", J. Math. Pure Appl. 48, 117–158 (1969).

6. J. D. Cole and L. P. Cook, Transonic Aerodynamics, Elsevier, Amsterdam (1986).

7. A. D. D. Craik, Wave Interactions and Fluid Flows, Cambridge University Press, Cambridge (1985).

8. M. S. Cramer and A. R. Seebass, "Focusing of a weak shock at an arete", J. Fluid. Mech. 88, 209–222 (1978).

9. D. G. Crighton, "Model equations for nonlinear acoustics", Ann. Rev. Fluid. Mech. 11, 11–33 (1979).

10. D. G. Crighton, "Basic theoretical nonlinear acoustics", in Frontiers in Physical Acoustics, Proc. Int. School of Physics "Enrico Fermi", Course 93, North–Holland, Amsterdam (1986).

11. R. K. Dodd, J. C. Eilbeck, J. D. Gibbon and H. C. Morris, Solitons and Nonlinear Wave Equations, Academic Press, London (1982).

12. J. Gibbons and Y. Kodama,"Integrable quasi–linear systems: generalized hodograph transformation", preprint.

13. J. P Giraud, "Acoustique geometrique bruit ballistique des avions supersoniques et focalisation", J. Mechanique 4, 215–267 (1965).

14. M. F. Hamilton, "Fundamentals and applications of nonlinear acoustics", in Nonlinear Wave Propagation in Mechanics, ed. T. W. Wright, AMD–77, 1–28 (1986).

15. E. Harabetian, "Diffraction of a weak shock by a wedge", Comm. Pure Appl. Math. 40, 849–863 (1987).

16. W. D. Hayes, "Similarity rules for nonlinear acoustic propagation through a caustic", NASA Rept. Sp–255, 87–120 (1968).

17. J. K. Hunter and J. B. Keller, "Weakly nonlinear high frequency waves", Comm. Pure Appl. Math. 36, 547–569 (1983).

18. J. K. Hunter, A. Majda, and R. R. Rosales."Resonantly interacting weakly nonlinear hyperbolic waves, II: several space variables", Stud. Appl. Math. 75, 187–226 (1986).

19. J. K. Hunter and J. B. Keller, "Caustics of nonlinear waves", Wave Motion 9, 429–443, (1987).

20. J. K. Hunter, "Transverse diffraction and singular rays", SIAM J. Appl. Math 48, 1–37 (1988).

21. J. Kevorkian and J. D. Cole, Perturbation Methods in Applied Mathematics, Springer–Verlag, New York (1981).

22. A. Kluwick, "The analytical method of characteristics", Prog. Aerospace Sc. 19, 197–313 (1981).

23. P. Lax, "Hyperbolic systems of conservation laws and the mathematical theory of shock waves", Conf. Board Math. Sci., SIAM, Philadelphia (1973).

24. I. P. Lee–Bapty and D. G. Crighton, "Nonlinear wave motions governed by the modified Burgers equation", Phil. Trans. R. Soc. Lond. A333, 173–209 (1987).

25. M. J. Lighthill, "Reflection at a laminar boundary layer of a weak steady disturbance to a supersonic stream neglecting viscosity and heat conduction". Quart. J. Mech. Appl. Math. 3, 303–325 (1950).

26. M. J. Lighthill, "Viscosity effects in sound waves of finite amplitude". in Surveys in Mechanics, eds G. K. Batchelor and R. M. Davis, 255–351. Cambridge University Press (1956).

27. D. Ludwig, "Uniform asymptotic expansions at a caustic". Comm. Pure Appl. Math 19, 215–250, (1966).

28. A. Majda, "Nonlinear geometric optics for hyperbolic systems of conservation laws", in Oscillation Theory, Computation, and Methods of Compensated Compactness, IMA Volume 2, 115–165, Springer–Verlag, New York (1986).

29. A. Majda and R. R. Rosales, "Resonantly interacting weakly nonlinear hyperbolic waves, I: a single space variable", Stud. Appl. Math. 71, 149–179 (1984).

30. A. Majda and R. R. Rosales, "Nonlinear mean field–high frequency wave interactions in the induction zone", to appear in SIAM J. Appl. Math.

31. A. Majda, R. R. Rosales, and M. Schonbek, "A canonical system of integro–differential equations in nonlinear acoustics", to appear in Stud. Appl. Math.

32. D. W. McLaughlin, G. Papanicolau and L. Tartar, "Weak limits of semilinear hyperbolic systems with oscillating initial data", Lecture Notes in Phys. 230, 277–289, Springer–Verlag (1985).

33. A. Nayfeh, "A comparison of perturbation methods for nonlinear hyperbolic waves", in Singular Perturbations and Asymptotics, eds. R. Meyer and S. Parter, 223–276, Academic Press, New York (1980).

34. J. J. C. Nimmo and D. G. Crighton, "Backlund transforms for nonlinear parabolic equations: the general results", Proc. Roy. Soc. London, Ser. A. 384, 381–401, (1982).

35. J. J. C. Nimmo and D. G. Crighton, "Nonlinear and diffusive effects in nonlinear acoustic propagation over long ranges", Phil. Trans. R. Soc. Lond. A. 320, 1–35 (1986).

36. F. Obermeier, "On the propagation of weak and moderately strong curved shock waves", J. Fluid. Mech. 129, 123–136 (1983).

37. L. A. Ostrovsky, "Short wave asymptotics for weak shock waves and solitons in mechanics", Int. J. Non–Linear Mechanics 11, 401–416 (1976).

38. R. Pego, "Some explicit resonanting waves in weakly nonlinear gas dynamics", to appear in Stud. Appl. Math.

39. C. Rogers and W. F. Shadwick, Backlund Transforms and Their Applications, Academic Press, New York (1982).

40. R. R. Rosales, "Canonical equations of long wave weakly nonlinear asymptotics", to appear in Proc. Canadian Appl. Math. Soc. Conf.in Cont. Mech. and Appl. Math. (1988).

41. O. V. Rudenko and S. I. Soluyan, Theoretical Foundations of Nonlinear Acoustics, Consultant Bureau, Plenum, New York (1977).

42. P. L. Sachdev, Nonlinear Diffusive Waves, Cambridge University Press, Cambridge (1987).

43. A. R. Seebass, "Nonlinear acoustic behavior at a caustic", NASA Rept. Sp–120, 87–120 (1971).

44. T. Taniuti et al, "Reductive perturbation method for nonlinear wave propagation", Prog. Theor. Phy. Suppl. 55 (1974).

45. L. Tartar, "Oscillations and asymptotic behaviour for two semilinear hyperbolic systems", in Dynamics of Infinite Dimensional Dynamical Systems, eds. S. N. Chaw and J. K. Hale, 341–356, Springer–Verlag, Berlin (1987).

46. G. B. Whitham, Linear and Nonlinear Waves, Wiley, New York (1974).

47. G. I. Zahalak and M. K. Myers, "Conical flow near singular rays", J. Fluid Mech. 63, 537–561 (1974).

Phase-Change Problem for Hyperbolic Heat Transfer Model

Dening Li
Department of Mathematics
University of Colorado at Boulder
Boulder, CO 80309 USA

Abstract

Phase-change problem is discussed for a hyperbolic heat transfer model under the traditional assumption that the temperatures on two sides of the interface are equal and given. The sufficient and necessary conditions are given for the local solution to exist and be unique. Global existence is discussed for some special case.

1 Introduction

As is well-known, the classical mathematical model for the heat transfer and diffusion phenomena is of parabolic type, in particular, the heat equation. These models are based upon the Fourier's law of heat conduction:

$$\vec{q} = -k\nabla u, \tag{1.1}$$

where $\vec{q}$ is the heat flux vector, k the thermal conductivity, u the temperature. In most cases, this kind of model works pretty well and gives satisfactory results. But one inherent shortcoming of the parabolic model is that it implies a physically unacceptable infinite propagation speed. This might be very important in certain models with large variations in temperature or large gradients of temperature.

In order to avoid this difficulty, it has long been proposed that instead of the Fourier law, one should assume that the heat flux responds to temperature gradient after a delay period of $\tau > 0$: i.e.

$$\vec{q}(t+\tau) = -k\nabla T(t). \tag{1.2}$$

Taking first order approximation one has the following relaxation relation:

$$\tau\vec{q}'(t) + \vec{q}(t) = -k\nabla T(t). \tag{1.3}$$

Replacing the Fourier law with this relation and combining with the law of conservation of energy:

$$c\rho u_t + \nabla \cdot \vec{q} = 0, \tag{1.4}$$

with ρ being the density, c the specific heat, one has the hyperbolic telegrapher's equation

$$\tau u_{tt} + u_t - \alpha^2 \triangle u = 0. \tag{1.5}$$

Here, $\alpha^2 = k/\rho c$ is the diffusivity.

In particular, for $\tau = 0$, we get the classical heat equation. There are already a lot of works about the relation between the classical heat equation and the telegrapher's equation. For the Cauchy problem or initial-boundary value problem, the solution of telegrapher's equation tends uniformly to the solution of classical heat equation as $\tau \to 0$.

On the other hand , for the classical heat equation, an interesting and important problem both in theoretical research and application is the famous Stefan problem. The Stefan problem consists of finding not only the temperature distribution $u(x,t)$, but also the surface along which a phase change occurs. It is only natural that one should study the problems of Stefan type for the hyperbolic heat transfer model.

In [5], Solomon and others gave a formulation of hyperbolic Stefan problem based upon the traditional assumption that the temperatures on two sides of phase change boundary are given and equal. Also in their paper, an explicit solution was given where the phase change front propagates faster than sound speed and consequently is physically unacceptable.

Partly in order to avoid this difficulty, several authors [1][2][4][6] suggested other formulation of phase change condition based upon the Rankine-Hugoniot conditions for the hyperbolic conservation laws. In particular, in this type of formulations, the temperature across the phase-change surface

is discontinuous. And in all these papers, the original question of Solomon and others in [5] remained unanswered.

In this paper, we want to study the Stefan problem for the hyperbolic heat model in the classical framework where the temperature is assumed to be given and continuous across the phase-change surface. The sufficient and necessary conditions are given for the local existence and uniqueness. In particular, for the example in [5], a natural mathematical explanation is given. Also, the global solution is discussed for some examples. But the conditions to guarantee the global existence are only sufficient ones. It remains open as to what extent these conditions could be relaxed. And also, we treat here only the case of one space dimension. For the high dimensional case, the only result now available is in [4], where a weak solution was given.

2 Local Solution

2.1 One Phase Problem

The one-phase Stefan problem for the hyperbolic heat transfer model consists in solving the following free boundary problem;

$$\begin{cases} \tau q_t + q + k u_x = 0, \\ c\rho u_t + q_x = 0. \end{cases} \quad x_0 < x < \phi(t),\ t > 0. \tag{2.1}$$

$$\begin{cases} u(x,t) = 0, \\ \rho H \phi'(t) = q(x,t). \end{cases} \quad on \quad x = \phi(t),\ t > 0. \tag{2.2}$$

where H is the latent heat, and $\phi(0) = 0$.

The boundary condition imposed on the fixed boundary $x = x_0 < 0$ may be one of the following:

1. Imposed temperature:

$$u(x_0, t) = u_{\#}(t),\ t > 0; \tag{2.3}$$

2. Imposed flux

$$q(x_0, t) = q_{\#}(t),\ t > 0; \tag{2.4}$$

3. Convective boundary condition

$$q(x_0,t) = h[u_\#(t) - u(x_0,t)],\ t > 0. \tag{2.5}$$

If $x_0 = 0$, then no initial condition is needed. If $x_0 < 0$, initial conditions should be given:

$$u(x,0) = u_0(x),\ q(x,0) = q_0(x),\ x_0 \le x \le 0. \tag{2.6}$$

Now for the case of $x_0 < 0$, we have the following result:

Theorem 2.1 *If $u_0, q_0 \in C^1(-\infty, 0]$, and satisfy the corresponding compatible condition at $x = 0, t = 0$. Then the problem (2.1)(2.2) coupled with any one of the boundary conditions in (2.3)-(2.5) has a unique local solution $(u, q, \phi) \in C^1 \times C^1 \times C^2$ if*

$$|q_0(0)| < \rho H \alpha(\tau)^{-\frac{1}{2}}. \tag{2.7}$$

The idea of the proof is to introduce the new variables

$$\bar{x} = x - \phi(t), \bar{t} = t \tag{2.8}$$

to transform the original free boundary problem into a fixed boundary problem, and then solve the transformed problem by integration along characteristics and linear iteration.

The condition (2.7) is also necessary in the sense that if it is not true, one would have either non-uniqueness or non-existence of the solution.

In particular, Theorem 2.1 explains the physically unacceptable explicit solution example given in [5] where (2.7) is not satisfied and consequently the solution is not unique in that case.

If $x_0 = 0$, then the situation is a little different from the previous case:

Theorem 2.2 *If $u_0, q_0 \in C^1(-\infty, 0]$, and satisfy the corresponding compatible condition at $x = 0, t = 0$. Then the problem (2.1)(2.2) coupled with any one of the boundary conditions in (2.4)(2.5) has a unique local solution $(u, q, \phi) \in C^1 \times C^1 \times C^2$ if*

$$0 < q_0(0) < \rho H \alpha(\tau)^{-\frac{1}{2}}. \tag{2.9}$$

The problem (2.1)(2.2)(2.3) is not well-posed.

The proof of this theorem is again achieved by integration along characteristics and linear iteration. In doing so, we make use of the theorem of Zhao about the well-posedness of boundary value problems for hyperbolic system in corner domain in [7] which extends the result of [3].

The non-well-posedness of the problem (2.1)(2.2)(2.3) follows from the fact that $q(0,0)$ is not uniquely determined. Consequently the solution is not unique.

2.2 Two phase problem

Similar to the one phase case, the two phase Stefan problem can be formulated as follows;

$$\begin{cases} \tau_1 \partial_t q_1 + k_1 \partial_x u_1 = 0, \\ c_1 \rho \partial_t u_1 + \partial_x q_1 = 0. \end{cases} \quad in \quad x < \phi(t). \tag{2.10}$$

$$\begin{cases} \tau_2 \partial_t q_2 + k_2 \partial_x u_2 = 0, \\ c_2 \rho \partial_t u_2 + \partial_x q_2 = 0. \end{cases} \quad in \quad x > \phi(t). \tag{2.11}$$

$$\begin{cases} u_1(x,t) = u_2(x,t) = 0, \\ \rho H \phi(t) = (q_1 - q_2)(x,t). \end{cases} \quad on \quad x = \phi(t). \tag{2.12}$$

$$\begin{cases} u_1(x,0) = u_{10}(x), \quad q_1(x,0) = q_{10}(x), \quad x < 0, \\ u_2(x,0) = u_{20}(x), \quad q_2(x,0) = q_{20}(x), \quad x > 0, \\ \phi(0) = 0. \end{cases} \tag{2.13}$$

Theorem 2.3 *If $u_{10}, q_{10}, u_{20}, q_{20} \in C^1$, and satisfy the corresponding compatible condition at $x = 0, t = 0$. Then the problem (2.11)-(2.13) has a unique local solution $(u_1, q_1, u_2, q_2, \phi)$, if*

$$|(q_{20} - q10)(0)| < min\{\rho H(\alpha_1^2/\tau_1)^{\frac{1}{2}}, \rho H(\alpha_2^2/\tau_2)^{\frac{1}{2}}. \tag{2.14}$$

The proof of this theorem is similar to the proof of theorem 2.1.

3 Global solution

For the global solution of the hyperbolic Stefan problems discussed in section 2, we'll consider only some special case.

For the one phase problem with the imposed temperature condition on the fixed boundary (2.1)-(2.3),(2.6), we have

Theorem 3.1 *If $u_0, q_0, u_\#$ are sufficiently smooth and corresponding compatible conditions are satisfied at $(0,0),(x_0,0)$. If in addition,*

$$u'_\#(t) \geq 0,$$

$$u'_0 \pm k\alpha\tau^{\frac{1}{2}} q'_0 < 0,$$

$$0 < \frac{c}{H}(u_0(0) + k\alpha\tau^{\frac{1}{2}} q_0(0)) < 1.$$

Then the problem (2.1)-(2.3),(2.6) has a unique global smooth solution (u, q, ϕ).

For the proof of this global result, we follow the approach of J. Greenberg in [2]. As usual, by linear transformations of the independent variables and the unknown functions, we can reduced the problem into the diagonal form. The global existence is proven if we can show that the free boundary will remain uniformly noncharacteristic for all times. This in turn can be achieved by the monotoneity argument similar to the one employed in [2].

Very similarly, the global existence of the two-phase Stefan problem can be stated and proven if the same kind of monotoneity of the initial data is assumed and the relaxation time τ and the diffusivity α in two phases are assumed to be equal.

References

1. A. Friedman & B. Hu: The Stefan problem for a hyperbolic heat equation. (preprint)

2. J. Greenberg: A hyperbolic heat transfer problem with phase changes. IMA J. Appl. Math., 38(1987) 1-21.

3. Li Ta-tsien & Yu Wen-ci: *Boundary Value Problems for Quasilinear Hyperbolic Systems,* Duke Univ. Math. Series V, 1985.

4. R.E. Showalter & N.J. Walkington: A hyperbolic Stefan problem. Quarterly Appl. Math. (4)45(1987),769-782.

5. A. Solomon, V.Alexiades, D. Wilson & J. Drake: On the formulation of a hyperbolic Stefan problem . Quarterly Appl. Math. 43(1985), 295-304.

6. A. Solomon, V. Alexiades, D. Wilson & J. Greenberg: A hyperbolic Stefan problem with discontinuous temperature. ORNL-6216, March 1986.

7. Zhao Yanchun: Boundary value problem for first order quasilinear hyperbolic systems. Chin. Ann. Math. 7A(1986).

Modifications to the Calculation of Fire Spread in Large Compartments

K. C. Heaton

Defence Research Establishment Valcartier

Abstract

Recently, it has become possible to model the progress of fires in large structures, such as buildings, with considerable accuracy. However, many of the physical processes involved in large fires, such as the rate of spread of the fire and the balance between convective and radiative heating and cooling effects, are difficult to model accurately from first principles.

In this work, an improved approximation for the flux onto burning objects from flames is derived and used to obtain an expression describing the rate of fire spread. The new formulations developed in this work are incorporated into a compartment fire model. Some results from a numerical solution of the equations governing fires for a specific case are presented.

1 Introduction

Fueled by recent events such as the fire in the King's Cross underground station, there has been considerable interest in modelling the spread of fires in the interiors of large structures such as buildings and large compartments. In general, there are two types of fire models: stochastic and deterministic. An example of the first type is the model originally developed at Worchester Polytechnic Institute to model the spread of fire in buildings, and subsequently substantially modified at DREV to model the spread of fires on board ships (Fitzgerald 1984). This model is a stochastic one, and hence one must specify the probabilities of thermal and structural failures of the walls, the probability of self-extinction of a fire within a given compartment, and the probability of success of attempts to extinguish the fire.

The other approach is a deterministic one in which the initial conditions such as fuel load, ignition temperatures, ventilation parameters, and dimensions of the compartment are specified, with the progress of the fire being modelled by simulating the physics of the fire. The deterministic models are further subdivided into two subclasses: computational fluid dynamics models and zone models. The CFD models divide the volume of interest, such as the interior of a compartment, into several thousand cells, and calculate the conditions within each cell based on e.g.

the gas species and temperatures, incident and emitted radiation, and air movement within each cell. This method has the disadvantage of being very computationally intensive, requiring the use of a supercomputer.

The second type of deterministic fire model is the zone model. In this class of models, the volume of interest is divided into two or more zones, and the model calculates average conditions within each zone. The most usual choice of zones is one in which the layer of hot gases which tend to accumulate near the ceiling and the lower layer of relatively cool gases are represented. CFC V, the fifth Harvard Computer Fire Code (Mitler 1985, Mitler and Emmons 1981), is a two zone model originally written to be used in the prediction of the spread of fires within buildings. Even in these relatively simple forms, the physical processes involved in burning require large numbers of calculations with all the problems attendant upon the solution of non-linear systems of equations.

In this paper, the methods used in CFC V to calculate the radiant flux from flames, flame spread and temperatures for objects are described. The modification to the algorithms for the calculation of the radiant flux are described, along with the adaptation of Quintiere's (1981) model of flame spread to a form suitable for incorporation into CFC are described, and numerical examples are presented.

2 Calculation of Heat Flux, Temperature and Flame Spread

2.1 Review of Previous Work

In CFC V, a flame is modelled by a cone of grey gas at a temperature of 1260° K, with radius r_f, height h_f, and semi-apex angle ψ. The power per steradian per unit volume, d^3Q, from an emitting element dV is given by

$$d^3Q = g\frac{dV}{4\pi}, \tag{1}$$

where g is given by

$$g = 4\kappa\sigma T_f^4, \tag{2}$$

and κ is the absorptivity of the flame gas in m^{-1}, T_f the flame temperature in °K and $\sigma = 5.67 \times 10^{-8}$ W/m^2 °K^4 the Steffan-Boltzmann constant. The flux per unit volume per steradian at a point P a distance ρ from the emitting element dV is given by

$$d^3\vec{\Phi} = \frac{g}{4\pi}\frac{dV}{\rho^2}\hat{\rho}, \tag{3}$$

as shown in Fig. 1. The flux normal to the surface, $d^3\Phi_n$, is given by

$$d^3\Phi_n = \frac{g}{4\pi}\frac{dV}{\rho^2}\hat{\rho}\cdot\hat{n}. \tag{4}$$

As can be seen from Fig. 1, for an element of volume located on a disc at a height x above the surface containing the point P and at a radial coordinate z,

$$\hat{n} \cdot \hat{\rho} = \frac{x}{\rho}, \tag{5}$$

and

$$\rho^2 = s^2 + z^2 - 2sz\cos\phi', \tag{6}$$

where $s = (x^2 + L^2)^{\frac{1}{2}}$ is the distance from the centre of the disc at x to P, L is the distance from the centre of the flame base to P, and ϕ' is the angle between s and z. Again referring to Fig. 1, if ϕ is the usual azimuthal coordinate, and θ' is the angle between s and L,

$$\begin{aligned} \vec{z} &= z\cos\phi\,\hat{z} + z\sin\phi\,\hat{y}, \\ \vec{s} &= s\cos\theta'\,\hat{z} - s\sin\theta'\,\hat{y}. \end{aligned} \tag{7}$$

Hence,

$$\begin{aligned} \cos\phi' &= \frac{\vec{s}\cdot\vec{z}}{|s|\,|z|} \\ &= \cos\theta'\cos\phi \\ &= \frac{L}{s}\cos\phi. \end{aligned} \tag{8}$$

Substituting eqs. (5), (6) and (8) into eq. (4) and using $dV = z\,dx\,dz\,d\phi$, one obtains

$$d^3\Phi_n = \frac{g}{4\pi}\frac{xz}{(x^2 + z^2 + L^2 - 2Lz\cos\phi)^{\frac{3}{2}}}\,dx\,dz\,d\phi \tag{9}$$

(Mitler 1978).

The normal flux from the whole flame at P, Φ_n, is then given by

$$\Phi_n = \frac{g}{4\pi}\int_{x_a}^{x_b}\int_0^{z(x)}\int_0^{2\pi}\frac{xz}{(x^2 + z^2 + L^2 - 2Lz\cos\phi)^{\frac{3}{2}}}d\phi\,dz\,dx, \tag{10}$$

where $z(x)$, the width of the cone as a function of height x, is given by

$$z(x) = r_f + (x_a - x)\tan\psi. \tag{11}$$

Usually, $x_a = 0$ and $x_b = h_f$; however, if the cone of flame extends into the layer of smoke at a height x_l above the firebase, it is (perhaps unrealistically) assumed that that portion of the flame within the layer produces a negligible contribution to the flux at P. Under those circumstances, $x_b = h_f - x_l$. If P is not on the same level as the fire base, x_a will take a non-zero value which is a function of the line of sight to P.

Evidently, the evaluation of eq. (10) by analytic methods is not tractable. In CFC V, eq. (10) is integrated by the expedient of setting $\cos\phi$ to its average value of $\overline{\cos\phi} = 0$.

Substituting $\cos\phi = \overline{\cos\phi} = 0$ into eq. (10), and integrating, one obtains that

$$\Phi_n(L) = \frac{g}{2}\left((L^2+x_b^2)^{\frac{1}{2}} - (L^2+x_a^2)^{\frac{1}{2}} - \cos^2\psi(s_b - s_a) + b\sin\psi\cos^2\psi\ln\left[\frac{s_a + x_a\cos\psi - r_f\sin\psi}{s_b + x_b\sec\psi - b\sin\psi}\right]\right), \tag{12}$$

where

$$\begin{aligned} b &= r_f + x_a\tan\psi, \\ s_a &= (L^2 + x_a^2 + r_f^2)^{\frac{1}{2}}, \\ s_b &= (L^2 + x_b^2 + [r_f - (x_b - x_a)\tan\psi]^2)^{\frac{1}{2}} \end{aligned} \tag{13}$$

(Mitler 1978). The error introduced by setting $\cos\phi = 0$ in eq. (10) is compensated for by multiplying eq. (12) by

$$f(d) = \begin{cases} 0.5068d^p, & d < 1 \\ 1 + \dfrac{0.3787}{0.4349 - d^2}, & d \geq 1 \end{cases} \tag{14}$$

where $d = \dfrac{L}{r_f}$ and $p = 2.825$ (Mitler 1978). For an optically thick flame, the modified normal flux, Φ_n', is given by

$$\Phi_n' = f(d)\,\Phi_n\,\frac{1 - e^{-\tau}}{\tau}, \tag{15}$$

where the effective optical depth, τ, through the source is given by

$$\tau = \frac{4}{\pi}\kappa r_f(1 + 0.84\zeta^2), \tag{16}$$

where $\zeta \equiv \dfrac{r_f}{L} \leq 1$.

For the case in which P lies within another fire, Φ_n' is given by

$$\Phi_n' = f(d)\,\Phi_n\,\frac{1 - e^{-\tau}}{\tau}e^{-\tau_1}, \tag{17}$$

where $\tau_1 = \kappa_1 r_{f_1}$, and κ_1 and r_{f_1} are the absorptivity and radius, respectively, of the flame at P. For the case of a flame radiating to its own base, an average value for Φ_n' , $\overline{\Phi}_b$, is used where

$$\overline{\Phi}_b = \sigma T^4\left(1 - e^{-0.7755\frac{\Phi_n(0)}{\sigma T^4}}\right). \tag{18}$$

In CFC V, the fire spread is modelled by a semi-empirical formulation which uses an expression for $\overline{\Phi}$,the average normal flux to the base of the fire, given by

$$\overline{\Phi} = \frac{1}{B}\left(1 - e^{-\frac{\dot{r}_f}{A}}\right), \tag{19}$$

where A is an experimentally determined fire spread parameter, $B = \frac{1}{\sigma T_f^4}$, and $\dot{r}_f = \frac{dr_f}{dt}$. By inverting eq. (19), one obtains

$$\begin{aligned} \dot{r}_f &= -A\ln(1-B\bar{\Phi}) \\ &\approx AB\bar{\Phi}\left(1+\frac{B\bar{\Phi}}{2}+\frac{B^2\bar{\Phi}^2}{3}\right), B\bar{\Phi} \ll 1. \end{aligned} \tag{20}$$

A value for $r_f(t)$ at each time step of the integration can then be obtained from

$$r_f(t_0+\Delta t) = r_f(t_0) + \int_{t_0}^{t_0+\Delta t} \dot{r}_f dt. \tag{21}$$

There are several problems associated with this method of determining r_f. The principal difficulty is that eq. (19) was derived under an implicit assumption that $\dot{r}_f \approx .01 r_f$. This assumption derives from a series of tests on polyurethane foam, and there is no particular guarantee that it is applicable for all circumstances or for other materials.

2.2 Evaluation of Flux and Fire Spread

It is possible partially to evaluate eq. (10) exactly by rearranging the order of the integrations,thusly:

$$\Phi_n = \frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b}\int_0^{z(x)} \frac{xz}{(x^2+z^2+L^2-2Lz\cos\phi)^{\frac{3}{2}}}\, dz\, dx\, d\phi. \tag{22}$$

Integrating eq. (22) with respect to z yields

$$\Phi_n = \sum_{i=1}^{4} \Phi_i \tag{23}$$

where

$$\Phi_1 = \frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b} \frac{xL\cos\phi(z(x)-L\cos\phi)}{(x^2+L^2\sin^2\phi)(x^2+z^2(x)+L^2-2Lz(x)\cos\phi)^{\frac{1}{2}}} dx\, d\phi, \tag{24}$$

$$\Phi_2 = -\frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b} \frac{x}{(x^2+z^2(x)+L^2-2Lz(x)\cos\phi)^{\frac{1}{2}}}\, dx\, d\phi, \tag{25}$$

$$\Phi_3 = \frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b} \frac{xL^2\cos^2\phi}{(x^2+L^2)^{\frac{1}{2}}(x^2+L^2\sin^2\phi)}\, dx\, d\phi, \tag{26}$$

$$\Phi_4 = \frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b} \frac{x}{(x^2+L^2)^{\frac{1}{2}}}\, dx\, d\phi. \tag{27}$$

Now, Φ_4 can be completely integrated, thusly:

$$\Phi_4 = \frac{g}{2}\left((x_b^2+L^2)^{\frac{1}{2}} - (x_a^2+L^2)^{\frac{1}{2}}\right). \tag{28}$$

The integrals for Φ_1, Φ_2 and Φ_3 can be integrated analytically only with respect to x. Φ_3 becomes

$$\begin{aligned}\Phi_3 \;=\; & -\frac{g}{4\pi}\frac{L}{2}\int_0^{2\pi}|\cos\phi| \\ & \ln\Big((x^2+L^2\sin^2\phi)^{-1}\Big[(\sqrt{2}\,|\sin\phi|)x((\cos^4\phi x^4+2L^2(-\sin^4\phi+2\sin^2\phi+1)x^2 \\ & +L^4(\sin^2\phi+1)^2)^{\frac{1}{2}}-\cos^2\phi\,x^2-L^2(\sin^2\phi+1))^{\frac{1}{2}} \\ & +(\sqrt{2}L((\cos^4\phi x^4+2L^2(-\sin^4\phi+2\sin^2\phi+1)x^2 \\ & +L^4(\sin^2\phi+1)^2)^{\frac{1}{2}}+\cos^2\phi\,x^2+L^2(\sin^2\phi+1))^{\frac{1}{2}} \\ & +(\cos^4\phi x^4+2L^2(-\sin^4\phi+2\sin^2\phi+1)x^2+L^4(\sin^2\phi+1)^2)^{\frac{1}{2}}+\sin^2\phi\,x^2+L^2\Big]\Big)\Big|_{x_a}^{x_b}d\phi.\end{aligned} \tag{29}$$

In order to proceed further with integration of Φ_1 and Φ_2 it is assumed that $z(x)$ is always an expression of the form

$$z(x)=a-b\,x, \tag{30}$$

Substituting eq. (30) into eq. (25) one obtains

$$\Phi_2=-\frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b}\frac{x}{((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}}\,dx\,d\phi. \tag{31}$$

Integrating eq. (31) with respect to x yields

$$\begin{aligned}\Phi_2 \;=\; & \frac{g}{4\pi}\int_0^{2\pi}-\frac{((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}}{(1+b^2)} \\ & +\frac{b(L\cos\phi-a)}{(1+b^2)^{\frac{3}{2}}}\ln(2\sqrt{(1+b^2)}((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}} \\ & +2(1+b^2)x+2b(L\cos\phi-a))\Big|_{x_a}^{x_b}d\phi.\end{aligned} \tag{32}$$

The substitution of eq. (30) into eq. (24) and the expression of the result in terms of partial fractions produces

$$\Phi_1=\Phi_{11}+\Phi_{12}, \tag{33}$$

where

$$\begin{aligned}\Phi_{11} \;=\; & -\frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b} \\ & \frac{L\cos\phi(L\cos\phi-a)x-bL^3\cos\phi\sin^2\phi}{(x^2+L^2\sin^2\phi)((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}}\,dx\,d\phi,\end{aligned} \tag{34}$$

$$\Phi_{12}=-\frac{g}{4\pi}\int_0^{2\pi}\int_{x_a}^{x_b}\frac{bL\cos\phi}{((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}}\,dx\,d\phi. \tag{35}$$

Equation (35) can be integrated directly, resulting in

$$\Phi_{12} = -\frac{g}{4\pi}\int_0^{2\pi}\frac{bL\cos\phi}{(1+b^2)^{\frac{1}{2}}} \tag{36}$$
$$\ln(2\sqrt{(1+b^2)}((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}$$
$$+2(1+b^2)x+2b(L\cos\phi-a))\Big|_{x_a}^{x_b}\,d\phi.$$

The integration of Φ_{11} requires some additional effort. Using the substitutions suggested by Gradshteyn and Ryzhik (1980, pgs. 80 -81), one finds that

$$\Phi_{11} = -\frac{g}{4\pi}\int_0^{2\pi}\frac{L\cos\phi}{2} \tag{37}$$
$$\ln\frac{\left|((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}-bx-L\cos\phi+a)\right|}{\left|((1+b^2)x^2+2b(L\cos\phi-a)x+L^2+a^2-2La\cos\phi)^{\frac{1}{2}}+bx+L\cos\phi-a)\right|}\Bigg|_{x_a}^{x_b}d\phi.$$

By substituting eqs. (28), (29), (32), (36), and (37) into eq. (23), one can integrate the resulting expression for Φ_n numerically with respect to ϕ and so obtain an exact expression for the flux at the point L.

There are four special cases which must be dealt with individually: that for which $L = 0$ and those for which $x = 0$ while $\phi = 0, \pi$, or 2π.

When $L = 0$, Φ_n is given by :

$$\Phi_n = \frac{g}{2}\left(x_b - \frac{ab}{(1+b^2)^{\frac{3}{2}}}\ln\left[\frac{\sqrt{(1+b^2)}((1+b^2)x_b^2-2bax_b+a^2)^{\frac{1}{2}}+(1+b^2)x_b-2ba))}{a(\sqrt{(1+b^2)}-b)}\right]\right.$$
$$\left.-\frac{((1+b^2)x_b^2-2bax_b+a^2)^{\frac{1}{2}}-a}{(1+b^2)}\right). \tag{38}$$

The integrands of eqs. (29), (32)-(37) are undefined when $x = 0$ at the same time as $\phi = 0$, π, 2π. When this arises during the numerical integration of these equations with respect to ϕ, the limit of $\sum_{i=0}^{3}\Phi_i$ as $x \to 0$ must be taken. Explicitly,

$$\lim_{x\to 0}\sum_{i=1}^{3}\Phi_i\Big|_{\substack{\phi=0\\ \phi=2\pi}} = \frac{g}{4\pi}\left\{-L\ln\left(\frac{L}{L-a}\right)+\frac{L-a}{(b^2+1)}\right. \tag{39}$$
$$\left.+\frac{1}{(b^2+1)^{\frac{3}{2}}}\left(b(b^2L+a)\ln(b+(b^2+1)^{\frac{1}{2}})\right)\right\},$$

and

$$\lim_{x\to 0}\sum_{i=1}^{3}\Phi_i\Big|_{\phi=\pi} = \frac{g}{4\pi}\left\{L\ln\left(\frac{L+a}{L}\right)-\frac{L+a}{(b^2+1)}\right. \tag{40}$$
$$\left.-\frac{1}{(b^2+1)^{\frac{3}{2}}}\left(b(b^2L-a)\ln((b^2+1)^{\frac{1}{2}}-b)\right)\right\}.$$

Finally, when the field point L is very much greater than r_f, eq. (22) can be expressed as

$$\Phi_n = \frac{g}{4\pi L^3}\int_{x_a}^{x_b}\int_0^{z(x)}\int_0^{2\pi} xz\left(1+\frac{3\cos(\phi)z}{L}+\frac{3}{2L^2}\left((5\cos^2\phi-1)z^2-x^2\right)+...\right)d\phi\,dz\,dx \tag{41}$$

The integration of eq. (41) yields

$$\begin{aligned}\Phi_n \;=\; & \frac{g}{4L^3}\left[x^2\left(\frac{a^2}{2}-\frac{2abx}{3}+\frac{b^2x^2}{4}\right)\right. \\ & +\frac{x^2}{L^2}\left(\frac{3b^4x^4}{16}-\frac{b^2x^4}{4}-\frac{9ab^3x^3}{10}+\frac{3abx^3}{5}\right. \\ & \left.\left.+\frac{27a^2b^4x^2}{16}-\frac{3a^2x^2}{8}-\frac{a^3bx}{2}+\frac{9ba^4}{16}\right)\right]_{x_a}^{x_b}\end{aligned} \tag{42}$$

Once Φ_n has been obtained, either numerically or by means of direct integration if possible, Φ_n' may be found by the application of eqs. (16)-(17) with $f(d) \equiv 1$, since there is no error arising from the use of an average value for $\cos\phi$ in the calculations described above. For the case of $L = 0$, $\tau = .2062994\,\kappa h_f$ which corresponds to the average optical depth at the height at which the volume of the cone has been divided into two equal parts.

In order to obtain the total normal flux, $\dot{q}''$, at a point P, one must sum not only the contributions to the flux from all flames visible at that point, but as well those from the walls, ceiling, and hot layer.

Once the flux $\dot{q}''$ is known, the surface temperature distribution across any object may be calculated using the one dimensional heat conduction equation,

$$\frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2}, \tag{43}$$

under the condition that

$$-k\left.\frac{\partial T}{\partial x}\right|_{x=x_a} = \dot{q}''(x_a,z,\Delta t) - h(T(x_a,z,\Delta t) - T(x_a,z,0)) \tag{44}$$

(Quintiere 1981). It should be noted at this point that the coordinate system of eqs. (43)-(44) is that of Fig. 1, and hence x is the vertical coordinate and z the horizontal.

The solution to eq.(43) is given by

$$\begin{aligned}T(x_a,z,\Delta t) - T(x_a,z,0) \;=\; & \frac{\sqrt{\alpha}}{k\sqrt{\pi}}\int_0^{\Delta t}\frac{\dot{q}''(x_a,z,s)}{\sqrt{\Delta t - s}}\,ds \\ & -\frac{\alpha h}{k^2}\int_0^{\Delta t}\dot{q}''(x_a,z,s)\exp\left(a(\Delta t-s)\right)\operatorname{erfc}\sqrt{a(\Delta t-s)}\,ds\end{aligned} \tag{45}$$

where ρ is now the density of the object, h the heat transfer coefficient, k the thermal conductivity, c the specific heat capacity, $\alpha = \frac{k}{\rho c}$ the thermal diffusivity, and

$$a = \alpha \left(\frac{h}{k}\right)^2 . \tag{46}$$

If $\Delta t = t - t_0$ is sufficiently small, one can write

$$\dot{q}''(x_a, z, t) = \dot{q}''(x_a, z, t_0) + \left.\frac{d\dot{q}''}{dt}\right|_{t=t_0} \Delta t \tag{47}$$

Setting $t_0 = 0$ in eq. (47) and substituting the result into eq. (45), one obtains an expression for the change, ΔT, in the temperature T across the surface of an object during a time interval Δt:

$$\begin{aligned} \Delta T \;=\; & \dot{q}''(x_a, z, 0) \left(\frac{\mathrm{erf}(\sqrt{a\Delta t})e^{a\Delta t}}{h} - \frac{e^{a\Delta t}}{h} + \frac{2\sqrt{\alpha \Delta t}}{\sqrt{\pi} k} - \frac{2\sqrt{a\Delta t}}{\sqrt{\pi} h} + \frac{1}{h} \right) \\ & + \left.\frac{d\dot{q}''}{dt}\right|_{t_0=0} \left(\frac{\mathrm{erf}(\sqrt{a\Delta t})e^{a\Delta t}}{ah} - \frac{e^{a\Delta t}}{ah} + \frac{4\sqrt{\alpha}\,(\Delta t)^{\frac{3}{2}}}{3\sqrt{\pi} k} \right. \\ & \left. - \frac{4\sqrt{a}\,(\Delta t)^{\frac{3}{2}}}{3\sqrt{\pi} h} + \frac{\Delta t}{h} - \frac{2\sqrt{\Delta t}}{\sqrt{\pi a} h} + \frac{1}{ah} \right) . \end{aligned} \tag{48}$$

Using eq. (48), one can evaluate the temperature across the surface of an object at as many points as desired for each value of Δt. If one knows the ignition temperature T_{ig} for an object, it is possible to calculate, using eq. (48), , at what time the temperature at each point exceeds T_{ig}, and so determine the present radius of the flame.

3 Numerical Methods

The basic equations governing the spread of fires used in CFC V have been documented in several places e.g. Mitler (1978), Mitler and Emmons (1981). They constitute a set of coupled linear and non-linear simultaneous algebraic equations, linear ordinary differential equations with respect to time, and one partial differential equation, that for the diffusion of heat into a solid.

Two methods of solution are used in CFC V: a successive substitution method and a Newton-Raphson method for use when successive substitution fails. The convergence criterion for the equations is that the scaled difference between successive iterations of the system of equations be less than a predetermined value ϵ. These methods of solution have not been altered in the modified version of CFC V described above. Presently, ϵ has been chosen to be 1×10^{-4}.

CFC V has been modified to incorporate the changes described in section 2.2. The original expression for the flux Φ_n, eq. (12) was replaced in the computer

programme by the sum of eqs. (28), (29),(32), (36), and (37) or by eq. (38) for $L = 0$ or eq. (42) for $L > 10r_f$.

The integrals with respect to ϕ were evaluated using the composite trapezoidal rule and Richardson extrapolation. That is, two estimates, I_{n_1} and I_{n_2} for $\sum_{i=1}^{3} \Phi_i$ were obtained using subintervals of $\frac{\pi}{n_1}$ and $\frac{\pi}{n_1}$, respectively, were obtained. The final estimate, I^*, for the integral was found by means of

$$I^* = I_{n_1} + \frac{I_{n_2} - I_{n_1}}{1 - \left[\frac{n_1}{n_2}\right]^2}. \tag{49}$$

In this case, $n_1 = 32$ and $n_2 = 64$.

In eq. (48), $\left.\frac{d\dot{q}''}{dt}\right|_{t_0=0}$ was approximated by $\frac{\dot{q}''(\Delta t) - \dot{q}''(0)}{\Delta t}$. The surface temperature of combustible objects was calculated by means of eq. (48). The rate of flame spread, $\dot{r}_f$, was estimated from the calculation of the temperatures at a series of points along the burning object. If z_i is the closest exterior point to the flame radius at which the temperature has been calculated, then,

$$\dot{r}_f \approx \frac{z_i - r_f(t_0)}{T(z_i, t_0 + \Delta t) - T_{ig}} \frac{\Delta T}{\Delta t}. \tag{50}$$

Equation (21) was then used to determine r_f.

4 Numerical Results and Discussion

For the purposes of example, it was decided to simulate a fire in a compartment with aluminium walls and ceilings, with dimensions of $9.14 \times 5.8 \times 2.4$ m. and which contained two combustible objects. There were two vents, one near the ceiling and the other near the floor, and a door with dimensions $1.8 \times .69$ m. Air was allowed to circulate freely through all three openings. Since most common items of furniture are composed largely of polyurethane foam, it was decided to treat the two objects as parallelepipeds made entirely of polyurethane. The dimensions of the two objects were chosen to be $4.5 \times .72 \times 1.73$ and $2.06 \times .58 \times 1.73$ m. They were located along adjacent walls with the separation between their centres being 2.73 m. A fire, whose initial radius was set to .037 m., was assumed to have been started at the centre of the larger object (hereafter called object 1) at $t = 0$. Since both objects possessed the same height, $x_a = 0$, $x_b = h_f$ and hence $a = r_f$ and $b = \tan\psi$ in eqs. (30) - (42). Following Mitler's (1978) work, it was decided to set $T_f = 1260°$ k and the initial value of ψ to 30°. The initial temperature of the ambient air was taken as $293°K$, as was that of the second (non-burning) object. Because the fire had just started, it was assumed that the surface temperature of the object 1 was ambient, except at those points located inside r_f, where the temperature was assumed to be $T_{ig} = 740°$ K.

Figure 2 shows the results of a calculation carried out using both the original version of CFC v and a modified version, incorporating the changes described in the previous section. As can be seen, for the first 16 seconds, both versions yield approximately the same results, after which time the modified version of CFC predicts a marked increase in the rate of flame spread. This rapid increase after 16 seconds seems to be largely an artefact of the calculation. In order to keep the computation times as short as possible, the changes in temperature, ΔT, were evaluated at only four points on each object. For ease of computation, CFC V simulates all objects by cylinders of the same height as the objects they represent, and whose radii R_0 are chosen so as to give each cylinder the same total surface area as the object it models. The four points, then, at which the temperatures were evaluated were $r_0 = 0$, $r_1 = .05$ m, $r_2 = \frac{R_0 - .05}{2}$ and $r_3 = R_0$. The sharp transition at 16 seconds occurs at the point at which the application of eq. (50) for the estimation of $\dot{r}_f$ is switched from r_1 to r_2, and hence it is reasonable to expect that much of the apparent increase in r_f is due to estimation errors. This could be improved by substantially increasing the number of points, at a considerable sacrifice in computation speed.

Figure 3 shows the averaged radiative flux from the flame on each object 1. As can be seen, where the two fire radii are nearly the same, the modified code predicts significantly less flux than the original. This difference seems to arise because the term represented by Φ_3 in our version was set to zero in the original version of CFC V. Since this term is always negative, it has the effect of significantly reducing the flux from the value predicted by the original version of CFC V.

In the original version of CFC V, the surface temperature of any burning object was always set to T_{ig} for all points on the object's surface. In our modified version of CFC V, the average temperature of an object was calculated by assuming that the centre of the fire maintained a constant temperature of T_{ig} and the temperature was a linear function from r_0 to r_3. Figure 4 shows the result of this calculation. Figure 5 shows the same calculation for object 2. Since this object never ignited, the linear approximation of the temperature variation across its surface should be rather more realistic than for object 1. The slightly uneven temperature rise in the modified version of CFC V arises because the contribution to the total flux from the walls of the compartment varies from moment to moment. Object 2 shows no temperature rise in the original version of CFC V because the flux on object 2 never rises above .01 W/m^2, below which level the flux is considered to be negligible in both versions of CFC V.

Finally, Figs. 6 and 7 show the surface temperature and flux, calculated with the modified version of CFC V, at r_0, r_1, r_2 and r_3. It was assumed that the temperature inside r_f would always be constant at T_{ig}. The calculations of the flux (except for $L = 0$ which is a special case) , was terminated when the fire reached that point, since eqs. (28), (29),(32), (36), (37) are not valid for $L \leq r_f$.

5 Concluding Remarks

In this work, we have discussed the modifications made to CFC V in order to improve the calculation of the flux to an object from a conical flame. It was then shown how the value of the flux could be used to determine the temperature variation across the surface of an object, and consequently the rate of flame spread across the object. Some numerical results were presented and compared with results from the unmodified version for identical initial conditions.

Considerable work remains to be done in order to improve the fidelity of this fire simulation. The number of points at which the temperature profile across a burning object is calculated should be increased, therby improving the accuracy of the computation of the flame spread, more accurate models for the flux from the flames to the walls and ceiling should be developed, as well as for the exchange of heat between the hot layer and objects in the room.

References

[1] Drysdale, D., *An Introduction to Fire Dynamics*, John Wiley and Sons, Toronto, 1985.

[2] Fitzgerald, R.H., *Building Fire Safety Evaluation*, Worcester Polytechnic Institute, Worcester, Massachusetts, 1983.

[3] Gahm, J.B., *Computer Fire Code VI*, Home Fire Project Technical Report No. 58, Division of Applied Sciences, Harvard University, 1983.

[4] Gradsthteyn, I.S. and Ryzhik, I.M. 1980, *Table of Integrals, Series, and Products*, Academic Press, Inc., New York

[5] Mitler, Henri E., *Comparison of Several Compartment Fire Models: An Interim Report*, U.S. Department of Commerce, NBS, NEL, Center for Fire Research, 1985. Technical Report No. 45, Division of Applied Sciences, Harvard University, 1981.

[6] Mitler, Henri E., *The Physical Basis for the Harvard Computer Fire Code*, Home Fire Project Technical Report No. 34, Division of Applied Sciences, Harvard University, 1978.

[7] Mitler, Henri E. and Emmons, Howard W., *Documentation for CFC V, the Fifth Harvard Computer Fire Code*, Home Fire Project

[8] Quintiere, J.G., *A simplified theory for generalising results from a radiant panel rate of flame spread apparatus*. Fire and Materials, 5, 1981

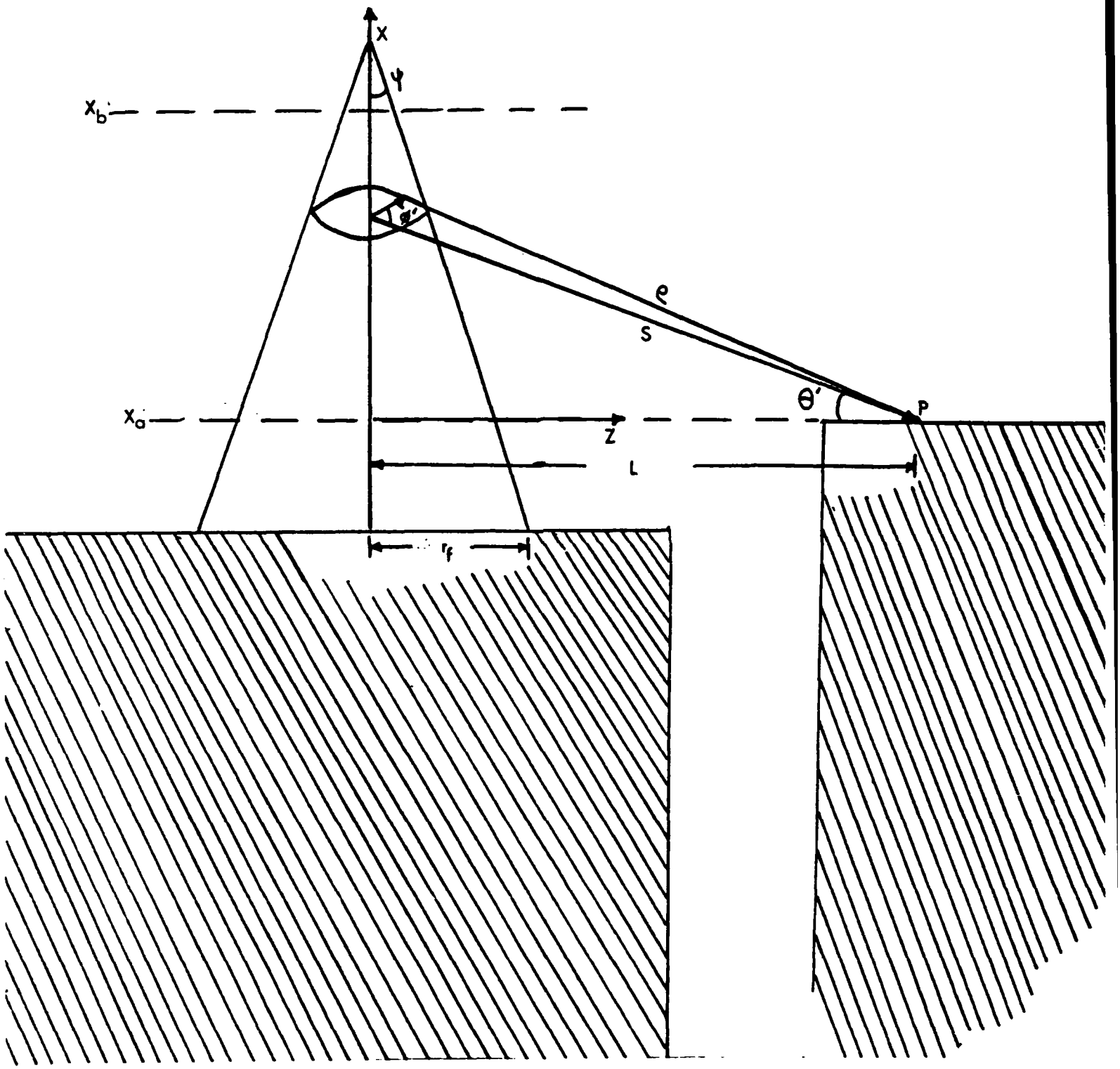
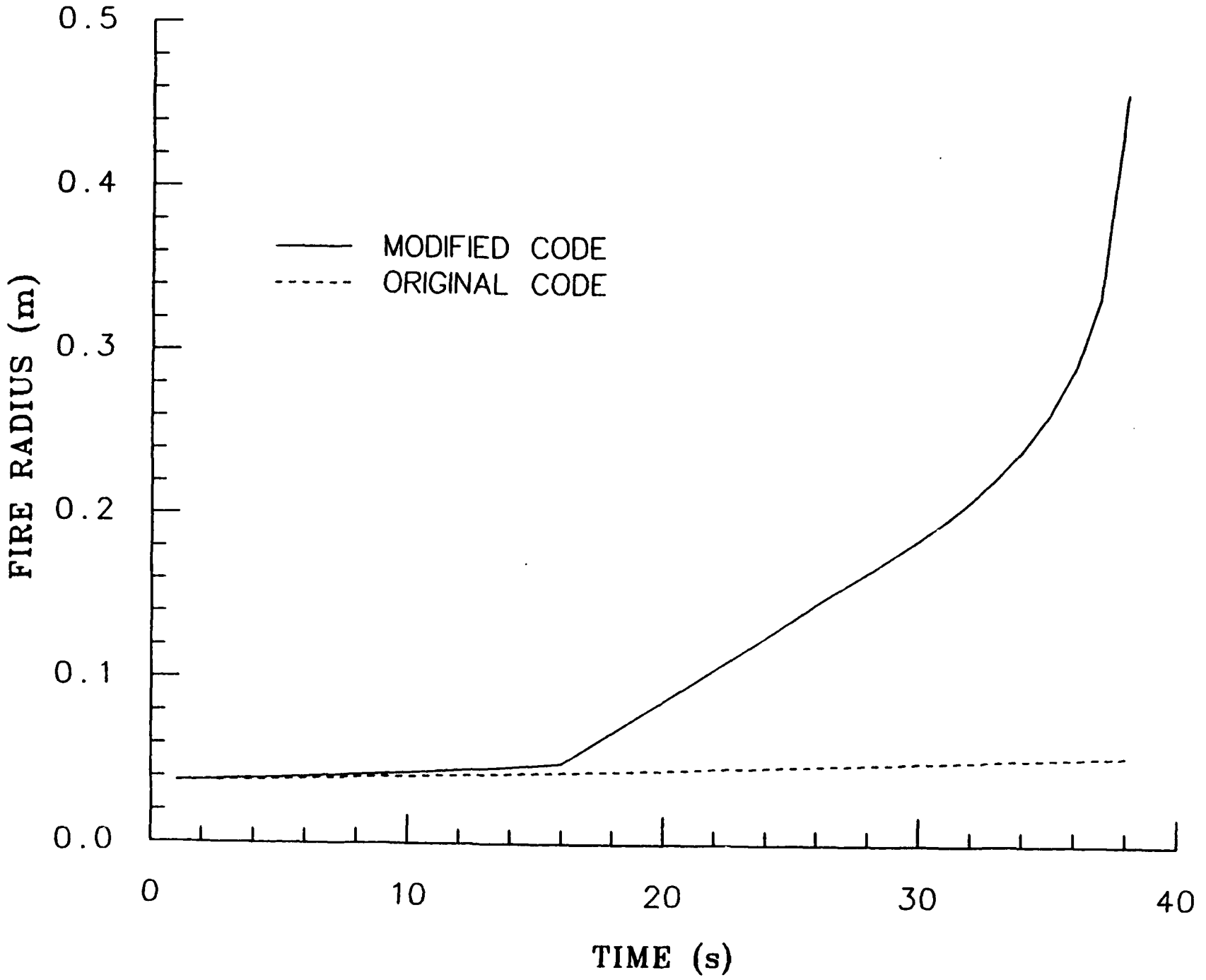


Figure 1: Fire Geometry

Fig. 2: Fire Radius on Object 1



590

Fig 3: Average Radiative Flux from Flame

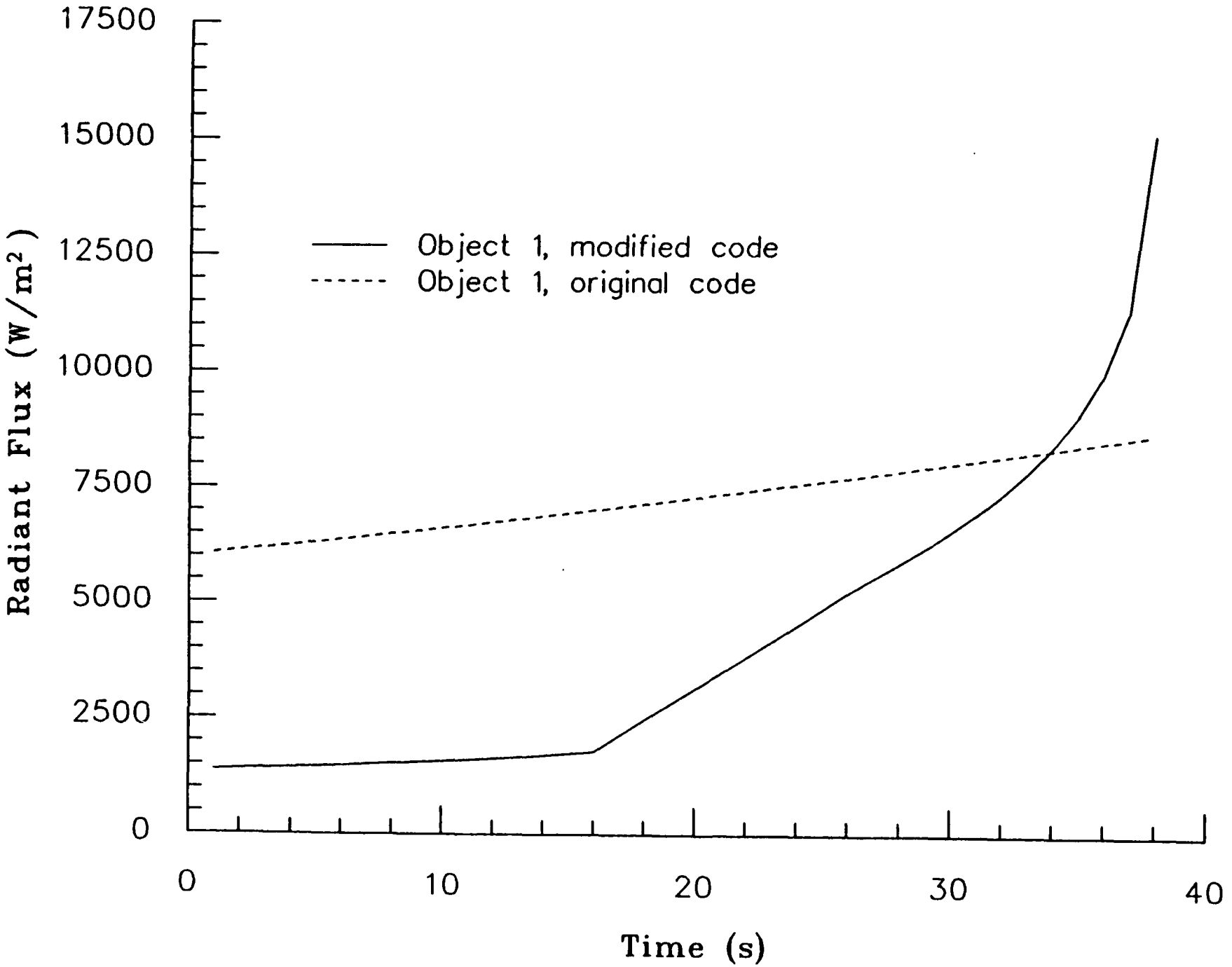


Fig. 4: Average Surface Temperature

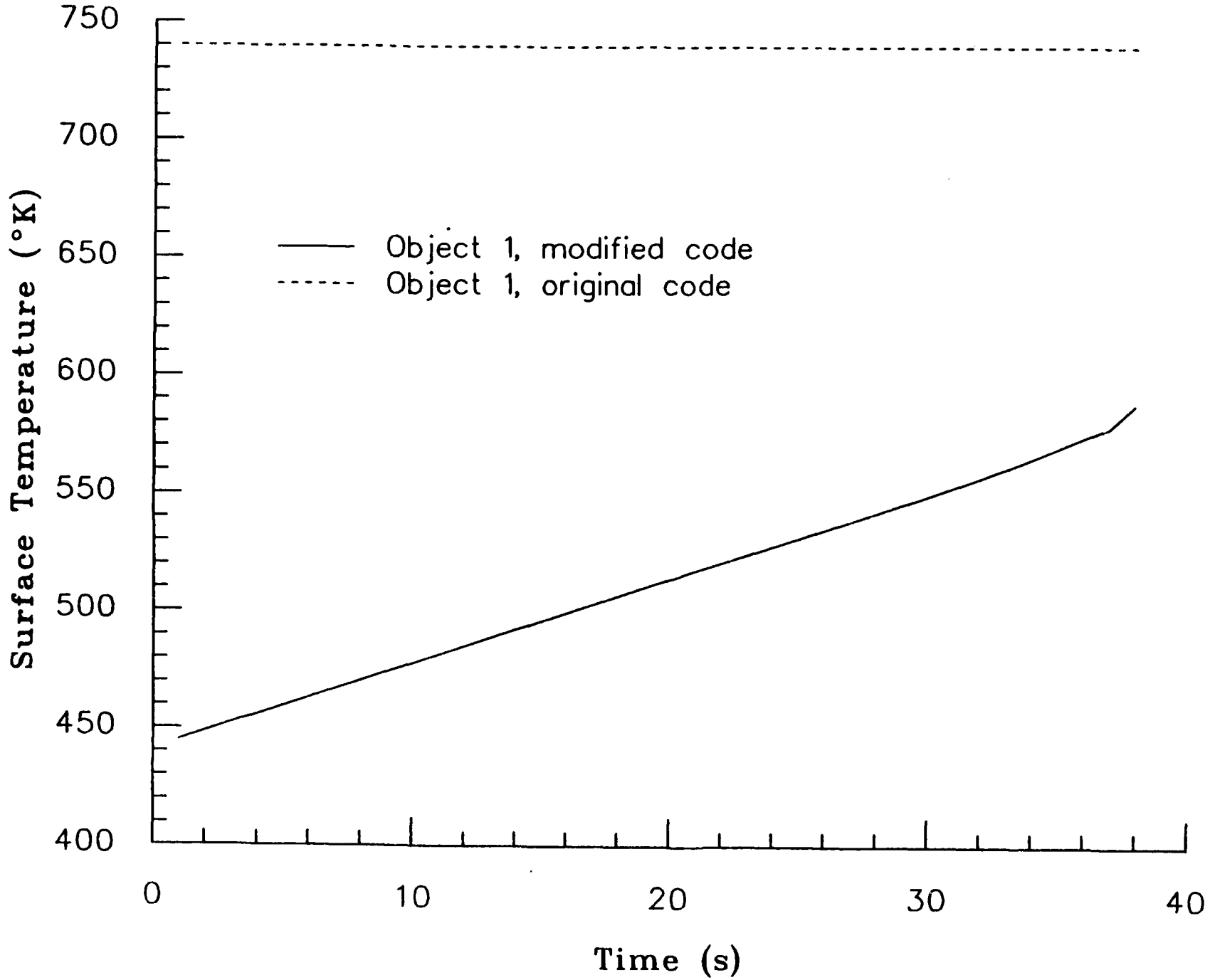


Fig 5: Average Surface Temperature

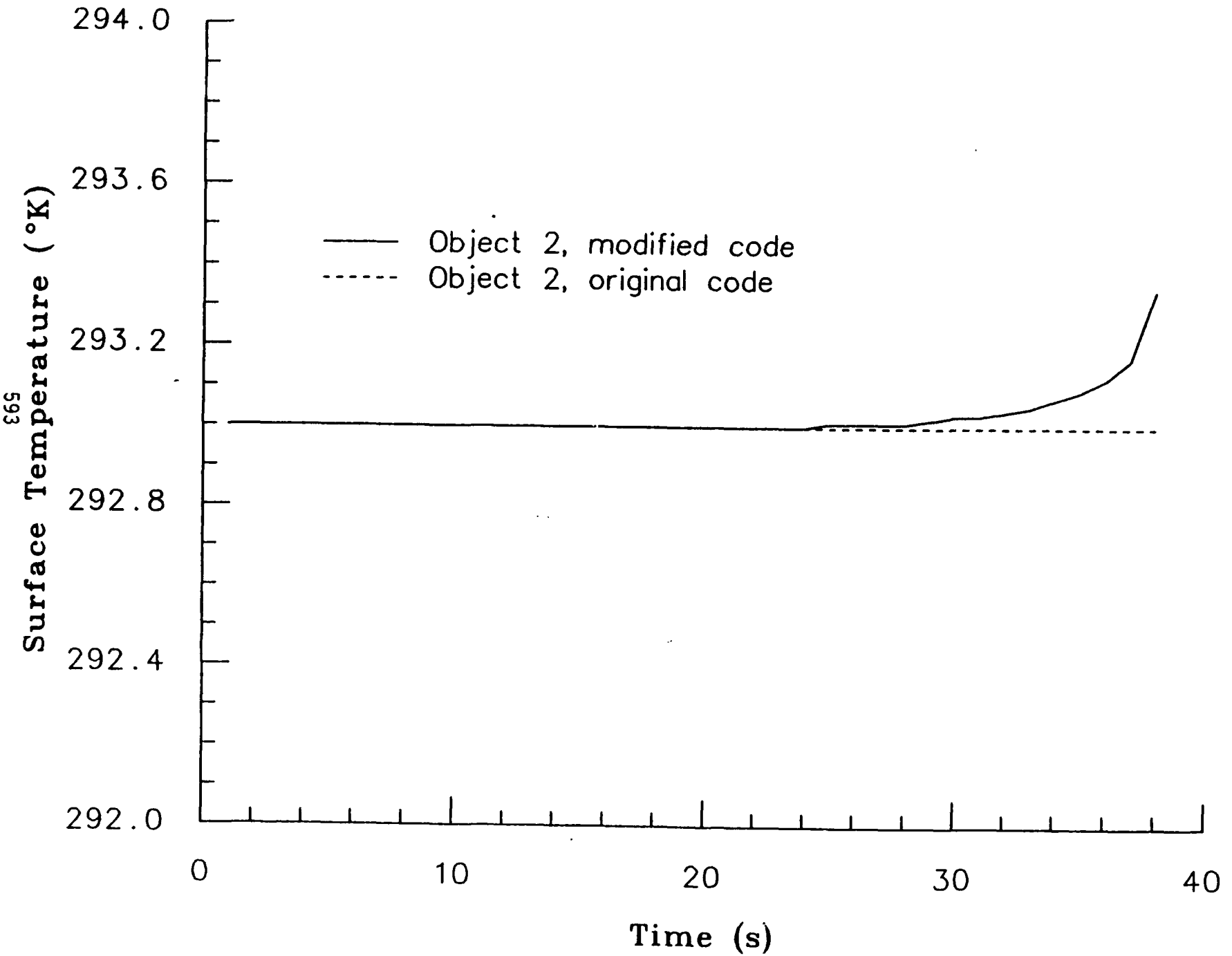


Fig. 6: Surface Temp. at Selected Points

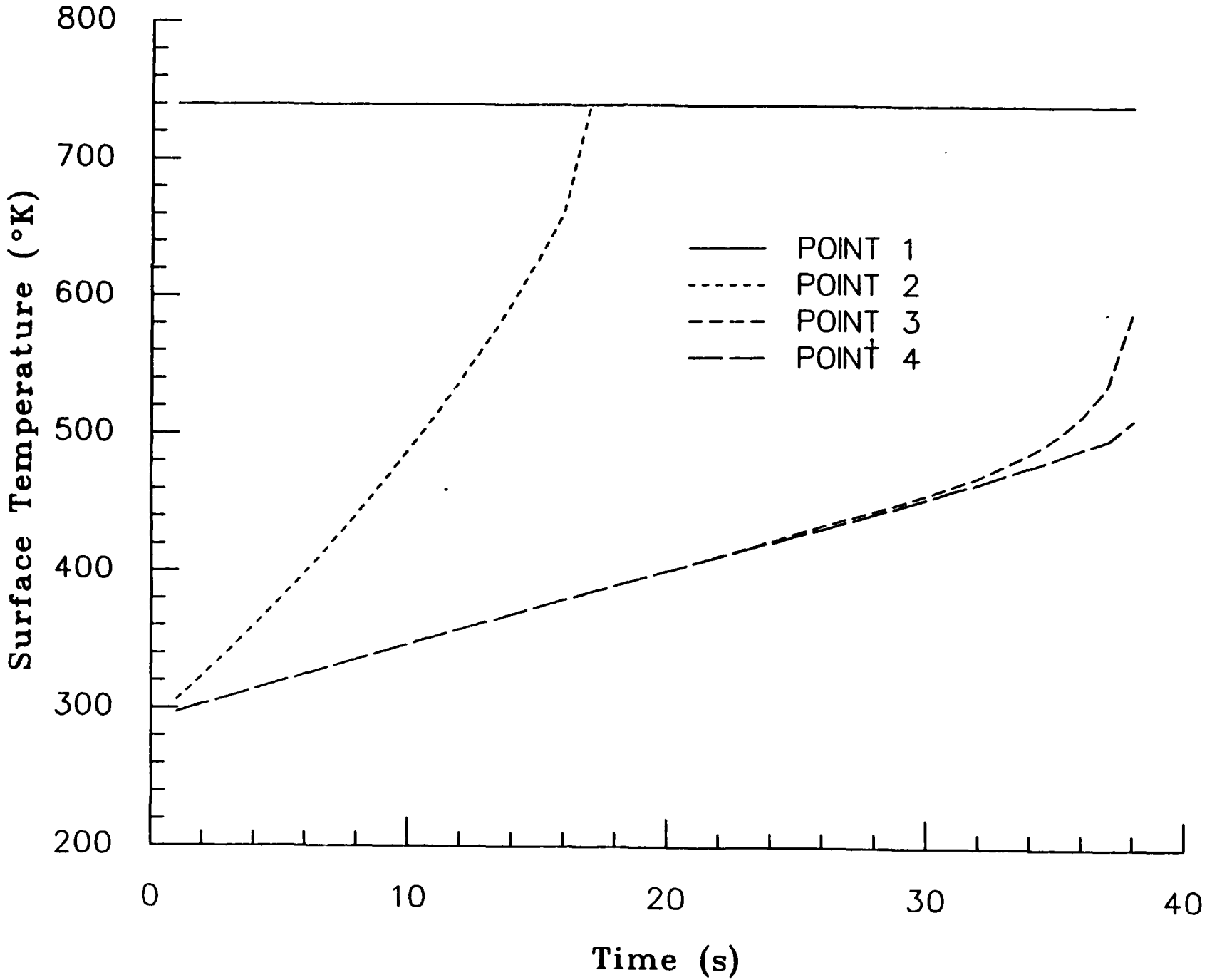
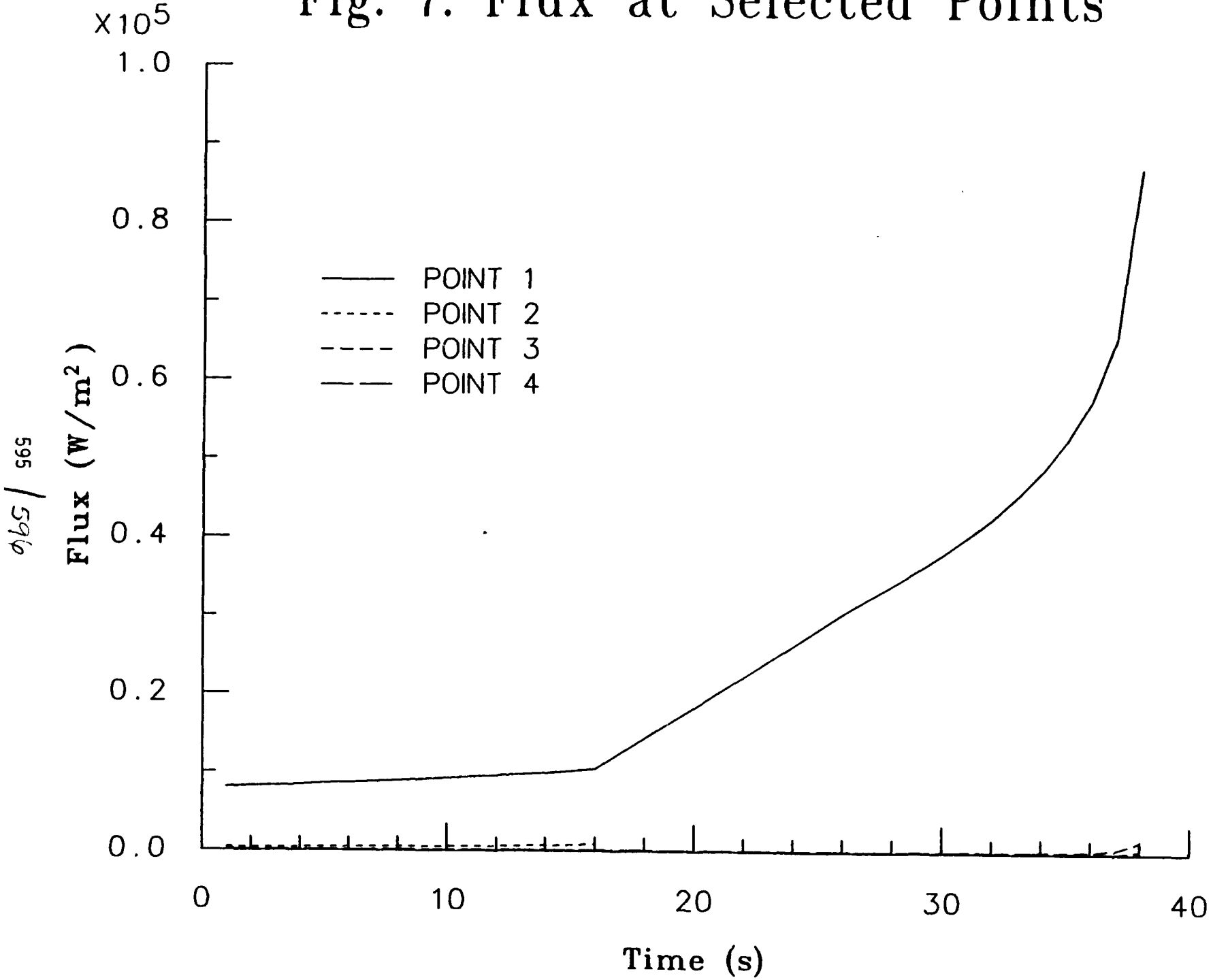


Fig. 7: Flux at Selected Points



ON THE NUMERICAL SOLUTION OF A SYSTEM OF PARTIAL DIFFERENTIAL EQUATIONS TO OBTAIN THE WIND FROM THE GEOPOTENTIAL FOR NUMBERICAL WEATHER PREDICTION AND ON RELATED MATHEMATICAL ASPECTS.

H. Baussus von Luetzow

U.S. Army Engineer Topographic Laboratories

ABSTRACT. The paper first discusses the numerical solution of a system of partial differential equations as optimal filter equations to obtain the wind field from the geopotential field under consideration of solution constraints. It then addresses the use of these equations in the case of available horizontal winds, constrained initialization in the sense of Sasaki, non-statistical smoothing and univariate and multivariate estimation and smoothing of meteorological variables, and a special problem of data assimilation. Finally, it outlines a non-hydrostatic prognostic approach in the case of highly accurate and dense initial meteorological fields. Although emphasizing the mathematical point of view including the need for parallel and large scale computing, the paper also endeavors to relate to the present state of the art in numerical weather prediction.

I. INTRODUCTION. Present operational numerical weather prediction models use the hydrostatic equation and require hydrodynamic and hydrostatic stability which, in the numerical solution process, is enforced by a convective adjustment. The horizontal grid length used is generally greater than 50 Km, and the number of vertical levels is usually 15 or less. In a prediction system with pressure p as the vertical independent coordinate, the dependent variables are the wind components u and v, the geopotential ϕ, the diabatic rate of heat dq/dt and the mixing ratio r, sometimes replaced by the relative humidity. Further considered are the saturation mixing ration r_s, the precipitation criterion δ, and the sea surface temperature T_s. The so-called primitive equations which incorporate the hydrostatic equation are the two equations of horizontal motion, the continuity equation, the thermodynamic equation, and the continuity equation for the mixing ratio. The latter may be supplemented by continuity equations for substances other than water. In global models, spherical coordinates are employed in the horizontal. Most models use normalized pressure $\sigma = p p_s^{-1}$ where p_s stands for surface pressure. Further, so-called spectral methods are generally used in major forecast centers to compute horizontal derivatives with a high accuracy. The determination of initial fields of dependent variables is the subject of objective analysis. The most important endeavor is the estimation of ϕ, u, and v at regular grid points from irregularly distributed geopotential and wind data by multivariate statistical methods and the geostrophic relationship between the wind and the geopotential. In the past, geopotential data have been considerably more numerous than wind data, and in the near future this situation is not expected to change. As a consequence, in order to use wind and geopotential data in the numerical integration process, filter equations have been developed classified as static, dynamic, and normal mode initialization. Unfortunately, the determination of the wind field from the geopotential field is unsatisfactory in the equatorial belt where wind estimates are presently obtained at about 2 levels from the movements of clouds and not in a desirable density. New

earth observing systems under development are expected to provide highly accurate and dense wind measurements over many areas of the globe. Denser and more accurate measurements of surface pressure, temperature, and humidity will be of utmost importance for numerical weather prediction.

Experience has shown that numerical humidity forecasts are only satisfactory for 1-2 days in contrast with better predictions of the other meteorological variables. This is partially due to the coarse representation of the humidity field which requires a relatively smaller grid of resolution, both horizontally and vertically. However, incorporation of a more detailed humidity field would result in a greater number of hydrodynamic-hydrostatic instabilities through the interaction of the continuity equation of water vapor with the thermodynamic equation. As shown by Baussus von Luetzow (1980), convection on the mesoscale, requiring a horizontal grid resolution of about 10Km, necessitates a more sophisticated equation for the vertical wind velocity and the application of the unmodified continuity equation in a coordinate system with z as the vertical coordinate. Parameterization of certain sub-grid processes, like moist cumulus convection, would also be more successful in the non-hydrostatic prediction system. Simultaneously, Baussus von Luetzow described a signal generation process approximately equivalent to the present hydrostatic forecast system, particularly for a period of several days. Significant in this respect is the statement by Ghil and Childress (1987) that the practical limit of usefulness of numerical weather forecasts is between 3-7 days. This limit could, however, be extended by nonhydrostatic forecasts in combination with a much denser and more accurate data base than presently available. The specification of lower and upper boundary values required for the numerical solution process presents additional difficulties. Their inaccuracies tend to degrade the forecast with increasing prediction time. As to an improvement of upper boundary values, Baussus von Luetzow's signal generation system, cited above, reduced to one level, has some potential value.

This paper addresses in section II the numerical solution of optimal filter equations with emphasis on the determination of the wind from the geopotential as the main effort. Section III contains some considerations about the incorporation of friction. The performance characteristics of new earth observing systems are shown in section IV. The use of filter equations as conditional equations, including constrained initialization, is the subject of section V. Section VI is concerned with a critique of objective analysis as practiced at this time. Relevant comments about a nonhydrostatic approach are made in section VII, and section VIII enumerates some pertinent conclusions. Throughout the paper, including the introduction, the author has endeavored to relate to the present state of the art in numerical weather prediction and to offer some new and/or relevant points of view.

II. NUMERICAL SOLUTION OF OPTIMAL FILTER EQUATIONS. According to Baussus von Luetzow (1971, 1980), the following system of diagnostic filter equations can be derived from the hydrostatic equations of motion without friction under partial use of the thermodynamic equation in a planar x,y,p,t-system:

$$\left(f+\frac{\partial u}{\partial y}\right)\frac{\partial^2\phi}{\partial x^2}+\left(\frac{\partial v}{\partial y}-\frac{\partial u}{\partial x}\right)\frac{\partial^2\phi}{\partial x\,\partial y}+\left(f-\frac{\partial v}{\partial x}\right)\frac{\partial^2\phi}{\partial y^2}$$
$$+\frac{\partial\omega}{\partial y}\frac{\partial^2\phi}{\partial x\,\partial p}-\frac{\partial\omega}{\partial x}\frac{\partial^2\phi}{\partial y\,\partial p}+\beta\left(\frac{\partial\phi}{\partial y}+2fu\right)-f^2\left(\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)=0 \qquad (1)$$

$$H_2 \Delta^2\omega + (f^2 + \Delta^2\phi)\frac{\partial^2\omega}{\partial p^2} - \frac{\partial^2\phi}{\partial x\,\partial p}\frac{\partial^2\omega}{\partial x\,\partial p} - \frac{\partial^2\phi}{\partial y\,\partial p}\frac{\partial^2\omega}{\partial y\,\partial p}$$

$$+ a_1\frac{\partial\omega}{\partial x} + a_2\frac{\partial\omega}{\partial y} + a_3\omega = F(x,y,p) \qquad (2)$$

In eq.(1), f is the Coriolis parometer and β stands for $\frac{\partial f}{\partial y}$. In eq. (2), $\omega = \frac{dp}{dt}$ is the generalized vertical velocity, and H_2 is the effective static stability. The terms a_1, a_2 and a_3 are functions of the geopotential and of H_2, and $F(x,y,p)$ is comprised of functionals involving u, v, ϕ, and spatial derivatives thereof, and of the radiation component of $\frac{dq}{dt}$.

Subsequently, Baussus von Luetzow (1988) showed that eqs. (1) and (2) are optimal filter equations or equilibrium solutions free of high frequency gravity-inertia waves and superior to normal mode initialization. Filter equations using spherical coordinates ϑ and λ which correspond to eqs. (1) and (2) can be derived as well.

Actually, the filtered variables in eqs. (1) and (2) should be designated by a symbol, eg., $\wedge$, to distinguish them from unfiltered variables. However, ϕ tends to be a non-fluctuating variable in the first place, and it is the primary purpose of eqs. (1) and (2) to determine filtered or relatively smooth winds primarily from the geopotential. Finally, the solution of the prognostic equations in a discrete manner implies the use of sufficiently smooth variables.

In the iterative, interactive solution of the system (1) and (2) it is necessary to observe the Helmholtz decomposition

$$u = u_1 + u_2 = -\frac{\partial\psi}{\partial y} + \frac{\partial\chi}{\partial x} \qquad (3a)$$

$$v = v_1 + v_2 = \frac{\partial\psi}{\partial x} + \frac{\partial\chi}{\partial y} \qquad (3b)$$

This decomposition applies only to filtered variables free of vertical transverse waves.

Using eqs. (3a) and (3b), vorticity and divergence can be expressed as

$$\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \Delta^2\psi \qquad (4a)$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = \Delta^2\chi = -\frac{\partial\omega}{\partial p} \qquad (4b)$$

Because of the dominance of the vorticity in horizontal motion, the first approximation to eq. (1) can be formulated as

$$\left(f^2 + \frac{\partial^2\phi}{\partial y^2}\right)\frac{\partial^2\psi^{(1)}}{\partial x^2} - 2\,\frac{\partial^2\phi}{\partial x\,\partial y}\frac{\partial^2\psi^{(1)}}{\partial x\,\partial y} + \left(f^2 + \frac{\partial^2\phi}{\partial x^2}\right)\frac{\partial^2\psi^{(1)}}{\partial y^2} + 2\beta f\frac{\partial\psi^{(1)}}{\partial y}$$

$$= f\Delta^2\phi + \beta\frac{\partial\phi}{\partial y} \qquad (5)$$

It is obvious that eq. (5) is not effective in the equatorial belt where the neglected terms involving the velocity potential χ approach those associated with the stream function ψ. Wind observations in the equatorial belt are thus indispensable for successful numerical weather prediction in the tropics. Additionally, the source function on the right side of eq. (5) and the partial derivatives of the geopotential are not strong in the equatorial belt.

If the smooth wind field is Known, the geopotential can be computed by means of eq. (1) without reference to the omega equation (2). Again, the source function $f^2\Delta^2\psi - 2\beta f u$ would be generally small in the equatorial belt. Because of the relatively loose functional relationship between wind and geopotential in the equatorial belt, they should be independently and as accurately as possible determined by measurements.

The solution criterion for the first approximation $\psi^{(1)}$ is obtained from eq. (5) as

$$\left(f^2 + \frac{\partial^2\phi}{\partial x^2}\right)\left(f^2 + \frac{\partial^2\phi}{\partial y^2}\right) - \left(\frac{\partial^2\phi}{\partial x \partial y}\right)^2 > 0 \qquad (6)$$

It is generally fulfilled, corresponds to the criteria applicable to the omega equation (2), and can be imposed if necessary. In this respect remember that ϕ is generally not free of "measurement" errors.

The solution criteria for the omega equations are

$$H_2 > 0 \qquad (7a)$$

$$f^2 + \Delta^2\phi > 0 \qquad (7b)$$

$$H_2\left(f^2 + \Delta^2\phi\right) - \frac{1}{4}\left[\left(\frac{\partial^2\phi}{\partial x \partial p}\right)^2 + \left(\frac{\partial^2\phi}{\partial y \partial p}\right)^2\right] > 0 \qquad (7c)$$

They can be imposed if required.

The filter equations (1) and (2) are applicable for a horizontal grid size $\Delta x = \Delta y \geq 100\ Km$ and are compatible with about 15 vertical levels.

Equation (5) may, for example, be solved for a square region 2000Km · 2000Km with $\Delta x = \Delta y = 200\ Km$. In this case, there are 81 $\psi^{(1)}$ - unKnowns. An effective solution process is the following:

(1) Formulation of finite difference equations for each interior grid point.
(2) Computation of functionals $\frac{\partial^2\phi}{\partial x^2}$, etc.
(3) Establishment of the matrix equation

$$A_{ij}\,\psi_{ij}^{(1)} = F_{ij} \qquad (8)$$

with $\psi^{(1)} = f^{-1}\phi$ on the boundary.
(4) Solution of eq. (8) as

$$\psi_{ij}^{(1)} = A_{ij}^{-1}\,F_{ij} \qquad (9)$$

but only for the central interior point where $i = m$, $j = m$.
(5) Equation(9)- type solutions for moving central points by translational shifts of the quadratic integration area by $l\Delta x$, $m\Delta y$ with $l, m = 1, 2, 3, \cdots$.
(6) Identification of about 25 square regions as an aggregate region.
(7) Parallel processing for separate aggregate regions.

(8) Improvement of initial $\psi_i^{(1)}$ – solutions by improved boundary values $\psi_{iB}^{(1)}$, resulting in a correction term

$$\delta\psi^{(1)} = A_{ij}^{-1}\,\delta F_{ij} \tag{10}$$

Under consideration of 9 intermediate levels, limited by a ground and a top level, the fundamental ω-integration region comprises 729 ω-unKnowns. The solution process to obtain $\omega^{(1)}$ is the following:
(1) Formulation of finite difference equations for each interior grid point.
(2) Computation of coefficient functionals and of $F(x,y,p)$ under utilization of $\psi^{(1)}$ and $\frac{dt_g}{dt}$ as radiation heat compatible with the grid resolution.
(3) Establishment of the matrix equation

$$A_{ijk}\,\omega_{ijk} = F_{ijk} \tag{11}$$

with $\omega = 0$ at the upper boundary and at the lateral boundaries and $\omega_g = c\,(u_g \frac{\partial\phi_g}{\partial x} + v_g \frac{\partial\phi_g}{\partial y})$ where c is a constant, u_g and v_g are the wind components at ground level, and ϕ_g is the geopotential of the ground commensurate with the horizontal grid resolution. As an approximation, u_g and v_g may be replaced by $-\frac{\partial\psi^{(1)}}{\partial y}$ and $\frac{\partial\psi^{(1)}}{\partial x}$ obtained at the lowest level.
(4) Solution of eq. (11) as

$$\omega_{ijk} = A_{ijk}^{-1}\,F_{ijk} \tag{12}$$

but only for central interior points where $i = m, j = m$, resulting in ω_{mmk} – determinations.
(5) Translational shifts and parallel processing for separate aggregate regions in accordance with those applicable to $\psi^{(1)}$ – determination.

The solution process to obtain improved stream functions $\psi^{(2)}$ is:
(1) Computation of $\chi^{(1)}$ from $\Delta^2\chi^{(1)} = -\frac{\partial\omega^{(1)}}{\partial p}\chi^{(1)}$.
(2) Computation of $u_2^{(1)}$ and $v_2^{(1)}$ from $u^{(1)}$ and $v^{(1)}$
(3) Computation of
(4) Determination of the omitted terms of eq. (1) by finite difference methods, yielding a right side corrective source function ΔF_{ij} in eq. (5) where $\psi^{(1)}$ is to be replaced by $\psi^{(2)}$.
(5) Computation of

$$\psi_{ij}^{(2)} - \psi_{ij}^{(1)} = \Delta\psi_{ij} \sim A_{ij}^{-1}\,\Delta F_{ij} \tag{13}$$

but only for central interior points.
(6) Use of eq. (13) for all computation grid points, i.e., with variable A_{ij}^{-1} and ΔF_{ij}.

If indicated by experimentation, an improved $\omega^{(2)}$ – solution may be attempted which can also be achieved in a differential form without new matrix inversions.

The solution of a system of linear equations with unKnowns of the order 10^3 can be accomplished without time constraints by the new generation of supercomputers which are a requisite for the timely and accurate solution of the omega equation. According to Elmer-DeWitt (1988), the Cray-3 will be released in 1989, soon followed by the Cray-4. IBM and AT&T Bell Laboratories are on the verge of introducing new parallel-processing computers, and Sandia National Laboratories has coaxed a 1024 - processing computer. Fortunately, the matrices to be inverted are essentially band matrices with numerous zero elements. Care has to be exercised to further the stability of the solution both in regard

to ellipticity and to accurate coefficient functions of the omega equation, notably of H_2. Additionally, round off errors have to be controlled. Whether application of Gauss' transformation, matrix splitting, and successive overrelaxation methods are promising, only experimentation can tell. If sufficiently accurate predicted ω-data are available at initialization time t_o, these might be used as lateral boundary values with a resulting decrease of the fundamental computational region.

III. INCORPORATION OF FRICTION. The inclusion of frictional terms F_u and F_v for the surface layer (0-100m) and D_u and D_v for the Prandtl layer (100-1000m) precludes a good determination of filtered winds from the geopotential by means of eqs. (1) and (2). The structure of these terms have been discussed, among others, by Kasahara (1977) and Corby, Gilchrist and Rowntree (1977). The frictional terms increase the divergence in comparison with the relative vorticity and tend to make the flow hydrodynamically unstable, particularly in low latitudes. Only in the stationary case $\frac{dV}{dt} = 0$, where V denotes the velocity vector, and under consideration of simplified frictional terms, approximate wind components may be computed from the geopotential. The general filter equations (1) and (2) are only fully effective in the free atmosphere.

IV. NEW EARTH OBSERVING SYSTEMS. The lack of a well synchronizeable, dense and accurate global data base has been the greatest drawback for numerical weather prediction. New earth observing systems under development and assumed to be operational in the foreseeable future can be expected, in conjunction with improved modified objective analysis, improved initialization and prognostic models, and supercomputers to revolutionize numerical weather prediction. The new systems of particular interest are the subject of the LAWS, HMMR and LASA Instrument Panel Reports (1987), published by the National Aeronautics and Space Administration. The performance characteristics of the above systems are:

LAWS - LASER ATMOSPHERIC WIND SOUNDER (DOPPLER LIDAR)
100 KM HORIZONTAL RESOLUTION
1 KM VERTICAL RESOLUTION
1-2 MS - 1 LOWER TROPOSPHERE
2-5 MS - 1 UPPER TROPOSPHERE
CLOUD COVER AND RAIN REMAIN OBSTACLES

HMMR - HIGH RESOLUTION MULTIFREQUENCY MICROWAVE RADIOMETER
IMPROVED TEMPERATURE PROFILES (±1K)
IMPROVED HUMIDITY PROFILES (±10 PERCENT)
HORIZONTAL RESOLUTION 10 KM
VERTICAL RESOLUTION 0.5 KM

LASA - LIDAR ATMOSPHERIC SOUNDER AND ALTIMETER
SURFACE PRESSURE (± 2MB)
VERTICAL PROFILES OF TEMPERATURE AND PRESSURE FROM THE STRATOSPHERE THROUGH THE TROPOSPHERE, TO THE GROUND TEMPERATURE IN TROPOSPHERE AND STRATOSPHERE.

The prognostic horizontal grid resolution should be smaller than 100 KM to

minimize forecast errors, especially in medium to long range predictions.

V. USE OF FILTER EQUATIONS AS CONDITIONAL EQUATIONS. The filter equations (1) and (2) will still be useful after the introduction of LAWS since cloud cover and rain present obstacles to wind measurements or accurate wind measurements.

The first filter equation or both filter equations may be used in connection with "measured" winds and geopotentials to obtain improved initial fields.

In a first application optimally smoothed LAWS wind measurements in the lower troposphere may be used to determine $\phi(u, v)$ by virtue of eq. (1). The "measured" optimally smoothed geopotential may be ϕ_m. An improved geopotential would then result as

$$\hat{\phi} = K_1 \phi(u, v) + K_2 \phi_m \qquad (14)$$

where K_1 and K_2 are regression coefficients.

In a second application, LAWS wind measurements in the upper troposphere may be improved and to a lesser extent the geopotential in the following sequence:

(1) $\phi(u, v)$ is determined from eq. (1).
(2) An improved $\hat{\phi}$ is calculated according to eq. (14).
(3) $\omega_c(\hat{\phi}, u, v, \frac{d\phi}{dt})$ is computed by means of eq. (2).
(4) It is then possible to formulate an improved $\hat{\omega} = m_1 \omega_c + m_2 \omega_m$ where m_1 and m_2 are regression coefficients and where $\omega_m = \omega_m(u, v, \omega_g)$.
(5) $\hat{u}_2$ and $\hat{v}_2$ are computed from $\frac{\partial \hat{\omega}}{\partial p}$.
(6) $\hat{\psi} = \hat{\psi}(\hat{\phi}, \hat{u}_2, \hat{v}_2, \hat{\omega})$ is determined from eq. (1).

The above procedure would simultaneously provide a multiple consistency check, and the integration domain for the solution of eqs. (1) and (2) could be reduced.

A third application would be the estimation of an improved geopotential from the "measured" geopotential $\phi_m = \phi_m(t_o)$ and from the computed geopotential $\phi_c = \phi_c(t_o)$ in the form $\hat{\phi} = n_1 \phi_m + n_2 \phi_c$ with n_1 and n_2 as regression coefficients. Potentially, a multiple regression approach could be used. Merging a measured and a computed field would only be warranted in the case of relatively large $\phi_m -$ errors which are not uniform because of the actual estimation of ϕ_m in the context of objective analysis, addressed in section VI. The above approach has been suggested by Hoffman and Kalnay (1983). The improved $\hat{\phi} -$ field can then be used for initialization by means of eqs. (1) and (2).

In constrained initialization according to Haltiner and Williams (1980), the integral

$$I = \iint_S \left[\alpha(\phi - \tilde{\phi})^2 + \beta(\nabla\psi + k \times \tilde{V})^2 + 2\lambda M\right] dS \qquad (15)$$

is minimized by a variational method. In eq. (15), α and β are generally latitude-dependent weights, $\tilde{\phi}$ and $\tilde{V}$ denote fields obtained from objective

analysis, λ is a variable Lagrangian multiplier, M is the classical truncated balance equation, and S is the integration area. Variation of I under neglect of nonlinear terms in M yields two differential equations which, together with M, permit the determination of λ and of improved fields of ϕ and ψ.

The above approach has some merit in the context of recent ϕ and V coverage and $\tilde{\phi}$ and $\tilde{V}$ estimation by objective analysis. It is not satisfactory in the case of sufficiently dense and accurate ϕ-fields and their use in eqs. (1) and (2). It is not required in future more accurate wind and geopotential determinations associated with new earth observing systems. Substitution of eq. (5) for the truncated classical balance equation M would result in somewhat improved ϕ and ψ solutions.

VI. OBJECTIVE ANALYSIS. Present objective analysis concentrates on the estimation of the geopotential and, secondarily, on the estimation of winds, using "measured" and generally not uniformly distributed data. The estimation of the geopotential can be characterized as

$$\delta\hat{\phi}_m = \sum_i a_{mi}(\delta\phi_i + e_{1i}) + \sum_j b_{mj}(\delta u_m + e_{2m}) + \sum_j c_{mj}(\delta v_m + e_{3m}) \qquad (16)$$

In eq. (16), the symbol δ denotes deviations from climatological means, e_1, e_2, and e_3 are uncorrelated measurement errors, and a_{mi}, b_{mj}, and c_{mj} are regression coefficients. The subscript m indicates estimation at point P_m. The multivariate estimation (16) is generally performed at one isobaric level although it can be extended to other, reasonably close isobaric levels. Covariance analysis of the geostrophic relationship permits the estimation of $\delta\hat{\phi}_m$ and of $\hat{\phi}_m = \bar{\phi}_m + \delta\hat{\phi}_m$ with $\bar{\phi}_m$ as the climatological mean. A recent review of methods of objective analysis has been made by Gustafsson (1981).

The approach (16) is not satisfactory in the case of future more accurate and more uniformly distributed "measured" winds and geopotentials and because the meteorological generation process is neither ergodic nor stationary. Newly developed advanced smoothing techniques such as those addressed by Adams, Willsky, and Levy (1984) would be more appropriate.

VII. NONHYDROSTATIC APPROACH. Highly accurate and dense measurements of pertinent meteorological variables, provided by new earth observing systems, in combination with advanced smoothing techniques might permit the replacement of the primitive equations with a nonhydrostatic prediction system as outlined by BAUSSUS von LUETZOW (1980). This system with z as the vertical coordinate has a more complicated diagnostic equation for the vertical wind component w and leaves the continuity equation invariant, i.e., introduces an additional degree of freedom. Application of the w-equation, however, requires improved condensation criteria and additionally incorporation of improved parameterization of moist cummulus convection to be highly effective. Only then can the inital humidity field be fully exploited. In agreement with Anthes, Kuo, Baumhefner, Errico, and Bettge (1985), the nonhydrostatic system would be able to cope with meso- scale phenomena including frontal and jetlike discontinuities, flows produced in response to small scale topographic forcing, and large-amplitude instabilities such as convective storms.

VIII. CONCLUSION.

(1) The wind field can be satisfactorily computed from the geopotential field in the free atmosphere and outside the equatorial belt by the numerical solution of a system of two partial differential equations, using supercomputers and parallel processing. This method is also promising in the case of Doppler Lidar failure.
(2) The optimal diagnostic filter equations permit the determination of improved geopotential and wind fields in the case of both uniform and dense "measured" wind and geopotential coverage, particularly in the upper troposphere.
(3) New earth observing systems with the capability to provide uniform, dense, and more accurate determinations of meteorological variables and advanced smoothing techniques permit the application of a nonhydrostatic prediction system which could fully exploit the availability of the initial humidity field.

REFERENCES

Adams, M., Willsky, A., and Levy, B. 1984. Linear Estimation of Boundary Value Stochastic Processes - Part I: The Role and Construction of Complementary Models. Transactions on Automatic Control, Vol. AC-29, No. 9.

Anthes, R., Kuo, Y., Baumhefner, D. Errico, R., and Bettge, T. 1985. Predictability of Mesoscale Atmospheric Motions. Advances in Geophysics, Vol. 28, Academic Press, New York.

Baussus von Luetzow, H. 1971, The Derivation and Potential of New Filter Equations for Numerical Weather Prediction. Research Note ETL-RH-71-3, U.S. Army Engineer Topographic Laboratories, Fort Belvoir, VA 22060-5546.

Baussus von Luetzow, H. 1980. On the Limitations and Improvement of Present Numerical Weather Prediction. Transactions, 25th Conference of Army Mathematicians, ARO Report 80-1, Army Research Office, Triangle Research Park, NC 27709-2211.

Baussus von Luetzow, H. 1988. Normal Mode and Superior Initialization for Numerical Weather Prediction and Related Aspects. Manuscript to be submitted for publication. U.S. Army Engineer Topographic Laboratories, Fort Belvoir, VA 22060-5546.

Corby, G., Gilchrist, A., and Rowntree, P. 1977. United Kingdom Meteorological Office Five-Level General Circulation Model. Methods in Computational Physics, Vol. 17, Academic Press, New York.

Elmer-Dewitt, P. 1988. Fast and Smart. Designers race to build the supercomputers of the future. TIME, Vol. 131, No. 13.

Ghil, M., and S. Childress. 1987. Topics in Geophysical Fluid Dynamics. Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics. Applied Mathematical Sciences, Vol. 60, Springer - Verlag, New York.

Gustaffson, N. 1981. A Review of Methods for Objective Analysis. Dynamic Meteorology. Data Assimilation Methods. Springer-Verlag, New York.

Haltiner, G., and R. Williams. 1980. Numerical Prediction and Dynamic Meteorology, Second Edition. John Wiley & Sons, New York.

Hoffman, R., and E. Kalnay. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus, Vol. 35A, Number2.

Instrument Panel Report HMMR, High-Resolution Multifrequency Microwave Radiometer. 1987. National Aeronautics and Space Administration, Washington, DC.

Instrument Panel Report LASA, Lidar Atmospheric Sounder and Altimeter. 1987. National Aeronautics and Space Administration, Washington, DC.

Instrument Panel Report LAWS, Laser Atmospheric Wind Sounder, 1987. National Aeronautics and Space Administration, Washington, DC.

Kasahara, A. 1977. Computational Aspects of Numerical Models for Weather Prediction and Climate Simulation. Methods in Computational Physics, Vol. 17, Academic Press, New York.

Literature used in addition to the above references:

Berg, L. 1986. Lineare Gleichungssysteme mit Bandstruktur. Carl Hanser Verlag, Muenchen.

Rice, J. and R. Boisvert. 1985. Solving Elliptic Problems Using ELL PACK. Springer - Verlag, New York.

Westlake, J. 1968. A Handbook of Numerical Matrix Inversion and Solution of Linear Equations. J. Wiley & Sons, New York.

COMPUTER ALGEBRA IMPLEMENTATION OF LIE TRANSFORMS FOR HAMILTONIAN SYSTEMS: APPLICATION TO THE NONLINEAR STABILITY OF L_4

Vincent T. Coppola and Richard H. Rand
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

Introduction

This work concerns Lie transforms, a method for obtaining approximate solutions to systems of differential equations. We apply the method to a general class of two degree of freedom Hamiltonian systems, viz., two coupled nonlinear oscillators with nonresonant frequencies. For systems in this class, we use Lie transforms to approximately reduce the system to an equivalent simpler system which is immediately solvable, i.e., a system with ignorable coordinates.

As an application of our results, we determine the nonlinear stability of the triangular points in the circular restricted three body problem. In doing so we corroborate a computation recently performed by Meyer and Schmidt [16]. Their computation was based on their own computer algebra program written in PL/I, whereas the present work is based on readily available utilities written in MACSYMA [19]. Moreover, while their computation was specifically performed for the problem at the triangular point L_4, the present work applies to a problem with arbitrary (symbolic) coefficients.

We begin by introducing the reader to Lie transforms. Then we show how the method may be applied to a particular class of problems, and finally we specialize the results to some examples, including the problem at L_4.

Lie Transforms

In this section we summarize the method of Lie transforms (see [8], [12], [15], [17]). This work is concerned with Hamiltonian systems, i.e. systems which are derivable from a single scalar function H, the Hamiltonian:

$$(1) \qquad \frac{dx_m}{dt} = \frac{\partial H}{\partial y_m}, \quad \frac{dy_m}{dt} = -\frac{\partial H}{\partial x_m}$$

where x_m and y_m are the dependent variables of the problem. $m = 1,\ldots,N$, where N is called the number of degrees of freedom. The method of Lie transforms generates a near-identity transformation from (x_m,y_m) to (X_m,Y_m) variables,

$$(2) \qquad \begin{aligned} x_m &= X_m + \begin{matrix}\text{quadratic terms} \\ \text{in } (X_k,Y_k)\end{matrix} + \begin{matrix}\text{cubic terms} \\ \text{in } (X_k,Y_k)\end{matrix} + \cdots \\ y_m &= Y_m + \begin{matrix}\text{quadratic terms} \\ \text{in } (X_k,Y_k)\end{matrix} + \begin{matrix}\text{cubic terms} \\ \text{in } (X_k,Y_k)\end{matrix} + \cdots \end{aligned}$$

which is canonical, i.e., which preserves the Hamiltonian form of the equations:

$$(3) \qquad \frac{dX_m}{dt} = \frac{\partial K}{\partial Y_m}, \quad \frac{dY_m}{dt} = -\frac{\partial K}{\partial X_m}$$

where $K = K(X_m,Y_m) = H(x_m,y_m)$ is the Hamiltonian in the new variables (called the Kamiltonian after Goldstein [11]).

The near-identity transformation is generated by first introducing a scaling parameter ϵ into the problem. Expanding H in a power series about the origin (assumed to be an equilibrium position),

$$(4) \qquad H = H_0(x_m, y_m) + \epsilon H_1(x_m, y_m) + \epsilon^2 H_2(x_m, y_m) + \cdots$$

where $H_n(x_m, y_m)$ is a polynomial of degree n+2. Then the near-identity transformation is generated by the associated Hamiltonian system

$$(5) \qquad \frac{dx_m}{d\epsilon} = \frac{\partial W}{\partial y_m}, \quad \frac{dy_m}{d\epsilon} = -\frac{\partial W}{\partial x_m},$$

in which ϵ plays the role of time. The transformation evolves in ϵ, starting with the initial conditions

$$(6) \qquad \epsilon = 0, \quad x_m = X_m, \quad y_m = Y_m.$$

The Hamiltonian W of eqs.(5), called the generating function, is also expanded in a power series in ϵ:

$$(7) \qquad W = W_1 + \epsilon W_2 + \epsilon^2 W_3 + \cdots,$$

where W_n is a polynomial of degree n+2. The point of this generating scheme is that the resulting transformation is canonical for any choice of the W_n's (see [8], [15]). The actual choice of these functions depends upon the problem at hand, but the main idea is to pick them so that the new Hamiltonian K is as simple as possible. We note that the parameter ϵ in this paper corresponds to $-\epsilon$ in [15] and [19].

The transformation is generated by expanding the variables (x_m, y_m) in Taylor series in ϵ and using the generating equations (5)-(7) to evaluate the coefficients:

$$(8) \qquad x_m = x_m\Big|_{\epsilon=0} + \frac{dx_m}{d\epsilon}\Big|_{\epsilon=0} \epsilon + \frac{d^2x_m}{d\epsilon^2}\Big|_{\epsilon=0} \frac{\epsilon^2}{2} + \cdots$$

(9) $$x_m\Big|_{\epsilon=0} = X_m$$

(10) $$\frac{dx_m}{d\epsilon}\Big|_{\epsilon=0} = \frac{\partial W}{\partial y_m}\Big|_{\epsilon=0} = \frac{\partial W_1}{\partial Y_m}$$

(11) $$\frac{d^2x_m}{d\epsilon^2}\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\frac{\partial W}{\partial y_m}\Big|_{\epsilon=0}$$

$$= \frac{\partial}{\partial\epsilon}\frac{\partial W}{\partial y_m} + \sum_j \frac{\partial^2 W}{\partial x_j \partial y_m}\frac{dx_j}{d\epsilon} + \frac{\partial^2 W}{\partial y_j \partial y_m}\frac{dy_j}{d\epsilon}\Big|_{\epsilon=0}$$

$$= \frac{\partial W_2}{\partial Y_m} + \sum_j \frac{\partial^2 W}{\partial x_j \partial y_m}\frac{\partial W}{\partial y_j} - \frac{\partial^2 W}{\partial y_j \partial y_m}\frac{\partial W}{\partial x_j}\Big|_{\epsilon=0}$$

$$= \frac{\partial W_2}{\partial Y_m} + \sum_j \frac{\partial^2 W_1}{\partial X_j \partial Y_m}\frac{\partial W_1}{\partial Y_j} - \frac{\partial^2 W_1}{\partial Y_j \partial Y_m}\frac{\partial W_1}{\partial X_j}$$

$$= \frac{\partial W_2}{\partial Y_m} + \{\frac{\partial W_1}{\partial Y_m}, W_1\}$$

where the Poisson or Lie bracket $\{f,g\}$ is given by

(12) $$\{f,g\} = \sum_j \frac{\partial f}{\partial X_j}\frac{\partial g}{\partial Y_j} - \frac{\partial f}{\partial Y_j}\frac{\partial g}{\partial X_j} .$$

The transformation is thus found to be given by

(13) $$x_m = X_m + \frac{\partial W_1}{\partial Y_m}\epsilon + \left[\frac{\partial W_2}{\partial Y_m} + \{\frac{\partial W_1}{\partial Y_m}, W_1\}\right]\frac{\epsilon^2}{2} + \cdots$$

and similarly,

$$(14) \qquad y_m = Y_m - \frac{\partial W_1}{\partial X_m}\epsilon - \left[\frac{\partial W_2}{\partial X_m} + \{\frac{\partial W_1}{\partial X_m}, W_1\}\right]\frac{\epsilon^2}{2} + \cdots$$

In order to obtain the transformed Kamiltonian K (cf. eq.(3)), the transformation (13),(14) is substituted into a power series expansion for the original Hamiltonian H:

$$(15) \qquad K(X_m, Y_m) = H(x_m, y_m)$$

$$= H_0(x_m, y_m) + \epsilon\, H_1(x_m, y_m) + \epsilon^2\, H_2(x_m, y_m) + \cdots$$

$$(16) \qquad H_0(x_m, y_m) = H_0(X_m + \frac{\partial W_1}{\partial Y_m}\epsilon + \cdots, Y_m - \frac{\partial W_1}{\partial X_m}\epsilon - \cdots)$$

$$= H_0\Big|_{\epsilon=0} + \frac{dH_0}{d\epsilon}\Big|_{\epsilon=0}\epsilon + \frac{d^2H_0}{d\epsilon^2}\Big|_{\epsilon=0}\frac{\epsilon^2}{2} + \cdots$$

$$(17) \qquad H_0\Big|_{\epsilon=0} = H_0(X_m, Y_m)$$

$$(18) \qquad \frac{dH_0}{d\epsilon}\Big|_{\epsilon=0} = \sum_j \frac{\partial H_0}{\partial x_j}\frac{dx_j}{d\epsilon} + \frac{\partial H_0}{\partial y_j}\frac{dy_j}{d\epsilon}\Big|_{\epsilon=0}$$

$$= \sum_j \frac{\partial H_0}{\partial x_j}\frac{\partial W}{\partial y_j} - \frac{\partial H_0}{\partial y_j}\frac{\partial W}{\partial x_j}\Big|_{\epsilon=0}$$

$$= \sum_j \frac{\partial H_0}{\partial X_j}\frac{\partial W_1}{\partial Y_j} - \frac{\partial H_0}{\partial Y_j}\frac{\partial W_1}{\partial X_j} = \{H_0, W_1\}$$

where the generating eqs.(5)-(7) have been used. This gives

$$(19)\qquad H_0(x_m,y_m) = H_0(X_m,Y_m) + \{H_0,W_1\}\,\epsilon + [\{\{H_0,W_1\},W_1\} + \{H_0,W_2\}]\,\frac{\epsilon^2}{2} + \cdots$$

This equation, which represents the expansion of H_0 under the near-identity transformation (13),(14), also holds for any of the H_n's, and in fact is valid for any function $f(x_m,y_m)$. Substitution of (19) and the corresponding eqs. on the other $H_n(x_m,y_m)$ into eq.(15) gives, after some simplification:

$$(20)\qquad K(X_m,Y_m) = K_0(X_m,Y_m) + K_1(X_m,Y_m)\,\epsilon + K_2(X_m,Y_m)\,\epsilon^2 + \cdots$$

where

$$(21)\qquad K_0 = H_0$$

$$(22)\qquad K_1 = H_1 + \{H_0,W_1\}$$

$$(23)\qquad K_2 = H_2 + \tfrac{1}{2}\{H_0,W_2\} + \tfrac{1}{2}\{K_1,W_1\} + \tfrac{1}{2}\{H_1,W_1\}$$

$$(24)\qquad K_3 = H_3 + \tfrac{1}{3}\{H_0,W_3\} + \tfrac{1}{3}\{K_1,W_2\} + \tfrac{1}{3}\{K_2,W_1\} + \tfrac{1}{6}\{H_1,W_2\} + \tfrac{2}{3}\{H_2,W_1\} + \tfrac{1}{6}\{\{H_1,W_1\},W_1\}$$

$$(25)\qquad K_4 = H_4 + \tfrac{1}{4}\{H_0,W_4\} + \tfrac{1}{4}\{K_1,W_3\} + \tfrac{1}{4}\{K_2,W_2\} + \tfrac{1}{4}\{K_3,W_1\} + \tfrac{1}{12}\{H_1,W_3\} + \tfrac{1}{4}\{H_2,W_2\} + \tfrac{3}{4}\{H_3,W_1\} + \tfrac{1}{12}\{\{H_1,W_1\},W_2\} + \tfrac{1}{24}\{\{H_1,W_2\},W_1\} + \tfrac{1}{4}\{\{H_2,W_1\},W_1\} + \tfrac{1}{24}\{\{\{H_1,W_1\},W_1\},W_1\}$$

In eqs.(21)-(25), the H_n and W_n are taken as functions of the variables X_m, Y_m.

So we see that the method of Lie transforms is nothing more than the introduction of the generating equations (5)-(7) into Taylor series expansions for the variables (x_m, y_m) and H. However, the transformation eqs. (e.g. (21)-(25)) can be generated much more efficiently than by the foregoing expansion method. There are several schemes for doing so (including the original method of Deprit [8] based on the "Lie triangle" and a method of Dragt and Finn [10] based on infinite products rather than infinite series), but we prefer the following method (see [15]), which is easily implemented on MACSYMA ([13], [14], [19]).

Define the operators L_n and S_n as follows:

$$L_n = \{\ , W_n\} \tag{26}$$

$$S_0 = \mathrm{Id}\ \text{(the identity operator)} \tag{27.1}$$

$$S_n = \frac{1}{n} \sum_{m=0}^{n-1} L_{n-m}\, S_m\ ,\ n = 1,2,3,\ldots \tag{27.2}$$

Then the near-identity transformation from (x_m, y_m) to (X_m, Y_m) variables is given by

$$x_m = \left[S_0 + \epsilon\, S_1 + \epsilon^2\, S_2 + \cdots\right] X_m \tag{28.1}$$

$$y_m = \left[S_0 + \epsilon\, S_1 + \epsilon^2\, S_2 + \cdots\right] Y_m \tag{28.2}$$

and the n^{th} term K_n of the Kamiltonian is given by the expression

$$(29) \qquad K_n = H_n + \frac{1}{n}\{H_0, W_n\} + \frac{1}{n}\sum_{m=1}^{n-1}\left[L_{n-m}\, K_m + m\, S_{n-m}\, H_m\right] ,$$

$$n = 2,3,4,\ldots$$

where the cases n = 0,1 are given by eqs.(21),(22).

Coupled Oscillators

In this work we shall apply the method of Lie transforms to two degree of freedom Hamiltonian systems in which H_0 has the special form:

$$(30) \qquad H_0 = \frac{1}{2}\left(p_1^{\,2} + \omega_1^{\,2}\, q_1^{\,2}\right) - \frac{1}{2}\left(p_2^{\,2} + \omega_2^{\,2}\, q_2^{\,2}\right)$$

where q_m and p_m are variables representing the displacement and momentum of oscillator m. For $\epsilon = 0$, the equations of motion corresponding to such a Hamiltonian become

$$(31) \qquad \dot{q}_m = p_m \text{ and } \dot{p}_m = -\omega_m^{\,2}\, q_m \text{ , or } \ddot{q}_m + \omega_m^{\,2}\, q_m = 0.$$

Thus when $\epsilon = 0$, the system has eigenvalues $\pm i\,\omega_1$, $\pm i\,\omega_2$, where $i = \sqrt{-1}$, and we change variables to eigencoordinates (x_m, y_m).

$$(32) \qquad q_m = \frac{x_m}{\omega_m} + i\,\frac{y_m}{2} \quad , \quad p_m = \frac{\omega_m\, y_m}{2} + i\, x_m$$

for which the eqs. of motion (31) and Hamiltonian (30) take the form

$$(33) \qquad \dot{x}_m = i\,\omega_m\, x_m \quad \text{and} \quad \dot{y}_m = -\,i\,\omega_m\, y_m \ ,$$

(34) $$H_0 = i\,\omega_1\,x_1\,y_1 - i\,\omega_2\,x_2\,y_2$$

In these coordinates, each H_n becomes a polynomial of degree n+2 in the four variables x_1, y_1, x_2, y_2. For example, there are 20 cubic monomials which form a basis for H_1:

(35) H_1 = linear combination of $\{x_1^3, x_1^2x_2, x_1^2y_1, x_1^2y_2, x_1x_2^2,$

$$x_1x_2y_1, x_1x_2y_2, x_1y_1^2, x_1y_1y_2, x_1y_2^2, x_2^3, x_2^2y_1, x_2^2y_2,$$

$$x_2y_1^2, x_2y_1y_2, x_2y_2^2, y_1^3, y_1^2y_2, y_1y_2^2, y_2^3\}$$

The number of basis monomials for H_2, H_3 and H_4 are:

Term	Degree	No.of basis monomials
H_1	3	20
H_2	4	35
H_3	5	56
H_4	6	84

We now come to the question of how to choose the generating functions W_n so as to best simplifiy the Kamiltonians K_n. At the n^{th} step of the method, K_n is given by eq.(29),

(36) $$K_n = \frac{1}{n}\{H_0, W_n\} + \text{terms which are already known}$$

Now with H_0 in the simplified form (34),

$$(37)\qquad \{H_0, W_n\} = \frac{\partial H_0}{\partial X_1}\frac{\partial W_n}{\partial Y_1} - \frac{\partial H_0}{\partial Y_1}\frac{\partial W_n}{\partial X_1} + \frac{\partial H_0}{\partial X_2}\frac{\partial W_n}{\partial Y_2} - \frac{\partial H_0}{\partial Y_2}\frac{\partial W_n}{\partial X_2}$$

$$= i\,\omega_1\left[Y_1\frac{\partial W_n}{\partial Y_1} - X_1\frac{\partial W_n}{\partial X_1}\right] - i\,\omega_2\left[Y_2\frac{\partial W_n}{\partial Y_2} - X_2\frac{\partial W_n}{\partial X_2}\right]$$

We want to choose W_n so that this linear partial differential operator on W_n cancels as many terms as possible in eq.(36). Each term to be cancelled will be of the form

$$(38)\qquad A\,X_1^{\,j}\,Y_1^{\,l}\,X_2^{\,r}\,Y_2^{\,s}$$

where A is a constant. In view of the linearity of (37), we choose W_n to be a sum of terms, one for each term (38) to be cancelled, of the form

$$(39)\qquad W_n = B\,X_1^{\,j}\,Y_1^{\,l}\,X_2^{\,r}\,Y_2^{\,s}$$

where B is an undetermined constant. Then

$$(40)\qquad \frac{1}{n}\{H_0, W_n\} = \frac{1}{n}\left[i\,\omega_1(l-j) - i\,\omega_2(s-r)\right] B\,X_1^{\,j}\,Y_1^{\,l}\,X_2^{\,r}\,Y_2^{\,s}$$

leading to the choice

$$(41)\qquad B = \frac{i\,A\,n}{\omega_1(l-j) - \omega_2(s-r)}\ ,\quad n = j+l+r+s-2.$$

Note that this scheme fails if the denominator of (41) vanishes. Assuming that the frequencies ω_1 and ω_2 are incommensurable (nonresonant), the denominator will vanish only if both

$$(42)\qquad l = j \text{ and } s = r\ .$$

Thus we cannot remove terms of the form

$$(X_1\ Y_1)^j\ (X_2\ Y_2)^r \quad . \tag{43}$$

This means that we can always reduce every such (nonresonant) problem to the form:

$$K_0 = H_0 = i\ \omega_1\ (X_1Y_1) - i\ \omega_2\ (X_2Y_2) \tag{44}$$

$$K_1 = 0 \tag{45}$$

$$K_2 = K_{2200}\ (X_1Y_1)^2 + K_{1111}\ (X_1Y_1)(X_2Y_2) + K_{0022}\ (X_2Y_2)^2 \tag{46}$$

$$K_3 = 0 \tag{47}$$

$$K_4 = K_{3300}\ (X_1Y_1)^3 + K_{2211}\ (X_1Y_1)^2(X_2Y_2) + K_{1122}\ (X_1Y_1)(X_2Y_2)^2 + K_{0033}\ (X_2Y_2)^3 \tag{48}$$

That is, every such nonresonant two degree of freedom problem can, to O(4), be reduced to only 7 coefficients. Note that in this case the resulting Hamiltonian is a function only of the "action" variables,

$$I_1 = i\ X_1\ Y_1 \quad \text{and} \quad I_2 = i\ X_2\ Y_2 \tag{49}$$

and hence both coordinates are ignorable and the system is immediately solveable to O(4). Such a system is said to be in Birkhoff normal form ([5], p.85).

By inspection of eq.(41), the foregoing scheme fails at special <u>resonant</u> values of ω_1 amd ω_2. In solving for W_n, resonant terms occur for integer values of k_1 and k_2 such that

(50) $$k_1 \omega_1 + k_2 \omega_2 = 0 \ , \ |k_1| + |k_2| \leq n + 2$$

In such cases additional non-removable terms occur. We shall not consider such resonant cases in this work.

Computer Algebra

The computation just described turns out to involve vast quantities of algebra. We used the computer algebra system MACSYMA ([18]) in order to do the computation more accurately and more efficiently than by hand. For example, the key formulas (12),(26),(27),(29) can be represented in MACSYMA via the following lines of code ([7], [19]):

```
POISSON(F,G):=
SUM(DIFF(F,X[I])*DIFF(G,Y[I])-DIFF(F,Y[I])*DIFF(G,X[I]),I,1,N)$

L(I,F):=POISSON(F,W[I])$

S(I,F):=(IF I=0 THEN F ELSE SUM(L(I-M,S(M,F)),M,0,I-1)/I)$

K[I]:=(H[I]+POISSON(H[0],W[I])/I
          +SUM(L(I-M,K[M])+M*S(I-M,H[M]),M,1,I-1)/I)$
```

In order to efficiently compute W_n by the formulas (39),(41), we use the MACSYMA tool called pattern matching. A rule named WSOLVE is defined as follows:

```
LET(X1^J*Y1^L*X2^R*Y2^S,
    X1^J*Y1^L*X2^R*Y2^S*%I*N/(W1*(L-J)-W2*(S-R)),WSOLVE)$
```

That is, replace the term $X_1^{j} Y_1^{l} X_2^{r} Y_2^{s}$ by $\dfrac{i\, n\, X_1^{j} Y_1^{l} X_2^{r} Y_2^{s}}{\omega_1(l-j)-\omega_2(s-r)}$.

When WSOLVE is applied to the "terms which are already known" on

the right hand side of eq.(36), the correct expression for W_n is automatically generated. Note that this rule is not applied to non-removable terms of the form (43).

One could hope to simply apply these formulas to the problem at hand, and to thereby automatically obtain the transformed Kamiltonian. Unfortunately, the size of the O(4) computation is too large to proceed directly; MACSYMA on a Symbolics 3670 runs out of space. E.g. from eq.(25) we see that the computation of K_4 involves the evaluation of the quantity $\{\{\{H_1,W_1\},W_1\},W_1\}$. The innermost Poisson bracket involves 20 terms for H_1 and 20 terms for W_1, i.e. 400 pairs which can be collected together into 35 terms (since there are 35 fourth degree basis monomials). These then need to be combined with the 20 terms of W_1 in order to evaluate the second Poisson bracket, i.e. 700 pairs which combine together into 56 terms. Next the third Poisson bracket combines the previous result with the 20 terms of W_1 to require the computation of 1120 pairs, which may be collected together into 84 terms.

In order to complete the computation, we broke it up into pieces, each of which was sufficiently small so as not to cause MACSYMA to encounter space problems. We shall refer to our strategy for treating such large computations as the method of telescoping compositions. As an example of this strategy, we once again consider the computation of the triple Poisson bracket $\{\{\{H_1,W_1\},W_1\},W_1\}$. We first compute $\{H_1,W_1\}$ and store the resulting 35 coefficients A_{jlrs} in a disk file. Next, instead of computing $\{\{H_1,W_1\},W_1\}$, we compute instead $\{A,W_1\}$, where A is a dummy polynomial with symbolic coefficients A_{jlrs}. Although we are eventually interested in identifying these coefficients with those we have stored in a disk file, we save that step for later. We store the resulting 56 coefficients B_{jlrs} of $\{A,W_1\}$ in a disk file. Next we compute $\{B,W_1\}$, where

now B is a dummy polynomial with symbolic coeffcients B_{jlrs}. This results in 84 coefficients which are known in terms of the B_{jlrs} coefficients. The latter are stored in a file and are known in terms of the A_{jlrs} coefficients, which are also stored in a disk file. At this point the computation of $\{\{\{H_1,W_1\},W_1\},W_1\}$ is complete, although it still remains to plug the values of the A_{jlrs} and B_{jlrs} coefficients into the final result.

For a complete listing of the programs, see [7].

Results

The results of this work take the form of expressions for the transformed Kamiltonian K in terms of the original Hamiltonian H. If we express H in x_m,y_m eigencoordinates defined by eqs.(32), then H_0 takes the canonical form (34), and the polynomials H_n of eq.(4) can be written as

$$H_n = \sum H_{jlrs}\, x_1^{\,j} y_1^{\,l} x_2^{\,r} y_2^{\,s} \quad , \quad n = j+l+r+s-2$$

where the H_{jlrs} are given constants. Then the coefficients K_{jlrs} in K_2 in eq.(46) are given by:

$$(51) \qquad K_{2200} = H_{2200}$$

$$+ i \left[\frac{-1}{\omega_2} H_{1101}\, H_{1110} + \frac{3}{\omega_1} \{H_{1200}\, H_{2100} + H_{0300}\, H_{3000}\} \right.$$

$$\left. - \frac{1}{2\omega_1+\omega_2} H_{0210}\, H_{2001} + \frac{1}{2\omega_1-\omega_2} H_{0201}\, H_{2010} \right]$$

$$(52)\qquad K_{1111} = H_{1111}$$

$$+ i \left[\frac{2}{\omega_1} \{H_{1011}\, H_{1200} + H_{0111}\, H_{2100}\} - \frac{4}{\omega_1+2\omega_2} H_{0120}\, H_{1002} \right.$$

$$- \frac{2}{\omega_2} \{H_{1101}\, H_{0021} + H_{1110}\, H_{0012}\} + \frac{4}{\omega_2+2\omega_1} H_{0210}\, H_{2001}$$

$$\left. - \frac{4}{2\omega_2-\omega_1} H_{0102}\, H_{1020} + \frac{4}{2\omega_1-\omega_2} H_{0201}\, H_{2010} \right]$$

$$(53)\qquad K_{0022} = H_{0022}$$

$$+ i \left[\frac{1}{\omega_1} H_{0111}\, H_{1011} - \frac{3}{\omega_2} \{H_{0003}\, H_{0030} + H_{0012}\, H_{0021}\} \right.$$

$$\left. + \frac{1}{2\omega_2+\omega_1} H_{0120}\, H_{1002} - \frac{1}{2\omega_2-\omega_1} H_{0102}\, H_{1020} \right]$$

The comparable coefficients in K_4 in eq.(48) were also found, but cannot be displayed here because they are too long. E.g., the ASCII files for K_{3300} and K_{0033} contain 164K characters, while those for K_{1122} and K_{2211} contain 468K characters. These expressions simplify greatly, however, in the special case in which H_1 and H_3 are identically zero. Since this special case occurs in frequently in sample problems, we give the associated coefficients of K_4 here:

(54) $$K_{3300} = H_{3300} - \frac{i\, H_{0301}\, H_{3010}}{\omega_2 - 3\,\omega_1} - \frac{i\, H_{1201}\, H_{2110}}{\omega_2 - \omega_1} - \frac{i\, H_{1210}\, H_{2101}}{\omega_2 + \omega_1} - \frac{i\, H_{0310}\, H_{3001}}{\omega_2 + 3\,\omega_1} + \frac{4\, i\, H_{0400}\, H_{4000}}{\omega_1} + \frac{4\, i\, H_{1300}\, H_{3100}}{\omega_1}$$

(55) $$K_{2211} = H_{2211} - \frac{9\, i\, H_{0301}\, H_{3010}}{\omega_2 - 3\,\omega_1} - \frac{i\,(3\, H_{1201} + 2\, H_{0112})\, H_{2110}}{\omega_2 - \omega_1} + \frac{9\, i\, H_{0310}\, H_{3001}}{\omega_2 + 3\,\omega_1} + \frac{i\,(3\, H_{1210} - 2\, H_{0121})\, H_{2101}}{\omega_2 + \omega_1} - \frac{2\, i\, H_{0202}\, H_{2020}}{\omega_2 - \omega_1} - \frac{2\, i\, H_{1021}\, H_{1201}}{\omega_2 - \omega_1} - \frac{2\, i\, H_{0220}\, H_{2002}}{\omega_2 + \omega_1} - \frac{2\, i\, H_{1012}\, H_{1210}}{\omega_2 + \omega_1} - \frac{2\, i\, H_{1102}\, H_{1120}}{\omega_2} + \frac{3\, i\, H_{0211}\, H_{3100}}{\omega_1} + \frac{3\, i\, H_{1300}\, H_{2011}}{\omega_1}$$

(56) $$K_{1122} = H_{1122}$$

$$- \frac{9\, i\, H_{0103}\, H_{1030}}{3\,\omega_2 - \omega_1} - \frac{i\,(3\, H_{0112} + 2\, H_{1201})\, H_{1021}}{\omega_2 - \omega_1}$$

$$- \frac{9\, i\, H_{0130}\, H_{1003}}{3\omega_2 + \omega_1} - \frac{i\,(3\, H_{0121} - 2\, H_{1210})\, H_{1012}}{\omega_2 + \omega_1}$$

$$- \frac{2\, i\, H_{0202}\, H_{2020}}{\omega_2 - \omega_1} - \frac{2\, i\, H_{0112}\, H_{2110}}{\omega_2 - \omega_1} + \frac{2\, i\, H_{0220}\, H_{2002}}{\omega_2 + \omega_1}$$

$$+ \frac{2\, i\, H_{0121}\, H_{2101}}{\omega_2 + \omega_1} + \frac{2\, i\, H_{0211}\, H_{2011}}{\omega_1} - \frac{3\, i\, H_{0013}\, H_{1120}}{\omega_2}$$

$$- \frac{3\, i\, H_{0031}\, H_{1102}}{\omega_2}$$

(57) $$K_{0033} = H_{0033}$$

$$- \frac{i\, H_{0103}\, H_{1030}}{3\,\omega_2 - \omega_1} + \frac{i\, H_{0130}\, H_{1003}}{3\,\omega_2 + \omega_1} - \frac{i\, H_{0112}\, H_{1021}}{\omega_2 - \omega_1}$$

$$+ \frac{i\, H_{0121}\, H_{1012}}{\omega_2 + \omega_1} - \frac{4\, i\, H_{0004}\, H_{0040}}{\omega_2} - \frac{4\, i\, H_{0013}\, H_{0031}}{\omega_2}$$

Arnold's Theorem

We are interested in applying the previous results to the determination of the stability of the equilibrium at the origin in a system of two nonlinear coupled oscillators in which H_0 has the form (30). Note that the linearized Hamiltonian differential equations (1) corresponding to $H = H_0$ have purely imaginary eigenvalues, and thus are inconclusive regarding stability. Moreover, because of the minus sign in (30), H_0 is not positive definite, and Lyapunov's direct method [7] cannot be used to determine stability.

For such cases, stability may be determined by appealing to a theorem of Arnold [4], which has been restated and reproved by Meyer and Schmidt [16]. The theorem, based on the existence of invariant tori in KAM theory [3], gives sufficient conditions for stability in nonresonant systems, in terms of the transformed Hamiltonian $K(I_1,I_2)$ which has been put in Birkhoff normal form, cf.(49). The terms K_n of eqs.(44)-(48) are thought of as functions of I_1 and I_2, $K_n(I_1,I_2)$. The theorem involves quantities D_n defined by

$$D_n = K_n(\omega_2,\omega_1). \tag{58}$$

From (44)-(48), the first two non-identically zero D_n's are D_2 and D_4:

$$D_2 = -\left(K_{2200}\,\omega_2^{\,2} + K_{1111}\,\omega_1\omega_2 + K_{0022}\,\omega_1^{\,2}\right) \tag{59}$$

$$D_4 = i\left(K_{3300}\,\omega_2^{\,3} + K_{2211}\,\omega_1\omega_2^{\,2} + K_{1122}\,\omega_1^{\,2}\omega_2 + K_{0033}\,\omega_1^{\,3}\right) \tag{60}$$

Arnold's theorem states that the origin is stable for

those parameter values for which $D_2 \neq 0$. In the case that $D_2 = 0$, stability is assured if $D_4 \neq 0$, and so on. I.e., the origin is stable if $D_{2n} \neq 0$ for some n.

Using the expressions (51)-(57) for the coefficients K_{jlrs}, expressions for D_2 and D_4 (the latter in the special case that $H_1=H_3=0$) may be obtained:

$$(61)\quad D_2 = -(\omega_2^2 H_{2200} + \omega_1\omega_2 H_{1111} \omega_1^2 H_{0022})$$

$$+ i [\omega_2 H_{1101} H_{1110} - \omega_1 H_{1011} H_{0111}$$

$$+ 2\omega_1 (H_{1110} H_{0012} + H_{1101} H_{0021})$$

$$- 2\omega_2 (H_{0111} H_{2100} + H_{1011} H_{1200})$$

$$+ \frac{3\omega_1^2}{\omega_2} (H_{0003} H_{0030} + H_{0021} H_{0012})$$

$$- \frac{3\omega_2^2}{\omega_1} (H_{3000} H_{0300} + H_{2100} H_{1200})$$

$$+ \frac{4\omega_2 + \omega_1}{2\omega_2 - \omega_1} \omega_1 H_{1020} H_{0102} + \frac{4\omega_2 - \omega_1}{2\omega_2 + \omega_1} \omega_1 H_{1002} H_{0120}$$

$$- \frac{4\omega_1 + \omega_2}{2\omega_1 - \omega_2} \omega_2 H_{2010} H_{0201} - \frac{4\omega_1 - \omega_2}{2\omega_1 + \omega_2} \omega_2 H_{2001} H_{0210}]$$

$$(62)\quad D_4 = i (\omega_2^3 H_{3300} + \omega_2^2\omega_1 H_{2211} + \omega_2\omega_1^2 H_{1122} + \omega_1^3 H_{0033})$$

$$- 2\omega_1\omega_2 (H_{1102} H_{1120} + H_{0211} H_{2011} + H_{0202} H_{2020}$$

$$+ H_{1021}\, H_{1201} + H_{1012}\, H_{1210} + H_{0220}\, H_{2002}$$

$$+ H_{0112}\, H_{2110} + H_{2101}\, H_{0121})$$

$$- 3\,\omega_2^{\,2}\,(H_{3100}\, H_{0211} + H_{1300}\, H_{2011})$$

$$- 3\,\omega_1^{\,2}\,(H_{0013}\, H_{1120} + H_{0031}\, H_{1102})$$

$$- 4\,\frac{\omega_1^{\,3}}{\omega_2}\,(H_{0004}\, H_{0040} + H_{0013}\, H_{0031})$$

$$- 4\,\frac{\omega_2^{\,3}}{\omega_1}\,(H_{4000}\, H_{0400} + H_{1300}\, H_{3100})$$

$$- \frac{\omega_2^{\,2}\, H_{1201}\, H_{2110}\,(\omega_2 + 3\,\omega_1)}{\omega_2 + \omega_1} - \frac{\omega_2^{\,2}\, H_{1210}\, H_{2101}\,(\omega_2 - 3\,\omega_1)}{\omega_2 - \omega_1}$$

$$- \frac{\omega_1^{\,2}\, H_{1021}\, H_{0112}\,(\omega_1 + 3\,\omega_2)}{\omega_1 + \omega_2} - \frac{\omega_1^{\,2}\, H_{1012}\, H_{0121}\,(\omega_1 - 3\,\omega_2)}{\omega_1 - \omega_2}$$

$$- \frac{\omega_2^{\,2}\, H_{0301}\, H_{3010}\,(\omega_2 + 9\,\omega_1)}{\omega_2 + 3\,\omega_1} - \frac{\omega_2^{\,2}\, H_{0310}\, H_{3001}\,(\omega_2 - 9\,\omega_1)}{\omega_2 - 3\,\omega_1}$$

$$- \frac{\omega_1^{\,2}\, H_{0103}\, H_{1030}\,(\omega_1 + 9\,\omega_2)}{\omega_1 + 3\,\omega_2} - \frac{\omega_1^{\,2}\, H_{0130}\, H_{1003}\,(\omega_1 - 9\,\omega_2)}{\omega_1 - 3\,\omega_2}$$

(assumes $H_1 = H_3 = 0$)

The expression for D_4 in the general case is too long to be included here, but is available on our computer for numerical evaluation.

Example 1

We consider a variation of the Henon-Heiles Hamiltonian where the linear oscillators are not at low-order resonance and are of different signs:

$$(63) \qquad H = \frac{1}{2}\,(p_1^{\,2} + \omega^2\, q_1^{\,2}) - \frac{1}{2}\,(p_2^{\,2} + q_2^{\,2}) + q_1^{\,2} q_2 - \frac{1}{3}\, q_2^{\,3}$$

Using the transformation to eigencoordinates given by eq.(32), H becomes

$$(64) \qquad \begin{aligned} H = {} & i\,\omega\, x_1 y_1 - i\, x_2 y_2 - \frac{1}{3}\, x_2^{\,3} + \frac{i}{24}\, y_2^{\,3} + \frac{i}{2\,\omega^2}\, x_1^{\,2} y_2 \\ & + \frac{1}{\omega^2}\, x_1^{\,2} x_2 - \frac{i}{2}\, x_2^{\,2} y_2 - \frac{1}{4}\, y_1^{\,2} x_2 - \frac{i}{8}\, y_1^{\,2} y_2 + \frac{1}{4}\, x_2 y_2^{\,2} \\ & + \frac{i}{\omega}\, x_1 x_2 y_1 - \frac{1}{2\,\omega}\, x_1 y_1 y_2 \quad , \quad \omega > 0 \end{aligned}$$

Then using eqs.(51)-(53), we find the K_2 coefficients to be

$$(65) \qquad K_{2200} = \frac{3 - 8\,\omega^2}{4\,\omega^2\,(4\,\omega^2 - 1)}$$

$$(66) \qquad K_{1111} = \frac{4\,\omega^2 + 1}{\omega\,(4\,\omega^2 - 1)}$$

$$(67) \qquad K_{0022} = -\frac{5}{12}$$

Using eq.(61), we find D_2 to be

$$(68) \qquad D_2 = \frac{20\,\omega^6 - 53\,\omega^4 + 12\,\omega^2 - 9}{12\,\omega^2\,(4\,\omega^2 - 1)}$$

We find that $D_2=0$ only for $\omega = \omega_c \simeq 1.5752078\ldots$ In order to determine the stability of the origin for $\omega = \omega_c$, we must

consider the D_4 condition. Because H contains only cubic nonlinear terms and because each cubic coefficient is simple, we are able to find the expression for D_4 algebraically. The coefficients for K_4 turn out to be

$$(69)\quad K_{3300} = \frac{i\,(1024\,\omega^8 + 768\,\omega^6 - 1632\,\omega^4 + 596\,\omega^2 - 51)}{48\,\omega^5\,(2\,\omega - 1)^3\,(2\,\omega + 1)^3}$$

$$(70)\quad K_{2211} = \frac{i\,(384\,\omega^{10} - 288\,\omega^8 + 16\,\omega^6 - 340\,\omega^4 + 159\,\omega^2 - 6)}{4\,\omega^4\,(1 - \omega^2)\,(2\,\omega - 1)^3\,(2\,\omega + 1)^3}$$

$$(71)\quad K_{1122} = \frac{i\,(4\,\omega^2 + 1)(320\,\omega^8 - 480\,\omega^6 + 360\,\omega^4 - 161\,\omega^2 + 6)}{12\,\omega^3\,(\omega^2 - 1)\,(2\,\omega - 1)^3\,(2\,\omega + 1)^3}$$

$$(72)\quad K_{0033} = -\frac{235\,i}{432}$$

and D_4 becomes

$$(73)\quad D_4 = \frac{(15040\,\omega^{16} - 72400\,\omega^{14} + 113172\,\omega^{12} - 77935\,\omega^{10} + 14491\,\omega^8 - 10188\,\omega^6 - 3096\,\omega^4 + 5175\,\omega^2 - 459)}{432\,\omega^5\,(\omega^2 - 1)\,(2\,\omega - 1)^3\,(2\,\omega + 1)^3}$$

So, at $\omega = \omega_c$, $D_4 = -0.19180289\ldots \neq 0$.

Thus by Arnold's theorem, the origin is nonlinearly stable. We note that this result does <u>not</u> apply to a small set of resonant values of ω which correspond to vanishing denominators in the algorithm (41). From eq.(50) with n = 2, we find the following resonant values of ω:

$$\omega = \{\tfrac{1}{3},\ \tfrac{1}{2},\ 1,\ 2,\ 3\}.$$

Example 2

This second example involves a spinning mass-spring system, which contains no odd powered terms in the Hamiltonian. Consider 4 identical springs, each attached at one end to the outer rim of a wheel of unit radius separated by 90°. The other end of each spring is attached to a unit mass which is free to move about its equilibrium position at the center, see Fig. 1. Let the Q_1-Q_2 axes rotate with the wheel with angular velocity $\omega > 0$ relative to an inertial frame. Each spring is unstretched when the mass is at the origin. The potential energy V_i for each spring under a deflection δ_i is taken to be

$$(74) \qquad V_i = \frac{1}{4} (\delta_i^2 + \mu \delta_i^4)$$

where the linear spring constant has been taken equal to $\frac{1}{2}$ and μ is a nonlinear spring constant. Then this system has the Hamiltonian

$$(75) \qquad H = \frac{1}{2} (P_1^2 + P_2^2) + \omega (P_1Q_2 - P_2Q_1) + V_1 + V_2 + V_3 + V_4$$

where P_i are momenta. Then upon taking the Taylor series of H about the origin, H becomes [7]

$$\begin{aligned} (76) \qquad H = {} & \frac{1}{2} (P_1^2 + P_2^2) + \omega (P_1Q_2 - P_2Q_1) + \frac{1}{2} (Q_1^2 + Q_2^2) \\ & + \frac{1}{8} [(4\mu + 1) Q_1^4 - 8 Q_1^2Q_2^2 + (4\mu + 1) Q_2^4] \\ & - \frac{1}{16} [Q_1^6 + 4 (\mu - 1) Q_1^4Q_2^2 + 4 (\mu - 1) Q_1^2Q_2^4 + Q_2^6] \\ & + \ldots \\ & = H_0 + H_2 + H_4 + \ldots \end{aligned}$$

Using the linear differential equations corresponding to H_0, we

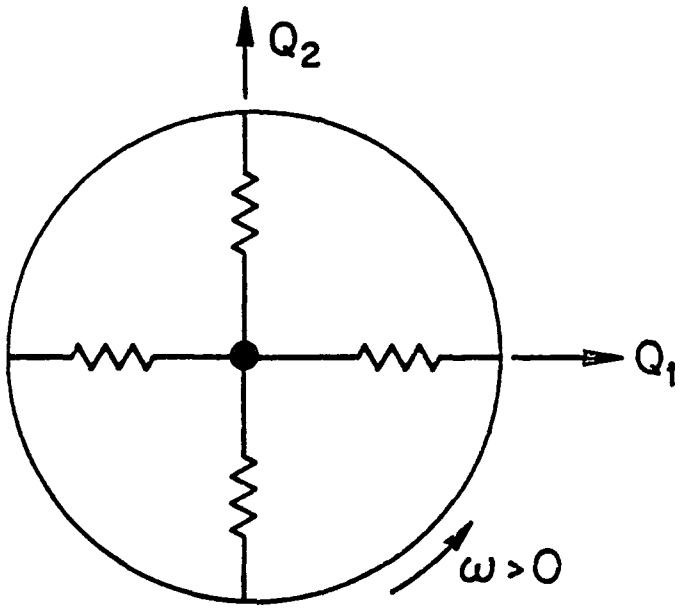


Fig. 1. Example 2 involves a spinning mass-spring system. The unit mass is restrained by 4 identical nonlinear springs. The Q_1-Q_2 axes are fixed to the wheel and rotate relative to an inertial frame.

find the characteristic equation to be

$$(77) \qquad \lambda^4 + 2\lambda^2 (1 + \omega^2) + (\omega^2 - 1)^2 = 0$$

which has eigenvalues $\lambda = \{\pm i (1-\omega), \pm i (1+\omega)\}$. From this we conclude that the equilibrium at the origin is elliptic for $\omega \neq 1$, i.e., comprised of two oscillators with frequencies $1 - \omega$ and $1 + \omega$ in the first approximation. Then using a canonical eigenvector transformation from (Q_m, P_m) to (x_m, y_m) gives [7]

$$(78) \qquad H_0 = i (1 - \omega) x_1 y_1 - i (1 + \omega) x_2 y_2$$

which is in the proper form for our analysis. After similarly transforming H_2 and H_4, we use eqs.(51)-(57) to find that

$$(79) \qquad K_{2200} = K_{0022} = \frac{1 - 12\mu}{32} \quad , \quad K_{1111} = \frac{1 - 12\mu}{8}$$

$$(80) \qquad K_{3300} = \frac{i [(576\mu^2 - 32\mu + 20)\omega^2 - (864\mu^2 - 48\mu + 30)\omega + 272\mu^2 - 56\mu - 15]}{1024 (\omega - 1)(2\omega - 1)}$$

$$(81) \qquad K_{2211} = \frac{3i [(1440\mu^2 - 48\mu + 58)\omega^2 - (720\mu^2 - 24\mu + 29)\omega - 48\mu^2 - 120\mu - 75]}{1024 \omega (2\omega - 1)}$$

$$(82) \qquad K_{1122} = \frac{3i [(1440\mu^2 - 48\mu + 58)\omega^2 + (720\mu^2 - 24\mu + 29)\omega - 48\mu^2 - 120\mu - 75]}{1024 \omega (2\omega + 1)}$$

$$(83) \qquad K_{0033} = \frac{i [(576\mu^2 - 32\mu + 20)\omega^2 + (864\mu^2 - 48\mu + 30)\omega + 272\mu^2 - 56\mu - 15]}{1024 (\omega + 1)(2\omega + 1)}$$

We find D_2 and D_4 using eqs.(61)-(62) to be

$$(84) \qquad D_2 = \frac{(12\,\mu - 1)\,(3\,\omega^2 - 1)}{16}$$

$$(85) \qquad D_4 = \frac{\left[(9792\,\mu^2 - 352\,\mu + 388)\,\omega^8 - (17744\,\mu^2 + 264\,\mu + 1213)\,\omega^6 + (9104\,\mu^2 + 232\,\mu + 657)\,\omega^4 - (1392\,\mu^2 + 216\,\mu + 207)\,\omega^2 - 144\,\mu^2 - 360\,\mu - 225\right]}{512\,\omega\,(\omega^2 - 1)\,(4\,\omega^2 - 1)}$$

Then $D_2 = 0$ for $\mu = \frac{1}{12}$ and $\omega^2 = \frac{1}{3}$, which are two lines in the μ-ω parameter plane. When $D_2 = 0$ we must check the D_4 condition. Consider the line $\mu = \frac{1}{12}$. The value of D_4 on this line is

$$(86) \qquad D_4(\mu = \tfrac{1}{12}) = \frac{60\,\omega^8 - 191\,\omega^6 + 104\,\omega^4 - 33\,\omega^2 - 36}{72\,\omega\,(\omega^2 - 1)\,(4\,\omega^2 - 1)}$$

which is zero only for $\omega = \omega_c \simeq 1.6241875\ldots$ Now consider the line $\omega^2 = \frac{1}{3}$. D_4 on this line becomes

$$(87) \qquad D_4(\omega^2 = \tfrac{1}{3}) = \frac{\sqrt{3}}{288}\,(336\,\mu^2 + 1064\,\mu + 661)$$

which is zero only for $\mu = \mu_{1,2} = -\frac{1}{84}\,(133 \pm 4\sqrt{238}) \simeq \{-0.84870,\ -2.31796\}$.

We now apply the stability theorem. First, note that we consider $\omega > 0$ and that for $\omega = 1$ the origin is not elliptic so that our analysis does not apply there. From eq.(50) with $n = 2$, we must also exclude $\omega = \{\frac{1}{3},\ \frac{1}{2},\ 2,\ 3\}$ from the analysis. Applying the D_2 condition, we find that the origin is stable

everywhere in the μ-ω parameter plane except possibly along the two lines $\mu = \frac{1}{12}$ and $\omega^2 = \frac{1}{3}$. On these lines the D_4 condition must be used. From eq.(50) with n = 4, we must now exclude $\omega = \{\frac{1}{5}, \frac{3}{5}, \frac{2}{3}, \frac{3}{2}, \frac{5}{3}, 5\}$ when $\mu = \frac{1}{12}$. Elsewhere on $\mu = \frac{1}{12}$, the origin is stable provided $\omega \neq \omega_c$; on $\omega^2 = \frac{1}{3}$, the origin is stable provided $\mu \neq \mu_{1,2}$. For the three points where $D_2 = D_4 = 0$, the D_6 condition must be used to prove nonlinear stability.

We note that for $\omega < 1$, stability of the origin can be independently proved by Lyapunov's direct method [7].

Application to the Problem of Three Bodies

The circular restricted three body problem is well-known to exhibit five equilibria in a rotating barycentric coordinate system [20]. L_1, L_2 and L_3 represent equilibrium positions of the third body, in which all three bodies are collinear. All three of these are unstable for all values of the mass ratio parameter μ. L_4 and L_5 represent equilibria where all three bodies sit at the vertices of an equilateral triangle. For values of $\mu > \mu_1 \simeq 0.0385208$, both these equilibria are unstable. For $\mu < \mu_1$, Alfriend [1,2] showed that the triangular points are unstable when $\mu = \mu_2$ and μ_3, special mass ratios which cause the linearized frequencies to be in the ratio of 1:2 and 1:3, respectively. For other values of $\mu < \mu_1$, stability of L_4 (and L_5) can be obtained by using Arnold's theorem. This was first done by Deprit and Deprit-Bartholome [9], who calculated D_2 by hand. The value they obtained,

$$D_2 = -\frac{36 - 541\,\omega_1^2\omega_2^2 + 644\,\omega_1^4\omega_2^4}{8\,(1 - 4\,\omega_1^2\omega_2^2)\,(4 - 25\,\omega_1^2\omega_2^2)} \tag{88}$$

is non-zero for all values of μ except for $\mu = \mu_c \simeq 0.0109136$. For $\mu = \mu_c$, Arnold's theorem requires the quantity D_4 be found. This computation was performed by Meyer and Schmidt, who found $D_4 \simeq -66.6$. The non-zero value of D_4 implies stability, by Arnold's theorem.

In what follows we shall apply the results obtained in this paper to confirm the previous computations of Deprit and Deprit-Bartholome [9] and Meyer and Schmidt [16].

The Hamiltonian for the circular restricted three-body problem about the equilibrium L_4 is :

$$H = \frac{1}{2}(P_1^2 + P_2^2) + P_1Q_2 - P_2Q_1 - \frac{(1-2\mu)}{2} Q_1 - \frac{\sqrt{3}}{2} Q_2 - \left[\frac{\mu}{\rho_1} + \frac{1-\mu}{\rho_2} \right] \tag{89}$$

where $$\rho_1^2 = Q_1^2 + Q_2^2 - Q_1 + \sqrt{3}\, Q_2 + 1$$

$$\rho_2^2 = Q_1^2 + Q_2^2 + Q_1 + \sqrt{3}\, Q_2 + 1$$

Expanding in a Taylor series about the origin, H becomes ΣH_n where H_n contains terms of order n+2 and H_0 is given by:

$$H_0 = \frac{1}{2}(P_1^2 + P_2^2) + P_1Q_2 - P_2Q_1 + \frac{1}{8} Q_1^2 - \frac{5}{8} Q_2^2 - \frac{3\sqrt{3}}{4} (1 - 2\mu)\, Q_1Q_2 \tag{90}$$

Then using the linearized differential equations corresponding to H_0, the characteristic equation for the system is found to be:

(91) $\lambda^4 + \lambda^2 + \frac{27}{16}(1-\tau^2) = 0$ where $\tau = 1 - 2\mu$

The eigenvalues λ have positive real parts for $\mu > \mu_1 = \frac{1}{2}(1 - \frac{\sqrt{69}}{9})$ implying the system is linearly unstable. For $\mu < \mu_1$, the system is critically stable having eigenvalues $\pm i\omega_1$ and $\pm i\omega_2$ where:

$$0 < \omega_2 < \frac{\sqrt{2}}{2} < \omega_1 < 1 \ , \ \omega_1^2 + \omega_2^2 = 1 \ ,$$

and $$\omega_1^2\omega_2^2 = \frac{27}{16}(1-\tau^2)$$

Using a canonical linear transformation (see [6]), $H_0(Q_m, P_m)$ is transformed into $H_0(q_m, p_m)$ which is of the form (30). Then following eqs.(30)-(34) we introduce the variables (x_m, y_m) and find the following components of K_2:

(92) $$K_{2200} = \frac{-\omega_2^2(124\,\omega_1^4 - 696\,\omega_1^2 + 81)}{144\,(2\omega_1^2 - 1)^2\,(\omega_2^2 - 4\omega_1^2)}$$

(93) $$K_{1111} = \frac{\omega_1\,\omega_2\,(64\,\omega_1^4 - 64\,\omega_1^2 - 43)}{6\,(2\omega_1^2 - 1)^2\,(\omega_2^2 - 4\omega_1^2)\,(\omega_1^2 - 4\omega_2^2)}$$

(94) $$K_{0022} = \frac{\omega_1^2(124\,\omega_1^4 + 448\,\omega_1^2 - 491)}{144\,(2\omega_1^2 - 1)^2\,(4\omega_2^2 - \omega_1^2)}$$

and the first stability condition becomes:

$$D_2 = \frac{-644\,\omega_1^8 + 1288\,\omega_1^6 - 1185\,\omega_1^4 + 541\,\omega_1^2 - 36}{8\,(2\omega_1^2 - 1)^2\,(\omega_2 - 2\omega_1)\,(\omega_2 + 2\omega_1)\,(2\omega_2 - \omega_1)\,(2\omega_2 + \omega_1)}$$

which is equivalent to the expression (88) found by Deprit and Deprit-Bartholome [9]. Then, on $0 < \mu < \mu_1$, $D_2 = 0$ only for:

$$\mu = \mu_c = \frac{3\sqrt{483} - \sqrt{2\sqrt{199945} + 3265}}{6\sqrt{483}} \simeq 0.0109136$$

At this value of $\mu = \mu_c$, the components of K_4 become:

(95) $K_{3300} \simeq 0.219259187\ i + 6.52\ E\text{-}37$

(96) $K_{2211} \simeq -7.79324843\ i + 3.74\ E\text{-}35$

(97) $K_{1122} \simeq 209.933620\ i + 2.35\ E\text{-}34$

(98) $K_{0033} \simeq 14.5264460\ i + 1.75\ E\text{-}34$

and D_4 becomes:

(99) $D_4 \simeq -66.6 - 4.27\ E\text{-}36\ i$

The very small real part of each K_{jlrs} and imaginary part of D_4 results from taking only a finite number of digits (40 in fact) in the numerical approximation. Because the real part is so much larger than the error term, the approximation $D_4 \simeq -66.6$ is accurate and $D_4 \neq 0$. Hence, at $\mu = \mu_c$, the Hamiltonian system is stable. These values for the coefficients of K_4 and D_4 agree with those obtained by Meyer and Schmidt [16].

Acknowledgement

The authors would like to thank Dr. Dieter Armbruster for his helpful advice with MACSYMA. This work was partially supported by grants from the National Science Foundation, AFOSR, and ARO through the Mathematical Sciences Institute at Cornell.

References

1. Alfriend, K.T. "The Stability of the Triangular Lagrangian points for commensurability of order two", *Celestial Mechanics*, Vol. 1, 1970, pgs. 351-359.
2. Alfriend, K.T. "The Stability of the Triangular Lagrangian points for commensurability of order three", *Celestial Mechanics*, Vol. 4, 1971, pgs. 60-77.
3. Arnold, V.I. Mathematical Methods of Classical Mechanics, Graduate Texts in Mathematics, Vol. 60, Springer-Verlag, NY, 1978.
4. Arnold, V.I. "The Stability of the Equilibrium Positions of a Hamiltonian System of Ordinary Differential Equations in the General Ellitptic Case", *Soviet Math Doklady*, Vol. 2, 1961, pgs. 247-249.
5. Birkhoff, G. Dynamical Systems, American Mathematical Society, Providence, RI, 1966.
6. Breakwell, J. and Pringle, R. "Resonances affecting motion near the Earth-Moon equilateral libration points" in Progress in Astronautics and Aeronautics: Methods in Astrodynamics and Celestial Mechanics, Vol. 17, Academic Press, NY, 1966.
7. Coppola, Vincent T. M.S. Thesis, Cornell University, 1988.
8. Deprit, A. "Canonical Transformations Depending on a Parameter", *Celestial Mechanics*, Vol. 1, 1969, pgs. 1-31.
9. Deprit, A. and Deprit-Bartholome, A. "Stability of the Triangular Lagrangian Points", *The Astronomical Journal*, Vol. 72, No. 2, 1967, pgs. 173-179.

10. Dragt, A. and Finn, J. M. "Lie Series and Invariant Functions for Analytic Symplectic Maps," *Journal of Mathematical Physics*, Vol. 17, 2215, 1976.

11. Goldstein, H. Classical Mechanics, 2nd ed., Addison-Wesley, Reading, Mass., 1980.

12. Hori, G. "Theory of General Perturbations with Unspecified Canonical Variables", *Publications of the Astronomical Society of Japan*, Vol. 18, No. 4, 1966, pgs. 287-296.

13. Len, J. "Nonlinear Parametric Excitation with Averaging and Lie Transform Methods", Ph.d. Thesis, Cornell University, 1987.

14. Len, J. and Rand, R. H. "Lie Transforms Applied to a Nonlinear Parametric Excitation Problem", *International Journal of Nonlinear Mechanics*, 1988 (to appear).

15. Lichtenberg, A.J. and Lieberman, M.A. Regular and Stochastic Motion, Applied Mathematical Sciences Vol. 38, Springer-Verlag, NY, 1983.

16. Meyer, K.R. and Schmidt, D.S. "The Stability of the Lagrange Triangular Point and a Theorem of Arnold", *Journal of Differential Equations*, Vol. 62, No. 2, 1986, pgs. 222-236.

17. Month, L. and Rand, R.H. "Bifucations of 4:1 subharmonics in the Nonlinear Mathieu Equation," *Mechanics Reasearch Communications*, Vol. 9(4), 1982, pgs. 233-240.

18. Rand, R.H. Computer Algebra in Applied Mathematics: An Introduction to MACSYMA, Research Notes in Mathematics, No. 94, Pitman Publishing, Boston, 1984.

19. Rand, R.H. and Armbruster, D. Perturbation Methods, Bifurcation Theory and Computer Algebra, Applied Mathematical Sciences, Vol.65, Springer-Verlag, NY, 1987.

20. Szebehely, V. Theory of Orbits: The Restricted Problem of Three Bodies, Academic Press, New York, 1967.

Symbolic Computation and Perturbation Methods Using Elliptic Functions

Vincent T. Coppola and Richard H. Rand
Department of Theoretical and Applied Mechanics
Cornell University
Ithaca, NY 14853

Abstract

We apply the method of averaging to first order in ϵ to the autonomous system

$$x'' + \alpha x + \beta x^3 = \epsilon g(x,x')$$

This involves perturbing off of Jacobian elliptic functions, rather than off of trignometric functions as is usually done. The resulting equations involve integrals of elliptic functions which are evaluated using a program written in the computer algebra system MACSYMA. The results are applied to the problem of finding limit cycles in the above differential equation.

Introduction

A limitation of most texts which treat nonlinear vibration problems by perturbation methods is that most problems involve perturbing off of the sine and cosine solutions of simple harmonic oscillators. For example, consider the nonlinear oscillator:

$$x'' + x = \epsilon \left[-x^3 + \frac{1}{2} x' + \frac{31}{10} x^2 x' - x'^3 \right], \quad \text{with } \epsilon = \frac{1}{10} \tag{1}$$

The usual approach to studying eq.(1) involves assuming that the parameter ϵ is asymptotically small, and perturbing off of the associated equation (for $\epsilon = 0$)

$$x'' + x = 0 \tag{2}$$

which has the general solution

$$x = C \cos (t + B) \tag{3}$$

The method of averaging [7-9,11-16] seeks a solution to eq.(1) when $\epsilon \neq 0$ in the form:

$$x = C(t) \cos \psi(t) \tag{4}$$

Variation of parameters and averaging over the unperturbed period 2π gives the usual formulas:

$$C' = \frac{\epsilon}{2\pi} \int_0^{2\pi} G(x,x') \sin \psi \, d\psi \tag{5.1}$$

$$\psi' = 1 + \frac{\epsilon}{2\pi C} \int_0^{2\pi} G(x,x') \cos \psi \, d\psi \tag{5.2}$$

in which eq.(1) has been written in the form

$$x'' + x + \epsilon G(x,x') = 0 \tag{6}$$

Evaluating eqs.(5) with $G(x,x') = x^3 - \frac{1}{2} x' - \frac{31}{10} x^2 x' + x'^3$ gives

$$C' = \frac{\epsilon}{80} C (C^2 + 20) \tag{7.1}$$

$$\psi' = 1 + \frac{3}{8} \epsilon C^2 \tag{7.2}$$

Nontrivial fixed points of eq.(7.1) are, in view of (4), periodic motions (limit cycles) of eq.(1). Since the only fixed point of (7.1) is C = 0, the method of averaging predicts that there are no limit cycles for eq.(1). This prediction is, however, erroneous! See Fig.1 which shows the results of numerically integrating eq.(1).

This embarrassing failure of averaging may be remedied in two ways. One may extend the averaging process to second order, i.e., include terms of $O(\epsilon^2)$. This involves combining the averaging process with a near-identity

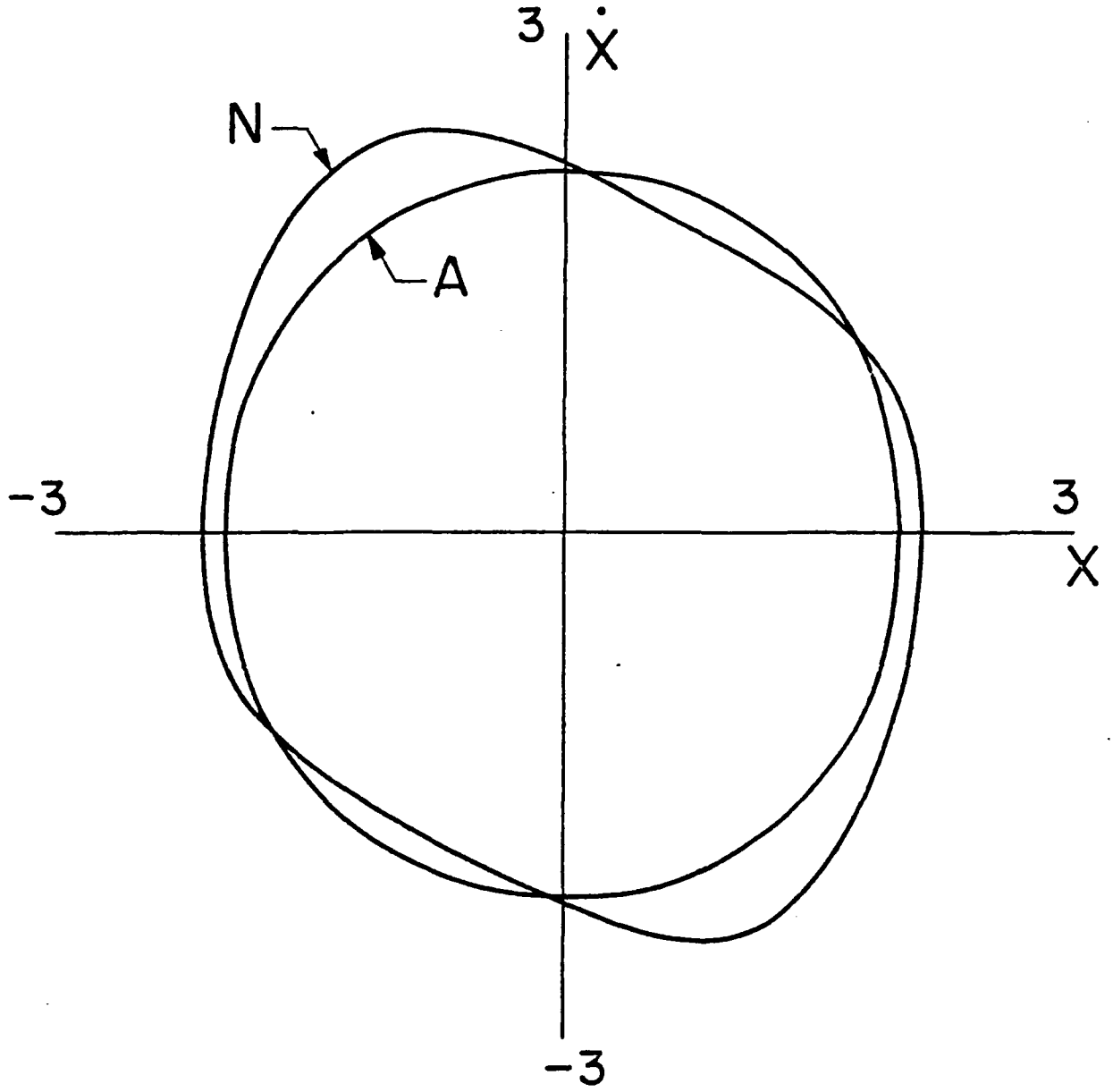


Fig.1. Limit cycle of eq.(1) obtained by numerical integration (N). Also shown is the analytic approximation (A) for the limit cycle obtained by using first order averaging utilizing elliptic functions, to be discussed later, see eq.(40). Note that first order averaging utilizing trigonometric functions fails to predict a limit cycle in this case, cf. eq.(7.1).

transformation of dependent variables. This route has been treated in [14], and computer algebra (MACSYMA) programs have been presented there to automate the process. Alternatively, one may stay with first order averaging, but follow the path presented in this paper.

In this paper we treat a class of problems which involve perturbing off of Jacobian elliptic functions. We consider the differential equation

$$x'' + \alpha x + \beta x^3 + \epsilon g(x,x') = 0 \ , \ \alpha > 0 \ , \ \beta > 0 \tag{8}$$

in which β is <u>not</u> assumed to be a small quantity. The unperturbed system is

$$x'' + \alpha x + \beta x^3 = 0 \tag{9}$$

which has the general solution

$$x = C \, cn(u,k), \tag{10}$$

$$u = \sqrt{\frac{\beta}{2k^2}} \, C \, t + u_o \quad \text{and} \quad k^2 = \frac{\beta C^2}{2(\alpha + \beta C^2)}$$

where cn(u,k) is a Jacobian elliptic function. We use the method of averaging implemented on MACSYMA to treat this type of problem. We compare results found using elliptic functions with those found using trigonometric functions. In particular we will return to eq.(1) later in this paper.

Although the method of averaging has been treated in numerous references (e.g. [7-9,11-16]), each deals almost exclusively with perturbations off of the simple harmonic oscillator, eq.(2). A few authors have considered

perturbations off of nonlinear systems using elliptic functions. Kuzmak [10] looked for periodic solutions in eq.(8) using a multiple scale method, where α and β are slowly varying parameters. Chirikov [3] studied resonance overlap in multiple harmonic excitations of eq.(8). Davis [4] investigated second order ordinary differential equations using elliptic functions. Cap [2] applied the method of averaging to perturbations of the mathematical pendulum. Greenspan and Holmes [6] and Guckenheimer and Holmes [7] have applied the Melnikov method to perturbations of eq.(8) where $\alpha < 0$. Nayfeh [12], Kevorkian & Cole [9] and Sanders & Verhulst [15] have also treated such problems. In most of these references the authors have reduced the problem to the evaluation of integrals which, through complicated algebraic manipulations, may often be expressed in terms of standard elliptic integrals. By using MACSYMA, we have been able to treat a large class of problems by efficiently evaluating the associated integrals.

We begin with a brief review of elliptic functions. Then we present a general treatment of averaging to systems of the form of eq.(8), and finally we apply the method to the problem of finding limit cycles in eq.(8).

Jacobian Elliptic Functions

Jacobian elliptic functions involve a collection of identities which are similar to those for trigonometric functions but are more complicated algebraically. The use of computer algebra makes manipulation of these identities easier, permitting investigations to proceed on problems which were previously avoided because of the quantities of algebra involved. All formulas and conventions concerning Jacobian elliptic functions in this paper are taken from Byrd and Friedman's Handbook of Elliptic Integrals for Engineers and Physicists [1].

We now offer a brief comparison of elliptic functions with the more familiar trigonometric functions. Corresponding to sin(u) and cos(u) are three fundamental elliptic functions sn(u,k), cn(u,k), and dn(u,k). Each of the elliptic functions depends on the modulus k as well as the argument u. These reduce to sin(u), cos(u), and 1 respectively, when k = 0. The sn and sin functions share common properties as do cn and cos. These are summarized in Table 1. The dn function has no trigonometric counterpart. Note that the elliptic functions sn and cn may be thought of as generalizations of sin and cos where their period depends on the modulus k.

The argument u is identified as the incomplete elliptic integral of the first kind which is usually denoted $F(\theta,k)$. This identification shows that u also depends on k. The value of k normally ranges from 0 to 1. The sn, cn, and dn functions are shown in Fig.2 for $k^2 = 1/2$.

Table 1

Function f

Property	sn(u,k)	sin(u)	cn(u,k)	cos(u)	dn(u,k)
Max. value	1	1	1	1	1
Min. value	-1	-1	-1	-1	$(1-k^2)^{1/2}$
Period	4 K(k)	2π	4 K(k)	2π	2 K(k)
Odd/Even	Odd	Odd	Even	Even	Even
df/du	cn dn	cos	-sn dn	-sin	$-k^2$ sn cn
$f \mid_{k=0}$	sin	sin	cos	cos	1

K(k) = complete elliptic integral of the first kind

$K(0) = \pi/2, \quad K(1) = +\infty$

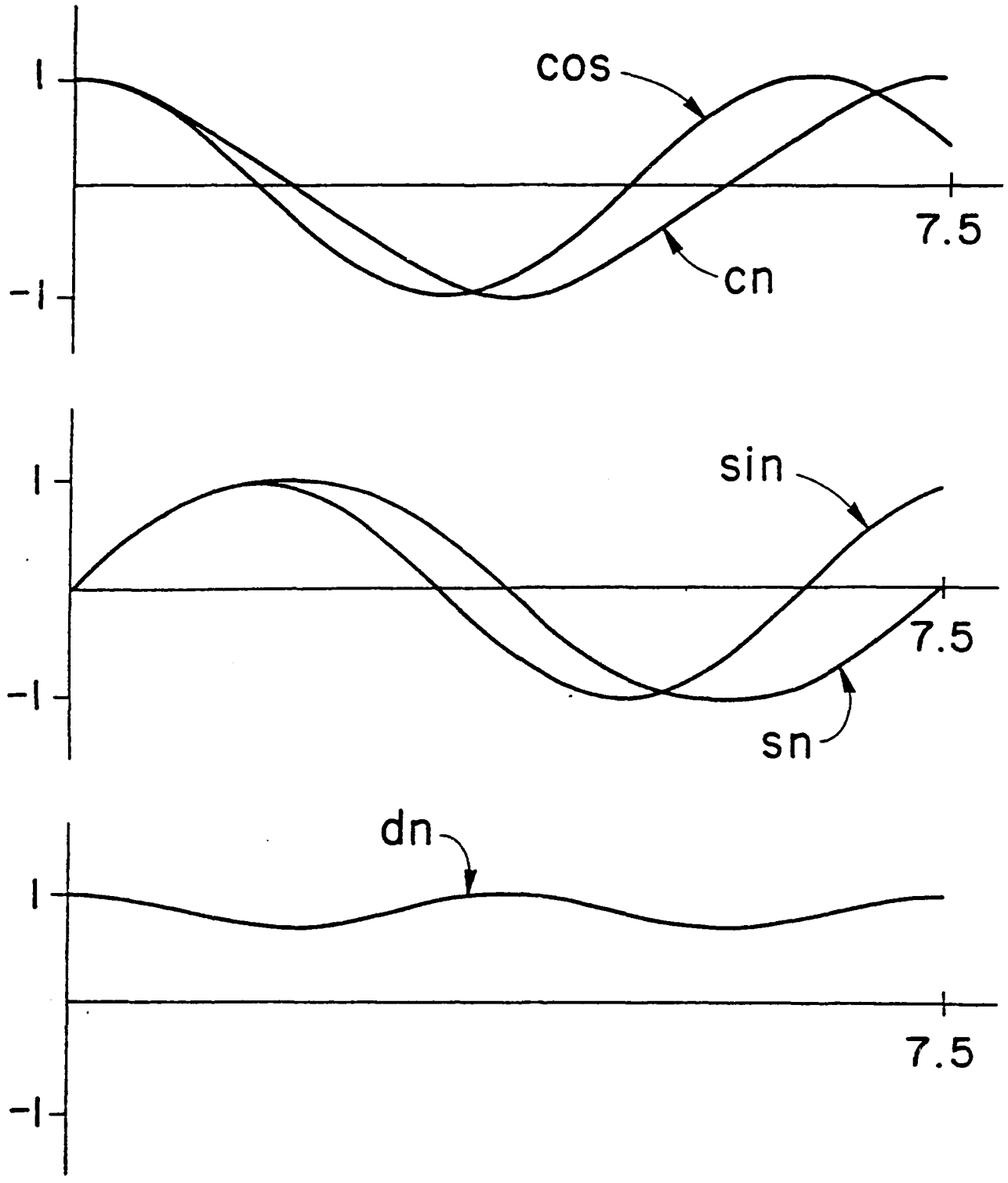


Fig.2. Comparison of elliptic functions for $k^2 = 1/2$ with trigonometric functions. The period of the elliptic functions is $4\,K(k^2 = 1/2) \simeq 7.416$. See Table 1.

The elliptic functions also satisfy the following identities which correspond to $\sin^2 + \cos^2 = 1$:

$$sn^2 + cn^2 = 1 \tag{11.1}$$

$$k^2\, sn^2 + dn^2 = 1 \tag{11.2}$$

$$1 - k^2 + k^2\, cn^2 = dn^2 \tag{11.3}$$

The Unperturbed Solution

We shall consider unperturbed systems of the form of eq.(9). We find the general solution by assuming the solution in the form

$$x = C\, cn(A\, t + B,k) = C\, cn(u,k) = C\, cn \tag{12}$$

where the argument is omitted for brevity and where A and C are positive constants. Substituting (12) into eq.(9) we find

$$[C\, A^2\, (2k^2 - 1) + \alpha\, C]\, cn + [C^3\, \beta - 2k^2\, A^2\, C]\, cn^3 = 0 \tag{13}$$

where we have used the relation

$$\frac{\partial^2 cn(u,k)}{\partial u^2} \equiv cn'' = (2k^2 - 1)\, cn - 2k^2\, cn^3 \tag{14}$$

For nontrivial solutions ($C \neq 0$), we find

(15.1) $$A^2 (1 - 2k^2) = \alpha$$

(15.2) $$C^2 \beta = 2k^2 A^2$$

We define three separate cases depending on the parameters α and β.

Case I: $\alpha \neq 0$, $\beta = 0$

From eqs.(15) we find that

(16.1) $$k = 0, \; A^2 = \alpha, \; C \text{ is undetermined}$$

which is the correct solution for the harmonic oscillator.

Case II: $\alpha \neq 0$, $\beta \neq 0$

From eqs.(15) we find that

(16.2) $$k^2 = \frac{A^2 - \alpha}{2 A^2} = \frac{\beta C^2}{2 (\alpha + \beta C^2)}, \qquad A^2 = \alpha + \beta C^2$$

For $C > 0$, the range of k^2 is $0 < k^2 < \frac{1}{2}$. We note that cases I and III are limiting cases of II that are recoverable by setting k^2 equal to zero or one-half.

Case III: $\alpha = 0,\ \beta \neq 0$

From eqs.(15) we find that

$$k^2 = \frac{1}{2}\ ,\ A^2 = \beta\, C^2 \tag{16.3}$$

The oscillator is purely nonlinear in this instance.

In all three cases, the origin is a center and the (x,x') phase space is filled with periodic orbits. For cases II and III, the period of an orbit depends on its amplitude, see Table 2. As the amplitude of the vibration approaches the origin ($C \to 0$), the period of oscillation increases to the value $T = 2\pi/\sqrt{\alpha}$, which becomes infinite for case III.

Table 2

Case	Period T
I	$\frac{2\pi}{\sqrt{\alpha}}$
II	$\frac{4\,K(k(C))}{\sqrt{\alpha + \beta\, C^2}}$
III	$\frac{4\,K(k^2=1/2)}{\sqrt{\beta}\, C}$

Since eq.(9) is Hamiltonian, it possesses the first integral

$$H = \frac{1}{2} x'^2 + \frac{1}{2} \alpha x^2 + \frac{1}{4} \beta x^4 \tag{17}$$

which provides another method for solving eq.(9). We define action-angle variables (J,φ) for this Hamiltonian [5,7] in order to provide more "natural" variables to be used later in setting up the averaging scheme. After some lengthy calculations, we find that

$$J = J(C) \tag{18.1}$$

$$4\,K(k)\,\varphi = A\,t + B = u \tag{18.2}$$

For simplicity, we take the variables (C,φ) as primitive.

It is interesting to note why the variable φ is preferred to B in deriving the averaging scheme. First note that although each orbit in phase space is orbitally stable [8,11,16], it is Lyapunov unstable. This is because the frequency of an orbit depends on its amplitude, and motions starting close together but on two different orbits eventually become far apart (in fact, out of phase), even though their orbits are close.

In the next section we derive the equations governing averaging based on the variables (C,φ), cf. eqs.(5) for the simple harmonic oscillator. In a similar fashion we could attempt to derive comparable equations based on using the phase B or the argument u of the unperturbed solution (cf.eq.(12)) instead of the angle variable φ. In doing so for (C,B), we would obtain equations of the form (before averaging):

$$C' = \epsilon\, f_1(C, At+B), \quad B' = -A't + \epsilon\, f_2(C, At+B) \tag{19}$$

in which f_2 turns out not to be periodic in t. The method of averaging requires that the variational equations be periodic. Thus eqs.(19) are unsuitable for averaging. The orbital stability of the solutions is reflected in the variational equation for C (or equivalently in the variational equation for A, since A and C are related algebraically, cf.(16.2)). The Lyapunov (phase) instability is reflected in the variational equation for B.

Similarly, choosing (C,u) as primitive variables, u = At+B, gives

$$C' = \epsilon\, f_1(C,u), \quad u' = A + \epsilon\, f_2(C,u) \tag{20}$$

in which f_2 is not periodic in u, so that eqs.(20) are again unsuitable for averaging.

However, setting $u = 4\,K(k(C))\,\varphi$, cf.(18.2), gives

$$C' = \epsilon\, f_1(C, 4\,K\,\varphi), \quad \varphi' = \frac{A}{4\,K} + \epsilon\, f_2(C, 4\,K\,\varphi) \tag{21}$$

in which both f_1 and f_2 are found to be periodic in φ and hence in the correct form for averaging. Thus the unperturbed solution can be written as

$$x = C\ cn(4\,K\,\varphi, k) \tag{22.1}$$

$$x' = C\,A\,cn'(4\,K\,\varphi, k) = C\sqrt{\alpha + \beta\,C^2}\ cn'(4\,K\,\varphi,\ k) \tag{22.2}$$

$$K = K(k), \quad k = k(C), \quad cn' \equiv \frac{\partial cn(u,k)}{\partial u} \tag{22.3}$$

which can be viewed as a generalized van der Pol transformation from (x,x') to (C,φ). In this way, (C,φ) constitute "natural" variables because they take intc account the change of period occurring from orbit to orbit in the unperturbed flow.

Variation of Parameters

In order to obtain a solution to eq.(8) when $\epsilon \neq 0$, we vary the parameters (C,φ) so that $C = C(t)$ and $\varphi = \varphi(t)$ in eqs.(22). Differentiating x in (22.1) and equating the result to (22.2), we obtain

$$(23)\quad \frac{dC}{dt}\left(cn + C\,cn'\,4\,\varphi\,K'\,k' + C\,\frac{\partial cn}{\partial k}\,k'\right) + C\,cn'\,4\,K\,\frac{d\varphi}{dt} = C\,A\,cn'$$

where primes denote differentiation with respect to the argument (the first argument in the case of cn). Differentiating eq.(22.2), we find

$$(24)\quad x'' = \frac{dC}{dt}\left[(A + A'\,C)\,cn' + 4\,C\,A\,K'\,k'\,\varphi\,cn'' + C\,A\,k'\,\frac{\partial cn'}{\partial k}\right] + 4\,C\,A\,K\,cn''\,\frac{d\varphi}{dt}$$

We substitute eqs.(24) and (22.1) into (8) and solve for dC/dt and dφ/dt. The result can be written in matrix notation as

$$(25)\quad \begin{bmatrix} & W & \end{bmatrix} \begin{bmatrix} \frac{dC}{dt} \\ \frac{d\varphi}{dt} \end{bmatrix} = \begin{bmatrix} C\,A\,cn' \\ -\epsilon\,g - \alpha\,C\,cn - \beta\,C^3\,cn^3 \end{bmatrix}$$

In solving these equations, we need to compute the determinant of W:

$$(26)\quad \det[W] = 4\,C\,K\,A\,(cn\,cn'' - cn'^2) - 4\,C^2\,K\,A'\,cn'^2$$

$$+ 4\,C^2\,A\,K\,k'(cn''\,\frac{\partial cn}{\partial k} - cn'\,\frac{\partial cn'}{\partial k})$$

$$= -4\,K\,C\,A = -4\,K\,C\,(\alpha + \beta\,C^2)^{1/2}$$

where we have used eq.(14), eq.(16.2), and the identities [1]:

$$(27.1)\qquad cn'^2 = (1 - cn^2)\,(1 - k^2 + k^2\,cn^2)$$

$$(27.2)\quad cn'\,\frac{\partial cn'}{\partial k} = \frac{1}{2}\,\frac{\partial}{\partial k}(cn'^2) = \frac{1}{2}\,\frac{\partial}{\partial k}\{(1 - cn^2)\,(1 - k^2 + k^2\,cn^2)\}$$

Note that the determinant of W is independent of φ. We then solve eq.(25) to find

$$(28.1)\qquad \frac{dC}{dt} = -\epsilon\,g\,(\alpha + \beta\,C^2)^{-1/2}\,cn'$$

$$(28.2)\quad \frac{d\varphi}{dt} = \frac{1}{4\,K}\,(\alpha + \beta\,C^2)^{1/2} + \epsilon\,g\,\frac{1}{4\,K\,C}\,(\alpha + \beta\,C^2)^{-1/2}$$

$$\times\left[cn - \frac{2\,\alpha}{2\,\alpha + \beta\,C^2}\left[Z\,cn' + \frac{\beta\,C^2}{2\,(\alpha + \beta\,C^2)}\,cn\,(1 - cn^2)\right]\right]$$

where $Z = Z(4\,K\,\varphi, k)$ denotes the Jacobi Zeta function (an odd 2 K periodic

function with zero mean) and all arguments are 4 K φ. In eq.(28.2), we have used [1]:

$$\frac{\partial cn}{\partial k} = \frac{cn'}{k\,(1-k^2)}\,[(1-k^2)\;4\;K\;\varphi - E(4\;K\;\varphi,k)] \tag{29.1}$$

$$- \frac{k}{(1-k^2)}\,cn\,(1-cn^2)$$

$$Z(4\;K\;\varphi,k) = E(4\;K\;\varphi,k) - 4\;\varphi\;E \tag{29.2}$$

where $E(4\;K\;\varphi,k)$ is shorthand notation for $E(\theta,k)$, the incomplete elliptic integral of the second kind (where $\theta = am(4\;K\;\varphi,k)$ and $am(u,k)$ is the elliptic amplitude function [1]) and $E = E(k)$ denotes the complete elliptic integral of the second kind.

We consider eqs.(28) in the three cases I,II,III separately:

Case I:

$$A^2 = \alpha,\; K(k=0) = \frac{\pi}{2},\; Z(u,k=0) = 0 \tag{30}$$

$$\frac{dC}{dt} = -\,\epsilon\,\frac{1}{\sqrt{\alpha}}\,g\;cn'(2\;\pi\;\varphi,k=0) = \epsilon\,\frac{1}{\sqrt{\alpha}}\,g\;\sin(2\;\pi\;\varphi) \tag{31.1}$$

$$\frac{d\varphi}{dt} = \frac{\sqrt{\alpha}}{2\;\pi} + \epsilon\,\frac{1}{2\;\pi\;C}\,\frac{1}{\sqrt{\alpha}}\,g\;\cos(2\;\pi\;\varphi) \tag{31.2}$$

which agrees with the perturbation equations off the linear oscillator, cf.eqs.(5) with $\psi = 2\pi\varphi$.

Case II:

Here the variables C and k are related. Since the modulus k is a natural elliptic function quantity, one could formulate eqs.(28) in terms of k and φ:

$$(32.1)\quad \frac{dk}{dt} = \frac{\alpha\sqrt{\beta}}{\sqrt{2}\,(\alpha + \beta C^2)^{3/2}} \frac{dC}{dt} = -\epsilon g \frac{\sqrt{\beta}}{\sqrt{2}\,\alpha} (1 - 2k^2)^2 \, cn'$$

$$(32.2)\quad \frac{d\varphi}{dt} = \frac{\sqrt{\alpha}}{4 K (1 - 2k^2)}$$

$$+ \epsilon g \frac{\sqrt{\beta}\,(1 - 2k^2)}{4\sqrt{2} K \alpha k} \left[cn - \frac{(1 - 2k^2)}{(1 - k^2)} \left[Z\, cn' + k^2 cn\,(1 - cn^2) \right] \right]$$

Note that this formulation breaks down for cases I and III.

Case III:

Eqs.(28) simplify to

$$(33.1)\quad \frac{dC}{dt} = -\epsilon \frac{1}{\sqrt{\beta}} \frac{1}{C} g \, cn'$$

$$(33.2)\quad \frac{d\varphi}{dt} = \sqrt{\beta} \frac{C}{4 K} + \epsilon g \frac{1}{\sqrt{\beta}} \frac{1}{4 K C^2} cn$$

We will formulate the averaging procedure in terms of (C, φ) and for case II, we will use k and C interchangeably (via eq.(32.1)).

The Averaging Procedure

While eqs.(28) are valid for any perturbation g, in this section we consider perturbations of the form $g = g(x,x')$, where g is a polynomial in x and x'. We write eqs.(28) in the form

$$C' = \epsilon F_1(C,\varphi) \tag{34.1}$$

$$\varphi' = \frac{1}{4K}(\alpha + \beta C^2)^{1/2} + \epsilon F_2(C,\varphi) \tag{34.2}$$

$$= \Omega(C) + \epsilon F_2(C,\varphi)$$

where the F_i, as given by eqs.(28), are periodic in φ.

We denote the averaged variables by $(\bar{C},\bar{\varphi})$. Then, the averaged equations become

$$\bar{C}' = \epsilon \bar{F}_1 + O(\epsilon^2) \tag{35.1}$$

$$\bar{\varphi}' = \Omega(\bar{C}) + \epsilon \bar{F}_2 + O(\epsilon^2) \tag{35.2}$$

where $\bar{F}_i$ are the mean values of F_i over one period of the unperturbed system:

$$\bar{F}_i = \frac{1}{T}\int_0^T F_i \, d\varphi = \frac{1}{4\bar{K}}\int_0^{4\bar{K}} F_i(\bar{C},\bar{u}) \, d\bar{u} \,, \tag{36}$$

where $\bar{u} = 4\bar{K}\bar{\varphi}$, $\bar{K} = K(\bar{k})$, $\bar{k} = k(\bar{C})$ as given by eqs.(16).

Computer Algebra Implementation Of The Averaging Scheme

We present a short summary of our implementation of the averaging scheme on the computer algebra system MACSYMA. The perturbation g is composed of a sum of terms of the form

$$x^n\, x'^m = C^{n+m}\, A^m\, cn^n\, cn'^m \tag{37}$$

each of which can be written as a sum of terms of the form

$$C^{n+m}\, A^m\, cn^{n+(m-1)/2}\, cn' \quad , \; m \text{ odd} \tag{38.1}$$

$$C^{n+m}\, A^m\, cn^{n+m/2} \quad , \; m \text{ even} \tag{38.2}$$

using eq.(27.1). It is therefore sufficient to consider g to be composed of a sum of terms of the form cn^m and $cn^m\, cn'$. By inspection of eqs.(28), we can make a list of all combinations of elliptic functions which can possibly occur in the integrands of eqs.(36), and their mean values. The integrands are listed in Table 3 and their mean values in Table 4.

Table 3. Terms occurring in F_i

Expression	Typical terms
F_1	cn^m, $cn^m\, cn'$
F_2	cn^m, $cn^m\, cn'$, $Z\, cn^m$, $Z\, cn^m\, cn'$

Table 4. Mean values of elliptic functions

Function	Mean Value
cn^m	D_m for m even
	0 for m odd
$cn^m\ cn'$	0
$Z\ cn^m$	0
$Z\ cn^m\ cn'$	0 for m even
	$-\frac{1}{m+1}\left[\left(1-k^2-\frac{E}{K}\right) D_{m+1} + k^2 D_{m+3}\right]$, m odd

where

$$D_0 = 1$$

$$D_2 = \frac{1}{k^2}\left(\frac{E}{K} - 1 + k^2\right)$$

$$D_m = \frac{1}{(m-1)\,k^2}\left[(m-2)\,(2k^2-1)\,D_{m-2} + (m-3)\,(1-k^2)\,D_{m-4}\right]$$

Armed with Table 4, one could find the averaged equations for a given perturbation g(x,x') by hand. This lengthy calculation, however, is much better suited to MACSYMA. The MACSYMA program which implements the foregoing averaging procedure is listed in the Appendix. As an example of its use, we next apply the method to the problem of finding limit cycles in eq.(8). We begin by returning to eq.(1), then we generalize the example.

Example: Eq.(1) revisited

If we write eq.(1) in the form

$$x'' + x + \frac{1}{10}x^3 = \epsilon\left[\frac{1}{2}x' + \frac{31}{10}x^2x' - x'^3\right], \quad \text{with } \epsilon = \frac{1}{10}$$

we identify $\alpha = 1$, $\beta = \frac{1}{10}$, $g = -\frac{1}{2}x' - \frac{31}{10}x^2x' + x'^3$. Substitution of these values into eqs.(28) and averaging gives (see sample run of our MACSYMA program in the Appendix):

$$(39) \qquad \bar{C}' = -\epsilon\,\frac{P(\bar{C})\,K - Q(\bar{C})\,E}{350\,\bar{C}\,K}, \qquad \bar{\varphi}' = \frac{(1 + \bar{C}^2/10)^{1/2}}{4\,K}$$

$$\text{where} \quad P(\bar{C}) = 5\,\bar{C}^6 + 447\,\bar{C}^4 + 10175\,\bar{C}^2 + 64700$$

$$Q(\bar{C}) = 594\,\bar{C}^4 + 11880\,\bar{C}^2 + 64700$$

$$\bar{k}^2 = \frac{\bar{C}^2}{2\,\bar{C}^2 + 20}$$

and where $K = K(\bar{k})$ and $E = E(\bar{k})$ represent complete elliptic integrals of the first and second kinds respectively. Numerical evaluation of the condition $\bar{C}' = 0$ gives the limit cycle amplitude $\bar{C} \cong 1.9861$. Then eq.(22.1) gives the following approximation for the limit cycle:

$$(40) \qquad x = 1.9861\ \mathrm{cn}(1.1808\ t,\ k=0.37608)$$

This approximation offers reasonable agreement with numerical integration of

eq.(1) for $\epsilon = 1/10$, see Fig.1. Note that first order averaging off of the simple harmonic oscillator failed to predict a limit cycle for this equation, cf. eq.(7.1).

Example: Limit Cycles in Eq.(8)

We investigate limit cycle solutions in systems of the form

$$x'' + \alpha x + \beta x^3 + \epsilon g = 0 \tag{41}$$

in which $g = \delta x' + \Sigma v_{ij} x^i x'^j$, where $2 \leq i+j \leq 4$

Using eq.(27.1), eq.(28.1), and Table 4, we find that the only terms that make nonzero contributions to C' are

$$\delta x', \; v_{21} x^2 x', \; v_{03} x'^3 \tag{42}$$

The condition for a limit cycle is that C' be zero, i.e., $\bar{F}_1 = 0$. This condition on the parameters δ, v_{21}, and v_{03} will then determine the limit cycle radius (if a limit cycle exists). The other ten terms in g do not influence the existence of a limit cycle (to $O(\epsilon)$). Therefore, we take a modified perturbation for g:

$$g = \delta x' + \rho x^2 x' + \eta x'^3 \tag{43}$$

Note also that $\delta = \rho = \eta = 0$ implies the existence of a family of closed orbits, and not limit cycles.

Eq.(45) can be viewed as a relationship between the limit cycle amplitude C and the parameters δ, ρ, and η. When any two of these parameters are zero, there can be no limit cycle. Eq.(45) is singular when the product $\eta\,\beta$ vanishes, i.e., the normally quadratic equation in C^2 becomes linear. When $\eta\,\beta = 0$, eq.(45) reduces to:

$$C^2 = -\frac{\delta\, V_1}{\rho\, V_2 + \eta\,\alpha\, V_3} \quad , \ \eta\,\beta = 0 \tag{46.1}$$

Eq.(46.1) can have at most one positive root, and hence there can be at most one limit cycle. For $\eta\,\beta \neq 0$, we solve eq.(45) to get

$$C^2 = \frac{-\alpha\,\eta\, V_3 - \rho\, V_2 \pm \sqrt{(\alpha\,\eta\, V_3 + \rho\, V_2)^2 - 4\,\beta\,\delta\,\eta\, V_1\, V_3}}{2\,\beta\,\eta\, V_3} \quad , \ \eta\,\beta \neq 0 \tag{46.2}$$

For cases I and III, C does not depend on k. Eqs.(46) are then explicit relations between the parameters and amplitude C for the existence of a limit cycle. For case II, however, C does depend on k so that eqs.(46) only implicitly define C (since V_i depends on k). Investigating eq.(46.2) numerically, we find C^2 can be made to have zero, one, or two positive roots for real C. A bifurcation occurs along the curve that is the intersection of eq.(45) with

$$\frac{d}{dC}\{\delta\, V_1 + \rho\, C^2\, V_2 + \eta\, C^2\,(\alpha + \beta\, C^2)\, V_3\} = 0 \tag{47}$$

We find F_1 and $\bar{F}_1$ to be (cf. eqs.(34).(36)):

$$(44.1) \qquad F_1 = - C \left[\delta \, cn'^2 + \rho \, C^2 \, cn^2 \, cn'^2 + \eta \, C^2 \, (\alpha + \beta \, C^2) \, cn'^4\right]$$

$$(44.2) \qquad \bar{F}_1 = - \bar{C} \left[\delta \, V_1 + \rho \, \bar{C}^2 \, V_2 + \eta \, \bar{C}^2 \, (\alpha + \beta \, \bar{C}^2) \, V_3\right]$$

where

$$V_1 = \text{mean of } cn'^2$$

$$= - \frac{1}{3 \, k^2 \, K} \left[K \, (k^2 - 1) + E \, (1 - 2k^2)\right]$$

$$V_2 = \text{mean of } cn^2 \, cn'^2$$

$$= - \frac{1}{15 \, k^4 \, K} \left[K \, (k^4 - 3k^2 + 2) - 2 \, E \, (k^4 - k^2 + 1)\right]$$

$$V_3 = \text{mean of } cn'^4$$

$$= \frac{1}{35 \, k^4 \, K} \left[K \, (8k^6 - 13k^4 + 3k^2 + 2) - 2 \, E \, (8k^6 - 12k^4 + 2k^2 + 1)\right]$$

We drop the bar notation for convenience here and in what follows. The value of k is related to C by eqs.(16). The V_i turn out to be positive for valid values of k. Ignoring the trivial case C = 0, a limit cycle exists ($\bar{F}_1$ becomes zero) for:

$$(45) \qquad \delta \, V_1 + \rho \, C^2 \, V_2 + \eta \, C^2 \, (\alpha + \beta \, C^2) \, V_3 = 0$$

Limit cycles on this curve are degenerate. As one moves across this bifurcation curve, two limit cycles coalesce at a finite non-zero radius. We continue the discussion by considering the limiting cases I and III:

Case I: Results for the linear oscillator

The values of V_i become indeterminate at $k = 0$. By taking limits we find (cf. eq.(46.1)):

$$V_1 = \frac{1}{2} \, , \; V_2 = \frac{1}{8} \, , \; V_3 = \frac{3}{8} \tag{48.1}$$

$$C^2 = - \frac{4 \, \delta}{\rho + 3 \, \alpha \, \eta} \tag{48.2}$$

This agrees with the solution found in [14] by perturbing off of the linear oscillator.

Case III: Results for the Purely Nonlinear Oscillator

We evaluate V_i and C^2 to be (cf. eqs.(46)):

$$V_1 = \frac{1}{3} \, , \; V_2 = .09139... \, , \; V_3 = \frac{1}{7} \tag{49.1}$$

$$C^2 = - \frac{1}{3 \, V_2} \frac{\delta}{\rho} \quad , \; \eta = 0 \tag{49.2}$$

$$C^2 = \frac{- 7 \sqrt{3} \, \rho \, V_2 \pm \sqrt{7} \sqrt{21 \, \rho^2 \, V_2^{\,2} - 4 \, \beta \, \delta \, \eta}}{2 \sqrt{3} \, \beta \, \eta} \, , \; \eta \neq 0 \tag{49.3}$$

We continue the discussion of this problem by considering the number of limit cycles which occur for given values of the parameters, i.e., the bifurcation set.

The Bifurcation Set

The cases $\eta = 0$ and $\eta \neq 0$ are considered separately. The latter case is then divided into the two cases $\beta = 0$ and $\beta \neq 0$.

Case $\eta = 0$

From eq.(46.1), we expect at most one limit cycle with amplitude C satisfying

(50) $$C^2 \frac{V_2}{V_1} = \mu_o$$

where $\mu_o = (-\frac{\delta}{\rho})$ is a parameter

We now compare limit cycle bifurcation curves for cases I, II, and III. Eq.(50) reduces in these instances to (cf. eqs.(48.1), (16.2), (49.1)):

(51.1) $$\text{Case I:} \quad \frac{1}{4} C^2 = \mu_o$$

(51.2) $$\text{Case II:} \quad (\frac{\alpha}{1 - 2k^2}) (\frac{2k^2}{\beta}) \frac{V_2(k)}{V_1(k)} = \mu_o$$

(51.3) $$\text{Case III:} \quad 0.27417\ldots\ C^2 = \mu_o$$

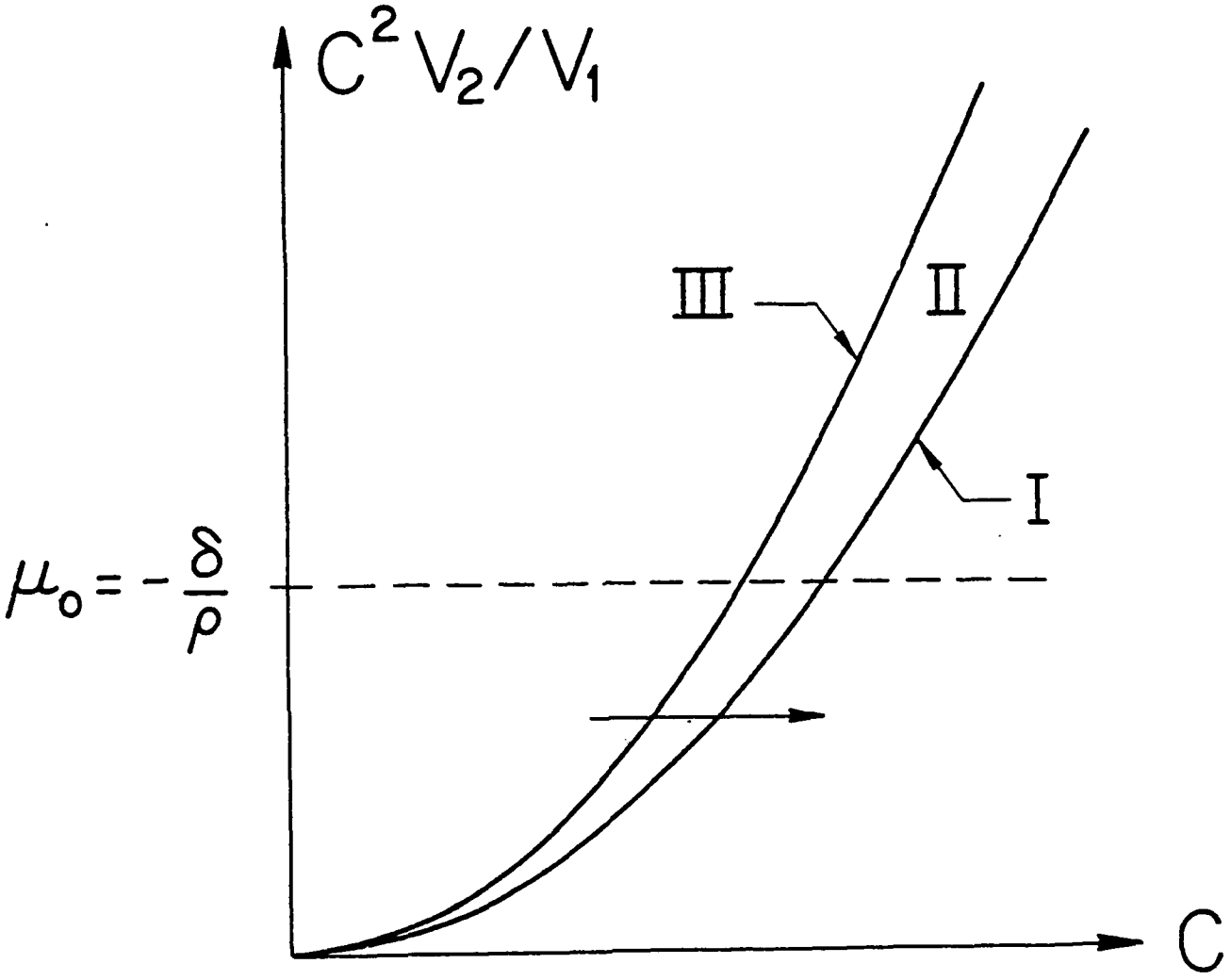


Fig.3. Limit cycle amplitude C for eqs.(41) and (43) with $\eta = 0$. C, shown as the abscissa, is determined by the intersection of a particular $C^2 V_2/V_1$ curve (which depends on α/β) with the straight line $\mu_0 = -\delta/\rho$. The arrow shows the increase in limit cycle amplitude C resulting from increasing α/β while holding μ_o fixed.

A graph of eqs.(51) appears in Fig.3. Note that cases I and III provide bounds for case II and that μ_0 is rather insensitive to the parameter ratio α/β. Numerical experiments support these analytical predictions.

Case $\eta \neq 0$

From eq.(45), we find

$$\mu_2 = \frac{C^2}{V_1} [V_2 \mu_1 + (\alpha + \beta C^2) V_3] \tag{52}$$

where $\mu_1 = \frac{\rho}{\eta}$ and $\mu_2 = (-\frac{\delta}{\eta})$ are parameters

Eq.(52) defines a family of straight lines in the (μ_1, μ_2) parameter plane with slopes and intercepts parameterized by α, β, and C. Both the slope and the μ_2-intercept have the value zero at C = 0, and increase as C increases.

Case $\eta \neq 0$, $\beta = 0$ (Case I)

Eq.(52) becomes (cf. eqs.(16.1),(48.1)):

$$\mu_2 = \frac{1}{4} C^2 [3 \alpha + \mu_1] \tag{53}$$

with μ_1-intercept at point P ($\mu_1 = -3 \alpha$, $\mu_2 = 0$) for all values of C. A graph of eq.(53) parameterized by C is given in Fig.4. There is one limit cycle in regions I and II; there are none in regions III and IV. The $\mu_2 = 0$ line is a

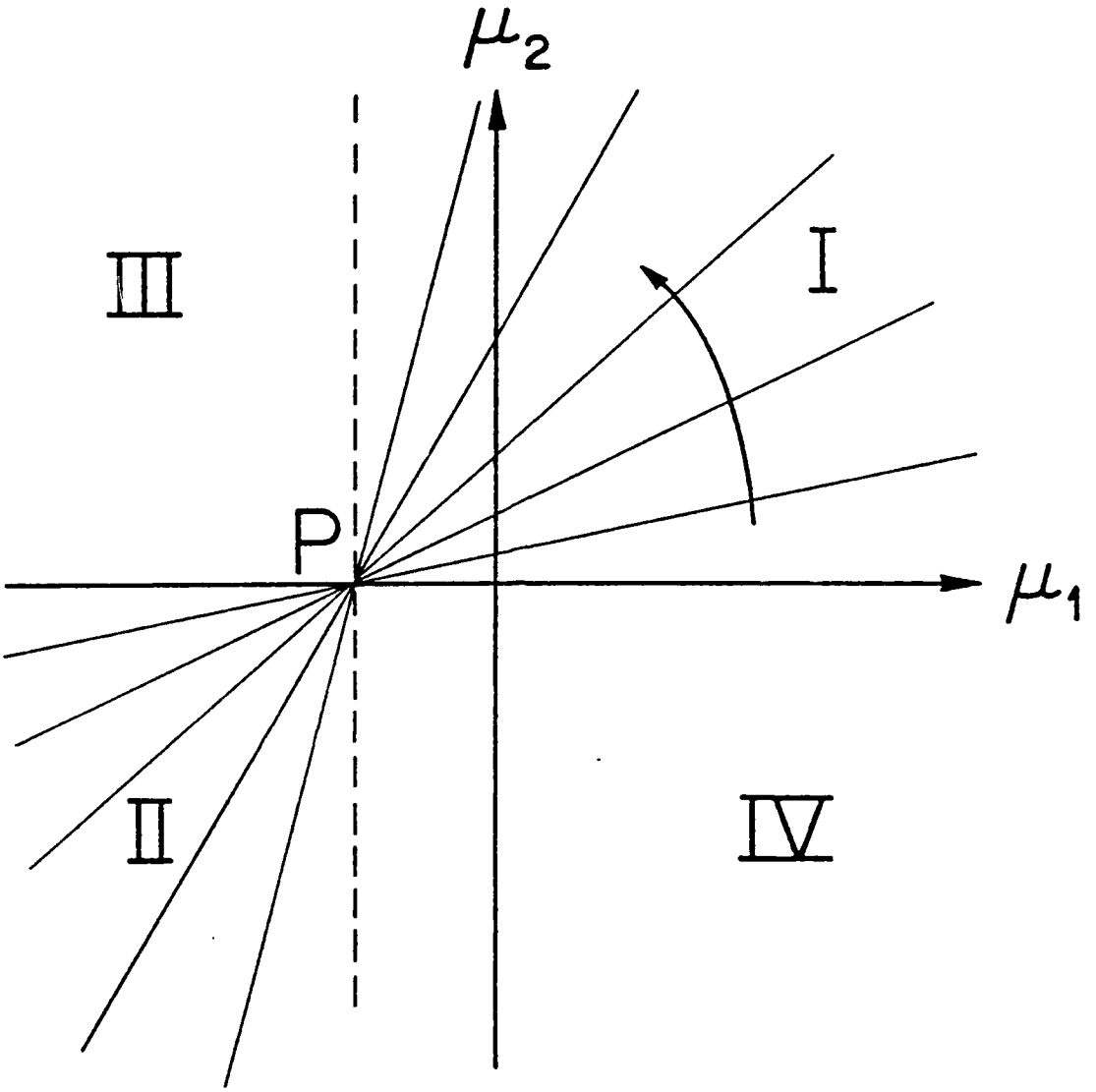


Fig.4. Limit cycles in eq.(41) for $\beta = 0$. The parameters μ_1 and μ_2 are defined by $\mu_1 = \rho/\eta$ and $\mu_2 = -\delta/\eta$, cf. eq.(53). Along each straight line there exists a limit cycle of fixed amplitude. Thus, in regions I and II there exists 1 limit cycle while in regions III and IV there are no limit cycles. The μ_1 axis corresponds to the limiting case of a limit cycle of zero amplitude (and, hence, a Hopf bifurcation occurs as one crosses the μ_1 axis.). The dashed line is $\mu_1 = -3\alpha$ and corresponds to limit cycles of infinite amplitude. The arrow shows the direction of increasing limit cycle amplitude.

Hopf bifurcation curve where a limit cycle is born at the origin. On the bifurcation line $\mu_1 = -3\alpha$, a limit cycle of infinite amplitude is predicted. Point P is a highly singular point: near P, the limit cycle amplitude is very sensitive to small changes in μ_1 and μ_2.

Case $\eta \neq 0$, $\beta \neq 0$ (Cases II and III)

The μ_1 intercept moves out from $\mu_1 = -3\alpha$ at $C = 0$ towards infinity as $C \to \infty$. With this information, we plot eq.(52), parameterized by C, in the (μ_1, μ_2) plane (see Fig.5). One limit cycle exists in region I, two in region II (where each point lies on exactly two intersecting lines), and none in region III. A degenerate limit cycle exists on the bifurcation curve between II and III. The μ_1 axis is a Hopf bifurcation curve where a limit cycle is born at the origin. Point P ($\mu_1 = -3\alpha$, $\mu_2 = 0$) is again a singular point where a degenerate limit cycle of zero amplitude exists. Near P, the sensitivity of the amplitude on μ_1 and μ_2 depends on the smallness of β.

The predictions of Fig.5 are in agreement with the results of numerical integration of the original differential equation (41).

A comparison of the linear analysis ($\beta = 0$, Fig.4) with the nonlinear analysis ($\beta \neq 0$, Fig.5) shows qualitatively different results. In both analyses, a perturbation term of the form $\epsilon v_{30} x^3$ does not contribute to determining the existence of a limit cycle. Yet for β small, the nonlinear analysis does not reduce to the linear one. The linear analysis fails to predict one limit cycle in region IV of Fig.4 and two limit cycles in part of region II for β small.

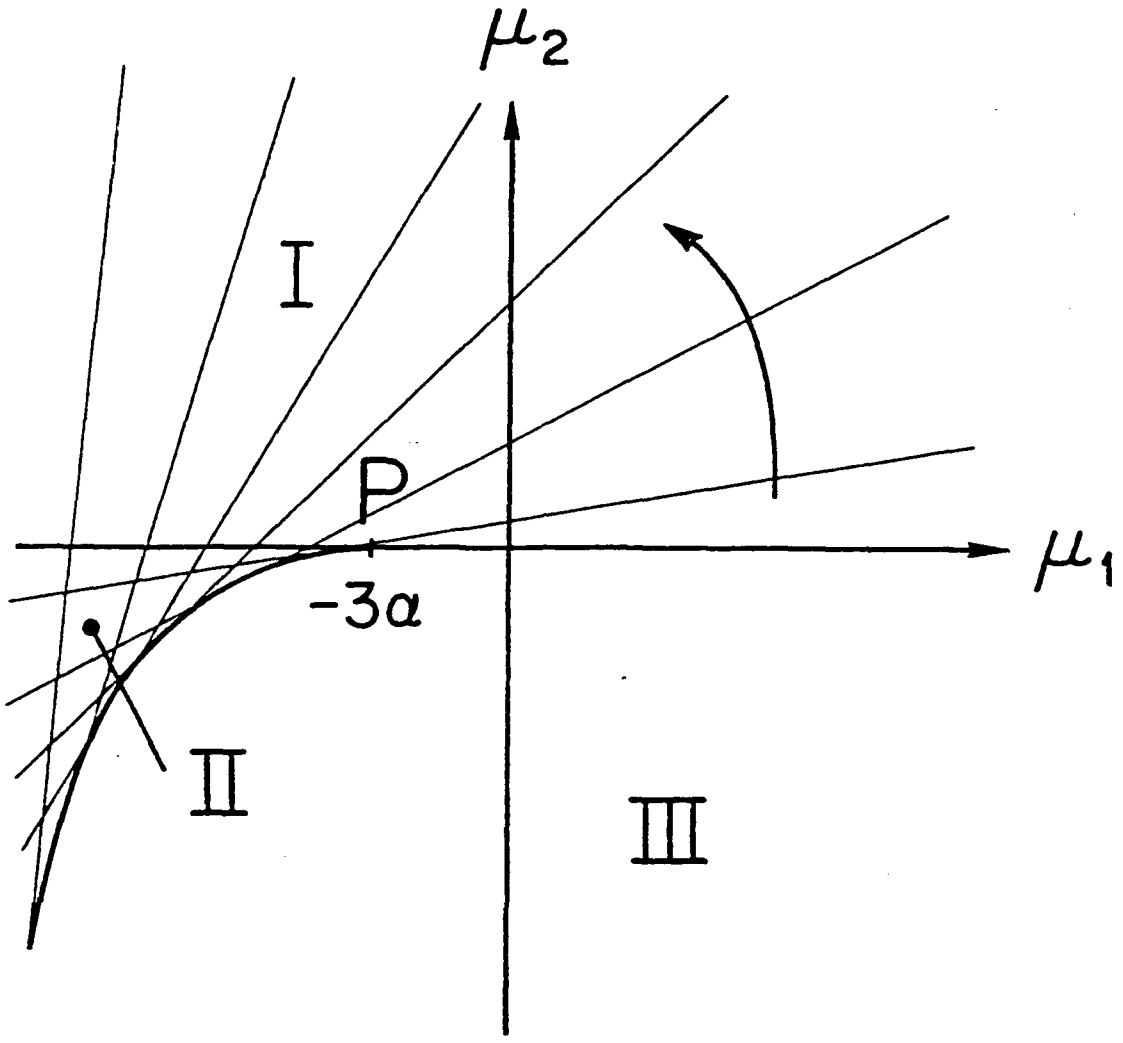


Fig.5. Limit cycles in eq.(41) for $\beta \neq 0$. The parameters μ_1 and μ_2 are defined by $\mu_1 = \rho/\eta$ and $\mu_2 = -\delta/\eta$, cf. eq.(52). Along each straight line there exists a limit cycle of fixed amplitude. Thus, in region I there exists 1 limit cycle; in region II there exists 2 limit cycles; and in region III there are no limit cycles. The μ_1 axis corresponds to the limiting case of a limit cycle of zero amplitude (and, hence, a Hopf bifurcation occurs as one crosses the μ_1 axis.). Along the curve separating region II from III two limit cycles coalesce. The arrow shows the direction of increasing limit cycle amplitude.

Numerical simulations confirm the nonlinear analysis. Eq.(1) provided an example with the following parameter values:

(54) $$\alpha = 1,\ \beta = \epsilon = 0.1,\ \delta = -0.5,\ \rho = -3.1,\ \eta = 1$$

in which the system belongs to region IV of Fig.4 and I of Fig.5. As we saw before, the analysis based on elliptic functions agreed with numerical integration, while the usual trigonometric approach failed to predict a limit cycle.

Another example is afforded by the parameter values:

(55) $$\alpha = 1,\ \beta = 2\ \epsilon = 0.1,\ \delta = 1,\ \rho = -4.6,\ \eta = 1$$

in which the system belongs to region II of Figs.4 and 5. A numerical simulation finds two limit cycles with amplitudes 1.93 and 2.93. Using eq.(52), the predicted values are 1.97 and 2.59, which compare well with the numerical integration values. The linear prediction eq.(48.2) predicts only one limit cycle with amplitude 1.58.

Conclusions

With the advent of computer algebra, perturbation analyses using elliptic functions can now be done almost as easily as those using trignometric functions. We have shown that perturbing off of elliptic functions will generally provide better quantitative and in some cases better qualitative results than a comparable perturbation off of trigonometric functions. In some problems, averaging off of elliptic functions (which contain an

amplitude-frequency dependence that trignometric functions lack) provides results at first order which can only be attained by averaging off of trigonometric functions to second order. In the case of limit cycles in eq.(41), first order trigonometric averaging gives qualitatively incorrect predictions if $\beta \neq 0$ and $\mu_1 < -3\alpha$, cf. Figs.4,5.

Related work in progress by the authors includes the extension of the averaging method off of elliptic functions to include terms of $O(\epsilon^2)$. This involves computing a near-identity transformation and is a generalization of second order averaging off of trigonometric functions (see [14].) Additional applications of the MACSYMA program have been made to the forced Duffing equation and to systems of the form of eq.(9) in which α and β are slowly varying functions of time. In particular, extensions of this work to problems in which α and β are not necessarily positive are in progress.

Acknowledgement

The authors wish to thank Dieter Armbruster, John Guckenheimer and Philip Holmes for valuable discussions. This work was partially supported by AFOSR and by ARO thru the Mathematical Sciences Institute at Cornell University.

Appendix. MACSYMA Computer Program Listing

```
/* ROUTINE TO PERTURB OFF X'' + AL X + 3E X^3 + E G(X,X') = 0  */

AVERAGE():=BLOCK([X,Y,XX,YY,EC,KC,AL,BE,G,F,FX2,FZ2,FI,FBAR,HI,D,CFLOW,PFLOW],
      PRINT("AVERAGING OF X'' + AL X + BE X^3 + EPS G(X,X',EPS*T"),
      PRINT(" "),AL:READ("ENTER AL:"),
      PRINT(" "),BE:READ("ENTER BE:"),
      PRINT(" "),PRINT("ENTER G(X,X') USING Y=X':"),
      G:READ(),
      PRINT(" "),PRINT("THE SOLUTION TO THE UNPERTURBED SYSTEM IS"),
      PRINT("X = C CN(4*KC(C)*PHI,K)"),
      PRINT("X' = C SQRT(AL + BE C^2) CN'(4*KC(C)*PHI,K)"),
      PRINT("WHERE 0 <= K^2 = BE C^2/2/(AL +BE C^2) <= 1/2"),
      PRINT("KC = COMPLETE ELLIPTIC INTEGRAL OF FIRST KIND"),
      PRINT("AND 4*KC(K)*PHI = SQRT(AL + BE C^2)*T+B"),PRINT(" "),
      PRINT("SEEK PERTURBED SOLUTION OF SAME FORM WHERE (C,PHI)"),
      PRINT("BECOME FUNCTIONS OF TIME"),
      PRINT(" "),

      /* X = C CN(4*KC*PHI)       */
      /* Y = X' = C SQRT(AL + BE C^2) CN'(4*KC*PHI)    */

      /* SYMBOLS */

      /* XX = CN FUNCTION                                          */
      /* YY = CN' FUNCTION (DERIVATIVE OF CN W.R.T. ARGUMENT) */
      /* ZZ = ZETA FUNCTION                                        */
      /* KC,EC = COMPLETE ELLIPTIC INTEGRALS OF 1ST,2ND KINDS */
      /* K  = MODULUS                                              */

      KILL(K),

      /* FOR SPECIAL CASES, K IS A NUMBER */

      IF AL = 0 THEN K:SQRT(1/2),
      IF BE = 0 THEN (K:0,KC:EC:%PI/2),

      /* REDUC ROUTINE TO REDUCE EXPRESSIONS TO FORMS:  CN^M  AND CN^M CNP */

      REDUC(EXPR):=BLOCK([EVEN,ODD,VAL],
            EVEN:EXPAND((EXPR+EV(EXPR,YY=-YY))/2),
            ODD:EXPAND((EXPR-EVEN)/YY),
            ODD:YY*EXPAND(EV(ODD,YY=SQRT((1-XX^2)*(1-K^2+K^2*XX^2)))),
            EVEN:EXPAND(EV(EVEN,YY=SQRT((1-XX^2)*(1-K^2+K^2*XX^2)))),
            VAL:EVEN+ODD
      ),
```

```
/* AVERAGING PROCEDURE */

G:EV(G,X=C*XX,Y=C*SQRT(AL+BE*C^2)*YY),

F[1]:-1/SQRT(AL+BE*C^2)*REDUC(G*YY),
F[2]:1/C/4/KC/SQRT(AL+BE*C^2)
            *REDUC(G*(XX-(1-2*K^2)/(1-K^2)*(ZZ*YY+K^2*XX*(1-XX^2)))),
IF K = 0 THEN F[2]:EV(F[2],ZZ=0),

F[1]:EV(F[1],YY=0),                          /* CN^M CNP TERMS HAVE NO MEAN */

FZ2:RATCOEF(F[2],ZZ),                        /* PICK OFF Z TERMS IN F[2] */
FX2:EXPAND(F[2]-FZ2*ZZ),                     /* PICK OFF X TERMS IN F[2] */
FZ2:EXPAND(EV(FZ2-EV(FZ2,YY=0),YY=1)),       /* Z CN^M TERMS HAVE NO MEAN */
FX2:EV(FX2,YY=0),                            /* CN^M CNP TERMS HAVE NO MEAN */

/* MEAN VALUE ROUTINE */

D[0]:1,
D[1]:0,
D[2]:1/KBAR^2*(EC/KC-1+KBAR^2),
D[3]:0,

D[II]:=RATSIMP(1/(II-1)/KBAR^2*((II-2)*(2*KBAR^2-1)*D[II-2]
               +(II-3)*(1-KBAR^2)*D[II-4])),

IF K = 0 THEN (D[2]:1/2,D[II]:=RATSIMP((II-1)/II*D[II-2])),
IF K = SQRT(1/2) THEN KBAR:SQRT(1/2),

/* FIND MEAN USING TABLE 4 */

HI:MAX(HIPOW(F[1],XX),HIPOW(FX2,XX),HIPOW(FZ2,XX)),
FOR II:1 THRU 2 DO FBAR[II]:0,
FOR II:0 THRU HI DO (
    FBAR[1]:FBAR[1]+RATCOEF(F[1],XX,II)*D[II],
    FBAR[2]:FBAR[2]+RATCOEF(FX2,XX,II)*D[II]
            -RATCOEF(FZ2,XX,II)/(II+1)*((1-KBAR^2-EC/KC)*D[II+1]
                                        +KBAR^2*D[II+3])
),

/* CHANGE RESULTS TO PRINTABLE FORM */

FOR II:1 THRU 2 DO FBAR[II]:EV(FBAR[II],ABS(C)=CBAR,C=CBAR,K=KBAR),

/* PRINT AVERAGED EQS */

CFLOW:EPS*FACTOR(FBAR[1]),
PFLOW:1/4/KC*EV(SQRT(AL+BE*CBAR^2),ABS(CBAR)=CBAR)+EPS*FACTOR(FBAR[2]),

DERIVABBREV:TRUE,KILL(KBAR),
VAL:[DIFF(CBAR(T),T)=CFLOW,DIFF(PHIBAR(T),T)=PFLOW,
     KBAR^2=BE*CBAR^2/2/(AL+BE*CBAR^2)],

PRINT("THE AVERAGED EQUATIONS ARE"),PRINT(" "),
PRINT(VAL),PRINT(" ")
)$
```

Here is a sample run based on the example discussed in the text, eq.(1):

```
(c6) AVERAGE()$

PERTURBATION OF X'' + AL X + BE X^3 + EPS G(X,X',EPS*T) = 0  BY AVERAGING

ENTER AL:
1;

ENTER BE:
1/10;

ENTER G(X,X') USING Y=X':
-Y/2-31*X^2*Y/10+Y^3;

THE SOLUTION TO THE UNPERTURBED SYSTEM IS
X = C CN(4*KC(C)*PHI,K) , X' = C SQRT(AL + BE C^2) CN'(4*KC(C)*PHI,K)
WHERE 0 <= K^2 = BE C^2/2/(AL +BE C^2) <= 1/2
KC = COMPLETE ELLIPTIC INTEGRAL OF FIRST KIND
AND 4*KC(K)*PHI = SQRT(AL + BE C^2)*T+B

SEEK PERTURBED SOLUTION OF SAME FORM WHERE (C,PHI) BECOME FUNCTIONS OF TIME

THE AVERAGED EQUATIONS ARE

                                    4     6              2     6
[cbar(t)  = - cbar eps (24 cbar  kbar  kc + 240 cbar  kbar  kc
        t

          4     4              2     4              4           4     2
 - 39 cbar  kbar  kc - 173 cbar  kbar  kc + 175 kbar  kc + 9 cbar  kbar  kc

          2     2              2           4             2
 - 561 cbar  kbar  kc - 175 kbar  kc + 6 cbar  kc + 494 cbar  kc

          4       6            2       6          4       4
 - 48 cbar  ec kbar  - 480 cbar  ec kbar  + 72 cbar  ec kbar

          2       4              4         4       2          2       2
 + 286 cbar  ec kbar  - 350 ec kbar  - 12 cbar  ec kbar  + 314 cbar  ec kbar

             2         4             2              4
 + 175 ec kbar  - 6 cbar  ec - 494 cbar  ec)/(1050 kbar  kc),

                      2
                   cbar
              sqrt(----- + 1)                    2
                    10                      cbar
phibar(t)  = ---------------, kbar  = --------------]
        t        4 kc                          2
                                            cbar
                                     20 (----- + 1)
                                          10

(VAX 8530 Time = 157 sec.)
```

The results of the program give the averaged equations in terms of both $\bar{C}$ (called cbar) and $\bar{k}$ (called kbar). The results are stored in the variable VAL: VAL[1] contains the $\bar{C}'$ equation, VAL[2] contains the $\bar{\varphi}'$ equation and VAL[3] contains the expression for $\bar{k}^2$ in terms of $\bar{C}^2$. The following command substitutes $\bar{k}$ in terms of $\bar{C}$, giving eq.(39) of the text:

```
(c7) FACTOR(EV(VAL[1],KBAR=SQRT(RHS(VAL[3]))));
```

$$(d7)\ \mathrm{cbar}(t)_t = -\mathrm{eps}\,(5\,\mathrm{cbar}^6\,\mathrm{kc} + 447\,\mathrm{cbar}^4\,\mathrm{kc} + 10175\,\mathrm{cbar}^2\,\mathrm{kc} + 64700\,\mathrm{kc} - 594\,\mathrm{cbar}^4\,\mathrm{ec} - 11880\,\mathrm{cbar}^2\,\mathrm{ec} - 64700\,\mathrm{ec})/(350\,\mathrm{cbar}\,\mathrm{kc})$$

(VAX 8530 Time = 3 sec.)

References

1. Byrd, P. and Friedman, M. Handbook of Elliptic Integrals for Engineers and Scientists, Springer-Verlag, Berlin, 1954.

2. Cap, F. F. "Averaging Methods for the Solution of Non-linear Differential Equations with Periodic Non-harmonic Solutions." *International Journal of Non-Linear Mechanics*, vol. 9, 1973, pp.441-450.

3. Chirikov, B. V. "A Universal Instability of Many Dimensioanl Oscillator Systems." *Physics Reports*, vol. 52, 1979, pp.263-379.

4. Davis, H. T. Introduction to Nonlinear Differential and Integral Equations, Dover, NY, 1962.

5. Goldstein, H. Classical Mechanics, second edition, Addison-Wesley, Reading, Mass., 1980.

6. Greenspan, B. and Holmes, P. "Repeated Resonance and Homoclinic Bifurcation in a Periodically Forced Family of Oscillators." *SIAM Journal of Mathematical Analysis*, vol. 15, no. 1, 1983, pp. 69-97.

7. Guckenheimer, J. and Holmes, P. Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields, Applied Mathematical Sciences vol. 42, Springer-Verlag, NY, 1983.

8. Hagedorn, P. Non-Linear Oscillations, Oxford University Press, NY, 1982.

9. Kervokian, J. and Cole, J. D. Perturbation Methods in Applied Mathematics, Applied Mathematical Sciences vol. 34, Springer-Verlag, NY, 1981.

10. Kuzmak, G. E. "Asymptotic Solutions of Nonlinear Second Order Differential Equations with Varaiable Coefficients." *P.M.M.* (English translation), vol. 23, no. 3, 1959, pp. 515-526.

11. Minorsky, N. Nonlinear Oscillations, Van Nostrand, NY, 1962.

12. Nayfeh, A. Perturbation Methods, Wiley-Interscience, NY, 1973.

13. Nayfeh, A. and Mook, D.T. Nonlinear Oscillations, John Wiley and Sons, NY, 1979.

14. Rand, R. H. and Armbruster, D. Perturbation Methods, Bifurcation Theory and Computer Algebra, Applied Mathematical Sciences vol. 65, Springer-Verlag, NY, 1987.

15. Sanders, J. A. and Verhulst, F. Averaging Methods in Nonlinear Dynamical Systems, Applied Mathematical Sciences vol.59, Springer-Verlag, NY, 1985.

16. Stoker, J. J. Nonlinear Vibrations, Wiley, NY, 1950.

The Effective Use of Computer Algebra Systems

Joel S. Cohen
Department of Mathematics and Computer Science
University of Denver
Denver, Colorado 80208

Abstract

In this paper we give an outline for an applied computer algebra course which describes the technical skills needed to effectively use a computer algebra system to solve symbolic mathematical problems in science and engineering.

1 Introduction.

A Computer Algebra System (CAS) is a powerful computer program which is able to manipulate and analyze symbolic mathematical expressions. Computer algebra systems are quite easy to use and it is easy for both students and professionals to learn to use the systems in a superficial way. For example, to compute an indefinite integral or find the closed form solution to an elementary differential equation, one needs to master only a few simple operations. However, to really understand the potential uses (and limitations) of these systems, a mathematical scientist [1] must have:

- An understanding of the kind of mathematical knowledge contained in a CAS and an understanding of the extent and reliability of this knowledge.
- Some knowledge of computer algebra programming techniques including recursion and list manipulation.
- The ability to formulate a symbolic mathematical problem in an algorithmic way and the ability to express the algorithm in terms of the mathematical operations and programming structures available in a computer algebra language.

[1] We use the term *mathematical scientist* as a generic term to represent mathematicians, computer scientists, physical scientists, engineers, statisticians, economists and others who use mathematical reasoning in their work.

- Some understanding about how a CAS works.
- An exposure to a collection of examples that illustrate the successful use of a CAS.
- A feeling for which symbolic calculations are best done by hand and which are best done by a CAS.

In this paper, we shall outline a course in applied computer algebra. The course outline describes the technical skills needed to effectively use a CAS to solve symbolic mathematical problems in science and engineering.

2 A Frustrating Example.

Let us begin by taking a critical look at an example that illustrates the use of a CAS. The problem we have chosen is from modern physics. It emphasizes many of the issues faced by the mathematical scientist who wishes to use a CAS to help solve a problem.

Consider the time dependent Schroedinger equation for the one electron atom

$$-\frac{\hbar^2}{2m_1}\left(\frac{\partial^2\psi_T}{\partial x_1^2}+\frac{\partial^2\psi_T}{\partial y_1^2}+\frac{\partial^2\psi_T}{\partial z_1^2}\right)-\frac{\hbar^2}{2m_2}\left(\frac{\partial^2\psi_T}{\partial x_2^2}+\frac{\partial^2\psi_T}{\partial y_2^2}+\frac{\partial^2\psi_T}{\partial z_2^2}\right)+V\psi_T=E_T\psi_T \tag{1}$$

where $\psi_T = \psi_T(x_1, x_2, y_1, y_2, z_1, z_2)$. If we make the following change of variables

$$\begin{aligned} x &= \frac{m_1x_1+m_2x_2}{m_1+m_2} \\ y &= \frac{m_1y_1+m_2y_2}{m_1+m_2} \\ z &= \frac{m_1z_1+m_2z_2}{m_1+m_2} \\ r &= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2} \\ \phi &= \arctan\left(\frac{y_2-y_1}{x_1-x_2}\right) \\ \theta &= \arccos\left(\frac{z_2-z_1}{r}\right), \end{aligned} \tag{2}$$

equation (1) is transformed to the form

$$\begin{aligned} &-\frac{\hbar^2}{2\,(m_1+m_2)}\left(\frac{\partial^2\psi_T}{\partial x^2}+\frac{\partial^2\psi_T}{\partial y^2}+\frac{\partial^2\psi_T}{\partial z^2}\right)-\frac{\hbar^2}{2\mu}\left\{\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi_T}{\partial r}\right)\right. \\ &\left.+\frac{1}{r^2\sin^2(\theta)}\frac{\partial^2\psi_T}{\partial\phi^2}+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi_T}{\partial\theta}\right)\right\}+V\psi_T=E_T\psi_T, \end{aligned} \tag{3}$$

where

$$\mu = \frac{m_1 m_2}{m_1 + m_2}. \tag{4}$$

The transformation appears in the text Eisberg([5], pp. 295-297) where the author remarks, "This is actually a quite tedious task, and so we present here only the results." This example is similar to the polar and spherical transformations of the Laplacian and other vector analysis operators which appear in texts on advanced engineering mathematics. This problem is more involved since there are six independent variables.

A number of years ago the author attempted to verify this transformation using a CAS. The point of the exercise was to illustrate to a class of mathematics, engineering and physical science students the power of a CAS to handle tedious but routine calculations. This problem is typical of the type of symbolic calculations a scientist or engineer might encounter in his work. Let us suppose this person is lucky enough to have a powerful work station with computer algebra software in the office and has had some experience with a CAS. Let us also suppose that he needs to verify this transformation and must make a decision whether to do this with pencil and paper or with a CAS. A number of questions of a philosophical nature come to mind:

- *Can this calculation be done by a CAS?* Since this calculation is straightforward, the answer is presumably "yes."

- *Should this calculation be done by a CAS or with paper and pencil?* There are number of important considerations here. First, there is the question of time and effort. Will using a CAS require less time than hand calculation? Next, there is a question of accuracy. This is particularly important if the final answer to the problem is unknown. Presumably, if the problem is accurately entered into the computer, if the user has chosen the appropriate commands, and if the CAS is free of bugs, the CAS should produce an accurate result. Finally, will important side effects of a pencil and paper calculation be lost? In the course of a derivation other important relationships often appear which would not be apparent if the calculation were done with a CAS.

- *Can this problem be done by a novice symbolic programmer or is expert knowledge of a CAS required?* Presumably, our hypothetical mathematical scientist is closer to the novice than the expert and is more interested in obtaining insight into the scientific problem than learning about computer algebra.

- *Is this calculation a direct application of the mathematical knowledge in a CAS or does it require a symbolic program which might include conditional statements, loops and subroutines?* Ideally, one would like to have a single command which takes (1) and (2) as input data, and returns (3) as a

result. If this command is not available, we must combine a number of commands into a symbolic program.

- *What algorithm should be used to perform the calculation? Can we imitate the steps found in a textbook?* A starting point for an algorithm is the transformation of the Laplacian from rectangular to polar coordinates found in many advanced engineering text books. (for example, see Kreyszig[8], pages 447-448). However, most textbook derivations are written to facilitate human understanding. They are not written with a computer algebra system in mind. For example, the Laplacian derivation cited above contains many local substitutions which make it understandable to the human reader. Is it necessary to follow the same approach with a CAS derivation?

- *How does one express the algorithm in terms of the operations and data structures available in a computer algebra language?* In the example considered here, the primary question is how to deal with the undefined function ψ_T. Most (but not all) computer algebra systems have two ways to express undefined relationships between variables. The first way represents an undefined relationship explicitly with an expression similar to $f(x, y)$. The other way declares that the symbol f depends on the symbols x and y, and then carries out all calculations with derivatives in this environment. Will both representations produce the desired result?

- *Can this calculation be done with any CAS? Will the same algorithm work with all systems?* Perhaps the mathematical scientist is wondering which of many available systems should be used. At the time this problem was tried, the author had these four systems available: MACSYMA[13], REDUCE[6], MAPLE[1], and muMath[10].

- *What input is required for this calculation?* Obviously, one has to input the original equation (1) and the transformation (2). However, the inverse transformation which expresses the variables $x_1, x_2, y_1, y_2, z_1, z_2$ in terms of the variables x, y, z, r, ϕ, θ is also required. Although, computer algebra systems are able to solve some systems of nonlinear equations, no CAS is able to invert the transformation (2). Therefore, the inverse transformation must also be input into the system.

- *Will a CAS produce the intended result or will the result appear in a mathematically equivalent form?* Although a CAS can perform some remarkable simplifications, it is often difficult to transform an expression to the exact form found in a textbook.

- *Is trigonometric simplification possible with a CAS?* Theoretically, it is impossible to do all the simplifications we may wish to do with a CAS. In this

case, however, the problem is primarily one of trigonometric simplification which involves repeated application of the identity $\sin^2 u + \cos^2 u = 1$.

- *How much CPU time will the calculation require?* Obviously, we should only perform the calculation if it is within the capabilities of our equipment.

- *How will we know if the answer is correct?* Since the problem is one of verification, we know what we are looking for. If we had not known the result, we could apply the inverse transformation to the result to see if we could obtain the original equation. In general, determining if a result is correct poses a difficult problem.

In retrospect, the transformation of (1) to (3) is not a particularly difficult calculation for a CAS. However, it would have been difficult to convince the author of this when he was trying to verify the result with a CAS. The question is not so much whether someone with an intimate knowledge of some CAS can obtain the result by pulling a few commands out of a hat. Rather, the question is whether a person with moderate knowledge of a CAS can obtain the result in a reasonable amount of time.

It would be interesting (in fact comical!) to review the path (including all the false starts) taken by the author to solve this problem with a CAS. We shall not go into all the technical details in this paper. Rather, we shall give some overall impressions of the experience.

- A preliminary analysis of the capabilities of the four systems indicated that the MACSYMA system had the greatest chance of success with this particular problem. Although it is not apparent from reading the manuals, the other three systems did not have the capability to apply the differentiation chain rule to abstract (undefined) functions [2]. Since this rule is needed in the derivation, these systems were dropped from consideration. Certainly, this limitation may be eliminated from these systems by modifying or even rewriting the differentiator.

- The problem took about one and a half weeks of the author's time and over 30 test files before a correct program was obtained.

- Although it was relatively easy to write a program to transform a two dimensional version of the problem, it was not so easy to modify the result for three dimensions, In fact, the obvious generalization of the two dimensional program to three dimensions originally produced a result which was nearly correct but contained a few superfluous terms which did not simplify to zero. Luckily, in this case, the final result was known. Otherwise this incorrect result may have been accepted as correct!

[2] In fact none of the systems handle the differentiation and integration of undefined functions in a satisfactory manner. See Cohen[2] and Wester[14].

- Originally, we ignored many of the local substitutions and simplifications found in textbook derivations for problems of this sort. Instead, we relied on a brute force approach which required an unacceptable amount of CPU time (up to four hours on a VAX 750). By performing some local substitutions we were eventually able to get the derivation down to 15 minutes of CPU time.

- To effectively use a CAS, it is essential to thoroughly understand the semantics of commands in the system . Unfortunately, a precise description of a command's function rarely appears in the system manual. In the course of trying the above example, significant differences were observed in the same command from system to system. A program's behavior in one system may be quite different from its behavior in another system.

3 A Course In Applied Computer Algebra.

We believe that the CAS experience described above is not unique. There is more to using a CAS than reading the manual, seeing a few examples and trying a few commands. Numerical analysts have always emphasized that using a numerical method in an inappropriate way can lead to disastrous results. We believe the same is true for symbolic methods. In the remainder of the paper, we describe an applied computer algebra course which is intended to address some of these issues.

There are currently many CAS courses being taught in the U.S., Canada and Europe. The range is from language courses, which describe how to use a particular CAS system to solve a variety of symbolic mathematical problems, to more advanced courses which concentrate on the mathematical background needed to develop efficient algorithms for computer algebra. To make an analogy with numerical computation, the former courses are similar to scientific programming courses which teach the mechanics of a programming language (FORTRAN or Pascal), and, in some cases, the use of numerical or statistical software packages. The advanced CAS course is similar to a course in numerical analysis which includes a theoretical discussion of numerical algorithms. In the numerical setting, numerical methods courses, which lie between these two extremes, serve to introduce a mathematical scientist to some of the issues and applications of numerical computation (for example, see Conte and deBoor[3] or Dorn [4]). These courses are taught in mathematics departments, computer science departments, and other science and engineering departments. The goal is to introduce the issues surrounding numerical computation and to emphasize its advantages and pitfalls. The applied computer algebra course we have in mind is the computer algebra analogue to the numerical methods course.

The following four premises underly the design of the course:

1. *The course is designed for mathematical scientists whose primary interests*

are not in computer algebra or even computer science. This population has an extremely diverse background in both mathematics and computing. A safe lower bound for backgrounds is to assume the usual two year freshman-sophomore mathematics sequence (through multivariable calculus, linear algebra and applied differential equations), plus some experience with numerical programming. Many in the audience will not have studied discrete mathematics, abstract algebra or programming concepts such as recursion and list manipulation. This limits the set of examples that can be used to demonstrate the capabilities of a CAS. It also means that mathematical concepts from these areas which are needed to write symbolic programs must be integrated into course material.

2. *The course is oriented toward algorithms rather than a particular CAS language.* Although the notion of an algorithm for a symbolic computation is implicit in traditional mathematics, it is not usually the focus of the subject. Techniques for solving a problem are usually not stated in the formal way that they are in numerical methods or computer programming. For example, although differentiation is formally a recursive process, it is usually described in an informal way in a mathematics textbook. Rather than emphasize a particular CAS language, the goal should be to develop the skills needed to create symbolic algorithms.

   Of course, it is important to include some programming in one or more CAS languages. However, computer algebra systems are evolving rapidly and new systems are being developed [3]. Rather than learn all the details and eccentricities of a particular language, it is more useful to learn the general principles of CAS languages and an approach to evaluate the capabilities of a particular language.

3. *Examples of CAS applications should be chosen to illustrate both the possibilities and limitations of computer algebra and should emphasize the role of computer algebra in the problem solving process.* The message of most written material currently available on computer symbolic computation tries to promote the field rather than give a balanced, realistic view of where a CAS can be used. In solving mathematical problems, a CAS is only of many tools that can help solve a problem. The important question is, what is the place for computer symbolic computation in the problem solving process? What is the role for the mathematical scientist and what is the role for the machine?

4. *Pencil and paper symbolic calculation is more important than ever.* Occasionally, it is said that computer algebra will revolutionalize the way we do mathematics by eliminating the need for much symbolic computation with

[3] During the past year, two new CAS systems, MATHEMATICA (see Wolfram[15]) and DERIVE([12]), have been introduced.

pencil and paper. We believe that this statement is misleading and gives a false picture of the role that computer algebra can play in the problem solving process. A CAS can perform many symbolic calculations which are ordinarily done by hand. However, we believe to effectively use a CAS, the mathematical scientist must be good at pencil and paper symbolic calculation and have a good understanding of the underlying mathematical concepts. This understanding is essential to recognizing situations where a CAS can be applied.

There is another more subtle reason to emphasize the importance of hand calculation. For some people there is a tendency to immediately try a computation with a CAS with the hope that the system will miraculously produce the intended result. Using a CAS in this way can be counterproductive. A more useful approach is to spend some time thinking about a problem with pencil and paper. Perhaps the problem can be put into a more convenient form which leads to a transparent solution or which provides some unexpected information.

Course Organization.

There are three important components to an applied computer algebra course:

1. An exploration of the capabilities of a CAS.
2. A discussion of symbolic programming techniques including recursion and list manipulation.
3. A discussion of some of the elementary algorithms which make a CAS work.

We shall discuss each component in greater detail in the following sections.

4 Exploring The Capabilities of a CAS.

Many mathematical scientists do not use a CAS because they do not understand its capabilities or believe it can be useful in their work. It is easy to understand where this feeling comes from. The manipulation of mathematical expressions is a difficult intellectual exercise which often requires insight as well as perseverance. Indeed, even those who have considerable experience with a CAS may find it difficult to describe what it can do or when it might be useful for a particular problem. There are some who even take the point of view that it is too difficult to precisely define what a CAS system does. They believe a user should simply try a problem with a CAS and see what happens.

The mathematical knowledge in a CAS is defined by the properties of the various mathematical operators in the system (expand, factor, limit, differentiate, integrate etc.) and the automatic simplification rules which are applied to

an expression. The capabilities of most mathematical operators can vary from system to system (sometimes dramatically) and may change significantly when a new version of a system is introduced. For example, some CAS systems have the capability to compute the limit of a function or a sequence. In the manual which accompanies the CAS, the semantic action of the *limit* operator is usually loosely described rather than precisely defined. This description will include a few isolated examples of how the operator is applied but little information about what to expect in non-trivial examples. If one were to scan through a typical text on applied mathematics, one would find that the limit operation is used in many different contexts — some very specific and some quite abstract. For which limit operations should we expect that a CAS will produce a reasonable result?

Fig. 1 shows the results of applying two computer algebra systems [4] to some limit problems encountered in undergraduate mathematics. Example 1 requires two applications of L'hospital's rule and is easily computed by both systems. Example 2 is similar but requires n applications of L'hospital's rule where n is undefined. Nevertheless, both systems are able to compute the result. However, MACSYMA requested additional information about x and n which in this case is extraneous. The limit in Example 3 requires another approach. The expression is the absolute value of the nth term of the series

$$\sum_{n=0}^{\infty} \frac{x^n}{n!}$$

which converges by the ratio test. Therefore,

$$\lim_{x \to \infty} \frac{x^n}{n!} = 0.$$

Neither system was able to compute this limit. Example 4 is the formal definition of the derivative for $\sin x$. Both systems are able to compute this limit but MACSYMA requested additional but extraneous information about the sin function. Example 5 is a similar calculation (the transformation $s = 1/t$ transforms the limit to the derivative of $\exp(x)$ at $x = 0$) but surprisingly MACSYMA is unable to compute the limit. Example 6 is the general definition of the derivative. However, neither system recognizes this fact even though both systems have some capability to work with undefined functions. Example 7 is a famous result due to Euler. The author was quite surprised to find that MAPLE can compute this limit. Example 8 is the Laplace transform of $\sin \omega t$. MACSYMA is able to compute this limit by requesting information about the sign of s. MAPLE was unable to compute the limit.

What is a user to make of this performance? Certainly all these limits may arise in the course of doing mathematics and are fair requests of a CAS. If the

[4] The systems are MACSYMA 309.6 and MAPLE 4.1. MACSYMA is a trademark of Symbolics Inc.

	Limit	MACSYMA	MAPLE
1	$\lim_{x\to\infty} \frac{x^2}{e^x} = 0$	0	0
2	$\lim_{x\to\infty} \frac{x^n}{e^x} = 0$	0	0
3	$\lim_{n\to\infty} \frac{\lvert x^n\rvert}{n!} = 0$	U	U
4	$\lim_{\Delta x\to 0} \frac{\sin(x+\Delta x)-\sin x}{\Delta x} = \cos x$	$\cos x$	$\cos x$
5	$\lim_{t\to\infty} t(\exp 1/t - 1) = 1$	U	1
6	$\lim_{\Delta x\to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x} = f'(x)$	U	U
7	$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j^2} = \frac{\pi^2}{6}$	U	$\frac{\pi^2}{6}$
8	$\lim_{x\to\infty} \int_0^x \sin\omega t \exp(-st)\, dt = \frac{\omega}{s^2+\omega^2},\ s>0$	$\frac{\omega}{s^2+\omega^2}$	U

Figure 1: Examples of limit operations in two computer algebra systems. The symbol U means "unable to compute."

person is familiar with the computer algebra field and has kept up with the work on limits, he might know about the difficulties which arise when programming a CAS to compute limits, and thus thus be willing to excuse the system when it fails to produce a result. Most users will not have this knowledge and may end up wondering just what a system can do and whether to trust the results. To effectively use a CAS, a user must have a clear idea about the semantic capabilities of mathematical operators in the system. The exploration of the semantic capabilities of one or more CAS systems is an important component of an applied computer algebra course.

We use the phrase *semantic capacity* to refer to the mathematical power of an operator in a computer algebra system. In many cases it is difficult to precisely define the semantic capacity since even the developer of a CAS may not know exactly what an operator can and cannot do. Nevertheless, we believe that this is an important question and it should be discussed even if it cannot be completely answered.

We do not believe it is practical or interesting to simply list in detail the capabilities of an operator in a CAS. The capacity of an operator may change when new versions of a CAS are released and it is difficult for a casual user to keep details of this sort in mind. An alternative approach is to discuss what a mathematical operator should *ideally* be able to do and compare this in a general way to what current systems are capable of doing. For example, consider the differentiation operator. Obviously, a CAS should be able to differentiate most specific functions no matter how complicated. However, in mathematical

calculation we also differentiate implicit functions, functions defined by integrals (including differentiation under the integral sign), functions defined by infinite series, and undefined functions (general $f(x)$, $g(x,y)$). We differentiate with respect to a variable, a function and even a differentiation symbol such as y'. We compute ordinary derivatives, partial derivatives and total derivatives. Is a CAS able to perform the differentiation operation correctly in all these cases? A carefully chosen list of exercises can help a user explore the semantic capacity of an operator.

Properly Posed Requests.

As with all forms of programming, the mathematical scientist should take care to ensure that the input to the system makes mathematical and computational sense. Informally speaking, an operation is said to be *properly posed* if:

- *The execution of the operation produces a meaningful mathematical expression.* It is quite easy to get a CAS to return absurd looking results. Therefore, the mathematical scientist must always carefully inspect the result returned by the system to ensure that it makes mathematical sense and satisfies all the explicit (and implicit) assumptions in a problem.
- *The CAS has all the information needed to perform the operation in an unambiguous manner.* A CAS is sometimes asked to perform a mathematical operation without all the necessary information. In this case, the system may not perform the operation, may request additional information, or may even return a result which is not entirely correct. To obtain a satisfactory result, the mathematical scientist must supply additional information to the system or modify the request to remove the ambiguity.

Fig. 2 is a MACSYMA session which illustrates a few instances of operations which are not properly posed. Line c1 assigns an equation to a variable d1. Line c2 substitutes the value $x = -3$ into d1. In response, MACSYMA returns the expression d2. We consider this substitution operation to be improperly posed since it returns an absurd expression even though it is perfectly legal in MACSYMA. The problem here is with the ambiguous use of the equal sign, which has a number of different uses in mathematics and therefore a number of different uses in a CAS. In the present context, no semantic meaning is assigned to the equality and the system is not aware that the last result is incorrect. A similar computation can by done in both MAPLE and muMath. The REDUCE system does not accept the substitution and returns an uninformative error message.

In statement c3 we request the system to evaluate the indefinite integral

$$\int x^n dx$$

```
(c1)  x^2+4=x-1;

                     2
(d1)                x  + 4 = x - 1

(c2) subst(x=-3,d1);

(d2)                       13 = - 4

(c3) integrate(x^n,x);

Is  n + 1  zero or nonzero?
nonzero;
                          n + 1
                         x
(d3)                     ------
                         n + 1
(c4) integrate(sin(omega*t)*exp(-s*t),t,0,inf);

Is  omega  positive, negative, or zero?

positive;

Is  s  positive, negative, or zero?

positive;
                           omega
(d4)                   -----------
                        2        2
                       s  + omega
```

Figure 2: A MACSYMA session demonstrating statements that are *improperly posed.*

where n does not depend on x but is otherwise undefined. This statment is improperly posed since the result depends on whether or not $n = -1$, which is unknown at this point. In this case, MACSYMA queries the user for more information about the value of n. We have informed the system that $n + 1 \neq 0$ and the system returns the appropriate result d3. The computation was tried on three other computer algebra systems (muMath, MAPLE and REDUCE). Each automatically assumed $n \neq -1$ and returned $x^n/(n+1)$.

In the next statement (c4), we ask MACSYMA to evaluate the integral

$$\int_0^\infty \sin \omega t \; e^{-st} dt.$$

This improper integral represents the Laplace transform of the function $\sin \omega t$. It converges when $s > 0$ and otherwise does not converge. Without information about the relationship between s and zero, the statement is improperly posed. MACSYMA realizes this fact and requests more information about s. MACSYMA also requests unnecessary information about the sign of *omega*. In both cases, we indicate that the variables are positive and MACSYMA returns the result (d4). The muMath, MAPLE and REDUCE systems are unable to evaluate this improper integral.

The question of when an operation is properly posed is an important aspect of operator capacity. It is important to understand that a CAS must occasionally make assumptions about the nature of variables in an expression and that the result produced by a system may not be correct in all situations.

Simplification Context.

For efficiency reasons it is unreasonable to expect a CAS to apply all its simplification rules during the course of a computation. The designer of a CAS must choose which simplification rules are appropriate for a particular operation. We use the term *simplification context* to refer to those simplification rules which are applied during the evaluation of a mathematical operator. The simplification context often determines the form of the output of an operator and in some cases determines whether or not a CAS can even perform an operation.

For a simple example, consider the MAPLE session in Fig. 3. At the first prompt, u is assigned a polynomial in x with coefficients which are polynomials in a. At the second prompt, we request that MAPLE determine the degree of u in x. The system returns the value 2 even though the the coefficient of x^2 simplifies to 0. In this example, the *expand* simplification rules apparently are not part of the simplification context of the degree operator.

The simplification context will often determine whether or not a CAS is able to perform an operation. For example, consider the indefinite integration

$$\int 2x \cos(x^2 + x) + cos(x^2 + x) \; dx = sin(x^2 + x) + C. \qquad (5)$$

```
> u := (a^2-1-(a+1)*(a-1))*x^2 + 2*x + 3;

                                  2                          2
                u := (a  - 1 - (a  - 1)(a - 1)) x + 2 x + 3

> degree(u,x);

                                  2
```

Figure 3: A MAPLE session demonstrating a *simplification context.*

If the integrand is first factored, the integral can be easily evaluated by making the substitution $u = x^2 + x$. In factored form, the four systems available to the author (MACSYMA, MAPLE, REDUCE and muMath) are able to evaluate the integral. However, in the form (5), only REDUCE was able to evaluate the integral. Apparently, factorization is part of the simplification context of the integration operator in REDUCE but not part of the simplification context of the operator in the other three systems.

Experiments with these computer algebra systems have shown that it can be difficult to determine exactly which simplification rules are applied during the the evaluation of an operator and at which point of the computation the rules are applied. Nevertheless, the simplification context is an important aspect of operator capacity and should be raised as an issue even if it can not be evaluated precisely.

Simplification of Mathematical Expressions With A CAS.

An important application of computer algebra is the simplification of involved mathematical expressions. It is easy to give examples which are difficult to simplify by hand but can be routinely simplified with a CAS(see [11]). Unfortunately, it is also easy to find simplifications which a CAS is unable to do. For most people, this is not particularly surprising since simplification with pencil and paper often requires considerable insight, clever substitutions and application of involved identities.

Although many simplifications involve concepts which most mathematical scientists consider elementary (factorization, expansion, trigonometric identities etc.), the simplification process is computationally quite complex. In fact, it is possible to show that it is theoretically impossible to find an algorithm that can perform all the simplifications we might hope to do with a CAS [5]. To complicate matters, the goal of simplification is difficult to define precisely. What is simple

[5] In the language of computer science, the simplification problem is recursively undecidable. For a discussion of this theorem see the paper by Moses[9]. This paper contains an interesting discussion of the simplification problem.

to one person may not be simple to another.

For a mathematical scientist the important question is "which simplifications are possible with CAS?" This question, which is not an easy one to answer, helps put in context the role of a CAS in the problem solving process. While a CAS usually cannot discover the sequence of steps needed for an involved derivation of a mathematical result, it can do many of the local simplifications which are encountered in the course the derivation. By combining a number of simplification commands, it is sometimes possible to write a program to verify all the details of an involved derivation.

Generally speaking, a CAS is good at simplifying expressions that contain explicit forms of elementary functions and involve complicated but straightforward manipulations. Most systems contain commands for expansion, some forms of substitution, simplification of rational expressions, radical simplification and transcendental simplification. If the simplification is not straightforward these systems are less useful. For example, none of the systems available to the author is able to perform the simplification

$$\log\tan(x+\frac{\pi}{4}) - \operatorname{arcsinh}(\tan 2x) = 0, \qquad -\frac{\pi}{4} < x < \frac{\pi}{4}$$

which requires a number of different transformations. In addition, simplification of expressions which contain indefinite sums or series (using the $\sum$ symbol or ellipses) and other expressions of a more abstract character are usually not possible with a CAS. For example, the MACSYMA system is able to solve the Bessel differential equation

$$x^2y'' + xy' + (x^2 - p^2)y = 0 \tag{6}$$

in terms of the Bessel functions $J_p(x)$ and $Y_p(x)$ or in terms of infinite series representations or these functions but is unable to perform the simplification which shows that the series

$$J_P(x) = \sum_{k=0}^{\infty} \frac{(-1)^k(\frac{x}{2})^{2k+p}}{k!(k+p)!} \tag{7}$$

satisfies differential equation (6) [6].

Each new user brings to a CAS a conceptual model of simplification which is based on experience with pencil and paper calculations. For many people this

[6] One must always be careful when making a claim that a particular CAS system is unable to carry out a certain computation. Computer algebra systems are complicated programs with many hundreds of commands and it is difficult for a person to know exactly what a CAS can and cannot do. When we say a CAS cannot perform a simplification we mean the author was unable to find one or two general purpose commands which are able to perform this simplification. Of course, it is possible to write a program with a large number of commands to perform this particular simplification. However, if one must go to all this trouble, the simplification might just as well be done with pencil and paper.

model may diverge radically from what is currently possible with a CAS. For example, some mathematical scientists may consider the simplification mentioned above which involves the series (7) fair game for a CAS and wonder why such a simple manipulation cannot be done. Indeed, glancing through an applied mathematics textbook, one finds many manipulations involving more general expressions of this type which also cannot be done with a CAS.

In order to develop confidence in the capabilities of a CAS, it is important to realistically assess what types of simplifications a CAS is able to do. A good starting point for selecting examples are the symbolic calculations found in elementary textbooks on trigonometry, algebra, calculus and differential equations [7]. Clearly, not all of these manipulations are appropriate for a CAS. By presenting these examples, we examine the types of manipulations a CAS is likely to encounter and raise important questions about the capabilities of these systems. In addition, we develop a connection between hand calculation and calculation with a CAS.

The Nature of Mathematical Knowledge in a CAS

The view of mathematics programmed into current computer algebra systems is reminiscent of the approach taken by mathematicians in the eighteenth century [8]. Like the mathematicians of that time, a CAS views the concepts of calculus primarily as an extension of the formal rules of algebra. Computer algebra systems have almost no knowledge about the underlying concepts of calculus such as rational and irrational numbers, the meaning of limits, the continuity of a function, the derivative of a function in terms of limits, the relationship between integrals and areas, or the convergence of infinite series. For example, in a CAS, the derivative is defined by a collection of transformation rules instead of with the limit definition [9].

For the most part, this lack of analytical knowledge does not hamper the use of a CAS. However, it does mean a CAS can occasionally produce an unexpected result. For example, consider the MACSYMA session in Fig. 4. At line c1, we ask MACSYMA to compute the power series representation for the expression $1/(1+x)$. The result is returned in d1 where the subscript "i1" has been generated by the system for this expression. At line c2, we ask MACSYMA to substitute $x = 1$ into both sides of the expression and simplify the sum. MAC-

[7] Knuth[7] illustrates some of the different types of manipulations found in mathematical reasoning.

[8] This refers to the user's view. Many of the algorithms which make a CAS work are part of twentieth century mathematics and computer science. Nevertheless, the primary applications of symbolic manipulation systems are to mathematics developed in the eighteenth and nineteenth centuries.

[9] However, it is interesting to note, that it is possible to find the derivative of most functions in a CAS using the limit definition. This is not because the system has an understanding of the limit process. It follows instead from the transformation rules for limits which use L'hospital's rule or a similar construction using derivatives for the calculation of limits involving indeterminate forms.

```
(c1) 1/(1+x)=powerseries(1/(1+x),x,0);

                                    inf
                                    ====
                              1     \          i1    i1
(d1)                        ----- =  >     (- 1)     x
                            x + 1   /
                                    ====
                                    i1 = 0

(c2) ev(d1,x=1,simpsum);
                                 1
(d2)                             - = undefined
                                 2

(c3) log(1+x)=powerseries(log(1+x),x,0);
                                      inf
                                      ====         i2   i2
                                      \       (- 1)    x
(d3)                  log(x + 1) = -   >      -----------
                                      /            i2
                                      ====
                                      i2 = 1

(c4) ev(d3,x=2,simpsum);
                                      inf
                                      ====         i2   i2
                                      \       (- 1)    2
(d4)                     log(3) = -    >      -----------
                                      /            i2
                                      ====
                                      i2 = 1
```

Figure 4: Convergence of series in MACSYMA.

SYMA tries to evaluate the series and realizes that it does not converge(d2). At line c3, we ask MACSYMA to find the power series representation for the function $\log(1+x)$. The result is returned in line d3 where the subscript "i2" has been generated by the system. At line c4, we ask MACSYMA to evaluate the series at $x = 2$ which is outside the interval of convergence. Since MACSYMA cannot evaluate this series it simply returns the series as a result. In this case, the series does not converge since the general term of the series does not converge to zero. Unfortunately, MACSYMA does not recognize this fact and returns a divergent series. This example emphasizes that mathematical reasoning is not just a matter of blind manipulation. Successful use of a CAS requires a good understanding of the underlying mathematics.

5 Symbolic Programming.

Computer algebra systems can be used in both an interactive mode and a programming mode. The interactive mode is illustrated by the examples in Fig. 2, Fig. 3 and Fig. 4. The programming mode makes it possible to implement mathematical algorithms in a high level programming language [10].

Like numerical computer programs [11], programs in a CAS language utilize assignment statements, conditional statements, loops and subprograms. Since there are differences between numerical programming and symbolic programming, it often takes time for a numerical programmer to feel comfortable with symbolic programming. These differences include:

- In a CAS language, variables can represent programming variables or mathematical variables in an expression.
- The primary data type in symbolic programming is the mathematical expression. The two most important data structures are lists and sets. Arrays are also used in symbolic programming but they do not play the essential role they play in numerical programming.
- Symbolic programs can utilize mathematical knowledge by invoking mathematical operators which analyze or manipulate mathematical expressions.
- Recursive programming techniques are often utilized (instead of loops) to solve symbolic mathematical problems. Many mathematical scientists (particularly FORTRAN programmers) may not be familiar with recursion.

[10] Some computer algebra systems (MAPLE and muMath) are written primarily in the high level computer language which comes with the system.

[11] We refer to programs written in languages such as FORTRAN, Pascal or C as numerical programs to distinguish them from programs written in the high level language of a CAS. Of course, these "numerical" languages have other data types and can be used for other types of programming. For the solution of mathematical problems, it is there numerical capabilities which are most important.

- Some computer algebra systems (MACSYMA, REDUCE, and MATHEMATICA) have pattern matching facilities which provide a way to add new simplification rules to a CAS. In some instances, this capability can eliminate the need for involved programs based on conditional statements, loops and recursion.

- In mathematical discussions there is a subtle distinction between the way the variable assignment and substitution operations are used. It is also important to determine when each of these two operations is appropriate in a symbolic program.

- Program efficiency (for CPU time and memory allocation) is an important issue for the symbolic programmer. Programs written in a CAS language can be unbearably slow. Reasons for this include:

    1. CAS languages are interactive rather than compiled. Each statement in a program must be translated each time it is used.
    2. Most arithmetic in a CAS is done with rational numbers, which can have an arbitrary number of digits, rather than real numbers, which have a fixed precision. This increases the CPU time for arithmetical operations.
    3. The algorithms to perform some mathematical operations (limits, integration, solution of ordinary differential equations, radical simplification, etc.) are time consuming.
    4. Automatic simplification rules are applied during the execution of each statement in a program.

  Programs in a CAS language can also require a large amount of computer memory. Reasons for this include:

    1. The storage of a mathematical expression requires much more computer memory than the storage of a real number in a numerical program.
    2. Some mathematical operations (expansion, differentiation, determinant calculation) often produce very large expressions which may eventually simplify to much smaller expressions. This phenomena, known as *intermediate expression swell*, can significantly increase the memory requirements for a program.

Programming examples and exercises are chosen to illustrate the programming techniques and data structures needed to implement symbolic mathematical algorithms in a CAS language. A good source for programming exercises is some of the elementary mathematical operators which already exist in a CAS. In the next section, we suggest a collection of exercises which illustrate some of the special problems associated with symbolic computation.

6 Algorithms For Symbolic Computation.

To effectively use a CAS, it is important to have some understanding of how a CAS works. By examining algorithms to perform symbolic computation, we clarify the important computational issues faced by the field and develop a sense of what manipulations are appropriate for a machine. Since the audience is composed of mathematical scientists who will use a CAS to solve problems rather than developers who design computer algebra systems, simple convincing algorithms are more appropriate than the most efficient algorithms currently available. The algorithms for the following operators illustrate many issues which arise in symbolic computation:

1. An operator *freeof(f,x)* which determines if an expression f contains a variable x.

2. An operator to find a list of all variables and function names which occur in a mathematical expression.

3. An operator to find the power set of a set.

4. An operator *degree(f,x)* which determines if an expression f is a polynomial in x, and, if so, returns the degree of the polynomial.

5. An operator *coefficient(f,x,n)* which determines the coefficient of x^n in a polynomial f.

6. An operator to perform polynomial division for polynomials with one or several variables.

7. An operator to compute the greatest common divisor of two polynomials with one or several variables using Euclid's algorithm.

8. An operator to find the square free factorization of a polynomial.

9. An operator to find the partial fraction expansion of a rational expression.

10. An operator to factor polynomials with one or several variables with integer coefficients using Kronecker's algorithm.

11. An operator to expand trigonometric expressions using the angle addition formulas and the multiple angle formulas.

12. An operator to reduce trigonometric expressions using the reduction formulas for products and powers of trigonometric expressions.

13. An operator to simplify all occurrences of i^n ($i^2 = -1$, n an integer) in an expression.

14. An operator to implement the rational substitution operation found in the MACSYMA system.

15. A differentiation operator.

16. An integration operator which applies the substitution method to perform integration.

17. An operator which determines the real and imaginary parts of a complex expression.

Most of these examples involve relatively simple symbolic programming techniques and none of them involves advanced mathematics. Nevertheless, a program to implement some of these operators can be surprisingly challenging to the beginning symbolic programmer.

We have purposely omitted from this list the more modern algorithms which make a CAS work efficiently. This material requires a strong background in modern algebra and is more appropriate for an advanced course in computer algebra. We have also purposely omitted more advanced areas of applied mathematics. Unless one is thoroughly familiar with some area of mathematics, it can be difficult to appreciate the point of an example. Once a mathematical scientist has some experience with symbolic programming, the techniques can be applied to more advanced problems.

7 Conclusion.

It is often said that computer algebra systems have the potential to revolutionize the way we do symbolic mathematics. Nevertheless, to date, only a fraction of the mathematical scientists who could profit by using this technology use a CAS in a significant way. We believe that this is partially due to the fact that many mathematical scientists do not understand the possibilities (and limitations) of computer symbolic computation and do not believe a CAS can help in their work. We also believe there is more to the effective use of a CAS than reading a system manual and seeing a few impressive examples. In this paper we have discussed an applied computer algebra course which can provide the technical background to effectively use this technology.

References

[1] Char, B. W., Geddes, K. O., Gonnet, G. H. and Watt, S. M., **Maple User's Guide**, WATCOM Publications Limited, Waterloo, Ontario, Canada, 1985.

[2] Cohen, J. S., *Differentiation of Undefined Functions with the Implicit and Inverse Function Theorems*, MACSYMA Newsletter, Volume IV, Number I, Symbolics Inc., January 1987.

[3] Conte, S. D. and de Boor, C, **Elementary Numerical Analysis, An Algorithmic Approach**, Third Edition, McGraw-Hill, 1980.

[4] Dorn, W.S. and McCracken, D., **Numerical Methods With Fortran IV Case Studies**, John Wiley and Sons, New York, 1972.

[5] Eisberg, R.M., **Fundamentals of Modern Physics**, John Wiley & Sons, 1961.

[6] Hearn, A. (editor) **Reduce Users Manual**, Version 3.2, The Rand Corporation, Santa Monica, CA 90406, April 1985.

[7] Knuth, D. E., *Algorithmic Thinking and Mathematical Thinking*, The American Mathematical Monthly, Vol. 92, No. 3, March 1985.

[8] Kreyszig, E., **Advanced Engineering Mathematics**, Third edition, John Wiley & Sons, 1972.

[9] Moses, Joel, *Algebraic Simplification, A Guide For the Perplexed*, Communications of the ACM, 14, No. 8, Aug. 1971, pp. 527-537.

[10] **muMath-83 Reference Manual**, The Soft Warehouse, Honolulu, Hawaii.

[11] Pavelle, Richard (Editor), **Applications of Computer Algebra**, Kluwer Academic Publishers, Boston, 1985.

[12] Stoutemyer, D. R. and Rich, A. D., **Derive, A Mathematical Assistant**, Computer Algebra Software Produced by Soft Warehouse Inc., Honolulu, 1988.

[13] **Vax Unix MACSYMA Reference Manual**, Symbolics Inc., Cambridge, MA, November 1985.

[14] Wester, M. and Steinberg S, *A Survey of Symbolic Differentiation Implementations*, Proceedings of the 1984 MACSYMA Users' Conference, General Electric, Schenectady, N.Y., July 1984, 330-335.

[15] Wolfram S., **Mathematica, A System for Doing Mathematics by Computer**, Addison-Wesley Publishing Company, 1988.

GROEBNER BASES

Moss Sweedler
Mathematical Sciences Institute
Cornell University
Ithaca NY 14853

ABSTRACT Groebner bases are remarkable sets of polynomials which permit effective manipulation of multivariate polynomials. In spirit, Groebner bases apply univariate polynomial techniques to multivariate polynomials. The theory and techniques which have grown up about Groebner bases are an important branch of computational algebra. While many of the techniques associated with Groebner bases are simple enough to be taught in high school, an undergraduate abstract algebra course is required to begin appreciatiating the algebra applications. Outside of algebra, Groebner bases have application to robotics, computational geometry, geometric theorem proving and other areas. Application to surface modeling and cryptography are under investigation. Groebner bases are remarkably poorly known within the algebra research community. This, despite the fact the associated algorithms are *high school algebra* [1] yet provide systematic answers to important questions which most algebraists have no other way to answer.

INTRODUCTION Groebner bases are the invention of Bruno Buchberger, [9]. The present importance of Groebner bases results from the conjunction of Buchberger's seminal work together with the body of techniques which have developed around his work. **Buchberger theory** is an appropriate name for the area. In the same way **Galois theory** refers to a body of techniques.[1] At present, pure mathematicians primarily use Groebner bases to compute examples. Inevitably, Buchberger theory, like Galois theory, will be freely used in proofs.[2]

The four cornerstones of Buchberger theory are:

LEADING TERMS **REDUCTION**

BASIS TEST **BASIS CONSTRUCTION**

The rest of the introduction airs algebra applications of Buchberger theory. Those unfamiliar with the concepts may still understand the sections: **1 LEADING TERMS, 2 REDUCTION, 3 GROEBNER BASIS TEST and CONSTRUCTION,** which are

---

[1]PROPAGANDA: Buchberger theory is as fundamental and more elementary than Galois theory and should be taught in advanced undergraduate algebra courses and the first year graduate algebra course.

[2]Which gets us to **Hironaka theory**. Ideal bases with special properties are not new. Among others, there are Ritt bases [50], [51] and standard bases [23] which are already used in proofs.

honestly elementary. The last harangue: **4 WHERE THE ACTION ISN'T**, passionately portrays the prevailing pitiful, paltry position of constructive algebra among North American academic algebraists.

ALGEBRA APPLICATIONS When using Groebner bases, one typically starts with a finite set of polynomials - and an ordering on the monomials of the polynomial ring - and constructs a **Groebner basis** for the ideal generated by the original polynomials. One customarily gets information from a Groebner basis by one of two methods:

I: simple **I**nspection of the Groebner basis

R: a constructive technique called **R**eduction.

Constructing the Groebner basis is generally tedious, i.e. computationally expensive, [38]. Reduction is much easier. Reduction has the *flavor* of the Euclidean algorithm and is occasionally described as: **The generalization of the Euclidean algorithm to several variables.**[3] Here are several algebra applications of Buchberger theory. Each application is preceded by **R** for **R**eduction or **I** for **I**nspection, according to how one gets information from the Groebner basis. We use the following notation: $A = R[X_1,\cdots,X_n]$ is a polynomial ring over the field R, a is an element of A and F is a finite subset of A. $\langle F\rangle$ is the ideal in A generated by F, $R[F]$ the subalgebra of A generated by F and $R(F)$ the subfield of $R(X_1,\cdots,X_n)$ generated by F.

R Determine if $a \in \langle F\rangle$.

I Determine if $a \in \sqrt{\langle F\rangle}$, the radical of $\langle F\rangle$.

R Determine if $a \in R[F]$.

I Determine if $a \in R(F)$.

I Determine a generating set for $(\langle F\rangle : a) = \{ b \in A \mid ba \in \langle F\rangle \}$.

I If J is another ideal in A with an explicit finite generating set, determine a generating set for $\langle F\rangle \cap J$.

I If $1 \le m < n$ and $B = R[X_1,\cdots,X_m]$, determine a generating set for the ideal: $B \cap \langle F\rangle$, in B.

I Find the relations among the elements of F.

I Determine $[\, R(X_1,\cdots,X_n) : R(F)\,]$, meaning the index if algebraic, the transcendence degree, if not.

R Be able to effectively work in $A/\langle F\rangle$ by having distinguished coset representatives in A for elements of $A/\langle F\rangle$. For any element of A be able to determine the distinguished coset representative to which it is equivalent.

There are aspects of Buchberger theory which have the spirit of **construction lines** in plain geometry. For example, in most of the above applications, one finds a Groebner basis for a cleverly chosen ideal in the ring: A-with-additional-indeterminates-adjoined. Although the techniques are elementary, they are tedious, when done by hand, for all

[3]See the end of section 2 for more about this.

but the smallest examples. A number of computer algebra systems -- Macaulay, MACSYMA, MAPLE, Mathematica, REDUCE, Scratchpad II, etc. -- are capable of executing various aspects of Buchberger theory.

Buchberger theory has many generalizations, for example: to free modules over polynomial rings [3], rings with suitable filtrations or valuations [53], [67], etc. Groebner bases for free modules allows effective computation of syzygies, free resolutions, Hilbert functions and more.

1 LEADING TERMS Portions of Buchberger theory are extensions of univariate polynomial techniques to multivariate polynomial rings. Univariate polynomials have a natural expression in terms of descending term degree. The degree of the largest non-zero term is the degree of the polynomial and plays a key role in univariate polynomial theory. The first difficulty with multivariate polynomials is the lack of a natural leading term. The first cornerstone of Buchberger theory is a method for recovering a notion of *leading term*. Buchberger's method, which we present here, involves orderings on the monomials of the polynomial ring. Specific orderings on sets of monomials have been used long before Buchberger's work. Particularly the lexicographic order. Buchberger isolated the needed properties of an abstract ordering.[4] One approach to generalizing Buchberger's work has been to develop alternative notions of *leading terms* not based on orderings of monomials.

1.1 DEFINITION A **multiplicative order** on monomials of a polynomial ring is a total order on the monomials satisfying:

1.1.a $1 \leq m$ for all monomials m

1.1.b if $m_1 \leq m_2$ then $m_1 m_3 \leq m_2 m_3$ for all monomials m_1, m_2, m_3

Lexicographic order is an easy example of a multiplicative order. In this order: $X^aY^b \cdots Z^c > X^dY^e \cdots Z^f$ if the left-most non-zero term of $(a - d, b - e, \cdots, c - f)$ is positive. The reverse lexicographic order - where $X^aY^b \cdots Z^c > X^dY^e \cdots Z^f$ if the right-most non-zero term of $(a - d, b - e, \cdots, c - f)$ is negative - is not a multiplicative order. Two other multiplicative orders:

1.2 Compare by total degree, break ties lexicographically.

1.3 Compare by total degree, break ties reverse lexicographically.

Mathematicians' usual initial reaction is that (1.2) and (1.3) must give essentially the same order, possibly after *renaming* the variables. However in three variables X, Y, Z :

$X > Y > Z$ and $XZ > Y^2$ in the (1.2) order

$X > Y > Z$ and $XZ < Y^2$ in the (1.3) order

This essential difference cannot be *renamed* away.[5] In one variable, the usual degree

[4]Not being a historian, I cannot say whether these properties had been isolated earlier.

[5]A few applications of Buchberger theory require the (1.3) order.

order is the unique multiplicative order on monomials. In general there are an infinite number of multiplicative orders. Often, the application for which one is using Buchberger theory, constrains the multiplicative order which may be used. The lexicographic order is an easily implemented multiplicative order for computer algebra systems. The lexicographic order is suitable for most, but not all, applications. However, other orders - in particular the (1.3) order - generally require less computation [6].

Once one has a multiplicative order, the univariate case may be imitated to some degree. For example, polynomials may be written with the monomials in descending order. The largest term - with non-zero coefficient - is dubbed the **leading term** of the polynomial. The coefficient (monomial, exponent) of the leading term of a polynomial is called **the leading coefficient (monomial, exponent) of the polynomial.**

Multiplicative orders have the important property of being **well orderings**. Consequently, processes such as reduction halt.

2 REDUCTION Let us begin with polynomials in one variable and the familiar process of dividing one polynomial, $f_0(X)$, by another, $g_0(X)$. The aim is to *evolve the process of reduction from the process of polynomial division.* Hence, the polynomial division process will be considered in detail. Suppose

$$f_0(X) = 6X^{17} + \text{lower degree stuff}$$

$$g_0(X) = 2X^{12} + \text{lower degree stuff}$$

First step: $6X^{17}/2X^{12} = 3X^5$; $f_1(X)$ is defined as $f_0(X) - 3X^5 * g_0(X)$. The step is imitated with $f_1(X)$ and $g_0(X)$, assuming $f_1(X)$ has degree at least as large as $g_0(X)$. Let us tweak the division process. Suppose at the second step $g_0(X)$ may be replaced by another polynomial $g_1(X)$. In other words, the first step is imitated with $f_1(X)$ and $g_1(X)$, assuming $f_1(X)$ has degree at least as large as $g_1(X)$. Suppose at each step the $g_\#(X)$ polynomial may be replaced by another polynomial. Suppose **S** is a given set of polynomials, where at each step[6] $g_\#(X)$ may be chosen as any polynomial in S. This process is the **reduction of $f_0(X)$ over S**. When must it *halt*? When an $f_i(X)$ has been reached which has smaller degree than all polynomials in S. Polynomial division yields a remainder which is uniquely determined by the divisor and the dividend. The example where $f_0(X) = X$ and $S = \{ X , X + 1 \}$ shows that the remainder can depend upon which elements of S are chosen as $g_\#(X)$'s.[7] In one variable, *halting* is apparently based on degree. The halting condition may be rephrased: the reduction process *halts* when an $f_i(X)$ has been reached whose lead monomial is not divisible by the lead monomial of any polynomial in S. The univariate reduction process is now easily generalized to polynomial rings in several variables, with a given multiplicative order.

---

[6]Including the first!

[7]Looking ahead: when S is a Groebner basis, the remainder of complete reduction is independent of how the $g_\#(X)$'s are chosen from S.

Let **A** be a polynomial ring with a given multiplicative order. As indicated toward the end of the previous section, the multiplicative order allows us to speak of the leading term, leading coefficient, leading monomial, etc. of a polynomial. Let f_0 be a polynomial in A and let S be a subset of A. The following, inductively defined, process is the **reduction of f_0 over S**.

1 If $f_i = 0$, halt.

2 If the leading monomial of f_i is **not** divisible by the leading monomial of any polynomial in S, halt.

3 If this step has been reached[8], there is a polynomial s_i in S whose lead monomial divides the lead monomial of f_i. Find a polynomial q_i where the lead term of q_i times the lead term of s_i equals the lead term of f_i.

4 Set $f_{i+1} = f_i - q_i s_i$.

5 GOTO step (**1**).

Conventionally q_i is chosen as the polynomial consisting of the single term whose coefficient is the lead coefficient of f_i divided by the lead coefficient of s_i and whose monomial is the lead monomial of f_i divided by the lead monomial of s_i. The coefficient division can be performed because the coefficients lie in a field.[9] The monomial division can be performed by the assumption stated at the start of step (**3**). More elaborate choices of q_i, may lead to computational optimization in the reduction process. Allowing general q_i's in the definition of the reduction process has advantages for developing the general theory. Further restrictions may always be placed on q_i's in implementations.

The $f_\#$'s which result are called **reductums** of f_0 (over S).[10] The reduction process always halts. The last f_i reached is called a **final reductum** of f_0 (over S). The final reductum is the generalization of the remainder in polynomial division. As noted earlier, the final reduction is not an invariant of f_0 and S. If f_i is a reductum of f_0 over S then $f_0 - f_i$ lies in the ideal generated by S.

Suppose S lies in an ideal I. It is easy to show the equivalence of:

The lead monomial of each non-zero element of I is divisible by the lead monomial of some element of S.

Each element of I has a reduction over S with final reductum zero.

For each element of I all reduction over S have final reductum zero.

---

[8]Meaning, has been reached on the current pass through the algorithm.

[9]Another realm of generalizations of Buchberger theory concerns weakening the requirement that the coefficient ring of the polynomial ring be a field.

[10]**NOTATION CONUNDRUM** Across the disciplines of abstract algebra, computational algebra and computer algebra there is great terminology disparity. For example, one author's **term** is another author's **monomial**. Our usage of **reduction** is more or less standard. Our usage of **reductum** is not. A polynomial **f** can be written: **lead term** + **lower stuff**, where **lower stuff** indicates the sum of terms other than the lead term. What we call **lower stuff** is frequently called the **reductum of f**.

2.1 DEFINITION **S is a Groebner basis for I** if the above conditions are satisfied.

A Groebner basis for an ideal generates the ideal. This prompts:

2.2 DEFINITION A set T is a Groebner basis if it is a Groebner basis for the ideal it generates.

The previous univariate example with $S = \{ X, X + 1 \}$ is an example of a set which is **not** a Groebner basis. In the univariate polynomial ring, a set is a Groebner basis if and only if it contains a principal generator for the ideal it generates. In fact, for polynomial rings in any number of variables, a subset of a principal ideal is a Groebner basis for the ideal if and only if the subset contains a principal generator for the ideal. Thus a singleton set is always a Groebner basis.

A fundamental application of Buchberger theory uses reduction for an ideal membership test.

2.3 THEOREM Let S be a Groebner basis for an ideal I and let *a* be an element of the polynomial ring. The following are equivalent:

a lies in I.

a has a reduction over S with final reductum zero.

All reductions of *a* over S have final reductum zero.

COMPLETE REDUCTION In the reduction process, only the lead term of the $f_{\#}$'s gets reduced by elements of S. In the **complete reduction process**, all terms of the $f_{\#}$'s get reduced by elements of S. The process halts with an f_i which is either zero or where no terms of f_i (have monomials which) are divisible by the lead monomial of an element of S. The complete reduction process always halts. When doing complete reduction of f over S, the final reductum may be referred to as a **complete reduction of f over S**.

2.4 THEOREM Let S be a Groebner basis and let *a* be an element of the polynomial ring. The complete reduction of *a* over S is unique. If T is a Groebner basis which generates the same ideal as S, the complete reduction of *a* over S equals the complete reduction of *a* over T.

Thus, complete reduction gives distinguished coset representatives. This allows effective computation in multivariate polynomial rings modulo an ideal. To compute modulo I, find a Groebner basis S for I.[11] Given *a* in the polynomial ring, the distinguished coset representative for *a* modulo I is the complete reduction of *a* over S.

Groebner bases (for an ideal) are not unique. Their cardinality is not unique and they generally are not minimal generating sets for the ideal. There is a notion of **reduced Groebner basis** which involves complete reduction. Ideals have unique reduced Groebner bases.[12]

---

[11] How to find a Groebner basis for an ideal comes later.

[12] Wellllllll, reduced Groebner bases consisting of monic polynomials are unique.

Presenting **reduction** as the multivariate polynomial analog of **univariate polynomial division** has pedagogical merits and relies on the following dictionary:

REDUCTION	**POLYNOMIAL DIVISION**
the set one reduces over	the divisor
the element being reduced	the dividend
the final reductum	the remainder
comparison by multiplicative order	comparison by degree

The analogy has limitations. With R a field, consider the polynomial ring R[X,Y] having the lexicographic multiplicative order with $X > Y$. The singleton set $\{ Y \}$ is a Groebner basis. Consider the reduction of X over $\{ Y \}$. X itself is the final reductum. The size of the final reductum is larger than every element of $\{ Y \}$. With reduction, one cannot be certain that the size of the final reductum will be smaller than elements of the set one reduces over. Translated to univariate polynomial division, this would be as if the remainder were not necessarily of lower degree than the divisor.

3 GROEBNER BASIS TEST and CONSTRUCTION: in which it is revealed how to test if a given set S is a Groebner basis and if S is not a Groebner basis, how to enlarge S to a Groebner basis generating the same ideal. We plead guilty to presenting the easiest material, the most leisurely. The pace quickens this section. The Groebner basis test involves a number of reductions over S. S is a Groebner basis, if and only if **all** the final reductums are zero. If one of the final reductums is not zero, S is not a Groebner basis. However, this non-zero final reductum is an element to be used to enlarge S to get closer to having a Groebner basis.

3.1 DEFINITION Let f and g be non-zero polynomials with lead monomials M_f and M_g respectively. Let $M_{f,g}$ be the monomial which is the least common multiple of M_f and M_g. One can find polynomials F and G where fF and gG each have lead monomial $M_{f,g}$ and fF has the same lead coefficient as gG. fF - gG is an **S-polynomial of the pair f and g**.

Notice that the lead terms of fF and gG cancel in the difference fF - gG. Thus, the lead monomial of an S-polynomial of the pair f and g is always lower than $M_{f,g}$.

F and G may each be chosen as polynomials consisting of single terms which are gotten as follows: the monomial of F is $M_{f,g}/M_f$ and the monomial of G is $M_{f,g}/M_g$. The coefficient of F is the lead coefficient of g and the coefficient of G is the lead coefficient of f. More elaborate choices of F and G may lead to computational optimization in the Groebner basis test and construction processes. Allowing general F and G in the definition of S-polynomials has advantages for developing the general theory. Further restrictions may always be placed on F and G in implementations.

This may not be correct but it would make good sense if the **S** in **S-polynomial** stands for **s**yzy**g**y. S-polynomials are the key to the Groebner basis test:

3.2 TEST THEOREM S is a set in a polynomial ring which has a multiplicative order.

The following are equivalent:

S is a Groebner basis.

For each pair of distinct elements $f,g \in S$, there is an S-polynomial of f and g which has a reduction over S with final reductum zero.

For each pair of distinct elements $f,g \in S$, all reductions over S of S-polynomials of f and g have final reductums equal to zero.

S is not assumed to be finite in the theorem. When S is finite, the theorem yields a constructive test whether S is a Groebner basis. If S is a singleton set, it automatically passes the test because there are no pairs to reduce. As mentioned before, the test underlies Groebner basis construction procedures. Here is one such procedure which begins with a finite set S and produces a finite Groebner basis for the ideal which S generates.

For pairs of distinct elements $f,g \in S$ form an S-polynomial of f and g and reduce the S-polynomial over S. If the final reductum is always zero, halt. S is a Groebner basis. On encountering a non-zero final reductum, let $T = S \cup \{$ the non-zero final reductum $\}$. Repeat with T in place of S.

Although this procedure always terminates with a finite Groebner basis for the ideal which S generates, the cardinality of the Groebner basis may be much larger than the cardinality of S. There are many optimizations which can be made. For example, as described, the procedure will duplicate many computations.

4 A NON-IDEAL APPLICATION So far, Groebner bases have been used in connection with ideals. We end with an application not concerning ideals.[13] The application is the matter of subalgebra membership determination and appears in [59]. Suppose $f,g_1, \cdots, g_r$ lie in the polynomial ring $R[X_1, \cdots, X_n]$. The question is to determine whether f lies in $R[g_1, \cdots, g_r]$, the subalgebra of $R[X_1, \cdots, X_n]$ generated by $\{ g_1, \cdots, g_r \}$. As part of the solution, we introduce additional variables. This is typical of Groebner basis applications and is what is meant by **construction lines** in the introduction. In this case, we introduce an additional variable T_i for each g_i.

Let A be the polynomial ring $R[X_1, \cdots, X_n, T_1, \cdots, T_r]$. Choose a multiplicative order on A where each X_i is larger than all monomials just involving $\{ T_1, \cdots, T_r \}$. For example, the lexicographic order has this property. Construct a Groebner basis G which generates the same ideal as $\{ g_1 - T_1, \cdots, g_r - T_r \}$.[14] Considering $R[X_1, \cdots, X_n] \subset A$, we may think of f as lying in A. Reduce f over G and let $h \in A$ be the final reductum. The answer to the subalgebra membership question is given by:

f lies in $R[g_1, \cdots, g_r]$ if and only if h lies in $R[T_1, \cdots, T_r]$.

If h lies in $R[T_1, \cdots, T_r]$, so that $h = h(T_1, \cdots, T_r)$ then $f = h(g_1, \cdots, g_r)$.

---

[13] Wellllllll, ideals creep in.

[14] In Groebner basis applications, the additional variables are often referred to as tag variables when they are used to tag other elements in the problem.

5 WHERE THE ACTION ISN'T North American academic mathematics departments! As a whole, North American algebraists, in academic mathematics departments, appear to have *computer anxiety* or *computation anxiety*. Conferences and bibliographies, which pertain to computational and computer algebra, have a remarkably low fraction of contributors from North American academic mathematics departments. In their place one finds algebraists from Europe, from industry and from computer science and other academic departments. North American algebraists, in academic mathematics departments, are appallingly ignorant of even the most elementary, yet relevant, developments in computational algebra.

BIBLIOGRAPHY

1 Abhyankar S.A. (1976) Historical ramblings in algebraic geometry and related algebra, Amer. Math. Monthly 83, 409-448.

2 Barkee B. Ecks J. (1989) BUCHBERGER THEORY: Techniques with Polynomials for Computer and Mathematical Applications, book in preparation.

3 Bayer D.A. (1982) The division algorithm and the Hilbert scheme, Ph.D. Thesis, Harvard.

4 Bayer D. Stillman M. (1987) A criterion for detecting m-regularity, Inventiones Math. 87, 1-11.

5 Bayer D. Stillman M. The design of Macaulay: A system for computing in algebraic geometry and commutative algebra, Proc. of the 1986 A.C.M. Symposium on Symbolic and Algebraic Computation (ed. B. Char) pp. 157-162.

6 Bayer D. Stillman M. (1987) Refining orders by the reverse lexicographic order, Duke Math. Journal, 55, 321-328.

7 Bayer D. Stillman M. On the complexity of computing syzygies. To appear in Journal of Symbolic Computation.

8 Boege W. Gebauer R. Kredel H. (1986) Some Examples for Solving Systems of Algebraic Equations by Calculating Groebner Bases, J. Symbolic Computation, 1, 83-98.

9 Buchberger B. (1965) An Algorithm for Finding a Basis for the Residue Class Ring of a Zero-Dimensional Polynomial Ideal, Ph.D. dissertation, Mathematical Institute, Univ. Innsbruck, Austria (in German).

10 Buchberger B. (1970) An algorithmical criterion for the solvability of algebraic systems ofequations, Aequationes Mathematicae 4, 374-382.

11 Buchberger B. (1976a) A theoretical basis for the reduction of polynomials to canonical forms, ACM-SIGSAM Bull. 39, 19-29.

12 Buchberger B. (1976b) Some properties of Groebner-base for polynomial ideal, ACM-SIGSAM Bull. 10/4, 19-24.

13 Buchberger B. (1983) A note on the complexity of constructing Groebner-bases, In: J.A. van Hulzen (ed.) Computer Algebra, Springer Lec. Notes in Comp. Sci. 162, 137-145.

14 Buchberger B. (1984) A critical-pair/completion algorithm for finitely generated ideal in rings, In: E. Boerger, G. Hasenjaeger, D. Roedding (eds.) Logic and Machines: Decision Problems and Complexity, Springer Lec. Notes in Comp. Sci. 171, 137-155.

15 Buchberger B. (1985) Groebner bases: An algorithmic method in polynomial ideal theory, In: (N.K. Bose ed.) Progress, directions and open problems in multi-dimensional systems theory. pp. 184-232. Dordrecht: Reidel Publ. Comp.

16 Carra-Ferro G. Groebner Bases and Differential Algebra, preprint.

17 Carra-Ferro G. Groebner Bases and Hilbert Schemes I, to appear in Journal of Symbolic Computation.

18 Carra-Ferro G. Gallo G. A Procedure to Prove Geometrical Statements, preprint.

19 Dube T. Mishra B. Yap C. (1986) Admissible orderings and bounds on Groebner normal form algorithm, Technical Report No. 258, Robotics Report No. 88, New York Univ. Dept. of Computer Science, Courant Institute of Mathematical Sciences, New York, (submitted to J. of Symbolic Computation).

20 Gebauer R. Moeller H.M. (1986) A Variant of Buchberger's Algorithm, submitted to Journal of Symbolic Computation.

21 Giusti M. (1986) A note on the complexity of constructing standard bases, Eurocal 85, Lecture Notes in Computer Science, no. 204, 411-412, Springer-Verlag.

22 Gianni P. Trager B. Zacharias G. (1984) Groebner bases and primary decomposition of polynomial ideals, Preprint.

23 Hironaka H. (1964) Resolution of singularities of an algebraic variety over a field of characteristic zero, Ann. Math. 79, 109-326.

24 Kandri-Rody A. (1985) Dimension of ideals in polynomial rings, Proc. Combinatorial Algorithms in algebraic structures, Otzenhausen, J. Avenhaus, K. Madlener Eds. Fachbereich Informatik, University of Kaiserslautern.

25 Kandri-Rody A. Kapur D. (1984a) Computing the Groebner basis of an ideal in polynomial rings over the integers, Proc. Third MACSYMA User's Conf. 436-451.

26 Kandri-Rody A. Kapur D. (1984b) An algorithm for computing the Groebner basis of a polynomial ideal over a Euclidean ring, J. Symbolic Computation, this issue.

27 Kandri-Rody A. Kapur D. (1983) On Relationship between Buchberger's Groebner Basis Algorithm and the Knuth-Bendix Completion Procedure, Unpublished GE CRD Technical Report, Nov. 1983, Schenectady, NY.

28 Kandri-Rody A. Kapur D. (1984b) Algorithms for computing Groebner basis of polynomial ideals over various Euclidean rings, Lecture Notes in Computer Science 174, EUROSAM 84, International Symposium on Symbolic and Algebraic Computation, Cambridge, England, 195-208.

29 Kapur D.K. (1986) Using Groebner bases to reason about Geometry problems, J. of Sym. Computation 2, p. 399-408.

30 Kutzler B. Stifter S. (1986) Automated geometry theorem proving using Buchberger's algorithm, CAMP-Publ.-Nr. 85-29.0, University of Linz; Proc. SYMSAC '86, Waterloo, Canada.

31 Lankford D.S. (1985) Abstract Properties and Applications of Generalized Groebner Bases, notes.

32 Lankford D.S. (1986) Generalized Groebner Bases: Theory and Applications, preprint.

33 Lauer M. (1976) Canonical representatives for residue class of a polynomial ideal, In: R.D. Jenks (ed.) Proc. 1976 ACM Symp. Symbolic and Algebraic Computation, 339-345.

34 Lazard D. (1983) Groebner bases, Gaussian elimination, and resolution of systems of algebraic equations, Proc. EUROCAL 83 Springer, L.N.C.S. 162, 146-156.

35 Lazard D. (1985) Ideal Bases and Primary Decomposition: Case of Two Variables, J. Symbolic Computation I, 261-270.

36 Lejeune-Jalabert M. (1984) Effectivite de Calculs Polynomiaux, Universite de Grenoble I, Institut Fourier, Laboratoire de mathematiques associe au C.N.R.S.

37 Macaulay F.S. (1927) Some properties of enumeration in the theory of modular systems, Proc. London Math. Soc. 26, 531-555.

38 Mayr E. Meyer A.R. (1982) The complexity of the Word Problems for Commutative Semigroups and Polynomial Ideals, Adv. in Math. 46 305-329, Report LCS/TM-199, M.I.T. Laboratory of Computer Science.

39 Mishra B. Yap C. (1987) Notes on Groebner Bases, Courant Institute, NYU.

40 Moeller H.M. (1985) On the computation of Groebner bases in commutative rings, Preprint.

41 Moeller H.M. Mora F. (1983) The computation of the Hilbert function, in EUROCAL '83, Computer algebra, Springer L.N.C.S. vol. 162, 157-167.

42 Moeller H.M. Mora F. (1984) Upper and lower bounds for the degree of Groebner bases, Proc. EUROSAM 84 Springer L.N.C.S. 172-183.

43 Moeller H.M. Mora F. (1986a) New constructive methods in classical ideal theory, J. Alg. 99.

44 Moeller H.M. Mora F. (1986b) Computational aspects of reduction strategies to construct resolutions of monomial ideals, Proc. AAECC, Springer L.N.C.S.

45 Mora F. (1982) An algorithm to compute the equations of tangent cones, Proc. EUROCAM 82, Springer, L.N.C.S. 144, 158-165.

46 Mora F. (1983). A constructive characterization of standard bases, Boll. U.M.I. Sez. D 2, 41-50.

47 Mora F. (1985) An algorithmic approach to local rings, Proc. EUROCAL '85, L.N.C.S. 204, 518-525, Springer-Verlag.

48 Mora F. (1986) Groebner bases for non commutative polynomial rings, Proceedings for the AAECC3 L.N.C.S. vol. 229, 353-362.

49 Pan L. On the D-Bases of polynomial ideals over principal ideal domains, Bell Communications Research, N.J. Preprint.

50 Ritt J.F. (1932) Differential Equations from the Algebraic Standpoint, Colloquium Publications of the American Mathematical Society, Vol. 14.

51 Ritt J.F. (1950) Differential Algebra, Colloquium Publications of the American Mathematical Society, Vol. 33.

52 Robbiano L. (1985) Term orderings on the polynomial ring, Proc. EUROCAL 85, II Springer L.N.C.S. 204, 513-517.

53 Robbiano L. (1986) On the theory of graded structures, J. Symbolic Computation, 2, 139-170.

54 Robbiano L. Sweedler M. Bases for Subalgebras, in preparation.

55 Robbiano L. Valla G. On set theoretic complete intersections in the projective space, Preprint.

56 Robbiano L. Valla G. (1983) Free resolutions for special tangent cones, Commutative Algebra, Proc. of the Trento Conference, Lect. Notes in Pure and Appl. Math. 84, Marcel Dekker, New York.

57 Schaller S. (1979) Algorithmic Aspects of Polynomial Residue Cass Rings, Ph.D. dissertation, University of Wisconsin at Madison.

58 Schreyer F.O. (1980) Die Berechnung von Syzygien mit dem verallgemeinerten Weierstrasschen Divisionsatz ... Diplomarbeit Hamburg.

59 Shannon D. Sweedler M. Using Groebner bases to determine algebra membership, split surjective algebra homomorphisms and determine birational equivalence, with David Shannon, to appear Journal of Symbolic Computation.

60 Shannon D. Sweedler M. Using Groebner bases to determine the algebraic or transcendental nature of field extensions within the field of rational functions, in draft.

61 Spear D. (1977) A constructive approach to commutative ring theory, Proc. MACSYMA User's Conf. 369-376.

62 Stillman M&M. (1986) Macaulay Users Manual.

63 Sturmfels B. An Algorithmic Proof of the Quillen-Suslin Theorem. Institute for Mathematics and its Appl. University of Minnesota, Preprint.

64 Sturmfels B. White N. Computing Combinatorial Decompositions of Rings, Notes.

65 Sturmfels B. White N. (1988) Efficient Computation of Fundamental Invariants - An Approach Using Buchberger's Groebner Bases Method, preprint.

66 Sturmfels B. White N. Groebner bases and invariant theory, Advances in Math. to appear.

67 Sweedler M. (1987) Ideal bases and valuation rings, draft.

68 Szerkeres G. (1952) A canonical basis for the ideals of a polynomial domain, Am. Math. Monthly 59/6, 379-386.

69 Trinks W. (1978) On B. Buchberger's method for solving algebraic equations, J. Number Theory 10, 475-488 (in German).

70 Winkler F. (1983) An algorithm for constructing detaching bases in the ring of polynomials over a field, Computer Algebra, LNCS 162, Springer-Verlag, 168-179.

71 Winkler F. (1984) On the complexity of the Groebner algorithm over K[x,y,z], Proc. EUROSAM 84, Cambridge, (ed. by J.Fitch) Springer Lecture Notes in Computer Science 174, 184-194.

72 Winkler F. (1987) A p-adic Approach to the Computation of Groebner Bases, Technical Report, RISC-Linz Series no. 87-1.0.

73 Zacharias G. (1978) Generalized Groebner Bases in Commutative Polynomial Rings, Bachelor's thesis, Dept. of Computer Science, Mass. Inst. Technology.

USING MACSYMA IN A GENERALIZED HARMONIC BALANCE METHOD FOR A PROBLEM OF FORCED NONLINEAR OSCILLATIONS

M. A. Hussain
General Electric Corporate R&D Center, Schenectady, NY

B. Noble
Brunell University, Uxbridge, UK

J. J. Wu
US Army Research Office, Research Triangle Park, NC

ABSTRACT

This paper will intoduce a generalized harmonic balance method and illustrate the use of symbolic computing to solve a class of nonlinear vibration problems. The program package MACSYMA is used in this demonstration.

First, a forced vibration problem with several different type of nonlinearities is given. An outline of the method will be described next. This will be followed by symbolic computing statements and programs which will crank out asymptotic solutions in routine manner. Solutions for a subharmonic and a superharmonic will be given. Thus the easy of obtaining results of these otherwise extremely complicated problems will be shown.

1. INTRODUCTION

Since the introduction of symbolic computation as a tool for mathematical analysis, it has been increasingly used for solutions of problems where laborious and repetitious mathematical manipulations are required. In particular, it is found extremely helpful for solutions of nonlinear differential equations in conjunction with various perturbation methods[1]. In this paper, we will introduce a generalized harmonic balance methods for solutions of a class of nonlinear vibration problems and demonstrate the use of MACSYMA, a very powerful and popular symbolic computation software package to obtain these solutions in a routine manner. We will consider the vibrational response of a nonlinear single degree-of-freedom system with quadratic and cubic nonlinearities governed by the equation

$$d^2u/dt^2+u+2\varepsilon\mu(du/dt)+\varepsilon\alpha_2u^2+\varepsilon^2\alpha_3u^3+\varepsilon\alpha_4(du/dt)^2 +\varepsilon^2\alpha_5u(du/dt)^2=2f\cos(\Omega t) \qquad (1)$$

where $\Omega=2+\varepsilon\sigma$. This same equation has been considered by Nayfeh[2], using the method of multiple scaling. Our objective is to obtain his key results by a method that avoids steps of the elimination of secular terms, repeated solutions of intermidiate differential equations and Nayfeh's reconstitution method and thus to demonstrate that the multiple scaling results can derived from our solution approach. Although we solve only this specific example,

the proposed method is quite general and can be used for any problem where the nonlinearities are polynomials in u and du/dt, e.g., see Nayfeh and Mook [2], and Nayfeh [3] and [4].

In both Nayfeh [3] and Nayfeh and Mook[2], Sect. 2.3.4, it is shown that the method of harmonic balance can lead to erroneous results if applied simply in a routine fashion. Quoting from p.61 of this later reference: ".....to obtain a consistent solution by using the method of harmonic balance one needs either to know a great deal about the solution a priori or to carry out enough terms in the solution and check the order of the coefficients of all the neglected harmonics, ... Therefore we prefer not to use this technique." In this paper, we avoid both of these objects by using the beginning steps of multiple scaling to tell the form of the solution (see equation (18) at the end of Section 2), which gives us the a priori information we require and also enables us to see clearly which harmonics has to be taken into account and which can be neglected.

2. FORM OF SOLUTION VIA MULTIPLE SCALING

As emphasized already, the key to the success of our varient of the harmonic balance method is to know the form of the solution, by which we mean dependence of the solution on the small quantity ε which is a measure of the nonlinearity. We do this by the multiple scaling approach but without getting involved in the laborous tasks of supression of the secular terms, obtaining the explicit solutions of the intermediate differential equations and reconstitution of the final solution.

To illustrate the point, the multiple scale method as applied to (1) assumed that (cf. Nayfeh [3], p.461):

$$u(t,\varepsilon)=u_0(T_0,T_1,....)+\varepsilon u_1(T_0,T_1,....) +\varepsilon^2 u_2(T_0,T_1,....)+.... \tag{2}$$

where

$$T_n=\varepsilon^n t, \quad n=0,1,2,.... \tag{3}$$

Using Nahfey's notation, $D_n=d/dT_n$, one has

$$\begin{aligned} d/dt &= D_0 + \varepsilon D_1 + \varepsilon^2 D_2 + ... \\ d^2/dt^2 &= D_0^2 + 2\varepsilon D_0 D_1 + \varepsilon^2(2D_0D_2 + D_1^2) + ... \end{aligned} \tag{4}$$

Substituting (2) and (4) in (1) and equating coefficients of like powers of ε, one obtains (Nayfeh [3], p.466, (32)-(34)):

$$D_0^2 u_0 + u_0 = 2f\cos(\Omega T_0) \tag{5}$$

$$D_0^2 u_1 + u_1 = -2D_0D_1u_0 - 2\mu D_0 u_0 - \alpha_2 u_0^2 - \alpha_4 (D_0 u_0)^2 \tag{6}$$

$$D_0^2 u_2 + u_2 = -(2D_0D_2 + D_1^2)u_0 - 2D_0D_1u_1 - 2\mu D_0 u_1$$

$$-2\mu D_1 u_0 - 2\alpha_2 u_0 u_1 - \alpha_3 u_0^3$$
$$-2\alpha_4 D_0 u_0 (D_1 u_0 + D_0 u_1) - \alpha_5 u_0 (D_0 u_0)^2 \qquad (7)$$

Equations (5)-(7) are obtained easily using the basic steps of the multiple scaling represented by (2) and (3), which are the only part of the multiple scaling used in the present approach. Yet, (5)-(7) are significant because they provide us the form of the solution desired as will be described below.

In order to save labor from carrying unnecessary terms (and there are many of them), one must keep track on the relative order of various terms. We begin by noting that u_k (k=0,1,2,....) in (2) are of order unity, or O(1).

Taking the case of subharmonic response for an example,

$$\Omega = 2 + \varepsilon\sigma \qquad (8)$$

where σ is of O(1) and is called the detuning parameter. Eqn. (5) can now be written as

$$D_0^2 u_0 + u_0 = fSA^2 + c.c. \qquad (9)$$

where

$$S = e^{i\varepsilon\sigma t}, \quad A = e^{it} \qquad (10)$$

and c.c. stands for the complex conjugate.

It is important to note that S is a slow varying function compared with A in the sense that while dA/dt is of O(1), dS/dt is of O(ε).

It is easily observed, from (9), that

$$u_0 = P_{01}(T_1, T_2, ...)e^{iT_0} + P_{02}(T_1, T_2, ...)e^{iT_0} + c.c.$$

or,

$$u_0 = P_{01}(\varepsilon t, \varepsilon^2 t, ...)A + P_{02}(\varepsilon t, \varepsilon^2 t, ...)A^2 + c.c. \quad (11)$$

with

$$A = e^{iT_0} = e^{it} \quad (12)$$

and that P_{01} and P_{02} are some functions of t, the specific forms of which are not the concern here. However, it is important to note that P_{01} and P_{02} are slow varying functions in t compared with A in the sense that while dA/dt is of O(1), the derivatives of P_{01} and P02 are of $O(\varepsilon)$. We shall also use the fact that

$$\bar{A} = e^{-it}, \quad \text{and} \quad A\,\bar{A} = 1 \quad (13)$$

Next we substitute (11) into the right hand side of (6) resulting a polynomial in A^k, k=0,1,2,3,4. Hence, it is easily observed that the solution of (6) can be written as

$$u_1 = P_{10} + (P_{11}A + P_{12}A^2 + P_{13}A^3 + P_{14}A^4 + c.c.) \quad (14)$$

where, again, P_k, k=0,1,2,3 and 4, are slow varying functions in t compared with A. Now, substituting (11) and (14) into (7) and going through a similar process as before, one can write easily

$$u_2 = P_{20} + (P_{21}A + P_{22}A^2 + P_{23}A^3 + P_{24}A^4 + P_{25}A^5 + P_{26}A^6 + c.c.) \quad (15)$$

Once again, P_{2k} (k=0,1,....,6) are slow varying compared with A.

Hence to obtain the form of the solution u, one substitutes (11), (14) and (15) in (2) and collects terms of same powers in A. The result is

$$u = \varepsilon U_0 + [(U_1A + U_2A^2) + \varepsilon(U_3A^3 + U_4A^4) + \varepsilon^2(U_5A^5 + U_6A^6) + c.c.] \quad (16)$$

where

$$U_0 = P_{10} + \varepsilon P_{20}, \qquad U_1 = P_{01} + \varepsilon P_{11} + \varepsilon^2 P_{21}$$
$$U_2 = P_{02} + \varepsilon P_{12} + \varepsilon^2 P_{22}, \quad U_3 = P_{13} + \varepsilon P_{23} \tag{17}$$
$$U_4 = P_{14} + \varepsilon P_{24}, \qquad U_5 = P_{25}, \quad U_6 = P_{26}$$

It is then also clear that $U_k (k=0,1,2,\ldots,6)$ are of O(1) and slow varying compared with A. The approximate solution (16) of u is good to the order of ε . To obtain a solution of u good to the order of ε, we shall drop U_5 and U_6 terms so that the final form of the solution to be used in this paper is

$$u = \varepsilon U_0 + [(U_1 A + U_2 A^2) + \varepsilon (U_3 A^3 + U_4 A^4) + c.c.] \tag{18}$$

In the next two sections, we shall derive expressions of Uk's with the help of the MACSYMA program for the subharmonic and superharmonic vibrations.

3. THE TRANSIENT SUBHARMONIC CASE

If one wishes to obtain the first two-term approximation, $u=u_0+\varepsilon u_1$, in the solution of (1) as a power series in ε, the procedure is then to substitute (18) in (1) and set to zero the coefficients of A^k, k=0,1,2,3 and 4. First, all the terms in (1) will be written in power series in ε, dropping those of $O(\varepsilon^3)$:

$$du/dt=(dU_1/dt+iU_1)A+(dU_2/dt+2iU_2)A^2 +[(dU_3/dt+3iU_3)A^3+(dU_4/dt+4iU_4)A^4]+c.c. \tag{19}$$

$$d^2u/dt^2=(d^2U_1/dt^2+2idU_1/dt-U_1)A+(d^2U_2/dt^2+4idU_2/dt-4U_2)A^2 +\varepsilon[(6idU_3/dt-9U_3)A^3+(8idU_4/dt-16U_4)A^4]+c.c. \tag{20}$$

$$u^2=2U_1\bar{U}_1+2U_2\bar{U}_2+[2\bar{U}_1U_2A+U_1{}^2A^2+2U_1U_2A^3+U_2{}^2A^4 +2\varepsilon(U_0U_1+\bar{U}_2U_3)A+c.c.] \tag{21}$$

Since u^3 appears with a coefficient of ε^2 in (1), one only needs to keep terms of O(1) in the expansion:

$$u^3=3U_1{}^2\bar{U}_2+3U_1{}^2U_2+[(3U_1{}^2\bar{U}_1+6U_1U_2\bar{U}_2)A+(3U_2{}^2\bar{U}_2+6U_1\bar{U}_1U_2)A^2 +(U_1{}^3+3\bar{U}_1U_2{}^2)A^3+3U_1{}^2U_2A^4+c.c.] \tag{22}$$

Similarly, one has

$$(du/dt)^2=2U_1\bar{U}_1+8U_2\bar{U}_2+2i[(U_1\ d\bar{U}_1/dt-dU_1/dt\ \bar{U}_1) +2(U_2\ d\bar{U}_2/dt-dU_2/dt\ \bar{U}_2)] +[(4\bar{U}_1U_2-2i\bar{U}_1\ dU_2/dt+4id\bar{U}_1/dt\ U_2+12\varepsilon\bar{U}_2U_3)A +(-U_1{}^2+2idU_1/dt\ U_1+6\varepsilon\bar{U}_1U_3+16\varepsilon\bar{U}_2U_4)A^2 +(-4U_1U_2+2idU_2/dt\ U_1+4idU_1/dt\ U_2+8\varepsilon\bar{U}_1U_4)A^3$$

$$+(-4U_2^2+4idU_2/dt\ U_2-6\varepsilon U_1U_3)A^4+c.c.] \tag{23}$$

$$u(du/dt)^2=3\bar{U}_1^2U_2+3U_1^2\bar{U}_2+[(U_1^2\bar{U}1+8U_1U_2\bar{U}_2)A$$
$$+(2U_1\bar{U}_1U_2+4U_2^2\bar{U}_2)A^2-U_1^3A^3-5U_1^2U_2+c.c.] \tag{24}$$

We now substitute (18) and (20)-(24) in (1), collect terms of like power of A^k, k=0,1,..,4 and then set the coefficients to zero. The resulting equations are: for A^0 coefficient,

$$\varepsilon[U_0+2\alpha_2(U_1\bar{U}_1+U_2\bar{U}_2)+2\alpha_4(U_1\bar{U}_1+4U_2\bar{U}_2)]=0 \tag{25}$$

for A^1,

$$2i(dU_1/dt+\varepsilon\mu U_1)+2\varepsilon(\alpha_2+2\alpha_4)\bar{U}_1U_2+d^2U_1/dt^2+2\varepsilon\mu dU_1/dt$$
$$+2i\varepsilon\alpha_4(2U_2\ d\bar{U}_1/dt-\bar{U}_1\ dU_2/dt)$$
$$+\varepsilon^2[2\alpha_2(U_0U_1+U_3\bar{U}_2)+3\alpha_3(U_1^2\bar{U}_1+2U_1U_2\bar{U}_2)$$
$$+12\alpha_4\bar{U}_2U_3+\alpha_5(U_1^2\bar{U}_1+8U_1U_2\bar{U}_2)]=0 \tag{26}$$

for A^2,

$$-3U_2+4idU_2/dt+\varepsilon[4i\mu U_2+(\alpha_2-\alpha_4)U_1^2]=fS \tag{27}$$

for A^3,

$$\varepsilon[-8U_3+2(\alpha_2-2\alpha_4)U_1U_2]=0 \tag{28}$$

and, finally, for A^4,

$$\varepsilon[-15U_4+(\alpha_2-\alpha_4)U_2^2]=0 \tag{29}$$

Equations (25)-(29) can again be conveniently put in a tabulated form as in TABLE I.

Since the end goal here is to obtain an equation which contains the information on the relationship between the amptitude and the frequency, and since we have five unknowns and five equations, we can reduce them to one equation with a single

TABLE I. TABULATED EQUATIONS INDICATING RELATIVE ORDER OF TERMS IN THE TRANSIENT, SUBHARMONIC CASE

	ε^0	ε^1	ε^2	RHS
A^0	0	$\varepsilon[U_0+2\alpha_2(U_1\bar{U}_1+U_2\bar{U}_2)$ $+2\alpha_4(U_1\bar{U}_1+4U_2\bar{U}_2)]$	***	0
A^1	0	$2i(dU_1/dt+\varepsilon\mu U_1)$ $+2\varepsilon(\alpha_2+2\alpha_4)\bar{U}_1U_2$	$d^2U_1/dt^2+2\varepsilon\mu dU_1/dt$ $+2i\varepsilon\alpha_4(2U_2\ d\bar{U}_1/dt$ $-\bar{U}_1\ dU_2/dt)$ $+2\varepsilon^2[\alpha_2(U_0U_1+\bar{U}_2U_3)$ $+6\alpha_4U_2U_3]$ $+\varepsilon^2[3\alpha_3(\bar{U}_1U_1^2+2U_1U_2\bar{U}_2)$ $+\alpha_5(\bar{U}_1U_1^2+8U_1U_2\bar{U}_2)$	0
A^2	$-3U_2$	$4idU_2/dt$ $+\varepsilon[4i\mu U_2+(\alpha_2-\alpha_4)U_1^2$	***	fS
A^3	0	$\varepsilon[-8U_3+2(\alpha_2-2\alpha_4)U_1U_2]$	***	0
A^4	0	$\varepsilon[-15U_4+(\alpha_2-4\alpha_4)U_2^2]$	***	0

Note that, RHS indicates the right hand side of the equation and *** indicates terms not needed for the present approximation.

unknown. Also from (18), it is clear that U_1 and U_2 are more significant compared with U_0, U_3 and U_4 in the sense that, in order to obtain a solution of u(t) accurate to $O(\varepsilon^2)$, while U_1 and U_2 must be accurate to $O(\varepsilon^2)$, U_0, U_3 and U_4 only need to be of $O(\varepsilon)$. This relative significance in order also affords us an easy way to solve these equations by an iterative procedure.

The most dominant term of the solution is U_2 in (27),

$$U_2=-(1/3)fS=-(1/3)fe^{i\varepsilon\sigma t} \tag{30}$$

In terms of U_1 and U_2, one has, to the first approximation,

$$U_0=-2(\alpha_2+\alpha_4)U_1\bar{U}_1-2(\alpha_2+4\alpha_4)U_2\bar{U}_2 \tag{31}$$

$$U_3=(1/4)(\alpha_2-2\alpha_4)U_1U_2 \tag{32}$$

$$U_4=(1/15)(\alpha_2-4\alpha_4)U_2^2 \tag{33}$$

And the differential equation to determine U_1 is

$$dU_1/dt=-\varepsilon[\mu U_1-i(\alpha_2+2\alpha_4)\bar{U}_1U_2] \tag{34}$$

To improve the solution to the next order of accuracy, one includes the $O(\varepsilon)$ terms in (27) for U_2 and the $O(\varepsilon^2)$ terms in (26) for the differential equation for U_1. Hence

$$U_2=-(1/3)fS+(1/3)\varepsilon[(\alpha 2-\alpha 4)U_1^2-(4/3)(i\mu-\sigma)fS] \tag{35}$$

$$\bar{U}_2=-(1/3)f\bar{S}+(1/3)\varepsilon[(\alpha 2-\alpha 4)\bar{U}_1^2+(4/3)(i\mu+\sigma)f\bar{S}] \tag{36}$$

Also obtaining

$$d^2U_1/dt^2=-\varepsilon[\mu dU_1/dt-i(\alpha_2+2\alpha_4)(U_2\ d\bar{U}_1/dt+\bar{U}_1\ dU_2/dt)] \qquad (37)$$

Equations (35)-(37) can be substituted in (26) to obtain an improved first order differential equation for U_1:

$$\begin{aligned}
&2i(dU_1/dt+\varepsilon\mu U_1)-(2/3)\varepsilon(\alpha_2+2\alpha_4)fS\bar{U}_1\\
&+\varepsilon^2\{[-\mu^2U_1+(2/9)f^2)3\alpha_3+4\alpha_5)-(1/18)f^2(5\alpha_2{}^2+12\alpha_2\alpha_4-12\alpha_4{}^2)]U_1\\
&+(1/3)(9\alpha_3+3\alpha_5-10\alpha_2{}^2-10\alpha_2\alpha_4-4\alpha_4{}^2)U_1{}^2\bar{U}_1\\
&-(4/9)i\mu(2\alpha_2+\alpha_4)fS\bar{U}_1+(1/9)\sigma(11\alpha_2+16\alpha_4)fS\bar{U}_1\}=0 \qquad (38)
\end{aligned}$$

It is not difficult to show that (38) is identical to Nayfeh's equation (81) in [1] by replacing Ω with $2+\varepsilon\sigma$, ω_0 with unity and U_1 with A. In this comparison, caution must be used however, in obtaining the expanded terms of $O(\varepsilon^2)$ in the last expression in Nayfeh's equation where, specifically, $\Omega=2+\varepsilon\sigma$ and $\Omega^2=(2+\varepsilon\sigma)^2=4+2\varepsilon\sigma$ should be used for the parameter $\Lambda=f/(1-\Omega^2)$ in the expansion before dropping the higher order terms too early.

The procedure as described hitherto is automated via a MACSYMA program in the following. Remarks are contained between symbols /* and */ in the program. It is noted that the equation obtained by seting the final expression (D56) to zero in the MACSYMA program is identical to Eq. (18) as it should be.

```
(C3) /*  A GENERALIZED HARMONIC BALANCE METHOD FOR FORCED NONLIEAR VIBRATIONS.  */
/*  THIS IS FOR THE SUBHARMONIC CASE.  */


/* Set up automatic truncation.  */
nnpred(n):=is(n>=3)$

(C4) matchdeclare(nn,nnpred)$

(C5) tellsimpafter(ep^nn,0)$

(C6) /*  Define functional dependence.  */
depends([u,ut,a,at,s,st],[t])$

(C7) /*  Define derivatives.  */
gradef(a,t,%i*a)$

(C8) gradef(at,t,-%i*at)$

(C9) gradef(s,t,%i*sigma*ep*s)$

(C10) gradef(st,t,-%i*sigma*ep*st)$

(C11) /*  Form of approximate solution.  */
uu:ep*u[0]+(u[1]*a+u[2]*a^2+ut[1]*at+ut[2]*at^2)
        +ep*(u[3]*a^3+u[4]*a^4+ut[3]*at^3+ut[4]*at^4)$

(C12) /*  Formula for simplification.  */
term_simplify(term):=block([ans],ans:ratsubst(1,a*at,term),ans:expand(ans),
                           ans:subst(0,at,ans),return(ans))$

(C13) /*  Calculate, simplify and collect all the terms.  */
t1:term_simplify(diff(uu,t,2)+uu)$

(C14) t2:term_simplify(2*ep*mu*diff(uu,t))$

(C15) t3:term_simplify(ep*alpha2*uu^2)$

(C16) t4:term_simplify(ep^2*alpha3*uu^3)$

(C17) t5:term_simplify(ep*alpha4*diff(uu,t)^2)$

(C18) t6:term_simplify(ep^2*alpha5*diff(uu,t)^2*uu)$

(C19) final:t1+t2+t3+t4+t5+t6$

(C20) /*
```

```
.
.
*/
/*  Formula for collecting  terms with same power of "A".  */
power_list(exp):=block([l0,l1,l2,l3,l4,m,tt,expl,apower],expl:expand(exp),m:nterms(expl),
                       l0:l1:l2:l3:l4:0,for i thru m do
                      (tt:part(expl,i),apower:hipow(tt,a),
                       if apower = 0 then l0:l0+tt else
                      (if apower = 1 then l1:l1+tt else
                      (if apower = 2 then l2:l2+tt else
                      (if apower = 3 then l3:l3+tt else
                      (if apower = 4 then l4:l4+tt))))),return([l0,l1,l2,l3,l4]))$

(C21) /*  Formula for collecting terms with same power of "ep" for esch A^i.  */
orderterm(exp):=block([expl,t1,index,l0,l1,l2,l3,l4],l0:l1:l2:l3:l4:0,expl:expand(exp),
                m:nterms(expl),for i thru m do (t1:part(expl,i),index:0,
                                for j from 0 thru 4
                do (index:index+derivdegree(t1,u[j],t),
                    index:index+derivdegree(t1,ut[j],t)),
                    index:index+hipow(t1,ep),
                    if index = 0 then l0:l0+t1 else
                   (if index = 1 then l1:l1+t1 else
                   (if index = 2 then l2:l2+t1 else
                   (if index = 3 then l3:l3+t1 else
                   (if index = 4 then l4:l4+t1))))),[l0,l1,l2,l3,l4])$

(C22) powerset:power_list(final)$

(C23) set[i]:=orderterm(powerset[i])$

(C24) /*  Formula for getting the complex conjugates.  */
complex_transform(exp):=block([ltemp,ltempr,lcomp,ans],
        ltemp:[ut[1]=utemp[1],ut[2]=utemp[2],ut[3]=utemp[3],ut[4]=utemp[4],at=atemp,st=stemp],
        ltempr:[utemp[1]=u[1],utemp[2]=u[2],utemp[3]=u[3],utemp[4]=u[4],atemp=a,stemp=s],
        lcomp:[u[1]=ut[1],u[2]=ut[2],u[3]=ut[3],u[4]=ut[4],a=at,s=st,%i=-%i],
        ans:ev(exp,ltemp),ans:ev(ans,lcomp),ans:ev(ans,ltempr))$

(C25) /*  List coefficients C[i][j] of the term (ep)^i * (A)^j.  */
list[i]:=[set[1][i],set[2][i],set[3][i],set[4][i],set[5][i]]$

(C26) /*  Right hand side column vector.  */
rhs1:[0,0,a^2*f*s,0,0]$

(C27) /*  Define unknowns in equations of the first iteration.  */
depend1:[0,0,u[2],0,0]$

(C28) depend2:[u[0],diff(u[1],t),u[3],u[4]]$

(C29) /*
```

```
*/
/*  Solve for u[2] in the first iteration.  */
first_iteration:solve(list[1]=rhs1,u[2])$

Dependent equations eliminated:  (1 5 4 2)

(C30) first_iteration_t:complex_transform(first_iteration)$

(C31) /*  Solve for dU[1]/dt, U[3] and U[4] in the first iteration.  */
sol_set1:[u[0]=0,diff(u[1],t)=0,first_iteration[1][1],u[3]=0,u[4]=0]$

(C32) /*  Setting up for the second iteration.  */
second_orderlist:factor(ev(list[2],first_iteration))$

(C33) /*  Coefficient of the term (ep)^1 * (A)^2.  */
u21:second_orderlist[3]$

(C34) second_orderlist:delete(u21,second_orderlist)$

(C35) /*  Use the result of first iteration.  */
u21:ev(u21,first_iteration)$

(C36) u21:ev(u21,diff)$

(C37) /*  U[2] of the second iteration.  */
newu2:rhs(first_iteration[1][1])+u21/(3*a^2)$

(C38) updatefirst_iteration:[u[2]=newu2]$

(C39) updatefirst_iteration_t:complex_transform(updatefirst_iteration)$

(C40) /*  Second iteration for dU[1]/dt, U[3] and U[4].  */
second_iteration:solve(second_orderlist,depend2)$

(C41) second_iteration:ev(second_iteration,first_iteration)$

(C42) second_iteration:ev(second_iteration,diff)$

(C43) second_iteration_t:complex_transform(second_iteration)$

(C44) /*  Solution set from the second iteration.  */
sol_set2:[second_iteration[1][1],second_iteration[1][2],u[2]=u21,second_iteration[1][3],
          second_iteration[1][4]]$

(C45) /*  First derivatives of the second iteration.  */
deriv_secondit:diff(second_iteration,t)$

(C46) deriv_secondit:ev(deriv_secondit,first_iteration)$

(C47) deriv_secondit:ev(deriv_secondit,diff)$

(C48) /*
```

```
*/

/*  Coefficient of the term (ep)^2 * (A)^1.  */
test1:list[3][2]$

(C49) /*  Form row vector in order of power of (ep).  */
quadratic_order(expr):=block([ans],ans:ev(expr,deriv_secondit),ans:ev(ans,second_iteration),
                    ans:ev(ans,second_iteration_t),   ans:ev(ans,diff),
                    ans:ev(ans,updatefirst_iteration), ans:ev(ans,updatefirst_iteration_t),
                    ans:ev(ans,diff),ans:ratsubst(1,s*st,ans),ans:ratsubst(1,a*at,ans),
                    ans:orderterm(ans),return(ans))$

(C50) check:quadratic_order(test1)$

(C51) check[3]$

(C52) /*  Coefficient of the term (ep)^1 *(A)^1.  */
test2:list[2][2]$

(C53) test2:ev(test2,updatefirst_iteration)$

(C54) /*  Final differential equation for U[1] including (ep)^2 terms and simplifications.  */
test3:test2+check[3]$

(C55) test3:ratsimp(test3/a)$

(C56) test3:ratexpand(test3);
```

$$
\begin{aligned}
(D56)\quad & \frac{16\,UT_1\,ALPHA4\,EP^2\,F\,S\,SIGMA}{9} + \frac{11\,UT_1\,ALPHA2\,EP^2\,F\,S\,SIGMA}{9} - \frac{4\,\%I\,UT_1\,ALPHA4\,EP^2\,F\,MU\,S}{9} - \frac{8\,\%I\,UT_1\,ALPHA2\,EP^2\,F\,MU\,S}{9} \\
& - \frac{4\,UT_1\,ALPHA4\,EP\,F\,S}{3} - \frac{2\,UT_1\,ALPHA2\,EP\,F\,S}{3} - U_1\,EP^2\,MU^2 + 2\,\%I\,U_1\,EP\,MU + \frac{8\,U_1\,ALPHA5\,EP^2\,F^2}{9} + \frac{2\,U_1\,ALPHA4^2\,EP^2\,F^2}{3} \\
& + \frac{2\,U_1\,ALPHA2\,ALPHA4\,EP^2\,F^2}{3} + \frac{2\,U_1\,ALPHA3\,EP^2\,F^2}{3} - \frac{5\,U_1\,ALPHA2^2\,EP^2\,F^2}{18} + U_1^2\,UT_1^2\,ALPHA5\,EP^2 - \frac{4\,U_1^2\,UT_1^2\,ALPHA4^2\,EP^2}{3} \\
& - \frac{10\,U_1^2\,UT_1^2\,ALPHA2\,ALPHA4\,EP^2}{3} + 3\,U_1^2\,UT_1^2\,ALPHA3\,EP^2 - \frac{10\,U_1^2\,UT_1^2\,ALPHA2^2\,EP^2}{3} + 2\,\%I\,\left(\frac{d}{dT}\,(U_1)\right)
\end{aligned}
$$

```
(C57) closefile();
```

4. THE SUPERHARMONIC CASE

Next, we will demonstrate the technique in obtaining the solution for a superharmonic case, where the frequency of the forcing function Ω is close to one-half of the natural frequency of the system. That is

$$\Omega = (1/2) + \varepsilon\sigma, \text{ or, } \quad 2\Omega = 1+2\varepsilon\sigma \tag{39}$$

By using (39) instead of (8) in Sectioin 2 of this paper, it is clear that all the steps remain unaltered except that S and A in (10) will be replaced by

$$S = e^{i\varepsilon\sigma t/2} \quad \text{and} \quad A = e^{it/2} \tag{40}$$

Thus we can conclude that (18) is still the form of solution of (1) for the superharmonic case with A given by (40). Now the task is simple. Substituting (18) in (1) with proper consideration of A in (40), one arrives at a similar set of equations and can be solved as before. Here, we shall use the following simple case of (1):

$$d^2u/dt^2+u+2\varepsilon\mu(du/dt)+\varepsilon\alpha_2u^2=fe^{i\Omega t}+c.c. \tag{41}$$

or,

$$d^2u/dt^2+u+2\varepsilon\mu(du/dt)+\varepsilon\alpha_2u^2=fSA+c.c. \tag{42}$$

Substituting (18) in (42) the following Table (Table II) of equations similar to Table I is obtained.

The final equation in U_2 from solving the above set of equations iteratively as before is the following:

$$2i(dU_2/dt+\varepsilon\mu U_2)+(16/9)\varepsilon\alpha_2f^2S^2$$
$$+\varepsilon^2\{[-\mu^2-(64/9)(23/15)\alpha_2{}^2f^2]U_2-(10/3)\alpha_2{}^2\bar{U}_2U_2{}^2$$

$$+(8/27)(5\sigma-13i\mu)\alpha_2 f^2 s^2\} = 0 \qquad (43)$$

Again, it is easily shown that (43) is identical to the result obtained by Nayfeh using the method of multiple scaling (equation (2.46) in [4]).

TABLE II. TABULATED EQUATIONS INDICATING RELATIVE ORDER OF TERMS IN THE TRANSIENT, SUPERHARMONIC CASE

	ε^0	ε^1	ε^2	RHS
A^0	0	$\varepsilon[U_0+2\alpha_2(U_1\bar{U}_1+U_2\bar{U}_2)$	***	0
A^1	$3U_1/4$	$i(dU_1/dt+\varepsilon\mu U_1)$ $+2\varepsilon\alpha_2\bar{U}_1U_2$	***	fS
A^2	0	$2idU_2/dt$ $+\varepsilon(2i\mu U_2+\alpha_2U_1^2)$	$d^2U_2/dt^2+2\varepsilon\mu dU_2/dt$ $2\varepsilon^2(U_0U_2+\bar{U}_1U_3+U_2\bar{U}_2)$	0
A^3	0	$\varepsilon(-5U_3/4+2\alpha_2U_1U_2)$	***	0
A^4	0	$\varepsilon(-3U_4+\alpha_2U_2^2)$	***	0

Note that, as in TABLE I, RHS indicates the right hand side of the equation and *** indicates terms not needed for the present approximation.

REFERENCES

[1] Richard H. Rand and Dieter Armbruster, Perturbation Methods, Bifurcation Theory and Computer Algebra, Springer-Verlag, 1987.

[2] A. H. Nayfeh and D. T. Mook, Nonlinear Oscillations, Wiley-Interscience, 1979.

[3] A. H. Nayfeh, The response of single degree of freedom systems with quadratic and cubic non-linearities to a subharmonic excitation, Journal of Sound and Vibration (1983), Vol. 89(4), pp.457-470.

[4] A. H. Nayfeh, Perturbation Methods in Nonlinear Dynamics, Lecture Notes in Physics: Nonlinear Dynamics Aspects of Particle Accelerators - Proceedings of the Joint US-CERN School on Particle Accelerators, Editors: J. M. Jowett, M. Month and S. Turner, Spring-Verlag, 1985, pp.238-314.

A Shared Memory Parallel FFT for Real and Even Sequences

William L. Briggs

Van Emden Henson

Mathematics Department
University of Colorado at Denver
Denver, Colorado 80202

Abstract

A compact symmetric FFT algorithm for real and even data is implemented on a shared memory parallel processing computer. The parallel implementation is complicated by the uneven distribution of work induced by splitting symmetric sequences. A performance model is developed to predict the amount of speed-up that may be expected as the number of processors is increased. Factors included in the model are the arithmetic operations, calls to the transcendental libraries, and overhead for the fork-join operations. Actual processing times are given for the real and even FFT. For fixed N, speed-up curves are shown for increasing numbers of processors, and are compared to the theoretical curves of the performance models. While the speed-up is excellent for long sequences, for short sequences the speed-up peaks at some intermediate number of processors.

Introduction

Since its introduction in 1965 [1], the Fast Fourier Transform (FFT) has become one of the most widely used algorithms of computational mathematics. Its enormous popularity is due largely to the fact that the FFT requires $O(N\log N)$ arithmetic operations to compute the transform of a complex vector of length N, instead of the $O(N^2)$ operations required to compute the transform as a matrix-vector product. The term "compact symmetric" refers to a family of FFT algorithms that uses minimal storage and arithmetic for data sequences possessing certain symmetries. The first such algorithm, generally attributed to Edson [2], computes the transform of a real vector using half the storage and half the operations used by the original FFT. It has long been known that further savings are possible when the data has additional symmetries, but with the exception of one little-publicized algorithm by Gentleman [3], such transforms were performed by pre- and post-processing of data for use with conventional FFTs [4]. In recent papers by Swarztrauber [5] and Briggs [6], compact algorithms were developed for sequences with real, even, odd, quarter wave even, and quarter wave odd symmetries.

The Cooley-Tukey Algorithm

Suppose a complex sequence, $\{x_n\} = \{x_0, x_1, \cdots x_{N-1}\}$, is given, for which the Discrete Fourier Transform is desired. For convenience, assume that the length of the sequence, N, is a power of two. The DFT of the sequence is given by

$$X_k = \sum_{n=0}^{N-1} x_n \omega_N^{nk} \qquad k=0,1, \cdots N-1$$

where $\omega_N = e^{\frac{-i2\pi}{N}}$. As a matrix-vector multiplication, the DFT requires $O(N^2)$ operations.

Suppose the sequence $\{x_n\}$ is split into two subsequences $\{y_n\}$ and $\{z_n\}$, whose elements are $y_n = x_{2n}$, and $z_n = x_{2n+1}$. Then the DFT can be written as

$$X_k = \sum_{n=0}^{\frac{N}{2}-1} y_n \omega_{\frac{N}{2}}^{nk} + \omega_N^k \sum_{n=0}^{\frac{N}{2}-1} z_n \omega_{\frac{N}{2}}^{nk} \qquad k=0,1, \cdots \frac{N}{2}-1$$

The two summations in this equation are themselves DFTs, of the subsequences $\{y_n\}$ and $\{z_n\}$ respectively, which are denoted $\{Y_k\}$ and $\{Z_k\}$. Therefore the first half of the desired transform is given by

$$X_k = Y_k + \omega_N^k Z_k \qquad k=0,1, \cdots \frac{N}{2}-1$$

The second half of the desired transform may be obtained by substituting $k + \frac{N}{2}$ for k and noting the periodicity of $\{Y_k\}$ and $\{Z_k\}$, which yields

$$X_{k+\frac{N}{2}} = Y_k - \omega_N^k Z_k \qquad k=0,1, \cdots \frac{N}{2}-1$$

These two formulas together make up the "butterfly relation", or combine formulas.

The Cooley-Tukey FFT algorithm proceeds by recursively splitting the input vector until eventually N sequences of length 1 are produced, which are their own DFTs. The butterfly relations may then be applied to build longer transforms from pairs of short ones. This process continues until finally two transforms of length $\frac{N}{2}$ are combined to form the length N transform of the original input sequence.

An Algorithm for Real and Even Data

An even sequence is one in which $x_n = x_{N-n}$. A sequence which is both real and even (E) can be shown to have a transform which is also an E sequence. Suppose the transform of an E sequence is desired. Following a process suggested by the Cooley-Tukey algorithm, the input sequence is split into its even and odd subsequences. These subsequences are split in turn, and this recursive process followed until length one sequences are produced. With each splitting the subsequences inherit certain symmetries from the parent sequence.

Suppose an E sequence $\{x_n\}$ is split into two subsequences, $\{y_n\}$, consisting of the elements with even-numbered indices, and $\{z_n\}$, the elements whose indices are odd. Then

$$y_n = x_{2n} = x_{N-2n} = y_{\frac{N}{2}-n}$$

and

$$z_n = x_{2n+1} = x_{N-2n-1} = z_{\frac{N}{2}-n-1}.$$

Therefore the subsequence $\{y_n\}$ is an E sequence, while the subsequence $\{z_n\}$ is real and and has a new symmetry called quarter wave even.

A quarter wave even sequence of length N is one in which $x_n = x_{N-n-1}$. If the DFT of a real quarter wave even (QE) sequence is considered, it can be shown that the transform has the symmetry $X_k = e^{\frac{i2\pi k}{N}} \overline{X}_k$. Therefore it is possible to represent both the real and imaginary parts of the sequence element X_k by a strictly real quantity, namely $\tilde{X}_k = e^{-\frac{i\pi k}{N}} X_k$. If a QE sequence is split into its even-numbered and odd-numbered elements, each of the resulting subsequences has no symmetry by itself (except that it is real), but taken together they have the intrasequence symmetry

$$y_{\frac{N}{2}-n-1} = x_{N-2n-2} = x_{2n+1} = z_n.$$

Thus a real QE sequence splits into two strictly real subsequences, one of which is redundant.

A strictly real sequence (R) is one in which each element is its own complex conjugate. Substituting this relationship into the definition for the DFT, it is easy to show that the transform of an R sequence $\{x_n\}$ has the conjugate symmetric property that $X_k = \overline{X}_{N-k}$. Splitting an R sequence produces two R sequences which are of half the length of the original. No additional symmetry is induced by the splitting of an R sequence.

A symmetric algorithm, proposed by Swarztrauber, is schematically illustrated in Figure 1. The input sequence is split, producing one E subsequence and one QE sequence. The E subsequence splits into another E and QE pair, while the QE sequence splits into two real R sequences, one of which is redundant. The splitting process continues, with each E generating an E and QE pair, each QE splitting into an R (and a redundant R), and each R sequence spliting into two more R sequences, until sequences of length one are produced (left of the bar in Figure 1). Note that no work has yet been performed, merely that data has been moved, and is now said to be in scrambled order.

Since the transform of a length one sequence is itself, the short transforms may now be combined into longer transforms, a process that may be applied recursively

until the transform of the full length input sequence is obtained. In order to do this recombination, it is necessary to have butterfly relations that combine the various symmetric transforms into transforms of longer symmetric sequences. By substituting the transform symmetries discussed above into the Cooley-Tukey combine formulas and simplifying, Swarztrauber derived these butterfly relations.

The Combine Formulas

If an R sequence $\{x_n\}$ has been split into its two subsequences, $\{y_n\}$ and $\{z_n\}$, the butterfly relations for constructing the transform, $\{X_k\}$, of the original sequence are

$$X_k = Y_k + \omega_N^k Z_k \qquad k=0,1,...\frac{N}{4}$$

and

$$\overline{X}_{\frac{N}{2}-k} = Y_k - \omega_N^k Z_k \qquad k=0,1,...\frac{N}{4}-1$$

These equations are called the "RtoR" combine formulas, because they combine the transforms of real sequences into the transform of a real sequence.

It was shown above that an E sequence $\{x_n\}$ splits into an E and a QE subsequence. Since $\{Y_k\}$, the transform of the E subsequence, is also real and even, and $\{Z_k\}$, the transform of the QE subsequence, can be represented by the strictly real sequence $\{\tilde{Z}_k\}$, the butterfly relations for combining the transforms of an E and a QE sequence are

$$X_k = Y_k + \tilde{Z}_k \qquad k=0,1,...\frac{N}{4}$$

and

$$X_{\frac{N}{2}-k} = Y_k - \tilde{Z}_k \qquad k=0,1,...\frac{N}{4}-1$$

These relations, which are together called an "EQE" type combination, produce the first $\frac{N}{2}+1$ values of the transform $\{X_k\}$. Since $\{X_k\}$ is real and even the remaining values may be obtained by symmetry.

It has been noted that in splitting a QE sequence $\{x_n\}$ one of the resulting subsequences is redundant. Therefore $\{X_k\}$ can be recovered entirely from the transform of $\{y_n\}$, which can be represented by a strictly real sequence $\{\tilde{Y}_k\}$, while the transform of $\{z_n\}$ need be neither computed nor stored. The butterfly relations by which the transform of a QE sequence can be formed from the transform of one of its R subsequences are

$$\tilde{X}_k = 2\mathrm{Re}[\omega_{\frac{N}{2}}^k Y_k] \qquad k=0,1,...\frac{N}{4}$$

and

$$\tilde{X}_{\frac{N}{2}-k} = -2\mathrm{Im}[\omega_{\frac{N}{2}}^k Y_k] \qquad k=0,1,...\frac{N}{4}-1$$

Thus $\frac{N}{2}$ real values are required to represent the N complex values of the transform of a real and quarter wave even sequence. This set of relations is called an "RRQE" combine.

The combine phase of the algorithm is shown schematically by the right side of Figure 1. In general, each pass has one EQE combination, followed by an RRQE, followed by a series of RtoR combinations. At the second to last pass, there are only the RRQE and EQE types, while the final pass involves only the EQE combination, performed on sequences of half the length of the original input. In practice none of the redundant R sequences are computed or stored. The program uses only $\frac{N}{2}+1$ storage locations. The algorithm begins with the input data in scrambled order, and proceeds through $\log_2 N$ passes, until the transform coefficients are produced in natural order.

The Inverted Algorithm

It is generally more convenient to have an algorithm which operates on the input data in natural order, producing the coefficients in scrambled order. Briggs [5] developed such algorithms for sequences which are real, quarter wave even, and quarter wave odd. Following his lead, an ordered-to-scrambled algorithm for an E sequence may be developed.

To derive this algorithm it is necessary to formally invert the Swarztrauber algorithm. Since an E sequence is itself the transform of another E sequence, the inverted algorithm can be thought of as following Figure 1 backwards, from the right side to the middle bar. Beginning with the transform of an E sequence in natural order, it is possible to "uncombine" it into the transforms of its E and QE subsequences. These in turn are uncombined into the transforms of their subsequences, and so on. After $\log_2 N$ passes through the data, length one transforms are produced, which are the transform coefficients, in scrambled order, of the original sequence.

To invert the Swarztrauber algorithm, it is necessary to formally invert all of the combine formulas. The EQE combine relations are easy to invert, and produce the uncombine relations

$$Y_k = \frac{1}{2}(X_k + X_{\frac{N}{2}-k}) \qquad k=0,1,...\frac{N}{4}$$

and

$$\tilde{Z}_k = \frac{1}{2}(X_k - X_{\frac{N}{2}-k}) \qquad k=0,1,...\frac{N}{4}-1$$

Inverting the RRQE combine formulas is a bit more tedious, leading to the uncombine relations

$$\mathrm{Re}[Y_k] = \frac{1}{2}(\tilde{X}_k \cos\frac{\pi k}{N} + \tilde{X}_{\frac{N}{2}-k} \sin\frac{\pi k}{N}) \qquad k=0,1,...\frac{N}{4}$$

and

$$\mathrm{Im}[Y_k] = \frac{1}{2}(\tilde{X}_k \sin\frac{\pi k}{N} - \tilde{X}_{\frac{N}{2}-k}\cos\frac{\pi k}{N}) \qquad k=0,1,...\frac{N}{4}-1$$

Inverting the RtoR combine formulas, and at the same time separating real and imaginary parts for storage in a real array, leads to the following four relations:

$$\mathrm{Re}[Y_k] = \frac{1}{2}(\mathrm{Re}[X_k] + \mathrm{Re}[X_{\frac{N}{2}-k}])$$

$$\mathrm{Im}[Y_k] = \frac{1}{2}(\mathrm{Im}[X_k] - \mathrm{Im}[X_{\frac{N}{2}-k}])$$

$$\mathrm{Re}[Z_k] = \frac{1}{2}\left\{(\mathrm{Re}[X_k] - \mathrm{Re}[X_{\frac{N}{2}-k}])\cos\frac{2\pi k}{N} - (\mathrm{Im}[X_k] + \mathrm{Im}[X_{\frac{N}{2}-k}])\sin\frac{2\pi k}{N}\right\}$$

$$\mathrm{Im}[Z_k] = \frac{1}{2}\left\{(\mathrm{Re}[X_k] - \mathrm{Re}[X_{\frac{N}{2}-k}])\sin\frac{2\pi k}{N} + (\mathrm{Im}[X_k] + \mathrm{Im}[X_{\frac{N}{2}-k}])\cos\frac{2\pi k}{N}\right\}$$

where the real parts of both $\{Y_k\}$ and $\{Z_k\}$ are calculated for $k = 0,1,...\frac{N}{4}$ and the imaginary parts of both sequences are calculated for $k = 0,1,...\frac{N}{4}-1$.

The remainder of this paper is concerned only with the ordered-to-scrambled algorithm, so the names EQE, RRQE, and RtoR are retained for these uncombine formulas.

Savings from the Compact Symmetric Algorithm

The data flow and storage of the ordered-to-scrambled algorithm are illustrated in Figure 2 for a real even sequence of length N=32. During each pass through the data, the first type of uncombine is an EQE. In the first pass this is the only type of uncombine. Beginning with the second pass, the RRQE type uncombine follows the EQE, and with each succeding pass there is one EQE, one RRQE, and all remaining uncombines are of type RtoR.

To analyze the algorithm, let the passes through the data be indexed $j = 0,1, \cdots \log_2 N - 2$. The last pass, $j = \log_2 N - 1$, is considered as a special case. The scalar multiplication by one-half occurs in all of the formulas, and may be performed at the end of the algorithm.

The backwards running index $(\frac{N}{2}-k)$ is important in the EQE uncombines. Because of it, individual EQE butterflies cannot be performed in place. However, pairs of EQE butterflies can be performed in place if performed together, as an EQE "unit". The j^{th} pass through the data requires $\frac{N}{8}(\frac{1}{2})^j$ such units, beginning with pass j=0. On the j^{th} pass $\frac{N}{4}(\frac{1}{2})^j$ RRQE butterflies are required, beginning with pass j=1. Beginning with the pass j=2, each RtoR sequence requires $\frac{N}{4}(\frac{1}{2})^j$ RtoR butterflies, and there are $2^{j-1}-1$ such sequences. The last pass through the data is considered separately because in this pass the sequences are all of length two

and all the butterfly types reduce to a butterfly which is identical to the EQE, and there are $\frac{N}{4}$ such butterflies.

Noting that each of the combine types requires a different amount of work, and using the counting arguments just listed, it is possible to compute a total operation count for the algorithm. The transform of an E sequence of length N using this algorithm requires $\frac{N}{2}+1$ storage locations, and the total number of real arithmetic operations (counting multiplications and additions equally) is $\frac{5}{4}N\log_2 N-2N$.

Performing the transform of an E sequence by placing the input sequence into the real part of a complex array and using a conventional FFT requires $2N$ storage locations and a real operation count of $5N\log_2 N$. Thus the compact symmetric FFT requires one fourth the storage as its conventional counterpart, and requires somewhat less than one fourth the arithmetic. Performing the same transform by traditional pre- and post-processing methods [4] utilizes $\frac{N}{4}$ storage locations and entails $\frac{5}{4}N\log_2 N+\frac{9}{2}N$ real operations, somewhat greater than the compact symmetric transform.

Parallel FFTs

Before proceeding to the problem of parallelizing symmetric sequences, it is useful to review some of the features of parallel FFTs for complex sequences. Many of the problems encountered in developing the parallel symmetric algorithms are similar to those that arise in parallelizing the conventional FFTs. Briggs [7] developed strategies for implementing FFTs on shared memory parallel processors.

The fundamental work unit of the FFT is the butterfly relation. During every pass through the data each butterfly relation can be performed *in-place* (without using an extra storage array) and independently of all other butterflies. It is at this level of the algorithm that parallelization may occur. The $\log_2 N$ passes through the data must be performed sequentially, so there is no parallelism at a coarser level.

The two basic strategies for parallelizing an FFT are called *scheduling-on-pairs* and *scheduling-on-ω*. In the former strategy, each processor is assigned independent butterflies to perform. Suppose there are p processors. The butterflies are passed out by giving the j^{th} processor the j^{th} butterfly, and every p^{th} butterfly thereafter. Prior to performing the butterfly, the processor must calculate the appropriate power of ω. At the end of each pass through the data, each processor must wait in a synchronization step until all other processors have completed that pass.

During each of the $\log_2 N$ passes through the data, the number of powers of ω required changes. At the k^{th} pass, there are 2^{k-1} distinct powers required. This fact gives rise to the scheduling-on-ω strategy. If each processor is assigned every p^{th} butterfly, as suggested above, then several processors may have to compute the same powers of ω as they stride through the data, an obvious duplication of effort.

This is unavoidable in the early passes of the algorithm, where $2^{k-1} < p$. During the later passes, where $2^{k-1} > p$, each processor is assigned all the butterflies corresponding to a given set of powers of ω. This strategy avoids the duplication of effort in having several processors compute the same sets of exponentials.

Briggs [7] implemented both of these parallization strategies on the Denelcor HEP computer. It was found that the scheduling-on-ω ran faster than scheduling-on-pairs by 25% (for small N) to 80% (for large N).

The Parallel E Algorithm

The parallel algorithm for computing the transform of an E sequence may now be developed. Only the ordered-to-scrambled case is considered, but the commentary and analysis extend readily to the scrambled-to-ordered case, although the computational details differ. Further, for shared memory computers, the extensions are immediate to parallel algorithms for real, real and odd, and real quarter wave (even or odd) sequences. For consideration of symmetric FFTs on distributed memory architectures, see Sweet [8], or Henson [9].

The basic assumption regarding the hardware is that the number of processors is small compared to the sequence lengths (coarse grained processing), and that all of the processors share a common memory. This assumes that there are no explicit communication costs in the algorithm. There will, however, be some overhead that must be paid for fork-join operations, and there will be some implicit communication cost in the form of memory bus contention. Since all the processors have equal access to all of the data, the algorithm is distributed among the processors.

At the beginning of the j^{th} pass through the data (j=0,1,...$\log_2 N-1$), copies of the subroutine are "forked" to the processors. The "units" of type EQE are then distributed as evenly as possible across the processors. If the current pass is not the first, RRQE butterflies are distributed as evenly as possible across the processors. After the second pass RtoR butterflies are required. The number of RtoR butterflies per sequence decreases with each pass, but the number of sequences increases. There are two cases that must be handled. If the number of RtoR butterflies per sequence is greater than the number of processors, the algorithm distributes the butterflies as evenly as possible across the processors, each of which strides through the sequences performing its designated butterflies. This mode is called *scheduling-on-butterflies.* If, however, there are fewer butterflies per sequence than processors, then the algorithm distributes the sequences across the processors as best it can, and each processor must compute all the butterflies for each of its sequences. This mode is called *scheduling-on-sequences.* At the end of each pass the processors are joined in a synchronization step. If the current pass is the last pass, all the butterflies are of the EQE type, and these are distributed across the processors.

Considering the second parallel strategy, there are two causes of decreased parallelism. The first is simple divisibility. When the number of processors does not divide the number of work units to be performed, there will be a time in which some processors are busy while others must wait. The second cause is the duplication of effort required when the algorithm switches from scheduling-on-butterflies to scheduling-on-sequences. In scheduling-on-butterflies, each processor need only calculate the one set of cosines for each set of butterflies it performs. While scheduled on sequences, however, all the processors must calculate all of the cosines for each sequence, implying duplication of effort.

Complexity of the Parallel Algorithm

To predict the speedup due to the parallel implementation consideration must be given to several factors: the changing amount of work of each uncombine type, the cost of the change from scheduling-on-butterflies to scheduling-on-sequences, the divisibility problems, and the cost of the fork-join operation. This leads to an analytic expression involving six terms: the cost of the fork-joins, the cost of the EQE units, the cost of the RRQE butterflies, the cost of the RtoR scheduled on butterflies, the cost of the RtoR scheduled on sequences, and the cost of the last pass through the data. This can be written:

$$T_p = T_f(p,N) + 4a \sum_{j=0}^{\log_2 N-2} \left\lceil \frac{\frac{N}{8}(\frac{1}{2})^j}{p} \right\rceil + A \sum_{j=1}^{\log_2 N-2} \left\lceil \frac{\frac{N}{4}(\frac{1}{2})^j}{p} \right\rceil$$

$$+ \sum_{j=2}^{LT-1} (2^{j-1}-1) \left[B_1 \left\lceil \frac{\frac{N}{4}(\frac{1}{2})^j}{p} \right\rceil + 2c \right]$$

$$+ B_2 \sum_{j=LT}^{\log_2 N-2} \left\lceil \frac{2^{j-1}-1}{p} \right\rceil \frac{N}{4}(\frac{1}{2})^j + 2a \left\lceil \frac{N}{4p} \right\rceil$$

where a is the cost of one real addition, c the cost of obtaining a cosine from the transcendental library, A the cost of an RRQE butterfly, B_2 the cost of an RtoR butterfly, and B_1 the cost of an RtoR butterfly without the cosines. LT is the index of the first pass through the data in which the RtoR portion must be scheduled on sequences, rather than butterflies.

The overhead for forking operations is given by the expression

$$T_f(p,N) = \alpha_1 + \beta_1(p-1) + (log_2 N - 1)(\alpha_2 + \beta_2(p-1))$$

where α_1 is the cost of the first fork on the first processor, β_1 is the cost per additional processor for the first fork. All succeeding fork calls have a cost of α_2 for the first processor and β_2 for each additional processor.

The complexity equation is difficult to analyze because of the least integer function which occurs in most of the terms. In cases where the number of processors is a power of two, the least integer functions are easily computed, and after some algebraic labor, the complexity equation reduces to

$$T_p = T_f(p,N) + \frac{N}{p}\left[\frac{B_1}{8}\log_2 N + \frac{c}{4}\log_2 p + \frac{(3a+c-B_1)}{2} + \frac{A}{8}\right]$$

$$+ (A+4a+2c)\log_2 p - 2c\log_2 N - A + 2c + B_1.$$

Regrouping the terms of this equation, the structure of the performance model consists of four terms:

$$T_p = T_f(p,N) + O(\frac{N}{p}\log_2 N) + O(\frac{N}{p}\log_2 p) + O(\log_2 p)$$

Each of the terms of this equation can be identified with the phenomenon it represents. The first, $T_f(p,N)$, is the overhead required to fork processes. The $O(\frac{N}{p}\log_2 N)$ term represents perfect speedup relative to the serial algorithm. The remaining two terms reflect decreased parallelism. The first, $O(\frac{N}{p}\log_2 p)$, is the amount of time spent in the duplication of effort caused by changing from RtoR scheduled on butterflies to RtoR scheduled on sequences. The last term, $O(\log_2 p)$, represents the amount of time spent in EQE and RRQE butterflies after the sequences become sufficiently short that there are fewer of these type butterflies than there are processors.

The Parallel Implementation

The algorithm taking ordered E data to scrambled coefficients was implemented on a Sequent Balance multiprocessor. The maximum number of processors available to one user was 23. All of the processors had access, through a common bus, to all of the data. To compute predicted performance curves, the timing parameters of the machine were obtained from the Sequent documentation, and then verified by experiment.

Timings of the actual transforms were obtained for sequences of various lengths, and speedups compared with the values predicted by the performance model. The results are shown in two separate charts: Figure 3 shows the performance characteristics of long sequences and Figure 4 displays those of shorter sequences.

For very long sequences, (N=32768, N=65536, N=131072), the implementation performed very well. There is good speedup throughout, with speedup generally increasing with increasing processors. The maximum speedup achieved was just over 17, occurring on the longest sequence when transformed on 21 processors. The

open circles represent the predicted speedups from the model. For long sequences, the actual speedup very nearly matches the model speedup.

On shorter sequences, (N=16384, N=8192, N=4096), the implementation performed less well, both from a standpoint of measured speedup alone, and when compared with the model speedup. In all cases, there is a significant decrease in the efficiency as the number of processors is increased, and on each curve there is an "optimal" number of processors, after which the transform requires more time to perform as the number of processors is increased. On a sequence of length 16384, for example, the best performance was achieved using 17 processors, resulting in a speedup of approximately 8. On an 8192 point sequence, however, the best results occurred with 9 processors, but achieved a speedup of only 5. Additionally, as sequence lengths become shorter, the actual performance differs more and more from the predicted curve. This may be attributed to two factors. First, the overhead of loop indexing is not included in the model. As sequences become shorter, the loop indexing represents an increasing fraction of the algorithm. A more significant factor is related to the memory management of the Sequent Balance. Essentially, as sequences become shorter, the memory accesses by the processors become more frequent, and the bus becomes saturated. (It should be noted that the new generation of Sequent multiprocessors, the Symmetry family, utilize a different memory management scheme designed specifically to eliminate this effect.)

Conclusions

The transform for real and even data is one member of a family of algorithms that efficiently compute the transforms of symmetric sequences. The serial versions of these compact symmetric algorithms provide a tremendous savings over the direct use of the complex FFT to transform these sequences. They also offer a savings over traditional pre- and post- processing algorithms, using the same total storage, but requiring somewhat fewer arithmetic operations.

The compact symmetric algorithms have straightforward extensions to shared memory parallel computers, and produce additional savings from parallelization. A major benefit is that FFTs are usually performed as part of some larger calculation, which in turn is made more efficient. This is especially true for many of the symmetric sequences, that arise in the direct solution of partial differential equations with various boundary conditions. Much recent research [10] has centered on improving the performance of these larger computations by implementing them on parallel machines. The utilization of the family of parallel compact symmetric FFTs should represent a significant contribution to that effort.

Acknowledgements

The authors would like to thank Roland A. Sweet for his support and assistance. The Argonne National Laboratory provided time on the Sequent Balance 21000 computer. This research was supported by National Science Foundation grant number DMS-8611325.

REFERENCES

[1] Cooley, J. W., and Tukey, J. W., (1965), *An Algorithm for the Machine Calculation of Complex Fourier Series*, Math. Comp., v. 19, pp. 297-301.

[2] Bergland, G., D., (1968), *A Fast Fourier Transform for Real Valued Series*, Comm. ACM, v. 11, pp. 703-710.

[3] Gentleman, W. M., (1972), *Implementing Clenshaw-Curtis Quadrature, Computing the Cosine Transformation*, Comm. ACM, v. 15, pp. 343-346.

[4] Cooley, J. W., Lewis, P. A. W., and Welsh, P. D., (1970), *The Fast Fourier Transform Algorithm: Programming Considerations in the Calculation of Sine, Cosine, and Laplace Transforms*, J. Sound Vibration, v. 12, pp. 315-337.

[5] Swartztrauber, P. N., (1986), *Symmetric FFTs*, Math. Comp., v. 47, pp. 323-346.

[6] Briggs, W. L., (1987), *Further Symmetries of In-Place FFTs*, SIAM Sci. and Stat. Comp., v. 8, pp. 644-655.

[7] Briggs, W. L., Hart, L. B., Sweet, R. A., and O'Gallagher, A., (1987), *Multiprocessor FFT Methods*, SIAM Sci. and Stat. Comp., v. 8, pp. s27-s43.

[8] Sweet, R. A., Porsche, J. A., and Henson, V. E., (in preparation), *A Fast Fourier Transform for Real Data on a Hypercube*

[9] Henson, V. E., (in preparation), *A Comparison of Symmetric FFTs on Shared and Distributed Memory Parallel Processors*

[10] Swarztrauber, P. N., and Sweet, R. A., (to appear), *Vector and Parallel Methods for the Direct Solution of Poisson's Equation*, J. Comp. Appl. Math.

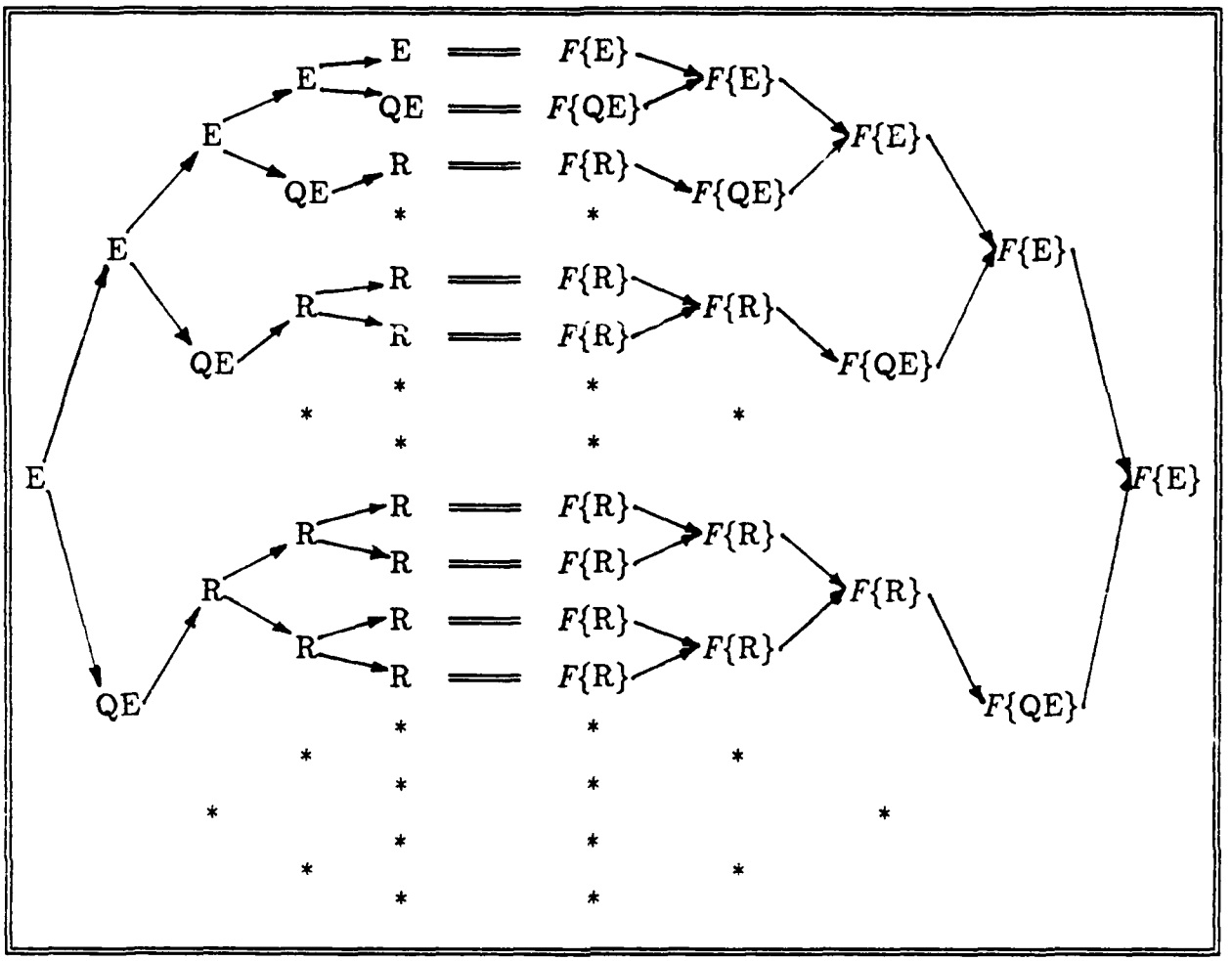


Figure 1

Schematic diagram of the Swarztrauber Algorithm. $F\{\ \}$ indicates the Discrete Fourier Transform. The asterisks represent the redundant R sequences which are not calculated or stored. The portion of the diagram to the left of the column of bars is the ordering phase, that to the right of the bars is the combine phase. The column immediately to the left of the column of bars is the E sequence in scrambled order, the final $F\{E\}$ on the right is the transform in natural order.

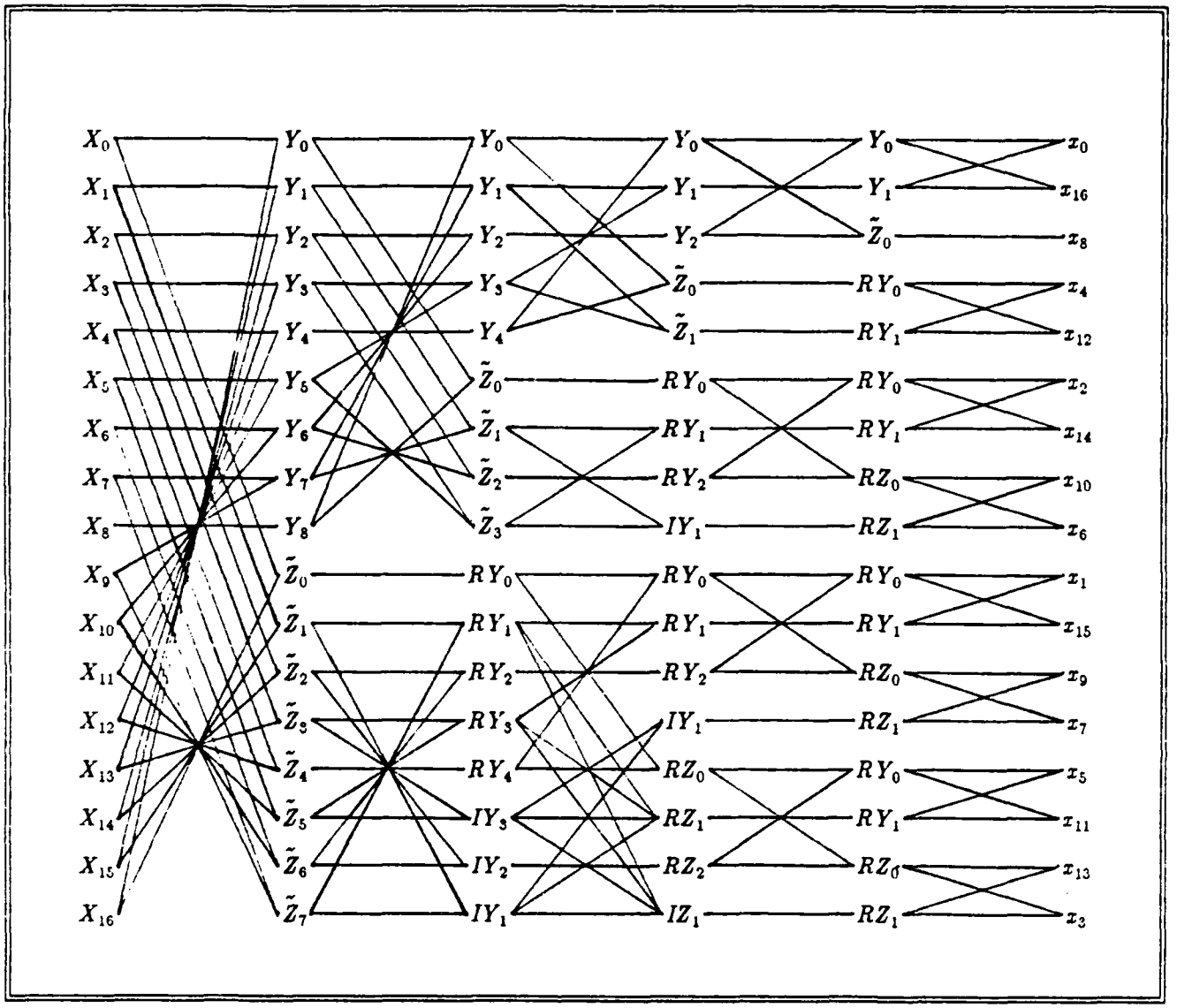


Figure 2

Storage and data flow diagram for compact real and even transform, with $N = 32$, taking ordered data to scrambled coefficients. R and I refer to the real and imaginary parts of the complex quantity. During each pass through the data, the first set of lines is the EQE uncombine, the second set is the RRQE uncombine, and all other sets are RtoR uncombines.

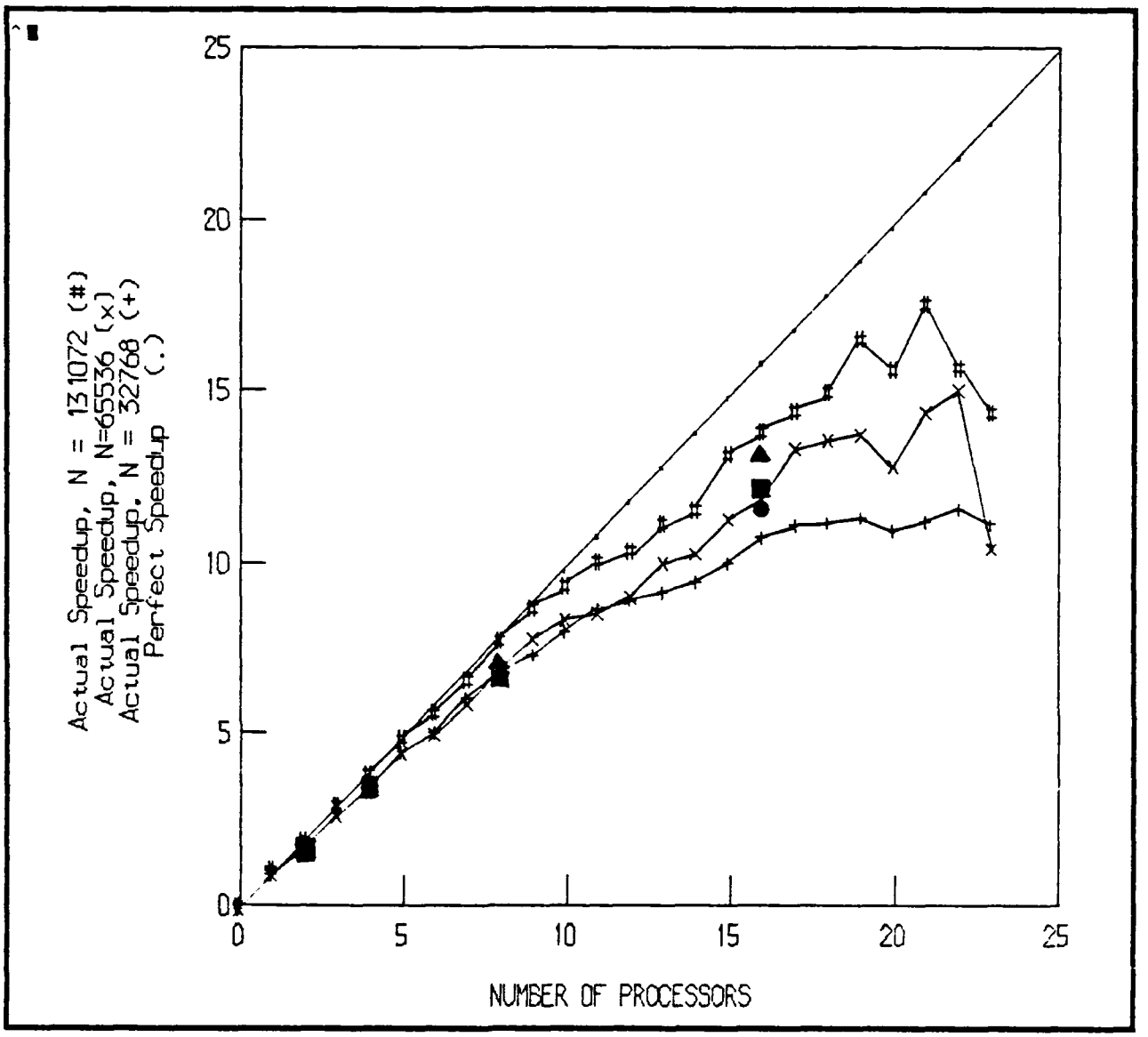


Figure 3

Measured Speedup for long sequences. Speedup curves are shown for $N = 131072$ (#), $N = 65536$ (x), and $N = 32768$ (+). Perfect speedup is represented by the diagonal line. Theoretical speedup is plotted at $p = 2, 4, 8$ for $N = 131072$ (solid triangle), $N = 65536$ (solid square), and $N = 32768$ (solid circle). At $p = 4$ only the square is plotted, as all three computed values fell within the size of the square.

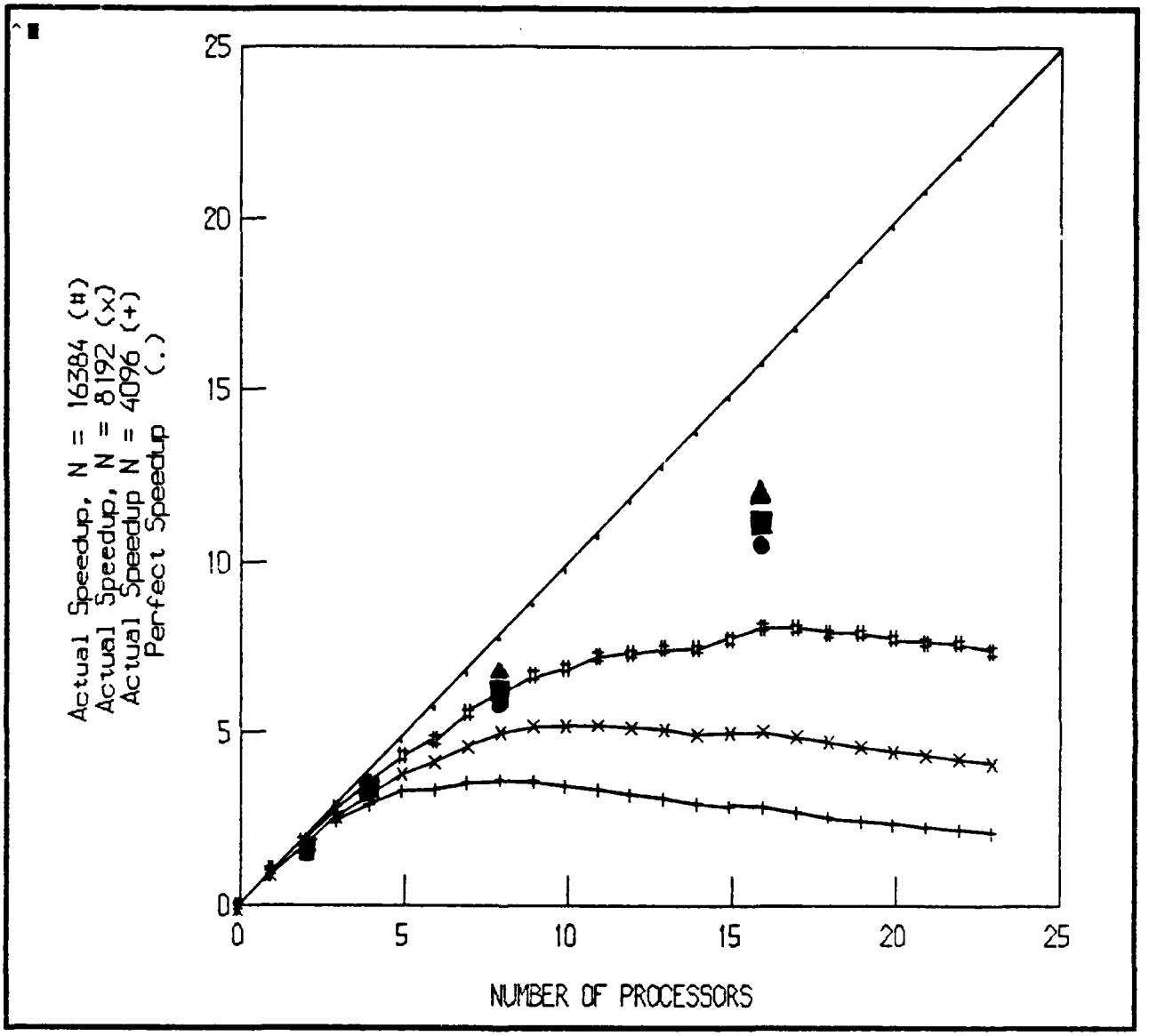


Figure 4

Measured Speedup for "short" sequences. Speedup curves are shown for $N = 16384$ (#), $N = 8192$ (x), and $N = 4096$ (+). Perfect speedup is represented by the diagonal line. Theoretical speedup is plotted at $p = 2, 4, 8$ for $N = 163842$ (solid triangle), $N = 8192$ (solid square), and $N = 4096$ (solid circle). At $p = 4$ only the square is plotted, as all three computed values fell within the size of the square.

Parallel Methods for Block Bordered Nonlinear Problems

Xiaodong Zhang, Richard Byrd, Robert Schnabel

Department of Computer Science
University of Colorado at Boulder
Boulder, Colorado 80309

Abstract. We discuss a group of parallel algorithms, and their implementations, for solving a special class of nonlinear equations arising in VLSI design, structural engineering and other areas. The class of sparsity occurring in these problems is called *block bordered* structure. We present the explicit method and several implicit methods for solving block bordered nonlinear problems, and give some mathematical analysis and comparisons of the two methods. Several variations and globally convergent modifications of the implicit method are also described. We present computational results on a sequential computer that help compare and justify the efficiency of the algorithms. Finally, the implementations on shared memory multiprocessors and local memory multiprocessors, are discussed.

1. Introduction.

The solution of a system of nonlinear equations is one of most basic and important problems encountered in many applications. The general form of a system of nonlinear equations is :

$$f_i(x_1, x_2, ..., x_n) = 0, \qquad i=1, ..., n. \tag{1.1}$$

Several parallel algorithms for solving (1.1) have been developed and implemented on some parallel computers. Newton's method is the main approach in those algorithms. Thus, most algorithms for solving (1.1) consist mainly of solving the linear Jacobian system. Many parallel algorithms have been developed for solving a linear system, such as parallel factorizations, parallel SOR method, parallel red and black method, parallel multicolor and so on (see e.g. Ortega and Voigt [1985]). One of the typical parallel Newton methods for solving (1.1) is called Newton-Jacobi(or Newton-SOR, or Newton-Gauss-Seidel), in which the main iteration is the Newton iteration for solving (1.1), and the inner loop is to solve the linear system iteratively by using the Jacobi method (or SOR method, or Gauss-Seidel method) (see e.g. O'Leary and White [1985], White [1986]). Fontecilla [1987] gives a parallel implementation of a different approach, the serial nonlinear Jacobi algorithm for (1.1). This algorithm is based on the same idea as the Jacobi algorithm

This research is partially supported by ARO contracts DAAL-03-k-0086 an AFOSR grant AFOSR-85-0251.

for solving linear systems of equations. The Jacobi (or SOR) is the primary iteration, and Newton iterations are used to approximately solve the *jth* block of equations for the *jth* block of variables in the inner loop where $j = 1, ..., m$ for $m \leq n$. This method is called the Jacobi-Newton method. Coleman and Li [1987] develop parallel algorithms for the solution of (1.1) on a message-passing multiprocessor computer with a distributed finite-difference Newton method, a multiple secant method and a rank-1 secant method.

In the case of very large nonlinear problems one cannot expect a single parallel algorithm to handle the all instances of the nonlinear problem (1.1) efficiently, but rather the algorithm must take into account the sparsity structure and other special characteristics of the problem. In fact, many nonlinear problems in the applications have their own special sparsity structure. Parallel algorithms taking advantage of the special structure can be much more efficient than the algorithms ignoring the special structure. In this paper we give a group of parallel algorithms and implementations for solving a special class of nonlinear equations arising in VLSI design, structural engineering and other areas. The class of sparsity occurring in these problems is called block bordered structure. In such a problem the n variables and equations may be grouped into $q+1$ subvectors, $x_1, ..., x_{q+1}$ and $f_1, ..., f_{q+1}$ such that the nonlinear system of equations has the form

$$\begin{aligned} f_i(x_i, x_{q+1}) &= 0; \qquad i = 1, ..., q \\ f_{q+1}(x_1, ..., x_{q+1}) &= 0 \end{aligned} \tag{1.2}$$

where

$$x_i \in R^{n_i}, \ f_i \in R^{n_i}, \quad i = 1, ..., q+1, \text{ and } \sum_{i=1}^{q+1} n_i = n.$$

The block bordered Jacobian matrix of (1.2) is as follows:

$$\begin{bmatrix} A_1 & & & & B_1 \\ & A_2 & & & B_2 \\ & & \ddots & & \vdots \\ & & & A_q & B_q \\ C_1 & C_2 & \cdots & C_q & P \end{bmatrix} \tag{1.3}$$

where

$$A_i = \frac{\partial f_i}{\partial x_i} \in R^{n_i \times n_i}, \ i = 1, ..., q,$$

$$B_i = \frac{\partial f_i}{\partial x_{q+1}} \in R^{n_i \times n_{q+1}}, \ i = 1, ..., q,$$

$$C_i = \frac{\partial f_{q+1}}{\partial x_i} \in R^{n_{q+1} \times n_i}, \ i = 1, ..., q,$$

$$P = \frac{\partial f_{q+1}}{\partial x_{q+1}} \in R^{n_{q+1} \times n_{q+1}}.$$

In VLSI design, P is a permutation matrix and the structure of B_i is in the form of

$$\begin{bmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_q \end{bmatrix} \tag{1.4}$$

i.e. only one B_i is nonzero in any given column of the right-hand border (see e.g. Rabbat et al [1979], [1980]). In addition, the equations f_{q+1} are linear. We will concentrate on systems with this special structure.

We will give a group of parallel algorithms for solving block bordered nonlinear system of form (1.2) which may be implemented on both shared memory multiprocessors, such as the Encore Multimax, and local memory multiprocessors, such as the Intel hypercube. In section 2, we give some background of the block bordered problems, and survey related work. Section 3 presents the explicit method and implicit method for solving the block bordered nonlinear problem, and gives some mathematical analysis and comparisons of the two methods. Several variations of the implicit method are also described. Experiments with the two methods on a sequential computer are given based on the analysis of the section. Global strategies for the different methods and their implementations, are given in section 4. Parallel versions of these methods are described in section 5. Our conclusions and some future research directions are summarized in section 6.

2. Background and related work.

Block bordered systems of equations having the form (1.2) arise in many areas of engineering and science, and a few algorithms have been developed to solve them. In structural engineering, models of large structures may be divided into q regions such that each region only interacts directly with neighboring regions. The x_i are the variables for each region, and the extra linking variables (the x_{q+1}) are introduced at the boundaries of the regions. The linking variables are tied together with an $q+1$st set of equations representing the interactions between the regions. Thus the equilibrium equations for such a model will be of the form (1.2). In addition, the Jacobian matrix is symmetric, i.e. $B_i = C_i$, and often the sizes of A_i are the same. One current parallel algorithm to solve the problem in the linear case (see e.g. Farhat and Wilson [1986]) is to let each processor hold the pair $\{A_i, B_i\}$ as well as f_i and x_i. Then the updates $\{x_i, i = 1, ..., q\}$ can be all performed concurrently by solving the subsystem in parallel:

$$A_i x_i^{k+1} = f_i - B_i x_{q+1}^k, \quad i = 1, \cdots q \tag{2.1}$$

and the components x_i are updated locally in each processor using sequential SOR iterations. It remains to update the unknowns associated with block P. This block is coupled to all the A_i terms. If its size is negligible compared to each of the sizes of the diagonal blocks, the overall algorithm will suffer a serialization for only a small amount of time. If not (and this usually the case for three dimensional structures), the updates of x_{q+1} may ruin the sought after speed-up. The algorithm is simple to implement, and efficient for

the special engineering structure problem because the problem is linear and the function of f_{q+1} is relatively small. Because the coefficient matrix (1.3) of the linear system is symmetric and positive definite, a parallel conjugate gradient method is also efficient for its solution (see e.g. Nour-Omid and Park [1986]).

Similar equations arise in the analysis of VLSI design, where the circuits may be subdivided. The concept of macromodeling the circuit is to decompose the circuit into subcircuits and to analyze them separately. Macromodeling of the circuit results in a system of nonlinear equations of the form (1.2). x_i $(i = 1, ..., q)$ and x_{q+1} in the Jacobian matrix are usually used to represent internal and input-output variables in each of the q independent subcircuits respectively. Here the equations involve voltages and currents, either of which between the subcircuits serve a linking role which results in the function f_{q+1}. Since each voltage or current is used only in one block of equations f_i plus possibly the bottom block f_{q+1}, the nonzero columns of B_i (and A_i) are disjoint and so the form (1.4) results. The size of the function f_{q+1} may be quite large.

Two nested sequential algorithms taking advantage of the structural properties of VLSI circuits have been developed by Rabbat et al [1979], [1980]. The multi-Newton method is to apply Newton's method to f_{q+1} of (1.2), where x_i, $(i = 1, ..., q)$ are implicitly determined by the f_i of (1.3), and another Newton method is applied to solve them in the inner loop. This is discussed further in section 3. Similarly, the Gauss-Seidel-Newton method is to apply the Gauss-Seidel method to f_{q+1} of (1.2), where x_i $(i = 1, ..., q)$ are implicitly determined by f_i of (1.2), and the Newton method is applied to solve them in the inner loop. These algorithms appear suitable for implementation on parallel computers, but to our knowledge this has not been done.

3. Explicit and implicit approaches to the problem

There are two basic ways in which Newton's method can be applied to (1.2). The explicit approach for solving (1.2) is related to Newton's method, which simply involves iteratively solving the linear system

$$J(X^k)\Delta X^k = -F(X^k) \qquad i = 1, ..., q \tag{3.1}$$

for ΔX^k, where $J(X^k)$ is the Jacobian of F, which has the block bordered structure of (1.3).

The implicit approach is to solve or approximately solve each of the q equations

$$f_i(x_i, x_{q+1}) = 0, \qquad i = 1, ..., q \tag{3.2}$$

for a fixed value of x_{q+1}. This would mean that each of the x_i is implicitly given by a function of x_{q+1}. The the whole problem (1.2) is then equivalent to solving

$$f_{q+1}(x_1(x_{q+1}), ..., x_q(x_{q+1}), x_{q+1}) = 0. \tag{3.3}$$

The Jacobian of this system is given by

$$\hat{f} = \frac{\partial f_{q+1}}{\partial x_{q+1}} - \sum_{i=1}^{q} \frac{\partial f_{q+1}}{\partial x_i}\left(\frac{\partial f_i}{\partial x_i}\right)^{-1}\frac{\partial f_i}{\partial x_{q+1}} \qquad i = 1, ..., q \tag{3.4}$$

or

$$\hat{J} = P - \sum_{i=1}^{q} C_i A_i^{-1} B_i \quad i = 1, ..., q \tag{3.4}$$

and we may solve (3.3) by Newton's method.

In this section we describe these two approaches and their relations to each other, and give some experimental results on a sequential computer.

3.1 Explicit method and implicit method

Newton's method applied to (1.2) in the explicit method consists of the following formulas at iteration k $(k = 0, 1, ...)$: from $f_i(x), i = 1, ..., q$,

$$\frac{\partial f_i}{\partial x_i^k} \Delta x_i^k + \frac{\partial f_i}{\partial x_{q+1}^k} \Delta x_{q+1}^k + f_i(x_i^k, x_{q+1}^k) = 0 \tag{3.1.1}$$

or equivalently

$$A_i \Delta x_i^k + B_i \Delta x_{q+1}^k + f_i(x_i^k, x_{q+1}^k) = 0 \tag{3.1.1}$$

and from $f_{q+1}(x_1^k, ..., x_q^k, x_{q+1}^k)$

$$\sum_{i=1}^{q} \frac{\partial f_{q+1}}{\partial x_i} \Delta x_i^k + \frac{\partial f_{q+1}}{\partial x_{q+1}} \Delta x_{q+1}^k + f_{q+1}(x_1^k, ..., x_q^k, x_{q+1}^k) = 0 \tag{3.1.2}$$

or equivalently

$$\sum_{i=1}^{q} C_i \Delta x_i^k + P \Delta x_{q+1}^k + f_{q+1}(x_1^k, ..., x_q^k, x_{q+1}^k) = 0. \tag{3.1.2}$$

Substituting (3.1.1) into (3.1.2), we obtain

$$(P - \sum_{i=1}^{q} C_i A_i^{-1} B_i) \Delta x_{q+1}^k = -f_{q+1}(x_1^k, ..., x_q^k, x_{q+1}^k) + \sum_{i=1}^{q} C_i A_i^{-1} f_i(x_i^k, x_{q+1}^k) \tag{3.1.3}$$

or

$$\hat{J} \Delta x_{q+1}^k = -f_{q+1}(x_1^k, ..., x_q^k, x_{q+1}^k) + \sum_{i=1}^{q} C_i A_i^{-1} f_i(x_i^k, x_{q+1}^k) \tag{3.1.3}$$

where $\hat{J}$ is given by (3.4). So

$$x_{q+1}^{k+1} = x_{q+1}^k + \Delta x_{q+1}^k \tag{3.1.4}$$

can be determined from (3.1.3), and

$$x_i^{k+1} = x_i^k + \Delta x_i^k \quad i = 1, ..., q \tag{3.1.5}$$

can be determined from (3.1.1).

In the implicit method, Newton's method is applied to (3.3), and gives

$$[\frac{\partial f_{q+1}}{\partial x_{q+1}} - \sum_{i=1}^{q} \frac{\partial f_{q+1}}{\partial x_i}(\frac{\partial f_i}{\partial x_i})^{-1} \frac{\partial f_i}{\partial x_{q+1}}] \Delta x_{q+1}^k + f_{q+1}(x_1^{k+1,0}(x_{q+1}), ..., x_q^{k+1,0}(x_{q+1}), x_{q+1}^k) = 0 \quad (3.1.6)$$

or

$$\hat{J} \Delta x_{q+1}^k + f_{q+1}(x_1^{k+1,0}(x_{q+1}^k), ..., x_q^{k+1,0}(x_{q+1}^k), x_{q+1}^k) = 0. \quad (3.1.6)$$

where $x_i^{k+1,0}(x_{q+1}^k)$ $(i = 1, ..., q)$ is implicitly determined by solving the nonlinear system

$$f_i(x_i^{k,j}, x_{q+1}^k) = 0 \quad (3.1.7)$$

for $x_i^{k,j}$. Here, j is the inner iteration number for solving (3.1.7) for x_i, and k is outer iteration number for solving (1.2). We use a second (or inner) Newton process on (3.1.7) to evaluate $x_i(x_{q+1})$, which yields

$$\frac{\partial f_i}{\partial x_i} \Delta x_i^{k,j-1} + f_i(x_i^{k,j-1}, x_{q+1}^k) = 0 \quad i = 1, ..., q, \ j = 1, 2, .. \quad (3.1.8)$$

or

$$\hat{A}_i \Delta x_i^{k,j-1} + f_i(x_i^{k,j-1}, x_{q+1}^k) = 0 \quad i = 1, ..., q, \ j = 1, 2, .. \quad (3.1.8)$$

where $\hat{A}_i = A_i$ if it is only evaluated once at the beginning of each outer iteration, else it may be evaluated up to j times. This second Newton process is at a lower level since x_{q+1} is determined from (3.1.6) and is held fixed in (3.1.8). Thus, $j = 1, 2, \cdots$, and k is fixed for the outer loop. Then

$$x_i^{k,j} = x_i^{k,j-1} + \Delta x_i^{k,j-1} \quad (3.1.9)$$

When $x_i^{k,j+1}$ exits from the inner loop, it is set to

$$x_i^{k+1,0} = x_i^{k,j}. \quad (3.1.10)$$

Then, x_{q+1} is determined from (3.1.6), and

$$x_{q+1}^{k+1} = x_{q+1}^k + \Delta x_{q+1}^k \quad (3.1.11)$$

3.2 Comparisons and analysis of the two methods

The following theorems show that the explicit and implicit methods are very closely related.

Theorem 1. If the function f_{q+1} of the nonlinear block bordered problem (1.2) is linear, then the equation solved for x_{q+1}^k $(k = 0, 1, \cdots)$ in the implicit method is equivalent to the one in the explicit method, except that the value of x_i that is used may be different.

Proof: Substituting (3.1.8) into (3.1.9), gives the implicit formula for solving the x_i $(i = 1, ..., q)$:

$$x_i^{k,j} = x_i^{k,j-1} - \hat{A}_i^{-1} f_i(x_i^{k,j-1}, x_{q+1}^k) \quad (3.2.1)$$

where $j = 1, 2, \cdots$, and k is fixed, and $\hat{A}_i = A_i$ if it is only evaluated once at the beginning, else it may be evaluated up to j times.

Substituting (3.2.1) to (3.1.6) with the condition of linear f_{q+1} gives

$$\hat{J}\Delta x_{q+1}^k = -f_{q+1}(x_1^{k,j-1}, \ldots, x_q^{k,j-1}, x_{q+1}^k) + \sum_{i=1}^{q} C_i \hat{A}_i^{-1} f_i(x_i^{k,j-1}, x_{q+1}^k). \tag{3.2.2}$$

(3.2.2) is equivalent to (3.1.3) which is the explicit formula, except different variables may be applied in the two formulas.

Theorem 2. If f_{q+1} is linear and only one Newton iteration is applied to solve $x_i^{k,j}$ $(i = 1, \ldots, q)$, i.e. $j = 1$, in the implicit method, then the steps Δx_{q+1}^k (for a fixed k) are identical in both methods.

Proof: This follows immediately by substituting $j = 1$ to (3.2.2):

$$\hat{J}\Delta x_{q+1}^k = -f_{q+1}(x_1^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k) + \sum_{i=0}^{q} C_i A_i^{-1} f_i(x_i^{k,0}, x_{q+1}^k) \tag{3.2.3}$$

which is identical to the explicit formula (3.1.3).

Theorem 3. If $-A_i^{-1}B_i\Delta x_{q+1}$ is added to the right hand side of (3.1.10), then the equation solved for Δx_i $(i = 1, \ldots, q)$ in the implicit method is equivalent to the one in the explicit method except that the value of x_i that is used may be different.

Proof: Adding $-A_i^{-1}B_i\Delta x_{q+1}$ to (3.1.10), and substituting (3.1.9) and (3.1.8) into (3.1.10), gives

$$x_i^{k+1,0} = x_i^{k,j-1} - A_i^{-1}[f_i(x_i^{k,j-1}, x_{q+1}^k) - B_i\Delta x_{q+1}^k] \tag{3.2.4}$$

which is equivalent to the explicit formula (3.1.1) is substituted by (3.1.5), except different variables may be applied in the two formulas.

Theorem 4. If f_{q+1} is linear and only one Newton iteration is applied to solve $x_i^{k,j}$ $(i = 1, \ldots, q)$. i.e. $j = 1$, in the implicit method, and the system is corrected by adding $-A_i^{-1}B_i\Delta x_{q+1}$ to x_i after each iteration, then the explicit method and implicit method are identical.

Proof: From *Theorem 2*, Δx_{q+1}^k are identical for the two methods. Substituting $j = 1$ into (3.2.4):

$$x_i^{k+1,0} = x_i^{k,0} - A_i^{-1}[f_i(x_i^{k,0}, x_{q+1}^k) - B_i\Delta x_{q+1}^k]. \tag{3.2.5}$$

which is identical to the explicit method and completes the proof.

The following theorems give the local convergence rates of the explicit and implicit methods. Theorem 5 results from standard theory. The proof of theorem 6 is given in Zhang [1989].

Theorem 5. Assume that $F(x)$ is continuously differentiable in an open convex set $D \in R^n$. Assume that there exists $x^* \in R^n$ such that $F(x^*) = 0$. $J(x^*)$ is nonlinear, and $J(x)$ is Lipschitz continuous in an open neighborhood containing x^*. Then the explicit Newton's method is locally quadratically convergent to x^*.

Theorem 6. Let the assumption of theorem 5 hold, and assume the addition that each $A_i(x^*)$ is nonsingular, and that each $A_i(x)$ is Lipschitz continuous in an open neighborhood containing x^*, then the implicit Newton method with one inner iteration per outer iteration ($j = 1$) is locally 2-step quadratically convergent to x^*.

3.3 A corrected implicit method

Theorem 3 and *Theorem* 4 indicate that the implicit method may obtain the same quadratic rate of convergence as the explicit method even if the inner iteration is solved inexactly, if a correction term is added to (3.1.12) after each iteration. The problem may be defined to find a correction term δ such that

$$f_i(x_i^{k+1,0} + \delta, x_{q+1}^k) \approx 0. \tag{3.3.1}$$

or

$$f_i(x_i^{k+1,0} + \delta, x_{q+1}^k + \Delta x_{q+1}^k) \approx 0. \tag{3.3.1}$$

(3.3.1) may be approximated by treating the function f_i to be linear, then

$$f_i(x_i^{k+1,0}, x_{q+1}^k) + A_i\delta + B_i\Delta x_{q+1}^k = 0. \tag{3.3.2}$$

The correction term δ is obtained from (3.3.2)

$$\delta = -A_i^{-1}[f_i(x_i^{k+1,0}, x_{q+1}^k) + B_i\Delta x_{q+1}^k] \tag{3.3.3}$$

After j inner iterations for solving $x_i^{k+1,0}$ for a fixed k, $f_i(x_i^{k+1,0}, x_{q+1}^k) \approx 0$. Thus we may make a further approximation for the correction term δ

$$\delta = -A_i^{-1}B_i\Delta x_{q+1}^k. \tag{3.3.4}$$

which is exactly the correction term we have used in *Theorem* 3.

We prove a lemma showing that one step of the corrected implicit method has a similar structure to one step of the explicit method.

Lemma 1. One step of the corrected implicit method with j inner iterations is identical to solving the following linear system of equations:

$$\begin{bmatrix} A_1 & & & B_1 \\ & A_2 & & B_2 \\ & & \ddots & \vdots \\ C_1 & C_2 & \cdots & C_q \quad P \end{bmatrix} \begin{bmatrix} \Delta x_1^k \\ \Delta x_2^k \\ \vdots \\ \Delta x_q^k \\ \Delta x_{q+1}^k \end{bmatrix} = - \begin{bmatrix} \sum_{i=0}^{j-1} f_1(x_1^{k,i}, x_{q+1}^k) \\ \vdots \\ \sum_{i=0}^{j-1} f_q(x_q^{k,i}, x_{q+1}^k) \\ f_{q+1}(x_1^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k) \end{bmatrix} \tag{3.3.5a}$$

and one step of the uncorrected implicit method with j inner iterations is identical to solving the following linear system of equations:

$$\begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_q \\ C_1 & C_2 & \cdots & C_q \quad \hat{P} \end{bmatrix} \begin{bmatrix} \Delta x_1^k \\ \Delta x_2^k \\ \vdots \\ \Delta x_q^k \\ \Delta x_{q+1}^k \end{bmatrix} = - \begin{bmatrix} \sum_{j=0}^{j-1} f_1(x_1^{k,j}, x_{q+1}^k) \\ \vdots \\ \sum_{j=0}^{j-1} f_q(x_q^{k,j}, x_{q+1}^k) \\ f_{q+1}(x_1^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k) \end{bmatrix} \tag{3.3.5b}$$

where $\hat{P} = \sum_{i=1}^{q} C_i A_i^{-1} B_i + P$.

Proof : Recall that x_i $(i = 1, \ldots, q)$ is calculated in the inner iterations using equations (3.1.8), (3.1.9), (3.1.10), and x_{q+1}^k is calculated following the inner iterations by (3.1.6).

Substituting $x_i^{k+1,0} = x_i^{k,0} + \sum_{j=0}^{j-1} \Delta x_i^{k,j}$ into (3.1.6) and using the linearity of f_{q+1}, we obtain

$$(P - \sum_{i=1}^{q} C_i A_i^{-1} B_i) \Delta x_{q+1}^k = -f_{q+1}(x_1^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k) + \sum_{i=1}^{q} C_i A_i^{-1} \sum_{j=0}^{j-1} f_i(x_i^{k,j}, x_{q+1}^k) \tag{3.3.6a}$$

or

$$\sum_{i=1}^{q} C_i [-A_i^{-1}(\sum_{j=0}^{j-1} f_i(x_i^{k,j}, x_{q+1}^k) + B_i \Delta x_{q+1})] + P \Delta x_{q+1}^k = -f_{q+1}(x_i^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k) \tag{3.3.6a}$$

where

$$-A_i^{-1}(\sum_{j=0}^{j-1} f_i(x_i^{k,j}, x_{q+1}^k) + B_i \Delta x_{q+1}^k) = \sum_{j=0}^{j-1} \Delta x_i^{k,j} - A_i^{-1} B_i \Delta x_{q+1}^k. \tag{3.3.6b}$$

The right hand side of (3.3.6b) is the corrected step of $x_i^{k+1,j} - x_i^{k,0}$ after the outer iteration is complete. Let the corrected step be Δx_i^k, (3.3.6a) becomes

$$\sum_{i=1}^{q} C_i \Delta x_i^k + P \Delta x_{q+1}^k = -f_{q+1}(x_1^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^{k,0}) \tag{3.3.7a}$$

The equation (3.3.6b) for solving the corrected Δx_i^k may be converted to

$$A_i \Delta x_i^k + B_i \Delta x_{q+1}^k = -\sum_{j=0}^{j-1} f_i(x_i^{k,j}, x_{q+1}^k). \tag{3.3.8a}$$

This completes the proof of the first part of the lemma.

From (3.3.6a) and (3.1.8), one step of uncorrected implicit method is equivalent to solving

$$A_i \Delta x_i^k = -\sum_{j=0}^{j-1} f_i(x_i^{k,j}, x_{q+1}^k) \tag{3.3.7b}$$

and

$$\sum_{i=1}^{q} C_i \Delta x_i^k + (\sum_{i=1}^{q} C_i A_i^{-1} B_i + P) \Delta x_{q+1}^k = -f_{q+1}(x_i^{k,0}, \ldots, x_q^{k,0}, x_{q+1}^k). \tag{3.3.8b}$$

which is equivalent to solve the linear systems of equation of (3.3.5.b).

From theorem 4 and 5, we view that the corrected implicit method with one inner iteration ($j = 1$) is locally quadratically convergent. Theorem 7 shows that this rate of convergence is retained if more inner iterations are used. Its proof will be given in Zhang [1989].

Theorem 7. Let the assumptions of theorem 6 hold. Then the corrected implicit Newton method with $j \geq 1$ inner iterations per outer iteration is locally quadratically convergent to x^*.

3.4 Some experiments on a sequential processor

We have tested the methods discussed in this section on several problems. Here we report results on a simple 20×20 nonlinear block bordered system of equations which has four 4×4 blocks, $A_1, \cdots, A_4$, and a 4×4 bottom block P which is a 4×4 matrix, and f_{q+1} linear. First, we compare the performance of the three methods when only one inner iteration ($j = 1$) is used in the uncorrected implicit and corrected implicit methods. All these experiments were run on Pyramid P90 computer.

Experiments with the three methods (j=1) outer iterations (seconds)		
explicit	implicit	corrected implicit
13 (0.44)	14 (0.40)	13 (0.40)

The explicit method and the corrected implicit method with $j = 1$ are identical (see *Theorem* 4.). Thus, the same number of iterations are used to converge to the solutions. The computing times are slightly different since the implementations of the two methods are different. The implicit method converges a little bit slower than the other two methods, which is reasonable since our analysis shows it has a 2-step quadratic convergence rate. (see *Theorem* 6).

Next we increased the number of inner iterations in the implicit and corrected implicit methods.

Experiments with the implicit method (j>1) outer iterations (seconds)				
j = 1	j = 2	j = 3	j = 4	j = 5
14 (0.40)	8 (0.34)	7 (0.40)	6 (0.44)	6 (0.54)

Experiments with the corrected implicit method (j>1) outer iterations (seconds)				
j = 1	j = 2	j = 3	j = 4	j = 5
13 (0.40)	8 (0.38)	6 (0.36)	6 (0.50)	5 (0.54)

The experimental results show that the number of outer iterations is sharply decreased when the number of inner iterations is greater than 1. However, the number of outer iterations does not decrease forever as j increases. There exists an optimal j for the number of outer iterations, or for the least time in both methods, but it is problem dependent. Our experiments also show that the corrected implicit method converges a little bit faster than the uncorrected implicit method when $j > 1$, which is consistant with our convergence analysis.

4. Globally convergent modifications of the corrected implicit method

The corrected implicit method was shown to be locally quadratically convergent in the last section. In this section, we will give conditions for the steps generated by explicit method and uncorrected implicit method to be descent directions. We will also discuss the combination of a globally convergent strategy for the corrected implicit method with a fast local strategy. We let $||.||$ denote the l_2 (Euclidean) norm.

4.1 The conditions for a descent direction

The basic idea of a global method for solving block bordered nonlinear problems is to choose a direction ΔX from the current point X^k in which F decreases initially, and a new point X^{k+1} in this direction from X^k such that $||F(X^{k+1})|| < ||F(X^k)||$. Such a direction is called a *descent direction*. Mathematically, ΔX is a descent direction from X^k if the directional derivative p of $||F||^2$ at X^k in the direction ΔX is negative,i.e. if

$$p = -F(X)^T J(X)\Delta X < 0. \tag{4.1.1}$$

If (4.1.1) holds, then it is guaranteed that for sufficiently small positive δ, $||F(X^k + \delta\Delta X)|| < ||F(X^k)||$. Given a descent direction ΔX^k, we set $X^{k+1} = X^k + \lambda_k \Delta X^k$ for some $\lambda_k > 0$ that makes $||F(X^{k+1})|| < ||F(X^k)||$, where λ_k is chosen by a line search strategy. (see e.g. Dennis and Schnabel [1983]). The following theorems indicate when the directions generated by the explicit, implicit and corrected implicit methods are descent directions.

Theorem 8. The step generated by the explicit method is a descent direction on the function $||F(X)||^2$.

Proof: Since the explicit method is a pure Newton method, the step will be a descent direction. This may be simply shown by

$$p = -F(X^k)^T J(X^k) J(X^k)^{-1} F(X^k) = -F(X^k)^T F(X^k) < 0$$

Recall from lemma 1 that one step of the uncorrected implicit method is identical to solving the linear system of equations (3.3.5b). From (3.3.5b), the directional derivative of $\|F(X)\|^2$ at X^k in the direction $\Delta X = (\Delta x_1^k, ..., \Delta x_q^k, \Delta x_{q+1}^k)$ is given by

$$p = F(X)^T J(X) J_{imp}^{-1} \hat{F} \tag{4.1.2}$$

where

$$\hat{F} = [\sum_{i=0}^{j-1} f_1(x_1^{k,i}, x_{q+1}^k), ..., \sum_{i=0}^{j-1} f_q(x_q^{k,i}, x_{q+1}^k), f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k)]$$

and J_{imp} is the Jacobi matrix in (3.3.5b). If only one inner iteration is applied ($j = 1$), (4.1.2) becomes

$$p = -F(X^k)^T J(X^k) J_{imp}^{-1} F(X^k) \tag{4.1.3}$$

The multiplication $J(X^k)J_{imp}^{-1}$ yields a full matrix which is given by

$$J(X_k)J_{imp}^{-1} = \begin{bmatrix} I - B_1 C_1 A_1^{-1} & -B_1 C_2 A_2^{-1} & \cdots & -B_1 C_q A_q^{-1} & B_1 \hat{P}^{-1} \\ -B_2 C_1 A_1^{-1} & I - B_2 C_2 A_2^{-1} & \cdots & -B_2 C_q A_q^{-1} & B_2 \hat{P}^{-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -B_q C_1 A_1^{-1} & \cdots & -B_q C_{q-1} A_{q-1}^{-1} & I - B_q C_q A_q^{-1} & B_q \hat{P}^{-1} \\ C_1 A_1^{-1}(I-P) & C_2 A_2^{-1}(I-P) & \cdots & C_q A_q^{-1}(I-P) & I \end{bmatrix} \tag{4.1.4}$$

When the element values of $B_i C_i A_i^{-1}$ $(i = 1, ..., q)$ are large enough, $J(X^k)J_{imp}^{-1}$ may not be positive definite, so it may occur that the step given by the uncorrected implicit method may not be a descent direction.

Now we consider the corrected implicit method with a line search on the inner iteration. If the Newton direction along $\Delta x_i^{k,l} = -A_i^{-1} f_i(x_i^{k,l}, x_{q+1}^k)$, $(l = 0, ..., j-1, i = 1, ..., q)$, in the inner iteration is a descent direction for $\|f_i(x_i^{k,l}, x_{q+1})\|^2$, a line search global strategy can be applied in the end of each inner iteration

$$\Delta x_i^{k,l+1} = x_i^{k,l+1} + \lambda_{l,i} \Delta x_i^{k,l}, \quad l = 0, ..., j-1, i = 1, ..., q.$$

where $\lambda_{l,i}$ is the distributed line search parameter for *ith* block in the $l+1$th iteration so that

$$\|f_i(x_i^{k,0}, x_{q+1}^k)\| \geq \|f_i(x_i^{k,1}, x_{q+1}^k)\| \geq \cdots \geq \|f_i(x_i^{k,j-1}, x_{q+1}^k)\| \quad i = 1, ..., q. \tag{4.1.5}$$

Theorem 9. If f_{q+1} is linear, and an inner line search satisfying (4.1.5) is required, then the necessary and sufficient conditions for the corrected implicit direction $\Delta x_i^{k,j-1}$, Δx_{q+1}^k, $(i = 1, ..., q)$ to be a descent direction from $x_i^{k,0}$, x_{q+1}^k, where $j = 1, 2, \cdots$ is the inner iteration number, are

(1) $j = 1$;

or (2) $j = 2$ and

$$\sum_{i=1}^{q}\lambda_{0,i}\,||f_i(x_i^{k,0}, x_{q+1}^k)||^2 + ||f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k)||^2 > \sum_{i=1}^{q}\lambda_{1,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,1}, x_{q+1}^k)$$

or (3) $j > 2$ and

$$\sum_{i=1}^{q}\lambda_{0,i}\,||f_i(x_i^{k,0}, x_{q+1}^k)||^2 + ||f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1})||^2 > \sum_{i=1}^{q}\sum_{l=1}^{j-1}\lambda_{l,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,l}, x_{q+1}^k)$$

Some sufficient conditions which are valid given (4.1.5) are:

(1) $j = 2$, and $\lambda_{0,i} \geq \lambda_{1,i} \quad i = 1, ..., q$;

or (2) $j > 2$, and

$$\sum_{i=1}^{q}\lambda_{0,i}\,||f_i(x_i^{k,0}, x_{q+1}^k)||^2 + ||f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1})||^2 > \sum_{i=1}^{q}\sum_{l=1}^{j-1}\lambda_{l,i}\,||f_i(x_i^{k,0}, x_{q+1}^k)||\,||f_i(x_i^{k,l}, x_{q+1}^k)||.$$

Proof : Based on (3.3.5a) in lemma 1, one step of the corrected implicit method with inner line search is identical to solving the following linear system of equations similar to (3.3.5a):

$$\begin{bmatrix} A_1 & & & B_1 \\ & A_2 & & B_2 \\ & & \ddots & \vdots \\ C_1 & C_2 & \cdots & C_q \quad P \end{bmatrix} \begin{bmatrix} \Delta x_1^k \\ \Delta x_2^k \\ \vdots \\ \Delta x_q^k \\ \Delta x_{q+1}^k \end{bmatrix} = - \begin{bmatrix} \sum_{l=0}^{j-1}\lambda_{l,1} f_1(x_1^{k,l}, x_{q+1}^k) \\ \vdots \\ \sum_{l=0}^{j-1}\lambda_{l,q} f_q(x_q^{k,l}, x_{q+1}^k) \\ f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k) \end{bmatrix} \tag{4.1.6}$$

The directional derivative of $||F(X)||^2$ at X^k in the direction $\Delta X = (\Delta x_1^k, ..., \Delta x_q^k, \Delta x_{q+1}^k)$ is given by

$$p = -F(X)^T \hat{F} \tag{4.1.7}$$

where

$$\hat{F} = [\sum_{l=0}^{j-1}\lambda_{l,1} f_1(x_1^{k,l}, x_{q+1}^k), ..., \sum_{l=0}^{j-1}\lambda_{l,q} f_q(x_q^{k,l}, x_{q+1}^k), f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k)]$$

Thus

$$p = -(p_0 + \sum_{i=1}^{q}\sum_{l=1}^{j-1}\lambda_{l,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,l}, x_{q+1}^k)) \tag{4.1.8}$$

where $p_0 = \sum_{i=1}^{q}\lambda_{0,i}\,||f_i(x_i^{k,0}, x_{q+1}^k)||^2 + ||f_{q+1}(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1})||^2 > 0$

Therefore, $p < 0$ if and only if

(1) $j = 1$;

or (2) $j = 2$, and

$$p_0 > \sum_{i=1}^{q} \lambda_{1,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,1}, x_{q+1}^k)$$

or (3) $j > 2$ and

$$p_0 > \sum_{i=1}^{q} \sum_{l=1}^{j-1} \lambda_{l,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,l}, x_{q+1}^k)$$

In order to derive some simpler sufficient conditions, the following approximation is substituted into (4.1.8):

$$\sum_{i=1}^{q} \sum_{l=1}^{j-1} \lambda_{l,i} f_i(x_i^{k,0}, x_{q+1}^k) f_i(x_i^{k,l}, x_{q+1}^k) \geq -\sum_{i=1}^{q} \sum_{l=1}^{j-1} \lambda_{l,i} \| f_i(x_i^{k,0}, x_{q+1}^k) \| \| f_i(x_i^{k,1}, x_{q+1}^k) \|$$

Then

$$p \leq -(p_0 - \sum_{i=1}^{q} \sum_{l=1}^{j-1} \lambda_{l,i} \| f_i(x_i^{k,0}, x_{q+1}^k) \| \| f_i(x_i^{k,l}, x_{q+1}^k) \|).$$

Therefore, $p < 0$ if

(1) $j = 2$ and $\lambda_{0,i} \geq \lambda_{1,i}$ $(i = 1, ..., q)$, since $\| f_i(x_i^{k,0}, x_{q+1}^k) \| > \| f_i(x_i^{k,1}, x_{q+1}^k) \|$ after the inner line search,

(2) $j > 2$ and

$$p_0 > \sum_{i=1}^{q} \sum_{l=1}^{j-1} \lambda_{l,i} \| f_i(x_i^{k,0}, x_{q+1}^k) \| \| f_i(x_i^{k,l}, x_{q+1}^k) \|.$$

A special case of theorem 9 is when $\lambda_{l,j} = 1$ $(l = 0, ..., j-1,\ i = 1, ..., q)$, that is, no line search is applied in the inner iteration. The necessary and sufficient conditions are then:

(1) $j = 1$;

or (2) $j = 2$ and

$$\| F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k) \|^2 + F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k)^T F(x_1^{k,1}, ..., x_q^{k,1}, x_{q+1}^k) > 0$$

or (3) $j > 2$ and

$$\| F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k) \|^2 + F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k)^T \sum_{l=1}^{j-1} F(x_1^{k,l}, ..., x_q^{k,l}, x_{q+1}^k) > 0$$

Some sufficient conditions are

(1) $j = 2$, and $\| F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k) \| \geq \| F(x_1^{k,1}, ..., x_q^{k,1}, x_{q+1}^k) \|$.

or (2) $j > 2$, and $\| F(x_1^{k,0}, ..., x_q^{k,0}, x_{q+1}^k) \| \geq \sum_{l=1}^{j-1} \| F(x_1^{k,l}, ..., x_q^{k,l}, x_{q+1}^k) \|$.

4.2 The implementation of the global strategy for the corrected implicit method

The idea of a globally convergent modification of the corrected implicit method is to try the method with step length one first at each iteration. If it seems to be taking a reasonable step -- $||F(x_1^{k+1,0}, ..., x_q^{k+1,0}, x_{q+1}^{k+1})||$ decreases sufficiently, then use it. If not, fall back on a step dictated by the line search method. Such a strategy will always end up using the full corrected implicit method step close to the solution and thus retain its fast local convergence rate. If the global method such as line search is chosen and incorporated properly, the method will also be globally convergent under appropriate conditions.

Our analysis shows that the conditions for descent direction without inner line search are stronger than the conditions with inner line search, since the latter is guaranteed to be satisfied when $j = 2$ and $\lambda_{0,i} \geq \lambda_{1,i}$ $(i = 1, ..., q)$. Our experiments show that the total number of iterations decreases most sharply at $j = 2$. Thus, we would choose to implement the global strategy with inner line search When $j > 2$, $(\Delta x_1^k, ..., \Delta x_q^k, \Delta x_{q+1}^k)$ is a descent direction and the line search can be applied at the end of the iteration if the condition (3) of theorem 9 is satisfied. If the condition is not satisfied, the line search is not applied at current the j point but the $j-1$ point which satisfies the conditions for the descent direction. Then a new iteration is started. The detail corrected implicit method with global line search strategy is given by:

Inner Newton step

1. set $j = 0$ and I_{in} = fixed inner iteration number (≥ 2).

2. Do

   (a) Solve $A_i(x_i^{k,0}, x_{q+1}^k)\Delta x_i^{k,j} = -f_i(x_i^{k,j}, x_{q+1}^k)$ for $\Delta x_i^{k,j}$ $(i = 1, ..., q)$;

   (b) *inner line search*: $x_i^{k,j+1} = x_i^{k,j} + \lambda_{l,j}\Delta x_i^{k,j}$ for some $\lambda_{l,j} > 0$ so that

   $$||f_i(x_i^{k,j+1}, x_{q+1}^k)|| < ||f_i(x_i^{k,j}, x_{q+1}^k)||, (i = 1, ..., q,\ l = 0, ..., j-1).$$

   (c) set $j = j + 1$

   (d) if $(j = 1)$ then

   set $\lambda_{1,i} = \lambda_{0,i}$ $\quad i = 1, ..., q$

   endif

Until $\{j = I_{in}$ or condition (3) is not true$\}$

if (condition (3) is not true)

Set $x_i^{k+1,0} = x_i^{k,j-1}, (i = 1, ..., q)$.

else

Set $x_i^{k+1,0} = x_i^{k,j}, (i = 1, ..., q)$.

endif

Main Newton step

3. Form $\hat{J} = P - \sum_{i=1}^{q} C_i(x_i^{k,0}, x_{q+1}^k) A_i^{-1}(x_i^{k,0}, x_{q+1}^k) B_i(x_i^{k,0}, x_{q+1}^k)$

4. Factorize $\hat{J} = L_{q+1} U_{q+1}$.

5. Solve $\hat{J}\Delta x_{q+1} = -f_{q+1}(x_1^{k+1,0}, ..., x_q^{k+1,0}, x_{q+1}^k)$ for Δx_{q+1}.

6. Do correction: $x_i^{k+1,0} = x_i^{k+1,0} - A_i^{-1} B_i \Delta x_{q+1}^k$, $(i = 1, ..., q)$

7. *outer line search*: $x_i^{k+1,0} = x_i^{k,0} + \lambda_k (x_i^{k+1,0} - x_i^{k,0})$, $(i = 1, ..., q)$,

   and $x_{q+1}^{k+1} = x_{q+1}^k + \lambda_k \Delta x_{q+1}$ for some $\lambda_{.k}$ and $\lambda_k > 0$

   so that $||F(x_1^{k+1,0}, ..., x_q^{k+1,0}, x_{q+1}^{k+1})|| < ||F(x_1^{k,0}, \cdots x_q^{k,0}, x_{q+1}^k)||$.

4.3 Summaries of the two types of the methods

We give the following summary based on our experimental comparisons and analysis of the explicit method and implicit methods.

(1) The implicit (uncorrected and corrected) methods requires more function evaluations per iteration than the explicit method since more than one inner iterations are applied, but possibly fewer total iterations.

(2) The corrected implicit method has an asymptotic convergence rate at least as fast as the explicit method since it retains quadratic convergence rate, and a little bit faster than the uncorrected implicit method. Both uncorrected and corrected implicit methods may speed up the convergence of the interior variables.

(3) A global strategy such as a line search can be applied to the corrected implicit method to ensure global convergence subject to limited restrictions.

(4) The implicit methods will be shown to have additional advantages on parallel computers in the next sections.

5. Parallel solutions to the problem

We will briefly discuss a range of possible strategies for parallelizing the structure of block bordered systems of nonlinear equations. What strategies are best depends on the nature of the nonlinearities and the sparsity structure of the problem, as well as on the characteristics of the parallel machine being used. We intend to implement these strategies and variations of them on both shared memory parallel machines, such as the Encore Multimax, and local memory parallel machines such as the Intel Hypercube.

5.1 Parallel explicit method

The explicit method involves iteratively solving the linear system (3.1). Thus the parallel method focuses on how to solve the block bordered linear system $J(X)\Delta X = -F(X)$ which is of the form:

$$\begin{bmatrix} A_1 & & & B_1 \\ & A_2 & & B_2 \\ & & \ddots & \vdots \\ & & A_q & B_q \\ C_1 & C_2 & \cdots C_q & P \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \vdots \\ \Delta x_q \\ \Delta x_{q+1} \end{bmatrix} = \begin{bmatrix} -f_1 \\ -f_2 \\ \vdots \\ -f_q \\ -f_{q+1} \end{bmatrix} \qquad (5.1.1)$$

in parallel.

Recall that $\Delta x_i \ \ (i = 1, ..., q)$ and Δx_{q+1} can be explicitly solved as follows:

$$\Delta x_i = -A_i^{-1} f_i - A_i^{-1} B_i \Delta x_{q+1}. \tag{5.1.2}$$

and

$$(P - \sum_{i=1}^{q} C_i A_i^{-1} B_i)\Delta x_{q+1} = -f_{q+1} + \sum_{i=1}^{q} C_i A_i^{-1} f_i. \tag{5.1.3}$$

Obviously, the q factorizations of $A_i = L_i U_i$ and the q solutions of $A_i^{-1} f_i$ and $A_i^{-1} B_i \ \ i = 1, ..., q$, may be performed concurrently. But the other operations do not decompose as obviously. Thus, the following basic operations are directly from (5.1.2) and (5.1.3):

1. factorize $A_i = L_i U_i \ \ (i = 1, ..., q)$ in parallel.

2. solve $A_i z_i = f_i \ \ (i = 1, ..., q)$ for $z_i = A_i^{-1} f_i$ in parallel.

3. solve $A_i w_i = B_i \ \ (i = 1, \ ..., q)$ for $w_i = A_i^{-1} B_i$ in parallel.

*4. form $\hat{f} = (P - \sum_{i=1}^{q} C_i w_i)$

*5. factorize $\hat{J} = L_{q+1} U_{q+1}$.

*6. solve $\hat{J} \Delta x_{q+1} = -f_{q+1} + \sum_{i=1}^{q} C_i z_i$ for Δx_{q+1}.

7. $\Delta x_i = -z_i - w_i \Delta x_{q+1} \ \ (i = 1, ..., q)$ in parallel.

8. $x_i^{k+1} = \Delta x_i^k + x_i^k \ \ (i = 1, ..., q)$ in parallel.

*9. $x_{q+1}^{k+1} = \Delta x_{q+1}^k + x_{q+1}^k$

The steps with stars requires some synchronizations on a shared memory multiprocessor, or some message-passing among the nodes on a local memory multiprocessor to parallelize. We will discuss those their implementations next.

On a shared memory multiprocessor data is stored in the shared memory where it can be accessed by all processors through an interconnection network. Step 1, 2, 3, 7 and 8 are independent data operations, and may be fully parallelized. The matrix multiplications and subtractions in step 4 and 9 are independent data operations on a shared memory multiprocessor, which may also be fully parallelized. Steps 5 and 6 are to solve a linear system of equations by first factoring $\hat{J}$ and then back solving for the variables of Δx_{q+1}. These operations involve dependent data operations, and synchronizations are required for the computations on a shared memory multiprocessor. Many parallel algorithms for LU decomposition and back solving on a shared memory multiprocessor have been developed. (see e.g. Jordan [1985]).

Thus, on a shared memory multiprocessor, the operations of the explicit method may fully be parallelized except step 5 and 6 which involve some of the synchronizations. Although the synchronizations

seem minor in comparison with the parallel operations in those two steps, the bottle-neck of explicit method, if any, will come from solving the bottom linear systems of equation at each iteration.

On a local memory multiprocessor, there is only local memory associated with each processor and data is passed among the processors through a connection network. Since data is not shared, a distributed data structure is associated with the parallel algorithm. In our application, processor p_i, $i = 0, ..., p-1$ will store the following data file:

Block A_i or a group of blocks A_i;

Block C_i or a group of blocks C_i;

Block B_i or a group of blocks B_i;

Blocks Δx_i and x_i or groups of blocks Δx_i and x_i;

An efficient LU factorization algorithm needs to minimize the communication costs among the processors, and keep all the processors working in parallel. Current fast parallel LU factorizations on local memory multiprocessors (see e.g. Moler [1986]) require the columns of the coefficient matrix to be evenly distributed among the processors. In order to keep all processors working efficiently, the matrix is distributed in following order: column j is in processor $(j-1) \bmod p$. This kind of storage is called *wrap* mapping. Thus, the columns of P matrix are distributed in wrap mapping among the processors. Δx_{q+1}, x_{q+1} and f_{q+1} are stored in the control processor, say p_0.

Based on this distributed data structure, steps 1, 2, 3, and 8 in the explicit algorithm are independent data operations without any data communications and may fully be parallelized in a local memory multiprocessor. Step 7 is also a independent data operation after Δx_{q+1} is broadcast from p_0. Since Δx_{q+1} and x_{q+1} are stored in p_0, step 9 is a single process in p_0. This sequential operation has minor effect to the parallel performance since the computation is small comparing with other parallel operations.

The columns of $\hat{J}$ are required to be distributed in wrap fashion for efficiently solving the linear system of equations on a local memory multiprocessor, and the columns of P are already distributed in wrap mapping among the processors. Thus, forming $\hat{J}$ in parallel in a local memory multiprocessor requires some message passing among the processors.

Step 6 in the explicit method involves solving a single (lower or upper) triangular linear system of equations in parallel. This would be hard in a local memory multiprocessor, and would be especially hard in the case where the matrix is distributed by columns instead of by rows. There has been some recent progress on this problem (see e.g. Romine & Ortega [1986], Li & Coleman [1986], [1987]). Li-Coleman's methods require the columns of the (upper or lower) triangular matrix be distributed to p processors in a wrap mapping. The computation is not perfectly parallelized, and the speedup increases as $\frac{n}{p}$ increases. We would use a distributed sequential method is applied to solve the triangular system without any extra communications when n is small. When n is large, we would apply Coleman's method to solve the

triangular system since the columns have already been stored in wrap fashion during the factorization.

Steps 4, 5, 6 in the explicit method involve data communications. Thus, the bottle-neck of the parallel explicit method is to form $\hat{J}$ and to solve the bottom system of linear equations.

5.2 Parallel implicit method

One portion of an iteration in the implicit method is to solve each of the q equations

$$f_i(x_i, x_{q+1}) = 0, \qquad i = 1, ..., q \tag{5.2.1}$$

in parallel for a fixed value of x_{q+1}. Newton's method may be used to solve (5.2.1). Then Δx_{q+1} is solved by

$$\hat{J}\Delta x_{q+1}^k + f_{q+1}(x_1^{k+1,0}(x_{q+1}^k), ..., x_q^{k+1,0}, x_{q+1}^k) = 0 \tag{5.2.2}$$

Based on (5.2.1) and (5.2.2), the parallel implicit method is given by:

Inner Newton step

1. $j = 0$.

2. Solve $A_i \Delta x_i^{k,j} = -f_i$ at $x_i^{k,j}$ points for $\Delta x_i^{k,j}$ in parallel.

   (a) $x_i^{k,j+1} = x_i^{k,j} + \Delta x_i^{k,j}$.

   (b) if $x_i^{k+1,j}$ is not "precise" enough, set $j = j + 1$, and goto step 2.a. Else continue.

   (c) Set $x_i^{k+1,0} = x_i^{k,j+1}$ in parallel.

3. solve $A_i w_i = B_i$ $(i = 1, ..., q)$ for $w_i = A_i^{-1}B_i$ in parallel.

Main Newton step

*4. Form $\hat{J} = P - \sum_{i=1}^{q} C_i W_i$

*5. factorize $\hat{J} = L_{q+1}U_{q+1}$.

*6. solve $\hat{J}\Delta x_{q+1} = -f_{q+1}(x_1^{k+1,0}, ..., x_q^{k+1,0}, x_{q+1}^k)$ for Δx_{q+1}.

7. Do correction: $x_i^{k+1,0} = x_i^{k+1,0} - W_i \Delta x_{q+1}^k$

8. $\Delta x_{q+1}^{k+1} = x_{q+1}^k + \Delta x_{q+1}$.

The data structures and operations of the implicit method are almost exactly same as the operations of the explicit method although they are different methods and have different performance. The implicit method is expected to have more inner iterations and less total iterations than the explicit method. However, the implementations of each of these steps on both shared memory and local memory multiprocessors are roughly as same as for the parallel explicit method described in section 5.1.

5.3 What can we gain from the implicit method in parallel

The analysis in section 3 indicates that the corrected implicit method converges at least as fast as the explicit method. If we assume the total computing time for solving a given nonlinear problem by the explicit method and corrected implicit method are identical on a sequential processor, then it is easy to see that the parallel corrected implicit method will be more efficient than the parallel explicit method on a parallel multiprocessor, especially on a local memory multiprocessor. As we know, the bottle-neck of the explicit or the implicit methods implemented on either type of multiprocessor is to (form the $\hat{J}$ and) solve the bottom system of equations which involves synchronizations or data communications. If the implicit method has more inner iterations and fewer total iterations than the explicit method, then the implicit method will form $\hat{J}$ and solve the bottom linear system of equations less times than the explicit method. The effects of reducing this bottle-neck on the parallel performance will be greater on a local memory multiprocessor than a shared memory multiprocessor, since the formation of $\hat{J}$ may fully be parallelized on a shared memory multiprocessor, and since the communications delays on a local memory processor are usually significantly larger than synchronization delays on a shared memory multiprocessor.

6. Conclusions and future work

We have studied three methods for solving block bordered nonlinear system of equations: explicit, implicit and corrected implicit methods. The following conclusions are obtained from our analysis and experiments:

(1) The corrected implicit method retains the quadratic convergence rate of the explicit method, and appears to converge a little faster in practice.

(2) The steps of the corrected implicit method are in descent directions under some limited conditions.

(3) Both the explicit method and the implicit methods should get reasonably good speedup on a shared memory multiprocessor. The implicit methods should gain more if the solution of $\hat{J}$ is expensive.

The next stage of this research will be to complete the implementations of the parallel methods on both a shared memory multiprocessor, the Encore Multimax, and a local memory multiprocessor, the Intel hypercube, and to study the performance of the methods when $\hat{J}$ is full, sparse and very sparse. A load balancing problem in a local memory multiprocessor may occur in the applications when the size of the diagonal blocks A_i are different. We also plan to study this issue.

The methods we have discussed have assumed that the Jacobian matrix of the block bordered nonlinear system (1.2) is available. However, in many practical applications, the Jacobian matrix is not given by a set of formulas, rather it is the output from some computational or experimental procedure. In this case, secant methods (such as Broyden's method) are often used to solve (1.1). (see e.g. Dennis and Schnabel [1983]). We also intend to develop a secant method for solving block bordered systems of nonlinear equations.

References

J. E. Dennis, R. Schnabel [1983], *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice-Hall, Englewood Cliffs, New Jerdey.

C. Farhat, Edward Wilson [1986], "Concurrent iterative solution of large finite element systems", Technical Report, Civil Engineering Department, University of California at Berkeley.

R. Fontecilla [1987], "A parallel nonlinear Jacobi algorithm for solving nonlinear equations", Technical Report, Computer Science Department, University of Maryland.

G. A. Geist, M. T. Heath [1986], "Matrix factorization on a hypercube multiprocessor", *Hypercube Multiprocessors 1986*, M. Heath ed., SIAM Publications, Philadelphia, PA.

H. F. Jordan [1985], "Parallel computation with the FORCE", ICASE Technical Report, No. 85-45, NASA.

G. Li, T. F. Coleman [1986], "A parallel triangular solver for a hypercube multiprocessor", Technical Report (TR86-787), Department of Computer Science, Cornell University.

G. Li, T. F. Coleman [1987], "A new method for solving triangular systems on distributed memory message-passing multiprocessors", Technical Report (TR87-812), Department of Computer Science, Cornell University.

C. Moler [1986], "Matrix computation on distributed memory multiprocessors", Technical Report, Intel Scientific Computers.

B. Nour-Omid, K.C. Park [1986], "Solving structural mechanics problems on a Caltech hypercube machine", Technical Report, Mechanical Engineering Department, University of Colorado at Boulder.

D.P. O'Leary, R.E. White [1985], "Multi-splittings of matrices and parallel solution of linear systems", *SIAM Journal of Alg. & Disc Math.*, No. 4, 1985, pp. 137-149.

J. M Ortega, R.G. Voigt [1985], "Solution of partial differential equations and Gauss-Seidel type iterative methods", *SIAM Review*, No. 2, 1985, pp. 149-240.

N. B. Rabbat, A. L. Sangiovanni-Vincentelli, H. Y. Hsieh [1979], "A multilevel Newton algorithm with macromodeling and latency for the analysis of large-scale nonlinear circuit in the time domain", *IEEE Transactions on Circuits and Systems*, No. 9, 1979, pp. 733-741.

N. B. Rabbat, A. L. Sangiovanni-Vincentelli [1980], "Techniques of time-domain analysis of LSI circuits", Technical Report, IBM T. J. Watson Research Center, RC 8351 (#36320), July, 1980.

C. H. Romine, J. M. Ortega [1986], "Parallel solution of triangular systems of equations", Technical Report, RM -86-05, Department of Applied Math., University of Virginia.

R.E. White [1986], "Parallel algorithms for nonlinear problems", *SIAM Journal of Alg. & Disc Math.*, No. 7, 1986, pp. 137-149.

X. Zhang [1988], "Parallel block SOR methods for solving Poisson equation on shared memory and local memory processors", *Proceedings of Third International Conference on Supercomputing*, Boston, May, 1988.

X. Zhang [1989], "Parallel methods for block bordered nonlinear problems", Ph.d disertations, Department of Computer Science, University of Colorado at Boulder, in preparation.

Fast and Stable Algorithms for Computing the Principal nth Root of a Complex Matrix and Their Applications to Mathematical Science and Control Systems

Leang-San Shieh and Jason Sheng-Horng Tsai

Department of Electrical Engineering
University of Houston
Houston, Texas 77004, U.S.A.

Norman P. Coleman

U.S. Army Armament Center
Dover, New Jersey 07801, U.S.A.

Preface

A complete study of the *principal* nth root of a complex matrix and associated matrix-valued functions is presented in this research monograph. This includes the development of techniques to compute the principal nth root of a matrix, study of associated matrix-valued functions, and their applications to mathematical sciences and control systems. First of all, a computationally *fast* and numerically more *stable* algorithm has been developed to compute the principal nth root of a complex matrix without explicitly utilizing its eigenvalues and/or eigenvectors. The principal nth root of a matrix is shown to be useful for the following: constructing the matrix-sign function and the (generalized) matrix-sector function; solving the matrix Lyapunov and Riccati equations; separating matrix eigenvalues relative to a circle, sector and a sector of a circle in the λ-plane; block-diagonalization (parallel decomposition) and block-triangularization (cascaded decomposition) of a general system matrix; generalizing the block-partial-fraction expansion of a rational matrix; and modelling a continuous-time system from the identified discrete-time model. Also, in this research monograph, new definitions and computational algorithms have been presented to determine the rectangular and polar representations of a complex matrix. Furthermore, their applications to control systems have been discussed. Finally, utilizing the developed algorithms, a multi-stage design procedure has been established to design discrete-time controllers to achieve pole-assignment in a specified region for a large-scale discrete-time multivariable system.

Acknowledgments

The authors wish to express their gratitude for the fruitful discussions with Dr. Jagdish Chandra, Director of Mathematics Science and Dr. Francis Dressel, U.S. Army Research Office, and Dr. Robert E. Yates, Director of the Guidance and Control Directorate, U.S. Army Missile Command.

This research monograph was financially supported in part by the U.S. Army Research Office, under contract DAAL-03-87-K0001, and the U.S. Army Missile Command, under contract DAAH-01-85-C-A111.

Two numbers appear near the bottom of the pages of this article. The top one coinsides with the page number in the auther's Table of Contents. The lower one denotes its position in this journal.

Table of Contents

Chapter 1

Introduction

Computational methods for finding the nth roots of some specific matrices have been proposed in [1,3,4,9] and [17]-[22]. Hoskins and Walton [4], using the Newton-Raphson algorithm, have derived a fast and stable method for computing the nth roots of positive-definite matrices. Based on a spectral decomposition technique obtained from the matrix-sign function [17] together with Hoskins-Walton algorithm [6], Denman *et al.* [18,19] have proposed an algorithm to compute the nth roots of real and complex matrices without prior knowledge of the eigenvalues and eigenvectors of matrices. However, in general, the computed nth root of a general matrix by using the above algorithms is not the principal nth root of the matrix. There are many applications of the principal nth root method to mathematical sciences and control systems such as these listed below:

1) to construct the matrix-sign function [9,17], the matrix-sector function [26,27] and the generalized matrix sector function, to solve the matrix Lyapunov and Riccati equations [1,17,23,24,25],
2) to separate matrix eigenvalues relative to a sector, circle and a sector of a circle in the λ-plane,
3) to achieve A-invariant space, the block-diagonalization (parallel decomposition) and block-triangularization (cascaded decomposition) of the system matrix,
4) to generalize block partial-fraction expansion of a rational matrix [12,13],
5) to model a continuous-time system from the identified discrete-time model,
6) to determine the rectangular and polar representations of a complex matrix, and
7) to develop the multi-stage design procedure for designing discrete-time controllers to achieve pole-assignment in a specific region for a large-scale discrete-time multivariable system.

Shieh *et al.* [20] first proposed an algorithm to compute the principal nth roots of complex matrices. To improve the convergence rate of the computational algorithm in [20], Tsay *et al.* [21] derived a fast algorithm using the matrix continued-fraction method to compute the principal nth roots of complex matrices. However, the above two algorithms [20,21] are not numerically stable. For example, for an ill-conditioned matrix such as a stiff matrix containing both large and small eigenvalues, the algorithms in [20,21] converge in the first few iterations and then diverge very quickly. To overcome this problem of numerical stability, Higham [22] and

Shieh *et al.* [29] have proposed fast and stable algorithms, respectively, for computing the principal square root of a complex matrix. Since the algorithms [22,29] are limited to compute the principal square root of a matrix only, we can not apply the algorithms to compute the principal nth root of a complex matrix when n is not the power of two.

Since there are so many applications of the principal nth root method to mathematical sciences and control systems, a computationally fast and numerically more stable algorithm has been developed to compute the principal nth root of a matrix without explicity utilizing its eigenvalues and/or eigenvectors. Moreover, some applications of the principal nth root method to mathematical sciences and control systems are presented in this research monograph.

The material in this research monograph is organized as follows.

In Chapter 2, based on the generalized continued-fraction method for finding the nth roots of real numbers, a fast computational method for finding the *principal* nth root of a complex matrix has been developed. Computational algorithms with high convergence rates are presented, and their global convergence properties are investigated from the viewpoint of systems theory.

In Chapter 3, rapidly convergent and more stable recursive algorithms for finding the principal nth root of a complex matrix have been developed. The developed algorithms significantly improve the computational aspects of finding the principal nth root of a matrix. Thus, the developed algorithms will enhance the capabilities of the existing computational algorithms such as the principal nth root algorithm, the matrix-sign algorithm and the matrix-sector algorithm for developing applications to control-system problems.

In Chapter 4, the matrix-sector function of A has been generalized to the matrix- sector function of $\mathrm{g}(A)$, where the complex matrix A may have a real or complex characteristic polynomial and $\mathrm{g}(A)$ is a matrix function of a conformal mapping. Based on the computationally fast and numerically more stable algorithms for computing the principal nth root of a complex matrix, rapidly convergent and more stable recursive algorithms for finding the matrix-sector function and the generalized matrix-sector function have been developed. Moreover, the generalized matrix-sector function of A is employed to separate the matrix eigenvalues relative to a sector, a circle, and a sector of a circle in a complex plane without actually seeking the characteristic polynomial and the matrix eigenvalues themselves. Also, the generalized matrix-sector function of A is utilized to carry out the block-diagonalization and block-triangularization of a system matrix, which are useful in developing applications to mathematical science and control-system problems.

In Chapter 5, fast computational methods are developed for finding the equivalent continuous-time state equations from discrete-time state equations. The computational methods utilize the direct-truncation method, the matrix continued-fraction method, and the geometric-series method in conjunction with the principal nth root of the discrete-time system matrix for quick determination of the approximations of a matrix-logarithm function. It is shown that the use of the principal nth root of a matrix enables us to enlarge the convergence region of the expansion of a matrix logarithm function and to improve the accuracy of the approximations of the matrix-logarithm function.

In Chapter 6, some new definitions of the real and imaginary parts and the associated amplitude and phase of a real or complex matrix have been defined. Computational methods, which utilize the properties of the matrix-sign function and the principal nth root of a complex matrix, are given for finding these quantities. A geometric-series method is newly developed for finding the approximation of the matrix-valued function of $\tan^{-1}(X)$, which is the principal branch of the arc tangent of the matrix X.

In Chapter 7, a multi-stage pseudo-continuous-time state-space method is developed for designing a large-scale discrete system, which does not exhibit a two- or multi-time scale structure explicity. The designed pseudo-continuous-time regulator places the eigenvalues of the closed-loop discrete system within the common region of a circle (concentric within the unit circle) and a logarithmic spiral in the complex z-plane, without explicitly utilizing the open-loop eigenvalues of the given system. The proposed method requires the solution of small order Riccati equations only at each stage of the design. The principal nth root method has been employed to obtain a multi-time scale structure for the proposed design method.

Conclusions are summarized in Chapter 8 and numerical examples are given at the end of each chapter to illustrate the concepts of the material presented in that respective chapter.

Chapter 2

A Fast Method for Computing the Principal *n*th Roots of Complex Matrices

Based on the generalized continued-fraction method for finding the nth roots of real numbers, this chapter presents a fast computational method for finding the *principal* nth roots of complex matrices. Computational algorithms with high convergence rates are developed, and their global convergence properties are investigated from the viewpoint of systems theory [21].

2.1 Introduction

Computational methods for finding the nth root of some specific matrices have been proposed in [1–7]. The matrix-sign function method [1,7], the matrix continued-fraction method [2,5,6], and the Newton-Raphson method [3,4] have successfully been used to determine the square roots of real and complex matrices. Applications of above methods have been made to solve systems problem, such as the matrix Lyapunov and Riccati equations, spectral factorization and solvents of matrix polynomials, etc. Recently, Hoskins and Walton [4] have proposed an accelerated, stable Newton-Raphson method for computing the nth root of a positive-definite matrix, whereas Denman and Leyva-Ramos [7] have used the extended matrix-sign function [8], which is a variant of the Newton-Raphson method [9], for finding the nth root of a positive-semidefinite matrix. However, the existing Newton-Raphson methods [4,7], in general, cannot be applied to determine the *principal* nth roots of complex matrices which may be positive or positive-semidefinite.

In this chapter, we shall extend the generalized continued-fraction method [10], which was developed for determining the nth root of a positive real number, to find the principal nth roots of a complex number and a complex matrix. Also, we shall establish a fast computational algorithm for determining the principal nth roots of complex matrices which may not be positive or positive semidefinite. Moreover, we shall investigate the global convergence properties of the proposed algorithm from the viewpoint of systems theory.

2.2 The Principal nth Roots of Complex Numbers

The principal nth root of a complex number is defined as follows.

Definition 2.2.1

Let $a = \rho e^{j\theta} \in C$, where $\rho,\ \theta \in R$ and $\rho \geq 0$, $\theta \in [\pi,\ -\pi)$. The principal nth root of a is defined as

$$\sqrt[n]{a} = \sqrt[n]{\rho} e^{j\theta/n}, \tag{2.1}$$

where the real number $\sqrt[n]{\rho}$ with $\sqrt[n]{\rho} \geq 0$ is the principal nth root of ρ. □

Based on the generalized continued fractions [10], a recursive algorithm with the help of matrix operations has been developed for finding the nth root of a positive real number and associated fractional powers of the positive real number. The algorithm is described below.

Consider a discrete state equation,

$$X(k+1) = HX(k), \quad X(0) = [1,0,0,\ldots,0,0]^T \in C^{n\times 1}, \tag{2.2a}$$

where

$$X(k) = [x_1(k), x_2(k), \ldots, x_n(k)]^T \in C^{n\times 1} \tag{2.2b}$$

and

$$H = \begin{bmatrix} 1 & a & a & \ldots & a & a \\ 1 & 1 & a & \ldots & a & a \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \ldots & 1 & a \\ 1 & 1 & 1 & \ldots & 1 & 1 \end{bmatrix} \in C^{n\times n}. \tag{2.2c}$$

The superscript T in (2.2) denotes the transpose operation on a vector. When a in (2.2c) is a positive real number, its determined fractional powers are

$$\lim_{k\to\infty} \frac{x_i(k)}{x_j(k)} = (\sqrt[n]{a})^{j-i} \quad \text{for} \quad 0 \leq i,j \leq n \text{ and } k \geq 1. \tag{2.3}$$

The correctness of the convergence values for the formulation in (2.3) has been proved in [10] via the continued-fraction approximation theory. In this section, we shall extend the results in (2.2) and (2.3) to include a complex number a with $\arg(a) \neq \pi$ and $a \neq 0$, and we shall investigate the convergence properties from the viewpoint of systems theory.

Consider the matrix H in (2.2c), which is the transpose of a k-circulant with $k = a$ [14, pp. 84–85] and can be expressed by

$$H = (D^{-1}F)\Lambda(D^{-1}F)^{-1}, \tag{2.4a}$$

where the matrix $\Lambda = \mathrm{diag}[p(\alpha), p(\alpha W), \ldots, p(\alpha W^{n-1})]$ with $\alpha = \sqrt[n]{a}$, $W = e^{j2\pi/n}$, and $p(x) = \sum_{i=0}^{n-1} x^i$, the matrix $D = \mathrm{diag}[1, \alpha, \ldots, \alpha^{n-1}]$, and the Fourier matrix F [14, p. 32] is

$$F = \frac{1}{\sqrt[2]{n}} \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & W^{-1} & W^{-2} & \ldots & W^{-n+1} \\ 1 & W^{-2} & W^{-4} & \ldots & W^{-2n+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W^{-n+1} & W^{-2n+2} & \ldots & W^{-(n-1)(n-1)} \end{bmatrix}. \qquad (2.4b)$$

Hence the eigenvalues of H, which are defined as λ_i for $i = 1, 2, \ldots, n$, are $p(\alpha W^{i-1})$ for $i = 1, 2, \cdots, n$, and their associated eigenvectors of H are $(D^{-1}F)e_i$, where e_i is the ith column of I_n. It also follows that the modal matrix of H, denoted by M, is $D^{-1}F$. Employing the similarity transformation,

$$X(k) = D^{-1}FX_d(k) = MX_d(k) \qquad (2.5a)$$

to (2.2a) yields

$$X_d(k+1) = \Lambda X_d(k), \quad X_d(0) = \frac{1}{\sqrt[2]{n}}[1, 1, \ldots, 1]^T. \qquad (2.5b)$$

The solution of (2.5b) is

$$\begin{aligned} X_d(k) &= \frac{1}{\sqrt[2]{n}}\left[p(\alpha)^k, P(\alpha W)^k, \ldots, p(\alpha W^{n-1})^k\right]^T \\ &= \frac{1}{\sqrt[2]{n}}\left[\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k\right]^T, \end{aligned} \qquad (2.5c)$$

and the solution of (2.2a) becomes

$$\begin{aligned} X(k) &= MX_d(k) \\ &= \frac{1}{n}\Bigg[\left(\sum_{l=1}^{n}\lambda_l^k\right), \left(\sum_{l=1}^{n}\lambda_l^k W^{-(l-1)}\right)\left(\sqrt[n]{a}\right)^{-1}, \\ &\quad \left(\sum_{l=1}^{n}\lambda_l^k W^{-2(l-1)}\right)\left(\sqrt[n]{a}\right)^{-2}, \ldots, \left(\sum_{l=1}^{n}\lambda_i^k W^{-(n-1)(l-1)}\right)\left(\sqrt[n]{a}\right)^{-n+1}\Bigg]^T \end{aligned}$$

$$\text{for} \quad k \geq 0. \qquad (2.5d)$$

Lemma 2.2.1

If $\arg(a) \neq \pi$ and $a \neq 0$, then $\lambda_1 \neq 0$ and $0 \leq |\lambda_l/\lambda_1| < 1$ for $n \geq l > 1$.

Proof

When $i = 1$, λ_1 $[= p(\alpha)]$ becomes

$$\lambda_1 = 1 + \sqrt[n]{a} + \left(\sqrt[n]{a}\right)^2 + \cdots + \left(\sqrt[n]{a}\right)^{n-1}.$$

If $a = 1$, then $\lambda_1 = n \neq 0$ and $\lambda_l = 0$ for $l > 1$. Thus, $|\lambda_l/\lambda_1| = 0$ for $n \geq l > 1$. On the other hand, if $a \neq 1$, then $\lambda_1 = (1 - a)/(1 - \sqrt[n]{a}) \neq 0$ and

$$\frac{\lambda_l}{\lambda_1} = \frac{1 - \sqrt[n]{a}}{1 - \sqrt[n]{a}W^{l-1}}.$$

Let $a = \rho e^{j\theta}$ and $\sqrt[n]{a} = \sqrt[n]{\rho}e^{j\theta/n}$. Thus, we have

$$\left|\frac{\lambda_l}{\lambda_1}\right|^2 = \frac{1 + \left(\sqrt[n]{\rho}\right)^2 - 2\cos\left(\frac{\theta}{n}\right)\sqrt[n]{\rho}}{1 + \left(\sqrt[n]{\rho}\right)^2 - 2\cos\left(\frac{\theta + 2\pi(l-1)}{n}\right)\sqrt[n]{\rho}}$$

$$= 1 - \frac{2\sqrt[n]{\rho}\left[\cos\left(\frac{\theta}{n}\right) - \cos\left(\frac{\theta + 2\pi(l-1)}{n}\right)\right]}{1 + \left(\sqrt[n]{\rho}\right)^2 - 2\cos\left(\frac{\theta + 2\pi(l-1)}{n}\right)\sqrt[n]{\rho}}.$$

Since $a \neq 0$, we have $\sqrt[n]{\rho} > 0$ and so the lemma is proved if

$$\cos\left(\frac{\theta}{n}\right) - \cos\left(\frac{\theta + 2\pi(l-1)}{n}\right) > 0.$$

It follows that $0 \leq |\lambda_l/\lambda_1|^2 < 1$ and $0 \leq |\lambda_l/\lambda_1| < 1$ provided that $\theta \neq \pi$. ■

Theorem 2.2.1

$\lim_{k\to\infty}[x_i(k)/x_j(k)] = (\sqrt[n]{a})^{j-i}$ for $i \neq j$, $1 \leq i, j \leq n$, $a \neq 0$, and $\arg(a) \neq \pi$.

Proof

From (2.5d), we obtain

$$\frac{x_i(k)}{x_j(k)} = \frac{\sum_{l=1}^{n} \lambda_l^k W^{-(l-1)(i-1)} \left(\sqrt[n]{a}\right)^{-(i-1)}}{\sum_{l=1}^{n} \lambda_l^k W^{-(l-1)(j-1)} \left(\sqrt[n]{a}\right)^{-(j-1)}}.$$

Since $\lambda_1 \neq 0$, we have

$$\frac{x_i(k)}{x_j(k)} = \left(\sqrt[n]{a}\right)^{j-i} \frac{\sum_{l=1}^{n} \left(\frac{\lambda_l}{\lambda_1}\right)^k W^{-(l-1)(i-1)}}{\sum_{l=1}^{n} \left(\frac{\lambda_l}{\lambda_1}\right)^k W^{-(l-1)(j-1)}}. \tag{2.6}$$

Using the result in Lemma 2.2.1 or $0 < |\lambda_l/\lambda_1| < 1$, we obtain

$$\lim_{k\to\infty} \frac{x_i(k)}{x_j(k)} = \left(\sqrt[n]{a}\right)^{j-i}.$$

■

Corollary 2.2.1

Having the state equation defined in (2.2), the principal nth root of a can be found as

$$\lim_{k\to\infty} \frac{x_i(k)}{x_{i+1}(k)} = \sqrt[n]{a} \qquad \text{for} \quad 1 \leq i < n \tag{2.7}$$

if $a \neq 0$ and $\arg(a) \neq \pi$. ■

Corollary 2.2.2

The pth power of the principal nth root of a can be found as

$$\lim_{k\to\infty} \frac{x_1(k)}{x_{p+1}(k)} = \left(\sqrt[n]{a}\right) \qquad \text{for} \quad n-1 \geq p \geq 1. \tag{2.8}$$

■

2.3 Recursive Algorithms and Their Global Convergence Properties

From (2.2), we can compute each state $x_i(k)$ as follows,

$$x_n(k) = \sum_{l=1}^{n} x_l(k-1), \tag{2.9a}$$

$$x_i(k) = x_{i+1}(k) + (a-1)x_i(k-1) \qquad \text{for} \quad i = n-1, n-2, \ldots, 1. \tag{2.9b}$$

The algorithm to compute the pth power of the principal nth root of a complex number a becomes

$$r_n^p \triangleq \left(\sqrt[n]{a}\right)^p = \lim_{k\to\infty} \frac{x_1(k)}{x_{p+1}(k)}. \tag{2.9c}$$

The direct use of (2.9) to compute $\left(\sqrt[n]{a}\right)^p$ may result in numerical overflow if the magnitude of any eigenvalue of H in (2.2) is larger than unity. However, the numerical difficulty may be overcome by normalizing $x_1(k)$ in (2.9) to be unity for all k.

To analyze the convergence rate of the algorithm in (2.9), we assume

$$\epsilon = \max\left\{\left|\frac{\lambda_l}{\lambda_1}\right|,\ l = 2, 3, \ldots, n\right\} \tag{2.10a}$$

and rewrite (2.6), with $i = 1$ and $j = 2$, as

$$\frac{x_1(k)}{x_2(k)} = \sqrt[n]{a}\,\Delta(k), \tag{2.10b}$$

where

$$\Delta(k) = \frac{\displaystyle\sum_{l=1}^{n}\left(\frac{\lambda_l}{\lambda_1}\right)^k}{\displaystyle\sum_{l=1}^{n}\left(\frac{\lambda_l}{\lambda_1}\right)^k W^{-(l-1)}}. \tag{2.10c}$$

Then, by using Lemma 2.2.1 and assuming k is sufficintly large, the error ratio $\left|\left(\sqrt[n]{a}\,\Delta(k) - \sqrt[n]{a}\right)/\sqrt[n]{a}\right|$ becomes

$$|\Delta(k) - 1| \le \frac{2(n-1)\epsilon^k}{1-(n-1)\epsilon^k} \tag{2.10d}$$

or

$$|\Delta(k) - 1| = O(\epsilon^k). \tag{2.10e}$$

Therefore, the algorithm in (2.9) has a linear convergence rate. The derivation of the convergence in (2.10) for the algorithm in (2.9) is similar to that of the Bernoulli-Aitken method [15,16], which is the well-known power method for finding the largest real or complex root of an algebraic equation.

The linear convergence of the algorithm in (2.9) is not realistic for practical computations. We shall now develop alternative algorithms with higher order convergence rates.

Lemma 2.3.1

From (2.2), if the first column of H^k is defined as

$$h_1 \triangleq [h_{1k}, h_{2k}, \ldots, h_{nk}]^T,$$

then

$$H^k = \begin{bmatrix} h_{1k} & ah_{nk} & ah_{(n-1)k} & \cdots & ah_{3k} & ah_{2k} \\ h_{2k} & h_{1k} & ah_{nk} & \cdots & ah_{4k} & ah_{3k} \\ h_{3k} & h_{2k} & h_{1k} & \cdots & ah_{5k} & ah_{4k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{(n-1)k} & h_{(n-2)k} & h_{(n-3)k} & \cdots & h_{1k} & ah_{nk} \\ h_{nk} & h_{(n-1)k} & h_{(n-2)k} & \cdots & h_{2k} & h_{1k} \end{bmatrix}. \tag{2.11}$$

Proof

Since H satisfies $H\eta_k^T = \eta_k^T H$ [14, p. 84], where

$$\eta_k = \begin{bmatrix} 0 & I_{n-1} \\ a & 0 \end{bmatrix},$$

Lemma 2.3.1 follows immediately, since $H^k \eta_k^T = \eta_k^T H^k$. ■

Lemma 2.3.2

The solution $X(k)$ of the state equation in (2.2) is the first column of H^k in (2.11).

Proof

Since $X(k) = HX(k-1) = H^k X(0)$ and $X(0) = [1, 0, \ldots, 0]^T$, the first column of H^k is $X(k)$. ■

Theorem 2.3.1

If the solution of the state equation in (2.2) at the kth step is $X(k)$, then

$$X(2k) = [x_1(2k), x_2(2k), \ldots, x_n(2k)]^T, \tag{2.12}$$

where

$$x_l(2k) = \sum_{i=1}^{l} x_i(k)x_{l+1-i}(k) + a\sum_{i=l+1}^{n} x_i(k)x_{n-i+l-1}(k) \qquad \text{for} \quad 1 \le l < n,$$

$$x_n(2k) = \sum_{i=1}^{n} x_i(k)x_{n+1-i}(k) \qquad \text{for} \quad k \ge 1.$$

Proof

From (2.2), we have $X(2k) = H^k X(k)$. Using the results in Lemmas 2.3.1 and 2.3.2 yields the result in Theorem 2.3.1. ■

From Theorem 2.3.1, we can establish a quadratic convergence algorithm for computing the principal nth roots of complex numbers.

Corollary 2.3.1

The convergence rate of the algorithm in (2.12) is quadratic.

Proof

Define $Z(k) \triangleq [z_1(k), z_2(k), \ldots, z_n(k)]^T = X(2^{k-1})$ for $k \ge 1$ and $Z(0) = X(0)$. From the algorithm in (2.12), we obtain

$$Z(k+1) = X(2^k) = H^{2^k}X(0) = H^{2^k}Z(0) = H_z(k)Z(k),$$

where $H_z(k) = H^{2^{k-1}}$.

Define $r_n^1(k) \triangleq z_1(k)/z_2(k)$ and $\epsilon \triangleq \max\{|\lambda_l/\lambda_1|,\ l = 2, 3, \ldots, n\}$. From (2.6), we have

$$r_n^1(k) = \sqrt[n]{a}\Delta(k),$$

where

$$\Delta(k) = \frac{\displaystyle\sum_{l=1}^{n}\left(\frac{\lambda_l}{\lambda_1}\right)^{2^{k-1}}}{\displaystyle\sum_{l=1}^{n}\left(\frac{\lambda_l}{\lambda_1}\right)^{2^{k-1}} W^{-(l-1)}}.$$

Similar to the derivation in (2.10), the error ratio $\left|\left(\sqrt[n]{a}\Delta(k) - \sqrt[n]{a}\right)/\sqrt[n]{a}\right|$ becomes

$$|\Delta(k) - 1| \leq \frac{2(n-1)\epsilon^{2^{k-1}}}{1-(n-1)\epsilon^{2^{k-1}}}$$

and

$$|\Delta(k+1) - 1| = O\left(|\Delta(k) - 1|^2\right) = O\left(\epsilon^{2^{k-1}}\right)$$

for large k.

Therefore, the algorithm in (2.12) converges quadratically. ■

Algorithms with higher order convergence rates are established below.

Theorem 2.3.2

Define $Z(k) \triangleq X(q^{k-1})$, where q is a positive integer with $q \geq 2$ and $k \geq 1$. Also, define a state equation,

$$Z(k+1) = H_z^{q-1}(k)Z(k), \qquad Z(0) = X(0), \tag{2.13a}$$

where $H_z(k) = H^{q^{k-1}}$ for $k \geq 1$ and $H_z(k)$ has the same structure as H^k in (2.11), having the first column $[z_1(k), z_2(k), \ldots, z_n(k)]^T$. Then, the algorithm for finding the pth power of the principal nth root of a,

$$\left(\sqrt[n]{a}\right)^p = \lim_{k\to\infty} \frac{z_i(k)}{z_j(k)} \tag{2.13b}$$

where $p = j - i$, has qth-order convergence rate.

Proof

Theorem 2.3.2 can be proven in a manner similar to Theorem 2.3.1 and Corollary 2.3.1. ■

2.4 The Principal nth Roots of Complex Matrices

The methods described in Sections 2.2 and 2.3 for computing the principal nth roots of complex numbers can be extended to compute the principal nth roots of complex matrices. The principal nth root of a complex matrix is defined below.

Definition 2.4.1

Let $A \in C^{m\times m}$, $\sigma(A) = \{\lambda_i,\ i = 1,2,\ldots,m\}$, $\lambda_i \neq 0$ and $\arg(\lambda_i) \neq \pi$. The principal nth root of A is defined as $\sqrt[n]{A} \in C^{m\times m}$, where n is a positive integer and

(a) $\left(\sqrt[n]{A}\right)^n = A$,

(b) each eigenvalue of $\sqrt[n]{A}$ is the principal nth root of each λ_i. □

To derive a fast algorithm for computing the principal nth roots of complex matrices, we extend the discrete-state equation in (2.2) to the *bloc*-discrete-state equation as follows,

$$\bar{X}(k+1) = G\bar{X}(k), \quad \bar{X}(0) = [I_m, 0_m, \ldots, 0_m]^T, \tag{2.14a}$$

where the matrix G is the transpose of a block-k-circulant [14], viz.,

$$G = \begin{bmatrix} I_m & A & A & \ldots & A & A \\ I_m & I_m & A & \ldots & A & A \\ I_m & I_m & I_m & \ldots & A & A \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ I_m & I_m & I_m & \ldots & I_m & A \\ I_m & I_m & I_m & \ldots & I_m & I_m \end{bmatrix} \in C^{nm\times nm}, \tag{2.14b}$$

and

$$\bar{X}(k) = \left[\bar{x}_1^T(k), \bar{x}_2^T(k), \ldots, \bar{x}_n^T(k)\right]^T \in C^{nm\times m}. \tag{2.14c}$$

Note that the state variables $x_i(k)$ in (2.2*b*) are of dimension 1×1, whereas the *block*-state variables $\bar{x}_i(k)$ in (2.14*c*) are of dimension $m \times m$. The characteristic polynomial matrix [14] of G can be determined as

$$D(\lambda) = \lambda^n I_m - {}_nC_1\lambda^{n-1}I_m + {}_nC_2\lambda^{n-2}(I_m - A) - {}_nC_3\lambda^{n-3}(I_m - A)^2 + \cdots$$

$$+ (-1)^{n-1}{}_nC_{n-1}\Lambda(I_m - A)^{n-2} + (-1)^n(I_m - A)^{n-1}, \tag{2.15a}$$

where $D(\lambda) \in C^{m\times m}[\lambda]$, $\lambda \in C$, and ${}_nC_i$ are the coefficients of a binomial expansion.

The block eigenvalues of G, which are also known as the solvents [11,12] of $D(\lambda)$, can be obtained from $D(\lambda)$ in (2.15*a*) as

$$D(\Lambda_i) = 0_m \qquad \text{for} \quad i = 1,2,\ldots,n, \tag{2.15b}$$

where $\Lambda_i \in C^{m\times m}$ are the block eigenvalues of G.

Let a set of complete solvents [11–13] of $D(\lambda)$ be

$$\Lambda_l = \sum_{i=1}^{n} [\sqrt[n]{A} W^{l-1}]^{i-1} \qquad \text{for} \quad 1 \le l \le n, \tag{2.16a}$$

where $W \triangleq e^{j2\pi/n}$. Then the block eigenvector [11] associated with Λ_l becomes

$$\hat{V}_l = \frac{1}{\sqrt[n]{n}} \left[I_m, \left((\sqrt[n]{A} W^{l-1})^{-1} \right)^T, \ldots, \left((\sqrt[n]{A} W^{l-1})^{-n+1} \right)^T \right]^T \in C^{nm \times m}, \tag{2.16b}$$

and the corresponding block-modal matrix [11] is

$$\bar{M} = [\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_n] \in C^{nm \times nm}. \tag{2.16c}$$

Thus, following the derivations in (2.4) and (2.5) and employing the properties of the block-k-circulant [14] in (2.14b), the block-state equation in (2.14a) can be transformed into a block-diagonalized state equation by using the following transformation [11],

$$\bar{X}(k) = \bar{M} \bar{X}_d(k). \tag{2.17}$$

The transformed block-state equation becomes

$$\bar{X}_d(k+1) = G_d \bar{X}_d(k), \tag{2.18a}$$

$$\bar{X}_d(0) = \frac{1}{\sqrt[n]{n}} [I_m, I_m, \ldots, I_m]^T \tag{2.18b},$$

where

$$G_d = \bar{M}^{-1} G \bar{M} = \text{block diag}[\Lambda_1, \Lambda_2, \ldots, \Lambda_n]. \tag{2.18c}$$

The solution of the block-state equation in (2.14a) is

$$\bar{x}_i(k) = \frac{1}{n} (\sqrt[n]{A})^{-i+1} \sum_{l=1}^{n} \Lambda_l^k W^{-(i-1)(l-1)} \qquad \text{for} \quad i = 1, 2, \ldots, n. \tag{2.18d}$$

Thus, we have the following result.

Theorem 2.4.1

Let $A \in C^{m\times m}$, $\sigma(A) = \{\lambda_i,\ i = 1,2,\ldots,m\}$, $\lambda_i \neq 0$, and $\arg(\lambda_i) \neq \pi$. Then,

$$\lim_{k\to\infty} \bar{x}_i(k)\bar{x}_j^{-1}(k) = \left(\sqrt[n]{A}\right)^{j-i} \qquad \text{for} \quad i \geq 1 \text{ and } j \leq n. \tag{2.19}$$

Proof

Since Λ_l and $\sqrt[n]{A}$ or $\left(\sqrt[n]{A}\right)^{-1}$ commute, (2.18) becomes

$$\bar{x}_i(k)\bar{x}_j^{-1}(k) = \left(\sqrt[n]{A}\right)^{j-i}\left[\sum_{l=1}^{n}\Lambda_l^k W^{-(i-1)(l-1)}\left(\sum_{l=1}^{n}\Lambda_l^k W^{-(j-1)(l-1)}\right)^{-1}\right].$$

Let γ_i for $1 \leq i \leq m$ be the eigenvalues of $\Lambda_l\Lambda_1^{-1}$, $l = 2,3,\ldots,n$. Then, from (2.16a), we obtain

$$\gamma_i = \frac{\sum_{j=1}^{n}\left(\sqrt[n]{\lambda_i}W^{(l-1)}\right)^{j-1}}{\sum_{j=1}^{n}\left(\sqrt[n]{\lambda_i}\right)^{j-1}} \qquad \text{for} \quad i = 1,2,\ldots,m.$$

From Lemma 2.2.1, we have $0 \leq |\gamma_i| < 1$ if $\lambda_i \neq 0$ and $\arg(\lambda_i) \neq \pi$, $1 \leq i \leq m$, and so

$$\lim_{k\to\infty}\Lambda_l^k\left(\Lambda_l^{-1}\right)^k = \lim_{k\to\infty}\left(\Lambda_l\Lambda_l^{-1}\right)^k = 0 \qquad \text{if} \quad l \neq 1.$$

Thus, we have

$$\lim_{k\to\infty}\bar{x}_i(k)\bar{x}_j^{-1}(k) = \left(\sqrt[n]{A}\right)^{j-i}.$$

■

Corollary 2.4.1

The principal nth root of a complex matrix $A \in C^{m\times m}$ with $\sigma(A) = \{\lambda_i,\ i = 1,2,\ldots,n\}$, $\lambda_i \neq 0$ and $\arg(\lambda_i) \neq \pi$, can be found as

$$\lim_{k\to\infty}\bar{x}_i(k)\bar{x}_{i+1}^{-1}(k) = \sqrt[n]{A} \qquad \text{for} \quad 1 \leq i \leq n.$$

■

Corollary 2.4.2

Given a complex matrix $A \in C^{m\times m}$ as defined in Corollary 2.4.1, the principal nth root of A is unique. ■

Corollary 2.4.3

The pth power of the principal nth root of a complex matrix A as defined in Corollary 2.4.1 can be obtained as

$$\lim_{k\to\infty} \bar{x}_1(k)\bar{x}_{p+1}^{-1}(k) = \left(\sqrt[n]{A}\right)^p \qquad \text{for} \quad 0 < p < n.$$

■

Following Theorems 2.4.1, 2.3.1 and Corollary 2.4.3, we can construct a quadratic convergence algorithm for computing the pth power of the principal nth root of a complex matrix A as follows.

Algorithm 2.4.1

Given:

A = an $m \times m$ complex matrix with eigenvalues $\lambda_i = \rho_i e^{i\theta_i}$, where $\rho_i \neq 0$ and $\theta_i \neq \pi$ for $i = 1, 3, \ldots, m$,
n = root index,
δ = error tolerance.

Find:

R_n^p = pth power of the principal nth root of a for $1 \leq p \leq n-1$.

Algorithm:

```
{Initialization}
for i := 1 to n do {Initialize the states X_i, i = 1,2,...,n}
  X_i := I_m;
R := 0_m; {Initialize the principal nth roots}
{Computation of the principal nth roots}
repeat
  for i := 1 to n do {Copy X_i to Y_i for i = 1,2,...,n}
    Y_i := X_i;
  for i := 1 to n do {Compute X_i, i = 1,2,...,n}
    begin
      X_i := X_i Y_1;
      for j := 2 to n do
        if j ≤ i then
          X_i := Y_j Y_{i-j+1} + X_i
        else
          X_i := A Y_j Y_{n-j+i+1} + X_i
    end;
  R_n^1 := X_1 X_2^{-1}; {Find the principal nth root}
  Δ := ||R - R_n^1||; {The norm of difference between the last and current
                       iteration of the principal nth root}
  if Δ > δ then {Error is not within the specified
tolerance}
    begin
      R := R_n^1; {Copy R_n^1 to R}
      for i := 2 to n do {Normalization}
        X_i := X_i X_1^{-1};
      X_1 := I_m
    end
until Δ ≤ δ;
{Compute the pth power of principal nth root, for 2 ≤ p ≤ n - 1}
for i := 2 to n - 1 do
  R_n^i := X_1 X_{i+1}^{-1};
```

■

When $\arg(\lambda_i) = \pi$, Algorithm 2.4.1 cannot directly be applied to compute $\left(\sqrt[n]{A}\right)^p$. The matrix A can be rotated by a small angle to give $\bar{A} = Ae^{-j\Delta\beta}$ (where $\Delta\beta$ is a small positive real angle) so that $\left(\sqrt[n]{A}\right)^p = \left(\sqrt[n]{\bar{A}}\right)^p e^{jp\Delta\beta/n}$.

Example 2.4.1

Given a complex matrix,

$$A = \begin{bmatrix} -1.25 + j3.25 & 2.50 + j3.50 & -2.75 - j3.25 & 6.25 + j0.75 \\ 4.00 + j1.75 & 6.00 - j1.50 & -6.50 + j2.75 & 6.00 - j7.75 \\ -2.25 + j1.75 & -0.50 + j2.50 & 0.25 - j2.75 & 3.25 + j2.25 \\ -2.00 - j1.50 & -3.00 - j1.00 & 4.00 + j0.50 & -4.00 + j1.50 \end{bmatrix},$$

it is desired to find $\sqrt[5]{A}$ with $n = 5$.

The matrix A has an eigenvalue -1 with $\arg(-1) = \pi$. Thus, Algorithm 2.4.1 can not be directly used for finding $\sqrt[5]{A}$. The matrix A is modified with the rotation angle $\Delta\beta = 5°$ (or $\pi/36$). The modified matrix $\bar{A}$ becomes

$$\bar{A} = \begin{bmatrix} -0.96199 + j3.34658 & 2.79553 + j3.26879 & -3.02279 - j2.99795 & 6.29158 + j0.20242 \\ 4.13730 + j1.39472 & 5.84643 - j2.01723 & -6.23559 + j3.30605 & 5.30171 - j8.24344 \\ -2.08892 + j1.93944 & -0.28021 + j2.53406 & 0.00937 - j2.76132 & 3.43373 + j1.95818 \\ -2.12312 - j1.31998 & -3.07574 - j0.73473 & 4.02836 + j0.14947 & -3.85408 + j1.84292 \end{bmatrix}.$$

Using Algorithm 2.4.1 with error tolerance 10^{-10}, we have the principal 5th root $\bar{A}$ with 10 iterations as

$$\sqrt[5]{\bar{A}} = \begin{bmatrix} 1.23828 + j0.78835 & 0.70104 - j0.13515 & -0.63288 + j0.13618 & 0.49772 - j0.83723 \\ 0.63025 - j0.40270 & 1.61288 - j0.72001 & -0.37929 + j1.37819 & -0.15461 - j1.93558 \\ 0.18516 + j0.32017 & 0.31079 - j0.29380 & 0.69469 + j0.48856 & 0.06699 - j0.98547 \\ -0.23397 + j0.10540 & -0.39025 - j0.15865 & 0.50842 - j0.22120 & 0.38842 + j0.42534 \end{bmatrix}.$$

Thus, the desired principal 5th root of A is given by

$$\sqrt[5]{A} = \begin{bmatrix} 1.22433 + j0.80985 & 0.70330 - j0.12290 & -0.63516 + j0.12512 & 0.51226 - j0.82841 \\ 0.63718 - j0.39163 & 1.62521 - j0.69175 & -0.40329 + j1.37136 & -0.12081 - j1.93799 \\ 0.17954 + j0.32336 & 0.31587 - j0.28833 & 0.68606 + j0.50061 & 0.08418 - j0.98415 \\ -0.23577 + j0.10130 & -0.38742 - j0.16543 & 0.51220 - j0.21229 & 0.38094 + j0.43205 \end{bmatrix}.$$

It is interesting to note that the eigenvalues of A are $-1, 2 + j1, 1 - j1, -1 + j0.5$, whereas the eigenvalues of $\sqrt[5]{A}$ are $0.80902 + j0.58779$, $1.0586 - j0.16766$, $1.1696 + j0.10877$, $0.87937 + j0.52186$. All eigenvalues of $\sqrt[5]{A}$ lie in $(-\pi/5, \pi/5]$.

Example 2.4.2

Given a complex matrix,

$$A = \begin{bmatrix} 1.20 - j0.10 & 0.60 - j1.30 & -1.35 + j0.05 & -0.95 - j0.65 \\ -0.70 + j0.10 & 0.40 - j1.20 & 0.35 + j0.45 & -0.55 - j0.35 \\ 1.10 + j1.20 & 1.05 - j1.15 & -1.55 - j1.10 & -1.10 - j1.95 \\ 0.80 + j0.10 & -0.35 - j0.45 & -0.40 - j0.05 & -0.55 - j0.60 \end{bmatrix},$$

it is desired to find $\sqrt[5]{A}$.

The matrix A has an eigenvalues 1.0 and a Jordan chain of length 3 with an eigenvalue -0.5-j1.0. The Jordan form of A can be found as

$$J_A = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & -0.5 - j1.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & -0.5 - j1.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & -05 - j1.0 \end{bmatrix}.$$

Using Algorithm 2.2.1 with error tolerance 10^{-10}, we have the principal 5th root of A with 7 iterations as follows,

$$\sqrt[5]{A} = \begin{bmatrix} 0.85676 + j0.00206 & 0.23415 + j0.02042 & -0.04496 - j0.29987 & 0.10000 - j0.02759 \\ -0.17671 - j0.00329 & 1.11192 - j0.21818 & -0.00296 - j0.00354 & 0.06362 - j0.14713 \\ -0.26275 + j0.28943 & 0.26571 + j0.16929 & 0.98433 - j0.64590 & 0.34128 - j0.10743 \\ 0.10531 + j0.14969 & 0.2219 - j0.12750 & -0.05265 - j0.07484 & 0.86423 - j0.35203 \end{bmatrix}.$$

It is interesting to note that the eigenvalues of $\sqrt[5]{A}$ are 1, 0.93908-j0.40468, 0.93908-j0.40468 and 0.93908-j0.40468. All eigenvalues of $\sqrt[5]{A}$ are lying within $-\pi/5$ and $+\pi/5$.

This example demonstrates that Algorithm 2.4.1 can be equally be used to find the principal nth roots of complex matrices having eigenvalues unity and/or Jordan chains with length greater than unity.

2.5 Conclusion

The generalized continued-fraction method developed for finding the nth roots of real numbers has been extended to determine the principal nth roots of complex matrices. Computational algorithms with high order convergence rates have been established for determination of the principal nth root and associated pth power of the principal nth root of a complex matrix. The global convergence properties of the proposed algorithms have been investigated from the viewpoint of systems theory.

Fast and Stable Algorithms for Computing the Principal *n*th Root of a Complex Matrix

This chapter presents rapidly convergent and more stable recursive algorithms for finding the principal *n*th root of a complex matrix. The developed algorithms significantly improve the computational aspects of finding the principal *n*th root of a matrix. Thus, the developed algorithms will enhance the capabilities of the existing computational algorithms such as the principal *n*th root algorithm, the matrix-sign algorithm and the matrix-sector algorithm for developing applications to control-system problems [61].

3.1 Introduction

Computational methods for finding the *n*th roots of some specific matrices have been proposed in [1,3,4,9] and [17]-[22]. Hoskins and Walton [4], using the Newton-Raphson algorithm, have derived a fast and stable method for computing the *n*th roots of positive-definite matrices. Based on a spectral-decomposition technique obtained from the matrix-sign function [17] together with Hoskins-Walton algorithm [4], Denman *et al.* [18,19] have proposed an algorithm to compute the *n*th roots of real and complex matrices without prior knowledge of the eigenvalues and eigenvectors of matrices. However, in general, the computed *n*th root of a general matrix by using the above algorithms is not the principal *n*th root of the matrix. The principal *n*th root of a matrix can be utilized to construct the matrix-sign function [9,17] and the matrix-sector function [26,27], to solve the matrix Lyapunov and Riccati equations [1,17,23,24,25] and to approximate some matrix-valued functions [28] etc. Shieh *et al.* [20] first proposed an algorithm to compute the principal *n*th roots of complex matrices. To improve the convergence rate of the computational algorithm in [20], Tsay *et al.* [21] derived a fast algorithm using the matrix continued-fraction method to compute the principal *n*th roots of complex matrices. However, the above two algorithms [20,21] are not numerically stable. For example, for an ill-conditioned matrix such as a stiff matrix containing both large and small eigenvalues, the algorithms in [20,21] converge in the first few iterations and then diverge very quickly. To overcome this problem of numerical stability, Higham

[22] and Shieh *et al.* [29] have proposed fast and stable algorithms, respectively, for computing the principal square root of a complex matrix. Since the algorithms [22,29] are limited to compute the principal square root of a matrix only, we can not apply the algorithms to compute the principal nth root of a complex matrix when n is not the power of two. In this chapter, we generalize the fast and stable algorithm in [29] to compute the principal nth root of a complex matrix and then extend the algorithm to compute the matrix-sector function.

This chapter is organized as follows: In Section 3.2, we summarize the fast algorithm for finding the principal nth root of a matrix. Next, fast and stable recursive algorithms for finding the principal nth root of a matrix are developed in Section 3.3. An illustrative example is given in Section 3.4, and the results are summarized in Section 3.5.

3.2 Summary of the Fast Algorithm for Finding the Principal nth Root of a Matrix

The fast algorithm [21] which was derived via the matrix continued-fraction method for finding the principal nth root of a complex matrix is summarized below.

Consider a block-discrete-state equation as

$$X(k+1) = H^{r-1}(k)X(k), \quad X(0) = [I_m, I_m, \cdots I_m]^T,$$

$$\text{for } k = 0, 1, 2, \cdots. \tag{3.1a}$$

Then, we have

$$\lim_{k\to\infty} X_i(k)X_{i+1}^{-1}(k) = \sqrt[n]{A} \quad \text{for } n \geq 2 \text{ and } i \in [1, n-1] \tag{3.1b}$$

where I_m denotes the identity matrix of dimension $m \times m$, and $H(k) \in C^{nm\times nm}$ is the tranpose of a block-K-circulant matrix with $K = A$ [21], viz.,

$$H(k) = \begin{bmatrix} X_1(k) & AX_n(k) & AX_{n-1}(k) & \cdots & AX_3(k) & AX_2(k) \\ X_2(k) & X_1(k) & AX_n(k) & \cdots & AX_4(k) & AX_3(k) \\ X_3(k) & X_2(k) & X_1(k) & \cdots & AX_5(k) & AX_4(k) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n-1}(k) & X_{n-2}(k) & X_{n-3}(k) & \cdots & X_1(k) & AX_n(k) \\ X_n(k) & X_{n-1}(k) & X_{n-2}(k) & \cdots & X_2(k) & X_1(k) \end{bmatrix} \in C^{nm\times nm}, \tag{3.1c}$$

$$X(k) = [X_1^T(k), X_2^T(k), \cdots, X_n^T(k)]^T \in C^{nm\times m}, \tag{3.1d}$$

$X_i(k) \in C^{m\times m}$, for $i = 1, 2, \cdots, n$, are block elements, and $r(\geq 2)$ is the convergence rate of the algorithm in (3.1). Note that $X_i(k)$ for $i = 1, 2, \cdots, n$ commutes with itself and with A.

The solution $X(k)$ of the block-state equation in (3.1) is the first block column of $H(k)$ in (3.1c). By taking the advantage of the K-circulant matrix, the algorithm with the quadratic convergence rate ($r = 2$) for computing the principal nth root of a complex matrix is given below.

Theorem 3.2.1 [21]

The solution of the block-state equation in (3.1) with the quadratic convergence rate ($r = 2$) at the kth step is $X(k)$, then we have

$$X(k) = [X_1^T(k), X_2^T(k), \cdots, X_n^T(k)]^T, \tag{3.2a}$$

where

$$X_l(k) = \sum_{i=1}^{l} X_i(k-1)X_{l+1-i}(k-1) + A\sum_{i=l+1}^{n} X_i(k-1)X_{n-i+l+1}(k-1)$$

$$\text{for } 1 \le l \le n-1, \tag{3.2b}$$

$$X_n(k) = \sum_{i=1}^{n} X_i(k-1)X_{n+1-i}(k-1) \quad \text{for } k \ge 1. \tag{3.2c}$$

Also, we obtain

$$\lim_{k\to\infty} X_i(k)X_j^{-1}(k) = (\sqrt[n]{A})^{j-i} \quad \text{for } i \ge 1 \text{ and } j \le n \tag{3.3a}$$

and

$$\lim_{k\to\infty} X_i(k)X_{i+1}^{-1}(k) = \sqrt[n]{A} \quad \text{for } 1 \le i \le n-1. \tag{3.3b}$$

The principal nth root of a matrix is unique. □

When the matrix A consists of any negative real eigenvalue (i.e., any $\arg(\sigma(A)) = \pi$), the algorithm in [21] cannot be directly applied to compute $\sqrt[n]{A}$. The matrix A can be rotated by a small angle to give $\hat{A} = Ae^{-j\delta\beta}$ (where $\delta\beta$ is a small positive real angle) so that $(\sqrt[n]{A})^p = (\sqrt[n]{\hat{A}})^p e^{jp\delta\beta/n}$.

3.3 Fast and Stable Algorithms for Finding the Principal nth Root of a Matrix

The purpose of this chapter is to generalize the fast and stable algorithm in [29] for computing the principal square root of a complex matrix to the fast and stable algorithm for computing the principal nth root of a complex matrix.

Premultiplying both sides of the block-state equation in (3.1a) with a matrix, block diag $[X_1^{-r}(k), X_1^{-r}(k), \cdots, X_1^{-r}(k)] \in C^{nm \times nm}$, and defining $X_1^{-r}(k)X_i(k+1) \triangleq \hat{X}_i(k+1)$ for $i = 1, 2, \cdots, n$ and $X_1(k)X_2^{-1}(k) \triangleq R(k)$, we obtain the normalized equivalent block-state equation in (3.1a) as

$$\hat{X}(k+1) = \hat{H}^{r-1}(k)\hat{X}(k), \tag{3.4a}$$

where

$$\hat{H}(k) = \begin{bmatrix} I_m & AR^{-n+1}(k) & AR^{-n+2}(k) & \cdots & AR^{-2}(k) & AR^{-1}(k) \\ R^{-1}(k) & I_m & AR^{-n+1}(k) & \cdots & AR^{-3}(k) & AR^{-2}(k) \\ R^{-2}(k) & R^{-1}(k) & I_m & \cdots & AR^{-4}(k) & AR^{-3}(k) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ R^{-n+2}(k) & R^{-n+3}(k) & R^{-n+4}(k) & \cdots & I_m & AR^{-n+1}(k) \\ R^{-n+1}(k) & R^{-n+2}(k) & R^{-n+3}(k) & \cdots & R^{-1}(k) & I_m \end{bmatrix} \tag{3.4b}$$

where $\hat{H}(k) \in C^{nm \times nm}$,

$$\hat{X}(k+1) = \left[\hat{X}_1^T(k+1), \hat{X}_2^T(k+1), \cdots, \hat{X}_n^T(k+1)\right]^T \in C^{nm \times m}, \tag{3.4c}$$

$$\hat{X}(k) = \left[I_m, \left(R^{-1}(k)\right)^T, \left(R^{-2}(k)\right)^T, \cdots, \left(R^{-n+1}(k)\right)^T\right]^T \in C^{nm \times m}, \tag{3.4d}$$

$$\text{with } R(k+1) = \hat{X}_1(k+1)\hat{X}_2^{-1}(k+1), \text{ and } \lim_{k \to \infty} R(k) = \sqrt[n]{A}. \tag{3.4e}$$

A recursive form can be obtained from (3.4a) by using the following definitions,

$$Y_j(k) \triangleq \hat{H}^{j-1}(k)\hat{X}(k), \tag{3.5a}$$

$$Y_{j-1}(k) \triangleq \hat{H}^{j-2}(k)\hat{X}(k), \tag{3.5b}$$

where

$$Y_j(k) \triangleq \left[Y_{1,j}^T(k), \left(R^{-1}(k)Y_{2,j}(k)\right)^T, \left(R^{-2}(k)Y_{3,j}(k)\right)^T, \cdots, \left(R^{-n+1}(k)Y_{n,j}(k)\right)^T\right]^T, \tag{3.5c}$$

the block vector $Y_j(k) \in C^{nm \times m}$ and the block elements $Y_{i,j}(k) \in C^{m \times m}$ for $i = 1, 2, \cdots, n$. Note that the subscript j in (3.5) denotes the index of the convergence rate. Then, from (3.5), we obtain the following recursive algorithm for $j = 2, 3, \cdots, r$ and any k,

$$Y_j(k) = T(k)Y_{j-1}(k), \quad Y_{i,1}(k) = I_m \quad \text{for } i = 1, 2, \cdots, n, \tag{3.5d}$$

where

$$T(k) = \begin{bmatrix} I_m & AR^{-n}(k) & AR^{-n}(k) & \cdots & AR^{-n}(k) \\ I_m & I_m & AR^{-n}(k) & \cdots & AR^{-n}(k) \\ I_m & I_m & I_m & \cdots & AR^{-n}(k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_m & I_m & I_m & \cdots & AR^{-n}(k) \\ I_m & I_m & I_m & \cdots & I_m \end{bmatrix} \in C^{nm \times nm}. \tag{3.5e}$$

Substituting $\hat{X}_1(k+1) = Y_{1,r}(k)$ and $\hat{X}_2(k+1) = R^{-1}(k)Y_{2,r}(k)$ into (3.5e), we have

$$R(k+1) = R(k)Y_{2,r}^{-1}(k)Y_{1,r}(k) \ , \quad R(0) = I_m \text{ for } k = 0, 1, 2, \cdots. \tag{3.6a}$$

Note that $R(k)$, $Y_{2,r}^{-1}(k)$ and $Y_{1,r}(k)$ commute with each other. Let us define $G(k) \triangleq AR^{-n}(k)$, and then from (3.6a), we obtain the following equation,

$$G(k+1) = G(k)[Y_{2,r}(k)Y_{1,r}^{-1}(k)]^n, \quad G(0) = A \quad \text{for } k = 0, 1, 2, \cdots. \tag{3.6b}$$

Expanding the matrix equation in (3.5d), we have

$$Y_{1,j}(k) = Y_{1,(j-1)}(k) + G(k)\big[Y_{2,(j-1)}(k) + Y_{3,(j-1)}(k) + \cdots + Y_{n,(j-1)}(k)\big],$$

$$Y_{2,j}(k) = Y_{1,(j-1)}(k) + Y_{2,(j-1)}(k) + G(k)\big[Y_{3,(j-1)}(k) + Y_{4,(j-1)}(k) + \cdots + Y_{n,(j-1)}(k)\big],$$

$$\vdots$$

$$\vdots$$

$$Y_{(n-1),j}(k) = Y_{1,(j-1)}(k) + Y_{2,(j-1)}(k) + \cdots + Y_{(n-1),(j-1)}(k) + G(k)Y_{n,(j-1)}(k),$$

$$Y_{n,j}(k) = Y_{1,(j-1)}(k) + Y_{2,(j-1)}(k) + \cdots + Y_{n,(j-1)}(k), \tag{3.7a}$$

or, in general form,

$$Y_{i,j}(k) = \sum_{p=1}^{i} Y_{p,(j-1)}(k) + G(k) \sum_{\ell=i+1}^{n} Y_{\ell,(j-1)}(k),$$

$$\text{for } j = 2,3,\cdots,r,\ i = 1,2,\cdots,n, \text{ and } \quad k = 0,1,2,\cdots. \tag{3.7b}$$

Combining the algorithms in (3.6a), (3.6b) and (3.7b), we obtain the desired algorithm for $k = 0,1,2,\cdots$ as follows,

$$Y_{i,j}(k) = \sum_{p=1}^{i} Y_{p,(j-1)}(k) + G(k) \sum_{\ell=i+1}^{n} Y_{\ell,(j-1)}(k)$$

$$\text{for } j = 2,3,\cdots,r, \text{ and } i = 1,2,\cdots,n, \tag{3.8a}$$

$$G(k+1) = G(k)\Big[Y_{2,r}(k)Y_{1,r}^{-1}(k)\Big]^n,\ G(0) = A,\ \lim_{k\to\infty} G(k) = I_m, \tag{3.8b}$$

$$R(k+1) = R(k)Y_{2,r}^{-1}(k)Y_{1,r}(k),\ R(0) = I_m,\ \lim_{k\to\infty} R(k) = \sqrt[n]{A}, \tag{3.8c}$$

where n denotes the index of the nth root of a matrix and r is the order of the desired convergence rate. Let r=2 and 3 in (3.8), respectively, we obtain the nth root algorithms as shown below.

When $r = 2$, (3.8) becomes

$$G(k+1) = G(k)\Big\{\big[2I_m + (n-2)G(k)\big]\big[I_m + (n-1)G(k)\big]^{-1}\Big\}^n,$$

$$G(0) = A, \quad \lim_{k\to\infty} G(k) = I_m, \tag{3.9a}$$

$$R(k+1) = R(k)\big[2I_m + (n-2)G(k)\big]^{-1}\big[I_m + (n-1)G(k)\big],$$

$$R(0) = I_m, \quad \lim_{k\to\infty} R(k) = \sqrt[n]{A}. \tag{3.9b}$$

When $r = 3$, we obtain

$$G(k+1) = G(k)\Big\{\big[3I_m + (\frac{n^2+5n-12}{2})G(k) + (\frac{n^2-5n+6}{2})G^2(k)\big]\times$$

$$\big[I_m + (\frac{n^2+3n-4}{2})G(k) + (\frac{n^2-3n+2}{2})G^2(k)\big]^{-1}\Big\}^n,$$

$$G(0) = A, \quad \lim_{k\to\infty} G(k) = I_m, \tag{3.9c}$$

$$R(k+1) = R(k)\Big[3I_m + (\frac{n^2+5n-12}{2})G(k) + (\frac{n^2-5n+6}{2})G^2(k)\Big]^{-1} \times$$

$$\Big[I_m + (\frac{n^2+3n-4}{2})G(k) + (\frac{n^2-3n+2}{2})G^2(k)\Big],$$

$$R(0) = I_m, \quad \lim_{k\to\infty} R(k) = \sqrt[n]{A}. \tag{3.9d}$$

Now, we list some commonly used pairs as shown below.
When $r = 2$ and $n = 2$, we have

$$G(k+1) = G(k)\Big\{[2I_m][I_m + G(k)]^{-1}\Big\}^2, \quad G(0) = A, \tag{3 10a}$$

$$R(k+1) = R(k)[2I_m]^{-1}[I_m + G(k)], \quad R(0) = I_m, \tag{3.10b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[2]{A}. \tag{3.10c}$$

When $r = 2$ and $n = 3$, we have

$$G(k+1) = G(k)\Big\{[2I_m + G(k)][I_m + 2G(k)]^{-1}\Big\}^3, \quad G(0) = A, \tag{3 11a}$$

$$R(k+1) = R(k)[2I_m + G(k)]^{-1}[I_m + 2G(k)], \quad R(0) = I_m, \tag{3.11b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[3]{A}. \tag{3.11c}$$

When $r = 2$ and $n = 4$, we have

$$G(k+1) = G(k)\Big\{[2I_m + 2G(k)][I_m + 3G(k)]^{-1}\Big\}^4, \quad G(0) = A, \tag{3 12a}$$

$$R(k+1) = R(k)[2I_m + 2G(k)]^{-1}[I_m + 3G(k)], \quad R(0) = I_m, \tag{3 12b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[4]{A}. \tag{3.12c}$$

When $r = 3$ and $n = 2$, we have

$$G(k+1) = G(k)\Big\{[3I_m + G(k)][I_m + 3G(k)]^{-1}\Big\}^2, \quad G(0) = A, \tag{3 13a}$$

$$R(k+1) = R(k)[3I_m + G(k)]^{-1}[I_m + 3G(k)], \quad R(0) = I_m, \tag{3.13b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[3]{A}. \tag{3.13c}$$

When $r = 3$ and $n = 3$, we have

$$G(k+1) = G(k)\Big\{[3I_m + 6G(k)][I_m + 7G(k) + G^2(k)]^{-1}\Big\}^3, \quad G(0) = A, \tag{3.14a}$$

$$R(k+1) = R(k)[3I_m + 6G(k)]^{-1}[I_m + 7G(k) + G^2(k)], \quad R(0) = I_m, \tag{3.14b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[3]{A}. \tag{3.14c}$$

When $r = 3$ and $n = 4$, we have

$$G(k+1) = G(k)\Big\{[3I_m + 12G(k) + G^2(k)][I_m + 12G(k) + 3G^2(k)]^{-1}\Big\}^4,$$

$$G(0) = A, \tag{3.15a}$$

$$R(k+1) = R(k)[3I_m + 12G(k) + G^2(k)]^{-1}[I_m + 12G(k) + 3G^2(k)],$$

$$R(0) = I_m, \tag{3.15b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[4]{A}. \tag{3.15c}$$

Some other cases are listed below.
When $r = 4$ and $n = 2$, we have

$$G(k+1) = G(k)\Big\{[4I_m + 4G(k)][I_m + 6G(k) + G^2(k)]^{-1}\Big\}^2,$$

$$G(0) = A, \tag{3.16a}$$

$$R(k+1) = R(k)[4I_m + 4G(k)]^{-1}[I_m + 6G(K) + G^2(k)],$$

$$R(0) = I_m, \tag{3.16b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[3]{A}. \tag{3.16c}$$

When $r = 4$ and $n = 3$, we have

$$G(k+1) = G(k)\Big\{[4I_m + 19G(k) + 4G^2(k)][I_m + 16G(k) + 10G^2(k)]^{-1}\Big\}^3,$$

$$G(0) = A, \tag{3.17a}$$

$$R(k+1) = R(k)[4I_m + 19G(k) + 4G^2(K)]^{-1}[I_m + 16G(k) + 10G^2(k)],$$

$$R(0) = I_m, \tag{3.17b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[3]{A}. \tag{3.17c}$$

When $r = 4$ and $n = 4$, we have

$$G(k+1) = G(k)\Big\{[4I_m + 40G(k) + 20G^2(k)][I_m + 31G(k) + 31G^2(K) + G^3(k)]^{-1}\Big\}^4,$$

$$G(0) = A, \tag{3.18a}$$

$$R(K+1) = R(k)[4I_m + 40G(K) + 20G^2(k)]^{-1}[I_m + 31G(k) + 31G^2(k) + G^3(k)],$$

$$R(0) = I_m, \tag{3.18b}$$

$$\lim_{k\to\infty} R(k) = \sqrt[4]{A}. \tag{3.18c}$$

Theorem 3.3.1

The principal nth root algorithm in (3.8) with the $r(\geq 2)$th-order convergence rate is numerically stable in the sense that the perturbations arising from the round-off errors at the kth iteration have only a bounded effect on succeeding iterates if no new round-off errors are introduced on succeeding iterates.

Proof

The convergence rate of the algorithm in (3.8) is the same as that in (3.1) because the algorithm in (3.8) is derived from the algorithm in (3.1). The numerical stability of the algorithm in (3.8) can be analyzed below.

Consider the principal nth root algorithms in (3.9), which has quadratic convergence rate ($r = 2$). Our objective is to show that the algorithm in (3.9) is numerically stable in the sense that perturbations arising from the rounding errors at the kth iteration do not lead to unbounded perturbations on succeeding iterates. Let the perturbed models be $\hat{G}(k)$ and $\hat{R}(k)$ and the associated round-off errors be $E(k)$ and $F(k)$, respectively. Hence by definition, $\hat{G}(k)$=G(k)+E(k) and $\hat{R}(k) = R(k) + F(k)$. Our purpose is to analyze how the error matrices $E(k)$ and $F(k)$ propogate at the $(k+1)$th stage. To simplify the analysis, we assume that no round-off errors occur when we compute $\hat{G}(k+1)$ and $\hat{R}(k+1)$ in the following equations,

$$\hat{G}(k+1) = \hat{G}(k)\Big\{[2I_m + (n-2)\hat{G}(k)][I_m + (n-1)\hat{G}(k)]^{-1}\Big\}^n, \tag{3.19a}$$

$$\hat{R}(k+1) = \hat{R}(k)[2I_m + (n-2)\hat{G}(k)]^{-1}[I_m + (n-1)\hat{G}(k)]. \tag{3.19b}$$

Substituting $\hat{G}(k)(= G(k) + E(k))$ and $\hat{R}(k)(= R(k) + F(k))$ into (3.19a) using the perturbation formula, we have

$$(D + \triangle)^{-1} = D^{-1} - D^{-1}\triangle D^{-1} + o(\| \triangle \|), \tag{3.20}$$

where the D and $\triangle$ are matrices and $o(\| \triangle \|)$ is the high order trivial term of $(\| \triangle \|)$. Omitting the high order trivial terms of $E(k)$ and $F(k)$ results in

$$G(k+1) + E(k+1) = [G(k) + E(k)]\Big\{\{2I_m + (n-2)G(k) + (n-2)E(k)\}\times$$

$$\{[I_m + (n-1)G(k)]^{-1} - [I_m + (n-1)G(k)]^{-1}(n-1)E(k)[I_m + (n-1)G(k)]^{-1}\}\Big\}^n \tag{3.21a}$$

$$= [G(k) + E(k)]\Big\{[2I_m + (n-2)E(k)][I_m + (n-1)G(k)]^{-1} - [2I_m + (n-2)G(k)]\times$$

$$[I_m + (n-1)G(k)]^{-1}(n-1)E(k)[I_m + (n-1)G(k)]^{-1} + (n-2)E(k)\times$$

$$[I_m + (n-1)G(k)]^{-1} - (n-2)E(k)[I_m + (n-1)G(k)]^{-1}(n-1)\times$$

$$E(k)[I_m + (n-1)G(k)]^{-1}\Big\}^n. \tag{3.21b}$$

When $k \to \infty$, $R(k) \to \sqrt[n]{A}$. Hence $G(k) = AR^{-1}(k) \to I_m$. Thus, (3.21$b$) becomes

$$G(k+1) + E(k+1) = [I_m + E(k)]\Big\{I_m - \frac{(n-1)}{n}E(k) + \frac{(n-2)}{n}E(k)\Big\}^n \tag{3.21c}$$

$$= [I_m + E(k)]\Big\{I_m - \frac{E(k)}{n}\Big\}^n, \tag{3.21d}$$

$$G(k+1) + E(k+1) = [I_m + E(k)][I_m - E(k)] = I_m. \tag{3.21e}$$

Substituting $k \to \infty$ to (3.9a), we obtain

$$G(k+1) = I_m. \tag{3.21f}$$

Thus, from (3.21e), we prove

$$E(k+1) = 0_m. \tag{3.21g}$$

Similarily, substituting $\hat{G}(k)$ and $\hat{R}(k)$ into (3.19b), we get

$$R(k+1) + F(k+1) = \left[R(k) + F(k)\right]\left[I_m + (n-1)G(k) + (n-1)E(k)\right] \times$$

$$\left\{\left[2I_m + (n-2)G(k)\right]^{-1} - \left[2I_m + (n-2)G(k)\right]^{-1}(n-2)E(k)\left[2I_m + (n-2)G(k)\right]^{-1}\right\} \tag{3.22a}$$

$$= \left[R(k) + F(k)\right]\Big\{\left[I_m + (n-1)G(k)\right]\left[2I_m + (n-2)G(k)\right]^{-1} - \left[I_m + (n-1)G(k)\right] \times$$

$$\left[2I_m + (n-2)G(k)\right]^{-1}(n-2)E(k)\left[2I_m + (n-2)G(k)\right]^{-1} +$$

$$(n-1)E(k)\left[2I_m + (n-2)G(k)\right]^{-1}\Big\}. \tag{3.22b}$$

Subtracting (14b) from (27b) and substituting $G(k) = I_m$ for $k \to \infty$ into (3.22b), we get

$$F(k+1) = F(k) + R(k)\Big[\frac{-(n-2)}{n}E(k) + \frac{(n-1)}{n}E(k)\Big] \tag{3.22c}$$

$$= F(k) + R(k)\frac{E(k)}{n}. \tag{3.22d}$$

The block-state equations in (3.21g) and (3.22d) with a null-system matrix and an identity-system matrix, respectively, are stable because the eigenvalues of the system matrices in (3.21g) and (3.22d) are zeros and ones, respectively. If we make a further assumption that no new round-off errors are introduced at the $(k+2)$th stage of the iterations, then (3.22d) becomes $F(k+2) = F(k+1) + R(k+1)E(k+1)/n = F(k+1)$. This suggests that the perturbations arising from the round-off errors at the kth iteration have only bounded effects on succeeding iterates. Thus, the algorithm in (3.9) is numerically stable provided that the above assumptions hold. In a similar manner, we can prove that the algorithm in (3.8) is numerically stable for $r \geq 3$. ■

One of the applications of the principal nth root of a matrix is in the derivation of the matrix-sector algorithm which in turn has many applications in solving control- system problems [26,27].

3.4 Illustrative Example

Example 3.4.1

Given a stiff matrix [22,29],

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0.01 & 0 & 0 \\ -1 & -1 & 100 & 100 \\ -1 & -1 & -100 & 100 \end{bmatrix}$$

where $\sigma(A) = \{0.01,\ 1,\ 100 \pm j100\}$, it is desired to find the $\sqrt[3]{A}$. The exact solution is

$$\sqrt[3]{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -0.792481 & 0.215444 & 0 & 0 \\ -0.013484 & -0.013483 & 5.032481 & 1.348449 \\ -0.048250 & -0.048173 & -1.348449 & 5.032481 \end{bmatrix},$$

where each eigenvalue of $\sqrt[3]{A}$ $(= \{0.215444,\ 1,\ 5.032481 \pm j1.348449\})$ is the principal cubic root of each $\sigma(A)$. Let us define the absolute error $e_a(k) \triangleq \|R(k) - \sqrt[3]{A}\|_\infty$, where $R(k)$ is the computed cubic root of A at the kth iteration. For this example, the upper limit for the iteration index k is taken as 30.

Applying the algorithm in [21] with $n = 3$ and $r = 2$ in (3.3), we have the result as shown in Table 3.1. We find that this algorithm converges in the usual sense at $k = 6$ with the $e_a(k) = 4.347 \times 10^{-7}$; however, it diverges very quickly. Therefore, this algorithm is numerically unstable.

Applying the algorithm with $n = 3$ and $r = 2$ in (3.11), we have the result as shown in Table 3.2. This algorithm converges at $k = 6$ with the $e_a(k) = 2.387 \times 10^{-15}$, then it remains invariant for $k \geq 6$. Employing the algorithm with $n = 2$ and $r = 3$ in (3.13), we obtain the result as shown in Table 3.3. This algorithm converges at $k = 5$ with the $e_a(k) = 6.610 \times 10^{-12}$, then it remains invariant at $e_a(k) = 6.577 \times 10^{-12}$ for $k \geq 6$. Using the algorithm with $n = 3$ and $r = 4$ in (3.17), we have the result as shown in Table 3.4. This algorithm converges at $k = 4$ with the $e_a(k) = 1.146 \times 10^{-12}$. Also, the relative error, $e_r(k) \triangleq \|R(k) - R(k-1)\|_\infty$, remains invariant at $1.3877787807814457E - 16$ for $k \geq 5$.

Therefore, the algorithms proposed in this chapter are numerically stable. Note that a high convergence-rate algorithm may not necessarily give the faster computational time.

3.5 Conclusion

Rapidly convergent and more stable recursive algorithms for finding the principal nth root of a matrix have been developed. The developed recursive algorithms can be applied to an ill-conditioned matrix containing large and small eigenvalues. By means of a perturbation analysis with suitable assumptions, it is shown that the proposed recursive algorithms are numerically more stable than the algorithms in [20,21,26]. The analysis of absolute numerical stability of the proposed algorithms has not been done in this chapter. The developed algorithms will enhance the capabilities of the existing computational algorithms such as the principal nth root

algorithm, the matrix-sign algorithm and the matrix-sector algorithm which in turn can be applied to many control-system problems.

k	$e(k)$
1	4.381223792317501
2	3.882609786095389
3	0.5926630610994657
4	0.1829671341223767
5	2.3013953289560130E-03
6	4.3471165761532760E-07
7	1.7435863254853623E-04
8	3.3132393959246582E-02
9	1.932622803044509
10	1202.571712161634
11	248467.2691452321
12	11959194.08924662
13	8210528857.322972
14	1837674003853.891
15	111162541995391.2
16	3.9045242057133462E+17
17	1.7048772918039683E+21
18	2.5431940655482168E+23
19	7.6820331894349227E+28

Table 3.1 Error analysis: the second-order numerically unstable algorithm

k	$e(k)$
1	4.381223792317501
2	2.424605807720003
3	0.2299970454407299
4	2.0625460001852391E-04
5	1.4119261315670428E-13
6	2.3869795029440866E-15
7	2.3869795029440866E-15
8	2.3869795029440866E-15
9	2.3869795029440866E-15
10	2.3869795029440866E-15
11	2.3869795029440866E-15
12	2.3869795029440866E-15
13	2.3869795029440866E-15
14	2.3869795029440866E-15
15	2.3869795029440866E-15
16	2.3869795029440866E-15
17	2.3869795029440866E-15
18	2.3869795029440866E-15
19	2.3869795029440866E-15
20	2.3869795029440866E-15
21	2.3869795029440866E-15
22	2.3869795029440866E-15
23	2.3869795029440866E-15
24	2.3869795029440866E-15
25	2.3869795029440866E-15
26	2.3869795029440866E-15
27	2.3869795029440866E-15
28	2.3869795029440866E-15
29	2.3869795029440866E-15
30	2.3869795029440866E-15

Table 3.2 Error analysis: the second-order numerically stable algorithm

k	$e(k)$
1	28.03572760609522
2	5.136272016363972
3	0.4574895834797120
4	6.2643637102149929E-04
6	6.5773046731277780E-12
7	6.5773046731277780E-12
8	6.5773046731277780E-12
9	6.5773046731277780E-12
10	6.5773046731277780E-12
11	6.5773046731277780E-12
12	6.5773046731277780E-12
13	6.5773046731277780E-12
14	6.5773046731277780E-12
15	6.5773046731277780E-12
16	6.5773046731277780E-12
17	6.5773046731277780E-12
18	6.5773046731277780E-12
19	6.5773046731277780E-12
20	6.5773046731277780E-12
21	6.5773046731277780E-12
22	6.5773046731277780E-12
23	6.5773046731277780E-12
24	6.5773046731277780E-12
25	6.5773046731277780E-12
26	6.5773046731277780E-12
27	6.5773046731277780E-12
28	6.5773046731277780E-12
29	6.5773046731277780E-12
30	6.5773046731277780E-12

Table 3.3 Error analysis: the third-order numerically stable algorithm

k	$e(k)$
1	3.882609786095389
2	0.732680674532077
3	3.8011418337163816E-04
4	1.1456591751668466E-11
5	1.1456588715902383E-11
6	1.1456585896976734E-11
7	1.1456590016944990E-11
8	1.1456594136913245E-11
9	1.1456598255881501E-11
10	1.1456602376849756E-11
11	1.1456606496818011E-11
12	1.1456610616786267E-11
13	1.1456614736754522E-11
14	1.1456618856722778E-11
15	1.1456622976691033E-11
16	1.1456627096659289E-11
17	1.1456631216627544E-11
18	1.1456635336595800E-11
19	1.1456639456564055E-11
20	1.1456643576532310E-11
21	1.1456647696500566E-11
22	1.1456651816468821E-11
23	1.1456655936437077E-11
24	1.1456660056405332E-11
25	1.1456664176373588E-11
26	1.1456668296341843E-11
27	1.1456672416310099E-11
28	1.1456676536278354E-11
29	1.1456680656246609E-11
30	1.1456684776214865E-11

Table 3.4 Error analysis: the fourth-order numerically stable algorithm

Fast and Stable Algorithms for Computing the Generalized Matrix-sector Function and the Separation of Matrix Eigenvalues

The matrix-sector function of A has been generalized to the matrix-sector function of $g(A)$, where the complex matrix A may have a real or complex characteristic polynomial and $g(A)$ is a matrix function of a conformal mapping. Based on the computationally fast and numerically stable algorithm for computing the principal nth root of a complex matrix, rapidly convergent and more stable recursive algorithms for finding the matrix-sector function and the generalized matrix-sector function have been developed in this chapter. Moreover, the generalized matrix-sector function of A is employed to separate the matrix eigenvalues relative to a sector, a circle, and a sector of a circle in a complex plane without actually seeking the characteristic polynomial and the matrix eigenvalues themselves. Also, the generalized matrix-sector function of A is utilized to carry out the block-diagonalization and block-triangularization of a system matrix, which are useful in developing applications to mathematical science and control-system problems [27,61].

4.1 Introduction

The matrix-sign function introduced by Robert [17] has been successfully applied to solve systems science and engineering problems [1,9,17,30], [33]-[37] such as the solutions of the matrix Lyapunov and Riccati equations and the separation of matrix eigenvalues relative to strips, trapezoids, and circles in the complex plane without actually seeking the characteristic polynomial and matrix eigenvalues themselves. The important features of the use of the matrix-sign function [9,17] to systems science and engineering problems are: (a) the matrix-sign functions preserve the eigenvectors of a complex matrix which may have a real or complex characteristic polynomial; (b) the associated matrix-sign algorithms converge quickly and the convergence speeds are independent of the dimension of the system.

The matrix-sign function of A, which may be considered as a matrix-2-sector function of A and can be expressed as $\mathrm{Sign}(A) = A[\sqrt[2]{A^2}]^{-1}$ where $\sqrt[2]{A^2}$ is the principal square root of a complex matrix A^2, has been extened to the matrix-sector function of A [26], which is a matrix-n-sector function of A and can be

expressed by $\text{Sector}_n(A) = A[\sqrt[n]{A^n}]^{-1}$ where $\sqrt[n]{A^n}$ is the principal nth root of A^n. One of the applications of the principal nth root of a matrix is in the derivation of the matrix-sector algorithm which in turn has many applications in solving control-system problem. The matrix-sector function of A has been used for the separation of the matrix eigenvalues relative to an open sector of a complex plane and for block-diagonalization of a system matrix.

The purposes of this chapter are : (a) derive fast and stable algorithms for computing the matrix-sector function; (b) the matrix-sector function of A is generalized to the matrix-sector function of $g(A)$ where $g(A)$ is the matrix function of a conformal mapping; (c) the generalized matrix-sector function of A is applied to the system matrix A for the separation of matrix eigenvalues relative to a sector, a circle, and a sector of a circle; (d) the generalized matrix-sector function of A is utilized for block-diagonalization and block-triangularization of the system matrix A.

4.2 Definition and Properties of the Matrix-sector Function

To develop fast and stable algorithms for computing the matrix-sector function, the generalized matrix-sector function and their associated functions with applications, we review the scalar- and matrix-sector functions in the following.

The scalar-n-sector function of λ is defined as follows.

Definition 4.2.1 [26]

Let $\lambda \in C$ be expressed by $\lambda = |\lambda| e^{j\theta}$, where $\lambda \neq 0$, $j = \sqrt[2]{-1}$, $\theta \in [0,\ 2\pi)$ and $\theta \neq 2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$. Then, the scalar-n-sector function of λ, defined as $\text{Sector}_n(\lambda)$ or $S_n(\lambda)$, is

$$\text{Sector}_n(\lambda) \triangleq S_n(\lambda)$$

$$\triangleq e^{j2\pi q/n} = \lambda / \sqrt[n]{\lambda^n} \quad \text{for} \quad q \in [0,\ n-1], \tag{4.1}$$

where λ lies inside the sector in C bounded by the sector angles $2\pi(q - \frac{1}{2})/n$ and $2\pi(q + \frac{1}{2})/n$, and $\sqrt[n]{\lambda^n}$ is the principal nth root of λ^n. When n=2, the scalar-sector function of λ becomes the sign function of λ [9,17], i.e.,

$$\text{Sector}_2(\lambda) \triangleq S_2(\lambda) = e^{jq\pi}$$

$$= \lambda / \sqrt[2]{\lambda^2} = \text{Sign}(\lambda) \quad \text{for} \quad q \in [0,\ 1]. \tag{4.2}$$

□

The matrix-sector function of A is defined as in the following.

Definition 4.2.2 [26,27]

Let $A \in C^{m\times m}, \sigma(A) = \{\lambda_i, i = 1,2,\cdots,m\}, \lambda_i \neq 0$ and $\arg(\lambda_i) \neq 2\pi(k + 1/2)/n$ for $k \in [0, n-1]$. In addition, let M be a modal matrix of A, i.e., $A = MJ_AM^{-1}$, where J_A is a matrix containing Jordan blocks of A. Then the matrix-sector function of A, denoted by $\text{Sector}_n(A)$ or $S_n(A)$, is defined as

$$\text{Sector}_n(A) = \mathrm{S}_n(A) = M\Big[\bigoplus_{i=1}^{m} \mathrm{S}_n(\lambda_i)\Big]M^{-1}, \tag{4.3}$$

where $\mathrm{S}_n(\lambda_i)$ is the scalar-n-sector function of λ_i. □

The matrix-sector function $\mathrm{S}_n(A)$ defined in Definition 4.2.2 can be expressed as

$$\mathrm{S}_n(A) = A(\sqrt[n]{A^n})^{-1}, \tag{4.4a}$$

where $\sqrt[n]{A^n}$ is the principal nth root of A^n. Also, the associated matrix-sign function, denoted by Sign(A) [9,17], becomes

$$\mathrm{S}_2(A) = A(\sqrt[2]{A^2})^{-1} = \mathrm{Sign}(A). \tag{4.4b}$$

Moreover, the partitioned matrix-sector function of A can be described as follows.

Definition 4.2.3 [26,27]

Let $A \in C^{m\times m}, \sigma(A) = \{\lambda_i, i = 1,2,\cdots,m\}, \lambda_i \neq 0$, and $\arg(\lambda_i) \neq 2\pi(p+1/2)n$ for $p \in [0, n-1]$. Also, let M be a modal matrix of A . Then, the qth matrix-n-sector function of A, denoted by $\mathrm{S}_n^{(q)}(A)$, is defined by

$$\mathrm{S}_n^{(q)}(A) = M\Big[\bigoplus_{i=1}^{m} \mathrm{S}_n^{(q)}(\lambda_i)\Big]M^{-1}, \tag{4.5}$$

where the qth scalar-n-sector function of λ_i, denoted by $\mathrm{S}_n^{(q)}(\lambda_i)$, is

$$\mathrm{S}_n^{(q)}(\lambda_i) = \begin{cases} 1, & \text{when } 2\pi(q-1/2)/n < \arg(\lambda_i) < 2\pi(q+1/2)/n \text{ for } q \in [0, n-1] \\ 0, & \text{otherwise.} \end{cases}$$

□

The qth matrix-n-sector function of A can be obtained by the following equation,

$$\mathrm{S}_n^{(q)}(A) = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{S}_n(A)e^{-j2\pi q/n}]^{i-1} \quad \text{for } q \in [0, n-1]. \tag{4.6}$$

Separation of matrix eigenvalues is one of the applications of the matrix-sector function in systems theory. For example, the number of eigenvalues of $A \in C^{m\times m}$, which lie within the sector angles $2\pi(q-1/2)/n$ and $2\pi(q+1/2)/n$, where $q \geq 0$ and $n \geq 1$, is trace($\mathrm{S}_n^{(q)}(A)$).

4.3 Fast and Stable Algorithm for Computing the Matrix-sector Function

One of the applications of the principal nth root of a matrix is in the derivation of the matrix-sector algorithm which in turn has many applications in solving control- system problems [26,27]. The fast and stable matrix-sector algorithm corresponding to the fast and stable principal nth root algorithm in (3.8) can be obtained by modifying (3.8), appropriately.

The direct use of the algorithm in (3.8) to compute $\sqrt[n]{A^n}$ and the matrix-sector function in (4.4a) where A is an ill-conditioned matrix may give numerically unstable results because it involves the computation of A^n which may be numerically unstable. To overcome this difficulty, we develop a fast and stable algorithm for computing the matrix-sector function in the following.

Defining $Q(k) \triangleq AR^{-1}(k)$ and $G(k) \triangleq A^n R^{-n}(k) = Q^n(k)$, and using $R(0) = I_m$ and $G(0) = A^n$, we obtain the simplied matrix-sector algorithm from the algorithm in (3.8) for $k = 0, 1, 2, \cdots$ as follows,

$$Y_{i,j}(k) = \sum_{p=1}^{i} Y_{p,(j-1)}(k) + Q^n(k) \sum_{\ell=i+1}^{n} Y_{\ell,(j-1)}(k),$$

$$Y_{i,1}(k) = I_m \quad \text{for } j = 2, 3, \cdots, r, \text{ and } i = 1, 2, \cdots, n, \tag{4.7a}$$

$$Q(k+1) = Q(k) Y_{2,j}(k) Y_{1,j}^{-1}(k), \quad Q(0) = A,$$

$$\lim_{k \to \infty} Q(k) = \mathrm{S}_n(A), \tag{4.7b}$$

where n denotes the index of the nth root of a matrix and r is the order of the desired convergence rate.

Corollary 4.3.1

The algorithm in (4.7) with the $r(\geq 2)$th-order convergence rate is numerically stable in the sense that at the kth iteration has only a bounded effect on succeeding iterates if no new errors are introduced on succeeding iterates.

Proof

The proof of Corollary 4.3.1 is similar to that in Theorem 3.3.1. ■

Some explicit forms of the algorithm in (4.7) are listed below.
When $r = 2$, (4.7) becomes

$$Q(k+1) = Q(k)\Big[2I_m + (n-2)Q^n(k)\Big]\Big[I_m + (n-1)Q^n(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k \to \infty} Q(k) = \mathrm{S}_n(A). \tag{4.8}$$

When $r = 3$, (4.7) becomes

$$Q(k+1) = Q(k)\Big[3I_m + \frac{n^2+5n-12}{2}Q^n(k) + \frac{n^2-5n+6}{2}Q^{2n}(k)\Big]\times$$

$$\Big[I_m + \frac{n^2+3n-4}{2}Q^n(k) + \frac{n^2-3n+2}{2}Q^{2n}(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_n(A). \tag{4.9}$$

Substituting $n = 2, 3$ and 4 into (4.8) and (4.9), we obtain the following results.

When $r = 2$ and $n = 2$, we have

$$Q(k+1) = Q(k)\Big[2I_m\Big]\Big[I_m + Q^2(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_2(A), \tag{4.10a}$$

or

$$Q(k+1) = \frac{1}{2}\big[Q^{-1}(k) + Q(k)\big],$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_2(A). \tag{4.10b}$$

Note that $Q_n(k) = Q_n^{-1}(k)$ for $n = 2$ only.

When $r = 2$ and $n = 3$, we have

$$Q(k+1) = Q(k)\Big[2I_m + Q^3(k)\Big]\Big[I_m + 2Q^3(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_3(A). \tag{4.11}$$

When $r = 2$ and $n = 4$, we have

$$Q(k+1) = Q(k)\Big[2I_m + 2Q^4(k)\Big]\Big[I_m + 3Q^4(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_4(A). \tag{4.12}$$

When $r = 3$ and $n = 2$, we have

$$Q(k+1) = Q(k)\Big[3I_m + Q^2(k)\Big]\Big[I_m + 3Q^2(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_2(A). \tag{4.13}$$

When $r = 3$ and $n = 3$, we have

$$Q(k+1) = Q(k)\Big[3I_m + 6Q^3(k)\Big]\Big[I_m + 7Q^3(k) + Q^6(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_3(A). \tag{4.14}$$

When $r = 3$ and $n = 4$, we have

$$Q(k+1) = Q(k)\Big[3I_m + 12Q^4(k) + Q^8(k)\Big]\Big[I_m + 12Q^4(k) + 3Q^8(k)\Big]^{-1},$$

$$Q(0) = A, \quad \lim_{k\to\infty} Q(k) = \mathrm{S}_4(A). \tag{4.15}$$

Note that the algorithm in (4.10*b*) is the commonly used matrix-sign algorithm [1,9,17] for $r = 2$. Comparing the algorithm in (4.8) with that in [26,27] for determining the matrix-sector function, it can be noted that the proposed algorithms do significantly improve the computational aspects of the existing algorithms.

4.4 Definition, Computational Algorithms and Applications of the Generalized Matrix-sector Funcion

In this section, the matrix-sector function has been generalized to the matrix-section function of $\mathrm{g}(A)$ where $\mathrm{g}(A)$ is the matrix function of a conformal mapping, and the fast and stable algorithms for computing the matrix-sector function are employed for finding the generalized matrix-sector function. Also, the generalized matrix-sector function of A is applied to the system matrix A for the separation of matrix eigenvalues relative to a sector, circle, and a sector of a circle. Furthermore, the generalized matrix-sector function of A is utilized for block-diagonalization and block-triangularization of the system matrix A.

The generalized scalar-n-sector function of λ can be defined below.

Definition 4.4.1

Let the function of a conformal mapping be $\lambda \mapsto \mathrm{g}(\lambda)$ which maps simple closed curves L_q in the λ-plane onto the boundaries of the n minor sectors bounded by sector angles $2\pi(q - \frac{1}{2})/n$ and $2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$ in the $\mathrm{g}(\lambda)$-plane. Thus, the whole λ-plane is separated into open regions $\tilde{C}_q$ by the L_q such that the domains $\tilde{C}_q$ for $q \in [0,\ n-1]$ in the λ-plane will be mapped into the domains $\tilde{D}_q$ bounded by the sector angles $2\pi(q - \frac{1}{2})/n$ and $2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$ in the $\mathrm{g}(\lambda)$-plane, respectively. Hence, the generalized scalar-sector function of λ with $\mathrm{g}(\lambda) \neq 0$ and $\arg[\mathrm{g}(\lambda)] \neq 2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$, denoted by $\mathrm{Sector}_n(\mathrm{g}(\lambda))$ or $\mathrm{S}_n(\mathrm{g}(\lambda))$, is

$$\mathrm{Sector}_n(\mathrm{g}(\lambda)) \triangleq \mathrm{S}_n(\mathrm{g}(\lambda))$$

$$= e^{j2\pi q/n} = g(\lambda)/\sqrt[n]{(g(\lambda))^n}, \tag{4.16}$$

where λ lies within $\tilde{C}_q$ and $g(\lambda)$ lies within $\tilde{D}_q$ bounded by the sector angles $2\pi(q - \frac{1}{2})/n$ and $2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$, and $\sqrt[n]{(g(\lambda))^n}$ is the principal nth root of $(g(\lambda))^n$. □

When $n = 2$ and $g(\lambda)$ is the bilinear transformation, $\lambda \mapsto g(\lambda) = (\lambda - \rho)(\lambda + \rho)^{-1}$, then $g(\lambda)$ maps the origin-centred circle of radius ρ in the λ-plane onto the imaginary axis of the $g(\lambda)$-plane. Also, $g(\lambda)$ maps $\tilde{C}_0$, the exterior of the circle in the λ-plane, into the open right-half $g(\lambda)$-plane $\tilde{D}_0$, which is the 0th sector containing the set of sector angles $(-\pi/2,\ \pi/2)$. Moreover, $g(\lambda)$ maps $\tilde{C}_1$, the interior of the circle in the λ-plane, into the open left-half $g(\lambda)$-plane $\tilde{D}_1$, which is the 1th sector containing the set of sector angles $(\pi/2,\ 3\pi/2)$.

The extension of the generalized scalar-sector function of $\lambda \in C$ to the generalized matrix-sector function of $A \in C^{m\times m}$ and its associated functions with applications can be stated below.

Theorem 4.4.1

Let the matrix function of a conformal mapping be $A \mapsto g(A)$ where $A \in C^{m\times m}$, $\sigma(A) = \{\lambda_i,\ i = 1, 2, \ldots, m\}$, $g(\lambda_i) \neq 0$, and $\arg[g(\lambda_i)] \neq 2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$. Then, the generalized matrix-n-sector function of A, denoted by $\text{Sector}_n(g(A))$ or $S_n(g(A))$, is

$$\text{Sector}_n(g(A)) \triangleq S_n(g(A)) = M\left[\bigoplus_{i=1}^{m} S_n(g(\lambda_i))\right] M^{-1}$$

$$= g(A)[\sqrt[n]{(g(A))^n}]^{-1}, \tag{4.17}$$

where the matrix M is the modal matrix of A, and $S_n(g(\lambda_i))$ is the generalized scalar- sector function of λ_i. Also, $\sqrt[n]{(g(A))^n}$ is the principal nth root of $(g(A))^n$, which has the properties that $(\sqrt[n]{(g(A))^n})^n = (g(A))^n$ and each eigenvalue of $\sqrt[n]{(g(A))^n}$ is the principal nth root of each $(g(\lambda_i))^n$.

The associated qth generalized matrix-sector function of A with $\arg(\lambda_i) \neq 0$ and $\arg[g(\lambda_i)] \neq 2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$, denoted by $S_n^{(q)}(g(A))$, is

$$S_n^{(q)}(g(A)) \triangleq M\left[\bigoplus_{i=1}^{m} S_n^{(q)}(g(\lambda_i))\right] M^{-1}$$

$$= \frac{1}{n}\sum_{i=1}^{n}[S_n(g(A))e^{-j2\pi q/n}]^{i-1} \in C^{m\times m} \quad \text{for } q \in [0,\ n-1], \tag{4.18a}$$

where the qth generalized scalar-sector function of λ_i is

$$S_n^{(q)}(g(\lambda_i)) = \begin{cases} 1 & \text{when } \lambda_i \in \tilde{C}_q \quad \text{for } q \in [0,\ n-1] \\ 0 & \text{otherwise.} \end{cases}$$

The complement of $S_n^{(q)}(g(A))$, denoted by $\bar{S}_n^{(q)}(g(A))$, is

$$\bar{S}_n^{(q)}(g(A)) \triangleq I_m - S_n^{(q)}(g(A)) \in C^{m\times m} \quad \text{for } q \in [0,\ n-1], \tag{4.18b}$$

where I_m designates the $m \times m$ identity matrix. □

The number of eigenvalues of A lying inside the domain $\tilde{C}_q$, denoted by N_q, is

$$N_q = \text{trace}[S_n^{(q)}(g(A))] \quad \text{for } q \in [0,\ n-1], \tag{4.19}$$

and the A-invariant subspace of $S_n^{(q)}(g(A))$, denoted by $S^{(q)}$, is

$$S^{(q)} \triangleq \text{ind}[S_n^{(q)}(g(A))] \in C^{m\times N_q} \quad \text{for } q \in [0,\ n-1], \tag{4.20}$$

where ind[·] in (4.20) designates the collection of the independent (abbreviation as ind) column vectors of the matrix [·]. The matrices $S^{(q)}$ for $q \in [0,\ n-1]$ can be used to construct a block-modal matrix, M_S, for carrying out the block-diagonalization of the system matrix A, i.e.,

$$M_S^{-1} A M_S = \text{block diag}[A_0, A_1, \ldots, A_{n-1}] \in C^{m\times m}, \tag{4.21a}$$

where

$$M_S = [S^{(0)}, S^{(1)}, \ldots, S^{(n-1)}] \in C^{m\times m}, \tag{4.21b}$$

and

$$A_q = (S^{(q)})^+ A (S^{(q)}) \in C^{N_q \times N_q},$$

$$(S^{(q)})^+ = [(S^{(q)})^* (S^{(q)})]^{-1} (S^{(q)})^* \in C^{N_q \times m},$$

$$\sigma(A_q) \subseteq \tilde{C}_q \quad \text{for } q \in [0,\ n-1]. \tag{4.21c}$$

The superscript * in (4.21c) designates the conjugate transpose.

The other A-invariant subspace of $S_n^{(q)}(g(A))$, defined as $V^{(q)}$, can be constructed as the collection of the independent row vectors of $S_n^{(q)}(g(A))$ and expressed as

$$V^{(q)} = \{\text{ind}[(S_n^{(q)}(g(A)))^T]\}^T \in C^{N_q\times m} \quad \text{for } q \in [0,\ n-1]. \tag{4.22}$$

Hence, the associated block-modal matrix, M_V, can be constructed and used for block-diagonalization of the system matrix A,

$$M_V A M_V^{-1} = \text{block diag}[\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_{n-1}] \in C^{m\times m}, \tag{4.23a}$$

where

$$M_V = [(V^{(0)})^T, (V^{(1)})^T, \ldots, (V^{(n-1)})^T]^T \in C^{m \times m}, \tag{4.23b}$$

and

$$\hat{A}_q = (V^{(q)}) A (V^{(q)})^+ \in C^{N_q \times N_q},$$

$$(V^{(q)})^+ = (V^{(q)})^* [(V^{(q)})(V^{(q)})^*]^{-1} \in C^{m \times N_q},$$

$$\sigma(\hat{A}_q) \subseteq C_q \quad \text{for } q \in [0,\ n-1]. \tag{4.23c}$$

Also, by combining the A-invariant subspaces of $S_n^{(q)}(g(A))$ in (4.20) and (4.22), a similarity transformation of the system matrix T can be constructed for block-triangularization of the system matrix A so that each submatrix of the block-triangularized system matrix contains the eigenvalues lying within each specified region of the λ-plane.

The similarity-transformation matrix and its inverse are

$$T = \begin{bmatrix} (\bar{S}^{(q)})^+ \\ V^{(q)} \end{bmatrix} \in C^{m \times m},$$

$$\bar{S}^{(q)} \triangleq \text{ind}[I_m - S_n^{(q)}(g(A))] \in C^{m \times (m - N_q)} \tag{4.24a}$$

and

$$T^{-1} = [\bar{S}^{(q)},\ (V^{(q)})^+]. \tag{4.24b}$$

The block-triangularized system matrix becomes

$$A_T = TAT^{-1} = \begin{bmatrix} A_R & \vdots & A_{RL} \\ \cdots & \cdots & \cdots \\ 0 & \vdots & A_L \end{bmatrix}, \tag{4.25}$$

where $A_R = (\bar{S}^{(q)})^+ A (\bar{S}^{(q)})$, $\sigma(A_R) \subseteq$ the complement of $\tilde{C}_q$, $A_L = (V^{(q)})\ A\ (V^{(q)})^+$, $\sigma(A_L) \subseteq \tilde{C}_q$, and $A_{RL} = (\bar{S}^{(q)})^+ A (V^{(q)})^+$.

Proof

When $g(A) \equiv A$, the various results in this theorem have been proved in [9,26,35,36,37]. The corresponding results for the generalized version of the matrix-sector functions can be proven in a similar manner. ■

Replacing A in (4.7) with $g(A)$, we can obtain the fast and stable algorithm for computing the generalized matrix-sector function $S_n(g(A))$ in (4.17). When $\arg[g(\lambda_i)] = 2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$, the matrix $g(A)$ shall be rotated by a small positive real angle $(\Delta\beta)$ as $g(\bar{A}) \triangleq g(Ae^{-j\Delta\beta})$, so that the algorithm in (4.7) can still be applied to compute $S_n(g(A))$.

Corollary 4.4.1

Let $g(A) = (A - \rho I_m)(A + \rho I_m)^{-1}$, where $A \in C^{m\times m}$, $\det(A + \rho I_m) \neq 0$, and $\sigma(A) \cap \hat{C} = \phi$ and $\hat{C}$ is a circle of radius ρ with center at the origin of the λ-plane . The qth generalized matrix-sector function of A with n=2 and q=0 becomes

$$S_2^{(0)}(g(A)) = \frac{1}{2}[I_m + S_2(g(A))] \tag{4.26a}$$

and the complement of $S_2^{(0)}(g(A))$, denoted by $\bar{S}_2^{(0)}(g(A))$, is

$$\bar{S}_2^{(0)}(g(A)) = I_m - S_2^{(0)}(g(A)) = S_2^{(1)}(g(A))$$

$$= \frac{1}{2}[I_m - S_2(g(A))]. \tag{4.26b}$$

The number of eigenvalues of A lying in the exterior of the circle of radius ρ is N_0 = trace$[S_2^{(0)}(g(A))]$, and that in the interior of the circle is N_1 = trace$[S_2^{(1)}(g(A))]$ = $m - N_0$.

Proof

The bilinear transform (a conformal mapping), $g(\lambda) = (\lambda - \rho)(\lambda + \rho)^{-1}$, maps the circle of radius ρ in the λ-plane onto the imaginary axis of the $g(\lambda)$-plane and the interior (exterior) of the circle into the open left-half (open right-half) $g(\lambda)$-plane. Hence, Corollary 4.4.1 can be proved using Definitions 4.2.1 and 4.4.1 and Theorem 4.4.1. ■

To determine the number of matrix eigenvalues lying inside the intersection of two specific regions ($\tilde{C}_0$ and $\tilde{C}_1$) in a complex plane and to determine the associated A-invariant subspace of the intersection region, we present the following important result.

Corollary 4.4.2

Let $S_{n_1}^{(q_1)}(g_1(A))$ and $S_{n_2}^{(q_2)}(g_2(A))$ be two associated generalized matrix-sector functions of A which can be expressed as

$$S_{n_1}^{(q_1)}(g_1(A)) = M[I_{m_1} \oplus 0_{m_2}]M^{-1} \in C^{m\times m} \tag{4.27a}$$

and

$$S_{n_2}^{(q_2)}(g_2(A)) = M[I_{m_3} \oplus 0_{m_4}]M^{-1} \in C^{m\times m}, \tag{4.27b}$$

then, we have

$$S_{n_1}^{(q_1)}(g_1(A)) \times S_{n_2}^{(q_2)}(g_2(A)) = M[I_{m_x} \oplus 0_{m_y}]M^{-1}, \tag{4.27c}$$

where

$$m_x = \text{trace}[S_{n_1}^{(q_1)}(g_1(A)) \times S_{n_2}^{(q_2)}(g_2(A))], \quad \text{and } m_y = m - m_x.$$

Let

$$M_1 \triangleq \text{ind}[S_{n_1}^{(q_1)}(g_1(A))] \ , \quad M_2 \triangleq \text{ind}[S_{n_2}^{(q_2)}(g_2(A))] \ ,$$

$$M_1^+ A M_1 \triangleq A_0 \ , \quad \text{and } M_2^+ A M_2 \triangleq A_1$$

with $\sigma(A_0) \subseteq \tilde{C}_0$ and $\sigma(A_1) \subseteq \tilde{C}_1$,

also let

$$M_x \triangleq \text{ind}[S_{n_1}^{(q_1)}(g_1(A)) \times S_{n_2}^{(q_2)}(g_2(A))].$$

Then, we have

$$M_x^+ A M_x \triangleq A_x, \quad \sigma(A_x) \subseteq \tilde{C}_0 \cap \tilde{C}_1. \tag{4.27d}$$

The number of eigenvalues lying within $\tilde{C}_0 \cap \tilde{C}_1$ is m_x.

Proof

Corollary 4.4.2 can be proved by using the fact that the generalized matrix-sector function of A preserves the eigenvectors of A. ■

For engineering applications, we are often interested in selecting a sector of a circle in the λ-plane due to the consideration of damping ratio, damping frequency and decaying rate, etc., of the system. The separation of matrix eigenvalues relative to a sector of a circle can be stated below.

Corollary 4.4.3

Let $A \in C^{m \times m}$ and $\sigma(A) = \{\lambda_i,\ i = 1, 2, \ldots, m\}$. Also, let $\sigma(A) \cap (\ell_i,\ \ell_{i+1}) = \phi$ where ℓ_i and ℓ_{i+1} are two straight lines emanating from the original of the λ-plane at angles $2\pi(q - \frac{1}{2})/n$ and $2\pi(q + \frac{1}{2})/n$ for $q \in [0,\ n-1]$. Moreover, let $\sigma(A) \cap \hat{C} = \phi$, where $\hat{C}$ is a circle of radius ρ centred at the origin. Then, the generalized matrix-sector function of A with respect to this sector and the circle of radius ρ, denoted by $\tilde{S}^{(\phi_q, \rho)}(A)$, is

$$\tilde{S}^{(\phi_q, \rho)}(A) \triangleq S_n^{(q)}(A) \times S_2^{(1)}(g_2(A)), \tag{4.28a}$$

where

$$g_2(A) = (A - \rho I_m)(A + \rho I_m)^{-1} \quad \text{for } q \in [0,\ n-1]. \tag{4.28b}$$

The complement of $\tilde{S}^{(\phi_q, \rho)}(A)$ is $I_m - \tilde{S}^{(\phi_q, \rho)}(A)$. The number of matrix eigenvalues lying inside the closed sector is

$$\tilde{N}_q = \text{trace}[\tilde{S}^{(\phi_q, \rho)}(A)] \tag{4.28c}$$

Proof

Corollary 4.4.3 can be proved by using Theorem 4.4.1, Corollaries 4.4.1 and 4.4.2. ■

4.5 Illustrative Example

Consider a system matrix A,

$$A = \begin{bmatrix} 10.5 - j1.5 & -5.5 + j11.5 & -3.5 - j4.5 & -20.5 + j1.5 \\ 5.5 - j8.0 & 1.5 + j11.0 & -4.1 - j0.8 & -11.0 + j11.5 \\ 21.5 + j3.0 & -16.5 + j17.0 & -4.5 - j8.0 & -34.0 - j10.5 \\ 12.5 - j2.5 & -5.5 + j11.5 & -3.5 - j4.5 & -22.5 + j2.5 \end{bmatrix}. \tag{4.29}$$

Find

(a) The number of matrix eigenvalues lying inside the sector of a circle with a radius, ρ ($=|\sqrt[m]{\det(A)}|$ for $m = 4$), and a sector angle, $\phi_q \in (3\pi/4,\ 5\pi/4)$.

(b) The block-triangularization of the system matrix A such that $\sigma(A_L)$ lie inside the sector of a circle and $\sigma(A_R)$ lie outside the sector of a circle.

(c) The block-diagonalization of the system matrix A such that $\sigma(A_0)$ lie inside the sector of a circle and $\sigma(A_1)$ lie outside the sector of a circle.

Solution

To find the number of matrix eigenvalues lying within the closed sector, we use Corollary 4.4.3. The geometric mean of the matrix eigenvalues is $\rho = |\sqrt[4]{\det(A)}| = 3.2517$. Since the sector angle $\phi_q = 5\pi/4 - 3\pi/4 = \pi/2$, we decompose the entire λ-plane into n ($=2\pi/\phi_q=4$) sectors. As a result, the number of the sector, q, equals to two. Thus, the $q(=2)$th matrix-sector function of A is

$$\mathrm{S}_n^{(q)}(A) = \mathrm{S}_4^{(2)}(A)$$

$$= \begin{bmatrix} 1.0 + j0.0 & 0.0 + j0.0 & 0.0 + j0.0 & 0.0 + j0.0 \\ 0.0 + j0.5 & 0.0 - j0.5 & 0.2 + j0.1 & 0.5 - j1.0 \\ -1.0 + j0.5 & 0.0 - j2.5 & 1.0 + j0.5 & 2.5 + j0.0 \\ 0.0 + j0.0 & 0.0 + j0.0 & 0.0 + j0.0 & 1.0 + j0.0 \end{bmatrix}.$$

To use (4.28), we compute $\mathrm{g}_2(A)$ and $\mathrm{S}_2^{(1)}(\mathrm{g}_2(A))$ as

$$\mathrm{g}_2(A) = (A - \rho I_m)(A + \rho I_m)^{-1}$$

$$= \begin{bmatrix} -7.1074 + j8.4017 & -1.7329 - j7.4696 & 3.3344 + j0.8008 & 9.0711 - j8.4017 \\ -0.0838 + j4.6235 & -4.6659 - j0.6559 & 1.6462 - j1.2026 & 0.6559 - j6.6297 \\ -9.3308 + j4.4560 & 3.8846 - j8.1785 & 1.8800 + j4.6235 & 13.9150 - j5.3180 \\ -4.9359 + j5.8681 & -1.7329 - j7.4696 & 3.3344 + j0.8008 & 6.8996 - j5.8681 \end{bmatrix},$$

and

$$\mathrm{S}_2^{(1)}(\mathrm{g}_2(A))$$

$$= \begin{bmatrix} 2.5 - j0.5 & -0.5 + j1.5 & -0.5 - j0.5 & -2.5 + j0.5 \\ 0.5 - j1.0 & 1.5 + j1.0 & -0.5 + j0.0 & -1.0 + j1.5 \\ 2.5 + j0.0 & -1.5 + j2.0 & 0.5 - j1.0 & -4.0 - j0.5 \\ 1.5 - j0.5 & -0.5 + j1.5 & -0.5 - j0.5 & -1.5 + j0.5 \end{bmatrix}.$$

Thus, the desired $\tilde{S}^{(\phi_q,\rho)}(A)$ in (4.28a) becomes

$$\tilde{S}^{(\phi_q,\rho)}(A) = S_4^{(2)}(A) \times S_2^{(1)}(g_2(A))$$

$$= \begin{bmatrix} 2.5 - j0.5 & -0.5 + j1.5 & -0.5 - j0.5 & -2.5 + j0.5 \\ 0.5 - j0.5 & 0.5 + j0.5 & -0.3 + j0.1 & -0.5 + j0.5 \\ 1.5 + j0.5 & -1.5 - j0.5 & 0.5 - j0.5 & -1.5 - j0.5 \\ 1.5 - j0.5 & -0.5 + j1.5 & -0.5 - j0.5 & -1.5 + j0.5 \end{bmatrix}. \tag{4.30}$$

The number of eigenvalues lying within the sector of a circle is

$$N_q = \text{trace}[\tilde{S}^{(\phi_q,\rho)}(A))] = 2. \tag{4.31}$$

It is interesting to note that $\sigma(A) = \{\lambda_1 = \lambda_2 = \text{-2+j1}, \lambda_3 = -10, \lambda + 4 = -1 + j2\}$ and the repeated eigenvalues, $\{\lambda_1, \lambda_2\}$, lie in the desirable sector. Since the characteristic polynomial of A is a complex polynomial, the test procedures due to Gutman and Jury [73] and Zeheb and Hertz [74] can not directly be applied to determine the N_q in (4.31).

To find the block-triangularization of A, we use Theorem 4.4.1. The computed A_T in (4.25) is

$$A_T = TAT^{-1} = \begin{bmatrix} A_R & \vdots & A_{RL} \\ \cdots & \vdots & \cdots \\ 0 & \vdots & A_L \end{bmatrix}$$

$$= \left[\begin{array}{cc:cc} -16.25 - j3.75 & 13.75 - j8.75 & -7.292 - j1.875 & 8.375 - j5.542 \\ -0.75 - j7.25 & 5.25 + j5.75 & -0.875 - j3.458 & 5.458 + j2.875 \\ \hdashline 0.00 + j0.00 & 0.00 + j0.00 & -2.000 + j1.000 & 0.000 + j0.000 \\ 0.00 + j0.00 & 0.00 + j0.00 & 0.000 + j0.000 & -2.000 + j0.000 \end{array}\right], \tag{4.32}$$

where

$$T = \begin{bmatrix} (\bar{S}^{(q)})^+ \\ \cdots \\ V^{(q)} \end{bmatrix}$$

$$= \begin{bmatrix} -0.25+j0.25 & -0.083-j0.250 & -0.417-j0.417 & -0.25+j0.25 \\ -0.25+j0.25 & 0.250-j0.083 & -0.083-j0.583 & -0.25+j0.25 \\ \cdots & \cdots & \cdots & \cdots \\ 2.50-j0.050 & -0.500+j1.500 & -0.500-j0.500 & -2.50+j0.50 \\ 0.50-j0.50 & 0.500+j0.500 & -0.300+j0.100 & -0.50+j0.50 \end{bmatrix}$$

and

$$T^{-1} = \left[\bar{S}_1^{(q)} \;\middle|\; (V^{(q)})^+ \right]$$

$$= \left[\begin{array}{cc:cc} -1.5+j0.5 & 0.5-j1.5 & 0.250+j0.250 & 0.250-j0.750 \\ -0.5+j0.5 & 0.5-j0.5 & -0.417+j0.417 & 2.083-j0.417 \\ -1.5-j0.5 & 1.5+j0.5 & 0.250+j0.083 & -0.583-j0.750 \\ -1.5+j0.5 & 0.5-j1.5 & -0.250-j0.250 & -0.250+j0.750 \end{array}\right].$$

Note that $\sigma(A_R) = \{\lambda_3,\ \lambda_4\}$, and $\sigma(A_L) = \{\lambda_1,\ \lambda_2\}$.

The block-diagonalization of A in (4.21a) is

$$M_S^{-1} A M_S = \text{block diag}[A_0, A_1]$$

$$= \left[\begin{array}{cc:cc} -2.0+j1.0 & 0.0+j0.0 & 0.0+j0.0 & 0.0+j0.0 \\ 0.0+j0.0 & -2.0+j1.0 & 0.0+j0.0 & 0.0+j0.0 \\ \cdots & \cdots & \cdots & \cdots \\ 0.0+j0.0 & 0.0+j0.0 & -16.25-j3.75 & 13.75-j8.75 \\ 0.0+j0.0 & 0.0+j0.0 & -0.75-j7.25 & 5.25+j5.75 \end{array}\right], \qquad (4.33)$$

where

$$M_S = [S^{(0)}, S^{(1)}]$$

$$= \left[\begin{array}{cc:cc} 2.5-j0.5 & -0.5+j1.5 & -1.5+j0.5 & 0.5-j1.5 \\ 0.5-j0.5 & 0.5+j0.5 & -0.5+j0.5 & 0.5-j0.5 \\ 1.5+j0.5 & -1.5-j0.5 & -1.5-j0.5 & 1.5+j0.5 \\ 1.5-j0.5 & -0.5+j1.5 & -1.5+j0.5 & 0.5-j1.5 \end{array}\right].$$

Note that $\sigma(A_0) = \{\lambda_1,\ \lambda_2\}$ and $\sigma(A_1) = \{\lambda_3,\ \lambda_4\}$.

4.6 Conclusion

The matrix-sector function of A has been generalized to the matrix-sector function of $g(A)$. Based on the computationally fast and numerically stable algorithms for computing the principal nth root of a matrix, rapidly and stable algorithms for computing the matrix-sector function and the generalized matrix-sector function have beed developed. The generalized matrix-sector function of A has been utilized to carry out the separation of matrix eigenvalues relative to a sector, circle and a sector of a circle in the λ-plane. Also, the generalized matrix-sector function of A has been employed for block-diagonalization and block-triangularization of the system matrix; these are are useful in developing applications to mathematical science [32] and control-system problems [31].

Determining Continuous-time State Equations from Discrete-time State Equations Via the Principal qth Root Method

Fast computational methods are developed for finding the equivalent continuous-time state equations from discrete-time state equations. The computational methods utilize the direct-truncation method, the matrix continued-fraction method, and the geometric-series method in conjunction with the principal qth root of the discrete-time system matrix for quick determination of the approximants of a matrix- logarithm function. It is shown that the use of the principal qth root of a matrix enables us to enlarge the convergence region of the expansion of a matrix-logarithm function and to improve the accuracy of the approximants of the matrix-logarithm function [28].

5.1 Introduction

The identification [38] of a continuous-time system using the sampled input-output data of the system often results in an equivalent discrete-time model. Hence, the conversion of the obtained discrete-time model to the original continuous-time system is necessary. Also, a given discrete-time system is often transformed into an equivalent continuous-time model so that the well-developed continuous-time approaches such as the frequency-domain techniques [39] can efficiently be applied to the transformed model for analysis and design of sampled-data control systems [40].

Let the discrete-time system be

$$\begin{aligned} x(kT+T) &= Fx(kT) + Gu(kT), \\ y(kT) &= Cx(kT), \end{aligned} \tag{5.1}$$

where $x \in R^n$, $u \in R^m$, $y \in R^p$, the constant matrices F, G, and C are of

appropriate dimensions, and T is the sampling period. The equivalent continuous-time model is described by

$$\dot{x}(t) = Ax(t) + Bu(t),$$

$$y(t) = Cx(t), \tag{5.2}$$

where $x(t) \simeq x(kT)$, and $u(t) = u(kT)$ for $t = kT$.

The relationships [41]-[44] between the matrices A (B) and F (G) are

$$A = \frac{1}{T}\ln(F) \tag{5.3a}$$

and

$$B = A[F - I_n]^{-1}G, \tag{5.3b}$$

where I_n denotes the $n \times n$ identity matrix.

The problem of finding the matrix A from the matrix F in (5.3a) has been considered by several authors [43]-[46]. the most commonly used method is the direct-truncation method. That is, the matrix-logarithm function in $\ln(F)$ with certain convergence conditions is expanded into a certain type of infinite power series. Then, the matrix A is obtained by truncating the infinite power series. The direct-truncation method is a simple method; however, the truncation error depends heavily upon the type of power-series expansion used and the number of terms taken. Harris [46] has proposed a method which converts a matrix-logarithm function into a scalar-logarithm function via a modal-decomposition technique. The nonuniqueness of the logarithm of complex eigenvalues and the requirements of the complicated computations of eigenvectors and associated and/or repeated eigenvalues with unknown multiplicity limit the practical use of Harris' method. It seems that other methods [43]-[45] are more effective and straightforward than Harris' method [46] when the matrix of interest is defective. Recently, Sinha and Lastman [44] have proposed a fixed-point recursive algorithm for computing the matrix A from the matrix F, which involves the approximation of the Taylor series expansion of $\exp(AT)$ with $|\sigma(AT)| \leq 0.5$, where $\sigma(AT)$ denotes the eigenspectrum of the matrix AT. Moreover, Puthenpura and Sinha [45] have proposed the matrix Chebyshev method for the approximation of the shifted matrix-logarithm function $\ln(I_n + X)$ with $0 < \sigma(X) < 1$, where the matrix $X = F - I_n$. Furthermore, Shieh *et al.* [43] have proposed a direct-truncation method, a matrix continued-fraction method and a geometric-series method for determining the matrix A from the matrix F. The above three methods [43] can be summarized as follows.

1) The direct-truncation method is as follows:

$$A = \frac{1}{T}\ln(F)$$

$$= \frac{2}{T}\left[R + \frac{1}{3}R^3 + \frac{1}{5}R^5 + \ldots + \frac{1}{n}R^n + \frac{1}{n+2}R^{n+2} + \ldots\right] \tag{5.4a}$$

$$\simeq \frac{2}{T}R \tag{5.4b}$$

$$\simeq \frac{2}{T}\left[R + \frac{1}{3}R^3\right] \tag{5.4c}$$

$$\simeq \frac{2}{T}\left[R + \frac{1}{3}R^3 + \frac{1}{5}R^5\right] \tag{5.4d}$$

$$\simeq \cdots$$

where

$$R = [F - I_n][F + I_n]^{-1}. \tag{5.4f}$$

2) The matrix continued-fraction method is as follows:

$$A = \frac{1}{T}\ln(F)$$

$$= \frac{2}{T}R\left[I_n + \frac{1}{3}N + \frac{1}{5}N^2 + \frac{1}{7}N^3 + \ldots\right] \tag{5.5a}$$

$$= \frac{2}{T}R[K_1 + N[K_2 + N[K_3 + N[K_4 + N[\ldots]^{-1}]^{-1}]^{-1}]^{-1}]^{-1} \tag{5.5b}$$

$$\simeq \frac{2}{T}R[K_1]^{-1} = \frac{2}{T}R \tag{5.5c}$$

$$\simeq \frac{2}{T}R[K_1 + N[K_2]^{-1}]^{-1} = \frac{2}{T}R\left[I_n - \frac{1}{3}R^2\right]^{-1} \tag{5.5d}$$

$$\simeq \frac{2}{T}R[K_1 + N[K_2 + N[K_3]^{-1}]^{-1}]^{-1}$$

$$= \frac{2}{T}R\left[I_n - \frac{4}{15}R^2\right]\left[I_n - \frac{3}{5}R^2\right]^{-1} \tag{5.5e}$$

$$\simeq \frac{2}{T}R[K_1 + N[K_2 + N[K_3 + N[K_4]^{-1}]^{-1}]^{-1}]^{-1}$$

$$= \frac{2}{T}R\left[I_n - \frac{11}{21}R^2\right]\left[I_n - \frac{6}{7}R^2 + \frac{3}{35}R^4\right]^{-1} \tag{5.5f}$$

$$\simeq \cdots$$

where $N = R^2$, the matrix quotients K_i $(= k_i I_n)$ is the ith diagonal matrix. The ith scalar k_i can be determined from the following Routh algorithm,

$$a_{1,1} = 1,$$

$$a_{1,j} = 0 \quad \text{for} \quad j = 2, 3, \ldots,$$

$$a_{2,j} = 1/(2j-1) \quad \text{for} \quad j = 1, 2, \ldots,$$

$$a_{i,j} = a_{i-2,j+1} - k_{i-2} a_{i-1,j+1} \quad \text{for} \quad j = 1, 2, \ldots \quad \text{and} \quad i = 3, 4, \ldots,$$

and

$$k_i = a_{i,1}/a_{i+1,1} \quad \text{for} \quad i = 1, 2, \ldots.$$

3) The geometric-series method is as follows:

$$A = \frac{1}{T}\ln(F)$$

$$= \frac{2}{T}\left[R + \frac{1}{3}R^3 + \ldots + \frac{1}{n+2}R^{n+2} + \sum_{i=2}^{\infty} \frac{1}{n\left(1+\frac{2i}{n}\right)} R^{n+2i}\right] \tag{5.6a}$$

$$\simeq \frac{2}{T}\left[R + \frac{1}{3}R^3 + \ldots + \frac{1}{n+2}R^{n+2} + \sum_{i=2}^{\infty} \frac{1}{n\left(1+\frac{2}{n}\right)^i} R^{n+2i}\right] \tag{5.6b}$$

$$= \frac{2}{T}\left\{R + \frac{1}{3}R^3 + \ldots + \frac{1}{n}R^n\left[I_n - \frac{1}{\left(1+\frac{2}{n}\right)}R^2\right]^{-1}\right\}$$

$$\text{for} \quad |\sigma(R^2)| < \left(1 + \frac{2}{n}\right) \tag{5.6c}$$

$$\simeq \frac{2}{T}R\left[I_n - \frac{1}{3}R^2\right]^{-1} \quad \text{for} \quad n = 1 \tag{5.6d}$$

$$\simeq \frac{2}{T}R\left[I_n - \frac{4}{15}R^2\right]\left[I_n - \frac{3}{5}R^2\right]^{-1} \quad \text{for} \quad n = 3 \tag{5.6e}$$

$$\simeq \frac{2}{T}R\left[I_n - \frac{8}{21}R^2 - \frac{4}{105}R^4\right]\left[I_n - \frac{5}{7}R^2\right]^{-1} \quad \text{for} \quad n = 5 \tag{5.6f}$$

$$\simeq \cdots.$$

The condition for the convergence [47] of the matrix series in (5.4a), (5.5a), and (5.6a) is $\mathrm{Re}(\sigma(F)) > 0$. Note that the matrix A ($\simeq (2/T)R = (2/T)[F - I_n][F + I_n]^{-1}$) in (5.4$b$) and (5.5$c$) can be obtained by using the bilinear-transform method or the Tustin method [42].

From (5.4c), we observe that if $\mathrm{Re}(\sigma(F)) > 0$, or all eigenvalues of the matrix F lie in the right-half complex plane, then $|\sigma(R)| < 1$ and $|\sigma(N)| = |\sigma(R^2)| \ll 1$. As a result, the first few terms of the infinite power series in (5.4a), (5.5a), and (5.6a) are dominant terms. The desirable matrix A can be obtained by taking the first few dominant terms of the infinite power series in (5.4a), or can be determined by taking the first few dominant matrix quotients (K_i) of the matrix continued-fraction expansion in (5.5b). Moreover, the desirable matrix A can be obtained by taking the first few dominant terms of the infinite power series and the associated geometric-series $R^n[I_n - R^2/(1 + 2/n)]^{-1}/n$ in (5.6c). However, in general, the eigenvalues of the matrix F are not available and all eigenvalues of the matrix F are not always lying in the right-half complex plane. Therefore, the use of the above three methods is not always efficient. The purpose of this note is to develop a computational method, which uses the principal qth root of a nonsingular matrix F (or $\sqrt[q]{F}$ for $q \geq 2$) [61] together with the methods in (5.4), (5.5), and (5.6), for placing all eigenvalues of the matrix $\sqrt[q]{F}$ in the right half plane and for quickly determining the matrix A from the matrix F.

5.2 Determining Continuous-time State Equations from Discrete-time State Equations Via the Principal qth Root Method

The property of the matrix $\sqrt[q]{F}$ for $q \geq 2$ can be utilized to derive the above three approximation methods in the following.

Rewriting (5.3a) gives

$$A = \frac{1}{T}\ln(F)$$

$$= \frac{1}{T}\ln(\sqrt[q]{F})^q \tag{5.7a}$$

$$= \frac{q}{T}\ln(\sqrt[q]{F}) \tag{5.7b}$$

$$= \frac{2q}{T}\left[\hat{R} + \frac{1}{3}\hat{R}^3 + \frac{1}{5}\hat{R}^5 + \dots\right], \tag{5.7c}$$

where the matrix $\sqrt[q]{F}$ is the *principal* qth root of the matrix F, and $\hat{R} = [\sqrt[q]{F} - I_n][\sqrt[q]{F} + I_n]^{-1}$. Thus, (5.4),(5.5), and (5.6) can be rewritten as

$$A = \frac{2q}{T}\left[\hat{R} + \frac{1}{3}\hat{R}^3 + \frac{1}{5}\hat{R}^5 + \dots\right] \tag{5.8a}$$

$$\simeq \frac{2q}{T}[\hat{R}] \tag{5.8b}$$

$$\simeq \frac{2q}{T}\left[\hat{R}+\frac{1}{3}\hat{R}^3\right], \tag{5.8c}$$

$$A = \frac{2q}{T}\hat{R}[K_1+[\hat{N}[K_2+\hat{N}[K_3+\hat{N}[\dots]^{-1}]^{-1}]^{-1}]^{-1} \tag{5.9a}$$

$$\simeq \frac{2q}{T}\hat{R}[K_1]^{-1} \tag{5.9b}$$

$$\simeq \frac{2q}{T}\hat{R}[K_1+\hat{N}[K_2]^{-1}]^{-1} \tag{5.9c}$$

$$\simeq \cdots$$

where $\hat{N}=\hat{R}^2$,

and

$$A \simeq \frac{2q}{T}\left\{\hat{R}+\frac{1}{3}\hat{R}^3+\ldots+\frac{1}{n}\hat{R}^n\left[I_n-\frac{1}{\left(1+\frac{2}{n}\right)}\hat{R}^2\right]^{-1}\right\}$$

$$\text{for} \quad n=1,3,5,\cdots. \tag{5.10}$$

The condition for the convergence of the infinite power series in (5.7) becomes $\arg(\sigma(F)) \neq \pi$. The eigenvalues of the matrix F lying on the negative real axis in the complex plane are excluded in the convergence condition due to nonuniqueness of the logarithm of negative real eigenvalues. Note that the convergence region of the modified infinite power series in (5.8), (5.9), and (5.10) has been greatly enlarged from the original $\mathrm{Re}(\sigma(F)) > 0$ in (5.4), (5.5), and (5.6) to $\arg(\sigma(F)) \neq \pi$ in (5.8), (5.9), and (5.10). When $q \geq 2$, all $\sigma(\sqrt[q]{F})$ lie inside the sector angle $(-\pi/q, +\pi/q]$ of the complex plane. Therefore, $\mathrm{Re}(\sigma(\sqrt[q]{F})) > 0$, $|\sigma(\hat{R})| < 1$, $|\sigma(\hat{R}^2)| \ll 1$ and $|\sigma(\hat{R}^2)| \ll (1+2/n)$. If $q \gg 2$, then $|\sigma(\hat{R})| \ll 1$. Thus, the desired matrix A can quickly be determined by taking the first few dominant terms of the righthand side of the equations in (5.8),(5.9), and (5.10).

When the eigenvalues of the nonsingular real matrix F, which may contain negative real eigenvalues, are not available, we can employ the algorithm in (3.8) with $F := \tilde{F} = Fe^{-j\Delta\theta}$ to compute $\sqrt[q]{\tilde{F}}$. If $\sqrt[q]{\tilde{F}}\, e^{j\Delta/q}$ is a complex matrix, then there exist negative real eigenvalues. Thus, the desirable real matrix A cannot be obtained by the proposed method. On the other hand, if $\sqrt[q]{\tilde{F}}\, e^{j\Delta\theta/q}$ is a real matrix, then $\arg(\sigma(F)) \neq \pi$ and the methods in (5.8), (5.9), and (5.10) can be applied to obtain the desirable real matrix A.

5.3 Illustrative Example

Let an unstable discrete-time system matrix F be

$$F = \begin{bmatrix} 30 & -100 \\ 50 & -70 \end{bmatrix} \quad \text{and} \quad \sigma(F) = \{-20 \pm j50\}. \tag{5.11}$$

The exactly equivalent continuous-time system matrix A is

$$A = \begin{bmatrix} 5.9375 & -3.9026 \\ 1.9513 & 2.0349 \end{bmatrix}, \text{ with } \sigma(A) = \{3.9862 \pm j1.9513\}, \text{ and } \quad T = 1. \tag{5.12}$$

Since $\mathrm{Re}(\sigma(F)) < 0$, the desirable matrix A obtained from (5.4), (5.5), and (5.6) results in poor approximations of $1/T \ln(F)$. However, the desirable matrix A can be obtained from (5.8), (5.9), and (5.10) as follows. The computed $\sqrt[4]{F}$ with q=4 is

$$\sqrt[4]{F} = \begin{bmatrix} 3.6627 & -2.5394 \\ 1.2697 & 1.1233 \end{bmatrix} \quad \text{with} \quad \sigma(\sqrt[4]{F}) = \{2.3930 \pm j1.2697\}.$$

Note that $\arg(\sigma(\sqrt[4]{F})) \in (-\pi/4, +\pi/4)$, $|\sigma(\sqrt[4]{F})| > 1$, and $\mathrm{Re}[\sigma(\sqrt[4]{F})] > 0$. The approximations of $q/T \ln(\sqrt[4]{F})$ with q=4 in (5.8) obtained by taking the first N dominant terms, defined as $\hat{A}_d^{(N)}$, are

$$\hat{A}_d^{(1)} = \begin{bmatrix} 5.4115 & -3.0958 \\ 1.5479 & 2.3157 \end{bmatrix},$$

$$\hat{A}_d^{(3)} = \begin{bmatrix} 5.9466 & -3.8945 \\ 1.9472 & 2.0521 \end{bmatrix}$$

and

$$\hat{A}_d^{(6)} = \begin{bmatrix} 5.9376 & -3.9029 \\ 1.9514 & 2.0347 \end{bmatrix}.$$

The associated errors $\|A - \hat{A}_d^{(N)}\|/\|A\|$ for N=1, 3, 6 are 1.4×10^{-1}, 3.2×10^{-3}, and 5.9×10^{-5}, rspectively. Also, the approximants of $q/T\ln(\sqrt[4]{F})$ with q=4 in (5.9) obtained by taking the first N dominant quotients, defined as $\hat{A}_m^{(N)}$, are $\hat{A}_m^{(1)} = \hat{A}_d^{(1)}$,

$$\hat{A}_m^{(2)} = \begin{bmatrix} 5.9281 & -3.8459 \\ 1.9229 & 2.0823 \end{bmatrix},$$

$$\hat{A}_m^{(3)} = \begin{bmatrix} 5.9399 & -3.9021 \\ 1.9510 & 2.0378 \end{bmatrix}$$

and

$$\hat{A}_m^{(4)} = \begin{bmatrix} 5.9378 & -3.9029 \\ 1.9514 & 2.0349 \end{bmatrix}.$$

The associated errors $\|A - A_m^{(N)}\|/\|A\|$ for $N = 1, 2, 3, 4$ are 1.4×10^{-1}, 1.3×10^{-2}, 4.4×10^{-4}, and 5.3×10^{-5}, respectively. Moreover, the approximations of $q/T\ln(\sqrt[q]{F})$ with q=4 in (5.10) obtained by taking the first N dominant terms, defined as $\hat{A}_g^{(N)}$, are $\hat{A}_g^{(1)} = \hat{A}_m^{(2)}$, $\hat{A}_g^{(2)} = \hat{A}_m^{(3)}$ and

$$\hat{A}_g^{(3)} = \begin{bmatrix} 5.9380 & -3.9030 \\ 1.9515 & 2.0350 \end{bmatrix}.$$

The associated errors $\|A - \hat{A}_g^{(N)}\|/\|A\|$ for N=1, 2, 3 are 1.3×10^{-2}, 4.4×10^{-4}, and 7.8×10^{-5}, respectively. From our experience, we have observed that the direct-truncation method in (5.8) often gives satisfactory approximations when q is a large number and the matrix continued-fraction method in (5.9) converges faster than the geometric-series method in (5.10) and the direct-truncation method in (5.8) when q is a small number.

5.4 Conclusion

New computational methods, which utilize the direct-truncation method, the matrix continued-fraction method, and the geometrix-series method together with the principal qth root of a discrete-time system matrix have been presented for quick modeling of the equivalent continuous-time state equations from the discrete-time state equations. The proposed method is useful for identifying a continuous-time system based on the observation of sampled input-output data and for design of sampled-data control systems.

Rectangular and Polar Representations of a Complex Matrix

This chapter presents some new definitions of the real and imaginary parts and the associated amplitude and phase of a real or complex matrix. Computational methods, which utilize the properties of the matrix-sign function and the principal nth root of a complex matrix, are given for finding these quantities. A geometric-series method is newly developed for finding the approximation of the matrix-valued function of $\tan^{-1}(X)$, which is the principal branch of the arc tangent of the matrix X. Several illustrative examples are presented [75].

6.1 Introduction

The definitions of the real and imaginary parts of a complex number in rectangular coordinates and the associated amplitude and phase of the complex number in polar coordinates are well known, and these have been commonly used in mathematical science and control system, such as complex variable analysis applied to linear control system. However, extensions of these definitions for a complex matrix (which may be defective) and their applications have not been generally investigated by researchers.

For simplicity of notation through out this chapter, let the matrix $\mathrm{Re}(A)$ be a real matrix which contains the real part of each element of the matrix A, and the matrix $\mathrm{Im}(A)$ be a real matrix which consists of the imaginary part of each element of the matrix A. Also, let the matrix $\sqrt[n]{A}$ denote the *principal* nth root of the matrix A and the matrix $\tan^{-1}(A)$ be the *principal branch* of the arc tangent of the matrix A. The detailed definitions of the matrices $\sqrt[n]{A}$ and $\tan^{-1}(A)$ are reviewed and stated, respectively, as follows.

Definition 6.1.1 [20,21]

Let the eigenspectrum of a nonsingular matrix $A \in C^{m\times m}$ be $\sigma(A) = \{\lambda_i,\ i = 1,2,\ldots,m\}$, $\lambda_i \neq 0$, and $\arg(\lambda_i) \neq \pi$.

(1) The principal nth root of A is denoted as $\sqrt[n]{A} \in C^{m\times m}$, where n is a

positive integer and is such that $(\sqrt[n]{A})^n = A$, and for every $\delta_i = \sqrt[n]{\lambda_i} \in \sigma(\sqrt[n]{A})$, $i = 1, 2, \ldots, m$, then $\arg(\delta_i) \in (-\pi/n, \pi/n)$, where $\sqrt[n]{\lambda_i}$ is the principal nth root of λ_i.

(2) The matrix $\tan^{-1}(A)$ has the property that $\sigma(\tan^{-1}(A)) = \{\tan^{-1}(\mathrm{Im}(\lambda_i)/\mathrm{Re}(\lambda))\} = \{\arg(\lambda_i) \in (-\pi, \pi),\ i = 1, 2, \ldots, m\}$.

□

Computational algorithms [21,61] are available for finding the principal nth roots of complex matrices, and computational methods for determining the matrix $\tan^{-1}(A)$ are proposed in this chapter. Note that previous algorithms [4,18,49] for finding the nth root of a matrix may not result in the principal nth root of the matrix.

The straightforward extension of the definitions of the real and imaginary parts and the amplititude and phase from a scalar to a matrix can be described as follows.

Let $A \in C^{m \times m}$ be a nonsingular matrix with $\sigma(A) = \{\lambda_i = \alpha_i + j\beta_i \text{ for } i = 1, 2, \ldots, m\}$ where $j = \sqrt{-1}$. Then the rectangular representation of a complex matrix A would be

$$A = \mathrm{Re}(A) + j\,\mathrm{Im}(A), \tag{6.1a}$$

and the polar representation of A would be

$$A = D\exp(j\phi) \tag{6.1b}$$

or

$$A = \exp(j\phi)D. \tag{6.1c}$$

where the matrix $D = [(\mathrm{Re}(A))^2 + (\mathrm{Im}(A))^2]^{1/2}$, and the matrix ϕ is either $\tan^{-1}((\mathrm{Re}(A))^{-1}(\mathrm{Im}(A)))$ or $\tan^{-1}((\mathrm{Im}(A)(\mathrm{Re}(A))^{-1})$. If the matrices $\mathrm{Re}(A)$ and $\mathrm{Im}(A)$ in (6.1) do not contain the modal matrix of A, $\sigma(\mathrm{Re}(A)) \neq \mathrm{Re}(\sigma(A))$ and $\sigma(\mathrm{Im}(A)) \neq \mathrm{Im}(\sigma(A))$; then, in general, the representation in (6.1b) or (6.1c) is not the polar representation of A because $|\sigma(D)| \neq |\sigma(A)|$ and $\sigma(\phi) \neq \arg(\sigma(A))$. Another important consequence would be $D\exp(j\phi) \neq \exp(j\phi)D$. In other words, the commutative property of matrix multiplication in the polar representation of a matrix in (6.1) is not preserved because the matrices $\mathrm{Re}(A)$, $\mathrm{Im}(A)$, D and ϕ do not contain the modal matrix of A. As a result, it is difficult to generalize a scalar-valued function to a matrix-valued function, and to develop complex variable approaches to the analysis of linear multivariable control systems.

Another popular rectangular representation [52] of the matrix A is

$$A = (A + A^*)/2 + j[(A - A^*)/2j], \tag{6.2}$$

where the asterisk superscript (for Hermitian) designates the conjugate transpose. If the matrix A is not a normal matrix [48,51], then the real part, $(A + A^*)/2$, and the imaginary part, $(A - A^*)/2j$, do not contain the modal matrix of A, and they do not commute. Therefore, the representation in (6.2) is not suitable to be used for defining the amplitude and phase of A.

A formal polar representation [51] of a nonsingular matrix A is

$$A = HU, \tag{6.3}$$

where the matrix H is a square root of the symmetric matrix (AA^*), and the matrix U is a unitary matrix having $U = H^{-1}A = \exp(j\phi)$ where $\phi = -j\ln(U)$. If the matrix A is not a normal matrix, then the matrices H and U do not contain the modal matrix of A. Also, $HU \neq UH$, $|\sigma(H)| \neq |\sigma(A)|$ and $\sigma(\phi) \neq \arg(\sigma(A))$. As a result, the representation in (6.3) is not suitable to be used for defining the real and imaginary parts of A. Hence the application of the polar representation in (6.3) to complex variable analysis and computational aspects is limited. For example, if the matrix A is a defective matrix, then $|\sigma(A)| \neq |\sigma(H)|$. As a result, the matrix H cannot be utilized to normalize the amplitude of the matrix A for reducing the computational error of A^k where k is a large positive integer. The need for the computation of A^k and its applications can be found in [9,26,36].

This chapter presents some new definitions of the real and imaginary parts and the amplitude and phase of a complex matrix. Procedures are given for computing these matrices, and several illustrative examples are presented. The aims of this chapter are primarily to develop theoretical tools rather than highly efficient computational algorithms.

This chapter is organized as follows: In Section 6.2, we define two different rectangular and polar representations of a matrix, and give illustrative examples. In Section 6.3, we develop computational procedures for finding the projected imaginary part (A_I), the projected real part (A_R), the amplitude (A_ρ) and the phase (A_θ) of the matrix A. An illustrative example is shown in Section 6.4, and the results are summarized in Section 6.5.

6.2 Rectangular and Polar Representations of a Matrix

Let us first define the rectangular and polar representations of a complex matrix, which may be a defective matrix [54], in the following way.

Definition 6.2.1

Consider a matrix $A \in C^{m\times m}$ with eigenspectrum and associated modal matrix,

$$\sigma(A) = \left\{ \lambda_i = \alpha_i + j\beta_i, \text{ for } i = 1,2,\ldots,k \right.$$
$$\left. \text{with multiplicity } m_i, \text{ and } \sum_{j=1}^{k} m_j = m,\ \alpha_i \neq 0 \right\}$$

and $M \in C^{m\times m}$, respectively. Then the complex matrix A, which may be a defective matrix with $|\arg(\sigma(A))| \neq \pi/4$ or $3\pi/4$, can be described in the *rectangular coordinates* as

$$A = MJM^{-1} = M[\mathrm{Re}(J)]M^{-1} + jM[\mathrm{Im}(J)]M^{-1} = \bar{A}_R + j\bar{A}_I, \qquad (6.4a)$$

where the matrix J is of Jordan form, the matrices $\bar{A}_R(\triangleq M[\mathrm{Re}(J)]M^{-1})$ and $\bar{A}_I(\triangleq M[\mathrm{Im}(J)]M^{-1})$ are the *real* and *imaginary* parts of the matrix A, respectively.

The *polar representation* of the complex matrix A is

$$A = \bar{A}_\rho \angle \bar{A}_\theta = \bar{A}_\rho \exp(j\bar{A}_\theta) = \exp(j\bar{A}_\theta)\bar{A}_\rho, \qquad (6.4b)$$

where the matrix,

$$A_\rho = (\bar{A}_R^2 + \bar{A}_I^2)^{1/2}, \tag{6.4c}$$

is defined as the *amplitude* of the matrix A, and the matrix,

$$\bar{A}_\theta = \tan^{-1}(\bar{A}_I \bar{A}_R^{-1}) = \tan^{-1}(\bar{A}_R^{-1} \bar{A}_I), \tag{6.4d}$$

is defined as the *phase* of the matrix A.

The *projected* real and imaginary parts of the matrix A can be computed in polar coordinates as

$$\bar{A}_R = \bar{A}_\rho \cos(\bar{A}_\theta) = \cos(\bar{A}_\theta)\bar{A}_\rho \tag{6.4e}$$

and

$$\bar{A}_I = \bar{A}_\rho \sin(\bar{A}_\theta) = \sin(\bar{A}_\theta)\bar{A}_\rho. \tag{6.4f}$$

□

Note that both matrices $\bar{A}_R$ and $\bar{A}_I$ contain the same modal matrix of A; therefore, the matrices $\bar{A}_R$ and $\bar{A}_I$ commute, and the matrices $\bar{A}_\rho$ and $\exp(j\bar{A}_\theta)$ also commute. When the matrix A is a defective matrix in which the nontrivial elements on the super-diagonal line of the Jordan matrix J may be complex numbers, the matrices $\mathrm{Re}(J)$ and $\mathrm{Im}(J)$ may not be diagonal matrices.

When the eigenvalues, associated eigenvectors and the structure of the Jordan matrix J of a defective matrix A in (6.4) are known, the amplitude matrix $\bar{A}_\rho$ in (6.4c) can be determined by finding the principal square root of the matrix, $\bar{A}_R^2 + \bar{A}_I^2$, via the algorithms developed in Chapter 3. However, the determination of the principal branch of $\tan^{-1}(\bar{A}_R^{-1}\bar{A}_I)$ for the phase matrix $\bar{A}_\theta$ in (6.4d) is rather more complicated than that of $\bar{A}_\rho$. An illustrative example is shown as follows.

Example 6.2.1

Consider a defective complex matrix A,

$$A = \begin{bmatrix} \alpha + j\beta & 1 \\ 0 & \alpha + j\beta \end{bmatrix} \quad \text{with} \quad \alpha \neq 0.$$

Following Definition 6.2.1 with $A = J$ and $M = I_2$ where I_2 is an 2×2 identity matrix, we obtain

$$\bar{A}_R = \mathrm{Re}(A) = \begin{bmatrix} \alpha & 1 \\ 0 & \alpha \end{bmatrix},$$

$$\bar{A}_I = \mathrm{Im}(A) = \begin{bmatrix} \beta & 0 \\ 0 & \beta \end{bmatrix},$$

$$A_\rho = (\bar{A}_R^2 + \bar{A}_I^2)^{1/2} = \begin{bmatrix} \sqrt{\alpha^2+\beta^2} & \frac{\alpha}{\sqrt{\alpha^2+\beta^2}} \\ 0 & \sqrt{\alpha^2+\beta^2} \end{bmatrix},$$

and

$$\bar{A}_\theta = \tan^{-1}(\bar{A}_R^{-1}\bar{A}_I) = \tan^{-1}(X),$$

where

$$X = \bar{A}_R^{-1}\bar{A}_I = \begin{bmatrix} \frac{\beta}{\alpha} & \frac{-\beta}{\alpha^2} \\ 0 & \frac{\beta}{\alpha} \end{bmatrix}.$$

Since the matrix X has Jordan canonical form, the matrix $\bar{A}_\theta$ can be determined by using the standard formula [51] (Gantmacher 1959, p.98) as follows,

$$\bar{A}_\theta = \tan^{-1}(X) = \begin{bmatrix} \tan^{-1}\left(\frac{\beta}{\alpha}\right) & z \\ 0 & \tan^{-1}\left(\frac{\beta}{\alpha}\right) \end{bmatrix} \quad \text{for } |\sigma(X)| = \left|\frac{\beta}{\alpha}\right| \leq 1,$$

where

$$z = -\frac{\beta}{\alpha}\left[1 - \left(\frac{\beta}{\alpha}\right)^2 + \left(\frac{\beta}{\alpha}\right)^4 - \left(\frac{\beta}{\alpha}\right)^6 + \ldots\right].$$

Note that the evaluation of $\tan^{-1}(\beta/\alpha)$ depends upon the signs of α and β and the determination of the infinite series z on the magnitude of $|\beta/\alpha|$. For example, when $\alpha = |\beta|$ or $|\arg(\sigma(A))| = \pi/4$, the infinite series z does not converge and becomes null or $-\beta/\alpha^2$. Hence, the matrix $\bar{A}_\theta$ is not the desired phase matrix. Thus, we conclude that when the matrix A is a defective matrix with any $|\arg(\sigma(A))| = \pi/4$ or $3\pi/4$ (i.e., $|\sigma(X)| = 1$), and/or $|\sigma(X)| > 1$, the direct use of the standard formula [51] (Gantmacher 1959, p. 98) for determining the above matrix-valued function of $\tan^{-1}(X)$ will not result in the desired phase matrix $\bar{A}_\theta$. A computational method will be developed in Section 6.3 to overcome the above difficulty and for determining the desired $\bar{A}_\theta$.

Let us define an additional notation, which will be used throughout this chapter.

Definition 6.2.2

Let the matrix $J \in C^{m\times m}$ be the Jordan matrix of a defective matrix $A \in C^{m\times m}$, and let the diagonal matrix $\Lambda \in C^{m\times m}$ contain only the diagonal elements of J, and $\sigma(A) = \sigma(J) = \sigma(\Lambda)$. Then, the matrix J_1 is defined as $J - \Lambda$, which is a matrix containing only the elements on the super-diagonal line of J. The nontrivial elements in J_1 may be complex numbers. □

Definition 6.2.1 cannot be utilized for finding the rectangular and polar representations of the matrix A when it is a defective matrix with $|\arg(A)| = \pi/4$ or

$3\pi/4$. To relax the constraint in Definition 6.2.1 and to develop a computational method for finding the rectangular and polar representations of the matrix A without actuallly knowing the eigenvalues, eigenvectors and the structure of the Jordan form, we define *alternative* rectangular and polar representations of the matrix as follows.

Definition 6.2.3

Let $A \in C^{m\times m}$ be a matrix, and let its eigenspectrum and associated modal matrix be $\sigma(A) = \{\lambda_i = \alpha_i + j\beta_i \text{ for } i = 1, 2, \ldots, k \text{ with multiplicity } m_i, \text{ and } \sum_{j=1}^{k} m_j = m,\ \alpha_i \neq 0\}$, and $M \in C^{m\times m}$, respectively. Then, the matrix A can be represented as

$$A = MJM^{-1} = M[\Lambda + J_1]M^{-1}$$

$$= (A_{Rd} + jA_{Id}) + A_1 = A_{\rho d}\exp(jA_{\theta d}) + A_1, \tag{6.5a}$$

where $A_{Rd} \triangleq M[\mathrm{Re}(\Lambda)]M^{-1}$, $A_1 = MJ_1M^{-1}$, $A_{Id} \triangleq M[\mathrm{Im}(\Lambda)]M^{-1}$, $A_{\rho d} \triangleq (A_{Rd}^2 + A_{Id}^2)^{1/2}$ and $A_{\theta d} \triangleq \tan^{-1}(A_{Rd}^{-1}A_{Id}) \triangleq \tan^{-1}(A_{Id}A_{Rd}^{-1})$, respectively. The polar and rectangular representations of the matrix A can be defined as follows,

$$A = [A_{\rho d} + A_1\exp(-jA_{\theta d})]\exp(jA_{\theta d})$$

$$= A_\rho\exp(jA_\theta) = \exp(jA_\theta)A_\rho = A_R + jA_I, \tag{6.5b}$$

where

$$A_\rho \triangleq A_{\rho d} + A_1\exp(-jA_\theta) = (A_R^2 + A_I^2)^{1/2},$$

$$A_\theta = A_{\theta d} = \tan^{-1}(A_{Id}A_{Rd}^{-1}) = \tan^{-1}(A_{Rd}^{-1}A_{Id})$$

$$= \tan^{-1}(A_IA_R^{-1}) = \tan^{-1}(A_R^{-1}A_I),$$

$$A_R \triangleq A_\rho\cos(A_\theta) = \cos(A_\theta)A_\rho,$$

and

$$A_I \triangleq A_\rho\sin(A_\theta) = \sin(A_\theta)A_\rho\ .$$

□

The matrices A_ρ, A_θ, A_R and A_I are defined as the *amplitude*, *phase*, *real part* and *imaginary parts* of the matrix A, respectively. Note that in general these matrices are different from those defined earlier, indicated by an overbar.

Note also that any additional lower subscript d of a matrix shown in the above definition denotes that the matrix is a nondefective matrix; also that $A_R \neq A_{Rd} + A_1$ and $A_I \neq A_{Id}$. A simple example follows.

Example 6.2.2

Consider a defective complex matrix A with $\arg(\sigma(A)) = \pi/4$,

$$A = \begin{bmatrix} 1+j1 & 1 \\ 0 & 1+j1 \end{bmatrix}.$$

Following Definition 6.2.3 with $A = J$, $M = I_2$, $\Lambda = \text{diag}(1+j1, 1+j1)$ and $J_1 = J - \Lambda$, we obtain

$$A_{Rd} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_{Id} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_1 = J_1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$A_{\rho d} = (A_{Rd}^2 + A_{Id}^2)^{1/2} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

The desired phase matrix A_θ and amplitude matrix A_ρ are

$$A_\theta = \tan^{-1}(A_{Rd}^{-1} A_{Id}) = \text{diag}[\tan^{-1}(1), \tan^{-1}(1)] = \text{diag}[\pi/4, \pi/4]$$

and

$$A_\rho = A_{\rho d} + A_1 \exp(-jA_\theta) = A_{\rho d} + A_1 \,\text{diag}\left[\exp\left(-j\frac{\pi}{4}\right), \exp\left(-j\frac{\pi}{4}\right)\right]$$

$$= \begin{bmatrix} \sqrt{2} & \exp\left(-j\frac{\pi}{4}\right) \\ 0 & \sqrt{2} \end{bmatrix}.$$

Also, the desired real matrix A_R and imaginary matrix A_I are

$$A_R = A_\rho \cos(A_\theta) = A_\rho \,\text{diag}\left[\cos\left(\frac{\pi}{4}\right), \cos\left(\frac{\pi}{4}\right)\right] = \begin{bmatrix} 1 & (1-j1)/2 \\ 0 & 1 \end{bmatrix}$$

and

$$A_I = A_\rho \sin(A_\theta) = A_\rho \,\text{diag}\left[\sin\left(\frac{\pi}{4}\right), \sin\left(\frac{\pi}{4}\right)\right] = \begin{bmatrix} 1 & (1-j1)/2 \\ 0 & 1 \end{bmatrix}.$$

Note that Definition 6.2.1 cannot be applied directly to determine the phase of A in Example 6.2.2 because $\arg(\sigma(A)) = \pi/4$. Also, note that the matrices A_ρ and $\exp(jA_\theta)$ commute, $\sigma(A_\rho) = |\sigma(A)|$, $\sigma(A_\theta) = \arg(\sigma(A))$, $\sigma(A_R) = \text{Re}(\sigma(A))$,

$\sigma(A_I) = \mathrm{Im}(\sigma(A))$, and the matrices A_ρ, A_θ, A_R and A_I contain the modal matrix of A.

6.3 Computational Method for Determining the Amplitude and Phase of a Matrix

The determination of the matrices A_ρ and A_θ can be accomplished by finding the eigenvalues and eigenvectors of the matrix A via high quality algorithms such as the QR algorithm [54] and LINPACK [50], etc., and by using the definitions shown in Section 6.2. In this section, a computational method is developed for finding the matrices A_ρ and A_θ without directly involving the eigenvalues and eigenvectors of A and the preknowledge of the structure of the Jordan matrix of A. The matrix-sign function [9,26] of A, which preserves the eigenvectors of a complex matrix (which may be defective), is used as a basis for the development. The (scalar-) sign function of a complex variable λ with $\mathrm{Re}(\lambda) \neq 0$ is defined by

$$\mathrm{Sign}(\lambda) = \lambda/\sqrt{\lambda^2} = \begin{cases} +1 & \text{when } \mathrm{Re}(\lambda) > 0 \\ -1 & \text{when } \mathrm{Re}(\lambda) < 0 \end{cases}, \tag{6.6}$$

where $\sqrt{\lambda^2}$ is the principal value of the square root of λ^2.

Following the definition in (6.6), the *matrix-sign function* of the matrix A is defined as $\mathrm{Sign}(A) = A(\sqrt[2]{A^2})^{-1}$ with $\mathrm{Re}(\sigma(A)) \neq 0$, where the matrix $\sqrt[2]{A^2}$ is the principal square root of a matrix A^2. The computational algorithm for finding $\mathrm{Sign}(A)$ can be found in Chapter 4.

It is well-known that the imaginary parts of the eigenvalues of A and the associated eigenvectors of A are invariant under the horizontal translation of A on the real axis with a real value γ, that is $\mathrm{Im}(\sigma(A - \gamma I_m))$, and $A - \gamma I_m = M[(\mathrm{Re}(J) - \gamma I_m) + j\mathrm{Im}(J)]M^{-1}$ where I_m is an $m \times m$ indentity matrix. When the real value γ is selected so that $\mathrm{Re}(\sigma(A - \gamma I_m)) = 0$, then the shifted matrix $A - \gamma I_m (\triangleq \hat{A}_I)$ contains only imaginary eigenvalues $(j\beta_i, i = 1, 2, \ldots, m)$ and A_1. Hence, the desired matrix A_{Rd} in (6.5) becomes $A - \hat{A}_I$. The matrix-sign function of A is utilized to determine the matrix $\hat{A}_I$ and the desired matrix A_{Rd} in (6.5) as $A - \hat{A}_I$. In order to get the desired matrix A_{Id} in (6.5), we multiply the matrix A by j and repeat the above procedure to compute a new matrix $\hat{A}_I$.

Thus, we have

$$A_{Id} = \hat{A}_I - jA, \tag{6.7a}$$

and the desired matrix A_1 in (6.5) becomes

$$A_1 = A - (A_{Rd} + jA_{Id}). \tag{6.7b}$$

An alternative representation of A_{Id} is

$$A_{Id} = -j\hat{A}_{Id}, \tag{6.7c}$$

where $\hat{A}_{Id} = jA_{Id}$.

Hence, the original matrix A can be represented as

$$A = (A_{Rd} + A_1) + jA_{Id}, \tag{6.7d}$$

and we have obtained the desired matrices A_{Rd}, A_1 and A_{Id} for use in Definition 6.2.3.

The computaional algorithm for finding $\hat{A}_I (= A - \gamma I_m)$ is listed as follows.

Algorithm 6.3.1

Given that A is a complex matrix of dimension $m \times m$ with eigenvalues $\lambda_i = \alpha_i + \beta_i$, $i = 1, 2, \ldots, m$, $|\lambda_i| \neq 0$; γ is a small positive value, $\gamma \leq \bar{\epsilon}$ where $\bar{\epsilon}$ is an acceptable error tolerance, find $\hat{A}_I$.

Algorithm:

$$A_0 = A - \{\mathrm{Re}[\mathrm{trace}(A)]/n\} \cdot I_m,$$

$$\gamma^+ = \mathrm{Re}\left\{ \frac{\mathrm{trace}\left[\dfrac{I_m + \mathrm{Sign}(A_k - \gamma I_m)}{2} \cdot A_k\right]}{\mathrm{trace}\left[\dfrac{I_m + \mathrm{Sign}(A_k - \gamma I_m)}{2}\right]} \right\},$$

$$\gamma^- = \mathrm{Re}\left\{ \frac{\mathrm{trace}\left[\dfrac{I_m - \mathrm{Sign}(A_k + \gamma I_m)}{2} \cdot A_k\right]}{\mathrm{trace}\left[\dfrac{I_m - \mathrm{Sign}(A_k + \gamma I_m)}{2}\right]} \right\},$$

$$A_{k+1} = A_k - \gamma^+ \left[\frac{I_m + \mathrm{Sign}(A_k - \gamma I_m)}{2}\right] - \gamma^- \left[\frac{I_m - \mathrm{Sign}(A_k + \gamma I_m)}{2}\right]$$

for $k = 0, 1, 2, \cdots,$

until

$$\mathrm{trace}\left[\frac{I_m + \mathrm{Sign}(A_k - \gamma I_m)}{2}\right] + \mathrm{trace}\left[\frac{I_m - \mathrm{Sign}(A_k + \gamma I_m)}{2}\right] = 0,$$

where γ^+, γ^- and γ are scalars chosen so that

$$0 < \gamma < \min\{|\,\mathrm{Re}(\lambda)|\;|\;\lambda \in \sigma(A_k), \mathrm{Re}(\lambda) \neq 0\}.$$

γ^+ is the arithmetic mean of $\{\mathrm{Re}(\lambda)\,|\,\lambda \in \sigma(A_k), \mathrm{Re}(\lambda) > 0\}$ and γ^- is the arithmetic mean of $\{\mathrm{Re}(\lambda)\,|\,\lambda \in \sigma(A_k), \mathrm{Re}(\lambda) < 0\}$.

The amplitude of A can be represented in terms of A_{Rd}, A_{Id} and A_1 as

$$A_\rho = (A_{Rd}^2 + A_{I_d}^2)^{1/2} + A_1 \exp(-jA_\theta) = A_{\rho d} + A_1 \exp(-jA_\theta). \tag{6.8}$$

Since all eigenvalues of $A_{\rho d}$ are positive real values, we can compute $A_{\rho d}$ via either the algorithm in [61] or the Newton-Raphson algorithm due to [54].

As we have discussed in Example 6.2.1, the determination of the phase of A with and/or without prior knowledge of the eigenvalues of A is not a simple matter. Based on the property of the principal nth root of a matrix shown in Definition 6.1.1, we propose a new method for finding the approximation of A_θ as follows.

Rewriting the matrix-valued function in (6.5) gives

$$A_\theta = \tan^{-1}(A_{Rd}^{-1}A_{Id})$$

$$= \tan^{-1}(A_{Rd}^{-1}(-j\hat{A}_{Id})) = \tan^{-1}(-jA_{Rd}^{-1}\hat{A}_{Id}) = \tan^{-1}(-jX), \tag{6.9}$$

where $X = A_{Rd}^{-1}\hat{A}_{Id}$, and $\sigma(X) = \{\bar{\lambda}_i = 0 + j(\beta_i/\alpha_i),\ \text{for } i = 1,2,\ldots,m\}$. The matrix-valued function A_θ in (6.8) can be represented by an infinte series as

$$A_\theta = \tan^{-1}(-jX) = -j\left[X + \frac{X^3}{3} + \frac{X^5}{5} + \frac{X^7}{7} + \cdots\right]$$

$$\text{as} \quad |\sigma(X)| \leq 1, \quad \text{and} \quad \text{Re}(\sigma(A)) \geq 0. \tag{6.10a}$$

Thus, if $|\sigma(X)| \ll 1$ and $\text{Re}(\sigma(A)) \geq 0$, the approximations of A_θ can be obtained by taking only the first several terms as

$$A_\theta \simeq jX \simeq -j\left[X + \frac{X^3}{3}\right] \simeq \cdots. \tag{6.10b}$$

Since not *all* $|\sigma(X)|$ are less than or equal to unity and/or $\text{Re}(\sigma(A)) \geq 0$, it is difficult to obtain A_θ via the direct-truncation method. To overcome the above difficulty and to guarantee the convergence of the infinite series in (6.10*a*), we determine the principal 4th root of A (denoted by $A^{(4)}$) and the associated matrices $A_{Rd}^{(4)}$, $A_{Id}^{(4)}$ and $A_1^{(4)}$ as follows,

$$A^{(4)} = \left(A_{Rd}^{(4)} + A_1^{(4)}\right) + jA_{Id}^{(4)}, \tag{6.11a}$$

where the matrices $A_{Rd}^{(4)}$ and $A_{Id}^{(4)}$ can be obtained by Algorithm 6.3.1 having the matrix A replaced by $A^{(4)}$. Thus, the matrix $A_{\rho d}^{(4)}$ becomes

$$A_{\rho d}^{(4)} = \left[\left(A_{Rd}^{(4)}\right)^2 + \left(A_{Id}^{(4)}\right)^2\right]^{1/2}, \tag{6.11b}$$

and the phase of $A^{(4)}$, denoted by $A_\theta^{(4)}$, can be fully represented by an infinite series as

$$A_\theta^{(4)} = \tan^{-1}\left(-j\left(A_{Rd}^{(4)}\right)^{-1}\hat{A}_{Id}^{(4)}\right) = \tan^{-1}(-j\hat{X}) = -j\hat{A}_\theta^{(4)}$$

$$= -j\left[\hat{X} + \frac{\hat{X}^3}{3} + \frac{\hat{X}^5}{5} + \frac{\hat{X}^7}{7} + \cdots\right], \quad |\sigma(\hat{X})| \leq 1, \text{ and } \mathrm{Re}(\sigma(A^{(4)})) \geq 0, \tag{6.12a}$$

where $\hat{X} \triangleq \left(A_{Rd}^{(4)}\right)^{-1}\hat{A}_{Id}^{(4)}$. From Definition 6.1.1, we see that all eigenvalues of $A^{(4)}$, or $\sigma\left(A^{(4)}\right)$, lie inside the sector in the λ-plane with sector angles $(\pi/4, -\pi/4)$; therefore, the convergence conditions in (6.12a) are always satisfied. Hence, the approximations of A_θ can be obtained by truncating the infinite series in (6.12a) as

$$A_\theta^{(4)} \simeq -j\hat{X} \simeq -j\left[\hat{X} + \frac{\hat{X}^3}{3}\right] \simeq \cdots. \tag{6.12b}$$

The desired matrices A_ρ and A_θ in (6.5) can be obtained as

$$A_\rho = \left(A_{\rho d}^{(4)}\right)^4 + A_1 \exp(-jA_\theta)$$

$$= \left[\left(A_{Rd}^{(4)}\right)^2 + \left(A_{Id}^{(4)}\right)^2\right]^2 + A_1 \exp(-jA_\theta) \tag{6.13a}$$

and

$$A_\theta = 4A_\theta^{(4)}, \tag{6.13b}$$

where $A_1 = A - \left(A_{Rd}^{(4)} + jA_{Id}^{(4)}\right)^4$.

It is well-known that the Taylor series for $\tan^{-1}(x)$ converges too slowly to be of much use in numerical computation when the argument x is close to unity. For example, calculating $\tan^{-1}(0.9)$ to five significant digits requires the first 29 terms of the Taylor approximation. A more sophisticated approximation of A_θ can be obtained via the following geometric-series method.

Rewriting (6.12a) yields

$$A_\theta^{(4)} = -j\left[\hat{X} + \frac{1}{3}\hat{X}^3 + \cdots + \frac{1}{n}\hat{X}^n + \frac{1}{n+2}\hat{X}^{n+2}\frac{1}{n+4}\hat{X}^{n+4}\frac{1}{n+6}\hat{X}^{n+6} + \cdots\right]$$

$$= -j\left[\hat{X} + \frac{1}{3}\hat{X}^3 + \cdots + \frac{1}{n}\hat{X}^n + \frac{1}{n+2}\hat{X}^{n+2} + \sum_{i=2}^{\infty}\frac{1}{n\left(1+\frac{2i}{n}\right)}\hat{X}^{n+2i}\right]. \tag{6.14a}$$

The weighting factor of the term $\hat{X}^{n+2i}$ in the infinite series in (6.14a) can be approximated by the following approximation,

$$\frac{1}{n\left(1+\frac{2i}{n}\right)} \simeq \frac{1}{n\left(1+\frac{2}{n}\right)^{i}}. \tag{6.14b}$$

Thus, we have

$$\hat{A}_{\theta}^{(4)} \simeq \hat{X} + \frac{1}{3}\hat{X}^3 + \cdots + \frac{1}{n}\hat{X}^n + \frac{1}{n+2}\hat{X}^{n+2} + \sum_{i=2}^{\infty} \frac{1}{n\left(1+\frac{2}{n}\right)^{i}}\hat{X}^{n+2i}$$

$$= \hat{X} + \frac{1}{3}\hat{X}^3 + \cdots + \frac{1}{n}\hat{X}^n \left[I_m + \frac{1}{\left(1+\frac{2}{n}\right)}\hat{X}^2 \frac{1}{\left(1+\frac{2}{n}\right)^{2}}\hat{X}^4 + \cdots \right]$$

$$= \hat{X} + \frac{1}{3}\hat{X}^3 + \cdots + \frac{1}{n}\hat{X}^n \left[I_m - \frac{1}{\left(1+\frac{2}{n}\right)}\hat{X}^2 \right]^{-1}$$

$$\text{for} \quad |\sigma(\hat{X}^2)| < \left(1+\frac{2}{n}\right). \tag{6.14c}$$

Note that $[I_m - \hat{X}^2/(1+2/n)]^{-1}$ is a geometric-series and it converges when $|\sigma(\hat{X}^2)| < (1+2/n)$, where $\hat{X} = \left(A_{Rd}^{(4)}\right)^{-1} \hat{A}_{Id}^{(4)}$.

Some approximations of $\hat{A}_{\theta}^{(4)}$ in (6.14c) for $n = 1, 3, 5$ are listed below,

$$\hat{A}_{\theta}^{(4)} \simeq \hat{X}(I_m - \tfrac{1}{3}\hat{X}^2)^{-1} \simeq \hat{X}(I_m - \tfrac{4}{15}\hat{X}^2)(I_m - \tfrac{3}{5}\hat{X}^2)^{-1} \tag{6.14d}$$

$$\simeq \hat{X}(I_m - \tfrac{8}{21}\hat{X}^2 - \tfrac{4}{105}\hat{X}^4)(I_m - \tfrac{5}{7}\hat{X}^2)^{-1}. \tag{6.14e}$$

Smaller values of $|\sigma(\hat{X})|$ result in better approximations of $\hat{A}_{\theta}^{(4)}$. Since $|\sigma(\hat{X})| \leq 1$, the maximum error will occur when $|\sigma(\hat{X})| = 1$. Let $\hat{X}$ in (6.14e) be unity, then $4\hat{A}_{\theta}^{(4)} \simeq j3.13333\,(\text{rad})$. Note that the exact solution of $\hat{A}_{\theta}$ is $j\pi = j3.14159\,(\text{rad})$. If we compute the principal square root of $A^{(4)}$ and use (6.14e), then we obtain $8\hat{A}_{\theta}^{(8)} = j3.14157\,(\text{rad})$ which is close to the exact solution.

The Taylor series of $\cos\left(A_\theta^{(4)}\right)$, $\sin\left(A_\theta^{(4)}\right)$ and $\exp\left(jA_\theta^{(4)}\right)$, which often give good approximations, are listed as follows,

$$\cos\left(A_\theta^{(4)}\right) = I_m - \frac{\left[A_\theta^{(4)}\right]^2}{2!} + \frac{\left[A_\theta^{(4)}\right]^4}{4!} - \frac{\left[A_\theta^{(4)}\right]^6}{6!} + \cdots, \tag{6.15a}$$

$$\sin\left(A_\theta^{(4)}\right) = A_\theta^{(4)} - \frac{\left[A_\theta^{(4)}\right]^3}{3!} + \frac{\left[A_\theta^{(4)}\right]^5}{5!} - \frac{\left[A_\theta^{(4)}\right]^7}{7!} + \cdots, \tag{6.15b}$$

and

$$\exp\left(A_\theta^{(4)}\right) = \cos\left(A_\theta^{(4)}\right) + j\sin\left(A_\theta^{(4)}\right). \tag{6.15c}$$

To determine $\cos(A_\theta)$ and $\sin(A_\theta)$ from $\cos\left(A_\theta^{(4)}\right)$ and $\sin\left(A_\theta^{(4)}\right)$, we apply the following formulas,

$$\cos(n\phi) = \frac{1}{2}\left\{(2\cos(\phi))^n - \frac{n}{1}(2\cos(\phi))^{n-2} + \frac{n}{2}\binom{n-3}{1}(2\cos(\phi))^{n-4} - \frac{n}{3}\binom{n-4}{2}(2\cos(\phi))^{n-6} + \cdots\right\} \tag{6.16a}$$

and

$$\sin(n\phi) = \sin(\phi)\left\{(2\cos(\phi))^{n-1} - \binom{n-2}{1}(2\cos(\phi))^{n-3} + \binom{n-3}{2}(2\cos(\phi))^{n-5} - \cdots\right\} \tag{6.16b}$$

where $\phi = A_\theta^{(n)}$.

If the first four dominant terms are used to approximate the infinite series in (6.15a) and (6.15b), we obtain

$$\cos(A_\theta) \simeq 8\left(\cos\left(A_\theta^{(4)}\right)\right)^4 - 8\left(\cos\left(A_\theta^{(4)}\right)\right)^2 + I_m, \tag{6.16c}$$

$$\sin(A_\theta) \simeq 4\sin\left(A_\theta^{(4)}\right)\cos\left(A_\theta^{(4)}\right) - 8\left(\sin\left(A_\theta^{(4)}\right)\right)^3\cos\left(A_\theta^{(4)}\right). \tag{6.16d}$$

Hence, we have

$$\exp(jA_\theta) = \cos(A_\theta) + j\sin(A_\theta). \tag{6.16e}$$

The maximum error will occur when $\theta = \pi$. In this case, we shall obtain the approximations of $\cos(\theta)$, $\sin(\theta)$ and $\exp(j\theta)$ by the above procedure as $\cos(\pi) \simeq 1$, $\sin(\pi) = 0 \simeq 4.6 \times 10^{-9}$ and $\exp(j\pi) = -1 \simeq -1 + j4.6 \times 10^{-9}$. The approximations are quite satisfactory. The procedures for determining the matrices A_ρ, A_θ, A_R and A_I are summarized in the following algorithm.

Algorithm 6.3.2

Step 1.

Compute the principal qth ($q \geq 4$) root of A, defined as $A^{(q)}$, via the algorithms in (3.9).

Step 2.

Find $\hat{A}_I^{(q)}$ via Algorithm 6.3.1 having the matrix A replaced by $A^{(q)}$ to obtain $A_{Rd}^{(q)}$ and $A_{Id}^{(q)}$ via procedures derived in Section 6.3.

Step 3.

Determine A_1, A_ρ and A_θ as follows,

$$A_\rho = A_{\rho d} + A_1 \exp(-jA_\theta) = \left[\left(A_{Rd}^{(q)}\right)^2 + \left(A_{Id}^{(q)}\right)^2\right]^{q/2} + A_1 \exp(-jA_\theta),$$

where

$$A_1 = A - \left[A_{Rd}^{(q)} + jA_{Id}^{(q)}\right]^q,$$

$$A_\theta = qA_\theta^{(q)},$$

$$A_\theta^{(q)} = -j\left[\hat{X} + \frac{\hat{X}^3}{3} + \frac{\hat{X}^5}{5} + \frac{\hat{X}^7}{7} + \cdots\right]$$

$$\text{for} \quad |\sigma(\hat{X})| \leq 1, \text{ and } \mathrm{Re}(\sigma(A^{(q)})) \geq 0$$

$$\simeq -j\hat{X}(I_m - \tfrac{1}{3}\hat{X}^2)^{-1} \simeq -j\hat{X}(I_m - \tfrac{4}{15}\hat{X}^2)(I_m - \tfrac{3}{5}\hat{X}^2)^{-1}$$

$$\simeq -j\hat{X}(I_m - \tfrac{8}{21}\hat{X}^2 - \tfrac{4}{105}\hat{X}^4)(I_m - \tfrac{5}{7}\hat{X}^2)^{-1},$$

where

$$\hat{X} = (A_{Rd}^{(q)})^{-1}(jA_{Id}^{(q)}),$$

$$\exp(-jA_\theta) = \cos(A_\theta) - j\sin(A_\theta).$$

The $\cos(A_\theta)$ and $\sin(A_\theta)$ can be obtained by using the approximations of the infinite series in (6.15) and (6.16).

Step 4.

Determine A_R and A_I as

$$A_R = A\rho\cos(A_\theta),$$
$$A_I = A_\rho\sin(A_\theta).$$

■

6.4 Illustrative Example

Consider a defective real matrix A as

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -4 & 8 & -8 & 4 \end{bmatrix} = MJM^{-1},$$

where

$$J = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 & 1 & 0 \\ \lambda_1 & 1 & \lambda_2 & 1 \\ \lambda_1^2 & 2\lambda_1 & \lambda_2^2 & 2\lambda_2 \\ \lambda_1^3 & 3\lambda_1^2 & \lambda_2^3 & 3\lambda_2^2 \end{bmatrix},$$

$$\lambda_1 = 1 + j1 \quad \text{and} \quad \lambda_2 = 1 - j1.$$

Find

A_ρ the amplitude of A,
A_θ the phase of A,
A_R the projected real part of A, and
A_I the projected imaginary part of A.

Solution

From Step 1 in Algorithm 6.3.2, we use the algorithm in (3.9) to obtain

$$A^{(4)} = \begin{bmatrix} 0.697246 & 0.424768 & -0.132240 & 0.026230 \\ -0.104920 & 0.907085 & 0.214929 & -0.027320 \\ 0.109282 & -0.323484 & 1.125649 & 0.105646 \\ -0.422587 & 0.954456 & -1.168657 & 1.548236 \end{bmatrix}.$$

It might be interesting to note that $\sigma(A^{(4)}) = \{\sqrt[4]{\lambda_1}, \sqrt[4]{\lambda_1}, \sqrt[4]{\lambda_2}, \sqrt[4]{\lambda_2}\} =$

$\{1.069554 + j0.2127475,\ 1.069554 + j0.2127475,\ 1.069554 - j0.2127475,\ 1.069554 - j0.2127475\}$, $\mathrm{Re}(\sigma(A)) > 0$, and $|\mathrm{Im}(\sigma(A))/\mathrm{Re}(\sigma(A))| < 1$. From Step 2 in Algorithm 6.3.2, we use Algorithm 6.3.1 to obtain

$$A_{Rd}^{(4)} = \mathrm{diag}[1.069554,\ 1.069554,\ 1.069554,\ 1.069554]$$

and

$$A_{Id}^{(4)} = \begin{bmatrix} j0.425495 & -j0.638243 & j0.319121 & -j0.106374 \\ j0.425495 & -j0.425495 & j0.212748 & -j0.106374 \\ j0.425495 & -j0.425495 & j0.425495 & -j0.212748 \\ j0.850990 & -j1.276485 & j1.276485 & -j0.425495 \end{bmatrix}.$$

Note that

$$\sigma\left(A_{Id}^{(4)}\right) = \{0.2127475,\ -0.2127475,\ 0.2127475,\ -0.2127475\} \subset R.$$

Thus, we obtain

$$A_{\rho d} = \left[\left(A_{Rd}^{(4)}\right)^2 + \left(A_{Id}^{(4)}\right)^2\right]^2 = \mathrm{diag}[1.414214,\ 1.414214,\ 1.414214,\ 1.414214].$$

From Step 3, we can compute A_θ with $q = 4$ and $\hat{X} = \left(A_{Rd}^{(q)}\right)^{-1}\left(jA_{Id}^{(q)}\right)$ as

$$A_\theta = 4A_\theta^{(4)} = \begin{bmatrix} j1.570796 & -j2.356194 & j1.178097 & -j0.392699 \\ j1.570796 & -j1.570796 & j0.785398 & -j0.392699 \\ j1.570796 & -j1.570796 & j1.570796 & -j0.785398 \\ j3.141593 & -j4.712389 & j4.712389 & -j1.570796 \end{bmatrix},$$

where

$$\sigma(A_\theta) = \{0.7853981,\ -0.7853981,\ 0.7853981,\ -0.7853981\} \subset R.$$

Also, we get

$$A_1 = A - \left[A_{Rd}^{(4)} + jA_{Id}^{(4)}\right]^4 = \begin{bmatrix} 1 & -2 & 1.5 & -0.5 \\ 2 & -3 & 2 & -0.5 \\ 2 & -2 & 1 & 0 \\ 0 & 2 & -2 & 1 \end{bmatrix},$$

and

$$A_\rho = A_{\rho d} + A_1 \exp(-jA_\theta) = \begin{bmatrix} 1.414214 & -0.707107 & 0.707109 & -0.353553 \\ 1.414214 & -1.414214 & 2.121320 & -0.707109 \\ 2.828427 & -4.242641 & 4.242641 & -0.707109 \\ 2.828427 & -2.828427 & 1.414213 & 1.414214 \end{bmatrix},$$

where

$$\sigma(A_\rho) = \{1.414214, 1.414214, 1.414214, 1.414214\} \subset R.$$

From Step 4, we obtain

$$A_R = A_\rho \cos(A_\theta) = \begin{bmatrix} 1 & -0.5 & 0.5 & -0.25 \\ 1 & -1 & 1.5 & -0.5 \\ 2 & -3 & 3 & -0.5 \\ 2 & -2 & 1 & 1 \end{bmatrix}$$

with $\sigma(A_R) = \{1, 1, 1, 1\} \subset R$, and

$$A_I = A_\rho \sin(A_\theta) = \begin{bmatrix} j1 & -j1.5 & j0.5 & -j0.25 \\ j1 & -j1 & j0.5 & -j0.5 \\ j2 & -j3 & j3 & -j1.5 \\ j6 & -j10 & j9 & -j3 \end{bmatrix}$$

with $\sigma(A_I) = \{1, -1, 1, -1\} \subset R$.

It can be shown that $A = A_R + jA_I = A_\rho \exp(jA_\theta)$, $A_\rho = [A_R^2 + A_I^2]^{1/2}$, and $A_\theta = \tan^{-1}(A_R^{-1}A_I) = \tan^{-1}(A_I A_R^{-1})$.

Note that $A_R \neq \mathrm{Re}(A) = A$, $A_I \neq \mathrm{Im}(A) = 0_4$, $A_\rho \neq [(\mathrm{Re}(A))^2 + (\mathrm{Im}(A))^2]^{1/2} = [A^2 + 0_4^2]^{1/2} = A$, and $A_\theta \neq \tan^{-1}[(\mathrm{Re}(A))^{-1}(\mathrm{Im}(A))] = \tan^{-1}(A^{-1}0_4) = 0_4$.

6.5 Conclusion

The amplitude and phase of a complex matrix and the projected real and imaginary parts of the complex matrix have been defined and computational methods for finding the above matrices have been proposed in this chapter. By utilizing the important property of the matrix-sign function that the associated matrix-sign functions of a shifted complex matrix preserve the eigenvectors of the original matrix, the algorithm for finding the principal nth root of a complex matrix has been employed for computing the amplitude and phase of the original complex matrix.

The newly developed geometric-series method can be utilized for finding the approximation of the matrix-valued function, $\tan^{-1}(X)$, where X is a matrix. Questions of computational cost have not, however, been considered in any detail. The applications of the developed amplitude and phase of a complex matrix to systems theory [32] are being investigated.

Application of the Principal *n*th Root Method to Large-scale Discrete Systems Design

A multi-stage pseudo-continuous-time state-space method is developed for designing large-scale discrete systems, which do not exhibit a two- or multi-time scale structure explicity. The designed pseudo-continuous-time regulator places the eigenvalues of the closed-loop discrete system within the common region of a circle (concentric within the unit circle) and a logarithmic spiral in the complex z-plane, without explicity utilizing the open-loop eigenvalues of the given system. The proposed method requires the solution of small order Riccati equations only at each stage of the design. An illustrative example is presented to demonstrate the effectiveness of the proposed procedures [78].

7.1 Introduction

Physical realizations of engineering systems result, in general, in large-scale models. In most cases, it is quite impractical to consider the analysis and design of the large-scale system model itself. Therefore, a necessity arises for decomposing the original system into decoupled subsystems, each with their own distinct characteristics, so that the resulting model has a completely decoupled multi-time scale structure. Some of the existing approaches for decomposition of large-scale systems are aggregation [55], multi-time scales [56] and model analysis [57]. However, most of these appear to be restricted to the continuous-time systems. The corresponding problem for large-scale discrete-time systems has received very little attention [58-60]. Mahmoud *et al.* [58] derived a matrix-norm condition for separating large-scale discrete-time systems into two-time scales without originally assuming the availability of such a structure. However, computationally, it might not be always be feasible to satisfy this condition. Shieh *et al.* [53] have developed an algebraic method based on the matrix-sign function [9] for separating the slow (dominant) modes from the fast (nondominant) modes (two-time scale structure) of a large-scale multivariable system (continuous and discrete). The matrix-sign function algorithm has been used for the following: block-diagonalization and block-triangularization [37] of a large-scale system, i.e., decomposing the system into parallel and cascaded struc-

tures; solving non-linear Riccati equations, which often appear in feedback design of systems based on linear quadratic theory; and model conversions of systems via the computation of the principal qth root of the system matrix [29,37]. Recently, fast and stable algorithms have been developed for the computation of the matrix-sign function [29] and for the computation of the principal qth root of a complex matrix [37] which in turn can be used for discrete-to-continuous model conversion. These algorithms will be utilized in the development of our multi-stage design procedure for designing discrete controllers with pole-assignment in a specified region of the complex z-plane.

The optimal linear quadratic (LQ) design method has several good properties. For instance, the closed-loop system is stable and has good robustness properties provided the weighting matrices satisfy certain positivity conditions [62]. The transient behavior of the closed-loop system is, however, difficult to determine since there is a complex relation between the weighting matrices and the closed-loop poles. This implies that the weighting matrices have to be determined through trial and error. Pole-placement methods have the advantage that the closed-loop poles can be specified. The drawback is the nonuniqueness of choice of feedback for multivariable systems. Further, it is too restrictive to place the poles in pre-determined locations [63], since for nonlinear systems the exact location of the closed-loop poles might be difficult to attain for each operational condition. Hence, in general, it would suffice to have the poles placed within a specified region. Also, the regional pole-assignment method is suited for tradeoffs between eigenvalue locations, actuator-signal magnitudes and requirements of robustness against large parameter variations, sensor failures, implementation accuracies, gain reduction, etc. [13]. In this chapter, we consider the common region of a circle and a logarithmic spiral in the z-plane (Fig. 7.2) for pole-assignment. This is equivalent to the sector region (hatched) in Fig. 7.1 in the s-plane. It is well-known that if the poles of a system lie within the above mentioned region(s), then the system responses converge at appropriate speed and any existing vibrating modes are well-damped.

The problem of designing feedback gains to optimally place all the poles of a closed-loop system within a specified region was first studied by Anderson and Moore [62], who used a shifted system matrix to obtain an optimal closed-loop system with its eigenvalues lying in the open left-hand side of a vertical line on the negative real axis. Shieh *et al.* [64,65] extended this idea to optimally place the poles within a vertical strip as well as a horizontal strip in the left-half plane. Kawasaki and Shimemura [66] propsed an iterative procedure to place the poles inside a hyperbola in the left-half plane, which is actually an approximation of the sector region shown in Fig. 7.1. In [67], a pseudo-continuous-time method has been developed to place the eigenvalues of a discrete system within the hatched region of Fig. 7.2. However, it involves the solution of full order Riccati equations, which could be computationally difficult for large-scale systems. The Luenbeger transformation, sometimes numerically unstable, is utilized to transform the full order discrete-time system to its equivalent canonical form so as to determine the pole-placement discrete-feedback gain. In this chapter, at each stage of the design, only reduced order Riccati equations need to be solved and also, in most cases, the transformation to the general canonical forms is avoided.

The material in this chapter is organized as follows: Section 7.2 contains a review of the results associated with the design of a linear quadratic regulator which would optimally place the closed-loop eigenvalues of a continuous-time system on or within the hatched region of Fig. 7.1. In Section 7.3, the method, using the matrix-sign function, for block-decomposing a large-scale discrete-time system into a multi-time scale structure is introduced. Then, a brief review of the model-conversion technqiues is given, following which a pseudo-continuous-time multi-stage design

procedure is presented for designing large-scale discrete systems decomposed in a multi-time scale structure, with pole-placement on or within the hatched region of Fig. 7.2. An illustrative example is given in Section 7.4 to demonstrate the effectiveness of the proposed design procedure and the conclusions are summarized in Section 7.5. Some computational algorithms are given in Appendix A.

7.2 Continuous-time Optimal Quadratic Regulators with Pole-placement

Consider the linear controllable continuous-time system described by

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0), \tag{7.1}$$

where $x(t)$ and $u(t)$ are the $n \times 1$ state vector and the $m \times 1$ input vector, respectively, and A and B are constant matrices of appropriate dimensions. Let the quadratic cost function for the system in (7.1) be

$$J = \int_0^\infty (x^T(t)Qx(t) + u^T(t)Ru(t))dt, \tag{7.2}$$

where the weighting matrices Q and R are $n \times n$ nonnegative-definite and $m \times m$ positive-definite symmetric matrices, respectively. The feedback-control law that minimizes the performance index in (7.2) is given by [62],

$$u(t) = -Kx(t) + \hat{r}(t) = -R^{-1}B^TPx(t) + \hat{r}(t), \tag{7.3}$$

where K is the feedback gain, $\hat{r}(t)$ is a reference input and P, a $n \times n$ nonnegative-definite symmetric matrix, is the solution of the Riccati equation,

$$PBR^{-1}B^TP - PA - A^TP - Q = 0_n \tag{7.4}$$

with (Q, A) detectable. The superscript T and the matrix 0_n denote the transpose and the $n \times n$ null matrix, respectively. Thus, the resulting closed-loop system becomes

$$\dot{x}(t) = (A - BK)x(t) + B\hat{r}(t). \tag{7.5}$$

The eigenvalues of $A - BK$, denoted by $\sigma(A - BK)$, lie in the open left-half plane of the complex s-plane. Our objective is to determine Q, R and K so that the closed-loop system in (7.5) has its eigenvalues on or within the hatched region of Fig. 7.1. The important results along with the design procedure to achieve the desired design are presented in the following.

Lemma 7.2.1 [62,67]

Let (A, B) be the pair of the given open-loop system in (7.1). Also, let $h \geq 0$ represent the prescribed degree of relative stability. Then, the eigenvalues of the closed-loop system $A - BR^{-1}B^TP$ lie to the left of the $-h$ vertical line with the matrix P being the solution of the Riccati equation,

$$PBR^{-1}B^TP - P(A + hI_n) - (A + hI_n)^TP = 0_n, \tag{7.6}$$

where the matrix I_n is an $n \times n$ identity matrix. ■

Theorem 7.2.1 [67]

Let the given stable system matrix $A \in \mathcal{R}^{n\times n}$ have eigenvalues $\hat{\lambda}_i^-$ $(i = 1,\ldots,n^-)$ lying in the open sector of Fig. 7.1 and the eigenvalues $\hat{\lambda}_i^+$ $(i = 1,\ldots,n^+)$ lying outside that sector, with $n = n^- + n^+$. Now, consider the two Riccati equations,

$$\hat{Q}BR^{-1}B^T\hat{Q} - \hat{Q}(-A^2) - (-A^2)^T\hat{Q} = 0_n \tag{7.7a}$$

and

$$PBR^{-1}B^TP - PA - A^TP - \hat{Q} = 0_n. \tag{7.7b}$$

Then, the closed-loop system,

$$A_c = A - rBK = A - rBR^{-1}B^TP, \tag{7.8}$$

will enclose the invariant eigenvalues $\hat{\lambda}_i^-$ $(i = 1,\ldots,n^-)$, and at least one additional pair of complex conjugate eigenvalues lying in the open sector of Fig. 7.1, for the constant gain r in (7.8) satisfying

$$r \geq \max\{\frac{1}{2}, \frac{b+\sqrt{b^2+ac}}{a}\}, \tag{7.9}$$

where $a = \mathrm{tr}[(BR^{-1}B^TP)^2]$, $b = \mathrm{tr}[BR^{-1}B^TPA]$ and $c = (1/2)\ \mathrm{tr}[BR^{-1}B^T\hat{Q}]$. ■

Remark 7.2.1

The steady state solutions of the Riccati equations in (7.6) and (7.7) can be found using the matrix-sign function techniques [9,23], and a brief review of this is given in the Appendix. ■

7.2.1 Continuous-time Design Procedure

Step 1.

Let the given continuous-time system be as in (7.1). Specify h so that the $-h$ vertical line on the negative real axis would represent the line beyond which the eigenvalues have to be placed in the sector of Fig. 7.1. Also, assign $A_0 = A$ and the positive-definite matrix R. Set $i = 1$. If the system is unstable, then solve (7.6) to obtain the closed-loop system $A_1 = A - r_0BR^{-1}B^TP_0 = A - r_0BK_0$, with $r_0 = 1$; else (stable system) go to *Step* 2 with $A_1 = A$, $P_0 = 0_n$ and $r_0 = 0$.

Step 2.

Solve (7.7*a*) for $\hat{Q}_i$ with $A := A_i$. Check if $\frac{1}{2}\ \mathrm{tr}[BR^{-1}B^T\hat{Q}_i]$ is zero? If it is equal to zero, go to *Step* 4 with $j = i$; else, continue and go to *Step* 3. Note that when $\frac{1}{2}\ \mathrm{tr}[BR^{-1}B^T\hat{Q}_i] = 0$, all eigenvalues of the matrix A_i lie on or within the open sector of Fig. 7.1.

Step 3.

Solve (7.7*b*) for P_i with $A := A_i$ and $\hat{Q} := \hat{Q}_i$. Then, the constant gain r_i can be evaluated using (7.9). The closed-loop system matrix is

$$A_{i+1} = A_i - r_i B R^{-1} B^T P_i = A_i - r_i B K_i. \tag{7.10a}$$

Set $i := i + 1$, and go to *Step* 2.

Step 4.

Check if tr $[(A_j + hI_n)]^+$ (sum of the eigenvalues to the right of the vertical line at $-h$) is zero? If it is equal to zero, go to *Step* 5 with $P_{j+1} = 0_n$ and $r_{j+1} = 0$; else, solve (7.6) for P_{j+1} with $A := A_j$ and obtain the closed-loop system $A_j - r_{j+1} B R^{-1} B^T P_{j+1} = A_j - r_{j+1} B K_{j+1}$, with $r_{j+1} = 1$ and $K_{j+1} = R^{-1} B^T P_{j+1}$.

Step 5.

The designed closed-loop system is

$$A_0 - B R^{-1} B^T \sum_{k=0}^{j+1} r_k P_k, \tag{7.10b}$$

and its eigenvalues lie in the hatched region of Fig. 7.1. Note that the above system matrix in (7.10*b*) is equal to the system matrix in (7.5), $A - BR^{-1}B^T\hat{P}$, where $\hat{P}$ is the solution of the Riccati equation in (7.4) with

$$Q = 2h(P_o + P_{j+1}) + \sum_{i=1}^{j} (\hat{Q}_i + \Delta r_i P_i B R^{-1} B^T P_i) r_i. \tag{7.10c}$$

In the above equation, $\Delta r_i = r_i - 1$, and the matrix R is as originally assigned. Also, the optimal continuous-time regulator can be given as

$$u(t) = -(\sum_{i=0}^{j+1} r_i K_i) x(t) + \hat{r}(t) = -Kx(t) + \hat{r}(t), \tag{7.10d}$$

where $\hat{r}(t)$ is any reference input and K is the desired state-feedback gain. ■

7.3 Pseudo-continuous-time Pole-placement Regulators

In this section, the block decomposition of a large-scale discrete-time system is considered first. In this context, the method based on the matrix-sign function [9] for block-diagonalizing a large-scale discrete system into a multi-time scale structure is discussed. Then, some of the existing model-conversion methods [28,68] for transforming a continuous-time (discrete-time) model to an equivalent discrete-time (continuous-time) model are reviewed. Finally, a pseudo-continuous-time state-space method for determining pole-placement digital regulators for eigenvalue-placement in a specific region (Fig. 7.2) is considered.

7.3.1 Block-diagonalization Via Matrix-sign Function

The definition of and an algorithm to compute the matrix-sign function are given in the Chapter 4. In the following, the results leading to the decomposition of a discrete system into a multi-time scale structure are presented.

Lemma 7.3.1 [53]

Consider a discrete-time system matrix, $G \in \mathcal{R}^{n\times n}$. The mapping $h(G) = (G-\rho I_n)(G+\rho I_n)^{-1}$, $\det[G+\rho I_n] \neq 0$, maps the circle of radius ρ in the discrete z-plane onto the imaginary axis of the $h(z)$-plane and the interior (exterior) of the circle into the open left-half (open right-half) $h(z)$-plane. ■

Definition 7.3.1 [53]

Let the eigenvalues of a discrete-time stable system matrix, $G \in \mathcal{R}^{n\times n}$, be $\lambda_i, i = 1,\ldots,n$. The nondominant modes of this system are the modes with $|\lambda_i| < \rho$, where ρ is a positive real number, while the dominant modes are those having $|\lambda_i| > \rho$, where $|(.)|$ represents the absolute value of $(.)$. If the eigenvalues of the original system are unknown, as in the case of a large-scale system, the poitive real number ρ can be chosen as $\rho = |\sqrt[n]{\det(G)}|$, which is the geomteric mean of the eigenvalues of G. If the given system G is unstable, then we choose $\rho = 1$. □

Theorem 7.3.1 [37]

Let $G \in \mathcal{R}^{n\times n}$ and $|(\sigma(G))| \cap \{\rho_i, i = 0,1,\ldots,k\} = \emptyset$, where $\sigma(G)$ represents the eigenspectrum of G, $\rho_i \in \mathcal{R}$, $i = 0,1,\ldots,k$ represent radii of circles concentric with the unit circle. Let a set of matrix-sign functions (see Chapter 4) be

$$\text{Sign}_{(\rho_i)}\,(h(G)) \triangleq \text{Sign}\left[(G-\rho_i I_n)(G+\rho_i I_n)^{-1}\right] \quad \text{for } i = 0,1,\ldots,k. \tag{7.11a}$$

Define

$$S_i \triangleq \text{ind}\,[\text{Sign}^{+}_{(\rho_{i-1},\rho_i)}\,(h(G))] \in \mathcal{R}^{n\times n_i}, \quad 1 \leq i \leq k, \tag{7.11b}$$

where ind(.) represents the collection of the linearly independent column vectors of (.), and

$$\text{Sign}^{+}_{(\rho_{i-1},\rho_i)}\,(h(G)) \triangleq \frac{1}{2}\left[\text{Sign}_{(\rho_{i-1})}\,(h(G)) - \text{Sign}_{(\rho_i)}\,(h(G))\right] \tag{7.11c}$$

with $\rho_0 = 0$, and $\text{sign}_{(0)}\,(h(G)) = I_n$. Assume that $n_i \neq 0$ for $1 \leq i \leq k$. Then, we have

$$G_R = M_s^{-1} G M_s = \text{block diag}\,[G_{Rk}, G_{R(k-1)},\ldots,G_{R1}], \tag{7.12a}$$

where M_s is the *right* block-modal matrix given by

$$M_s = [S_k, S_{k-1},\ldots,S_1], \tag{7.12b}$$

and

$$G_{Ri} = S_i^{+} G S_i \in \mathcal{R}^{n_i\times n_i} \quad \text{for } 1 \leq i \leq k. \tag{7.12c}$$

$S_i^+ \in \mathcal{R}^{n_i \times n}$ is the *left* inverse of S_i and is defined as $S_i^+ \triangleq (S_i^T S_i)^{-1} S_i^T$. ■

7.3.2 Model Conversions

Consider the system governed by the continuous-time state equation (as in (7.1)), i.e.,

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0). \tag{7.13}$$

If we approximate $u(t)$ as a piecewise input function,

$$u(t) = u(kT) \ \text{ for } kT \leq t < (k+1)T, \tag{7.14}$$

where T is the sampling period, then we can write the equivalent discrete-time model as

$$x(k+1) = Gx(k) + Hu(k),\ x(0), \tag{7.15a}$$

where

$$G = \exp(AT) \quad \text{and} \quad H = [G - I_n]A^{-1}B. \tag{7.15b}$$

If the input function $u(t)$ is not a piecewise constant, a better formulation of the input matrix H can be obtained according to the nature of $u(t)$. In general, the matrices G and H can be determined exactly from the matrices A and B, and the input function $u(t)$ in (7.14) using the eigenvalue and eigenvector approach [68]. However, for computational purposes, approximations are required for obtaining G and H matrices without involving the eigenvalues explicitly. There are a number of methods available [18] to evaluate approximately G and H given in (7.15). The simplest one of them is the truncation of the infinite series of $\exp(AT)$ [68] which results in a good approximation when $T \ll 1$. A popular method for determining G and H approximately is the Pade approximation method [28,68]. Some of the approximations obtained using this method are listed below,

$$G \simeq [I_n - \frac{1}{2}AT]^{-1}[I_n + \frac{1}{2}AT] \triangleq G_3 \tag{7.16a}$$

$$\simeq [I_n - \frac{1}{2}AT + \frac{1}{12}(AT)^2]^{-1}[I_n + \frac{1}{2}AT + \frac{1}{12}(AT)^2] \triangleq G_5 \tag{7.16b}$$

and

$$H \simeq T[I_n - \frac{1}{2}AT]^{-1}B \triangleq H_3 \tag{7.17a}$$

$$\simeq T[I_n - \frac{1}{2}AT + \frac{1}{12}(AT)^2]^{-1}B \triangleq H_5. \tag{7.17b}$$

It can be noted that the matrices G_3 in (7.16a) and H_3 in (7.17a) correspond to the popular Tustin approximation (bilinear transformation) [70]. The matrices G_5

and H_5, when used with even large sampling periods, provide good approximations. The use of scaling and squaring method [68] as shown below, along with one of the above approximations, would result in better approximations,

$$G \simeq [e^{AT/m}]^m \ , \quad m \text{ is a power of two.} \tag{7.18}$$

Now, given a discrete-time model as in (7.15), an equivalent continuous-time model in (7.13) can be obtained by using the following equations,

$$A = \frac{1}{T} \ln (G), \quad \text{and} \quad B = A[G - I_n]^{-1} H. \tag{7.19}$$

As before, the matrix A can be obtained from its discrete equivalent G exactly by using the eigenvalue and eigenvector approach. It can also be obtained approximately by truncating the infinite power series of the matrix-logarithm function ln(G), subject to certain convergence conditions. Shieh *et al.* [28] have proposed a direct-truncation method and a matrix continued-fraction method for determining A from G. The commonly used approximation for $\frac{1}{T} \ln (G)$, obtained using the matrix continued-fraction method, is

$$A = \frac{1}{T} \ln (G) \simeq \frac{2}{T} R \simeq \frac{2}{T} R[I_n - \frac{4}{15} R^2][I_n - \frac{3}{5} R^2]^{-1}, \tag{7.20}$$

where $R = [G - I_n][G + I_n]^{-1}$. The matrix-series approximations obtained from truncation or continued fractions converge when Re $(\sigma(G)) > 0$, where $\sigma(G)$ represents the eigenvalues of G. In general, the eigenvalues of the matrix G are not available, and they do not always lie in the right-half of the complex z-plane. In order to satisfy the convergence condition, the principal qth root of the matrix G [28,29,61] can be made use of. Shieh *et al.* [29] and Tsai *et al.* [61] have recently developed a fast and stable algorithm for computing the principal qth root of a general complex matrix. This is listed in Chapter 3. The eigenvalues of $\sqrt[q]{G}$ lie in the right-half of the complex z-plane, i.e., Re $(\sigma(\sqrt[q]{G})) > 0$, for $q \geq 2$. Therefore, instead of G the principal qth root of G can be used in determining an approximation for A. In this case, the matrix equation (7.19) becomes

$$A = \frac{1}{T} \ln (G) = \frac{q}{T} \ln (\sqrt[q]{G}). \tag{7.21}$$

As a result, the matrix R in equation (7.20) would become $R := [\sqrt[q]{G} - I_n][\sqrt[q]{G} + I_n]^{-1}$, and the constant factor $2/T$ would be replaced by $2q/T$. The condition for the convergence of the power series of ln $(\sqrt[q]{G})$ becomes arg $(\sigma(G)) \neq \pi$ and det $(G) \neq 0$, which is a much less restrictive condition.

7.3.3 Pseudo-continuous-time Multi-stage Design Procedure

Let the given large-scale discrete-time system with appropriate sampling period T be

$$x(k+1) = \bar{G}x(k) + \bar{H}u(k), \ x(0). \tag{7.22}$$

Also, let the dimension of the system be n and the number of inputs be m. The objective is to first decompose the system into a multi-time scale structure, using techniques based on the matrix-sign function, then design each decomposed subsystem using model conversions and with eigenvalue placement in the hatched region of Fig. 7.2, and finally determine the digital regulator for the whole large-scale system.

Step 1.

Set $i = 1$, $G := \bar{G}$, $H := \bar{H}$ and the feedback gain $\bar{K} = 0_{m \times n}$.

Step 2.

Now, specify a positive real scalar ρ_i (see Definition 7.3.1) and find a transformation matrix $M_1^{(i)}$ such that the matrix G can be block-diagonalized into the following form,

$$G := (M_1^{(i)})^{-1} G M_1^{(i)} = \text{block diag } [G_c, \hat{G}_i, \bar{G}_i], \tag{7.23a}$$

where $G_c \in \mathcal{R}^{(n-n_i)\times(n-n_i)}$ represents a block, which has already been designed or does not need to be designed, and the matrices $\hat{G}_i \in \mathcal{R}^{\hat{n}_i \times \hat{n}_i}$ and $\bar{G}_i \in \mathcal{R}^{\bar{n}_i \times \bar{n}_i}$, with $n_i = \hat{n}_i + \bar{n}_i$, contain eigenvalues less than and greater than (in absolute value) ρ_i, respectively. The transformation matrix $M_1^{(i)}$ is given by

$$M_1^{(i)} = \text{block diag } \left[I_{n-n_i}, (S_2, S_1)\right], \tag{7.23b}$$

where $S_1 \in \mathcal{R}^{n \times \bar{n}_i}$ and $S_2 \in \mathcal{R}^{n \times \hat{n}_i}$ are as defined in (7.11) with respect to the matrix-sign function of the matrix G_i, where $G_i :=$ block diag $[\hat{G}_i, \bar{G}_i]$, $i > 1$ and $G_i := G$, $i = 1$. Using $M_1^{(i)}$, transform H as

$$H := (M_1^{(i)})^{-1} H = [H_c^T, \hat{H}_i^T, \bar{H}_i^T]^T. \tag{7.23c}$$

The dimensions of the matrices H_c, $\hat{H}_i$ and $\bar{H}_i$ are $(n - n_i) \times m$, $\hat{n}_i \times m$ and $\bar{n}_i \times m$, respectively. Accumulate the transformations in $M_1^{(i)} := M_1^{(i-1)} M_1^{(i)}$.

Step 3.

The subsystem considered for design at this stage is $(\bar{G}_i, \bar{H}_i)$. Transform the above discrete-time system into an equivalent continuous-time system, $(\bar{A}_i, \bar{B}_i)$, using the principal qth root techniques [29.61] and apply the design procedure given in Section 7.2 to design this continuous-time system. Let the immmediate optimal closed-loop continuous-time system be $(\bar{A}_{c_i}, \bar{B}_i)$.

Step 4.

Transform the designed continuous-time system into an equivalent discrete-time system, $(\bar{G}_{c_i}, \bar{H}_{c_i})$, using techniques discussed earlier in this section.

Step 5.

If $\bar{H}_i$ is invertible (nonsingular), then the discrete-time feedback gain for this design stage is given by

$$\bar{K}_i = (\bar{H}_i)^{-1} (\bar{G}_i - \bar{G}_{c_i}). \tag{7.24}$$

The dimension of $\bar{K}_i$ is $m \times \bar{n}_i$. When $\bar{H}_i$ is non-square, then the feedback gain $\bar{K}_i$ can be found through appropriate coordinate transformations [69] and other manipulations [65,70].

Step 6.

Update the feedback gain $\bar{K}$ and the system matrix G, respectively, as

$$\bar{K} := \bar{K} + [0_{m\times(n-\bar{n}_i)}, \bar{K}_i](M_1^{(i)})^{-1}, \tag{7.25}$$

$$G := G - H[0_{m\times(n-\bar{n}_i)}, \bar{K}_i] = \begin{bmatrix} \tilde{G}_i & W_i \\ 0_{\bar{n}_i\times(n-\bar{n}_i)} & \bar{G}_{c_i} \end{bmatrix}, \tag{7.26}$$

where $\tilde{G}_i$ = block diag $[G_c, \hat{G}_i]$, $W_i = -[H_c^T, \hat{H}_i^T]^T \bar{K}_i$ and the dimensions of the matrices $\tilde{G}_i$ and W_i are $(n - \bar{n}_i) \times (n - \bar{n}_i)$ and $(n - \bar{n}_i) \times \bar{n}_i$, respectively.

Step 7.

Block-diagonalize the partially designed system G and move the last block of G in (7.26) (viz., $\bar{G}_{c_i}$) to the first, via a transformation matrix $M_2^{(i)}$ which is given as

$$M_2^{(i)} = \begin{bmatrix} L_i & I_{n-\bar{n}_i} \\ I_{\bar{n}_i} & 0_{\bar{n}_i\times(n-\bar{n}_i)} \end{bmatrix}, \quad (M_2^{(i)})^{-1} = \begin{bmatrix} 0_{\bar{n}_i\times(n-\bar{n}_i)} & I_{\bar{n}_i} \\ I_{n-\bar{n}_i} & -L_i \end{bmatrix}. \tag{7.27a}$$

The matrix L_i ($\in \mathcal{R}^{(n-\bar{n}_i)\times\bar{n}_i}$) can be solved from the following Lyapunov equation [37], [58]-[60],

$$\tilde{G}_i L_i - L_i \bar{G}_{c_i} + W_i = 0_{(n-\bar{n}_i)\times\bar{n}_i}. \tag{7.27b}$$

The transformed system is

$$G := (M_2^{(i)})^{-1} G M_2^{(i)} = \begin{bmatrix} \bar{G}_{c_i} & 0_{\bar{n}_i\times(n-\bar{n}_i)} \\ 0_{(n-\bar{n}_i)\times\bar{n}_i} & \tilde{G}_i \end{bmatrix}, \tag{7.28a}$$

$$H := (M_2^{(i)})^{-1} H = [\bar{H}_i^T, (\tilde{H}_i - L_i \bar{H}_i)^T]^T \tag{7.28b}$$

where $\tilde{H}_i = [H_c^T, \hat{H}_i^T]^T$. Accumulate the transformations in $M_1^{(i)} := M_1^{(i)} M_2^{(i)}$.

Step 8.

Set $i := i + 1$. If $i > k$ (k is the number of time-scales), then stop; else, go to *Step* 2.

The digital regulator is

$$u(k) = -\bar{K} x(k) + r(k) \tag{7.29}$$

with $r(k)$ as any reference input, would place the eigenvalues of the system in (7.22) within the hatched region of Fig. 7.2. Also, when the sampling period T is sufficiently small, the digital regulator can be considered as a suboptimal discrete-time regulator because of the approximations involved in the inputs and the various model conversions, although the equivalent continuous-time regulator is optimal. ■

7.4 Illustrative Example

Consider an unstable discrete-time system in (7.22) with

$$\bar{G} = \begin{bmatrix} -0.357 & -0.657 & -0.146 & 0.119 & -0.041 \\ 1.675 & -0.460 & 0.335 & 0.000 & 0.167 \\ -0.075 & 0.146 & 0.360 & -0.593 & 0.009 \\ 0.376 & 0.263 & 0.518 & 0.280 & 0.016 \\ -0.882 & 0.252 & -0.176 & 0.000 & -0.070 \end{bmatrix},$$

$$\bar{H} = \begin{bmatrix} 0.689 & 0.283 \\ 0.240 & -0.387 \\ -0.339 & 0.332 \\ 0.063 & 0.020 \\ -0.126 & 0.268 \end{bmatrix}, \tag{7.30a}$$

and $\sigma(\bar{G}) = \{-0.46 \pm j1.005,\ 0.3276 \pm j0.5103,\ 0.0179\}$ for $T = 0.5$.

The location of the poles of $\bar{G}$ in the discrete z-plane are shown in Fig. 7.2 and it is seen that except for the one at 0.0179, which is to be kept invariant, the rest of the poles lie outside the region of interest. The objective is to design the discrete-time system in (7.30a) with multi-time scale decomposition and pole-assignment within the specified region in the z-plane. The pseudo-continuous-time design procedure given in Section 7.3.3 will be used to achieve the desired design.

Since the given system is unstable, the first step is to block decompose the system into its stable and unstable parts. Assign $\rho_1 = 1$ (represents the unit circle). The transformation matrix $M_1^{(1)}$, found using the matrix-sign function technique given in Section 7.3.1, which block-diagonalizes $\bar{G}$ is given by

$$M_1^{(1)} = [S_2, S_1]$$

$$= \begin{bmatrix} \begin{pmatrix} -0.0464 & -0.0551 & -0.2090 \\ 0.0003 & -0.0002 & 0.0003 \\ 0.2325 & 0.0124 & 1.0464 \\ 0.4189 & 0.0222 & 0.0837 \\ 0.0002 & 0.5269 & 0.0000 \end{pmatrix} & \begin{pmatrix} 1.0463 & 0.0551 \\ -0.0003 & 1.0002 \\ -0.2325 & -0.0124 \\ -0.4189 & -0.0222 \\ -0.0002 & -0.5269 \end{pmatrix} \end{bmatrix}, \tag{7.30b}$$

where $S_2 \in \mathcal{R}^{5\times 3}$ and $S_1 \in \mathcal{R}^{5\times 2}$ can be found from (7.11). The transformed matrices, using M_1, corresponding to $\bar{G}$ and $\bar{H}$ in (7.30a), are

$$G := (M_1^{(1)})^{-1}\bar{G}M_1^{(1)} = \text{block diag } [\hat{G}_1, \bar{G}_1]$$

$$= \begin{bmatrix} \begin{pmatrix} 0.5827 & -0.0002 & 1.1480 \\ -0.0002 & 0.0179 & -0.0001 \\ -0.2835 & 0.0000 & 0.0726 \end{pmatrix} & 0_{2\times 3} \\ 0_{3\times 2} & \begin{pmatrix} -0.4601 & -0.6027 \\ 1.6744 & -0.4601 \end{pmatrix} \end{bmatrix}, \qquad (7.31a)$$

$$H := (M_1^{(1)})^{-1}\bar{H} = \begin{bmatrix} \hat{H}_1 \\ \bar{H}_1 \end{bmatrix}$$

$$= \begin{bmatrix} \begin{pmatrix} 0.8464 & 0.3328 \\ 0.0007 & 0.1218 \\ -0.3739 & 0.3208 \end{pmatrix} \\ \begin{pmatrix} 0.6087 & 0.3761 \\ 0.2400 & -0.3870 \end{pmatrix} \end{bmatrix}. \qquad (7.31b)$$

The eigenspectra of the diagonal blocks in (7.31a) are $\sigma(\hat{G}) = \{0.3276 \pm j0.5103, 0.0179\}$ and $\sigma(\bar{G}) = \{-0.46 \pm j1.005\}$. The unstable subsystem $(\bar{G}_1, \bar{H}_1)$ is to be designed at this stage. The equivalent continuous-time subsystem is found using the principal qth root of $\bar{G}_1$ ($q = 4$) (the algorithm in Chapter 3). Note that since the eigenvalues of $\bar{G}_1$ are in the right-half z-plane, the well-known bilinear transformation for model conversion will not converge. The continuous-time subsystem is

$$\bar{A}_1 = \begin{bmatrix} 0.1996 & -2.4000 \\ 6.6679 & 0.1997 \end{bmatrix}, \quad \bar{B}_1 = \begin{bmatrix} 0.9993 & -0.0003 \\ -1.6667 & -1.6648 \end{bmatrix} \qquad (7.32a)$$

with $\sigma(\bar{A}_1) = \{0.1996 \pm 4.0006\}$. Assign $h = 1.1$, i.e., the eigenvalues of the closed-loop system should lie to the left the vertical line at -1.1 on the negative real axis in the s-plane, and $R = I_2$. To achieve the necessary design, we follow the steps of the continuous-time design procedure in Section 7.2.1. Let $A = \bar{A}_1$ and $B = \bar{B}_1$. Solving the Riccati equation in (7.6) with $(A + hI_2, B)$, we have

$$P_0 = \begin{bmatrix} 2.250 & -0.038 \\ -0.038 & 0.509 \end{bmatrix}, \qquad K_0 = R^{-1}\bar{B}_1^T P_0 = \begin{bmatrix} 2.311 & -0.887 \\ 0.062 & -0.848 \end{bmatrix}. \qquad (7.32b)$$

The resulting closed-loop system is

$$A_1 = A - BK_0 = \begin{bmatrix} -2.110 & -1.515 \\ 10.623 & -2.690 \end{bmatrix}, \qquad (7.32c)$$

and $\sigma(A_1) = \{-2.3996 \pm j4.0006\}$. Note that $|(\text{Re } \sigma(A_1))| > 1.1$. Now, solving the Riccati equation in (7.7a) with $(-A_1^2, B)$, we have

$$\hat{Q}_1 = \begin{bmatrix} 31.469 & 1.284 \\ 1.284 & 2.494 \end{bmatrix}, \qquad \text{and } (1/2)\text{ tr } [BR^{-1}B^T\hat{Q}_1] = 20.49 \neq 0. \qquad (7.32d)$$

Solving the Riccati equation in (7.7b) with (A_1, B) and $\hat{Q}_1$, we obtain

$$P_1 = \begin{bmatrix} 4.093 & 0.073 \\ 0.073 & 0.326 \end{bmatrix}, \qquad K_1 = R^{-1}B^T P_1 = \begin{bmatrix} 3.968 & -0.471 \\ -0.123 & -0.543 \end{bmatrix}. \tag{7.32e}$$

From (7.9), the constant gain $r_1 = 0.6385$. Therefore, the closed-loop system is

$$A_2 = A_1 - r_1 B K_1 = \begin{bmatrix} -4.641 & -1.214 \\ 14.714 & -3.768 \end{bmatrix}, \qquad \hat{Q}_2 = 0_2, \tag{7.32f}$$

and $\sigma(A_2) = \{-4.2046 \pm j4.2046\}$. Note that all the eigenvalues lie on the boundary of the hatched region in Fig. 7.1, tr $[(A_2 + hI_2)^+] = 0$ and $(1/2)$ tr $[BR^{-1}B^T\hat{Q}_2] = 0$, where $\hat{Q}_2$ solved from (7.7a) with respect to $(-A_2^2, B)$. This verifies that the desired design has been achieved for the subsystem in (7.32a). Let us denote this closed-loop subsystem by $\bar{A}_{c_1} = \bar{A}_1 - \bar{B}_1(K_0 + r_1 K_1)$. Now, we transform this continuous-time system into its equivalent discrete-time system, $\bar{G}_{c1}$, given by

$$\bar{G}_{c1} = \begin{bmatrix} -0.0729 & -0.0304 \\ 0.3686 & -0.0510 \end{bmatrix}. \tag{7.33a}$$

The eigenspectrum corresponding to this system matrix is $\sigma(\bar{G}_{c1}) = \{-0.0619 \pm j0.1053\}$. Note that this complex conjugate pair is inside the hatched region of Fig. 7.2. The discrete-time feedback gain for this stage is

$$\bar{K}_1 = (\bar{H}_1)^{-1}(\bar{G}_1 - \bar{G}_{c1}) = \begin{bmatrix} 1.047 & -1.152 \\ -2.725 & 0.343 \end{bmatrix}. \tag{7.33b}$$

Using this feedback gain, the updated system is given by

$$G := G - H[0_{2\times 3}, \bar{K}_1] = \begin{bmatrix} \bar{G}_1 & W_1 \\ 0_{3\times 2} & \bar{G}_{c1} \end{bmatrix}$$

$$= \begin{bmatrix} \begin{pmatrix} 0.5827 & -0.0002 & 1.1480 \\ -0.0002 & 0.0179 & -0.0001 \\ -0.2835 & 0.0000 & 0.0726 \end{pmatrix} & \begin{pmatrix} 0.0204 & 0.8610 \\ 0.3311 & -0.0410 \end{pmatrix} \\ 0_{2\times 3} & \begin{pmatrix} -0.0729 & -0.0304 \\ 0.3686 & -0.0510 \end{pmatrix} \end{bmatrix}. \tag{7.33c}$$

The updated feedback gain $\bar{K}$ is

$$\bar{K} := \bar{K} + [0_{2\times 3}, \bar{K}_1](M_1^{(1)})^{-1}$$

$$= \begin{bmatrix} 1.047 & -1.152 & 0.209 & 0.000 & 0.104 \\ -2.725 & 0.343 & -0.544 & 0.000 & -0.272 \end{bmatrix}. \tag{7.33d}$$

The solution of the Lyapunov equation in (7.27*b*) for $i = 1$ and $\bar{n}_i = 2$ is

$$L_1 = \begin{bmatrix} 2.893 & -2.029 \\ -0.449 & 0.787 \\ -2.322 & 0.293 \end{bmatrix}. \tag{7.33e}$$

Thus the transformation matrix $M_2^{(1)}$ that block-diagonalizes G in (7.33*c*) and swaps the blocks $\bar{G}_{c1}$ and $\bar{G}_1$ is given by

$$M_2^{(1)} = \begin{bmatrix} L_1 & I_3 \\ I_2 & 0_{2\times 3} \end{bmatrix}. \tag{7.33f}$$

The transformed system is now given by

$$G := (M_2^{(1)})^{-1} G M_2^{(1)} = \begin{bmatrix} \bar{G}_{c1} & 0_{2\times 3} \\ 0_{3\times 2} & \bar{G}_1 \end{bmatrix} \tag{7.34a}$$

and

$$H := (M_2^{(1)})^{-1} H = \begin{bmatrix} \begin{pmatrix} 0.6087 & 0.3761 \\ 0.2400 & -0.3870 \end{pmatrix} \\ \begin{pmatrix} -0.4280 & -1.5404 \\ 0.0853 & 0.5952 \\ 0.9688 & 1.3074 \end{pmatrix} \end{bmatrix}. \tag{7.34b}$$

The accumulated transformation becomes $M_1^{(1)} := M_1^{(1)} M_2^{(1)}$.

Now, we proceed to the second stage of design which consists of designing the stable dominant poles of the original discrete-time system in (7.30*a*). Choose $\rho_2 = e^{-hT} = 0.57695$. The transformation matrix M_1 which block-diagonalizes the block $\bar{G}_1$ in (7.33*c*) while preserving the block $\bar{G}_{c1}$ is given by (as in (7.23*b*))

$$M_1^{(2)} = \begin{bmatrix} I_2 & 0_{2\times 3} \\ 0_{3\times 2} & (S_2, S_1) \end{bmatrix} \tag{7.34c}$$

with

$$[S_2, S_1] = \begin{bmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \end{bmatrix}, \tag{7.34d}$$

where $S_2 \in \mathcal{R}^{3\times 2}$ and $S_1 \in \mathcal{R}^{3\times 1}$ can be found from (7.11) with respect to $\bar{G}_1$ and ρ_2. The transformed matrices G and H are

$$G := (M_1^{(2)})^{-1} G M_1^{(2)} = \text{block diag } [\bar{G}_{c1}, \hat{G}_2, \bar{G}_2]$$

$$= \begin{bmatrix} \begin{pmatrix} -0.0729 & -0.0304 \\ 0.3686 & -0.0510 \end{pmatrix} & 0_{2\times 1} & 0_{2\times 2} \\ 0_{1\times 2} & (0.0179) & 0_{1\times 2} \\ 0_{2\times 2} & 0_{2\times 1} & \begin{pmatrix} 0.5827 & 1.1480 \\ -0.2835 & 0.0726 \end{pmatrix} \end{bmatrix}, \qquad (7.35a)$$

$$H := (M_1^{(2)})^{-1} H = \begin{bmatrix} H_1 \\ \hat{H}_2 \\ \bar{H}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \begin{pmatrix} 0.6087 & 0.3761 \\ 0.2400 & -0.3870 \end{pmatrix} \\ (0.0847 \quad 0.5943) \\ \begin{pmatrix} -0.4279 & -1.1540 \\ 0.9687 & 1.3073 \end{pmatrix} \end{bmatrix}. \qquad (7.35b)$$

Again, the accumulated transformation becomes $M_1^{(2)} := M_1^{(1)} M_1^{(2)}$. The subsystem to be designed in this stage is $(\bar{G}_2, \bar{H}_2)$. Following the same procedures as in the first stage, we obtained the designed discrete subsystem as

$$\bar{G}_{c2} = \begin{bmatrix} 0.4272 & 1.1292 \\ -0.1513 & -0.1608 \end{bmatrix} \qquad (7.35c)$$

with $\sigma(\bar{G}_{c2}) = \{0.1332 \pm j0.2905\}$. Again, note that these eigenvalues are within the hatched region of Fig. 7.2. The discrete-time feedback gain for this stage is

$$\bar{K}_2 = (\bar{H}_2)^{-1}(\bar{G}_2 - \bar{G}_{c2}) = \begin{bmatrix} 0.0000 & 0.4117 \\ -0.1008 & -0.1265 \end{bmatrix}. \qquad (7.35d)$$

The updated feedback gain is

$$\bar{K} := \bar{K} + [0_{3\times 2}, \bar{K}_2](M_1^{(2)})^{-1}$$

$$= \begin{bmatrix} 2.000 & -1.274 & 0.812 & -0.229 & 0.199 \\ -2.828 & 0.175 & -0.671 & -0.181 & -0.272 \end{bmatrix}. \qquad (7.35e)$$

The eigenvalues of $G - H[0_{2\times 3}, \bar{K}_2]$, with G and H as in (7.35a) and (7.35b), are $\{-0.0619 \pm j0.1053,\ 0.1332 \pm j0.2905,\ 0.0179\}$. Note that all of them are within the hatched region of Fig. 7.2, and the nondominant eigenvalue of the open-loop system at 0.0179 is not designed. Therefore, the closed-loop discrete-time system is

$$\bar{G}_c = \bar{G} - \bar{H}\bar{K}$$

$$= \begin{bmatrix} -0.9374 & 0.1709 & -0.5157 & 0.3282 & -0.1013 \\ 0.0997 & -0.0864 & -0.1197 & -0.0151 & 0.0139 \\ 1.5432 & -0.3440 & 0.8582 & -0.6105 & 0.1669 \\ 0.3063 & 0.3397 & 0.4803 & 0.2981 & 0.0089 \\ 0.1284 & 0.0445 & 0.1062 & 0.0197 & 0.0280 \end{bmatrix}. \qquad (7.36a)$$

The pseudo-continuous-time pole-placement regulator is given by

$$u(k) = -\bar{K}x(k) + r(k), \qquad (7.36b)$$

where $\bar{K}$ is the total feedback gain as in (7.35e) and $r(k)$ is any reference input.

7.5 Conclusion

The design of large-scale discrete-time systems, which do not exhibit a two- or multi-time scale structure explicitly, has been considered in this chapter. It has been shown that a large-scale discrete system can be decomposed into a completely decoupled multi-time scale structure (block-diagonalization) using the techniques based on the matrix-sign function, without explicitly utilizing the open-loop eigenvalues of the given system. A pseudo-continuous-time state-space method, based on model conversions, has been developed for methodically designing each subsystem (corresponding to one-time scale), with eigenvalue-placement in a desired region of the complex z-plane. The model conversions and various other computations can be achieved using fast and stable algorithms based on the principal qth root of the system matrix and the matrix-sign functions. When the sampling period T is sufficiently small, the designed discrete controller is suboptimal while its associated continuous-time controller is optimal with respect to certain weighting matrices. The proposed method requires the solution of Riccati equations of small order only at each stage of the design. Transformation to general canonical form so as to determine the discrete feedback gain can be avoided in most cases. The developed state-space method can be used to design multivariable digital control systems, for determining the state-feedback pole-placement controllers; whereas, the existing pseudo-continuous-time frequency-domain method [71] can only be applied to design single-variable digital control systems for obtaining the cascaded controllers.

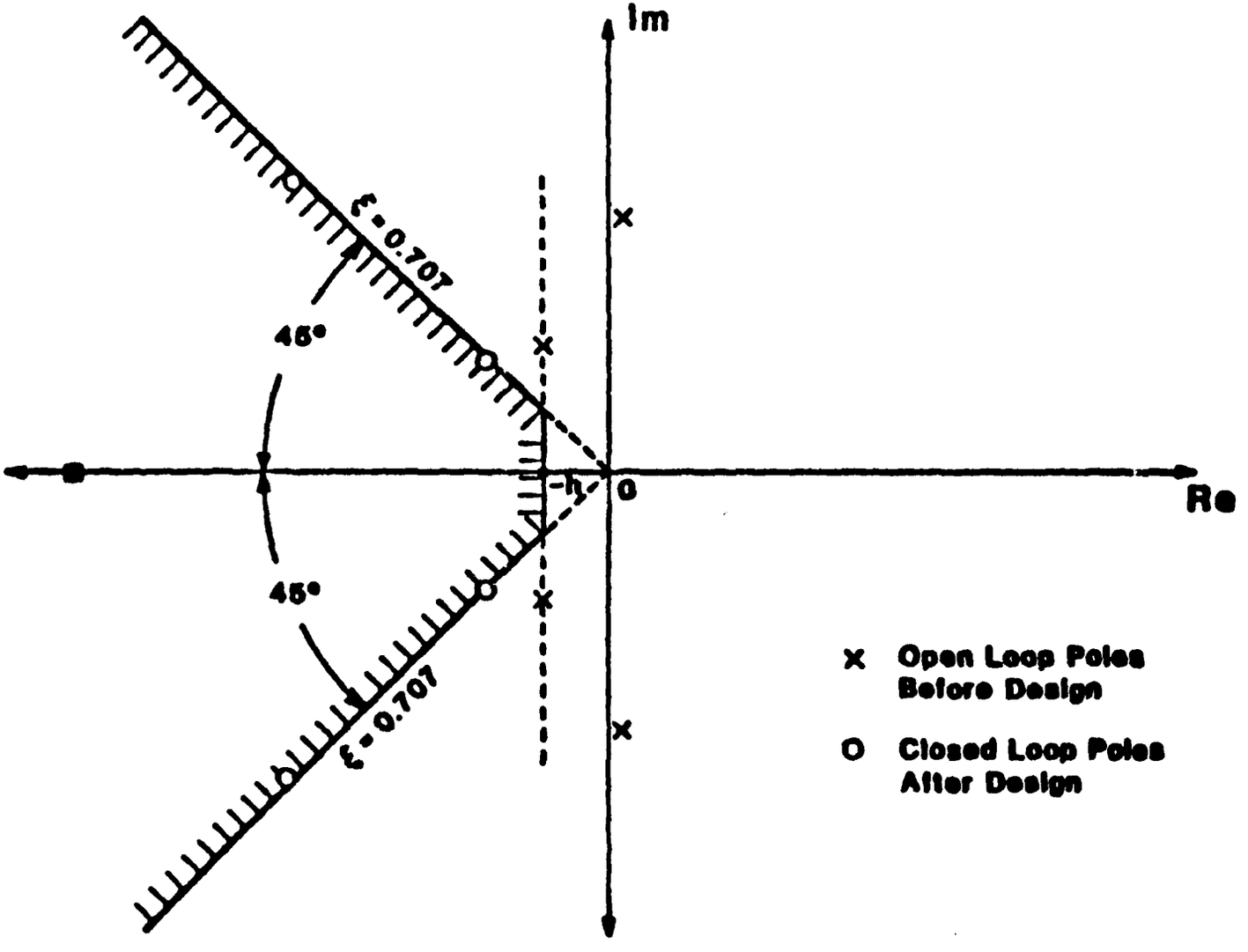


Fig. 7.1 The region of interest in the continuous-time s-plane.

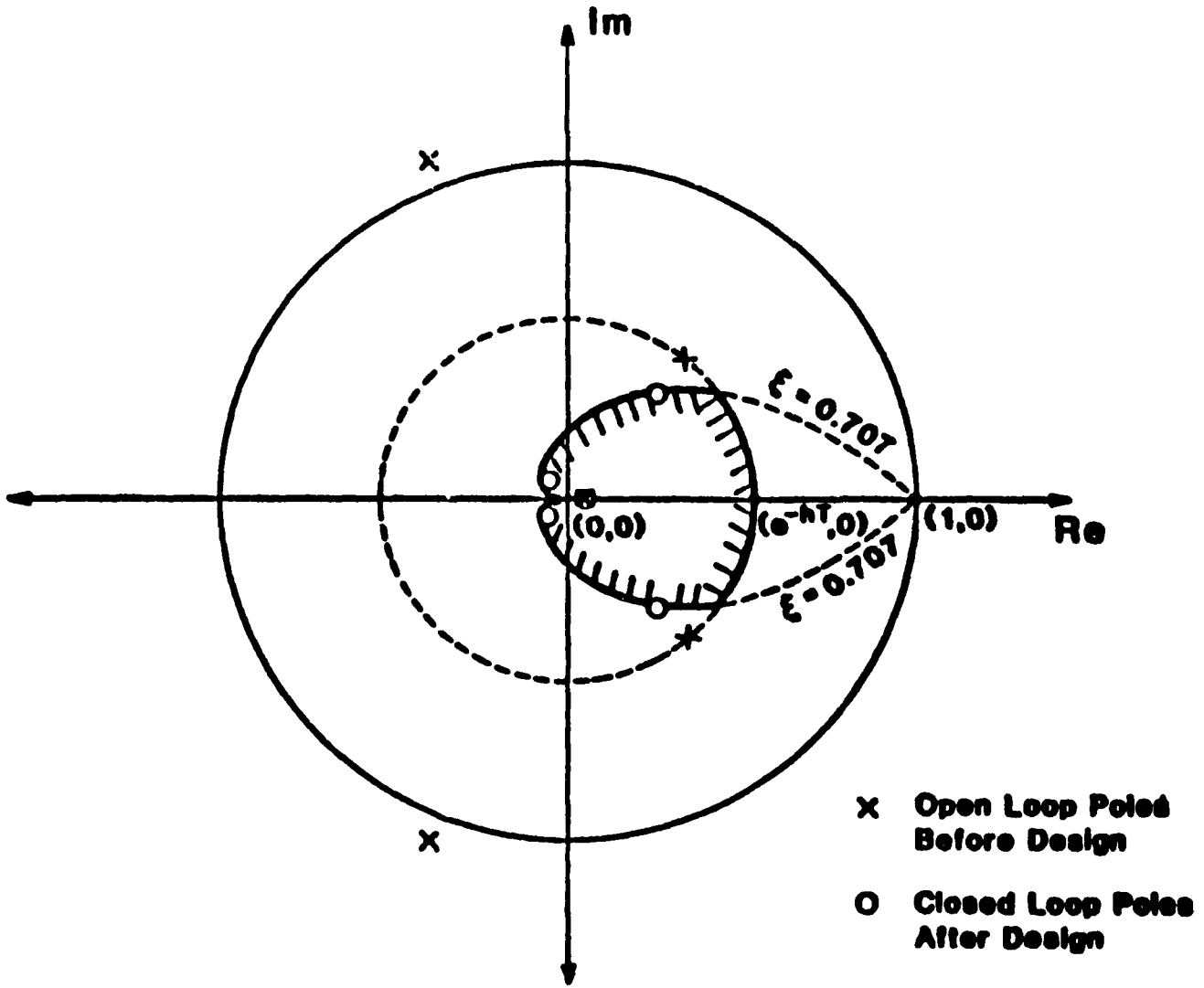


Fig. 7.2 The region of interest in the discrete-time z-plane.

Conclusions

A complete study of the *principal* nth root of a complex matrix and associated matrix-valued functions is presented in this research monograph. This includes the development of techniques to compute the principal nth root of a matrix, study of associated matrix-valued functions, and their applications to mathematical sciences and control systems.

In Chapter 2, the generalized continued-fraction method developed for finding the nth roots of real numbers has been extended to determine the principal nth roots of complex matrices. Computational algorithms with high order convergence rates have been established for determination of the principal nth root and the associated pth power of the principal nth root of a complex matrix. The global convergence properties of the proposed algorithms have been investigated from the viewpoint of systems theory.

Rapidly convergent and more stable recursive algorithms for finding the principal nth root of a matrix have been developed in Chapter 3. The developed recursive algorithms can be applied to an ill-conditioned matrix containing large and small eigenvalues. By means of a perturbation analysis with suitable assumptions, it is shown that the proposed recursive algorithms are numerically more stable than the algorithms in [20,21,26]. The analysis of absolute numerical stability of the proposed algorithms has not been done in this research monograph. The developed algorithms will enhance the capabilities of the existing computational algorithms such as the principal nth root algorithm, the matrix-sign algorithm and the matrix-sector algorithm which in turn can be applied to many control-system problems.

In Chapter 4, the matrix-sector function of A has been generalized to the matrix-sector function of $\mathbf{g}(A)$. Based on the computationally fast and numerically stable algorithms for computing the principal nth root of a matrix, fast and stable algorithms for computing the matrix-sector function and the generalized matrix-sector function have beed developed. The generalized matrix-sector function of A has been utilized to carry out the separation of matrix eigenvalues relative to a sector, circle and a sector of a circle in the λ-plane. Also, the generalized matrix sector function of A has been employed for block-diagonalization and block-triangularization of the system matrix, which are useful in developing applications to mathematical science [32] and control-system problems [31].

New computational methods, which utilize the direct-truncation method, the

matrix continued-fraction method, and the geometrix-series method together with the principal qth root of a discrete-time system matrix have been presented in Chapter 5 for quick modeling of the equivalent continuous-time state equations from the discrete-time state equations. The proposed method is useful for identifying a continuous-time system based on the observation of sampled input-output data and for design of sampled-data control systems.

The amplitude and phase of a complex matrix and the projected real and imaginary parts of the complex matrix have been defined and computational methods for finding the above matrices have been proposed in Chapter 6. By utilizing the important property of the matrix-sign function that the associated matrix-sign functions of a shifted complex matrix preserve the eigenvectors of the original matrix, the algorithm for finding the principal nth root of a complex matrix has been employed for computing the amplitude and phase of the original complex matrix. The newly developed geometric-series method can be utilized for finding the approximation of the matrix-valued function, $\tan^{-1}(X)$, where X is a matrix. Questions of computational cost have not, however, been considered in any detail. The applications of the developed amplitude and phase of a complex matrix to systems theory [32] are being investigated.

The design of large-scale discrete-time systems, which do not exhibit a two- or multi-time scale structure explicitly, has been considered in Chapter 7. It has been shown that a large-scale discrete system can be decomposed into a completely decoupled multi-time scale structure (block-diagonalization) using the techniques based on the matrix-sign function, without explicitly utilizing the open-loop eigenvalues of the given system. A pseudo-continuous-time state-space method, based on model conversions, has been developed for methodically designing each subsystem (corresponding to one-time scale), with eigenvalue-placement in a desired region of the complex z-plane. The model conversions and various other computations can be achieved using fast and stable algorithms based on the principal qth root of the system matrix and the matrix-sign functions. When the sampling period T is sufficiently small, the designed discrete controller is suboptimal while its associated continuous-time controller is optimal with respect to certain weighting matrices. The proposed method requires the solution of Riccati equations of small order only at each stage of the design. Transformation to general canonical form so as to determine the discrete feedback gain can be avoided in most cases. The developed state-space method can be used to design multivariable digital control systems, for determining the state-feedback pole-placement controllers; whereas, the existing pseudo-continuous-time frequency-domain method [71] can only be applied to design single-variable digital control systems for obtaining the cascaded controllers.

References

[1] Denman, E.D. and Beavers, A.N., "The matrix-sign function and computations in systems," *Applied Mathematics and Computation*, vol. 2, pp. 63-94, 1976.

[2] Shieh, L.S., Patel, C.G. and Chow, H.Z., "Additional properties and applications of matrix continued-fraction," *International Journal of Systems Science*, vol. 8, pp. 97-109 , 1977.

[3] Hoskins, W.D. and Walton, D.J., "A fast method of computing the square root of a matrix," *IEEE Transactions on Automatic Control*, vol. AC-23, pp. 494-495, 1978.

[4] Hoskins, W.D. and Walton, D.J., "A fast, more stable method for computing the pth roots of positive-definite matrices," *Linear Algebra and Its Applications*, vol. 26, pp. 139-163, 1979.

[5] Repperger, D.W., "On covariance propagation using matrix continued fractions," *International Journal of Systems Science*, vol. 10, pp. 913-925, 1979.

[6] Shieh, L.S. and Chahin, N., "A computer-aided method for the factorization of matrix polynomials," *International Journal of Systems Science*, vol. 12, pp. 305-323, 1981.

[7] Denman, E.D. and Leyva-Ramos, J., "Spectral decomposition of a matrix using the generalized sign matrix," *Applied Mathematics and Computation*, vol. 8, pp. 237-250, 1981.

[8] Matteys, R.I., "Stability analysis via the extended matrix-sign function," *Proceedings of the Inst. Elec. Engrs.*, vol. 125, pp. 241-243, 1978.

[9] Shieh, L.S., Tsay, Y.T. and Yates, R.E., "Some properties of matrix-sign functions derived from continued fractions," *IEE Proceedings Part D*, vol. 130, pp. 111-118, 1983.

[10] Khovanskii, A.N., *The application of continued fractions and their generalizations to problems in approximation theory.* Noordhoff: Netherlands, pp. 182-202, 1963.

[11] Shieh, L.S. and Tsay, Y.T., "Block modal matrices and their applications to multivariable control systems," *IEE Proceedings Part D*, vol. 129, pp. 41-48, 1982.

[12] Tsay, Y.T. and Shieh, L.S., "Irreducible divisors of λ-matrices and their applications to multivariable control systems," *International Journal of Control*, vol. 37, pp. 17-36, 1983.

[13] Tsay, Y.T., Shieh, L.S., Yates, R.E. and Barnett, S., "Block partial-fraction expansion of a rational matrix," *Linear and Multilinear Algebra*, vol. 11, pp. 225-241, 1982.

[14] Davis, P.J., *Circulant Matrices*. New York: Wiley, 1979.

[15] John, Fritz, *Lectures on Advanced Numerical Analysis*, New York: Gordon and Breach, pp. 46-50, 1967.

[16] Buckingham, R.A., *Numerical Methods*. London: Pitman, pp. 277-297, 1957.

[17] Roberts, J.D., "Linear model reduction and solution of the algebraic Riccati equation by use of the sign function," *CUED/B-Control/TR13 Report, Cambridge University*, 1971; also, *International Journal of Control*, vol. 32, pp. 677-687, 1980.

[18] Denman, E.D. and Leyva-Ramos, J., "Computation of roots of real and complex matrices," *Applied Mathematics and Computation*, vol. 16, pp. 213-228, 1985.

[19] Denman, E.D., "Roots of real matrices," *Linear Algebra and Its Applications*, vol. 36, pp. 133-139, 1981.

[20] Shieh, L.S., Tsay, Y.T. and Yates, R.E., "Computation of the principal nth roots of complex matrices," *IEEE Transactions on Automatic Control*, vol. AC-30, pp. 606-608, 1985.

[21] Tsay, Y.T., Shieh, L.S. and Tsai, J.S.H., "A fast method for computing the principal nth root of complex matrices," *Linear Algebra and Its Applications*, vol. 76, pp. 205-221, 1986.

[22] Higham, N.J., "Newton's method for the matrix square root," *Mathematics of Computation*, vol. 46, pp. 537-549, 1986.

[23] Bierman, G.J. "Computational aspects of the matrix-sign function solution to the ARE," *Proceedings of the 23rd Conference on Decision and Control*, pp. 514-519, 1984.

[24] Gardiner, J.D. and Laub, A.J., "A generalization of the matrix-sign function solution for algebraic Riccati equations," *Proceedings of the 24th Conference on Decision and Control*, pp. 1233-1235, 1985.

[25] Shieh, L.S., Wang, C.T. and Tsay, Y.T., "Fast suboptimal state-space self-tuner for linear stochastic multivariable systems," *IEE Proceedings Part D*, vol. 130, pp. 143-154, 1983.

[26] Shieh, L.S., Tsay, Y.T. and Wang, C.T., "Matrix-sector functions and their applications to system theory," *IEE Proceedings Part D*, vol. 131, pp. 171-181, 1984.

[27] Shieh, L.S., Tsai, J.S.H. and Yates, R.E., "The generalized matrix-sector function and the separation of matrix eigenvalues," *IMA Journal of Mathematical Control & Information*, vol. 2, pp. 251-258, 1985.

[28] Shieh, L.S., Tsai, J.S.H. and Lian, S.R., "Determining continuous-time state equations from discrete-time state equations via the principal qth root method," *IEEE Transactions on Automatic Control*, vol. AC-31, pp. 454-457, 1986.

[29] Shieh, L.S., Lian, S.R. and McInnis, B.C., "Fast and stable algorithms for computing the principal square root of a complex matrix," *IEEE Transactions on Automatic Control*, vol. AC-32, pp. 820-823, 1987.

[30] Attarazadeh, F., "Relative stability test for continuous and sampled-data control systems using the generalized sign matrix," *IEE Proceedings Part D*, vol. 129, pp. 189-198, 1982.

[31] Barnett, S., *Matrices in Control Theory.* New York: Van Nostrand Reinhold, 1971; Section Edition, Malabar: Krieger, 1984.

[32] Barnett, S., *Polynomials and Linear Control Systems.* New York: Marcel Dekker, 1983.

[33] Howlands J.L., "The sign matrix and the separation of matrix eigenvalues," *Linear Algebra and Its Applications*, vol. 49, pp. 221-229, 1983.

[34] Mattheys, R.L., "Stability analysis via the extended matrix-sign functions," *IEE Proceedings*, vol. 125, pp. 241-249, 1978.

[35] Shieh, L.S., Chen, C.H. and Yates, R.E., "Separation of matrix eigenvalues and block-diagonalization of a system matrix," *U.S. Army Research Report, Department of Electrical Engineering, University of Houston*, 1983.

[36] Shieh, L.S. and Tsay, Y.T., "Algebra-geometric approach for the model reduction of large-scale multivariable systems," *IEE Proceedings Part D*, vol. 131, pp. 23-30, 1984.

[37] Shieh, L.S., Tsay, Y.T., Lin, S.W. and Coleman, N.P., " Block-diagonalization and block-triangularization of a matrix via the matrix-sign function," *International Journal of Systems Science*, vol. 15, pp. 1203-1213, 1984.

[38] Sinha, N.K. and Kuszta, B., *Modelling and Identification of Dynamic Systems.* New York: Van Nostrand Reinhold, 1983.

[39] Houpis C.H. and Lamont, G.B., *Digital Control Systems.* New York: McGraw-Hill, 1985.

[40] Koupis, C.H., "Refined design method for sampled-data control systems: The pseudo-continuous-time control-system design," *IEE Proceedings Part D*, vol. 132, pp. 69-74, 1985.

[41] Frankline, G.F. and Power, J.D., *Digital Control of Dynamical Systems*. Reading: Addison-Wesley, 1980.

[42] Cadzow, J.A., *Discrete-time Systems*. Englewood Cliffs: Prentice-Hall, 1973.

[43] Shieh, L.S., Wang, H. and Yates, R.E., "Discrete-continuous model conversion," *Applied Mathematics on Modelling*, vol. 4, pp. 449-455, 1980.

[44] Sinha, N.K. and Lastman, G.J., "Transformation algorithm for identification of continuous-time multivariable systems from discrete data," *Electron. Lett.*, vol. 17, pp. 779-780, 1981.

[45] Puthenpura, S. and Sinha, N.K., "Transformation of continuous-time model of a linear multivariable system from its discrete-time model," *Electron. Lett.*, vol. 20, pp. 737-738, 1984.

[46] Harris, E.L., "Using discrete models with continuous design packages," *Automatica*, vol. 15, pp. 97-99, 1979.

[47] Wall, H.S., *Analytic Theory of Continued Fractions*. New York: Van Nostrand-Reinhold, 1948.

[48] Bellman, R., *Introduction to Matrix Analysis*. New York: McGraw-Hill, p. 226, 1970.

[49] Bjorck, A. and Hammarling, S., "A Schur method for the square root of a matrix," *Linear Algebra and Its Applications*, vol. 52/53, pp. 127-140, 1983.

[50] Dongarra, J.J., Bunch, J.R., Moler, C.B. and Stewart, G.W., *LINPACK User's Guide*. Philadelphia: SIAM, 1979.

[51] Gantmacher, F.R., *Theory of Matrices*. New York: Chelsea, vol. 1, 1959.

[52] Householder, A.S., *The Theory of Matrices in Numerical Analysis*. New York: Dover, 1964.

[53] Atkinson, K.E., *An Introduction to Numerical Analysis*. New York: Wiley, 1978.

[54] Wilkinson, J.H., *The Algebraic Eigenvalue Problem*. Oxford: Clarendon, 1965.

[55] Aoki, M., "Control of large-scale dynamic systems by aggregation," *IEEE Transactions on Automatic Control*, vol. AC-13, pp. 246-253, 1968.

[56] Kokotovic, P.V., O'Malley, R.E. and Sannuti, P., "Singular perturbations and order reduction in control theory - an overview," *Automatica*, vol. 12, pp. 123-132, 1976.

[57] Porter, B. and Crossley, R., *Modal Control - Theory and Applications.* London: Taylor and Francis, 1972.

[58] Mahmoud, M.S., Chen, Y. and Singh, M.G., "On eigenvalue-assignment in discrete systems with fast and slow models," *International Journal of Systems Science*, vol. 16, pp. 61-70, 1985.

[59] Mahmoud, M.S., Chen, Y. and Singh, M.G., "Discrete two-time scale systems," *International Journal of Systems Science*, vol. 17, pp. 1187-1207, 1986.

[60] Naidu, D.S. and Price, D.B., "Time-scale synthesis of a closed-loop discrete optimal control system," *Journal of Guidance, Control & Dynamics*, vol. 32, pp. 417-422, 1987.

[61] Tsai, J.S.H., Shieh, L.S. and Yates, R.E., "Fast and stable algorithms for computing the prinicpal nth root of a complex matrix and the matrix-sector function," *International Journal of Computers and Mathematics with Applications*, 1988 (in press).

[62] Anderson, B.D.O. and Moore, J.B., *Linear Optimal Control.* Englewood Cliffs: Prentice-Hall, 1971.

[63] Ackermann, J., *Sampled-data Control Systems.* Berlin: Mercedes-Druck, p. 234, 1985.

[64] Shieh, L.S., Dib, H.M. and McInnis, B.C., "Linear quadratic regulators with eigenvalue-placement in a vertical strip," *IEEE Transactions on Automatic Control*, vol. AC-31, pp. 241-243, 1986.

[65] Shieh, L.S., Dib, H.M. and Ganesan, S., "Linear quadratic regulators with eigenvalue-placement in a horizontal strip," *International Journal of Systems Science*, vol. 19, pp. 1279-1290, 1987.

[66] Kawasaki, N. and Shimemura, E., "Determination of quadratic weighting matrices to locate poles in a specified region," *Automatica*, vol. 19, pp. 557-560, 1983.

[67] Shieh, L.S., Dib, H.M. and Ganesan, S., "Continuous-time quadratic regulators and pseudo-continuous-time quadratic regulators with pole placement in a specific region," *IEE Proceedings Part D*, vol. 134, pp. 338-346, 1987.

[68] Moler, C. and Loan, C.V., "Nineteen dubious ways to compute the exponential of a matrix," *SIAM Rev.*, vol. 20, pp. 801-836, 1978.

[69] Chen, C.T., *Linear System Theory and Design.* New York: Holt, Reinhart and Winston, 1984.

[70] VanLandingham, H.F., *Introduction to Digital Control Systems.* New York: Macmillan, p. 287, 1985.

[71] Houpis, C.H., "Refined design method for sampled-data control systems: the pseudo-continuous-time control-system design," *IEE Proceedings Part D*, vol. 132, pp. 69-74, 1985.

[72] Shimemura, E. and Fujita, M., "A design method for linear state-feedback systems possessing integrity based on a solution of a Riccati-type equation," *International Journal of Control*, vol. 42, pp. 887-899, 1985.

[73] Gutman, S. and Jury, E.I., "A general theory for matrix root-clustering in subregions of the complex plane," *IEEE Transactions on Automatic Control*, vol. AC-26, pp. 853-862, 1981.

[74] Zeheb, E. and Hertz, D., "Complete root distribution with respect to parabolas and some results with respect to hyperbolas and sectors," *International Journal of Control*, vol. 36, pp. 517-526, 1982.

[75] Shieh, L.S., Tsai, J.S.H. and Coleman, N.P., "The rectangular and polar representations of a complex matrix," *International Journal of Systems science*, vol. 18, no. 10, pp. 1825-1838, 1987.

Solution of Riccati Equation Via Matrix-sign Function

The Riccati equation for the controllable continuous-time system (A, B) with weighting matrices $Q(\geq 0)$ and $R(> 0)$ is given by

$$PBR^{-1}B^TP - A^TP - PA - Q = 0. \qquad (A.1a)$$

The steady state solution of this Riccati equation, $P(\geq 0)$ with (Q, A) detectable, can be easily computed using the properties of the matrix-sign function [9,23]. Consider the Hamiltonian associated with the given system,

$$H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix}. \qquad (A.1b)$$

The following algorithm can be utilized to obtain the solution P,

$$H_{k+1} = (1/2)\,[H_k + H_k^{-1}],\ H_0 = H, \qquad \text{and}$$

$$\lim_{k\to\infty} H_k = \text{Sign}(H). \qquad (A.2a)$$

Let

$$\text{Sign}^+(H) \triangleq (1/2)[I_{2n} + \text{Sign}(H)]. \qquad (A.2b)$$

Construct a block-modal matrix X as

$$X = [\text{ind}\,(\text{Sign}^+(H)),\ \text{ind}\,(I_{2n} - \text{Sign}^+(H))] \triangleq \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \qquad (A.3a)$$

where ind(.) represents the collection of the linearly independent column vectors of (.). Then, we have

$$P = X_{22}(X_{12})^{-1} = X_{21}(X_{11})^{-1}. \qquad (A.3b)$$

To alleviate the problems of computing H_k^{-1}, the Hamiltonian can be transformed into a symmetric form as follows [23],

$$\hat{H} = \hat{J}H = \begin{bmatrix} 0_n & -I_n \\ I_n & 0_n \end{bmatrix} H = \begin{bmatrix} Q & A^T \\ A & -BR^{-1}B^T \end{bmatrix}. \qquad (A.4a)$$

Then, the algorithm in (A.2) becomes

$$\hat{H}_{k+1} = (1/2)\,[\hat{H}_k + \hat{J}\hat{H}_k^{-1}\hat{J}],\ \hat{H}_0 = \hat{J}H, \qquad \text{and}$$

$$\lim_{k\to\infty}(-\hat{J}\hat{H}_k) = \text{Sign}(H). \tag{A.4b}$$

The computation of the inverse of the symmetric matrix $\hat{H}_k$ is much simpler than computing the inverse of H_k. The Riccati solution P is again given by (A.3).

The HK Singular Value Decomposition

L. Magnus Ewerbring and Franklin T. Luk

School of Electrical Engineering
Phillips Hall
Cornell University
Ithaca, New York 14853, USA

Abstract

In this paper we consider a generalization of the singular value decomposition (SVD) that involves three matrices. We show how the decomposition can be used in important applications such as weighted least squares, and present a new computational procedure based on an implicit SVD method for a triple matrix product. Our algorithm is well suited for parallel implementation.

Keywords: Singular value decomposition, weighted least squares, Jacobi methods, parallel computing

1. Introduction

In this paper we develop a new algorithm for computing the HK-singular value decomposition (HK-SVD). The paper is organized as follows. Section 1 presents a description of the problem, its relation to the generalized singular value decomposition (GSVD), and an application in which the HK-SVD provides a powerful solution. What follow in Section 2 are an implicit algorithm for computing the SVD of a product of three matrices and a new HK-SVD algorithm in which the implicit method is embedded. A summary and some final remarks conclude the paper in Section 3.

Notations. We make the standard choice to represent column vectors by bold lower case roman characters, matrices by upper case roman characters, and scalars by either greek letters or roman letters with subscripts, as elements in vectors and matrices. In addition, the following notations are used:

$\mathbf{0}_{n \times p} \rightarrow$ an $n \times p$ block matrix with all zero elements

$I_p \rightarrow$ a p-dimensional identity matrix

$A^{i,i+1} \rightarrow$ a 2×2 matrix formed by intersecting rows and columns i and $i+1$ of A

1.1. HK-SVD. Given three matrices A $(n \times p)$, H $(n \times n)$ and K $(p \times p)$, where $n \geq p$ and both H and K are symmetric positive definite, we wish to find two transformations Y and Z such that

$$Y^{-1}AZ = D \quad , \tag{1.1}$$

where

$$Y^T H Y = I_n \quad \text{and} \quad Z^T K Z = I_p \quad ,$$

and D $(n \times p)$ is diagonal (Van Loan [10]). We say that the matrices Y and Z are H-orthogonal and K-orthogonal, respectively.

A straightforward way [10] to compute the HK-SVD is to first determine the Cholesky factorizations:

$$H = R_H^T R_H \quad , \quad K = R_K^T R_K \quad , \tag{1.2}$$

where R_H and R_K are upper triangular matrices, and then find an SVD of the product $R_H A R_K^{-1}$:

$$U^T (R_H A R_K^{-1}) V = D \quad , \tag{1.3}$$

where U and V are orthogonal, and D is diagonal. The two transformations Y and Z are given by

$$Y = R_H^{-1} U \quad , \quad \text{and} \quad Z = R_K^{-1} V \quad . \tag{1.4}$$

We will present a new algorithm for finding the HK-SVD via equation (1.3) without any explicit matrix inversions or product formations.

1.2. Weighted Least Squares. The HK-SVD is useful in finding the solution to a weighted least squares problem:

$$||A\mathbf{x}-\mathbf{b}||_H = \min \quad \text{s.t.} \quad ||\mathbf{x}||_K = \min \ . \tag{1.5}$$

The M-vector norm is defined by

$$||\mathbf{y}||_M^2 \equiv \mathbf{y}^T M\mathbf{y} \ ,$$

where M denotes a symmetric and positive definite matrix. We may reformulate the problem as

$$||A\mathbf{x}-\mathbf{b}||_H = ||R_H(A\mathbf{x}-\mathbf{b})||_2 = ||R_H A R_K^{-1}(R_K\mathbf{x}) - R_H\mathbf{b}||_2 = \min \ ,$$

subject to

$$||R_K\mathbf{x}||_2 = \min \ .$$

The procedure is to compute an SVD:

$$U^T(R_H A R_K^{-1})V = D \ , \tag{1.6}$$

and solve the simple problem:

$$||D\mathbf{w}-\mathbf{f}||_2 = \min \quad \text{s.t.} \quad ||\mathbf{w}||_2 = \min \ , \tag{1.7}$$

where

$$\mathbf{w} = V^T R_K\mathbf{x} = Z^{-1}\mathbf{x} \quad \text{and} \quad \mathbf{f} = U^T R_H\mathbf{b} = Y^{-1}\mathbf{b} \ ,$$

with Y and Z as defined in (1.4). We see that the HK-SVD provides an easy solution to the weighted least squares problem (1.5).

1.3 Previous Work. Our new algorithm is based on an implicit GSVD method (Paige [9] and Luk [5]), which computes an SVD of a $p \times p$ product AB^{-1}, without explicitly forming the product and without inverting B.

The SVD of a matrix product finds applications in many areas. For instance, it can be used in control theory to compute system balancing transformations (cf. Moore [8], Heath et al. [3] and Laub et al. [4]). That is, we find a contragradient transformation P to diagonalize two given symmetric positive definite matrices A and B:

$$P^T AP = P^{-1}BP^{-T} = \Lambda \ .$$

The transformation thus solves the generalized eigenvalue problem:

$$AB\mathbf{x} = \lambda^2 \mathbf{x} \quad .$$

One way to find P is to compute the Cholesky decomposition of B, i.e., $B = R_B^T R_B$, and then calculate an eigenvalue decomposition of the symmetric matrix

$$U^T (R_B A R_B^T) U = \Lambda^2 \quad . \tag{1.8}$$

We get the transformation as

$$P = R_B^T U \Lambda^{-1/2} \quad .$$

Despite the similarity of equation (1.8) and (1.3), since A is symmetric positive definite here, we may compute the Cholesky decomposition of $A = R_A^T R_A$ and find an SVD of $R_A R_B^T$ [3], [4]. For equation (1.3), however, A is not symmetric, and so we must consider an SVD of three matrices even when $H = K$.

2. New Algorithm

In this section we derive a new algorithm for finding the HK-SVD via equation (1.3). First, we consider the special product:

$$EFG^{-1} \quad , \tag{2.1}$$

where E, F and G are all $p \times p$ and upper triangular. We assume further that G^{-1} exists. The special case of (2.1) where $E = I_p$ reduces to the GSVD problem for the two matrices F and G.

In a Jacobi SVD algorithm we solve a sequence of 2×2 problems by finding rotation parameters to annihilate off-diagonal elements. An important issue is the order of elimination. Luk [6] chooses the *odd–even* ordering and *outer–rotations* for an efficient parallel implementation. The convergence of this scheme has been proved (Luk and Park [7]), and the algorithm implemented on a massively parallel machine (Ewerbring and Luk [1], [2]). We define the odd and even index sets by

$$Odd\text{–}set \rightarrow \{1, 3, 5, \cdots, p-1\}$$

$$Even\text{-}set \rightarrow \{2, 4, 6, \cdots, p-2\}$$

assuming that p is even. For an odd p, we define

$$Odd\text{-}set \rightarrow \{1, 3, 5, \cdots, p-2\}$$
$$Even\text{-}set \rightarrow \{2, 4, 6, \cdots, p-1\}.$$

2.1. GSVD. The GSVD of F and G is computed via an SVD of the matrix

$$C = FG^{-1}. \tag{2.2}$$

The procedure [5], [9] determines orthogonal transformations U, V and Q so that the two resulting matrices U^TFQ and V^TGQ have parallel rows, i.e.,

$$U^TFQ = D \cdot V^TGQ,$$

where D is some diagonal matrix. We can easily check that

$$U^T(FG^{-1})V = D \quad , \tag{2.3}$$

which is just an SVD of C.

The special advantage of Luk's approach [5] is that it preserves the upper triangular structure of both F and G. Indeed, the two matrices G^{-1} and C are also upper triangular. Consider a transformation in the $(i,i+1)$ plane, and denote by $M^{i,i+1}$ the 2×2 matrix formed by intersecting rows $i, i+1$ and columns $i, i+1$ of a $p \times p$ matrix M. Being triangular, the two matrices G and C satisfy these special relations:

$$(G^{-1})^{i,i+1} = (G^{i,i+1})^{-1},$$
$$C^{i,i+1} = F^{i,i+1}(G^{-1})^{i,i+1}.$$

The nonsingularity of $G^{i,i+1}$ follows trivially from the nonsingularity and the triangularity of G. We have thus proved that

$$C^{i,i+1} = F^{i,i+1}(G^{i,i+1})^{-1}, \tag{2.4}$$

the key condition for an implicit computation of an SVD of C. So, let $U^{i,i+1}$ and $V^{i,i+1}$ denote rotations for a 2×2 SVD :

$$(U^{i,i+1})^T C^{i,i+1} V^{i,i+1} = S \ ,$$

where S is diagonal. We have

$$(U^{i,i+1})^T F^{i,i+1} = S \cdot (V^{i,i+1})^T G^{i,i+1} \ ,$$

i.e., the two rows of $(U^{i,i+1})^T F^{i,i+1}$ and $(V^{i,i+1})^T G^{i,i+1}$ are parallel. Hence we can find one single rotation, say $Q^{i,i+1}$, that will triangularize both 2×2 matrices $F^{i,i+1}$ and $G^{i,i+1}$ [5], [9].

How do these transformations affect the $p \times p$ upper triangular matrices F and G? We have

$$F \leftarrow U_{i,i+1}^T \, F \, Q_{i,i+1} \ ,$$
$$G \leftarrow V_{i,i+1}^T \, G \, Q_{i,i+1} \ ,$$

where $U_{i,i+1}$, $V_{i,i+1}$ and $Q_{i,i+1}$ denote appropriate $p \times p$ rotations in the $(i, i+1)$-plane. Note that both $p \times p$ matrices $U_{i,i+1}^T F$ and $V_{i,i+1}^T G$ have only one non-zero subdiagonal element each, in the $(i+1, i)$-position. These two extraneous elements are annihilated by the same rotation $Q_{i,i+1}$.

2.2. Algorithm PSVD. We extend the GSVD algorithm to the general case where $E \neq I_p$. First, consider the product (2.1). Define

$$C = EFG^{-1} \quad \text{and} \quad H = EF \ , \tag{2.5}$$

even though we never intend to explicitly form either product. Once again, focus on a 2×2 problem that lies on the diagonal:

$$C^{i,i+1} = E^{i,i+1} F^{i,i+1} (G^{i,i+1})^{-1}$$
$$= H^{i,i+1} (G^{i,i+1})^{-1} \ .$$

We find two rotations, say $U^{i,i+1}$ and $V^{i,i+1}$, to diagonalize the matrix $C^{i,i+1}$. The rotations are applied to $H^{i,i+1}$ and $G^{i,i+1}$:

$$H^{i,i+1} \leftarrow (U^{i,i+1})^T H^{i,i+1} \ ,$$
$$G^{i,i+1} \leftarrow (V^{i,i+1})^T G^{i,i+1} \ .$$

From previous discussions we learn that we can find one rotation $Q^{i,i+1}$ to restore both matrices to triangular forms:

$$H^{i,i+1} \leftarrow (U^{i,i+1})^T H^{i,i+1} Q^{i,i+1} ,$$

$$G^{i,i+1} \leftarrow (V^{i,i+1})^T G^{i,i+1} Q^{i,i+1} .$$

Naturally, we want to rotate $E^{i,i+1}$ and $F^{i,i+1}$ individually, and not their product $H^{i,i+1}$:

$$E^{i,i+1} \leftarrow (U^{i,i+1})^T E^{i,i+1} ,$$

$$F^{i,i+1} \leftarrow F^{i,i+1} Q^{i,i+1} .$$

The fact that $H^{i,i+1}$ stays upper triangular means that another single rotation $W^{i,i+1}$ can be applied to maintain the triangularity of both $E^{i,i+1}$ and $F^{i,i+1}$:

$$E^{i,i+1} \leftarrow (U^{i,i+1})^T E^{i,i+1} W^{i,i+1} , \tag{2.6}$$

and

$$F^{i,i+1} \leftarrow (W^{i,i+1})^T F^{i,i+1} Q^{i,i+1} .$$

We summarize our algorithm as follows.

Algorithm PSVD.

do until convergence

alternate between $i \in$ *Odd-set* and *Even-set*
begin
{ $U_{i,i+1}$ and $V_{i,i+1}$ are "outer rotations" }
determine $U_{i,i+1}$ and $V_{i,i+1}$ to
annihilate $c_{i,i+1}$ and $c_{i+1,i}$;

$E \leftarrow U_{i,i+1}^T E$; $G \leftarrow V_{i,i+1}^T G$;

$U \leftarrow U U_{i,i+1}$; $V \leftarrow V V_{i,i+1}$;

find $Q_{i,i+1}$ to zero out $h_{i+1,i}$ and $g_{i+1,i}$;

$F \leftarrow F Q_{i,i+1}$; $G \leftarrow G Q_{i,i+1}$;

find $W_{i,i+1}$ to zero out $e_{i+1,i}$ and $f_{i+1,i}$;

$E \leftarrow E W_{i,i+1}$; $F \leftarrow W_{i,i+1}^T F$;

end. □

By convergence we mean that the matrix C has converged to a diagonal form $D = diag(\gamma_i)$, with

$$\gamma_i = e_{ii}\hat{f}_{ii}/g_{ii} \quad .$$

The matrices of left and right singular vectors are given by U and V, respectively.

2.3. Algorithm HK-SVD. As described in (1.3), for computing the HK-SVD, we need to find an SVD of the matrix product

$$C = R_H A R_K^{-1} \quad . \tag{2.7}$$

To make use of the implicit algorithms of Section 2.2, we must reduce C to a product of upper triangular matrices. To accomplish this, compute the QR decomposition (QRD) of A:

$$A = Q_A R_A \quad ,$$

where

$$R_A = \begin{bmatrix} \hat{R}_A \\ \mathbf{0}_{(n-p)\times p} \end{bmatrix} \quad ,$$

and $\hat{R}_A$ denotes a $p \times p$ upper triangular matrix. We get

$$R_H A R_K^{-1} = (R_H Q_A) R_A R_K^{-1} \quad .$$

Another QRD is performed, this time on the matrix product $R_H Q_A$:

$$R_H Q_A = Q_{\hat{R}} R_{\hat{H}} \quad . \tag{2.8}$$

Thus, the problem has been reduced to that of finding an SVD of the product

$$\hat{C} = R_{\hat{H}} R_A R_K^{-1} \quad , \tag{2.9}$$

where $R_{\hat{H}}$ is $n \times n$, R_A is $n \times p$ and R_K is $p \times p$. So, we need to handle the different dimensions. For $n > p$, the last $n - p$ rows and columns of $R_{\hat{H}}$ can be discarded because the last $n - p$ rows of R_A are zero. Hence, set

$$E \leftarrow \hat{I}_p^T R_{\hat{H}} \hat{I}_p \quad , \quad F \leftarrow \hat{I}_p^T \hat{R}_A \quad , \quad G \leftarrow R_K \quad , \tag{2.10}$$

where

$$\hat{I}_p = \begin{bmatrix} I_p \\ \mathbf{0}_{(n-p)\times p} \end{bmatrix} ,$$

and compute the diagonalization of the product EFG^{-1} using Algorithm PSVD. Finally, set the $n \times n$ matrix of left singular vectors to be

$$Q_{\hat{R}} \begin{bmatrix} U & \mathbf{0}_{p\times(n-p)} \\ \mathbf{0}_{(n-p)\times p} & I_{n-p} \end{bmatrix} ,$$

to account for the QRD of (2.8). We thus obtain our new algorithm:

Algorithm HK-SVD.

compute Cholesky factorizations:

$$H = R_H^T R_H; \quad K = R_K^T R_K;$$

compute QR decomposition of A:

$$A = Q_A R_A;$$

transform the matrix R_H:

$$R_H \leftarrow R_H Q_A;$$

compute QR decomposition of R_H:

$$R_H = Q_{\hat{H}} R_{\hat{H}};$$

set

$$E \leftarrow \hat{I}_p^T R_{\hat{H}} \hat{I}_p \quad , \quad F \leftarrow \hat{I}_p^T \hat{R}_A \quad , \quad G \leftarrow R_K \quad ;$$

use Algorithm PSVD to find an SVD of EFG^{-1}. □

3. Final Remarks

This paper presents an implicit algorithm for computing the SVD of a product of three matrices. The algorithm plays an integral role in the new

method in Section 2.3 for computing the HK-SVD. The applicability of the algorithm was exemplified in the solution to a weighted least squares problem, which, for instance, arises in a specific aircraft problem.

All problems in the paper call for the diagonalization of a product of three, not necessarily symmetric, matrices. The extension of our methods to a product of more matrices is straightforward. Although we assume that the inverses in (2.1) exist, our algorithms can easily be adapted for rank deficiency by using matrix adjoints (cf. Paige [9]).

Our new algorithm was simulated on a VAX 11/750 using MATLAB. It is well for a massively parallel computer; Ewerbring and Luk [1], [2], presented implementations of the SVD and GSVD methods described in this paper on the 65,536 processor Connection Machine.

Acknowledgements

This work was supported in part by the SDIO/IST and managed by the Army Research Office under contract DAAL 03-86-K0109. L.M. Ewerbring acknowledges travel support by the Mathematical Sciences Institute of Cornell University.

References

[1] L. M. Ewerbring and F. T. Luk, *Computing the singular value decomposition on the Connection Machine*, Proc. Internat. Workshop on SVD and Signal Processing, Les Houches, France 1987.

[2] L. M. Ewerbring and F. T. Luk, *Almost linear time matrix operations on the Connection Machine*, Proc. SPIE, High Speed Computing, vol. 880 (1988), pp. 198-205.

[3] M.T. Heath, A.J. Laub, C.C. Paige and R.C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM J. Sci. Statist. Comput., vol. 7 (1986), pp. 1147-1159.

[4] A. J. Laub, M. T. Heath, C. C. Paige and R. C. Ward, *Computation of system balancing transformations and other applications of simultaneous*

diagonalization algorithms, IEEE Trans. Automatic Control, vol. AC-32, No. 2 (1987), pp. 115-122.

[5] F.T. Luk, *A parallel algorithm for computing the generalized singular value decomposition*, J. Parallel Distrib. Comput., vol. 2 (1985), pp. 250-260.

[6] F.T. Luk, *A triangular processor array for computing singular values*, Lin. Alg. Applics., vol. 77, (1986), pp. 259-273.

[7] F.T. Luk and H. Park, *A proof of convergence for two parallel Jacobi SVD algorithms*, IEEE Trans. Computers, vol. C-37 (1988), to appear.

[8] B. Moore, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automatic Control, vol. AC-26, No. 1. (1981), pp. 17-32.

[9] C.C. Paige, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., vol. 7 (1986), pp. 1126-1146.

[10] C.F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., vol. 13 (1976), pp. 76-83.

THE ADMISSIBILITY OF A GENERALIZATION OF A^*

James W. Lark, III and Chelsea C. White, III
Department of Systems Engineering
University of Virginia
Charlottesville, Virginia 22901

ABSTRACT. We present an algorithm, called A^G, for finding the least-cost path from start node to goal node set in an OR-graph, where arc costs are scalar-valued and the cost of each path is the sum of the concomitant arc costs. Search is guided by a set, H, of real-valued functions on the node set. If $H = \{h: \ell \leq h\}$ for given function ℓ, then A^G essentially becomes A^*. If H is bounded, then successors of the newly expanded node may not be placed on OPEN. We address the issue of admissibility. A new concept, the completeness of a heuristic set with respect to a path in the graph, is introduced.

INTRODUCTION. In this paper, we present a generalization of A^*, which we call A^G. The key characteristic that distinguishes A^G from A^* is that knowledge used to guide A^G is represented by a set of heuristic functions, or a heuristic set, rather than by a single heuristic function (or more precisely, a specially structured heuristic set induced by a single heuristic function). A key result of this characteristic is that it may not be necessary to place on OPEN all the successor nodes of a node chosen for expansion. A possible implication of this result is that the OPEN set will tend to be smaller and hence easier to store and to sort.

There are at least three reasons for allowing knowledge to be represented by a set of heuristic functions in order to guide search. First, more information about the perfect heuristic may be available than just a lower bound, and this information may be such that it can be represented by set inclusion. Second, it seems reasonable that more (or better) information for search guidance would not degrade the quality of the search procedure, although this may not always be true in general (White and Harrington, 1980). Third, upper and lower bound information has proven very useful in action elimination algorithms for Markov decision processes (e.g., Puterman and Shin, 1982), a problem formulation of particular interest to us.

The outline of this paper and its results follow the basic outline of Section 3.1 (Pearl, 1984). We begin by defining the problem of interest and

Acknowledgement: This research has been supported by U.S. Army Research Office and the National Science Foundation.

setting terminology. The A^G algorithm is presented in Section 2. Termination and the completeness of a heuristic set, a new concept, are the topics of Section 3. Section 4 is concerned with admissibility.

Future research will involve comparing A^G with different admissible heuristic sets and investigating the computational significance of A^G.

1. PROBLEM DEFINITION and TERMINOLOGY

Let N represent the countable set of nodes in the OR graph. The set $A \subseteq N \times N$ is the set of directed arcs. Node $s \in N$ represents the start node; the finite set $\Gamma \subseteq N$, having generic element γ, represents the goal node set. We let $G = (N,A,s,\Gamma)$ designate the graph under consideration.

Let SCS: $N \to 2^N$ be the successor set function, where SCS(n) represents the set of all nodes $n' \in N$ such that $(n,n') \in A$. We assume throughout that SCS(n) is finite for all $n \in N$.

A path $P = (n_1,\ldots,n_K)$ is a sequence of nodes such that $n_{k+1} \in SCS(n_k)$ for all $k = 1,\ldots,K-1$. Let P(n,S) be the set of all finite length, acyclic paths from $n \in N$ to $S \subseteq N$. Notationally, if S is a singleton, i.e., if $S = \{n'\}$, then we will write $P(n,S) = P(n,n')$.

The function c: $A \to R$ is the arc cost function; the cost assigned to a path is assumed to be the sum of the concomitant arc costs. Throughout, we assume that there is a constant $\delta > 0$ such that $\delta \leq c(\alpha)$ for all $\alpha \in A$. Notationally, we will often replace $c(\alpha)$ with $c(n,n')$, where $\alpha = (n,n')$.

The problem objective is to find a minimum cost path in $P(s,\Gamma)$. Let $P^*(n,S) \subseteq P(n,S)$ represent the set of all optimal, i.e., minimal cost, paths from $n \in N$ to $S \subseteq N$. Thus, we seek a path in $P^*(s,\Gamma)$.

Heuristic information will prove useful in meeting our objective. We assume that this information is represented in set form. Specifically, let H be the set of all real-valued functions on N. We call a given subset $H \subseteq H$ the heuristic set. We will assume that search for a path in $P^*(s,\Gamma)$ is guided by a given heuristic set. This is in contrast to the heuristic search procedure A^*, which assumes that search is guided by a given heuristic function, i.e., an element, rather than a subset, of H.

Several functions in H will prove to be important in developments to follow. Let g be the current path cost function, where g(n) represents the cost of the current path from s to n and where $g(s) = 0$. The function g^* is such that $g^*(n)$ represents the minimal cost of paths going from s to n. For given heuristic set H, let ℓ, the lower bound function of H, be defined as

$\ell(n) = \inf\{h(n):h \in H\}$ for all $n \in N$. Define f as $f(n) = g(n) + \ell(n)$ for all $n \in N$. Also, let h^* represent the perfect heuristic function, which must satisfy the following dynamic program:

$$h^*(n) = \min \{c(n,n') + h^*(n'):n' \in SCS(n)\}$$

$$h^*(\gamma) = 0, \ \gamma \in \Gamma$$

$$h^*(n) = \infty \text{ if } SCS(n) \text{ is empty.}$$

We let C^* represent the minimal cost of paths going from s to Γ. Thus, $C^* = h^*(s)$. Note that our objective is to determine a path in $P(s,\Gamma)$ having cost C^*.

Let U: $A \times 2^H \rightarrow R$ be called the node expansion function, which we define as:

$$U(n,n',H) = \sup \{h(n) - h(n'):h \in H\}.$$

2. THE A^G ALGORITHM

We now state the A^G algorithm:

0. Initialization. Set OPEN equal to the set containing only the start node and set CLOSED to the empty set.

1. If OPEN is empty, then terminate with failure.

2. Remove from OPEN and place on CLOSED a node n for which $f(n) = g(n) + \ell(n)$ is minimum with respect to all nodes in OPEN.

3. If n is a goal node, then trace through backpointers from n to s to determine the solution path and terminate successfully.

4. If n is not a goal node, generate its successors. If n has no successors, then go to Step 5. Otherwise, for all successors n' of n, compute $U(n,n',H)$.

   a. If $n' \notin \text{OPEN} \cup \text{CLOSED}$ and $U(n,n',H) \geq c(n,n')$, then add n' to OPEN and add a backpointer from n' to n.

   b. If $n' \notin \text{OPEN} \cup \text{CLOSED}$ and $U(n,n',H) < c(n,n')$, then go to Step 5.

   c. If $n' \in \text{OPEN} \cup \text{CLOSED}$ and $U(n,n',H) \geq c(n,n')$, then direct its pointers along the path yielding the lowest g(n') and put n' on OPEN if pointer adjustment was required.

d. If $n' \in OPEN \cup CLOSED$ and $U(n,n',H) < c(n,n')$, then go to Step 5.

5. Increment the iteration counter and go to Step 1.

Step 4b represents the major new feature of A^G, relative to A^*. Justification for this step is as follows. Assume $h^* \in H$, a condition on H that we will later refer to as admissibility. Then, $U(n,n',H) < c(n,n')$ implies that $h^*(n) - h^*(n') < c(n,n')$, or equivalently

$$h^*(n) < c(n,n') + h^*(n').$$

It then follows from the dynamic programming equation describing h^* that n' is not the minimizing element in SCS(n) and hence n' is not on an optimal path from n to Γ. Thus, in searching for a path in $P^*(s,\Gamma)$, it will never be useful to consider a path in $P(s,\Gamma)$ containing arc (n,n').

The heuristic function providing guidance to A^* is said to be admissible if it represents a lower bound on the perfect heuristic. It is therefore natural to think of heuristic functions and lower bound functions as being analogous. Let $H = \{h \in H : \ell \leq h\}$ be the heuristic set induced by the lower bound function ℓ. Then $U(n,n',H) \geq c(n,n')$ for all $(n,n') \in A$, and A^G essentially becomes A^*. Thus, we consider the concept of a heuristic set to be a generalization of the concept of a heuristic function and hence that A^G is a generalization of A^*.

3. TERMINATION and COMPLETENESS

Assumptions on SCS and c insure the following result. Proof is a straightforward adaptation of the concomitant result for A^* (see pp. 76-77 in Pearl, 1984).

THEOREM 1. A^G terminates after a finite number of iterations.

We now present a sufficient condition for A^G to be complete, i.e., to terminate with a path in $P(s,\Gamma)$, assuming $P(s,\Gamma)$ is not empty.

DEFINITION. The heuristic set H is complete with respect to the path $(n_1,\ldots,n_K)$ if $U(n_k,n_{k+1},H) \geq c(n_k,n_{k+1})$ for all $k = 1,\ldots,K-1$.

Let H be the heuristic set induced by the bounded lower bound function ℓ. Then $U(n,n',H) \geq c(n,n')$ for all $(n,n') \in A$, and hence the heuristic set induced by any bounded lower bound function is trivially complete with respect to any path in the graph. We remark that this fact eliminates the need to define complete heuristic functions for the A^* algorithm.

THEOREM 2. For infinite graph G, assume that the heuristic set H is complete with respect to a path in $P(s,\Gamma)$. Then A^G is complete on G.

Proof: The completeness of H insures that at least one node from at least one solution path is always OPEN prior to termination. The result then follows as for A^*; e.g., see the proof of Theorem 1, p. 77, in Pearl, 1984.

□

4. ADMISSIBILITY

We now present a condition which will insure that A^G is admissible, i.e., A^G will terminate with a path in $P^*(s,\Gamma)$, assuming $P^*(s,\Gamma)$ is nonempty.

DEFINITION: The heuristic set H is admissible if $h^* \in H$.

Note that if H is admissible, then $\ell \leq h^*$, where ℓ is the lower bound function induced by H. An important relationship between heuristic set admissibility and completeness is now presented.

LEMMA 1. Assume that the heuristic set H is admissible. Then H is complete with respect to every path in $P^*(s,\Gamma)$.

Proof: Let $(n_1,\ldots,n_K) \in P^*(s,\Gamma)$. Then, there exists an $h \in H$, namely h^*, such that:

$$h^*(n_k) = c(n_k,n_{k+1}) + h^*(n_{k+1})$$

for all $k = 1,\ldots,K-1$, and hence

$$h^*(n_k) - h^*(n_{k+1}) = c(n_k,n_{k+1})$$

for all $k = 1,\ldots,K-1$. Therefore, for all $k = 1,\ldots,K-1$, $U(n_k,n_{k+1},H) \geq c(n_k,n_{k+1})$. □

Let $P^c(n,S,H)$ be the set of all paths in $P(n,S)$ for which H is complete. The following result is then a corollary to Lemma 1.

LEMMA 2. Assume H is admissible. Then, $P^*(s,\Gamma) \subseteq P^c(s,\Gamma,H) \subseteq P(s,\Gamma)$.

We observe that if $H = \{h \in H: \ell \leq h\}$ for some bounded heuristic ℓ, then $P^c(s,\Gamma,H) = P(s,\Gamma)$, and if H contains only the perfect heuristic, then $P^c(s,\Gamma,H) = P^*(s,\Gamma)$.

LEMMA 3. Let H be complete with respect to path $P \in P^*(s,n")$, where $n"$ is not necessarily a node in Γ.

(a) If there exists a shallowest node, n', on P in OPEN, then $g(n') = g^*(n')$. Furthermore, all ancestors, n^a, of n' on P are on CLOSED and are such that $g(n^a) = g^*(n^a)$.

(b) If there does not exist a shallowest node on P in OPEN, then all nodes, n, on P are in CLOSED and are such that $g(n) = g^*(n)$.

Lemma 3 indicates that A^G has already found the optimal pointer-path to n' (along the path in $P^*(s,n")$) and that this pointer-path will remain unaltered throughout the search.

Proof: By induction. We will show that for all iterations, either

i. there exists a shallowest node, n', on P in OPEN, $g(n') = g^*(n')$, and all ancestors of n' on P, n^a, are on CLOSED and are such that $g(n^a) = g^*(n^a)$, or

ii. there does not exist a shallowest node on P in OPEN.

We begin by proving an intermediate result: Assume $P \cap \text{OPEN} = \phi$; then $P \subseteq \text{CLOSED}$ and $g(n) = g^*(n)$ for all $n \in P$. Note that $P \cap \text{OPEN} = \phi$ cannot hold initially, since A^G places s on OPEN at the beginning of iteration 1. Observe, however, that at the beginning of iteration 2, $s \in \text{CLOSED}$ and $g(s)$

$= g^*(s) = 0$. More generally, assume $n \in P \cap CLOSED$, n is an ancestor of n", and that $g(n) = g^*(n)$. Since n is on CLOSED, n has been expanded. Since H is complete with respect to P, $n' \in P \cap SCS(n)$ is placed on OPEN $\cup$ CLOSED. But since $P \cap OPEN = \phi$, $n' \in$ CLOSED. The optimality of P implies $g(n') = g(n) + c(n,n') = g^*(n) + c(n,n') = g^*(n')$. The intermediate result then follows by induction.

Consider iteration 1. Node s is the shallowest node on OPEN, $g(s) = g^*(s) = 0$, and s has no ancestors. So the result holds at iteration 1.

Assume the result holds at iteration k. If there does not exist a shallowest node on P in OPEN, then by the above intermediate result, all nodes on P are on CLOSED and are not candidates for pointer path readjustment. Therefore, all nodes on P will remain on CLOSED, and hence there will continue to be no shallowest node on $P \cap OPEN$. Thus the result holds for iteration k + 1.

Assume there does exist a shallowest node $n \in P$ in OPEN such that $g(n) = g^*(n)$. Furthermore, assume all ancestors n^a of n on P are such that $n^a \in$ CLOSED and $g(n^a) = g^*(n^a)$. If n is not expanded, n remains the shallowest node on $P \cap OPEN$, since all ancestors of n are not candidates for pointer path readjustment. Hence, the result holds for iteration k + 1.

Assume n is expanded and n is an ancestor of n". (If n = n", then the result holds trivially.) Since H is complete with respect to P, $n' \in P \cap SCS(n)$ will be placed on OPEN $\cup$ CLOSED. Prior to the expansion of n, three cases are possible: (i) $n' \notin$ OPEN $\cup$ CLOSED, (ii) $n' \in$ OPEN, and (iii) $n' \in$ CLOSED.

Assume $n' \notin$ OPEN $\cup$ CLOSED. Then n' will be placed on OPEN, becoming the new shallowest node on OPEN, and $g(n') = g(n) + c(n,n') = g^*(n')$. Hence the result holds for iteration k + 1.

Assume $n' \in$ OPEN. Then n' will remain on OPEN, becoming the new shallowest node on OPEN, and pointer path readjustment may have to take place in order to insure that $g(n') = g^*(n')$. Hence the result holds for iteration k + 1.

Assume $n' \in$ CLOSED. If pointer path readjustment is required, n' is placed on OPEN, becoming the new shallowest node on OPEN, and $g(n') = g^*(n')$. Hence the result holds for iteration k + 1.

If $n' \in$ CLOSED and pointer path readjustment is not required, then n' remains on CLOSED and $g(n') = g^*(n')$. Since n' was on CLOSED, it has been expanded. Use of induction, the completeness of H with respect to P, the

finite length of P, and the optimality of P guarantees either (α) or (β), where:

(α) There is a descendant of n' on P $\cap$ OPEN, n^d, such that $g(n^d) = g^*(n^d)$ and each ancestor of n^d on P that is a descendent of n', n^+, is such that $g(n^+) = g^*(n^+)$. Hence the result holds for iteration k + 1.

(β) Each descendent of n' on P, n^+, is on CLOSED and is such that $g(n^+) = g^*(n^+)$. Hence P $\cap$ OPEN = ϕ, and the result holds for iteration k + 1. □

The following example indicates that if H is not complete with respect to a path in $P^*(s,n")$, n" not necessarily in Γ, then Lemma 3 may not hold, even if H is admissible.

EXAMPLE 1: Let $N = \{s,n_1,\ldots,n_5,\gamma\}$ and $\Gamma = \{\gamma\}$. The sets SCS(•), the cost structure c(•,•), and the resulting function $h^*(\cdot)$, are given in Table 1. Let $H = (\bar{h},h^*)$, where $\bar{h}(n) = h^*(n)$ for all $n \in N$ except n_4. Let $\bar{h}(n_4) = 0$. We note that H is admissible. Let $OPEN_k$ and $CLOSED_k$ be the OPEN and CLOSED sets at the beginning of the kth iteration of A^G. Then:

$$OPEN_1 = \{s\} \qquad CLOSED_1 = \phi$$
$$OPEN_2 = \{n_1\} \qquad CLOSED_2 = \{s\}$$
$$OPEN_3 = \{n_3\} \qquad CLOSED_3 = \{s,n_1\}$$
$$OPEN_4 = \{n_4,n_5\} \qquad CLOSED_4 = \{s,n_1,n_3\}.$$

Node n_2 was not placed on OPEN during iteration 1 because $U(s,n_2,H) < c(s,n_2)$. We note that

$$g(n_4) = c(s,n_1) + c(n_1,n_3) + c(n_3,n_4) = 3,$$

whereas

$$g^*(n_4) = c(s,n_2) + c(n_2,n_4) = 2.$$ □

n	SCS(n)	$h^*(n)$
s	$\{n_1, n_2\}$	9/2
n_1	$\{n_3\}$	7/2
n_2	$\{n_4\}$	4
n_3	$\{n_4, n_5\}$	5/2
n_4	$\{n_5\}$	3
n_5	$\{\gamma\}$	1
γ	ϕ	0

n \ n'	c(n,n') n_1	n_2	n_3	n_4	n_5	γ
s	1	1				
n_1			1			
n_2				1		
n_3				1	3/2	
n_4					2	
n_5						1

TABLE 1: Data for Example 1.

LEMMA 4. Assume H is admissible and that path $P \in P^*(s,\Gamma)$. Then at any time before A^G terminates, there exists an OPEN node n' on P such that $g(n') = g^*(n')$ and $f(n') \leq C^*$.

Proof: The admissibility of H and Lemma 1 imply that H is complete with respect to path P. Assume there does not exist a node $n' \in P \cap$ OPEN. Then by Lemma 3b, all nodes on P are on CLOSED, including a goal node. But a goal node on CLOSED implies that A^G has terminated, which is a contradiction. Therefore, there exists a shallowest node, n', on P in OPEN. By Lemma 3a, $g(n') = g^*(n')$. Since $P \in P^*(s,\Gamma)$ and since H is admissible, $f(n') = g(n') + \ell(n') = g^*(n') + \ell(n') \leq g^*(n') + h^*(n') \leq C^*$. □

THEOREM 3. Assume the heuristic set H is admissible. Then, A^G is admissible.

Proof: Assume there exists an optimal path $P \in P^*(s,\Gamma)$ with cost C^*. Since H is admissible, then by Lemma 1 H is complete with respect to P. By Lemma 4, at any time before A^G terminates there exists a node $n' \in$ OPEN $\cap$ P such that $g(n') = g^*(n')$ and $f(n') \leq C^*$. Therefore, A^G cannot terminate until it has expanded a goal node γ.

At the time A^G selects γ for expansion, there exists a node $n' \in$ OPEN $\cap$ P such that $g(n') = g^*(n')$ and $f(n') \leq C^*$. Thus, for A^G to choose γ for expansion, $f(\gamma) \leq f(n') \leq C^*$. Hence, $f(\gamma) = C^*$, and so A^G has found an optimal path. □

REFERENCES

White, C.C., and Harrington, D.P., "Application of Jensen's Inequality for Adaptive Suboptimal Design," Journal of Optimization Theory and Applications, Vol. 32, pp. 89-99, 1980.

Puterman, M.L., and Shin, M.C., "Action Elimination Procedure for Modified Policy Iteration Algorithms," Operations Research, Vol. 30, pp. 301-318, 1982.

Pearl, J., Heuristics: Intelligent Strategies for Computer Problem Solving. Addison-Wesley, Reading, MA, 1984.

A QR Factorization Algorithm with Controlled Local Pivoting*†

Christian H. Bischof

Department of Computer Science
Cornell University
Ithaca, New York 14853

Abstract. This paper presents a new parallel version of the Householder algorithm with column pivoting for computing the QR factorization of a matrix. In contrast to the standard algorithm we employ a local pivoting scheme that allows for efficient implementation of the algorithm on a parallel machine, in particular one with a distributed architecture. An inexpensive but reliable incremental condition estimator is used to control the selection of pivot columns by obtaining cheap estimates for the smallest singular value of the currently created upper triangular matrix R. Numerical experiments show that the local pivoting strategy behaves about as well as the traditional global pivoting strategy. They also show the advantages of incorporating the controlled pivoting strategy into the traditional QR algorithm to guard against the known pathological cases.

1 Introduction

One of the standard problems in numerical linear algebra is the solution of the linear least squares problem

$$\min \|Ax - b\|_2 \tag{1}$$

where A is an $m \times n$ ($m \geq n$) matrix. The common way to approach this problem [12,17,19] is via a *QR factorization*

$$AP = QR \tag{2}$$

of A. Here P is an $n \times n$ permutation matrix, Q is an $m \times n$ matrix with orthogonal columns (i.e. $Q^TQ = I_n$) and R is an upper triangular $n \times n$ matrix. If A is a dense matrix, Q is

*This work was supported by the U.S. Army Research Office through the Mathematical Science Institute of Cornell University, by the Office of Naval Research under contract N00014-83-K-0640 and by NSF contract CCR 86-02310.

†A preliminary version of this paper was published in the proceedings of the 3rd International Conference on Supercomputing, Steve and Lana Karthashev, Eds.

usually computed by a sequence of *Householder Transformations*

$$H = I - 2u\,u^T.$$

Choosing

$$u = \frac{x + \mathrm{sign}(x_1)\,\|x\|_2\,e_1}{\|x + \mathrm{sign}(x_1)\,\|x\|_2\,e_1\|_2} \tag{3}$$

we can reduce a given vector x to a multiple of the canonical unit vector e_1 since

$$(I - 2u\,u^T)\,x = -\mathrm{sign}(x_1)\,\|x\|_2\,e_1.$$

If A has full rank, we can avoid exchanging columns when computing the QR factorization, i.e. P in (2) will be the identity. If the rank of A is not known, we can employ column pivoting [3]. The idea is to choose as next column always the one that has the highest residual with respect to the subspace spanned by the columns that were selected before. The hope is that in the resulting QR factorization (2) of A the ill-conditioning of A will reveal itself by a small trailing subblock of R: if $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ are the singular values of A and we partition R into

$$\begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix} \tag{4}$$

with an $r \times r$ lower right hand block R_{22} then it is easy to show [12, page 19] that

$$\sigma_{n-r+1}(A) \leq \|R_{22}\|_2.$$

While there are counterexamples (see section 5) where the column pivoting strategy fails to reveal ill-conditioning of A, it works well in practice.

Another alternative is to compute the *Singular Value Decomposition* (SVD)

$$A = U\Sigma V^T \tag{5}$$

of A. Here U and V are orthogonal matrices whose columns are the left– and right–singular vectors of A, respectively. $\Sigma = \mathrm{diag}(\sigma_i)$ contains the singular values of A. The SVD is at least twice as expensive to compute as the QR factorization and for that reason the QR factorization with column pivoting is usually preferred.

There is also a middle path between QR factorization and SVD. As was pointed out originally in [11] we can use the singular vector corresponding to the smallest singular value to find a permutation P that will guarantee a small r_{nn} if $\sigma_n(A)$ is small. Chan [4] and Foster [9] extend this idea to higher dimensions. Their idea is to first compute any QR factorization of A and then "peel off" the small singular values of R one after the other by computing an appropriate singular vectors at each step. Let us from now on assume that A

has r small singular values and that there is a well-defined gap between σ_{n-r} and σ_{n-r+1}. It is shown in [11] that a well-defined gap is necessary to make a sensible decision on the numerical rank of A. Then Chan proves that if r is not too large his algorithm will compute a "rank-revealing QR factorization" in the sense that R_{22} in (4) is guaranteed to be small.

On a single processor the Householder QR factorization without pivoting requires $O(mn^2)$ flops, column pivoting requires an additional n^2 flops and the rank-revealing QR algorithm requires an additional $3rn^2$ flops on average [4]. So the computational complexity of these algorithms is comparable on a single-processor machine.

The situation is quite different on a multiprocessor machine especially if it is based on a distributed architecture. The Householder QR algorithm without pivoting can be very efficiently parallelized simply by pipelining the computation [1,20]. So one processor can still be busy finishing a previous update while another already computes the next Householder vector. The introduction of column pivoting makes pipelining impossible since all processors have to synchronize to select the next pivot column. Hence the Householder QR algorithm essentially proceeds in a lockstep fashion which results in a serious loss of efficiency on machines that previously could profit from the pipelining.

In Chan's algorithm the steps after the initial QR factorization are hard to parallelize. For each of the r small singular values of A, the algorithm computes an approximate singular vector via inverse iteration. On average this requires two iteration steps [4] and hence the solution of four triangular equation systems per small singular value. Although much progress has been made recently in solving triangular equation systems on distributed architectures [7, 14, 13, 18] this problem can by no means be parallelized as efficiently as the inital QR factorization. In addition the permutation deduced from the singular vector destroys the upper triangular shape of R which then has to be restored by a sequence of Givens rotations. Again that is essentially a sequential process that is hard to parallelize [7]. Apart from their sequential nature, an inherent difficulty in parallelizing the equation solving and QR update steps is that the computational work is of the same order of magnitude as the amount of data it involves. That is we have to perform $O(n^2)$ flops using $O(n^2)$ data. Since R is distributed throughout the system it is hard to mask the communication overhead with the little arithmetic work to be performed. So the post-processing of R can end up being a good part of the overall computation time on a parallel machine.

In this paper we suggest a new QR decomposition algorithm that avoids these penalties and can be efficiently parallelized. By using a *local* pivoting strategy we are able to pipeline the computation and at the same time identify the set of columns of A responsible for its ill-conditioning. In section 2 we outline the pipelined Householder QR algorithm and motivate the local pivoting strategy. Section 3 introduces a condition estimator that allows us to monitor the numerical soundness of the local pivoting strategy and presents numerical results showing its robustness. Section 4 combines these ideas into an effective parallel algorithm for determining the numerical rank of A and thus solving the linear least squares problem.

Numerical results obtained by simulating the parallel algorithm are presented in section 5. We summarize our results and outline possible directions of future research in section 6.

2 The Householder QR Algorithm with Local Column Pivoting

The Householder QR algorithm without pivoting processes the columns of A in their natural order from left to right. If we have a parallel machine, it is natural to group the processors into a logical ring and deal out columns in a round-robin fashion. This technique staggers the computation across the processors and guarantees a load balanced computation. It allows simple static assignment of data to processors and is for the most part synchronized by the flow of data between processors. Due to these attractive characteristics, the pipelining technique has been widely used [10,13,16]. If special vector hardware can be exploited, several Householder matrices can be bundled together by using the WY factorization [2,21] to arrive at a block pipelined algorithm [1].

In contrast the QR algorithms with traditional column pivoting at each step chooses the column that has the highest residual with respect to the subspace spanned by the columns already selected. This residual is easy to compute and can be updated cheaply as new columns are selected [8]. But in the parallel setting, the selection of the pivot column introduces a synchronisation point. Each processor can easily choose its local candidate pivot column by considering only the columns that are assigned to it. Choosing the global pivot column on the other hand requires that each processor either makes its local pivot information known to all other processors or that a designated processor collects all the local pivot informations. So global pivoting essentially forces the program into a lockstep mode that may severly curtail performance.

The easiest way out of this dilemma is to forgo global pivoting altogether and content oneself with local pivoting. A simplified version of the resulting algorithm for a ring of processors is given in Figure 1. We distribute columns of A to processors in a round-robin fashion. To be precise, let us assume that we have p processors $proc_0, \ldots, proc_{p-1}$ and that a_j is the jth column of A. Then processor $proc_i$ receives columns a_j where

$$i = (j-1) \bmod p.$$

This is commonly referred to as the *column wrap mapping*. The array C is local to each processor and contains the $cols_k$ columns assigned to processor $proc_i$ ($\sum_{i=0}^{p-1} cols_k = n$) . p_{left} and p_{right} designate the left and right neighbour of $proc_k$, respectively.

$$u \leftarrow genhh(x)$$

returns u as defined by (3) and

$$A \leftarrow apphh(u, A)$$

processor $proc_k$

$lcnt \leftarrow 0$; {counter for HH vectors generated in $proc_k$}
$gcnt \leftarrow 0$; {counter for HH vectors generated globally}
foreach $i \in \{1, \ldots, cols_k\}$ **do**
  $perm_i \leftarrow (k+1) + (i-1)p$; { wrap mapping}
  $res_i \leftarrow \|c(:,i)\|_2$;
end foreach
if $(k = 0)$ **then** {determine first pivot column}
  $lcnt \leftarrow gcnt \leftarrow 1$;
  determine first pivot column, send it it p_{right} and
  update all other columns as shown in main loop below.
end if
while $(lcnt < cols_k)$ **do** {main loop}
  receive u from p_{left} ; $gcnt \leftarrow gcnt + 1$;
  if (u not generated by p_{right}) **then** send u to p_{right} **end if**
  if $(k = gcnt \bmod p)$ **then** { my turn to generate next HH vector }
    $lcnt \leftarrow lcnt + 1$;
      { complete enough of $H(u)$ update to determine next pivot column }
    $z \leftarrow c(gcnt{:}m, lcnt{:}cols_k)^T u$;
    $c(gcnt{+}1, lcnt{:}cols_k) \leftarrow c(gcnt{+}1, lcnt{:}cols_k) - 2\,u(2{:}m{-}gcnt{+}1)\,z$;
    $res_i \leftarrow \sqrt{res_i^2 - C(gcnt,i)^2}, i \in \{lcnt, \ldots, cols_k\}$
    Let $pvt \in \{lcnt, \ldots, cols_k\}$ be such that res_{pvt} is maximal
      { guarded pivoting strategy will be inserted here }
    $c(gcnt{:}m, pvt) \leftarrow c(gcnt{:}m, pvt) - 2\,u(1)\,z^T$;
    $\hat{u} \leftarrow genhh(C(gcnt{+}1{:}m, pvt))$; $gcnt \leftarrow gcnt + 1$;
    if $(gcnt < n)$ **then** send $\hat{u}$ to p_{right} **end if**
    $c(gcnt{:}m, lcnt{:}pvt{-}1) \leftarrow c(gcnt{:}\ m, lcnt{:}pvt{-}1) - 2\,u(2{:}m{-}gcnt{+}2)z^T$;
      { complete $H(u)$ update }
    $c(gcnt{:}m, pvt{+}1{:}cols_k) \leftarrow c(gcnt{:}\ m, pvt{+}1{:}cols_k) - 2\,u(2{:}m{-}gcnt{+}2)z^T$;
      { complete $H(\hat{u})$ update }
    $c(:, pvt) \leftrightarrow c(:, lcnt)$; $perm_{lcnt} \leftrightarrow perm_{pvt}$; $res_{pvt} \leftarrow res_{lcnt}$;
    $c(gcnt{:}m, lcnt{+}1{:}cols_k) \leftarrow apphh(\hat{u}, c(gcnt, lcnt{+}1{:}cols_k))$;
    else { apply $H(u)$ update }
      $c(gcnt{:}m, lcnt{+}1{:}cols_k) \leftarrow apphh(u, C(gcnt{:}m, lcnt{+}1{:}cols_k))$;
    end if
  $res_i \leftarrow \sqrt{res_i^2 - c(gcnt,i)^2}, i \in \{lcnt+1, \ldots, cols_k\}$
end while

Figure 1: The Pipelined QR Algorithm with Local Pivoting

returns $H(u)\,A$. The vector *perm* is used to store the permutation matrix. If $perm(i) = l$, then the lth column of A has been permuted into the ith position. The vector *res* contains the residuals that the columns not yet chosen have with respect to the other columns already selected. To save space, "HH" is used as an abbreviation for "Householder".

It is worth pointing out that a processor that has to generate a new pivot column completes only as much of the previous Householder update as is necessary to update *res* and to determine the next pivot column. This is important since we want to avoid that other nodes are idle waiting for a new Householder vector to arrive.

The problem with the strictly local pivoting strategy is obviously reliability. As a pathological example, assume that all columns in processor 1 are nearly equal. As a result, processor 1 will make bad choices after it has generated the very first Householder vector. The resulting upper triangular matrix R will be very ill-conditioned but will not necessarily have a small lower right hand block. So in order for the local pivoting strategy to be reliable, we have to *guard against choosing nearly dependent pivot columns.*

3 An Incremental Estimator for the Smallest Singular Value of a Triangular Matrix

To guard against choosing "bad" pivot columns, we have to monitor the smallest singular value $\sigma_{min}(R_i)$ where R_i is the leading $i \times i$ upper triangular matrix generated after applying i Householder transformations to A. The exact computation of $\sigma_{min}(R_i)$ by inverse iteration for example is too expensive, especially since a good order-of-magnitude estimate suffices for our purposes.

A common idea underlying condition estimators [5,6] is to exploit the implication

$$Rx = d \implies \frac{1}{\sigma_{min}(R)} = \|R^{-1}\|_2 \geq \frac{\|R^{-1}d\|_2}{\|d\|_2} = \frac{\|x\|_2}{\|d\|_2}$$

by generating a large norm solution x to a moderately sized right hand side d and then to use

$$\hat{\sigma}_{min}(R) := \frac{\|d\|_2}{\|x\|_2}$$

as an estimate for $\sigma_{min}(R)$. The hope is that x will be an approximate singular vector corresponding to the smallest singular value and that as a consequence $\hat{\sigma}_{min}(R)$ will not be too much of an over-estimate of $\sigma_{min}(R)$. Our choice of algorithms for an condition estimator is severely restricted by the fact that it is not feasible to access the previously generated R when we want to decide on the suitability of a new pivot column. To be more precise, given a good estimate $\hat{\sigma}_{min}(R)$ defined by a large norm solution x to $Rx = d$ and a new column

$\begin{pmatrix} v \\ \gamma \end{pmatrix}$ of R, we want to obtain a large norm solution y to

$$R'y = \begin{pmatrix} R & v \\ 0 & \gamma \end{pmatrix} y = \hat{d}$$

without accessing R again. None of the condition estimators surveyed by Higham [15] has that property, but the two-norm condition estimator suggested by Cline, Conn and Van Loan [5] can be modified to conform to those restrictions. The idea then is the following:

Given x such that $R^T x = d$ with $\|d\|_2 = 1$, find $s := \sin\varphi$ and $c := \cos\varphi$ such that $\|y\|_2$ is maximized where $y = \begin{pmatrix} z \\ \delta \end{pmatrix}$ solves

$$\begin{pmatrix} R^T & 0 \\ v^T & \gamma \end{pmatrix} y = \begin{pmatrix} sd \\ c \end{pmatrix}. \tag{6}$$

We here exploit the fact that R' and R'^T have identical singular values. An easy calculation shows that maximizing $\|y\|_2$ is equivalent to maximizing

$$\Phi(\varphi) = s^2\beta - 2\alpha sc \tag{7}$$

where

$$\alpha = v^T x \text{ and } \beta = \gamma^2 x^T x + \alpha^2 - 1.$$

Taking derivatives in (7) and setting $\eta = \beta/(2\alpha)$ we find two possible solutions:

$$s_{1,2} = \frac{1}{\sqrt{1+\mu_{1,2}}}$$

where

$$\mu_{1,2} = \eta \pm \sqrt{1+\eta^2}.$$

The corresponding cosine values are

$$c_{1,2} = s_{1,2}\,\mu_{1,2}.$$

To choose between the two possibilites, we compute $\Phi(s_1)$ and $\Phi(s_2)$ and choose the sine/cosine pair that results in the greater value for Φ. For the special case $\alpha = 0$ we obtain $c_1 = 1$, $s_1 = 0$ and $c_2 = 0$ $s_2 = 1$. The new approximate singular vector y as defined by (6) is then given by setting

$$z := sx \text{ and } \delta := \frac{c - s\alpha}{\gamma}.$$

The resulting estimate for the smallest singular value $\sigma_{min}(R')$ of R' is

$$\hat{\sigma}_{min}(R') = \frac{1}{\|y\|_2}.$$

From this description it is clear that this condition estimator satisfies our algorithmic constraints. Given a current R_i we only need to save the current solution x and its norm $\|x\|_2$ to arrive at an estimate for $\sigma_{min}(R_{i+1})$. Furthermore the calculation is inexpensive. For a $k \times k$ matrix R_i we only need $2k$ flops to arrive at an estimate for $\sigma_{min}(R_{i+1})$. So altogether it costs only n^2 flops to run this condition estimator alongside the generation of an $n \times n$ triangular matrix.

To assess the accuracy of our condition estimator, we performed the suite of tests suggested by Higham [15]. Three different types of test matrices are employed. In each test, upper triangular matrices R were generated by computing the QR factorization of various $n \times n$ matrices A for $n = 10, 25, 50$ both with and without column pivoting.

Test1 (see Table 1): The elements of A were chosen as random numbers from the uniform distribution on $[-1, 1]$. Fifty matrices were generated for each n. As observed by Higham, this type of matrix usually is well-conditioned. Over the whole test the minimum, maximum and average values of the two-norm condition number $\kappa_2(A) = \sigma_1/\sigma_n$ were $21, 1.4 \cdot 10^4$ and $2.0 \cdot 10^3$ respectively.

Test2 (see Tables 2 and 3) and *Test 3*: In these tests we used random matrices A with preassigned singular value distributions $\{\sigma_i\}$. Random orthogonal matrices U and V were generated using the method of Stewart [22] and then A was formed as in (5). For each value of n and each singular value distribution, fifty matrices were generated by choosing different matrices U and V. For test 2 we chose the exponential distribution

$$\sigma_i = \alpha^i, \ 1 \leq i \leq n$$

where α is determined by $\kappa_2(A)$. For test 3, we chose the sharp-break distribution

$$1 = \sigma_1 = \cdots = \sigma_{n-1} > \sigma_n = \frac{1}{\kappa_2(A)}.$$

The figures given in Tables 1–3 are the ratios

$$\hat{\sigma}_{min}(R)/\sigma_{min}(R) \geq 1$$

The first number in each pair is the maximum ratio over the fifty matrices and the second is the average ratio. All results were rounded to two significant digits. For Test 3 we observed a ratio of 1.0 (i.e. the estimate had at least two correct figures) in all cases. These results show that our conditon estimator produces indeed good estimates. We overestimate $\sigma_{min}(R)$

Table 1: *Results of Test* 1

pivoting	$n = 10$	25	50
no	2.2/1.4	6.6/2.4	7.0/3.1
yes	2.3/1.5	3.2/2.0	3.9/2.6

Table 2: *Results of Test* 2 *without Pivoting*

κ_2	$n = 10$	25	50
10	1.8/1.3	1.7/1.4	1.6/1.4
10^3	3.0/1.9	2.5/2.0	3.2/2.2
10^6	8.1/1.9	6.3/2.6	4.2/2.8
10^9	6.1/2.2	5.9/3.0	5.2/3.2

Table 3: *Results of Test* 2 *with Pivoting*

κ_2	$n = 10$	25	50
10	1.6/1.3	1.6/1.4	1.7/1.4
10^3	2.2/1.5	2.3/1.8	2.5/2.0
10^6	2.8/1.5	3.4/2.1	3.4/2.5
10^9	2.4/1.6	3.3/2.2	4.3/2.7

only by a small factor and the results vary only little with condition number, matrix size and singular value distribution. Pivoting increases the accuracy of the condition estimator and we can confidently expect similar accuracy when applying this estimator to matrices R generated by the local pivoting strategy.

4 The QR Algorithm with Controlled Local Pivoting

With the condition estimator we now have the tool to insure the numerical stability of the local pivoting strategy. Using the same notation as in the algorithm of Figure 1 processor k now can check whether $c(:,j)$ is a reasonable choice for the next pivot column before computing $\hat{u}$. Assuming that processor k knows the current estimate x as well as $\|x\|_2$ for the current upper triangular matrix R_{gcnt}, all that is needed for the next condition estimator step is the last column $\begin{pmatrix} v \\ \gamma \end{pmatrix}$ of R_{gcnt+1}. But

$$v = c(1 : gcnt, j)$$

has already been computed and from the definition of u and res it follows immediately that

$$\gamma = -\text{sign}(c(gcnt + 1, j))\, res_j.$$

So all the information for the next condition estimator step is readily at hand and we can compute a new approximate singular vector y for R_{gcnt+1}.

With

$$\alpha = \max_{1 \le i \le n} \|a_i\|_2$$

being the norm of the largest column of A, we then take

$$\hat{\sigma}_{min}(R_{gcnt+1}) = \frac{1}{\eta_1 \|y\|_2} \tag{8}$$

as an estimate for the smallest singular value of R_{gcnt+1} and

$$\hat{\kappa}\,(R_{gcnt+1}) = \eta_2\, \alpha \|y\|_2 \tag{9}$$

as an estimate for the true condition number of R_{gcnt+1}. The scaling factors η reflect the trust we have in the accuracy of our estimates. Based on the numerical results of section 3 we recommend $\eta_1 = 3$ and $\eta_2 = 10$. The choice of η_2 reflects the fact that in general the norm of the largest column is a good estimator for the largest singular value of a matrix.

Comparing the estimates (8) or (9) against a chosen threshold we will then accept or reject a candidate pivot column. The exact threshold depends heavily on the application, in

particular the accuracy of the initial data. If the data is accurate to machine precision ϵ, a candidate pivot will in general be rejected if $\hat{\sigma}_{min}(R_{gcnt+1}) = O(1/\epsilon)$.

If the candidate pivot column is rejected, processor k has exhausted its supply of "reasonable" columns and from then on it will only apply Householder vectors generated by other processors to its remaining columns. If on the other hand we accept the candidate pivot column, then processor k will actually compute $\hat{u}$, send $(\hat{u}, y, \|y\|_2)$ to its right neighbour and then proceed as in Figure 1. It should be noted that y and $\|y\|_2$ have to be forwarded only to the processor that will generate the next Householder vector, while $\hat{u}$ will eventually be known to all processors. So the propagation of the condition estimator results will result in only a minor increase in data traffic.

This scheme continues until no processor has any acceptable pivot candidate left. Assuming that altogether we generated $\hat{n} = n - \hat{r}$ Householder vectors, we have at this point computed the incomplete QR factorization

$$AP = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ 0 & \hat{A} \end{pmatrix} \tag{10}$$

where Q_1 is $m \times \hat{n}$, Q_2 is $m \times (m - \hat{n} + 1)$ and $Q = [Q_1, Q_2]$ is orthogonal. R_{11} is upper triangular of size $\hat{n} \times \hat{n}$ and $\hat{A}$ is of size $(m - \hat{r} + 1) \times \hat{r}$. Our controlled pivoting strategy gives us an estimate for $\sigma_{min}(R_{11})$ and further we know that adding any of the leftover $\hat{r}$ columns of AP would result in a decrease of the smallest singular value below our chosen threshold. So we have good reason to assume that $\hat{r}$ is the dimension of the numerical null space of A. Then we can set $\hat{A}$ in (10) to zero and use the resulting truncated QR factorization to solve the least squares problem (1).

5 Numerical Experiments

To assess the numerical behavior of the proposed local pivoting scheme, we simulated the parallel algorithm using PRO-MATLAB and compared it with the traditional QR factorization algorithm with global column pivoting. Various 50×50 matrices were generated and the local pivoting strategy simulated on 8 processors.

For tests 1 to 3 we generated 50 random matrices for each singular value distribution $\{\sigma_i\}$. For all matrices the largest and smallest singular values were 1 and 10^{-9} respectively.

Break 1 Distribution: $\sigma_1 = \ldots = \sigma_{49} = 1$; $\sigma_{50} = 10^{-9}$.

Break 9 Distribution: $\sigma_1 = \ldots = \sigma_{41} = 1$; $\sigma_{42} = \ldots = \sigma_{50} = 10^{-9}$.

Exponential Distribution: $\sigma_i = \alpha^i$; $\alpha = (10^{-9})^{-\frac{1}{49}}$.

Table 4: min/avg/max Values of the Condition Numbers of R using Local and Global Pivoting

Distribution	break 1	break 9	exponential
$\kappa_{par}(R)$	2.8 /4.8 /9.5	4.0 /12 /180	5.0e6 /8.8e7 /1.9e7
$\kappa_{trad}(R)$	2.1 /2.7 /3.6	3.4 /4.4 /6.1	4.2e6 /7.2e6 /1.3e7
$\kappa_{opt}(R)$	1.0	1.0	9.5e6

Setting the rejection threshold for the smallest singular value to 10^{-7} and discounting the estimate for the smallest singular value (8) by a factor of $\eta_1 = 3$ we reject a candidate pivot column in the parallel algorithm if

$$\frac{1}{\|y\|_2} = \hat{\sigma}_{min}(R_{gcnt+1}) \leq 3 \cdot 10^{-7}.$$

For the traditional QR factorization algorithm we use the last diagonal entry of R_{gcnt+1} as estimate and reject a candidate pivot column if

$$| r_{gcnt+1,gcnt+1} | \leq 3 \cdot 10^{-7}.$$

Table 4 shows the condition number $\kappa(R)$ of the upper triangular matrices R generated by controlled local pivoting and by traditional column pivoting on those matrices. Letting σ_{cut} be the smallest singular value greater than 10^{-7} then the optimal value we can achieve for $\kappa(R)$ is $\kappa_{opt}(R) = 1/\sigma_{cut}$. Furthermore let $\kappa_{par}(R)$ be the condition number resulting form the parallel scheme and $\kappa_{trad}(R)$ the condition number resulting from the traditional column pivoting scheme. For $\kappa_{par}(R)$ and $\kappa_{trad}(R)$ observed minimum, average and maximum values are displayed. These results show that guarded local pivoting is about as effective as full column pivoting in generating a well-conditoned R — even if the number of local columns is fairly small.

For the sharp break distributions there is a well-defined gap between the singular values before and after the acceptance threshold and both local and global column pivoting identify the numerical nullspace correctly in all cases. As already pointed out earlier, the determination of numerical rank becomes problematic if there is no well-defined gap between singular values that are considered "large" and "small". The exponential distribution is such a problematic case. There are 39 singular values that are larger than 10^{-7} but there is no well-defined break. To be exact:

$$\sigma_{37} = 2.4 \cdot 10^{-7},\ \sigma_{38} = 1.6 \cdot 10^{-7},\ \sigma_{39} = 1.04 \cdot 10^{-7} \text{and}\ \sigma_{40} = 6.8 \cdot 10^{-8}\ .$$

Table 5: Frequency of Accepting Columns for the Exponential Distribution

No. of columns accepted	36	37	38
local pivoting	14	30	6
global pivoting	5	42	3

The column pivoting strategy reflects this difficulty in accepting less than 39 columns and the observed results are displayed in Table 5. It shows that even for an ill-defined problem the guarded local pivoting scheme is very reliable in that it leans towards a small underestimate of the dimension of the range space of A.

Our last example shows the advantage resulting from integrating the incremental condition estimator with the global pivoting scheme. Let

$$A_n = \mathrm{diag}(1, s, s^2, \ldots, s^{n-1}) \begin{pmatrix} 1 & -c & \cdots & \cdots & -c \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & -c \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} + \mathrm{diag}(n\epsilon, (n-1)\epsilon, \ldots, \epsilon)$$

where $c^2 + s^2 = 1$ and ϵ is the machine precision. A_n is very ill-conditioned, but although each leading principal submatrix $A_k (k \leq n)$ is also ill-conditioned, there is a well-defined gap between σ_n and σ_{n-1}. As an example we have $\sigma_{49} = 1.2 \cdot 10^{-3}$ and $\sigma_{50} = 3.7 \cdot 10^{-12}$ for $n = 50$ and $c = 0.5$. This is a well-known example where the QR factorization with column pivoting fails since even in floating point arithmetic the matrix is its own QR factorization but no trailing block of R is small to reveal its ill-conditioning. In this example both the local and global pivoting schemes select the the columns in their natural order. However the incremental condition estimator integrated into the parallel scheme detects the ill-conditioning of the leading principal submatrices A_k — it never overestimates the smallest singular value by a factor of more than 1.5. So while the column pivoting scheme fails the incremental condition estimator insures that failure will not go unnoticed. Given its negligible extra cost this suggest the usefulness of incorporating the incremental condition estimator into the global column pivoting scheme.

The matrix A_n is also an example where the local pivoting scheme performs better than the global one. Let $\tilde{A}_{50}$ be the same matrix as A_{50} except that the order of columns has been reversed. For the global column pivoting scheme this permutation is without consequences and it fails. The parallel scheme simulated on 8 processors on the other hand correctly

identifies the numerical nullspace of $\tilde{A}_{50}$. While this is an exceptional occurence due to the special structure of A_n, it is nonetheless surprising since intuitively one would expect the global pivoting strategy to always perform better than the local one.

6 Conclusions

This paper presented a new variant of the Householder QR algorithm with column pivoting. In that context we introduced a new incremental condition estimator that allowed us to update the estimate for the smallest singular value of the upper triangular matrix R as new columns were added to R. The update required only $O(n)$ flops and the saving of $O(n)$ words between successive steps. Despite its small computational cost, experiments with a variety of matrices demonstrated the reliability of the condition estimation algorithm. This condition estimator made it possible to implement a strictly local pivoting scheme for the QR factorization by guarding against an improper choice of pivot columns. Numerical experiments show that the local pivoting scheme performs by and large as well as global pivoting. There even exist cases where the local pivoting scheme succeeds while the global pivoting scheme fails.

We also gave an example showing the usefulness of integrating the incremental condition estimator with the traditional global column pivoting strategy. The n^2 flops for the condition estimator might be a worthwhile investment to guard against the pathological cases that are not revealed by the traditional QR factorization algorithm with column pivoting.

We are currently investigating the effect of a dynamic threshold for the acceptance or rejection of a pivot column. Starting with a relative conservative threshold and relaxing it as the computation proceeds is likely to result in better conditioned leading submatrices of R. The penalty is a possible loss of effiency as processors might have to reconsider the same column as pivoting candidate at a later stage of the algorithm.

References

[1] Christian Bischof. A pipelined block QR decomposition algorithm. In Garry Rodrigue, editor, *Proceedings of the 3rd SIAM Conference on Parallel Processing for Scientific Computing*. SIAM Press, 1988. *to appear.*

[2] Christian H. Bischof and Charles F. Van Loan. The WY representation for products of Householder matrices. *SIAM Journal on Scientific and Statistical Computing*, 8:s2–s13, 1987.

[3] P. A. Businger and G. H. Golub. Linear least squares solution by Householder transformation. *Numerische Mathematik*, 7:269–276, 1965.

[4] Tony F. Chan. Rank revealing QR factorizations. *Linear Algebra and its Applications*, 88/89:67–82, 1987.

[5] A. K. Cline, A. R. Conn, and C.-F. Van Loan. *Generalizing the LINPACK Condition Estimator*, volume 909 of *Lecture Notes in Mathematics*, pages 73–83. Springer Verlag, Berlin, 1982.

[6] A. K. Cline, C. B. Moler, G. W. Stewart, and J. H. Wilkinson. An estimate for the condition number of a matrix. *SIAM Journal on Numerical Analysis*, 16:368–375, 1979.

[7] Thomas F. Coleman and Guangye Li. Solving systems of nonlinear equations on a message-passing multiprocessor. Technical Report CS–87–887, Cornell University, November 1987.

[8] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *LINPACK Users' Guide.* SIAM Press, 1979.

[9] L. V. Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra and its Applications*, 74:47–71, 1986.

[10] George A. Geist and Michael T. Heath. Parallel Cholesky factorization on a hypercube multiprocessor. Technical Report ORNL–6190, Oak Ridge National Laboratory, 1985.

[11] G. H. Golub, V. Klema, and G. W. Stewart. Rank degeneracy and least squares problems. Technical Report TR–456, Dept. of Computer Science, University of Maryland, 1976.

[12] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1983.

[13] Michael T. Heath and Charles H. Romine. Parallel solution of triangular systems on distributed-memory multiprocessors. Technical Report ORNL/TM–10384, Oak Ridge National Laboratory, 1987.

[14] Michael T. Heath and Charles H. Romine. Parallel solution of triangular systems on distributed-memory multiprocessors. Unpublished Manuscript, 1988.

[15] Nicholas J. Higham. A survey of condition number estimation for triangular matrices. *SIAM Journal on Scientific and Statistical Computing*, 29(4):575–598, 1987.

[16] Ilse Ipsen, Youcef Saad, and Martin Schultz. Dense linear systems on a ring of processors. *Linear Algebra and its Applications*, 77:205–239, 1986.

[17] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems.* Prentice Hall, 1974.

[18] Guangye Li and Thomas F. Coleman. A new method for solving triangular systems on distributed memory message-passing multiprocessors. Technical Report 87-812, Cornell University, 1987.

[19] Åke Björck. *Handbook of Numerical Analysis*, volume 1, chapter Least Squares Methods. North Holland, 1987.

[20] Alex Pothen and Padma Raghavan. Distributed orthogonal factorization: Givens and Householder algorithms. Technical Report CS-87-24, The Pennsylvania State University, 1987.

[21] Robert Schreiber and Charles Van Loan. A storage efficient WY representation for products of Householder transformations. Technical Report CS-87-864, Cornell University, 1987.

[22] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17:403-409, 1980.

A Parallel Algorithm for Nonlinear Equations

Rodrigo Fontecilla
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742

ABSTRACT

In this paper, we present a parallel algorithm for the solution of systems of nonlinear equations. The algorithm is primarily based on the serial nonlinear Jacobi algorithm. Different parallel implementations are discussed. In particular, a block form is presented for the case when the number of processors is small in comparison to the number of variables. A straightforward implementation is given for the solution of unconstrained minimization problems. Some numerical experiments run on an Encore/Multimax with 20 processors are presented.

A parallel algorithm for nonlinear equations

1.Introduction. In this paper we present a parallel implementation of the serial nonlinear Jacobi algorithm for the solution of systems of nonlinear equations. The serial algorithm is a particular case of the more general SOR–Newton algorithms. The algorithm is based on the same idea as the Jacobi algorithm for solving linear systems of equations and thus, it suffers from the same drawback as its counterpart for linear systems, namely, its slow rate of convergence. The serial algorithm was first presented by Wegge [23], later analyzed by Schechter [20,21] and Voigt [22] and the most recent implementation given by Dennis and Walker [6]. For a detailed overview of SOR–Newton methods see Ortega and Rheinboldt [17].

The algorithm was discarded as a viable way of solving nonlinear systems of equations and replaced by more efficient methods such as Newton–like methods. However, these latter methods do not lend themselves in a straightforward manner to a parallel environment. Consider, for instance, Newton's method for solving

$$F(x) = 0 \quad \text{with} \quad F: \mathbf{R}^n \rightarrow \mathbf{R}^n . \tag{1.1}$$

The iterative scheme follows:

Step 0. Get x_0. Set $k=0$.

Step 1. Solve $F'(x_k)s_k = -F(x_k)$ (1.2.a)

Step 2. Update $x_{k+1} = x_k + s_k$ (1.2.b)

Step 3. Set $k=k+1$. Go to step 1.

The linear system of equations (1.2.a) that arises at each iteration could be solved in parallel; see for instance [19]. However, the amount of message passing involved at each iteration remains a bottleneck. Moreover, if the initial guess x_0 is far away from the solution, a globally convergent technique such a line search or a trust region must be imple-

mented. Such global techniques must be run in parallel, otherwise several processors will be idle while such computation occurs.

Other parallel algorithms have been proposed in the literature for solving (1.1). The one dimensional case $n=1$ have been analyzed in [7] and [14]. Different parallel methods were proposed where the emphasis resided in doing concurrent function evaluations. In [1], Baudet presents an excellent study of asynchronous iterative methods for multiprocessors. He uses the contraction mapping iteration and present the convergence criteria for these methods. However, his numerical experiments were performed on linear functions only. Bojanczyk in [2] uses an asynchronous Newton method where the function $F(x)$ and the Jacobian of F are calculated in parallel. Since the Jacobian evaluation takes much longer than the function evaluation, Newton steps are taken using a fixed Jacobian until the processor calculating the new Jacobian finishes. He shows that this parallel method will be at most four times faster than the serial method no matter how many processors are used in the computations. More recently, White [24] presented a parallel nonlinear Newton–SOR algorithm. In there, the main iteration is the Newton iteration for solving (1.1), then the linear system is solved using the Gauss–Seidel method with the multi–splittings techniques developed in [15]. He shows convergence of the method and presents some numerical results on a serial computer.

The algorithm we are proposing lends itself in a straightforward fashion to a parallel implementation. In particular, the bigger the dimension of the problem the higher the speedup that can be attained. The main characteristics of the parallel algorithm are the following. Firstly, one need not solve a linear system of equations at each iteration if the dimension of the problem is less than or equal to the number of processors available. If the dimension of the problem is bigger than the number of processors available, then

small systems of linear equations are solved in each processor. In this way the message passing is decreased considerably at each iteration. Secondly, a globally convergent technique can be easily implemented in parallel since each processor is solving a different system of nonlinear equations. Moreover, such global procedure need not be the same in each processor; a line search approach could run in certain processors while a trust region could run on others. Another important feature of the parallel algorithm is that functions evaluations are implicitly done in parallel. Thus, considerable savings are obtained over a serial solver.

The main drawback of the serial algorithm is its slow rate of convergence, linearly convergent. With the numerical results obtained we show that this drawback can be circumvented using a parallel implementation.

The work in this paper is presented in the following fashion. In Section 2, we describe the serial algorithm along with its main convergence results. In Section 3 different parallel implementations of the serial algorithm are presented. Emphasis is given to the case where the dimension of the problem is bigger than the number of processors available. In Section 4, we briefly discuss the use of the parallel algorithm to solve unconstrained minimization problems. In Section 5 some numerical results obtained on the Encore/Multimax located at Argonne National Laboratory are presented. Finally, in Section 6 we present future work and draw some conclusions.

2. The serial nonlinear Jacobi algorithm. Consider the following system of nonlinear equations

$$F(x) = 0 \tag{2.1}$$

where F: $\mathbf{R}^n \rightarrow \mathbf{R}^n$ is a continuously differentiable function in an open subset Ω of $\mathbf{R}^n$.

We will assume there exist $x_* \in \Omega$ such that $F(x_*)=0$ and $[F'(x_*)]^{-1}$ exists.

The nonlinear Jacobi algorithm is a particular case of a more general class of algorithms, the nonlinear–SOR algorithms. These algorithms are based in the following idea. A basic step of the nonlinear Gauss–Seidel iteration is to solve the *ith* equation

$$f_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, z, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0 \tag{2.2}$$

for z, and to set

$$x_i^{(k+1)} = z_{m_i}^{(k)} \tag{2.3}$$

where m_i correspond to the number of inner steps performed in solving (2.2). Thus, in order to obtain $x^{(k+1)}$ from $x^{(k)}$, we solve successively the n one–dimensional nonlinear equations (2.2) for $i=1,...,n$. More generally, we may set

$$x_i^{(k+1)} = x_i^{(k)} + w_k(z_{m_i}^{(k)} - x_i^{(k)}) \tag{2.4}$$

in order to obtain a nonlinear SOR method where w_k is a parameter varying with k.

In an analogous fashion, the *kth* stage of the nonlinear Jacobi iteration may be defined by solving the equations

$$f_i(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, z, x_{i+1}^{(k)}, \ldots, x_n^{(k)}) = 0 \quad i=1,...,n \tag{2.5}$$

for z and setting $x_i^{(k+1)} = z_{m_i}^{(k)}$ for $i=1,...n$.

Notice that the above methods have meaning only if the equations (2.2) or (2.5) have unique solutions in some specific domain under consideration. Conditions must be given to ensure that this is true.

We now restrict ourselves to the nonlinear Jacobi iteration. The iterative method used to solve (2.5) plays the role of a secondary iteration, while the Jacobi (SOR) method is the primary iteration. There are different ways of implementing the algorithm based

on the number of inner steps m_i taken to obtain the solution of (2.5) and the iterative method used to solve (2.5). Namely, if $m_i=1$, only one step is carried out to obtain $x_i^{(k+1)}$ from $x_i^{(k)}$ and if Newton's method is used to solve (2.5), we get the one–step Jacobi–Newton method. In this case, $x_i^{(k+1)}$ is given by

$$x_i^{(k+1)} = x_i^{(k)} - \frac{f_i(x^{(k)})}{b_i^{(k)}} \tag{2.6}$$

where $b_i^{(k)}=\frac{\partial f_i}{\partial x_i}(x^{(k)})$ and $x^{(k)}=(x_1^{(k)}, \ldots, x_n^{(k)})$. Notice that the starting point for Newton's method, and for any other iterative method used to solve (2.5), is $z_o^{(k)}=x_i^{(k)}$. It is worth noticing that the one–step Jacobi–Newton method generates the same iterates as the one–step Newton–Jacobi method in which Newton's method is used to solve (1.1) and one step of the Jacobi algorithm is used to solve the linear system (1.2.a).

One can also use a secant method for solving (2.5) as suggested by Wegge [23] and obtain the one–step Jacobi–secant iteration. We just substitute the partial derivative in (2.5) by

$$b_i^{(k)} = \frac{f_i(x^{(k)}) - f_i(x^{(k)} + [x_i^{(k-1)} - x_i^{(k)}]e_i)}{x_i^{(k)} - x_i^{(k-1)}} \tag{2.7}$$

where e_i denotes the ith column of the identity matrix. One could also use the more recent secant implementation proposed by Dennis and Walker [6] in which $b_i^{(k)}$ is allowed to be equal to $b_i^{(k-1)}$ in some particular instances.

Theoretically, more than one inner step when solving (2.5) does not improve the rate of convergence of the algorithm (see [22]). Our numerical experiments show this is the case in the majority of the test problems. However, for certain problems we obtained faster convergence by using more than one inner step.

We now present an informal view of two local convergence results for the serial nonlinear Jacobi algorithm. The first one can be found in [17, Theorem 10.3.5], the second one is due to Dennis and Walker [6, Theorem 4.1]. For both we assume x_o is in a neighborhood of x_* and F satisfies the standard conditions stated at the beginning of this section. Let

$$F'(x) = D(x) - L(x) - U(x) \tag{2.8}$$

be the decomposition of $F'(x)$ into its diagonal, strictly lower-, and strictly upper-triangular parts and assume that $D_*=D(x_*)$ is nonsingular. Let $U_*=U(x_*)$ and $L_*=L(x_*)$. The first result treats the general nonlinear Jacobi iteration where (2.5) is solved by no specific method. The result says essentially that if $\rho(I - D_*^{-1}F'(x_*)) < 1$, then the sequence x_k converges to x_* r-linearly. The second result, which deals with the one-step Jacobi-secant iteration, states that if $\rho(I - D_*^{-1}F'(x_*)) < 1$, then the sequence x_k converges to x_* q-linearly.

An important feature of the nonlinear Jacobi (SOR) method is that it can be extended to block form. Partition x as $x=(x^1, \ldots, x^m)$, with $x^i \in \mathbf{R}^{l_i}$, and group correspondingly, the components f_i of F into mappings $F_i : \mathbf{R}^n \rightarrow \mathbf{R}^{l_i}$, for $i=1,\ldots,m$. Then solving

$$F_i((x^1)^k, \ldots, (x^{i-1})^k, (z), (x^{i+1})^k, \ldots, (x^m)^k) = 0\,, \quad i=1,\ldots,m$$

for (z) describes a nonlinear block Jacobi process in which a complete iteration requires the solution of m nonlinear systems of dimension l_i, $i=1,\ldots,m$. This approach will be used extensively in the parallel implementation to follow.

3. Parallel implementation. For the purpose of this presentation we can assume a parallel computer with or without a shared memory, with p processors and each processor able to sustain a heavy load of floating point computations. Such is the case of the most current parallel computers such as the N–Cube, Encore/Multimax, the Sequent/Balance, the BBN Butterfly and the Alliant to name just a few. We must also assume that a way of transferring data among processors exist, such as the Monitor System [12], the Domino System [16] or the DPUP System [10] among others.

The parallel implementation of the nonlinear Jacobi algorithm is straightforward from (2.5). Each processor will be assigned an index i and for such index it will have the task of computing the solution of (2.5). In this way, the parallel implementation allows the user to work with a different iterative method to solve (2.5) on each processor; the method may be a secant or a Newton method. Moreover, if in (2.5) we are solving a one–dimensional problem, one may use a bisection method combined with Newton's or secant method, such as Brent's algorithm [3]. The global technique to ensure convergence when far away from the solution may also vary among processors and the number of inner steps m_i used for each subproblem (2.5) may also vary. In this way, the parallel algorithm gives the user great flexibility in deciding which implementation to use for a particular problem.

If we try to solve (2.5) on each processor to a given tolerance, we might get idle processors waiting for the most nonlinear functions to converge. In order to deal with this problem, a fixed number of inner steps is allowed in each processor. In this way, a processor stops computing if either a given tolerance is reached, or if the the given number of inner steps is attained. It is worth noticing that in some of the test problems we obtained better convergence results by using more than one inner step.

The most interesting case to consider is when $p \ll n$. In this case we will solve (1.1) by using the nonlinear block Jacobi algorithm as presented in the previous section. We evenly load all the processors with a partition of the components f_i of F (say F_i). In this way, each processor will solve a nonlinear system of equations, thus diminishing the overhead created by the communication among processors since each processor will perform a considerable amount of floating point computation. The dimensions of each subsystem may be different. At this point, one can use a standard serial nonlinear solver in each processor (i.e., MINPACK). Once more a fixed number of inner steps might be appropriate to avoid having idle processors at each iteration.

4. Numerical experiments. All the experiments were performed in an Encore/Multimax located at Argonne National Laboratory at the Advanced Computing Research Facility. The Encore/Multimax has 20 processors with 20 Mbytes of memory. Each processor is a National Semiconductor 32032 chip set running at 10 MHz. The processors are connected via a 64-bit wide bus with a data transfer of 100 Mbytes per second. The operating system is UNIX™.

As a synchronization and communication system among processors we used the FORTRAN version of the Monitors macros developed by Lusk and Overbeek [12]. This system allows one to set up a pool of tasks which are solved in parallel by the processors.

The test problems we used were selected from the standard set of problems in Garbow, Hillstrom and Moré [9]. There are fourteen problems which are systems of nonlinear equations. They are presented with numbers from 1 to 14. We have kept the same numeration to denote those problems. They are:

1. Rosenbrock function
2. Powell singular function
3. Powell badly scaled function
4. Wood function
5. Helical valley function
6. Watson function
7. Chebyquad function
8. Brown almost–linear function
9. Discrete boundary valued function
10. Discrete integral equation function
11. Trigonometric function
12. Variably dimensioned function
13. Broyden tridiagonal function
14. Broyden banded function.

The first five test functions are of dimensions 2,4,2,4,3, respectively, while the remaining test functions are of variable dimension. For more information on these problems see [9]. For problems with variable dimension we decided to run them with dimensions 4, 8, 16, and 32. To problem number 5, the Helical valley function, we added one extra function, $f_4(x)=x_4$ to get convergence with the nonlinear Jacobi method. The nonlinear Jacobi method was unable to solve problems 6,7,8 and 12.

We decided to compare the parallel nonlinear Jacobi algorithm with the best nonlinear equation solver, MINPACK. Minpack's algorithm is well suited for solving small and medium size problems with expensive function evaluations. The set of test problems chosen are of small and medium size, however, their functions are not expensive to evaluate. On the other hand, the nonlinear Jacobi algorithm was designed for large problems and therefore, its parallel implementation will not perform as well in this particular set of problems. We must keep in mind that the main purpose of the numerical experiments is to study different parallel implementations of a linearly convergent algorithm and the comparison of its performance against a quadratically convergent algorithm such as the one in MINPACK. It is not our intention to claim that our algorithm is superior to the MINPACK algorithm. The numerical results will allow us to pinpoint synchronization

bottlenecks in the parallel implementation, possible drawbacks due to a lack of reliability, and the advantage of using this parallel algorithm in problems with a particular structure.

We used MINPACK on the Encore/Multimax to solve the same set of problems. We must emphasize that MINPACK was successful in all fourteen problems with all the different dimensions except number 11. In MINPACK we used the double precision version of HYBRD which solves systems of nonlinear equations by a modification of Powell's hybrid method. In this subroutine the Jacobian is approximated by a forward–difference approximation. For all problems we used the same initial points as presented in Garbow, Hillstrom and Moré [9]. The tolerance for convergence was set at 10^{-8} and it was used to check the stopping criteria. We stopped if either the maximum number of iterations (100) was attained, or if the relative error between two consecutive iterates is less or equal than the tolerance.

Let us now focus our attention on the implementation of the parallel algorithm and the numerical results obtained in the set of problems mentioned above. We implemented the nonlinear block Jacobi algorithm as presented in Section 2. Partition x as $x=(x^1, \ldots, x^m)$, with $x^i \in \mathbf{R}^{l_i}$, and group correspondingly, the components f_i of F into mappings $F_i: \mathbf{R}^n \rightarrow \mathbf{R}^{l_i}$, for $i=1,\ldots,m$. Then solve

$$F_i((x^1)^k, \ldots, (x^{i-1})^k, (z), (x^{i+1})^k, \ldots, (x^m)^k) = 0\,, \quad i=1,\ldots,m \tag{4.1}$$

for (z). This requires the solution of m nonlinear systems of dimension l_i, $i=1,\ldots,m$ at each iteration. Each of these subsystems is solved by a different processor using the same subroutine from MINPACK as mentioned above. As we pointed out in Section 3 in order not to have idle processors during the computation a predetermined number of inner

steps are allowed in each call to MINPACK. We did runs with 1, 5, and 10 inner steps. We stopped solving the sub–problem if either the tolerance was achieved or if the maximum number of inner steps was performed. The dimensions of the subproblems were equal; however, one can use different dimensions for different subproblems.

For a given dimension of a problem we experimented with different partitions. For instance, for a problem of dimension 32 we ran 5 different partitions: with 2 blocks of dimension 16; with 4 blocks of dimension 8; with 8 blocks of dimension 4; with 16 blocks of dimension 2; with 32 one–dimensional blocks. For each given partition we used one processor per block.

One additional advantage of the parallel algorithm is that function evaluations are implicitly performed in parallel. This is because when solving (4.1) only the functions involved in this group need to be evaluated. Therefore, great savings in time are obtained over the serial algorithm and over MINPACK.

For timing the experiments we used the FORTRAN function *etime* (UNIX^{TM}) which returns elapsed runtime in seconds. It has an array of two elements as argument. In the first element it returns user time and in the second element it returns system time. In all the experiments we performed we only used the first element: the user time.

Let us now discuss the numerical results. The entire set of numerical results can be found in [8]. We present here a sample of the most interesting problems. Each table shows the following columns:

NPRO number of processors

NBLK number of blocks used in the partition on subproblems

NSTP number of inner steps used to solve the subproblem

NFEV number of functions evaluations

INFO stopping message

TIME elapsed runtime. System time is not accounted for.

S_p speedup over the serial algorithm

S_p' speedup over MINPACK

E_p efficiency of the parallel implementation.

The number of processors in NPRO can vary from 1 (serial algorithm running) to 20 (maximum number of processors available on the Encore/Multimax). MINPACK results are always located in the first row of each table: NPRO is one and NBLK is one. When NPRO is equal to one and NBLK is different from one we are using the serial nonlinear Jacobi algorithm. The number of inner steps NSTP can be either 1, or 5, or 10. We only show the optimal case in the tables. NSTP=1 is often enough to get good and fast convergence, although for some problems more than one inner step was necessary. The number of function evaluations are calculated by counting each function evaluation f_i. Thus, for MINPACK we multiplied the number of function evaluations by n (dimension of the problem). There are several stopping messages in MINPACK; however, the only two we came across were 1 for a successful run and 4 when the iteration is not making good progress. We use INFO=2 whenever the maximum number of iterations is attained. All timings are given in seconds in TIME. For any given parallel algorithm there are three numbers which give an idea on how well the parallel algorithm is performing. The speedup S_p is defined by

$$S_p = \frac{\textit{running time for the serial algorithm}}{\textit{running time for the parallel algorithm using p processors}},$$

where the serial algorithm is the nonlinear Jacobi method using different block–partitions and the parallel algorithm is the parallel version of the Jacobi algorithm using the same block–partition. Obviously this number can never be bigger than the number of processors used in the computation. In order to know how much faster is the parallel algorithm *vis a vis* the best serial algorithm we calculate S_p' which is defined by

$$S_p' = \frac{\textit{running time of the best serial algorithm}}{\textit{running time of the parallel algorithm using p processors}},$$

where the best serial algorithm is MINPACK. Whenever in this column we find a zero it means that MINPACK was faster than the current combination of processor, blocks and steps. Whenever we find in this column ∞ it means that MINPACK failed to converge and the parallel algorithm was successful. This only occurs in problem 11.

In order to know the efficiency of our parallel implementation we calculate E_p, the efficiency of the algorithm, defined by

$$E_p = \frac{S_p}{\textit{number of processors } p}.$$

This number can never be bigger than one. Furthermore, the numbers in the last three columns of tables I through III have been rounded to two decimal places.

On table IV we have summarized all the results of the parallel Jacobi algorithm and the corresponding results from MINPACK. For each given problem we included the combination of processors, number of blocks and steps that gave the best timing. In column MPACK we present the MINPACK timing and in JAC–P we present the timing for the parallel algorithm. In the last column we present the number S_p'. In table V we present the analogous data for the serial nonlinear Jacobi algorithm. Columns JAC–S and JAC–

P represent the timings for the serial and the parallel algorithms respectively. The last column corresponds to the speedup obtained.

Problem: 9 Dimension: 8

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	1	-	112	1	0.400	-	-	-
1	2	1	3424	1	5.400	-	-	-
1	4	1	4168	1	8.150	-	-	-
1	8	1	6184	1	17.300	-	-	-
2	2	1	3424	1	2.867	1.88	0.	0.94
2	4	1	4168	1	4.350	1.87	0.	0.94
2	8	1	6184	1	9.333	1.85	0.	0.93
4	4	1	4168	1	2.483	3.28	0.	0.82
4	8	1	6184	1	5.200	3.33	0.	0.83
8	8	1	6184	1	3.267	5.30	0.	0.66

Table I

Problem: 10 Dimension: 32

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	1	–	1216	1	12.783	–	–	–
1	2	1	6112	1	41.450	–	–	–
1	4	1	3552	1	22.317	–	–	–
1	8	1	2048	1	12.867	–	–	–
1	16	1	1632	1	10.967	–	–	–
1	32	1	1312	1	9.883	–	–	–
2	2	1	6112	1	21.017	1.97	0.	0.99
2	4	1	3552	1	11.233	1.99	1.14	0.99
2	8	1	2048	1	6.450	1.99	1.98	1.00
2	16	1	1632	1	5.500	1.99	2.32	1.00
2	32	1	1312	1	5.033	1.96	2.54	0.98
4	4	1	3552	1	5.783	3.86	2.21	0.96
4	8	1	2048	1	3.317	3.88	3.85	0.97
4	16	1	1632	1	2.833	3.87	4.51	0.97
4	32	1	1312	1	2.617	3.78	4.88	0.94
8	8	1	2048	1	1.750	7.35	7.30	0.92
8	16	1	1632	1	1.500	7.31	8.52	0.91
8	32	1	1312	1	1.400	7.06	9.13	0.88
16	16	1	1632	1	0.917	11.96	13.94	0.75
16	32	1	1312	1	0.833	11.86	15.35	0.74
20	32	1	1312	1	1.117	8.85	11.44	0.44

Table II

Problem: 14 Dimension: 32

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	1	-	1632	1	16.850	-	-	-
1	2	1	4896	1	10.250	-	-	-
1	4	1	2848	1	4.783	-	-	-
1	8	1	2048	1	3.417	-	-	-
1	16	1	1632	1	3.383	-	-	-
1	32	1	1440	1	4.400	-	-	-
2	2	1	4896	1	5.183	1.98	3.25	0.99
2	4	1	2848	1	2.417	1.98	6.97	0.99
2	8	1	2048	1	1.767	1.93	9.54	0.97
2	16	1	1632	1	1.750	1.93	9.63	0.97
2	32	1	1440	1	2.317	1.90	7.27	0.95
4	4	1	2848	1	1.283	3.73	13.13	0.93
4	8	1	2048	1	0.950	3.60	17.74	0.90
4	16	1	1632	1	0.950	3.56	17.74	0.89
4	32	1	1440	1	1.233	3.57	13.67	0.89
8	8	1	2048	1	0.567	6.03	29.72	0.75
8	16	1	1632	1	0.567	5.97	29.72	0.75
8	32	1	1440	1	0.700	6.29	24.07	0.79
16	16	1	1632	1	0.367	9.22	45.91	0.58
16	32	1	1440	1	0.467	9.42	36.08	0.59
20	32	1	1440	1	0.850	5.18	19.82	0.26

Table III

MINPACK Timing Table

PROB	DIM	NPRO	NBLK	NSTP	MPACK	JAC-P	S_p'
1	2	2	2	5	0.267	0.033	8.09
2	4	4	4	1	1.100	0.576	1.94
3	2	2	2	1	2.317	0.25	9.27
4	4	4	4	10	2.333	0.317	7.36
5	4	2	2	1	0.783	0.267	2.93
8	4	2	2	1	0.267	1.750	0.
9	4	2	2	1	0.133	0.783	0.
9	8	4	4	1	0.400	2.483	0.
9	16	4	4	1	1.483	9.633	0.
9	32	2	2	1	6.483	-	-
10	4	4	4	1	0.167	0.167	1.
10	8	8	8	1	0.533	0.217	2.46
10	16	16	16	1	2.300	0.35	6.57
10	32	16	32	1	12.783	0.833	15.35
11	4	2	2	1	-	0.717	∞
11	8	8	8	1	-	0.717	∞
11	16	16	16	1	-	0.783	∞
11	32	16	32	1	-	2.000	∞
13	4	4	4	1	0.267	0.317	0.
13	8	4	4	1	0.750	0.367	2.04
13	16	8	8	1	2.767	0.450	6.15
13	32	16	16	1	10.967	0.617	17.77
14	4	4	4	1	0.383	0.117	3.27
14	8	8	8	1	1.45	0.150	9.67
14	16	8	8	1	4.700	0.250	18.80
14	32	16	16	1	16.85	0.367	45.91

Table IV

Jacobi Timing Table

PROB	DIM	NPRO	NBLK	NSTP	JAC-S	JAC-P	S_p
1	2	2	2	5	0.05	0.033	1.52
2	4	4	4	1	1.7	0.567	3.0
3	2	2	2	1	0.383	0.250	1.53
4	4	4	4	5	1.017	0.317	3.21
5	4	2	2	1	0.35	0.267	1.31
8	4	4	4	1	5.667	1.867	3.04
9	4	4	4	1	2.817	0.883	3.19
9	8	8	8	1	17.300	3.267	5.30
9	16	4	4	1	32.567	9.633	3.38
9	32	-	-	-	-	-	-
10	4	4	4	1	0.517	0.167	3.10
10	8	8	8	1	1.100	0.217	5.07
10	16	16	16	1	3.267	0.350	9.33
10	32	16	16	1	10.967	0.917	11.96
11	4	2	2	1	1.350	0.717	1.88
11	8	8	8	1	4.250	0.717	5.93
11	16	16	16	1	8.350	0.783	10.66
11	32	16	32	1	25.417	2.000	12.71
13	4	4	4	1	0.933	0.317	2.94
13	8	8	8	1	2.05	0.417	4.92
13	16	8	16	1	4.400	0.617	7.13
13	32	16	32	1	10.083	1.117	9.03
14	4	4	4	1	0.350	0.117	2.99
14	8	8	8	1	0.700	0.150	4.67
14	16	16	16	1	1.850	0.283	6.54
14	32	16	32	1	4.4	0.467	9.42

Table V

We have added several figures to illustrate the behavior of the parallel algorithm. The Figures are in logarithmic scale. In Figure 1, we notice that no matter how many processors we use to solve Problem 9, n=8, MINPACK is always faster than the parallel algorithm. In Figure 2, Problem 10 with dimension 16, we notice that MINPACK is faster than the serial algorithm but using two processors the parallel Jacobi becomes faster

than MINPACK. In Figures 3 and 4, a common paradigm in parallel computation occurs. Once one starts using too many processors on a problem, one starts running slower. As we see in Figures 3 and 4 the optimal number of processors seems to be around 15 on Problems 13 and 14 with $n=32$; already using 16 processors represents a lost in speedup. In Figure 5, we present for Problem 10 with $n=32$, the timings for all the different partitions. We notice that using 16 or 32 block partitions we get almost identical results. In all figures we notice the linear speedup characteristic of a parallel algorithm.

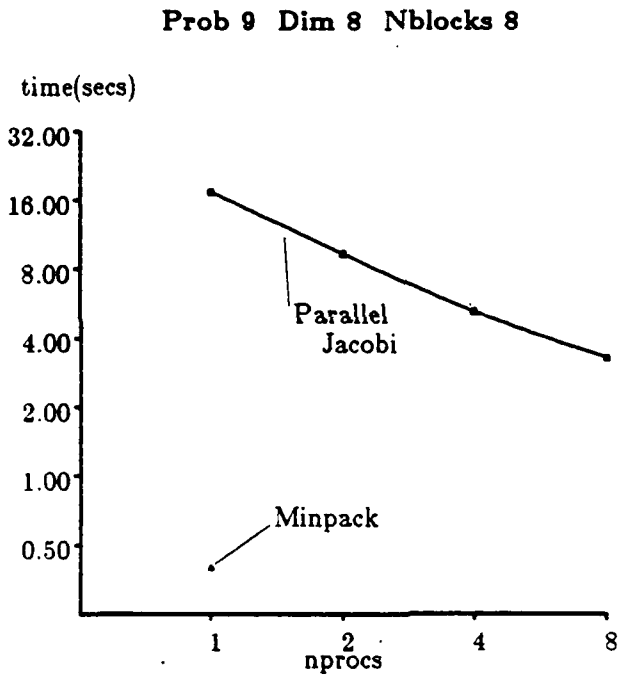


Figure 1. For this problem the parallel Jacobi algorithm is never faster than MINPACK. The higher speedup is 5.3 using 8 processors.

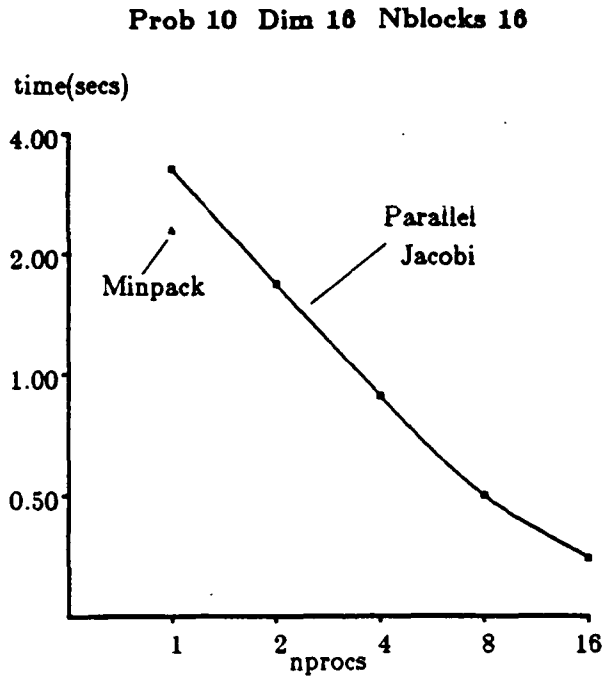


Figure 2. For this problem the parallel Jacobi algorithm is 6.57 times faster than MINPACK. The higher speedup is 9.33 using 16 processors.

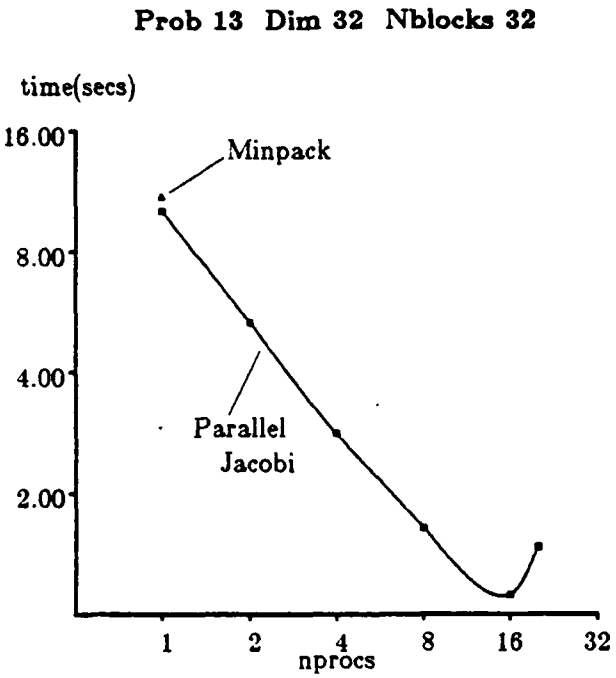


Figure 3. For this problem the parallel Jacobi algorithm is 9.82 times faster than MINPACK. The higher speedup is 9.03 using 16 processors.

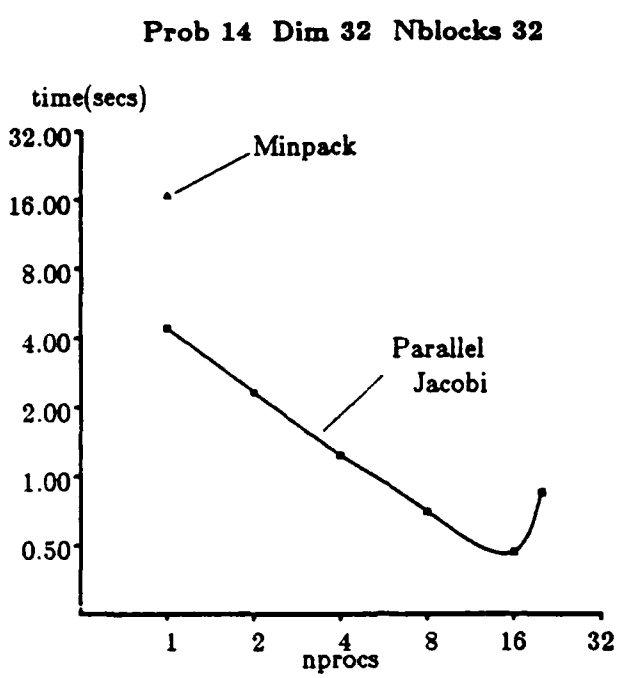


Figure 4. For this problem the parallel Jacobi algorithm is 36.08 times faster than MINPACK. The higher speedup is 9.42 using 16 processors.

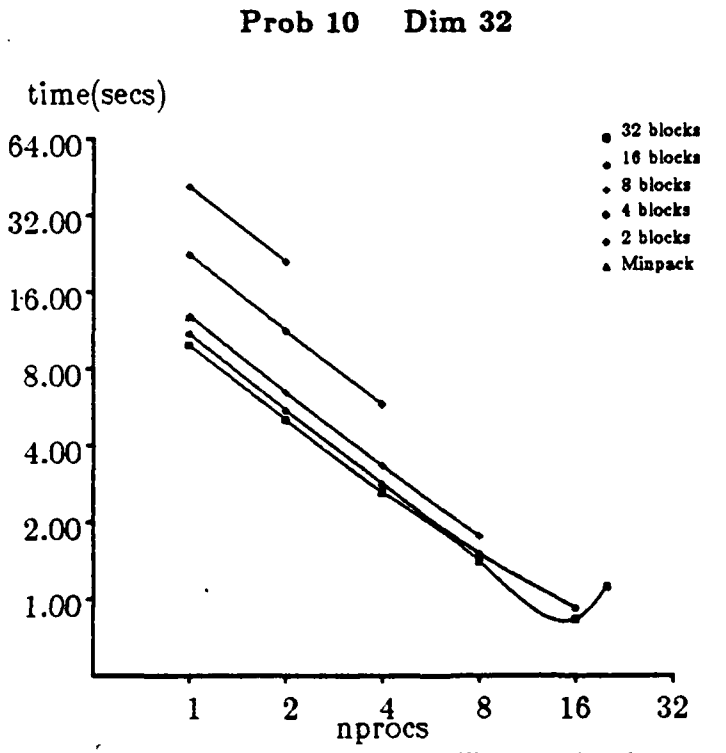


Figure 5. We are using different set of block partitions with different number of processors.

We were able to run 10 problems out of the standard 14 problems for nonlinear equations in [9]. We can say that the nonlinear Jacobi algorithm is 71% reliable on this set of test problems. In fact, this is the main disadvantage of the algorithm. On the other hand, in most cases whenever the method converges, the parallel implementation outperforms MINPACK. As we pointed out earlier this is partly due to the fact that function evaluations on this set of test problems are not expensive. As we can see in table 4 only in problem 9 does MINPACK outperform the parallel algorithm. The parallel algorithm is considerably faster than MINPACK; in particular, as the dimension increases the parallel algorithm seems to work better. The outstanding performance of the parallel Jacobi algorithm on problems 13 and 14 is due to the particular structure of the Jacobian, tridiagonal and banded respectively. The nonlinear Jacobi performs extremely well on problems whose Jacobians have a particular structure centered around the diagonal, tridiagonal or banded Jacobians. Nevertheless, the performance in general is quite interesting and extremely promising. On this set of problems we obtained on average an speedup of 10. The speedups in table 5 show the considerable improvement that the parallel implementation produced over the standard serial algorithm. We obtain a high of 11.96 using 16 processors in problem 10 with dimension 32.

In only two problems we set the parameter w_k, as defined in (2.4), to a value different from 1.0 in order to get faster convergence. On Problem 8, w_k=0.6 and on Problem 10, w_k=0.9. In table 6 we present the results of problem 10 with dimension 32 using the standard value w_k=1.0. We notice there is a considerable gain on speedup for a small change on this parameter.

Problem: 10 Dimension: 32

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	1	-	1216	1	12.783	-	-	-
1	2	1	5504	1	37.683	-	-	-
1	4	1	3904	1	24.683	-	-	-
1	8	1	2944	1	18.450	-	-	-
1	16	1	2272	1	15.050	-	-	-
1	32	1	1952	1	14.817	-	-	-
2	2	1	5504	1	18.967	1.99	0.	0.99
2	4	1	3904	1	12.383	1.99	1.03	1.00
2	8	1	2944	1	9.300	1.98	1.37	0.99
2	16	1	2272	1	7.650	1.97	1.67	0.98
2	32	1	1952	1	7.567	1.96	1.69	0.98
4	4	1	3904	1	6.383	3.87	2.00	0.97
4	8	1	2944	1	4.783	3.86	2.67	0.96
4	16	1	2272	1	3.950	3.81	3.24	0.95
4	32	1	1952	1	3.917	3.78	3.26	0.95
8	8	1	2944	1	2.550	7.24	5.01	0.90
8	16	1	2272	1	2.117	7.11	6.04	0.89
8	32	1	1952	1	2.133	6.95	5.99	0.87
16	16	1	2272	1	1.217	12.37	10.50	0.77
16	32	1	1952	1	1.217	12.18	10.50	0.76
20	32	1	1952	1	1.283	11.55	9.96	0.58

Table 6

Additional experiments.

We try to make the method more robust and reliable by using multi-splitting techniques such as the ones developed by O'Leary and White [15] for linear systems. The main idea behind this approach is to be able to use more information from the Jacobian at each iteration by using more processors to perform additional computations.

One idea is to use a card-dealer technique to assign each function f_i to each processor. In this way, we will be able to use part of the Jacobian matrix which lies outside the diagonal. This particular block- partition could be used concurrently with other block-

partitions. Such partitions could be made of blocks with different dimensions. At each iteration we get several solution vectors depending on the number of partitions we have used. In order to use all this information we take as our next iterate a convex combination of all the solution vectors at each iteration. In the following figure we show two different partitions running concurrently on a Jacobian of dimension six using three processors for each partition. Let us say that at the *kth* iteration partition one gives the solution vector x_1^k and partition two gives x_2^k, then our next iterate $x^{k+1} = x_1^k/2 + x_2^k/2$.

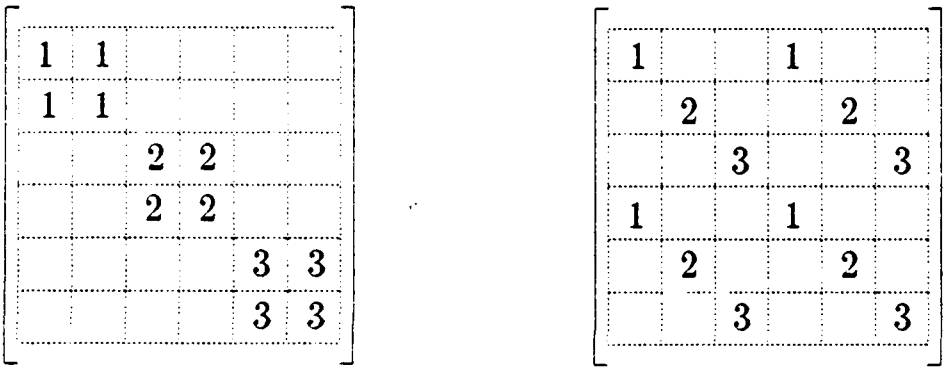


Figure 6. Different partitions running concurrently at each iteration.

We decided to test this idea on problem 9 with dimension 16. We decided to use a standard partition of 4 blocks each of dimension 4 and a second partition of 4 blocks with dimensions 5, 5, 5 and 1.

Problem: 9 Dimension: 16

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	8	1	14576	1	47.183	-	-	-
2	8	1	14576	1	27.083	1.74	0.	0.87
4	8	1	14576	1	14.233	3.32	0.	0.83
8	8	1	14576	1	9.017	5.23	0.	0.65

Table 7

The method is still slower than MINPACK. However, if we compare the results of table 7 above and table 9 in [8] we notice that we succeed in getting convergence using 8 processors and the method performs faster than the best case in [8].

We have also experimented with the following idea. Every other iteration we use a different partition with the same number of processors. We may also choose to do this every two or five iterations. One of the partitions may use only one block in which case we will be doing a MINPACK step every other iteration. We tried out this idea on problem 9 with dimension 8. One partition has 8 blocks with dimension 1 each, the other partition has 2 blocks with dimensions 7 and 1. We change partitions every five iterations. The results follow.

Problem: 9 Dimension: 8

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	8	1	2255	1	4.700	-	-	-
2	8	1	2255	1	2.967	1.58	0.	0.79
4	8	1	2255	1	2.200	2.14	0.	0.53
8	8	1	2255	1	1.800	2.61	0.	0.33

Table 8

We also tried a partition of 8 blocks with dimension 1 each and a partition of one block with dimension 8. Hence, we are doing a MINPACK step every five iterations.

Problem: 9 Dimension: 8

NPRO	NBLK	NSTP	NFEV	INFO	TIME	S_p	S'_p	E_p
1	8	1	702	1	1.25	-	-	-
2	8	1	702	1	0.867	1.44	0.	0.72
4	8	1	702	1	0.683	1.83	0.	0.46
8	8	1	702	1	0.583	2.14	0.	0.27

Table 9

With this technique we are able to perform more than 4 times faster than using the standard partition procedure (see [8]). Although MINPACK is still faster (0.400 secs) we have decreased considerably the execution time. It is interesting to note in tables 8 and 9 that the speedups S_p are not as big as in [8]. This was predictable since every five iterations we can have up to seven processors idling.

5. Conclusions and future work. The parallel implementation of the block nonlinear Jacobi algorithm has given us better results than we expected. It has given us a way to solve systems of nonlinear equations in parallel. To our knowledge this is the first time a parallel algorithm for solving this type of problems has had such a performance in a real parallel computer.

It is interesting to notice that the main idea behind the algorithm is the fundamental idea behind some powerful parallel algorithms, namely, divide and conquer. Furthermore, it is worth noticing that although the algorithm is only linearly convergent it performs faster than a quadratically convergent algorithm on certain problems with a particular structure and on certain other problems where the function evaluations are not expensive.

The main disadvantage of the algorithm is its lack of reliability. With only a 71% of success rate it cannot be thought as a way of solving nonlinear equations. However, there are several ways of improving the convergence of the method at no additional cost. Along these lines some preliminary tests were presented at the end of the previous section. Other approaches are currently being tested and will be part of a future report.

Some more testing is certainly necessary. In particular, we will try to study the behavior of the algorithm using initial points that are farther away from the solution. We will also implement the algorithm to solve unconstrained minimization problems.

One of the advantages of using the Monitors macros is that they are portable and therefore, the code which is running on the Encore/ Multimax will run on any other parallel computer where the macros have been installed. This is the case, for instance, for the Alliant FX/8 located at Argonne National Laboratory. The Alliant is a machine which is more suitable for numerical computations because it allows one to use concurrency and vectorization in each processor. We decided to start our experiments on the Encore/Multimax because this machine has 20 processors in contrast with the Alliant which has only 8 processors. The numerical results of these experiments on the Alliant will appear on a forthcoming report.

Acknowledgment: The author wishes to thank the Advanced Computing Research Faciiity at Argonne National Laboratory for the use of their parallel computers which was fundamental to carry out this research. Also he wishes to thank E.L. Lusk and R.A. Overbeek for teaching him the use of Monitors. The author is indebted to Allen Goldberg and his NRL group from whom the author received partial support on this research. The author is grateful to Dianne O'Leary who read the first draft of the manuscript and made numerous observations and changes which made the paper more readable.

References

[1] G.M. Baudet, *Asynchronous Iterative Methods for Multiprocessors*, J. of the ACM, 25, 1978, 226–244.

[2] A. Bojanczyk, *Optimal Asynchronous Newton method for the solution of nonlinear equations*, J. of the ACM, 31, 1984, pp. 792–803.

[3] R. Brent, *Algorithms for minimization without derivatives*, Prentice Hall, Englewood Cliffs, NJ, (1973).

[4] C. Bryan, *On the convergence of the method of nonlinear simultaneous displacements*, Rend. Circ. Mat. Palermo 13, pp. 177–191, (1964).

[5] Chazen and Miranker, "Chaotic Relaxation," Lin Alg and Applics 2, 1969, 199–222.

[6] J.E. Dennis Jr. and H.F. Walker, *Local convergence theorems for quasi-Newton methods*, TR 476–(141–171–163)–1, Rice Univ., Math. Sci. Dept., Houston, Tx., (1983).

[7] M.A. Franklin and I.N. Katz, *Two strategies for root finding on multiprocessor systems*, SIAM J. Sci. Stat. Comp., 6, 2, pp. 314–333, (1985).

[8] R. Fontecilla, *A parallel nonlinear Jacobi algorithm for solving nonlinear equations*, TR. 1807, Dept. of Computer Science, Univ. of Maryland, College Park, Maryland, (1987).

[9] B.S. Garbow, K.E. Hillstrom and J.J. Moré, *Testing unconstrained optimization software*, ACM Trans. Math. Soft., 7, 1, pp. 17–41, (1981).

[10] T.J. Gardner, I.M. Gerard, C.R. Mowers, E.N. Nemeth and R.B. Schnabel, *DPUP: A distributed processing utilities package*, TR. CS–CU–337–86, Univ. of Colorado, Boulder, Colorado, (1986).

[11] H. Lieberstein, *A numerical test case for the nonlinear overrelaxation algorithm*, Math. Res. Center Rept. 122, Univ. of Wisconsin, Madison, Wisconsin, (1960).

[12] E.L. Lusk and R.A. Overbeek, *Use of Monitors in FORTRAN: A tutorial on the Barrier, Self-Scheduling Do-Loop, and Askfor Monitors*, TR. ANL–84–51, Argonne National Laboratory, Argonne, Illinois, (1984).

[13] W.L. Miranker, *A survey of parallelism in numerical analysis*, Siam Rev., 13, 4, pp. 524–547, (1971).

[14] W.L. Miranker, *Parallel methods for solving equations*, Math. and Comp. in Simulation, XX, pp. 93–101, (1978).

[15] D.P. O'Leary and R.E. White, *Multi-splittings of matrices and parallel solution of linear systems*, SIAM J. Alg. Disc. Meth., 6, 4, pp. 630–640, (1985).

[16] D.P. O'Leary, G.W. Stewart and R. van de Geijn, *DOMINO A message passing environment for parallel computation*, TR. 1648, Univ. of Maryland, College Park, Maryland, (1986).

[17] J.M Ortega and W.C. Rheinboldt, *Iterative solutions of nonlinear equations in several variables*, Academic Press, New York, (1970).

[18] J.M. Ortega and M. Rockoff, *Nonlinear difference equations and Gauss-Seidel type iterative methods*, SIAM J. Numer. Anal. 3, pp. 497–513, (1966).

[19] J.M. Ortega, R.G. Voigt, *Solution of partial differential equations on vector and parallel computers*, SIAM Rev., 27, 2, pp. 149–240, (1985).

[20] S. Schechter, *Iteration methods for nonlinear problems*, Trans. Amer. Math. Soc. 104, pp. 179–189, (1962).

[21] S. Schechter, *Relaxation methods for convex problems*, SIAM J. Numer. Anal. 5, pp. 601–612, (1968).

[22] R. Voigt, Rates of convergence for iterative methods for nonlinear systems of equations, Ph.D. Diss., Univ. of Maryland, College Park, Maryland, (1969).

[23] L. Wegge, *On a discrete version of the Newton-Raphson method*, SIAM J. Numer. Anal. 3, pp. 134–142, (1966).

[24] R.E. White, *Parallel algorithms for nonlinear problems*, SIAM J. Alg. Disc. Meth., 7, 1, pp. 137–149, (1986).

STOCHASTIC MODELING FOR IMPROVED WEAPON PERFORMANCE

Mr. James Cantor
Mr. Steve Carchedi
Mr. Bruce Gibbs
of the
Computational Engineering Company
Laurel, Maryland

Mr. John Groff
Mr. Anthony Baran
of the
Ballistic Research Laboratory
Aberdeen Proving Ground,
Maryland

Mr. Herbert Cohen
of the
Army Materiel Systems Analysis Activity
Aberdeen Proving Ground, Maryland

1.0 ABSTRACT

This paper discusses the work jointly performed by the U.S. Army Material Systems Analysis Activity, the Computational Engineering Company and the Ballistic Research Laboratory in the application of time series analysis and modern control theory to characterize armored vehicle weapon tube flexure. The motivation for performing this work stems from the fact that gun bend related errors have a significant affect on fire-on-the-move delivery accuracy. Hence the ability to predict the precise location of the weapon's muzzle as a function of time in terms of it's past and current history as well as other sensor measurements could significantly enhance weapon system accuracy.

Previous efforts to develop muzzle flexure prediction algorithms have generally relied on purely deterministic techniques. That is gun flexure was mathematically characterized by deterministic differential equations that were a function of such parameters as weapon angle position, rate and acceleration, linear acceleration, and bending of the gun tube. In the case of gun dynamics this approach tended to be unsuitable for practical implementation because:

- o mathematically they may be extremely complicated,
- o they do not account for modeling and measurement uncertainties, and
- o they lack the robustness of being adaptive.

This study discusses the preliminary work that has been performed to develop practical algorithms that address the above problems. The overall approach was to:

o apply time series analysis techniques to strain gage and other test data obtained from the M1 Combat Tank mounting a 105mm weapon system and tested over a special Aberdeen Proving Ground Test Course,

o develop auto-regression/moving average (AR/MA) models of the test data to characterize dynamic weapon flexure, and

o convert the AR/MA models to adaptive Kalman filter prediction algorithms.

The paper concludes with a discussion of future modeling and field testing necessary to refine the existing Kalman filter/predictor algorithms and to incorporated a physical model into the filter structure.

2.0 INTRODUCTION

Since gun barrel bending contributes significantly to the total projectile error budget, efforts to predict the precise location of the gun muzzle as a function of time in terms of its past history could significantly improve the accuracy of the weapon system. Previous efforts by the U.S. Army Ballistic Research Lab (BRL) to develop precision aim techniques (PAT) have used a deterministic approach. Specifically, the gun motion was assumed to be described as a function of gun turret angles, angular rates and accelerations, tank vertical acceleration and bending (and bending rates) of the gun tube. The differential equation used to predict the position of the gun muzzle at projectile exit was derived using simple geometry and the equation for the fundamental bending mode. Although the deterministic approach is promising, it has not performed well in field tests at longer (e.g., 20 milliseconds) in-bore times.

An alternate approach is to use only strain gauge (gun tube deflection) and servo error data and model the gun deflection as a Markov process: a linear system driven by white noise. It is this stochastic approach which was investigated here.

In this paper we describe an adaptive model identification algorithm for predicting gun deflection as the projectile leaves the tube. The adaptability is necessary because of the potential great variability in the gun motion due to tank velocity or variation in terrain (e.g., surfaced road to rough ground). Efficient operation is desirable as the algorithm could possibly serve as the basis of a real-time gun inhibit algorithm.

The paper is in six sections. In Section 3, a discussion of the data utilized for modeling is given. In Section 4, the technical approach is outlined and the method used for evaluating the algorithm is described in Section 5. The paper is concluded in Section 6 by a brief discussion and suggestions for future investigations.

3.0 TEST DATA

The available data was obtained from strain gauge and digital control transformer (DCT) sensors mounted on the gun tube of a heavy tank in wide use by the Army. The strain gauge measured the gun tube deflection while the DCT measured the angle of the gun tube with respect to the turret (see Figure 1). Both sensors were sampled at 250 Hz.

criterion of minimizing the prediction error. A good summary of ARMA modeling techniques can be found in [1].

4.1 Data Analysis

The data was first carefully examined to determine general characteristics and to identify statistically (locally) stationary segments. By comparing the data to a schematic of the bump course, segments representative of various physical situations or environments could be selected. These segments provided the means to investigate the spectral content as well as evaluate the eventual design (see Section 4).

Computation of data power spectra using the periodogram and the maximum entropy method (MEM) was performed to identify the dominant spectral bands and the bandwidth of these spectra. This analysis proved useful in relating the observed spectral content to the physical effects as well as in the determination of the appropriate model and approximate model order. Further, an important conclusion based on the spectral analysis is that tank speed had little effect on observed spectral frequencies. An estimate of the power spectrum of the muzzle error using MEM for a segment consisting primarily of small bumps is given Figure 2.

4.2 Identification

It is assumed that an ARMA (p,q) model is sufficiently general to model both the DCT and strain gauge data. In equation (1), a_i, i=1,2,...,p and b_i, i=1,2,...,q denote respectively the autoregressive (AR) and moving average (MA) coefficients, p and q are the AR and MA orders, and w(k) is a zero mean, unit variance Gaussian white noise sequence.

$$z(k) = -\sum_{i=1}^{p} a_i z(k-i) + \sum_{i=1}^{q} b_i w(k-i) + b_0 w(k) \qquad (1)$$

The use of the model for prediction therefore initially requires estimation of the orders and coefficients. The autoregressive order (p) was estimated using a technique due to Cadzow [2] based on determining the effective rank of an associated overdetermined ARMA autocorrelation matrix. (The term overdetermined refers to AR and MA orders selected for estimation which are much larger than the true unknown orders.) As far as could be determined, no similar technique is available for estimating the MA order and for this reason, Cadzow's suggestion of simply setting the MA order equal to AR order was implemented.

As is well known [3]-[4], estimation of the AR coefficients under a least squares criterion results in a linear system of equations to be solved. However, the MA estimation is a nonlinear system. For this reason, the basic approach to coefficient estimation was to approximate the ARMA process by a large order autoregression. (Note that any ARMA process can be represented by an AR process of possibly infinite forder.) This viewpoint was adopted due to the severe computational constraints. The technique

(4 milliseconds).

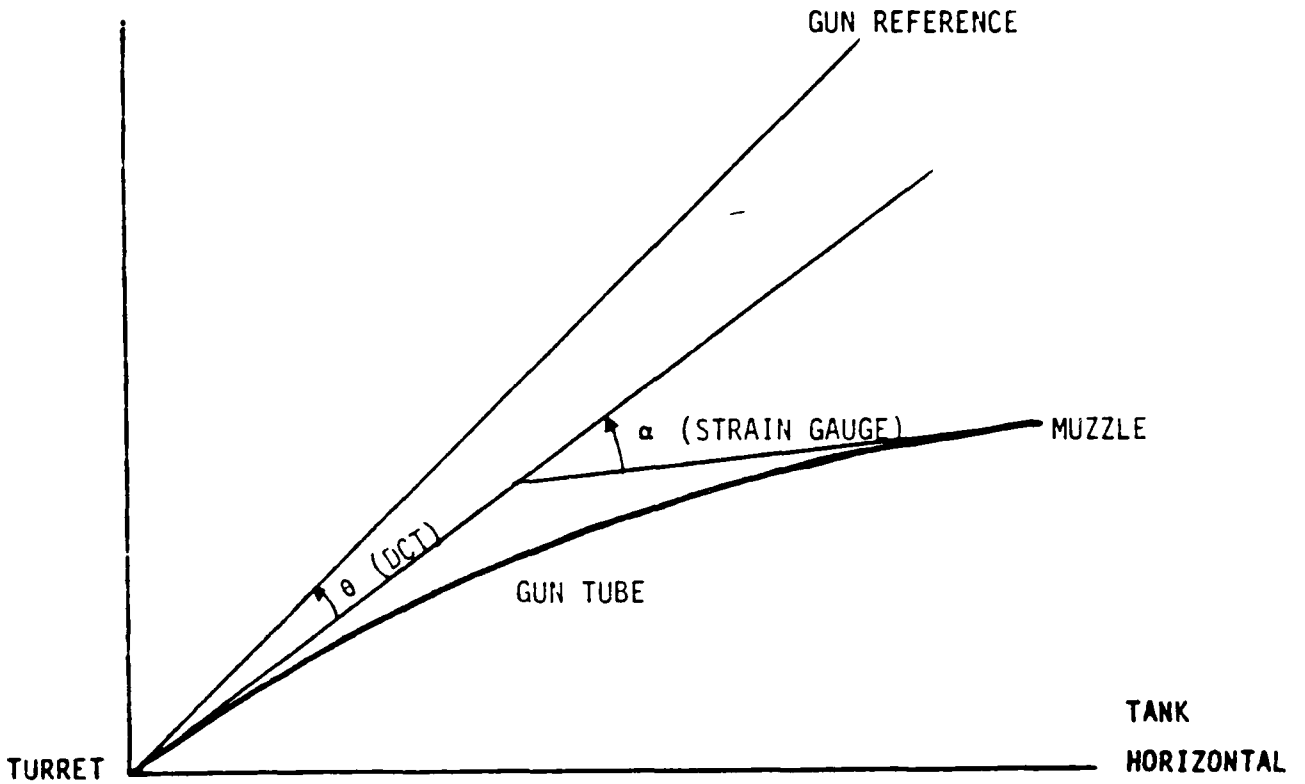


Figure 1. Geometry of Gun Tube Deflection

There were four tests available for analysis. Each test was conducted on the Profile IV bump course at Aberdeen Proving Ground at speeds of 5, 15, 22, and 30 miles per hour (one test at each speed). The course consists of approximately 460 feet of triangular and small wooden bumps up to 12 inches high with gravel lead-in and exit areas. The Profile IV course is considered to be one of the most severe tests of a tank's ability to point the gun accurately while traversing rough terrain.

4.0 TECHNICAL APPROACH

We next describe the technical approach implemented for strain gauge, DCT, and resultant muzzle error identification and prediction. The basic approach was to model the data as a Gauss-Markov process. Specifically, based on the spectral analysis (discussed in 4.1), it was assumed that the data was best modeled as an autoregressive moving average (ARMA) model (defined below) whose parameters are selected under the

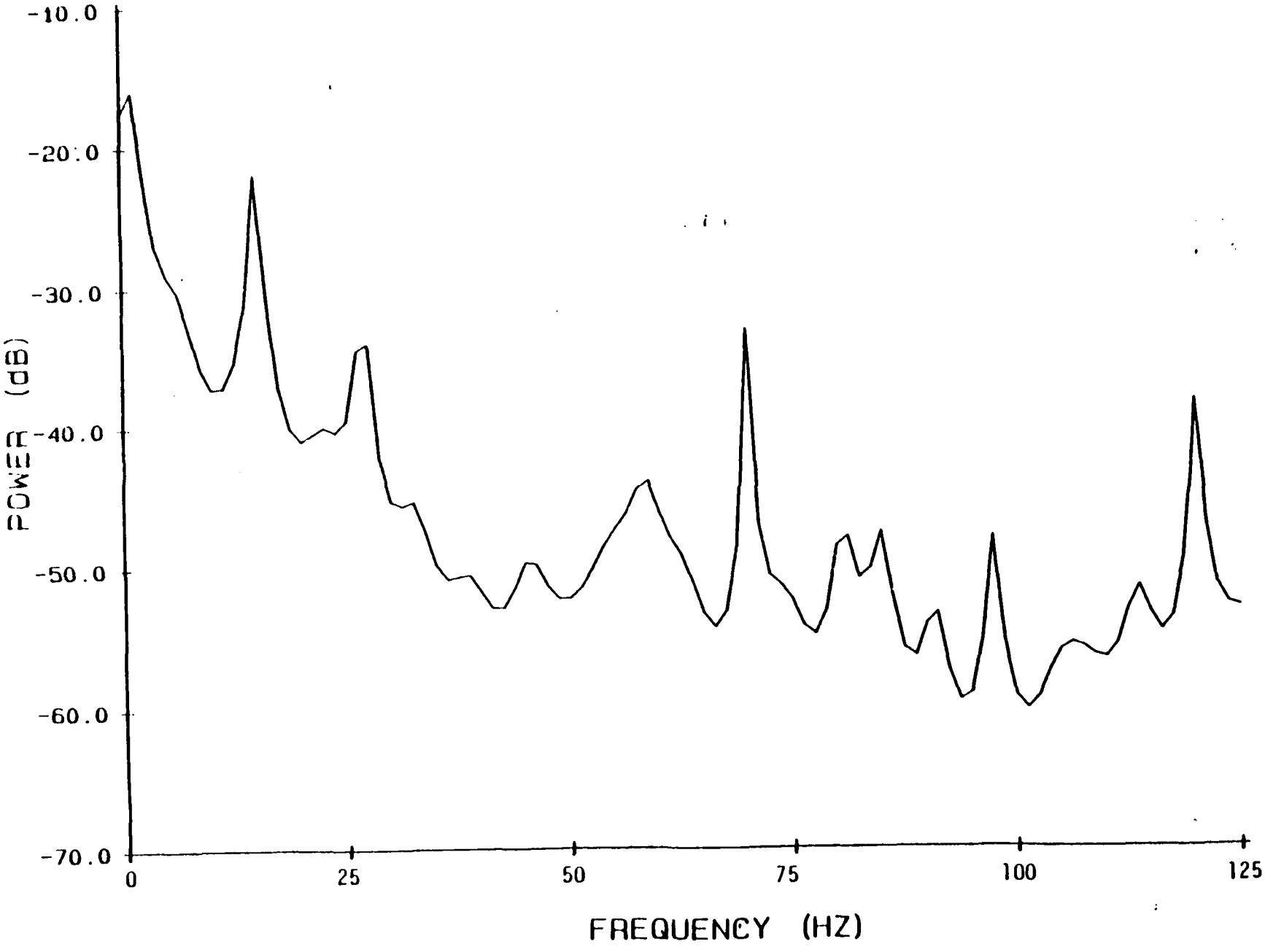


Figure 2. Muzzle Error Power Spectrum for 5 mph Test (Maximum Entropy (MEM) Technique)

implemented essentially follows an approach described by Graupe et al. [5] which involves estimating the coefficients of a high-order AR model and transforming it to a lower order ARMA Model. The AR coefficients were computed from MEM utilizing Burg's algorithm [6].

4.3 Prediction.

As discussed, accurate firing of the tank requires prediction of the muzzle error at some future time instant. The length of prediction step is dependent on, for example, the type of round, and the length of the gun, etc. This problem can be stated more formally as the optimal prediction of the ARMA process at step $k + n$ based on data up to step k.

Because of its many desirable features, the prediction method employed was the Kalman filter [7]. To utilize the Kalman filter, it was first necessary to convert the ARMA process to state space form. By defining the state equation

$$x(k+1) = A\,x(k) + B\,w(k) \tag{2}$$

where

$$A: = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & \\ -a_2 & 0 & 1 & & \\ \vdots & & & \ddots & \\ & & & & 1 \\ -a_{p-1} & 0 & \cdots & & 0 \\ -a_p & & & & \end{bmatrix}, \quad B = \begin{bmatrix} b_1 - a_1 b_0 \\ b_2 - a_2 b_0 \\ \vdots \end{bmatrix}$$

then the observation equation

$$z(k) = [1, 0, \ldots, 0]\; x(k) + b_0\, w(k) \tag{3}$$

describes the ARMA process. In the above, the estimated orders and coefficients are utilized. Note that from (3-2)-(3-3), the process noise is correlated with the observation noise. In order to avoid the increased complexity incurred in the correlated noise case, an equivalent augmented system was implemented which removed the "measurement" noise. In any case, the Kalman filter recursively computes the conditional expectation:

$$\hat{x}(k|k) = E[x(k)|z(0),\ldots,z(k)]$$

which is the minimum variance estimate (the estimate which produces the smallest

variance of the difference of the state and the estimate based on the observations). The ARMA estimate is immediate from (3-3)

$$\hat{z}(k\,|k) \;=\; [1,0,\ldots,0]\; x(k\,|k)$$

and because the state matrix A does not depend on time, it can be shown that

$$\hat{x}(k+n\,|k) \;=\; A^{n}\; \hat{x}(k\,|k) \tag{4}$$

The methodology was applied to the cases n = 3 (12 ms) and n = 5 (20 ms) . It is important to emphasize that while equation 4 is optimal, the quality of the estimate deteriorates as n grows large.

4.4 Adaptive Estimation

Implementation of the prediction algorithm, shown in Figure 3 consists of estimating the ARMA order and coefficients using data during a "training" interval followed by prediction for a short interval following the training. By training continuously, the algorithm provides an adaptive algorithm for prediction. The approach was considered not only for its simplicity, but also for its (comparatively) small computational burden. An estimate of the computational burden was made for a simplified (reduced order) version of the algorithm and it appears that it can operate in real-time on a DEC MicroVax II. However, to realistically measure the true computational requirements, the algorithm was extensively evaluated.

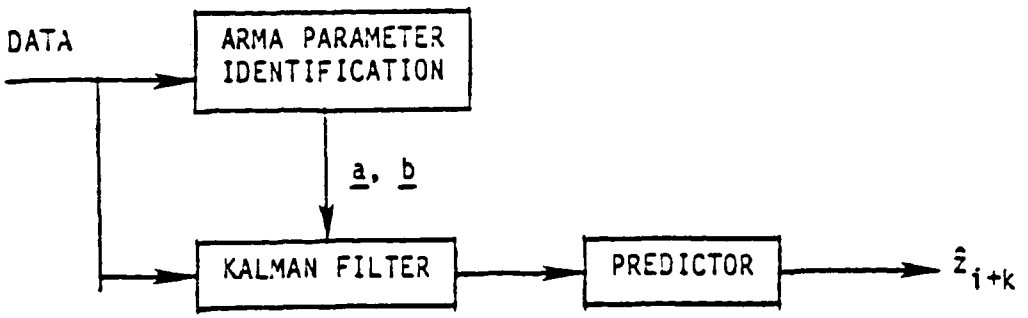


Figure 3. Adaptive Filtering Prediction Method

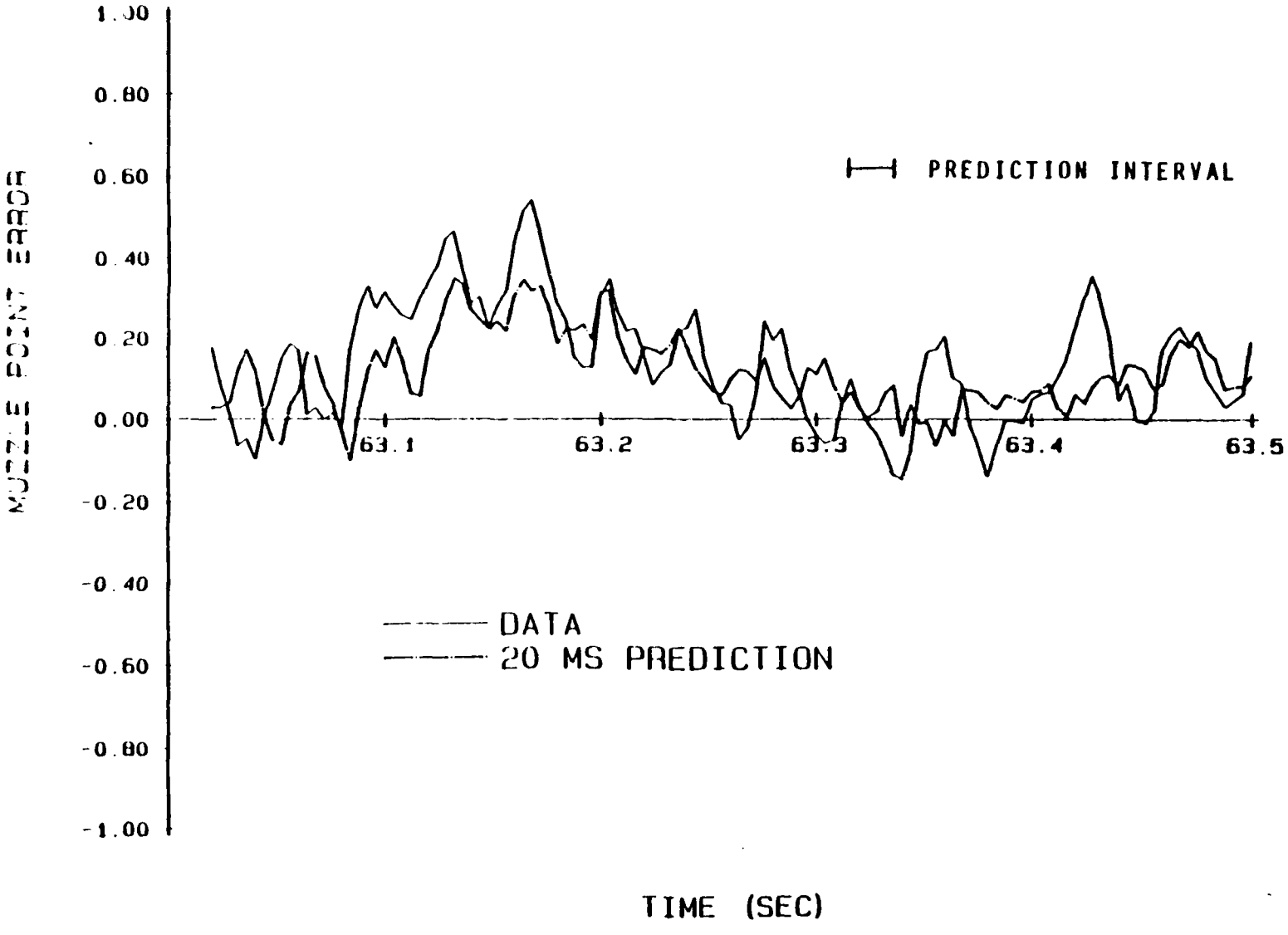


Figure 4. Muzzle Pointing Error and Prediction

5.0 ALGORITHM EVALUATION

The implementation of the algorithm required selection of certain "variables" such as the number of autoregression coefficients (N) to use in the approximation or the length of the training interval. Due to the requirement of computational efficiency, it was of interest to set variables providing acceptable performance while yielding the shortest possible run-time. Initially, values which resulted in good identification and prediction were chosen, then the values were altered in a systematic manner until a "minimal" set was obtained.

5.1 Experimental Baseline

To evaluate the performance as well as the limitations of the ARMA approach, the methodology was applied to a variety of representative data segments from the four tests. The segments were selected to provide typical (modeling and prediction over similar data), as well as atypical (modeling and prediction over different data) conditions. To be more precise, by choosing a variety of training and prediction interval combinations, the approach was tested under different physical "scenarios" associated with the tank traversing different portions of the track. Since the track is composed of regions consisting of primarily small bumps or large bumps, six different combinations were identified. For example, one combination resulting in a typical condition is training and prediction over data consisting primarily of large bumps. An atypical condition would result from training over large bumps and prediction over a segment consisting of small bumps. The ability of the algorithm to predict the data for a typical case is shown in figure 4.

To evaluate the quality of the muzzle pointing error, the sample standard deviation (RMS) of the error residual sequence e(k),

$$e(k) := z(k) - z(k|k-n)$$

was computed. The error RMS was computed both over the entire segment and data and only over the zero crossings: the points at which the prediction is within 1/10 milliradian band of zero. This latter statistic is important, as only at the predicted zero crossing will the gunner will be allowed to fire. For comparison, the RMS of the data over the entire segment was also computed.

Results for the baseline set of experiments show that the algorithm provides an average error reduction of 32% for the typical and 23% for the atypical segments. The most dramatic reductions often occur at the higher speeds.

5.2 Reduction of Algorithm Run Time

Since the algorithm achieved suitable performance on the baseline segments, values of the algorithm were next individually varied to result in shorter run times. Specifically, the effect of reducing the training interval, the order of the autoregressive approximation (N), and the estimated AR and MA orders was measured. In order to

quantify the effect of the various changes, the run time of each subroutine of the algorithm was calculated with a timing program. Execution time is most sensitive to ARMA order as it is directly related to the Kalman filter computation, often the most numerically expensive portion of the algorithm. Examples of results utilizing reduced values are given in Table 1. A typical run which used two seconds of data to establish the model followed by prediction over one second of data required approximately 4 to 5.5 seconds on a DEC MicroVax II.

Table 1. Results of Variable Testing (2 sec training)

SPEED MPG	CASE*	RAW DATA RMS	BEST MODEL PREDICTION ERROR	REDUCED ORDER PREDICTION ERROR
5	A	0.3mr	0.2mr (15)	0.2m4 (3)
	B	0.3	0.2 (1)	0.2 (1)
	C	0.2	0.2 (7)	0.2 (1)
15	A	0.3	0.3 (1)	0.3 (1)
	B	0.7	0.3 (4)	0.3 (1)
	C	0.5	0.4 (1)	0.4 (1)
30	A	0.6	0.4 (17)	0.4 (8)
	B	1.0	0.3 (9)	0.3 (4)
	C	0.9	0.4 (3)	0.4 (3)

*Cases:

A: Train on Small Bumps, Predict on Small Bumps
B: Train on Large Bumps, Predict on Large Bumps
C: Train on Small Bumps, Predict on Large Bumps

6.0 DISCUSSION AND FUTURE WORK

Considering the limited scope of this study (restriction of ARMA models and limited data types), the results are quite encouraging. For the baseline, the adaptive 20 millisecond ARMA predictor was able to reduce the total muzzle pointing error usually from 20% to 60% for the expected operational conditions (typical scenarios) and even for the atypical cases there was often a small to moderately large reduction. Usually, errors are between 0.1 and 0.4 milliradians. By reducing the training interval and forcibly decreasing the order of the model, a version of the algorithm with comparable performance was obtained which appears capable of real-time operation on commercially

available microprocessors. It is almost certain that VHSIC technology will make real-time operation feasible.

Since the ARMA work was completed an alternative technique, Canonical Variate Analysis (CVA), was applied to the strain gage test data. The CVA approach is a method for identifying the observable dynamics, or states, from empirical data. The algorithm is automatic and completely "data driven" no apriori modeling is required. A brief overview of the method is provided here.

The CVA algorithm provides directly from the data a state-space representation of the underling system generating the data. That is, CVA estimates all relevant quantities of a system of the form:

$$x(k+1) = T\,x(k) + G\,u(k) + w(k) \tag{5}$$

$$y(k) = H\,x(k) + A\,u(k) + B\,w(k) + v(k)$$

Where:

y(k) = is the output,
x(k) = is the state of the system
T = is the system transition matrix,
u(k) = is an input vector, and

w(k) and v(k) are independent white noise processes with covariance matrices Q and R, and G, H, A, & B are dynamic matrices. The salient difference between the ARMA modeling and CVA is that both plant, w(k), and measurement ,v(k), noise sources are estimated.

Figure 5 depicts the CVA modeling and prediction process. The model obtained through CVA techniques is inherently adaptive and robust in that as changes occur in the driving forces, a new model, very possibly of different order can be determined. Under the assumption the modeling procedure can be performed quickly and frequently, a very accurate representation of the observable dynamics is consistently available. As previously shown in section 1, an optimal (minimum mean square error) prediction of the state is immediate from the Kalman Filter.

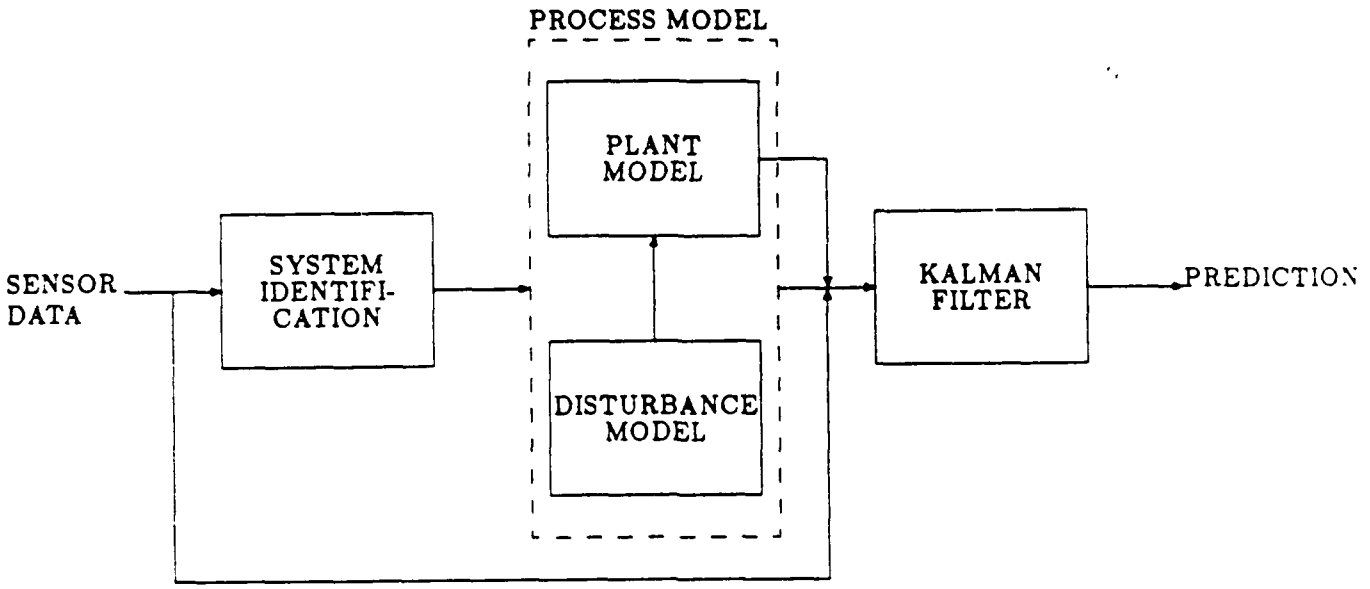


Figure 5. Dynamic Modeling and Prediction Using CVA

Figure 6 depicts a comparison of CVA prediction versus that of the optimum ARMA model. Also shown is the actual muzzle motion. The superior prediction capability of the CVA is evident.

Although the ARMA and CVA approaches show real promise, further work is required to refine the techniques and to investigate alternate forms of the adaptive filter. In particular, there is potential for great improvement if the ARMA or CVA modeling is augmented with physical modeling of the gun tube/turret and if additional data such as accelerometer/gyro or gunner servo error is

7.0 REFERENCES

[1] H. Cohen, "Methodology for Stochastic Modeling," U.S. Army Systems Analysis Activity, Technical Report No. 410, January 1985.

[2] J.A. Cadzow, "Spectral Estimation: an Overdetermined Rational Model Equation Approach," Proc. IEEE, Vol. 70, No. 9, Sept. 1982.

[3] G.E. Box and G.M. Jenkins, Time Series Analysis Forecasting and Control, San Francisco, California, Holden-Day 1970.

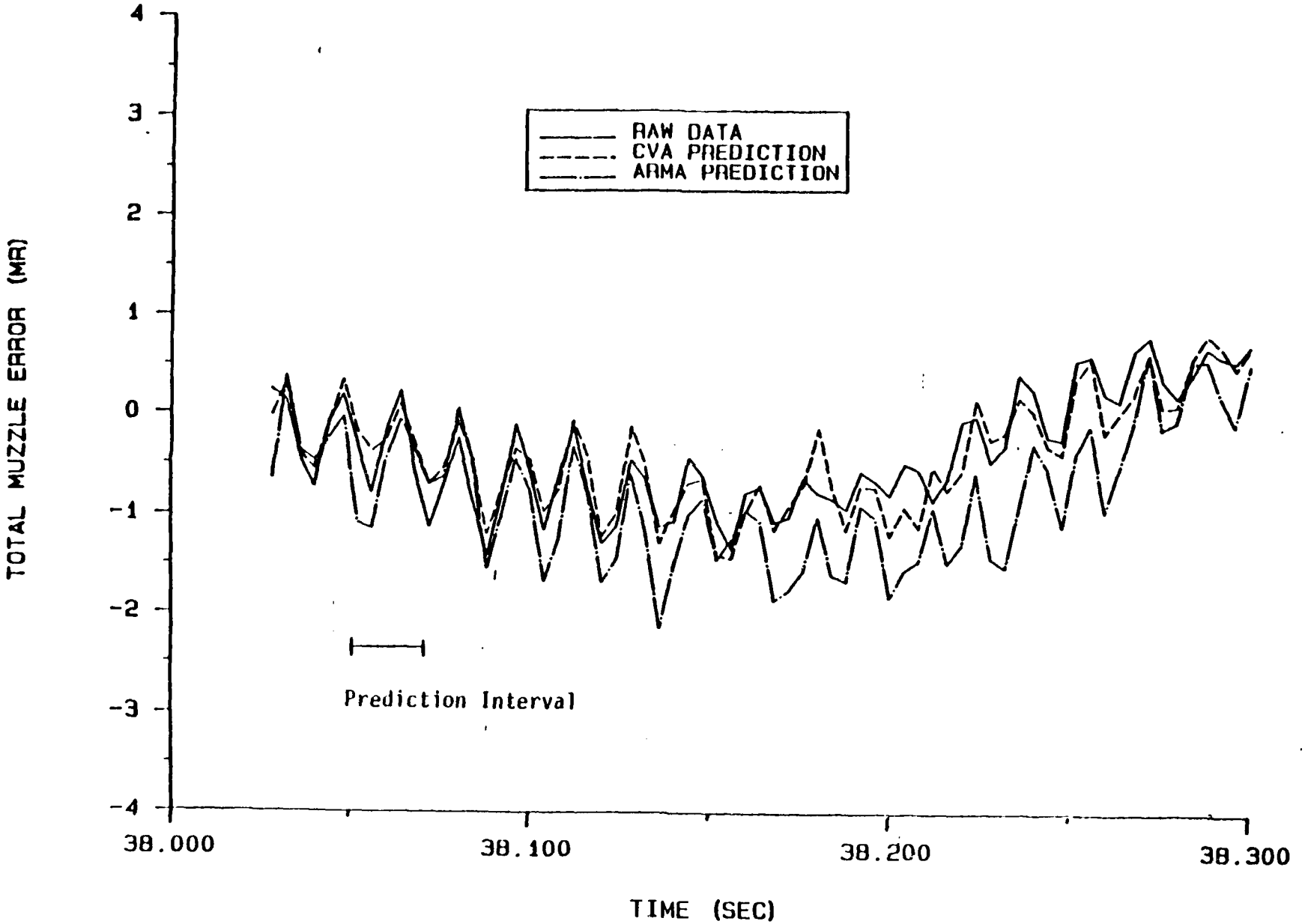


Figure 6. Comparison of CVA and ARMA for 20 Millisecond Prediction of Muzzle Pointing Error

[4] J. Makhoul, "Linear Prediction: A Tutorial Review" Proceedings of the IEEE, Vol. 63, No. 4, pp.561-580, April 1975.

[5] D. Graupe, D.J. Krause, and J.B. Moore, "Identification of Auto regressive Moving Average Parameters of Time Series," IEEE Trans. Automatic Control, AC-20, pp. 104-107, Feb. 1975.

[6] T.J. Ulrych and T.N. Bishop, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," Rev. Geophysics and Space phys., Vol. 13, pp. 183-200, Feb. 1975.

[7] B.D.O. Anderson and J.B. Moore, Optimal Filtering. Prentice Hall, New Jersey, 1979.

A NON-RECTANGULAR SAMPLING PLAN FOR ESTIMATING STEADY-STATE MEANS

Peter W. Glynn

Department of Operations Research
Stanford University
Stanford, CA 94305

Abstract

The method of multiple replicates is frequently used by simulators to estimate the steady-state mean of a stochastic simulation. One important advantage of this approach is that it is easily adapted to a parallel computer. Unfortunately, the method of multiple replicates is quite sensitive to contamination by "initial bias." In this paper, a new type of sampling plan is described. It retains the replication flavor, yet attenuates the bias problem. It is shown that the new method reduces mean square error relative to conventional multiple replicates for problems in which the "initial transient" decays slowly.

Keywords: Simulation, replication, mean square error, parallel computation.

Introduction

Let $Y = (Y(n) : n \geq 0)$ be a real-valued stochastic sequence corresponding to the output of a stochastic simulation. We assume that Y is ergodic, in the sense that there exists a finite (deterministic) constant r such that

$$\frac{1}{n}\sum_{i=0}^{n-1} Y(i) \Rightarrow r$$

as $n \to \infty$. The steady-state simulation problem concerns the question of estimating the parameter r efficiently, and providing confidence intervals for r.

Basically, two alternative approaches for dealing with this problem have been studied in the literature. One approach is known as the method of multiple replicates. The idea here is to generate m independent replicates of the process Y. Each replicate is simulated for t time units. The advantage of this method is that it gives rise to independent observations; this significantly simplifies the problem of producing confidence intervals for r. Furthermore, given access to a parallel computing environment, one can assign each independent replicate to a different processor. Thus, the method of multiple replicates is well suited to parallel computation.

A disadvantage of this approach is that each of the m independent replicates is contaminated by initial bias. This initial bias arises from the fact that each of the m replicates is initiated with an initial condition that is atypical of the steady-state of the system. If we view the first s time units of each replicate as representing an "initial transient" for the system, this analysis suggests that ms time units of the total time simulated are contaminated by initial bias. If m is large, we find that the method of multiple replicates devotes a significant amount of computation to generation of highly biased observations. This is, of course, undesirable.

In response to this, we can consider sampling plans in which only one observation of Y is generated. Such a strategy is known in the literature as a single replication method. Here, only the first s time units of the simulation are significantly biased, and there is no magnification effect by the parameter m. On the other hand, construction of confidence intervals for r is now complicated by the fact that all the observations collected are autocorrelated. Furthermore, it is now a non-trivial task to make an assignment of parallel processors that will significantly speed up the simulation.

Note that the method of multiple replicates involves factoring a computer time budget T into m replicates, each of length $t = T/m$. If we view the data of the i'th replicate as being assigned to the i'th row of a matrix, we obtain a rectangular $m \times t$ matrix which summarizes the data generated by the simulation. Consequently, we refer to the method of multiple

replicates as a rectangular sampling plan for estimating steady-state means (see Figure 1). Of course, a single replicate method is the special case of a rectangular scheme in which the data corresponds to a $1 \times T$ row vector.

In this paper, we consider these rectangular methods in greater detail. We also propose and analyze a new non-rectangular sampling scheme, which attempts to offer an advantageous compromise between the methods of single and multiple replicates.

The organization of this paper is as follows. Section 2 provides reasonably complete mean square error analysis of conventional rectangular sampling plans. In Section 3, the non-rectangular plan is introduced and studied. Section 4 offers some conclusions.

2. Rectangular Sampling Plans

We start by describing the traditional method of replication for solving the steady-state simulation problem. To simplify the discussion that follows, we will assume that in x units of computer time, precisely x time units of the process Y can be simulated. Thus, given a total computer time budget of size T, we can implement a rectangular sampling plan in the following way:

1.) Choose the number m of independent replicates. (If $m = 1$, this is a single replication method.)

2.) Choose the (deletion) parameter s, from the interval $[0, T/m]$. (The first s time units of each replication will be deleted from the set of observations.)

3.) Generate m independent copies $Y_1, Y_2, \ldots, Y_m$ of the process Y. Each copy is simulated over the interval $[0, T/m]$.

4.) Set $t = \lfloor T/m \rfloor$ and compute the estimator

$$\hat{Y}(m, s, T) = \frac{1}{m(t-s)} \sum_{i=1}^{m} \sum_{j=s+1}^{t} Y_i(j).$$

We will now consider the mean square error (MSE) of the estimator $\hat{Y}(m, s, T)$. The MSE criterion is often viewed as the most important quantitative measure of the quality of an estimator. We start with the well known MSE decomposition formula

$$\text{MSE}(\hat{Y}(m, s, T)) = \text{var}\, \hat{Y}(m, s, T) + (\text{bias}\, \hat{Y}(m, s, T))^2. \tag{2.1}$$

By using the independence of the replicates, we observe that

$$\text{var}\, \hat{Y}(m, s, T) = \frac{1}{m} \text{var}\, \frac{1}{t-s} \sum_{j=s+1}^{t} Y(j), \tag{2.2}$$

$$\text{bias}\, \hat{Y}(m, s, T) = \frac{1}{t-s} \sum_{j=s+1}^{t} EY(j) - r. \tag{2.3}$$

A Rectangular Sampling Plan

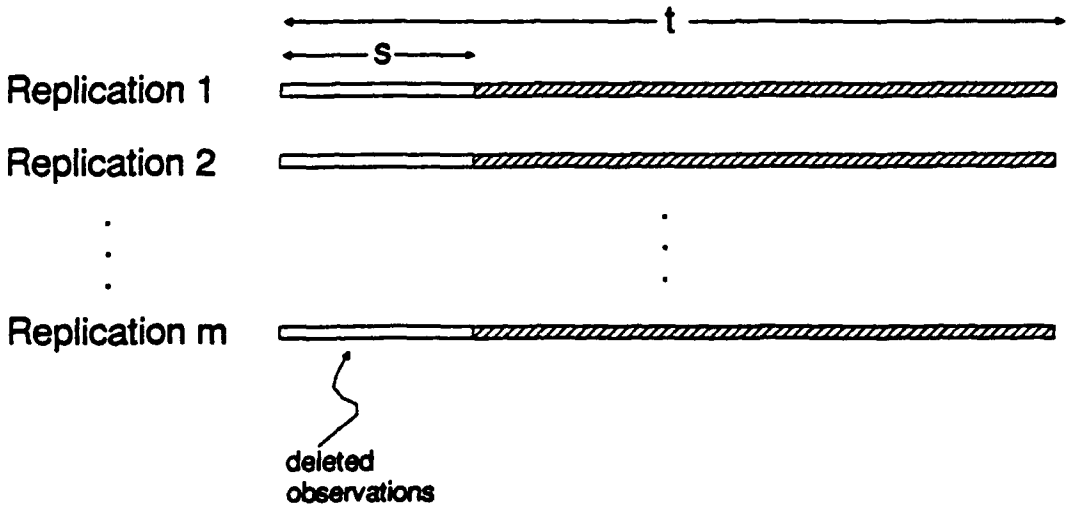


Figure 1

The Non-rectangular Sampling Plan

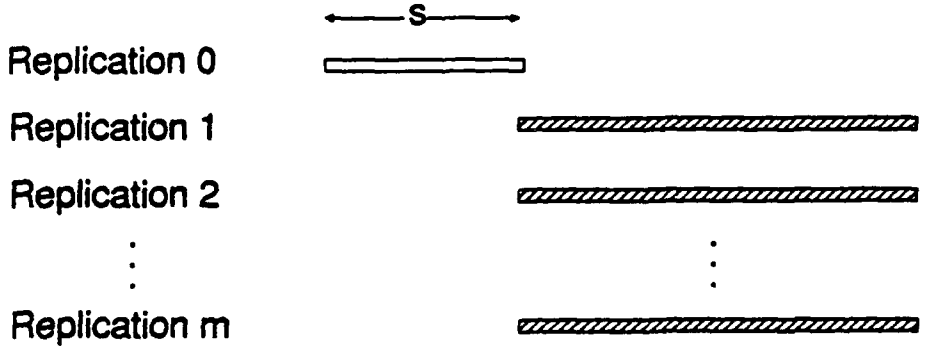


Figure 2

In order to analyze the terms appearing on the right-hand sides of (2.2) and (2.3), we will assume that $Y(n)$ can be expressed as a real-valued functional of a time-homogeneous Markov chain $X(n)$, so that $Y(n) = f(X(n))$ for some real-valued f defined on the state space S of X. The set S may be discrete or continuous. Continuous state space is particularly convenient in analysis of discrete-event simulations. The generalized semi-Markov process (GSMP) view of discrete-event systems shows that very general discrete-event simulations may be expressed in the form $Y(n) = f(X(n))$ with X Markov, provided that we permit continuous state space.

For $x \epsilon S, u \geq 1$, let $v(x,u)$ be the conditional variance defined by

$$v(x,u) = E\left\{\left(\frac{1}{u}\sum_{j=0}^{u-1} Y(j)\right)^2 | X(0) = x\right\} - \left(E\left\{\frac{1}{u}\sum_{j=0}^{u-1} Y(j) | X(0) = x\right\}\right)^2.$$

Similarly, let $b(x,u)$ be the conditional bias given by

$$b(x,u) = E\left\{\frac{1}{u}\sum_{j=0}^{u-1} Y(j) | X(0) = x\right\} - r.$$

Let $\mu(\cdot) = P\{X(0) \epsilon \cdot\}$ be the initial distribution of X. The Markov property permits us to re-express (2.3) as

$$\text{bias } \bar{Y}(m,s,T) = E_\mu b(X(s+1), t-s), \tag{2.4}$$

where $E_\mu(\cdot)$ denotes the expectation operator conditional on $X(0)$ having distribution μ.

To obtain a similar expression for the variance term (2.2) requires more care. We first apply the well known variance decomposition formula

$$\begin{aligned} \text{var}\,\frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) = E\,\text{var}\left\{\frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) | X(s+1)\right\} \\ + \text{var}\,E\left\{\frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) | X(s+1)\right\}. \end{aligned} \tag{2.5}$$

Clearly, we have

$$\text{var}\left\{\frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) | X(s+1)\right\} = v(X(s+1), t-s),$$

$$E\left\{\frac{1}{t-s}\sum_{j=s+1}^{t} Y(y) | X(s+1)\right\} = b(X(s+1), t-s).$$

Plugging these expressions into (2.5) yields

$$\text{var}\,\frac{1}{t-s}\sum_{j=s+1}^{t} Y(y) = E_\mu v(X(s+1), t-s) + \text{var}_\mu b(X(s+1), t-s), \tag{2.6}$$

where $\mathrm{var}_\mu(\cdot)$ denotes the variance operator conditional on $X(0)$ having distribution μ.

Suppose that X is a positive recurrent Markov chain possessing a unique invariant probability distribution π. A large class of such chains has the property that under suitable regularity conditions,

$$\sup_{h \in \mathcal{H}} |E_\mu h(X(s)) - E_\pi h(X(0))| = O(e^{-\alpha s})$$

for some $\alpha > 0$, where $\mathcal{H}$ is some appropriately defined family of real-valued functions $h : S \to \mathbb{R}$. (See NUMMELIN (1984), p. 120, for an example of such a theorem.) Assuming that the functions $v(\cdot, u)$, $b(\cdot, u)$, $b^2(\cdot, u) \in \mathcal{H}$ for all $u \geq 1$, we obtain the relations

$$E_\mu v(X(s+1), t-s) = E_\pi V(X(0), t-s) + O(e^{-\alpha s}), \tag{2.7}$$

$$E_\mu b(X(s+1), t-s) = E_\pi b(X(0), t-s) + O(e^{-\alpha s}), \tag{2.8}$$

$$E_\mu b^2(X(s+1), t-s) = E_\pi b^2(X(0), t-s) + O(e^{-\alpha s}), \tag{2.9}$$

where the constants implicit in each of the "big Oh" terms are independent of t.

Furthermore, for such a recurrent Markov chain, it is typically the case that the steady-state mean r can be expressed in the form $r = E_\pi f(X(0))$. As a consequence of the stationarity of X under initial distribution π, it is evident that $E_\pi Y(n) = r$ for $n \geq 0$ and hence $E_\pi b(X(0), t-s) = 0$. Thus, (2.8) can be simplified to

$$E_\mu b(X(s+1), t-s) = O(e^{-\alpha s}). \tag{2.10}$$

Combining (2.9) and (2.10), we obtain

$$\mathrm{var}_\mu b(X(s+1), t-s) = E_\pi b^2(X(0), t-s) + O(e^{-\alpha s}). \tag{2.11}$$

(Again, the constants implicit in (2.10) and (2.11) are independent of t.)

Combining (2.6), (2.7), and (2.11), we obtain the expression

$$\mathrm{var} \frac{1}{t-s} \sum_{j=s+1}^{t} Y(j) = E_\pi v(X(0), t-s) + E_\pi b^2(X(0), t-s) + O(e^{-\alpha s}).$$

Repeating the variance decomposition (2.6) under $\mathrm{var}_\pi(\cdot)$, we find that

$$\mathrm{var}_\pi \frac{1}{t-s} \sum_{j=s+1}^{t} Y(j) = E_\pi v(X(0), t-s) + E_\pi b^2(X(0), t-s)$$

and hence

$$\mathrm{var} \frac{1}{t-s} \sum_{j=s+1}^{t} Y(j) = \mathrm{var}_\pi \frac{1}{t-s} \sum_{j=s+1}^{t} Y(j) + O(e^{-\alpha s}). \tag{2.12}$$

To simplify (2.12), we again use the fact that X is stationary under initial distribution π. Set $Y_c(n) = Y(n) - r$,

$$\sigma^2 = E_\pi Y_c(0)^2 + 2\sum_{k=1}^{\infty} E_\pi Y_c(0)Y_c(k)$$

$$\eta = 2\sum_{k=1}^{\infty} kE_\pi Y_c(0)Y_c(k).$$

Under appropriate summability hypotheses (see, for example, p. 172 of BILLINGSLEY (1968)), we can use the stationarity to write

$$\text{(2.13)} \qquad \text{var}_\pi \frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) = \frac{\sigma^2}{t-s} - \frac{\eta}{(t-s)^2} - \frac{2}{t-s}\sum_{k=t=s}^{\infty}\left(1 - \frac{k}{t-s}\right) E_\pi Y_c(0)Y_c(k).$$

Note that

$$\text{(2.14)} \qquad |E_\pi Y_c(0)Y_c(k)| \le \int_S |f(x)| \cdot |E_x Y_c(k)| \cdot \pi(dx),$$

where $E_x(\cdot)$ is the expectation operator conditional on $X(0) = x$. We now observe that $E_x Y(k) = E_x f(X(k)) - E_\pi f(X(0))$. Appropriate regularity hypotheses on X permit us to assert that

$$\text{(2.15)} \qquad \sup_{x \in S} |E_x f(X(k)) - E_\pi f(X(0))| = O(e^{-\beta k})$$

for some $\beta > 0$. (See p. 122 of NUMMELIN (1984) for a typical such result.) Substituting this relation in (2.14) yields

$$E_\pi Y_c(0)Y_c(k) = O(e^{-\beta k}).$$

We may therefore conclude that

$$\text{(2.16)} \qquad \sum_{k=u}^{\infty}\left(1 - \frac{k}{u}\right) E_\pi Y_c(0)Y_c(k) = O(e^{-\beta' u})$$

for $0 < \beta' < \beta$. Substitution of (2.16) into (2.13) shows that

$$\text{(2.17)} \qquad \text{var}_\pi \frac{1}{t-s}\sum_{j=s+1}^{t} Y(j) = \frac{\sigma^2}{t-s} - \frac{\eta}{(t-s)^2} + O(e^{-\beta'(t-s)}).$$

Combining (2.1), (2.2), (2.4), (2.10), (2.12), and (2.17), we obtain the important relationship

$$\text{(2.18)} \qquad \text{MSE}(\bar{Y}(m,s,T)) = \frac{1}{m}\left(\frac{\sigma^2}{t-s} - \frac{\eta}{(t-s)^2}\right) + O(e^{-\alpha s}) + \frac{1}{m}O(e^{-\beta'(t-s)}),$$

where the implicit constants appearing in the "big OH" terms are independent of m, s, and T.

To gain further insight into (2.18), we consider the typical situation, in which the deletion point s is small relative to the length t of each replicate. Furthermore, in order to simplify the discussion, we assume that $mt = T$ (exactly). Then,

$$\frac{1}{m}\frac{\sigma^2}{t-s} = \frac{\sigma^2}{T} + \frac{\sigma^2 ms}{T^2} + \frac{1}{T}O\left(\frac{s^2m^2}{T^2}\right), \quad \text{and} \tag{2.19}$$

$$\frac{1}{m}\frac{\eta}{(t-s)^2} = \frac{m\eta}{T^2} + \frac{m}{T^2}O\left(\frac{ms}{T}\right). \tag{2.20}$$

Combining (2.18) through (2.20), we obtain the approximation

$$\mathrm{MSE}(\bar{Y}(m,s,T)) \approx \frac{\sigma^2}{T} + \frac{\sigma^2 ms}{T^2} - \frac{m\eta}{T^2}. \tag{2.21}$$

Viewing s and m as design parameters for the simulation, we see that (2.21) suggests that the deletion parameter s should be small. On the other hand, if s is chosen too small, difficulties can arise in the "big Oh" terms appearing in (2.18). This recommendation corresponds to intuition.

As for the number of replications m, m should be chosen small (for example, a single replicate method should be considered) whenever $\sigma^2 s > \eta$. For reasonable values of s, this inequality will typically be valid. Thus, mean square error favors using a small number of replicates. This differs from the conclusion reached by KELTON (1986) in his analysis of "replication splitting" schemes for simulation of autoregressive sequences. The arguments there show that using a large number of replicates can reduce the variance of the steady-state estimator when the autoregressive sequence is positively correlated (i.e. $\eta > 0$). In our current setting, we judge our estimators via mean square error (as opposed to variance). Since our error criterion explicitly considers the loss in estimator efficiency due to bias (variance does not measure bias), it is not surprising that our conclusions differ. Of course, if s is small (i.e. bias is not a major problem), (2.21) supports using a large number of replicates when $\eta > 0$,

To illustrate the above points, we calculate the mean square error of $\bar{Y}(m,s,T)$ when $m = T^p (0 \le p < 1)$ and $s = T^q (0 < q < 1-p)$, in which case $t = T^r$, where $r = 1-p$. We find that

$$\mathrm{MSE}(\bar{Y}(m,s,T)) = \frac{\sigma^2}{T} + \frac{\sigma^2}{T^{2-p-q}} - \frac{\eta}{T^{2-p}} + O(T^{2p+2q-3}). \tag{2.22}$$

Assuming that $p + q < 1/2$, (so that the "big oh" term is small) we find that relation (2.17) confirms the previous discussion. Both p and q should be chosen small, in accordance with our previous recommendations.

3. A Non-Rectangular Sampling Plan

The idea behind the sampling plan to be described in this section is that we try to avoid expending a significant fraction of the computer time budget on generation of

highly biased observations. As discussed in the Introduction, the initial bias problem is of particular concern when the method of multiple replicates is used, since the amount of contaminated data is proportional to the number of replicates. On the other hand, the method of multiple replicates enjoys several significant advantages: ease of construction of confidence intervals and development of parallel simulation schemes. Our goal here is to develop a method that has a multiple replicate flavor and yet avoids the initial bias difficulties that are associated with conventional multiple replicate methods.

As in Section 2, we assume that the output sequence Y takes the form $Y(n) = f(X(n))$ for some time-homogeneous Markov chain X, and real-valued function f. The following algorithm employs one simulation of length s to generate an initial condition which is reasonably typical of the steady-state. This initial condition is then used to generate m conditionally independent replicates (each of length t) from the output sequence Y. Thus, the effort to generate a "good" initial condition is amortized over the m replicates. In terms of observations generated, this sampling plan is non-rectangular (see Figure 2).

The non-rectangular sampling plan can be summarized as follows.

1.) Given the computer time budget T, choose the number m of (conditionally independent) replicates, and the deletion parameter s $(0 \le s \le T)$.

2.) Generate one copy Y_0 of the sequence Y to time s.

3.) Using the initial condition $X_0(s)$ (X_0 is the Markov chain corresponding to Y_0), generate m copies $Y_1, \ldots, Y_m$ of Y to time $t-1$, where $t = \lceil (T-s)/m \rceil$.

4.) Compute the estimator

$$\tilde{Y}(m,s,T) = \frac{1}{mt} \sum_{i=1}^{m} \sum_{j=0}^{t-1} Y_i(j)$$

We now turn to computing the mean square error of $\tilde{Y}(m,s,T)$. As in Section 2,

$$\text{MSE}(\tilde{Y}(m,s,T)) = \text{var}\, \tilde{Y}(m,s,T) + (\text{bias}\, \tilde{Y}(m,s,T))^2. \tag{3.1}$$

Using the fact that $Y_i(\cdot) \stackrel{D}{=} Y(\cdot + s)$ ($\stackrel{D}{=}$ denotes equality in distribution), we find that

$$\text{bias}\, \tilde{Y}(m,s,T) = E_\mu b(X(s),t).$$

From (2.8), it therefore follows that

$$\text{bias}\, \tilde{Y}(m,s,T) = O(e^{-\alpha s}). \tag{3.2}$$

To handle the variance term appearing on the right-hand side of (3.1), we again use the variance decomposition method:

$$\text{var}\, \tilde{Y}(m,s,T) = \text{var}\, E\{\tilde{Y}(m,s,T)|X_0(s)\} + E\, \text{var}\{\tilde{Y}(m,s,T)|X_0(s)\}. \tag{3.3}$$

It is easily seen (use the fact that $Y_1, \ldots, Y_m$ are independent and identically distributed, conditional on $X_0(s)$) that

$$E\{\tilde{Y}(m,s,T)|X_0(s)\} = b(X_0(s),t) \quad \text{a.s.}, \tag{3.4}$$

$$\text{var}\{\tilde{Y}(m,s,T)|X_0(s)\} = \frac{1}{m}v(X_0(s),t) \quad \text{a.s.} \tag{3.5}$$

Combining (3.3) through (3.5), we get

$$\text{var}\,\tilde{Y}(m,s,T) = \frac{1}{m}E_\mu v(X(s),t) + \text{var}_\mu b(X(s),t). \tag{3.6}$$

As in Section 2, we obtain

$$\text{var}\,\tilde{Y}(m,s,T) = \frac{1}{m}E_\pi v(X(0),t) + E_\pi b^2(X(0),t) + O(e^{-\alpha s}) \tag{3.7}$$

(use (2.7), (2.8), and (2.9)). Recall that

$$\text{var}_\pi \frac{1}{t}\sum_{j=0}^{t-1} Y(j) = E_\pi v(X(0),t) + E_\pi b^2(X(0),t).$$

(see Section 2). Plugging into (3.7), we get

$$\text{var}\,\tilde{Y}(m,s,T) = \frac{1}{m}\text{var}_\pi \frac{1}{t}\sum_{j=0}^{t-1} Y(j) + \left(\frac{m-1}{m}\right) E_\pi b^2(X(0),t). \tag{3.8}$$

The first term on the right-hand side of (3.8) was analyzed in (2.17). For the second term, note that

$$b(x,t) = \frac{1}{t}b(x) - \frac{1}{t}\sum_{k=t}^{\infty}(E_x Y(k) - r),$$

where

$$b(x) = \sum_{k=0}^{\infty}(E_x Y(k) - r).$$

From (2.15), it is evident that

$$\sup_{x\in S}|b(x,t) - \frac{1}{t}b(x)| = O(e^{-\beta t}). \tag{3.9}$$

Consequently, we obtain the inequality

$$b(X(0),t) \le \frac{1}{t}b(X(0)) + O(e^{-\beta t}). \tag{3.10}$$

Since $E_\pi Y(k) = r$, the expectations $E_\pi b(X(0),t)$ and $E_\pi b(X(0))$ both vanish. From (3.10), we therefore get

$$E_\pi b^2(X(0),t) \le \frac{1}{t^2}E_\pi b^2(X(0)) + O(e^{-\beta t})E_\pi|b(X(0))| + O(e^{-2\beta t}).$$

A similarly derived lower bound yields the formula

$$E_\pi b^2(X(0),t) = \frac{1}{t^2} E_\pi b^2(X(0)) + O(e^{-\beta t}). \tag{3.11}$$

Let $b = E_\pi b^2(X(0))$. To simplify the following discussion, assume $t = (T-s)/m$ (exactly). Combining (2.13), (3.8), and (3.11), we obtain the important relationship

$$\mathrm{MSE}(\tilde{Y}(m,s,T)) = \frac{1}{m}\left(\frac{\sigma^2}{t} - \frac{\eta}{t^2}\right) + \left(\frac{m-1}{m}\right)\frac{b}{t^2} + O(e^{-\gamma t}) + O(e^{-\alpha s}), \tag{3.12}$$

where $\gamma = \min(\beta, \beta')$ and the (implicit) constants in the "big oh" terms are independent of m, s, and T. Expressing t in terms of m, s, and T, we get

$$\frac{\sigma^2}{mt} = \frac{\sigma^2}{T} + \frac{\sigma^2 s}{T^2} + \frac{1}{T} O\left(\frac{s^2}{T^2}\right), \tag{3.13}$$

$$\frac{\eta}{mt^2} = \frac{m\eta}{T^2} + \frac{m}{T^2} O\left(\frac{s}{T}\right), \quad \text{and} \tag{3.14}$$

$$\left(\frac{m-1}{m}\right)\frac{b}{t^2} = \frac{(m-1)mb}{T^2} + \frac{m(m-1)}{T^2} O\left(\frac{s}{T}\right), \tag{3.15}$$

assuming that s is small relative to T. Combining (3.12) through (3.15), we obtain the approximation

$$\mathrm{MSE}(\tilde{Y}(m,s,T)) \approx \frac{\sigma^2}{T} + \frac{\sigma^2 s}{T^2} - \frac{m\eta}{T^2} + \frac{m(n-1)b}{T^2}. \tag{3.16}$$

We now compare the mean square error of our non-rectangular sampling plan with that of a rectangular plan having the same computer time budget T, number of replications m, and deletion parameter s. Comparing (3.16) to (2.21), we see that $\mathrm{MSE}(\tilde{Y}(m,s,T)) \leq \mathrm{MSE}(\bar{Y}(m,s,T))$ when

$$\sigma^2 ms \geq \sigma^2 s + b(m^2 - m).$$

We shall shortly show that $b \geq \sigma^2$. Thus, $\tilde{Y}(m,s,T)$ beats $\bar{Y}(m,s,T)$ when $sm \geq s + m^2$. This will typically occur when s is large relative to m. Thus, we can expect $\tilde{Y}(m,s,T)$ to have smaller MSE than $\bar{Y}(m,s,T)$ whenever s must be chosen relatively large, in order to remove initial bias.

We can illustrate this point when $m = T^p$ $(0 \leq p < 1)$ and $s = T^q$ $(0 < q < 1)$. Then, if $p + q < 1/2$,

$$\mathrm{MSE}(\tilde{Y}(m,s,T)) = \frac{\sigma^2}{T} + \frac{\sigma^2}{T^{2-q}} - \frac{\eta}{T^{2-p}} + \frac{b}{T^{2-2p}} + O\left(\frac{1}{T^2}\right). \tag{3.17}$$

Comparing (3.17) to (2.22), we find that $\mathrm{MSE}(\tilde{Y}(m,s,T)) \leq \mathrm{MSE}(\bar{Y}(m,s,T))$ when $p < q$, as was suggested above.

We conclude this section by showing $b \geq \sigma^2$. We first observe that $b(x)$ solves Poisson's equation

$$b(x) - E_x b(X(1)) = f_c(x),$$

where $f_c(x) = f(x) - r$. Additionally, $E_\pi b(X(0)) = 0$. Then,

$$\sum_{k=0}^{n-1} f_c(X(k)) = \sum_{k=1}^{n+1} D_k + b(X(0)) - b(X(n+1))$$

where $D_k = b(X(k)) - E\{b(X(k))|X(k-1)\}$ are martingale differences. Note that if $X(0) \stackrel{D}{=} \pi$, we can apply the martingale central limit theorem (see p. 205 of BILLINGSLEY (1968)) to conclude that

$$n^{-1/2} \sum_{k=0}^{n-1} f_c(X(k)) \Rightarrow \lambda N(0,1) \tag{3.18}$$

where $\lambda^2 = E_\pi D_1^2$. (The function $b(\cdot)$ is bounded under (2.15).) If the left-hand side of (3.18) is appropriately uniformly integrable, then

$$n^{-1} \mathrm{var}_\pi \sum_{k=0}^{n-1} f_c(X(k)) \to \lambda^2 \tag{3.19}$$

as $n \to \infty$. But

$$\mathrm{var}_\pi \sum_{k=0}^{n-1} f_c(X(k)) = n^2 \cdot \mathrm{var}_\pi \frac{1}{n} \sum_{j=0}^{n-1} Y(j).$$

From (2.17) and (3.19), it follows that $\lambda^2 = E_\pi D_1^2 = \sigma^2$. But D_1 is orthogonal to $b(X(0))$, being a martingale difference, and hence

$$E_\pi b(X(1))^2 = E_\pi D_1^2 + E_\pi(E\{b(X(1))|X(0)\}^2).$$

Since $b = E_\pi b(X(0))^2$, it is evident that $b \geq \sigma^2$.

4. Conclusions

The non-rectangular sampling plan introduced in this paper has a lower mean square error than that of the corresponding rectangular plan that involves an equivalent amount of computer time, when the "initial transient" decays slowly. This, of course, is precisely the setting in which the method of multiple replicates exhibits its poorest behavior (relative to a single replicate method). Thus, the non-rectangular plan described here is most beneficial in precisely those problems for which multiple replicates is typically most ineffective.

It should be clear that the replication component of this non-rectangular plan is well-suited to parallel computation. However, the generation of the initial condition $X_0(s)$ is

not easily adapted to the parallel setting. This aspect of the sampling plan described here deserves further attention.

Finally, it should be mentioned that a great deal of empirical work remains to be done in understanding the advantages and limitations of this non-rectangular method, when applied to "real world" problems.

References

1. BILLINGSLEY, P. (1968). *Convergence of Probability Measures*, John Wiley, New York.
2. KELTON, W. D. (1986). Replication splitting and variance for simulating discrete-parameter stochastic processes. *Operations Research Letters* 4, 275–279.
3. NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-negative Operators.* Cambridge University Press, Cambridge, UK.

Covariance Analysis for Split Plot and Split Block Designs and Computer Packages

Walter T. Federer and Michael P. Meredith

Mathematical Sciences Institute

Cornell University, Ithaca, N.Y. 14853

ABSTRACT. Covariance analysis for data from experiments designed in a split plot or split block design is mostly ignored in statistical literature. When it is considered, it is often done incorrectly and/or incompletely. This is especially true for computer packages. A discussion of what should be done, what is or can be done with computer packages, and a possible solution to the problems is given. The proposed solution is to obtain computer output for a particular package such as SAS, GENSTAT, BMDP, etc. and to annotate the output explaining which computations have been performed, which have not, and which are still needed. If an incorrect or useless procedure has been given, it is so stated. A short description of annotated computer outputs prepared to date is given. Annotated computer outputs for five packages for principal component analyses, and for three packages for covariance in a split plot design have been prepared. Two technical reports and an annotated computer output have been written for cluster analysis. Copies of these reports are available from the Mathematical Sciences Institute.

COVARIANCE ANALYSIS FOR SPLIT PLOT AND SPLIT BLOCK DESIGNS AND COMPUTER PACKAGES

Walter T. Federer and Michael P. Meredith
337 Warren Hall, Biometrics Unit
Cornell University
Ithaca, NY 14853

BU-974-M*

May, 1988

1. INTRODUCTION. Split plot and split block designs appear to be rather mystifying to many individuals. They apparently are not cognizant of the many and varied forms these designs may take, the philosophical nature, concepts, and usage of the several error mean squares that are required, and the nature and use of covariance analyses for these designs. Since the computational procedure for an analysis of variance (ANOVA) for orthogonal split plot and split block designs are trivial, many individuals feel that the concepts are also simple. Computational procedures for an ANOVA do not explain concepts contrary to some opinions.

Yates (1937) described *one* type of split plot design as an example of a *class* of designs. Unfortunately this one type of split plot design is described as THE split plot design in almost all of statistical literature, especially in textbooks. Federer (1955, 1975, 1977) described some variations, some misconceptions, and possible population structures for these designs. With regard to the last point, a glaring omission in statistics textbooks is the failure to include any discussion of population structure for even the simplest of experiment designs. This necessarily raises the question about meaningful inferences when the population is undefined and undescribed.

---

* In the Technical Report Series of the Biometrics Unit, Cornell University, Ithaca, N.Y. 14853.

When analyses of covariance (ANCOVA) are attempted, the confusion continues. This becomes strikingly evident in outputs for computer packages purporting to give such analyses for any but the simplest of experiment designs (See, e.g., Federer, 1955, Federer *et al.* 1979, 1987a, 1987b, 1987c, and Searle *et al.* 1982a, 1982b, 1982c). The concept of a separate regression for each error mean square is lacking in a number of computer packages. Hence, if a package does supply output for means adjusted for a covariate, the adjusted means given are often incorrect. The fact that there may be as many regression coefficients as there are error mean squares appears not to be understood. Since many regression coefficients can be and are computed in an ANCOVA, it is important to understand which ones are to be used for adjusting means for covariates and why.

Herein we shall discuss only ANCOVA for three specific designs, i.e.

(i) the standard split plot design where the whole plot treatments are in a randomized complete block design and split plot treatments are randomized within each whole plot,

(ii) a split-split plot design which is the one in (i) except that the split plot is further split to have whole plot treatments, split plot treatments, and split-split plot treatments, and

(iii) a split block design or two-way whole plot design where each set of treatments are in a randodmized complete block design arrangement.

In addition, a list of available annotated computer outputs (ACOs) is given in the last section.

2. Split Plot Experiment Designs. The almost universal split plot experiment design discussed in statistics textbooks is the one wherein the whole plot treatments are in a randomized complete block design and the split plots are completely randomized within each whole plot. Denote this as the standard design. However, Federer (1955, 1975) has pointed out that there is a vast variety of split plot experiment designs which are used in practice. There are many different experiment designs for whole plot treatments as well as for split plot treatments. Also, almost all statistics textbooks confine their discussion to an ANOVA for the standard split plot design with no discussion of an ANCOVA or of an ANOVA for nonorthogonal situations. Computer packages such as SAS, GENSTAT, BMDP,

and others are set up to provide computations for nonorthogonal situations but a full description and use of computer output computations is lacking, resulting in a need for annotating computer output (ACO). S. R. Searle and several co-workers have been very active in this area. A list of ACOs prepared by this group is given later in the paper. It should be noted that Searle is currently updating a number of previously prepared ACOs.

In order to keep this paper relatively short, only the standard (or usual) split plot experiment design will be considered in detail. Many response models may be used for the vast variety of experiments designed as a split plot but we shall confine ourselves to the linear model in Federer (1955). Let the ijk*th* observation Y_{ijk} with an associated covariate Z_{ijk} be represented as follows:

$$Y_{ijk} = \mu + \rho_j + \tau_i + \delta_{ij} + \alpha_k + \alpha\tau_{ik} + \beta_1(\bar{Z}_{ij.} - \bar{Z}...) + \beta_2(Z_{ijk} - \bar{Z}_{ij.}) + \varepsilon_{ijk}, \quad (1)$$

where μ is an overall mean effect, τ_i is the effect of the i*th* whole plot treatment, α_k is the effect of the k*th* subplot treatment, $\alpha\tau_{ik}$ is the interaction effect for the ik*th* combination of whole plot treatment i and split plot treatment k, ρ_j is a random block effect distributed with mean zero and variance σ^2_ρ, δ_{ij} is a random whole plot error effect distributed with mean zero and variance σ^2_δ, ε_{ijk} is a random split plot error effect distributed with mean zero and variance σ^2_ε, $\bar{Z}_{ij.}$ is the mean of the covariate for the ij*th* whole plot, $\bar{Z}...$ is the over-all mean of the covariate (i.e., the usual dot and bar notation), $i = 1,\ldots, a$, $j = 1, \ldots, r$, $k = 1, \ldots, s$, β_1 is a whole plot linear regression coeficient of the Y whole plot residuals on the Z whole plot residuals, and β_2 is a split plot linear regression of the Y split plot residuals on the Z split plot residuals. Note that using estimates of β_1 and β_2, i.e, $\hat{\beta}_1$ and $\hat{\beta}_2$, to adjust means is the correct thing to do. The purpose of using covariates is to reduce the variation in observed Y variable means by measuring and using an associated covariate. The reduction must then occur in the error or residual line in the ANOVA. We have encountered individuals who did not use this regression to adjust treatment means but used another regression, e.g., on the total line in the ANCOVA. This is incorrect and possible with present computer packages by eliminating the effect of the covariate first.

In some situations, the formulation of the response model as in (1) is inappropriate. Although (1) could be appropriate for one variable or for

one investigation it may not be for another. Also, as formulated (1) has two error effects, the δ_{ij} and ϵ_{ijk}. When the whole plot treatments, e.g., represent a random sample of treatments from a population, then the τ_i are distributed with mean zero and variance σ^2_τ. An appropriate error term for the fixed split plot treatment effects α_k would be the whole plot by split plot treatment interaction mean square. The $\alpha\tau_{ik}$ would have $E_{i,k}\ \alpha\tau_{ik} = 0$ and variance $\sigma^2_{\alpha\tau}$. Likewise in an ANCOVA, the appropriate regression for split plot treatment means would be computed from the interaction line rather than the error (b) line (see Table 1). In other situations, the split plot treatments or both split plot and whole plot treatments could be considered as a random sample of treatments and the effects would be random rather than fixed effects. Appropriate modifications in ANOVA and ANCOVA would be required for both situations.

A response model for variable Y is formulated and then an ANCOVA as in Table 1 is appropriate for a single covariate Z related to the variable Y in a linear manner. Note that the relation between Y and Z could be polynomial or nonlinear in nature. The number of covariates, say c, may exceed one. This situation may be handled as a straight-forward extension but we shall not consider these additional complexities. For response model equation (1), the ANCOVA is given in Table 1. The sums of products are computed in the usual manner. For example, $T_{yz} = \sum\sum\sum_{ijk} Y_{ijk}\ Z_{ijk}$, $A_{yz} = \sum_i \sum_j \hat{\delta}_{yij}\ \hat{\delta}_{zij}$, where $\hat{\delta}_{yij}$ is the residual for the variable Y alone and $\hat{\delta}_{zij}$ is the residual for the variable Z alone, and $B_{yz} = \sum_i \sum_j \sum_k \hat{\epsilon}_{yijk}\ \hat{\epsilon}_{zijk}$, where the $\hat{\epsilon}_{hijk}$ are the computed split plot residuals for variable h = y, z. The above computations would still hold even for non-orthogonal experiment designs. The mean squares in ANCOVA are obtained by dividing by the appropriate degrees of freedom. If, in addition to an ANCOVA, it is desired to obtain F-statistics, the ratios $W'_{yy}(ar-r-a)$ / $A'_{yy}(a-1)$, $S'_{yy}[a(r-1)(s-1)-1]$ / $B'_{yy}(s-1)$, and $I'_{yy}[a(r-1)(s-1)-1]$ / $B'_{yy}(a-1)(s-1)$ may be computed. Given that the δ_{ij} and ϵ_{ijk} are NIID, the probability of obtaining a larger F-statistic may be obtained from prepared tables or computer programs. Even if normality does not hold, the probabilities will be approximately correct for most situations.

Table 1. ANCOVA for equation (1) for a split plot experiment design [1]

Source of Variation	Degrees of Freedom (df)	Sums of Products YY	YZ	ZZ	df	Adjusted Sums Of Squares
Total	ars	T_{yy}	T_{yz}	T_{zz}		
Correction for Mean	1	M_{yy}	M_{yz}	M_{zz}		
Block	(r-1)	R_{yy}	R_{yz}	R_{zz}		
Whole Plot = W	(a-1)	W_{yy}	W_{yz}	W_{zz}		
Error (a)	(a-1)(r-1)	A_{yy}	A_{yz}	A_{zz}	(ar-a-r)	$A_{yy}-A_{yz}^2 / A_{zz} = A'_{yy}$
Split Plot = S	(s-1)	S_{yy}	S_{yz}	S_{zz}		
S X W	(a-1)(s-1)	I_{yy}	I_{yz}	I_{zz}	(as-a-s)	$I_{yy}-I_{yz}^2 / I_{zz} = I'_{yy}$
Error (b)	a(r-1)(s-1)	B_{yy}	B_{yz}	B_{zz}	a(r-1)(s-1)-1	$B_{yy}-B_{yz}^2 / B_{zz} = B'_{yy}$
Whole Plot (adj. for $\hat{\beta}_1$)	(a-1)	$W_{yy} - \frac{(W_{yz} + A_{yz})^2}{W_{zz} + A_{zz}} + \frac{A_{yz}^2}{A_{zz}} = W'_{yy}$				
Split Plot (Adj. for $\hat{\beta}_2$)	(s-1)	$S_{yy} - \frac{(S_{yz} + B_{yz})^2}{S_{zz} + B_{zz}} + \frac{B_{yz}^2}{B_{zz}} = S'_{yy}$				
S X W (Adj. for $\hat{\beta}_2$)	(a-1)(s-1)	$I_{yy} - \frac{(I_{yz} + B_{yz})^2}{I_{zz} + B_{zz}} + \frac{B_{yz}^2}{B_{zz}} = I'_{yy}$				

[1] The various mean squares may be obtained by dividing by the appropriate degrees of freedom.

The various Y means adjusted for the covariate Z are:

$$\bar{Y}_{i..}(adj.) = \bar{Y}_{i..} - \hat{\beta}_1(\bar{Z}_{i..} - \bar{Z}_{...}) = \bar{Y}'_{i..} ,$$

$$\bar{Y}_{..k}(adj.) = \bar{Y}_{..k} - \hat{\beta}_2(\bar{Z}_{..k} - \bar{Z}_{...}) = \bar{Y}'_{..k},$$

and

$$\bar{Y}_{i.k}(adj.) = \bar{Y}_{i.k} - \hat{\beta}_1(\bar{Z}_{i..} - \bar{Z}_{...}) - \hat{\beta}_2(\bar{Z}_{i.k} - \bar{Z}_{i..}) = \bar{Y}'_{i.k} ,$$

where $\hat{\beta}_1 = A_{yz} / A_{zz}$, $\hat{\beta}_2 = B_{yz} / B_{zz}$, and the usual dot notation is used for the various means.

Estimated variances of a difference between two adjusted means for $i \neq i'$ and $k = k'$ are:

Variance of a difference between two adjusted whole plot treatment means

$$V\left(\bar{Y}'_{i..} - \bar{Y}'_{i'..}\right) = \left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right)\left[\frac{2}{sr} + \frac{\left(\bar{Z}_{i..} - \bar{Z}_{i'..}\right)^2}{A_{zz}}\right] .$$

Variance of a difference between two adjusted split plot treatment means

$$V\left(\bar{Y}'_{..k} - Y'_{..k'}\right) = \hat{\sigma}^2_\varepsilon\left[\frac{2}{ar} + \frac{\left(\bar{Z}_{..k} - \bar{Z}_{..k'}\right)^2}{B_{zz}}\right] .$$

Variance of a difference between two adjusted split plot treatment means for the same whole plot treatment

$$V\left(\bar{Y}'_{i.k} - \bar{Y}'_{i.k'}\right) = \hat{\sigma}^2_\varepsilon\left[\frac{2}{r} + \frac{\left(\bar{Z}_{i.k} - \bar{Z}_{i.k'}\right)^2}{B_{zz}}\right] .$$

Variance of a difference between two adjusted whole plot treatment means for the same split plot treatment

$$V\left(\bar{Y}'_{i.k} - \bar{Y}'_{i'.k}\right) = \frac{2}{r}\left(\hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_\delta\right) + \left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right)\frac{\left(\bar{Z}_{i..} - \bar{Z}_{i'..}\right)^2}{A_{zz}}$$

$$+ \hat{\sigma}^2_\varepsilon\frac{\left(\bar{Z}_{i.k} - \bar{Z}_{i'.k} - \bar{Z}_{i..} + \bar{Z}_{i'..}\right)^2}{B_{zz}} .$$

$$\left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right) = A'_{yy} / (ar-a-r),\ \hat{\sigma}^2_\varepsilon = B'_{YY} / [a(r-)(s-1)-1] ,$$

$$\text{and } \hat{\sigma}^2_\delta = \left[\left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right) - \hat{\sigma}^2_\varepsilon\right] / s .$$

$\hat{\sigma}^2_\varepsilon$ is associated with $a(r-1)(s-1)-1$ degrees of freedom, $\left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right)$ is associated with $ar-r-a$ degrees of freedom, and the degrees of freedom for the last variance above are approximated as the degrees of freedom f associated with

$$t_\alpha(f) = \frac{(s-1)\left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right) t_\alpha\,(ar-r-a) + \hat{\sigma}^2_\varepsilon t_\alpha\,[a(r-1)(s-1)-1]}{(s-1)\left(\hat{\sigma}^2_\varepsilon + s\hat{\sigma}^2_\delta\right) + \hat{\sigma}^2_\varepsilon}$$

where t_{α} (f) is the tabulated value of the t-statistic at the α percentage level for f degrees of freedom. This approximation underestimates the degrees of freedom for this variance (see Cochran and Cox, 1950, and Grimes and Federer, 1984).

Given the above variances, one may now use a multiple range procedure to compare individual pairs of means. Some authors (e.g. Cochran and Cox, 1950) consider that there is a correlation between the split plot experimental units. Hence, the whole plot expected error mean square would be given as σ^2 and the split plot error would be written as $\sigma^2(1-\rho) = \sigma^2_{\varepsilon}$ where the correlation ρ is equal to $s\sigma^2_{\delta} / \sigma^2$. Although this formulation is useful for many situations it is not of universal application; e.g. when measurement error or competition exists between split plot experimental units but not between whole plot experimental units. Statistical modeling for any investigation should be carefully considered.

3. Split-Split Plot Experiment Designs. For this class of designs, various experiment designs may be used for whole plot treatments, for split plot treatments, and for split-split plot treatments. However, we shall confine our remarks to a single member of this class, i.e., the whole plot treatments are arranged in a randomized complete blocks design, the split plot treatments are randomly allocated to the split plot experimental units within each whole plot unit, and the split-split plot treatments are randomly assigned to the split-split plot experimental units within each split plot experimental unit. There will be r randomizations for the a whole plot treatments, ra randomizations for the s split plot treatments, and ras randomizations for the p split-split plot treatments. The

treatment design considered here is a three factor factorial with asp combinations, but it should be noted that other treatment designs are possible. The factors are assumed to be fixed effects to simplify presentation.

One possible response model for the above experiment and treatment design for a variable Y with a covariate Z is:

$$Y_{hijk} = \mu + \rho_h + \tau_i + \delta_{hi} + \beta_1 (\bar{Z}_{hi..} - \bar{Z}_{....}) + \alpha_j + \alpha\tau_{ij} + \varepsilon_{hij}$$
$$+ \beta_2 (\bar{Z}_{hij.} - \bar{Z}_{hi..}) + \gamma_k + \gamma\tau_{ik} + \alpha\gamma_{jk} + \alpha\gamma\tau_{ijk} + \pi_{hijk}$$
$$+ \beta_3 (Z_{hijk} - \bar{Z}_{hij.}) \quad , \qquad (2)$$

where the first nine effects are as defined for equation (1), γ_k is the effect of the k*th* split-split plot treatment, $\gamma\tau_{ik}$ is a two-factor interaction effect for combination ik, $\alpha\gamma_{jk}$ is a two-factor interaction effect for combination jk, $\alpha\gamma\tau_{ijk}$ is a three-factor interaction effect for combination ijk, π_{hijk} is a random error effect associated with split-split plot experimental unit hijk and distributed with mean zero and variance σ^2_π, β_3 is a linear regression coefficient of the split-split plot Y residuals on the corresponding Z residuals, h = 1, ..., r, i = 1, ..., a, j = 1, ..., s, and k = 1, ..., p. An ANCOVA for this design and response model is given in Table 2.

The various adjusted means are computed as:

$$\bar{Y}_{.i..}(adj.) = \bar{Y}_{.i..} - \hat{\beta}_1(\bar{Z}_{.i..} - \bar{Z}_{....}) = \bar{Y}'_{.i..} \quad ,$$

$$\bar{Y}_{..j.}(adj.) = \bar{Y}_{..j.} - \hat{\beta}_2(\bar{Z}_{..j.} - \bar{Z}_{....}) = \bar{Y}'_{..j.} \quad ,$$

$$\bar{Y}_{...k}(adj.) = \bar{Y}_{...k} - \hat{\beta}_3(\bar{Z}_{...k} - \bar{Z}_{....}) = \bar{Y}'_{...k} \quad ,$$

$$\bar{Y}_{.ij.}(adj.) = \bar{Y}_{.ij.} - \hat{\beta}_1(\bar{Z}_{.i..} - \bar{Z}_{....}) - \hat{\beta}_2(\bar{Z}_{.ij.} - \bar{Z}_{.i..}) = \bar{Y}'_{.ij.} \quad ,$$

$$\bar{Y}_{.i.k}(adj.) = \bar{Y}_{.i.k} - \hat{\beta}_1(\bar{Z}_{.i..} - \bar{Z}_{....}) - \hat{\beta}_3(\bar{Z}_{.i.k} - \bar{Z}_{.i..}) = \bar{Y}'_{.i.k} \quad ,$$

$$\bar{Y}_{..jk}(adj.) = \bar{Y}_{..jk} - \hat{\beta}_2(\bar{Z}_{..j.} - \bar{Z}_{....}) - \hat{\beta}_3(\bar{Z}_{..jk} - \bar{Z}_{.j..}) = \bar{Y}'_{..jk} \quad ,$$

Table 2. ANCOVA for equation (2) for a split-split plot experiment design [1]

Source of Variation	df	Sums of Products yy	yz	zz	df	Adjusted Sums Of Squares
Total	rasp	T_{yy}	T_{yz}	T_{zz}		
Correction for Mean	1	M_{yy}	M_{yz}	M_{zz}		
Block	(r-1)	R_{yy}	R_{yz}	R_{zz}		
Whole Plot = W	(a-1)	W_{yy}	W_{yz}	W_{zz}		
Error (a)	(a-1)(r-1)	A_{yy}	A_{yz}	A_{zz}	(ar-r-a)	$A_{yy}-A_{yz}^2 / A_{zz} = A'_{yy}$
Split Plot = S	(s-1)	S_{yy}	S_{yz}	S_{zz}		
S X W	(a-1)(s-1)	I_{yy}	I_{yz}	I_{zz}		
Error (b)	a(r-1)(s-1)	B_{yy}	B_{yz}	B_{zz}	a(r-1)(s-1)-1	$B_{yy}-B_{yz}^2 / B_{zz} = B'_{yy}$
Split-Split Plot = P	(p-1)	P_{yy}	P_{yz}	P_{zz}		
W X P	(a-1)(p-1)	Q_{yy}	Q_{yz}	Q_{zz}		
S X P	(p-1)(s-1)	U_{yy}	U_{yz}	U_{zz}		
W X S X P	(a-1)(p-1)·(s-1)	V_{yy}	V_{xy}	V_{zz}		
Error (c)	as(r-1)(p-1)	C_{yy}	C_{yz}	C_{zz}	as(r-1)(p-1)-1	$C_{yy}-C_{yz}^2 / C_{zz} = C'_{yy}$
W(adj. for $\hat{\beta}_1$)	(a-1)					$W_{yy} - (W_{yz} + A_{yz})^2 / (W_{zz} + A_{zz}) + A_{yz}^2 / A_{zz} = W'_{yy}$
S(adj. for $\hat{\beta}_2$)	(s-1)					$S_{yy} - (S_{yz} + B_{yz})^2 / (S_{zz} + B_{zz}) + B_{yz}^2 / B_{zz} = S'_{yy}$
SXW(adj. for $\hat{\beta}_2$)	(a-1)(s-1)					$I_{yy} - (I_{yz} + B_{yz})^2 / (I_{zz} + B_{zz}) + B_{yz}^2 / B_{zz} = I'_{yy}$
P(adj. for $\hat{\beta}_3$)	(p-1)					$P_{yy} - (P_{yz} + C_{yz})^2 / (P_{zz} + C_{zz}) + C_{yz}^2 / C_{zz} = P'_{yy}$
WXP(adj. for $\hat{\beta}_3$)	(a-1)(p-1)					$Q_{yy} - (Q_{yz} + C_{yz})^2 / (Q_{zz} + C_{zz}) + C_{yz}^2 / C_{zz} = Q'_{yy}$
SXP(adj. for $\hat{\beta}_3$)	(p-1)(s-1)					$U_{yy} - (U_{yz} + C_{yz})^2 / (U_{zz} + C_{zz}) + C_{yz}^2 / C_{zz} = U'_{yy}$
WXSXP(adj. for $\hat{\beta}_3$)	(a-1)(p-1)(s-1)					$V_{yy} - (V_{yz} + C_{yz})^2 / (V_{zz} + C_{zz}) + C_{yz}^2 / C_{zz} = V'_{yy}$

[1] The various mean squares may be obtained by dividing by the appropriate degrees of freedom.

and

$$\bar{Y}_{.ijk}\,(\text{adj.}) = \bar{Y}_{.ijk} - \hat{\beta}_1(\bar{Z}_{.i..} - \bar{Z}_{....}) - \hat{\beta}_2(\bar{Z}_{.ij.} - \bar{Z}_{.i..})$$

$$- \hat{\beta}_3(\bar{Z}_{.ijk} - \bar{Z}_{.ij.}) = \bar{Y}'_{.ijk} \quad ,$$

where $\hat{\beta}_1 = A_{yz} / A_{zz}$, $\hat{\beta}_2 = B_{yz} / B_{zz}$, and $\hat{\beta}_3 = C_{yz} / C_{zz}$.

Estimated variances of a difference between two means adjusted for a covariate for $i \neq i'$, $j \neq j'$, $E_a = A'_{yy} / (ar-r-a)$, $E_b = B'_{yy} / [a(r-1)(s-1)-1]$, and $\hat{\sigma}^2_{\pi} = C'_{yy} / [as(r-1)(p-1)-1]$ are given below:

Variance of a difference between two whole plot treatment adjusted means

$$V(\bar{Y}'_{.i..} - \bar{Y}'_{.i'..}) = E_a \left[\frac{2}{rsp} + \frac{(\bar{Z}_{.i..} - \bar{Z}_{.i'..})^2}{A_{zz}} \right] .$$

Variance of a difference between two split plot treatment adjusted means

$$V(\bar{Y}'_{..j.} - \bar{Y}'_{..j'.}) = E_b \left[\frac{2}{arp} + \frac{(\bar{Z}_{..j.} - \bar{Z}_{..j'.})^2}{B_{zz}} \right] .$$

Variance of a difference between two split-split plot treatment adjusted means

$$V(\bar{Y}'_{...k} - \bar{Y}'_{...k'}) = \hat{\sigma}^2_{\pi} \left[\frac{2}{ars} + \frac{(\bar{Z}_{...k} - \bar{Z}_{...k'})^2}{C_{zz}} \right] .$$

Variance of a difference between two adjusted means for combinations ij and ij'

$$V(\bar{Y}'_{.ij.} - \bar{Y}'_{.ij'.}) = E_b \left[\frac{2}{rp} + \frac{(\bar{Z}_{.ij.} - \bar{Z}_{.ij'.})^2}{B_{zz}} \right] .$$

Variance of a difference between two adjusted means for combinations ij and i'j

$$V(\bar{Y}'_{.ij.} - \bar{Y}'_{.i'j.}) = \frac{2}{rp} [\hat{\sigma}^2_{\delta} + \hat{\sigma}^2_{\varepsilon}] + E_a \frac{(\bar{Z}_{.i..} - \bar{Z}_{.i'..})^2}{A_{zz}}$$

$$+ E_b \frac{(\bar{Z}_{.ij.} - \bar{Z}_{.i..} - \bar{Z}_{.i'j.} + \bar{Z}_{.i'..})^2}{B_{zz}} .$$

Variance of a difference between two adjusted means for combinations ik and ik'

$$V(\bar{Y}'_{.i.k} - \bar{Y}'_{.i.k'}) = \hat{\sigma}^2_{\pi} \left[\frac{2}{rs} + \frac{(\bar{Z}_{.i.k} - \bar{Z}_{.i.k'})^2}{C_{zz}} \right] .$$

Variance of a difference between two adjusted means for combinations ik and i'k

$$V(\bar{Y}'_{.i.k} - \bar{Y}'_{.i'.k}) = \frac{2(s\hat{\sigma}^2_\delta + \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_\pi)}{rs} + E_a \frac{(\bar{Z}_{.i..} - \bar{Z}_{.i'..})^2}{A_{zz}}$$

$$+ \frac{\hat{\sigma}^2_\pi (\bar{Z}_{.i.k} - \bar{Z}_{.i..} - \bar{Z}_{.i'.k} + \bar{Z}_{.i'..})^2}{C_{zz}} .$$

Variance of a difference between two adjusted means for combinations ijk and ijk'

$$V(\bar{Y}'_{.ijk} - \bar{Y}'_{.ijk'}) = \frac{2}{r}(\hat{\sigma}^2_\pi) + \hat{\sigma}^2_\pi \frac{(\bar{Z}_{.ijk} - \bar{Z}_{.ijk'})^2}{C_{zz}} .$$

Variance of a difference between two adjusted means for combinations ijk and ij'k

$$V(\bar{Y}'_{.ijk} - \bar{Y}'_{ij'k}) = \frac{2}{r}(\hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_\pi) + \frac{E_b}{B_{zz}} (\bar{Z}_{.ij.} - \bar{Z}_{.ij'.})^2$$

$$+ \frac{\hat{\sigma}^2_\pi}{C_{zz}} (\bar{Z}_{.ijk} - \bar{Z}_{.ij.} - \bar{Z}_{.ij'k} + \bar{Z}_{.ij'.})^2 .$$

Variance of a difference between two adjusted means for combinations ijk and i'jk

$$V(\bar{Y}'_{.ijk} - \bar{Y}'_{.i'jk}) = \frac{2}{r}(\hat{\sigma}^2_\delta + \hat{\sigma}^2_\varepsilon + \hat{\sigma}^2_\pi) + \frac{E_a}{A_{zz}} (\bar{Z}_{.i..} - \bar{Z}_{.i'..})^2$$

$$+ \frac{E_b}{B_{zz}} (\bar{Z}_{.ij.} - \bar{Z}_{.i..} - \bar{Z}_{.i'j.} + \bar{Z}_{.i'..})^2$$

$$+ \frac{\hat{\sigma}^2_\pi}{C_{zz}} (\bar{Z}_{.ijk} - \bar{Z}_{.ij.} - \bar{Z}_{.i'jk} + \bar{Z}_{.i'j.})^2 .$$

Note that $V(\bar{Y}'_{.ijk} - \bar{Y}_{.i'jk}) = V(\bar{Y}'_{.ijk} - \bar{Y}'_{.i'j'k}) = V(\bar{Y}'_{.ijk} - \bar{Y}'_{.i'jk'})$ $= V(\bar{Y}'_{.ijk} - \bar{Y}'_{.i'j'k'})$ and that $V(\bar{Y}'_{.ijk} - \bar{Y}'_{.ij'k}) = V(\bar{Y}'_{.ijk} - \bar{Y}_{.ij'k'})$.

Most variances above without the covariate were given by Federer (1955). Also, the expected values of E_a and E_b are $\sigma^2_\pi + p\sigma^2_\varepsilon + ps\sigma^2_\delta$ and $\sigma^2_\pi + p\sigma^2_\varepsilon$, respectively. Estimates of variance components σ^2_δ, σ^2_ε, and $E(\hat{\sigma}^2_\pi) = \sigma^2_\pi$ are needed to compute the fifth, seventh, ninth, and tenth variances above.

The degrees of freedom for these variances need to be approximated as they were in the previous section. Also note that $ps(\hat{\sigma}^2_\pi + \hat{\sigma}^2_\epsilon + \hat{\sigma}^2_\delta) = s(p-1)\hat{\sigma}^2_\pi + (s-1)E_b + E_a$ and $p(\hat{\sigma}^2_\pi + \hat{\sigma}^2_\epsilon) = (p-1)\hat{\sigma}^2_\pi + E_b$.

4. Split Block Experiment Design. The experiment design considered here is denoted as a split block design. It has also been called a two-way whole plot and a strip trial design. This design has received no attention in statistical textbooks with an exception being Federer (1955). It does occur frequently in practice but sometimes is not analyzed correctly as a split block design. The member of this class of designs we shall discuss will be for a two-factor factorial treatment design with the levels of one factor being applied perpendicularly across all levels of the second factor within each replicate or complete block. The levels of each factor will have the same design for our example, that is a randomized complete block design. (The levels of one factor could be in a randomized complete block design and the levels of the second factor could be in a latin square, balanced incomplete block, or other experiment design.) Note that there will be r separate randomizations for the levels of each of the factors. The number of levels of factor one is a and the number of levels of the second factor is b, resulting in an a × b factorial treatment design.

A response model equation as given in Federer (1955) for a variable Y and a covariate Z is:

$$Y_{hij} = \mu + \rho_h + \alpha_i + \delta_{hi} + \gamma_j + \pi_{kj} + \alpha\gamma_{ij} + \epsilon_{hij} + \beta_1(\bar{Z}_{hi\cdot} - \bar{Z}_{\cdots}) + \beta_2(\bar{Z}_{h\cdot j} - \bar{Z}_{\cdots}) + \beta_3(Z_{hij} - \bar{Z}_{hi\cdot} - \bar{Z}_{h\cdot j} + \bar{Z}_{\cdots}) , \quad (3)$$

where μ is a general mean effect, ρ_h is the h*th* block effect, which has mean zero and variance σ^2_ρ, α_i is the effect of the i*th* level of factor one, say A, γ_j is the effect of the j*th* level of factor two, say B, δ_{hi} is a random error effect for the hi*th* whole plot for factor A and has mean zero and variance σ^2_δ, π_{hi} is a random error effect for the hj*th* whole plot for factor B and has mean zero and variance σ^2_π, $\alpha\gamma_{ij}$ is the interaction effect for the ij*th* combination of levels of factors A and B, ϵ_{hij} is a random error effect associated with the hij*th* subplot for the

A × B interaction and has mean zero and variance σ_ε^2, β_1 is the linear regression of Y whole plot residuals on the Z whole plot residuals for factor A, β_2 is the linear regression of the Y whole plot residuals on the Z whole plot residuals for factor B, and β_3 is the linear regression of Y subplot residuals on 2 subplot residuals.

An ANCOVA for response model (3) is given in Table 3. For this design and for fixed effects for the a × b factorial, there are three error variances and three error regressions. Given that the error effects are NIID, the usual F statistics may be used if desired. The adjusted means are given by:

$$\bar{Y}_{.i.}(\text{adjusted}) = \bar{Y}_{.i.} - \hat{\beta}_1(\bar{Z}_{.i.} - \bar{Z}_{...}) = \bar{Y}'_{.i.} \quad ,$$

$$\bar{Y}_{..j}(\text{adjusted}) = \bar{Y}_{..j} - \hat{\beta}_2(\bar{Z}_{..j} - \bar{Z}_{...}) = \bar{Y}'_{..j} \quad ,$$

and

$$\bar{Y}_{.ij}(\text{adjusted}) = \bar{Y}_{.ij} - \beta_1(\bar{Z}_{.i.} - \bar{Z}_{...}) - \beta_2(\bar{Z}_{..j} - \bar{Z}_{...}) - \hat{\beta}_3(\bar{Z}_{.ij} - \bar{Z}_{.i.} - \bar{Z}_{..j} + \bar{Z}_{...}) = \bar{Y}'_{.ij} \quad ,$$

where the $\hat{\beta}$s are defined in Table 3.

Estimated variances of a difference between adjusted means are given below for $i \neq i'$, $j \neq j'$:

$$V(\bar{Y}'_{.i.} - \bar{Y}'_{.i'.}) = E_a\left[\frac{2}{rb} + \frac{(\bar{Z}_{.i.} - \bar{Z}_{.i'.})^2}{A_{zz}}\right] \quad ,$$

$$V(\bar{Y}'_{..j} - \bar{Y}'_{..j'}) = E_b\left[\frac{2}{ab} + \frac{(\bar{Z}_{..j} - \bar{Z}_{..j'})^2}{B_{zz}}\right] \quad ,$$

$$V(\bar{Y}'_{.ij} - \bar{Y}'_{.ij'}) = \frac{1}{r}(\hat{\sigma}_\pi^2 + \hat{\sigma}_\varepsilon^2) + \frac{E_b}{B_{zz}}(\bar{Z}_{..j} - \bar{Z}_{..j'})^2 + \frac{E_c}{C_{zz}}(\bar{Z}_{.ij} - \bar{Z}_{.ij'} - \bar{Z}_{..j} + \bar{Z}_{..j'})^2 \quad ,$$

$$V(\bar{Y}'_{.ij} - \bar{Y}'_{.i'j}) = \frac{1}{r}(\hat{\sigma}_\delta^2 + \hat{\sigma}_\varepsilon^2) + \frac{E_a}{A_{zz}}(\bar{Z}_{.i.} - \bar{Z}_{.i'.})^2 + \frac{E_c}{C_{zz}}(\bar{Z}_{.ij} - \bar{Z}_{.i'j} - \bar{Z}_{.i.} + \bar{Z}_{.i'.})^2 \quad ,$$

and

$$V(\bar{Y}'_{\cdot ij} - \bar{Y}'_{\cdot i'j'}) = \frac{1}{r}(\hat{\sigma}^2_\delta + \hat{\sigma}^2_\pi + \hat{\sigma}^2_\varepsilon) + \frac{E_a}{A_{zz}}(\bar{Z}_{\cdot i\cdot} - \bar{Z}_{\cdot i'\cdot})^2 + \frac{E_b}{B_{zz}}(\bar{Z}_{\cdot\cdot j} - \bar{Z}_{\cdot\cdot j'})^2$$

$$+ \frac{E_c}{C_{zz}}(\bar{Z}_{\cdot ij} - \bar{Z}_{\cdot i'j'} - \bar{Z}_{\cdot i\cdot} + \bar{Z}_{\cdot i'\cdot} - \bar{Z}_{\cdot\cdot j} + \bar{Z}_{\cdot\cdot j'})^2 ,$$

where $E_a = A'_{yy}/(ar-a-r) = \hat{\sigma}^2_\varepsilon + b\hat{\sigma}^2_\delta$, $E_b = B'_{yy}/(br-b-r) = \hat{\sigma}^2_\varepsilon + a\hat{\sigma}^2_\pi$,

and $E_c = C'_{yy}/[(a-1)(b-1)(r-1)-1] = \hat{\sigma}^2_\varepsilon$.

The degrees of freedom for the last three variances need to be approximated by the method previously given or by some other appropriate approximation (See e.g., Grimes and Federer, 1984).

Table 3. ANCOVA for equation (3) for a split block experiment design.

Source of variation	df	Sum of products	df	Adjusted sums of squares
Total	rab	$T_{yy}\ T_{yz}\ T_{zz}$		
Correction for mean	1	$M_{yy}\ M_{yz}\ M_{zz}$		
Replicate = R	(r-1)	$R_{yy}\ R_{yz}\ R_{zz}$		
Whole plot A	(a-1)	$W_{yy}\ W_{yz}\ W_{zz}$		
Error (a)	(r-1)(a-1)	$A_{yy}\ A_{yz}\ A_{zz}$	$(ra-a-r)A_{yy} - \frac{A^2_{yz}}{A_{zz}} = A'_{yy}$	
Whole A adjusted for $\hat{\beta}_1 = A_{yz}/A_{zz}$			$(a-1)W_{yy} - \frac{(W_{yz}+A_{yz})^2}{W_{zz}+A_{zz}} + \frac{A^2_{yz}}{A_{zz}} = W'_{yy}$	
Whole plot B	(b-1)	$S_{yy}\ S_{yz}\ S_{zz}$		
Error (b)	(b-1)(r-1)	$B_{yy}\ B_{yz}\ B_{zz}$	$(rb-b-r)B_{yy} - \frac{B^2_{yz}}{B_{zz}} = B'_{yy}$	
Whole plot B adjusted for $\hat{\beta}_2 = B_{yz}/B_{zz}$			$(b-1)S_{yy} - \frac{(S_{yz}+B_{yz})^2}{S_{zz}+B_{zz}} + \frac{B^2_{yz}}{B_{zz}} = S'_{yy}$	
A X B	(a-1)(b-1)	$I_{yy}\ I_{yz}\ I_{zz}$		
Error (ab)	(r-1)(a-1)(b-1)	$C_{yy}\ C_{yz}\ C_{zz}$	$(r-1)(a-1)(b-1)-1\ C_{yy} - \frac{C^2_{yz}}{C_{zz}} = C'_{yy}$	
Interaction adjusted for $\hat{\beta}_3 = C_{yz}/C_{zz}$			$(a-1)(b-1)I_{yy} - \frac{(I_{yz}+C_{yz})^2}{I_{zz}+C_{zz}} + \frac{C^2_{yz}}{C_{zz}} = I'_{yy}$	

5. Some Comments. Since formulas for many of the above adjusted means and variances do not appear in statistical literature, it was deemed appropriate to include them here. As can be seen from the analyses for relatively simple designs from each of the three classes, there are a variety of formulas for adjusted means and variances of differences between two adjusted means. The more complex members of each class may have 5, 10, 15, or 20 error mean squares and the same number of regression coefficients. Experiments are conducted wherein some of the factors are arranged in split blocks and others in split plot arrangements. Many different designs may be used for the different factors (See e.g., Federer, 1955, 1975). The most complex experiment design encountered is described by Federer and Farden (1955), where there are several split plot and several split block arrangements with a total of 75 error mean squares and 203 lines in the ANOVA.

One method of aiding investigators with ANOVAs and ANCOVAs of complexly designed experiments is to ascertain how much of a statistical analysis can be obtained with computer packages such as SAS, BMDP, GENSTAT, SPSS, and others. Then, the output can be annotated, i.e. an explanation is appended to the computer output describing what has been computed and how to use the results. Annotated computer outputs for two different split plot designs with a covariate have been completed for SAS, BMDP, and GENSTAT (see Federer *et al.* 1987a, 1987b, 1987c). In addition to these covariance analyses, annotated computer outputs have been prepared for principal component analysis from five computer packages and the mixture method of cluster analysis on SAS. A listing of these is given in Appendix A. A second list of material available from the Biometrics Unit is given in Appendix B.

The analyses have been described for a single covariate. Noting that $A_{yy} - A_{yz}^2/A_{zz} = A_{yy}(1-r_{yz}^2) = A'_{yy}$, one may simply use $A_{yy}(1-R^2) = A'_{yy}$ when there are several covariates and where R^2 is the squared multiple correlation coefficient computed on the error line. If the relationship between a covariate Z and Y is curvilinear, it may be possible to use some function of Z, e.g. log Z, $\sqrt{Z}$, 1/Z, which makes the relation linear. If this can be accomplished both computations and interpretations are simplified.

A simplification of the estimated variances for differences of means has been given by Yates (1934) and Finney (1946). Instead of computing the

quantities $(\bar{Z}_{.i.} - \bar{Z}_{.i'.})^2 / A_{zz}$ and $(\bar{Z}_{..j} - \bar{Z}_{..j'})^2 / B_{zz}$, e.g, for each pair of means, one may compute a single variance by using $W_{xx}/(a-1)A_{zz}$ and $S_{xx}/(s-1)B_{zz}$, respectively. The quantity $W_{xx}/(a-1)$ is an average of all pairs ii' of $(\bar{Z}_{.i.} - \bar{Z}_{.i'.})^2$. This simplication and approximation considerably reduces the number of computations for large a and/or s. For the quantities $(\bar{Z}_{.ij} - Z_{.ij'} - \bar{Z}_{.i.} + \bar{Z}_{.i'.})^2$ and $(\bar{Z}_{.ij} - \bar{Z}_{.i'j} - \bar{Z}_{..j} + \bar{Z}_{..j.})^2$ it is suggested that $I_{xx} / (a-1)(s-1)B_{zz}$ be used if it is desired to compute only a single variance.

6. LITERATURE CITED.

Cochran, W. G. and G. M. Cox (1950). *Experimental Designs*, (2nd edition in 1957). John Wiley and Sons, Inc., New York, pp. 219-224.

Federer, W. T. (1955). Experimental Design - Theory and Application, Macmillan Co., New York (republished by Oxford and IBH Publishing Co., New Delhi, India, 1964, 1974).

Federer, W. T. (1975) The misunderstood split plot. In *Applied Statistics* (editor, R. P. Gupta) North-Holland Publishing Co., Amsterdam, Oxford, pp. 1-39.

Federer, W. T. (1977). Sampling, blocking, and model considerations for split plot and split block designs. Biometrical Journal *19*, 181-200.

Federer, W. T. and C. A. Farden (1955). Analysis of variance set-up for Joint Project 69. BU-67-M in the Technical Report Series of the Biometrics Unit, Cornell University.

Federer, W. T., Z. D. Feng, and N. J. Miles-McDermott (1987a). Annotated computer output for split plot design: GENSTAT. Annotated Computer Output '87-4, Mathematical Sciences Institute, Cornell University, 201 Caldwell Hall, Ithaca, N.Y. 14853.

Federer, W. T., Z. D. Feng, and N. J. Miles-McDermott (1987b). Annotated computer output for split plot design: BMDP 2V. Annotated Computer Output '87-5, Mathematical Sciences Institute, 201 Caldwell Hall, Ithaca, N.Y. 14853.

Federer, W. T., Z. D. Feng, M. P. Meredith, and N. J. Miles-McDermott (1987c). Annotated computer output for split plot design: SAS GLM. Annotated Computer Output '87-8, Mathematical Sciences Institute, Cornell University, 201 Caldwell Hall, Ithaca, N.Y. 14853.

Federer, W. T. and H. V. Henderson (1979). Covariance analysis of designed experiments x statistical packages: An update. Proc., Computer Science and Statistics, 12th Annual Symposium on the Interface, pp 228-235.

Finney, D. J., (1946). Standard errors of yields adjusted for regression on an independent measurement. Biometrics *2*, 53-55.

Grimes, B. A. and W. T. Federer (1984). Comparison of means from populations with unequal variances. In *W. G. Cochran's Impact on Statistics* (editors P. S. R. S. Rao and J. Sedransk), John Wiley and Sons, Inc., New York, pp. 353-374.

Searle, S. R., G. F. S. Hudson, and W. T. Federer (1982a). Annotated computer output for covariance (ACO: COV-GENSTAT). BU-782-M in Technical Report Series of the Biometrics Unit, Cornell University, July.

Searle, S. R., G. F. S. Hudson, and W. T. Federer (1982b). Annotated computer output for covariance (ACO: COV-SAS). BU-783-M in the Technical Report Series of the Biometrics Unit, Cornell University, July.

Searle, S. R., G. F. S. Hudson, and W. T. Federer (1982c). Annotated computer output for covariance (ACO: COV-SPSS). BU-784-M in the Technical Report Series of the Biometrics Unit, Cornell University, July.

Yates, F. (1934). A complex pig-feeding experiment. Journal of Agricultural Science. *24*, pp. 511-534.

Yates, F. (1937). The design and analysis of factorial experiments. Imperial Bur. Soil Sci., Tech Communication *35*, pp. 1-95.

Appendix A MSI ANNOTATED COMPUTER OUTPUT

ORDER FORM

1. COVARIANCE ANALYSIS FOR SPLIT PLOT DESIGN

	Office Ref.		
SAS	ACO #87-8	___copies at $ 5 each+	$____
BMDP 2V	ACO #87-5	___copies at $ 5 each+	$____
GENSTAT	ACO #87-4	___copies at $ 5 each+	$____

2. PRINCIPAL COMPONENT ANALYSIS

SAS	ACO #86-1	___copies at $ 5 each+	$____
SYSTAT	ACO #87-6	___copies at $ 5 each+	$____
BMDP	ACO #87-7	___copies at $ 5 each+	$____
SPSS-X	ACO #87-2	___copies at $ 5 each+	$____
GENSTAT	ACO #87-3	___copies at $ 5 each+	$____

3. CLUSTER ANALYSIS (Mixture Method)

TEXT	TR #86-38	___copies at $ 5 each+	$____
SAS	TR #87-5	___copies at $ 5 each+	$____
(Comparing 2 Clustering Methods to the Mixture Model Method)			
SAS	ACO #87-1	___copies at $ 5 each+	$____
(Annotated Computer Output for SAS, above)			

TOTAL.....$____

+One copy is free for U.S. Army Personnel upon request.

Send Check (payable to Cornell University) to:
Mathematical Consulting Liaison Group
Mathematical Sciences Institute
294 Caldwell Hall
Cornell University
Ithaca, New York, 14853, U.S.A.

The above order is to be sent to:

_____ (please print)

_____

_____

_____

NOTE: Orders will be mailed only after funds are received. This is our only invoice.

Appendix B

ANNOTATED COMPUTER OUTPUT (ACO)
ORDER FORM

Second Edition: ACO_2, 1988-9

[i] Analysis of Variance — Office Reference

BMDP2V.....(due Aug.'88).... .....___copies at $12 each $____
GENSTAT.ANOVA.....(due Apr.'88)....962.....___copies at $12 each $____
SAS GLM.....949.....___copies at $12 each $____
SAS HARVEY.(First Edition).....659.....___copies at $ 5 each $____
SPSSX ANOVA.....955.....___copies at $12 each $____

[ii] Variance Component Estimation

$ACO\sigma^2$: BMDP-V.....(due Feb.'89).... .....___copies at $12 each $____
$ACO\sigma^2$: SAS HARVEY.(First Edition).....723.....___copies at $ 5 each $____
$ACO\sigma^2$: SAS RANDOM.....(due June'88).... .....___copies at $12 each $____
$ACO\sigma^2$: SAS GLM VARCOMP..(due June'88)..... .....___copies at $12 each $____

First Edition: ACO COV, 1982

[iii] Analysis of Covariance

Text.....780.....___copies at $20 each $____
BMD(P1V, P2V, P4V).....781.....___copies at $ 5 each $____
GENSTAT (ANOVA).....782.....___copies at $ 5 each $____
SAS (GLM and HARVEY).....783.....___copies at $10 each $____
SPSS (ANOVA, MANOVA).....784.....___copies at $10 each $____

Other Publications Available

1. SOLUTIONS MANUAL to Searle's *LINEAR MODELS*.....___copies at $ 7 each $____

2. SOLUTIONS MANUAL to Searle's *MATRIX ALGEBRA USEFUL FOR STATISTICS*.....___copies at $ 7 each $____

3. NOTES ON VARIANCE COMPONENTS by S.R. Searle.....___copies at $ 7 each $____

4. PROCEEDINGS: *STATISTICAL DESIGN THEORY & PRACTICE, CONFERENCE IN HONOR OF W.T. FEDERER*, 1986.....___copies at $20 each $____

5. EXERCISES FOR SIMPLE REGRESSION
   Program REGDATA.....___copies at $ 5 each $____
   List of 100 Data Sets.....___copies at $ 5 each $____

6. *BIBLIOGRAPHY ON EXPERIMENT AND TREATMENT DESIGN - PRE-1968* by W.T. Federer and L.N. Balaam.....___copies at $10 each $____

TOTAL.....$____

7. *EXPERIMENTAL DESIGN* by W.T. Federer.....___copies at $13 each $____
   (check payable to W.T. Federer)

Send check (payable to Cornell University) to:
Biometrics Unit
336 Warren Hall
Cornell University
Ithaca, New York, 14853, U.S.A.

The above order is to be sent to:

_____ (please print)

_____

_____

_____

NOTE: Orders will be mailed only after funds are received. This is our only invoice.

3/88

ALTERNATIVES TO HYPOTHESIS TESTING INCLUDING A MAXIMUM LIKELIHOOD ESTIMATE TECHNIQUE

NATHANAEL ROMAN
ARMY MATERIEL TEST AND EVALUATION DIRECTORATE
WHITE SANDS MISSILE RANGE, NEW MEXICO

1. ABSTRACT

Hypothesis testing is often used to make decisions with respect to random data.

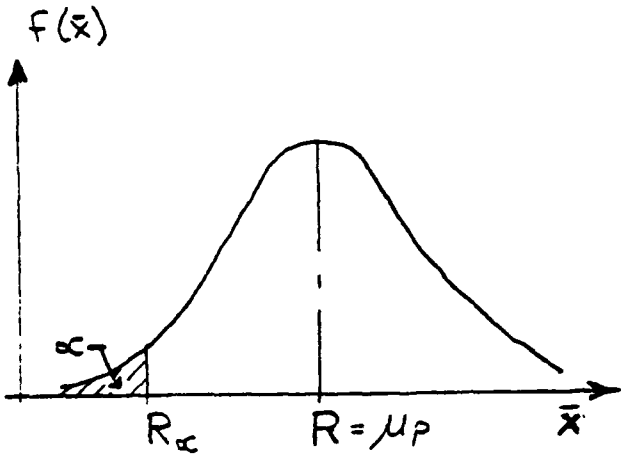
In Army air defense system specifications, criteria for hypothesis testing are rarely defined. Therefore, selection of pass/fail and risk criteria is arbitrary. Criteria are often chosen that tend to minimize the contractor's or developer's risk and to maximize the system user's risk. Hence, selection of hypothesis test criteria may compromise specification performance standards and the system user's interests during the test and evaluation process.

Alternatives to hypothesis testing are provided that use the specification performance standard itself as the pass/fail criterion for decision making and that directly indicate the risk associated with any resulting conclusion. The alternative approach includes a maximum likelihood estimate technique that compares two random variables.

2. CONTRACTOR'S VIEWPOINT

The contractor asserts or assumes that the system was built such that the population mean equals the requirement. Hence, the system was neither overdesigned nor underdesigned. The sample distribution of sample means $f(\bar{x})$ would then be as illustrated in Figure 1. In this approach, the contractor accepts a risk α that potentially maximizes the system user's risk. The contractor's risk is the probability of <u>rejecting</u> a system that <u>meets</u> the requirement, a Type I error. Typical values of α range from 0.10 through 0.01. If the contractor's assertion is accepted, then a one-sided hypothesis test based on this approach would accept the null hypothesis that $\mu_p > R$ for any $\mu_{\bar{x}} > R_\alpha$ and would reject the null hypothesis for any $\mu_{\bar{x}} < R_\alpha$.

Sample Distribution
of Sample Means



R = Requirement

N = Sample Size

μ_p = Population Mean

$\mu_{\bar{x}}$ = Mean Value of Sample Means ($\bar{x}$)

σ_p = Population Standard Deviation

$\sigma_{\bar{x}}$ = Standard Deviation of Sample Means
$= \sigma_p / \sqrt{N}$

α = Level of Significance (one-sided)

$$= \int_{-\infty}^{R_\alpha} f(\bar{x})\, d\bar{x} = F(R_\alpha)$$

R_α = A Sample Mean Determined by α

$f(\bar{x})$ = power density function in $\bar{x}$

Figure 1

3. SYSTEM USER'S VIEWPOINT

However, if the contractor's assertion is incorrect (i.e., $\mu_p < R$), and if the sample mean distribution based on actual data represents the actual population (i.e., $\mu_{\bar{x}} = \mu_p < R$ and $\sigma_{\bar{x}} = \sigma_p/\sqrt{N}$), then the user's risk is:

$$\int_{R_\alpha}^{\infty} f(\bar{x})\, d\bar{x} \quad \text{for} \quad R_\alpha < \mu_{\bar{x}} = \mu_p < R$$

For the conditions given, the user's risk is the probability of <u>accepting</u> a system that <u>does not meet</u> the requirement, a Type II error. The smaller the α, the smaller the R_α, and the greater the user's risk. The user's risk ranges between 0.5 and 1.0 minus α.

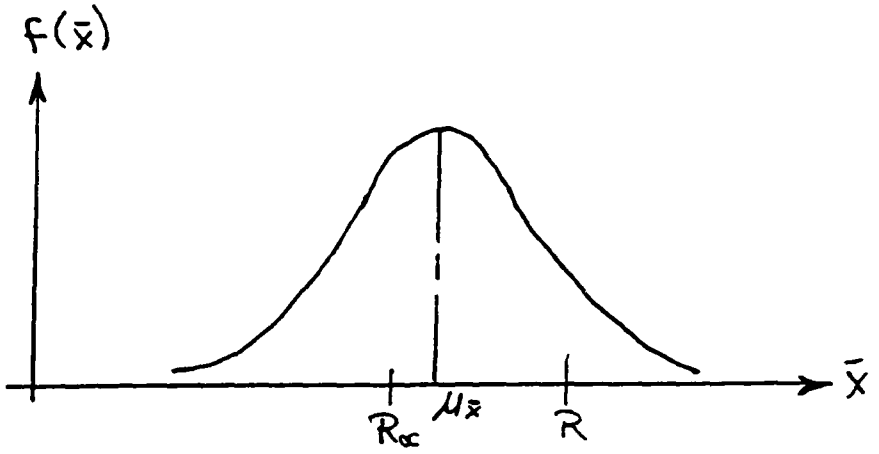


Figure 2

4. RECOMMENDED ANALYTICAL ALTERNATIVE FOR ASSESSING REQUIREMENT

System performance analysts must assess whether the specification requirement was met or not met. If the analyst assumes that the data are representative of the population (i.e., $\mu_{\bar{x}} = \mu_p$ and $\sigma_{\bar{x}} = \sigma_p/\sqrt{N}$), then from Figure 2 and for $\mu_{\bar{x}} < R$, the probability of $\bar{x} < R$ is:

$$\int_{-\infty}^{R} f(\bar{x})\, d\bar{x} = \text{Probability or confidence of not meeting the requirement} \qquad (2)$$

The user's risk is:

$$\int_{R}^{\infty} f(\bar{x})\, d\bar{x} = 1 - \int_{-\infty}^{R} f(\bar{x})\, d\bar{x} \qquad (3)$$

$$= \text{Probability of } \bar{x} \text{ meeting or exceeding the requirement}$$

This approach of using R itself as the basis for pass or fail decisions, and therefore as an integration limit, resembles the approach from the

contractor's viewpoint in Figure 1 for $\alpha = 0.5$. The underlying assumptions in each approach, however, are different. In the recommended approach, the user's risk will be between 0.5 (i.e., $\mu_{\bar{x}} = \mu_p$ just below R) and 0.0 (i.e., $\mu_{\bar{x}} = \mu_p << R$) when $\mu_{\bar{x}} < R$. When $\mu_{\bar{x}} = \mu_p > R$, the contractor's risk will be between 0.5 (i.e., $\mu_{\bar{x}} = \mu_p = R$) and 0.0 (i.e., $\mu_{\bar{x}} = \mu_p >> R$). See Figure 3.

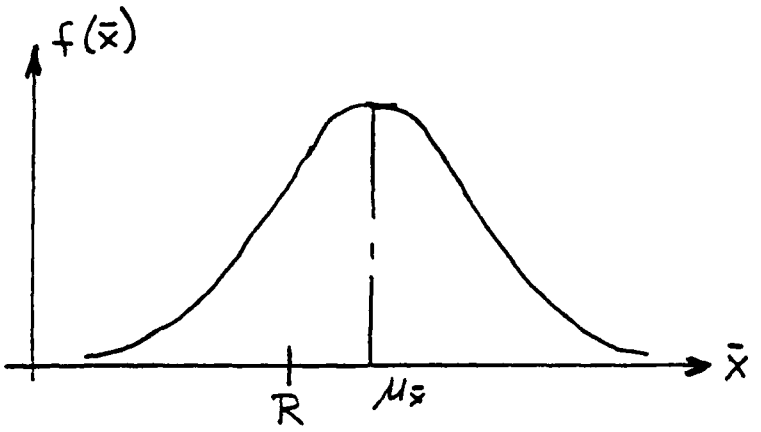


Figure 3

For $\mu_{\bar{x}} = \mu_p > R$ and $\sigma_{\bar{x}} = \sigma_p/\sqrt{N}$,

$$P\{\bar{x} > R\} = \int_R^{\infty} f(\bar{x})\, d\bar{x} \qquad (4)$$

$$= \text{Probability of meeting or exceeding the requirement}$$

$$\text{Contractor's Risk} = \int_{-\infty}^{R} f(\bar{x})\, d\bar{x} \qquad (5)$$

When pass/fail criteria are not completely specified in a specification, the hypothesis test approach to decision making usually puts the system user at a severe disadvantage with respect to risk, since α is usually chosen arbitrarily small. If the hypothesis test technique is applied using the contractor's assumption that $\mu_p = R$, then the corresponding user's risk [equation (1)] and the probability of not meeting the requirement [equation (2)] should be quantified as well when $\mu_{\bar{x}} < R$.

5. HYPOTHESIS TEST FOR COMPARING TWO SAMPLE MEANS

Two sets of data for two independent random variables are often compared to decide whether their sample means the same or whether there is a trend. Hypothesis testing is often used to make this decision. A sample distribution of sample means may also be compared with the contractor's assumed distribution of Figure 1 where the variances are equal.

Null Hypothesis: $\mu_{p_1} = \mu_{p_2}$

Alternate Hypothesis: $\mu_{p_1} \neq \mu_{p_2}$

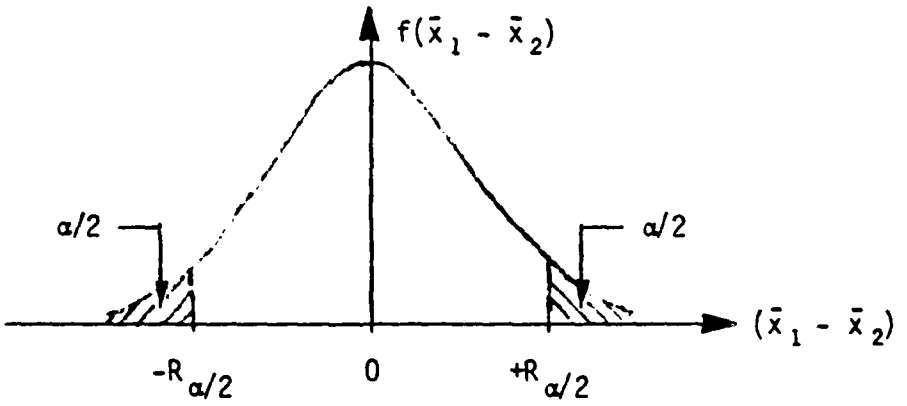


Figure 4

For $\mu_{\bar{x}_1-\bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$ and $\sigma^2_{\bar{x}_1-\bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_{p_1}^2}{N_1} + \frac{\sigma_{p_2}^2}{N_2}$,

if $|\mu_{\bar{x}_1} - \mu_{\bar{x}_2}| \leq |R_{\alpha/2}|$, then accept the null hypothesis; otherwise, reject it and accept the alternate hypothesis. As before, the choice of critical region size is entirely arbitrary, since it is not usually contained in a system specification. Hence, a more direct approach for comparing the two random variables is desired.

6. FIRST ALTERNATIVE FOR COMPARING TWO SAMPLE DISTRIBUTIONS OF SAMPLE MEANS

From Figure 5, a measure of the dissimilarity of $f(\bar{x}_1)$ and $f(\bar{x}_2)$, where $\mu_{\bar{x}_1} > \mu_{\bar{x}_2}$, is:

$$\int_0^{\infty} f(\bar{x}_1 - \bar{x}_2)\ d(\bar{x}_1 - \bar{x}_2) \tag{6}$$

The above equation provides the probability that $\bar{x}_1 > \bar{x}_2$. The risk in this conclusion is:

$$\int_{-\infty}^{0} f(\bar{x}_1 - \bar{x}_2)\ d(\bar{x}_1 - \bar{x}_2) = 1 - \int_0^{\infty} f(\bar{x}_1 - \bar{x}_2)\ d(\bar{x}_1 - \bar{x}_2) \tag{7}$$

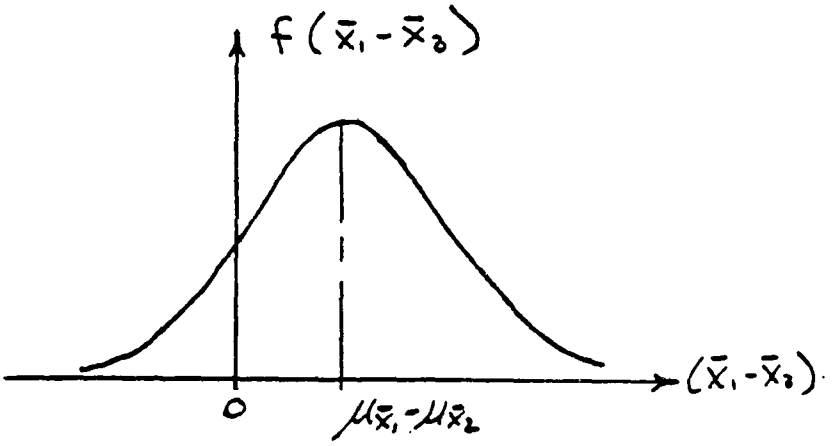


Figure 5

7. SECOND ALTERNATIVE FOR COMPARING TWO SAMPLE DISTRIBUTIONS OF SAMPLE MEANS USING A MAXIMUM LIKELIHOOD ESTIMATE TECHNIQUE.

Another method for comparing two random variables involves finding a real value c such that the joint probability that $\bar{x}_1 > c$ and $\bar{x}_2 < c$ is maximum.

The maximized joint probability then becomes a measure of the similarity or dissimilarity of $\bar{x}_1$ and $\bar{x}_2$. The maximized joint probability will typically range from 0.25 when $\mu_{\bar{x}_1} = \mu_{\bar{x}_2}$ through 1.0 when $\mu_{\bar{x}_1} >> \mu_{\bar{x}_2}$.

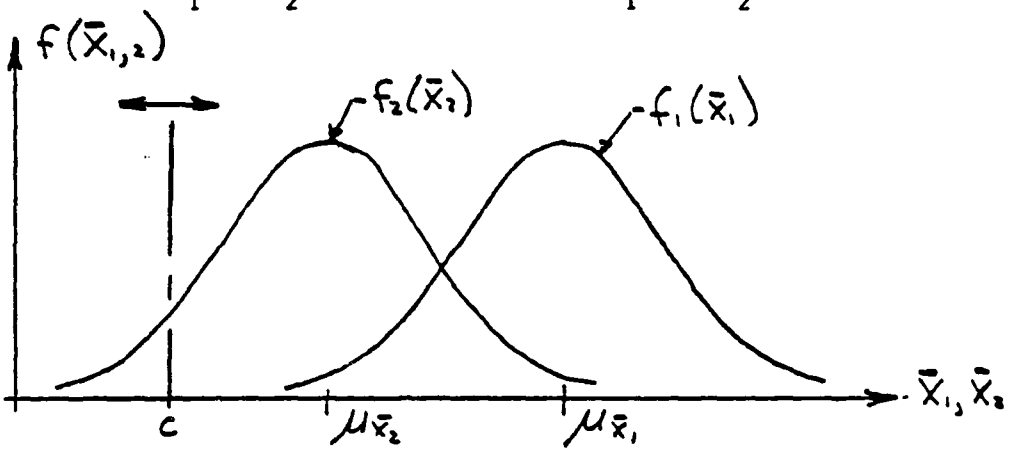


Figure 6

Find the value of c such that $P\{\bar{x}_1 > c, \bar{x}_2 < c\}$ is maximum:

$$P\{A\} = P(\bar{x}_1 > c) = \int_c^\infty f_1(\bar{x}_1)\, d\bar{x}_1 = 1 - F_1(c) \tag{8}$$

$$P\{B\} = P(\bar{x}_2 < c) = \int_{-\infty}^c f_2(\bar{x}_2)\, d\bar{x}_2 = F_2(c) \tag{9}$$

$$P\{AB\} = P\{A\} \cdot P\{B\} \tag{10}$$

(i.e., A and B are independent).

Let x be a binary random variable where:

$$p = P\{AB\}$$

$$q = 1 - p$$

$$= P\{\bar{A}B + A\bar{B} + \bar{A}\bar{B}\}$$

$$L = f\,[x;\ p(c)] = p^{x}\,(1 - p)^{1-x} \quad \text{where } x = 0,\ 1$$

$$\text{Ln } L = x\ \text{Ln } p + (1 - x)\ \text{Ln } (1 - p)$$

$$\frac{d(\text{Ln } L)}{dc} = x \cdot \frac{dp/dc}{p} - (1 - x)\,\frac{dp/dc}{(1 - p)} = 0$$

$$\frac{dp}{dc}\left[\frac{x}{p} - \frac{(1 - x)}{(1 - p)}\right] = 0$$

$$\therefore \frac{dp}{dc} = 0$$

$$p = P(\bar{x}_1 > c,\ \bar{x}_2 < c) = [1 - F_1(c)] \cdot [F_2(c)]$$

$$\frac{dp}{dc} = -\frac{dF_1(c)}{dc} \cdot F_2(c) + [1 - F_1(c)]\,\frac{dF_2(c)}{dc} = 0$$

$$\frac{f_2(c) \cdot \int_{c}^{\infty} f(\bar{x}_1)\ d\bar{x}_1}{f_1(c) \cdot \int_{-\infty}^{c} f_2(\bar{x}_2)\ d\bar{x}_2} = 1 \qquad (11)$$

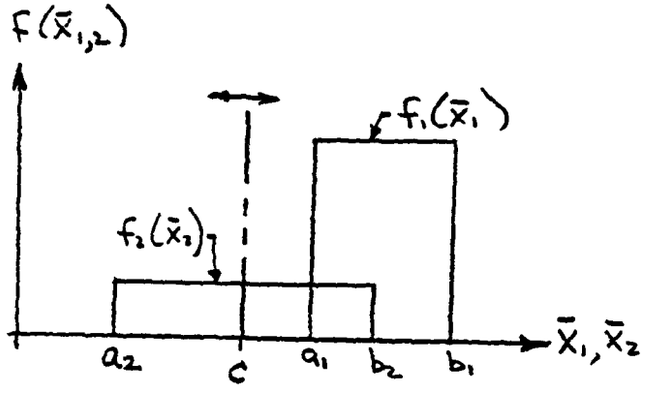
This equation is solved for c which maximizes p.

The second derivative of p with respect to c is found to determine whether a relative maximum or minimum is obtained.

$$\frac{d^2p}{dc^2} = \frac{df_2(c)}{dc} \cdot \int_{c}^{+\infty} f_1(\bar{x}_1)\ d\bar{x}_1 - 2\,f_1(c) \cdot f_2(c) - \frac{df_1(c)}{dc} \int_{-\infty}^{c} f_2(\bar{x}_2)\ d\bar{x}_2 \qquad (12)$$

c is substituted into this equation to confirm a maximum if $(d^2p/dc^2) < 0$, or a minimum if $(d^2p/dc^2) > 0$. Note that in equation (12) the integrals, $f_1(c)$ and $f_2(c)$ are always positive.

8. EXAMPLE 1 FOR COMPARING TWO SAMPLE DISTRIBUTIONS OF SAMPLE MEANS



$$f_1(\bar{x}_1) = \frac{1}{b_1 - a_1} = \frac{1}{B_1} \quad a_1 \leq x_1 \leq b_1$$

$$= 0 \text{ elsewhere}$$

$$f_2(\bar{x}_2) = \frac{1}{b_2 - a_2} = \frac{1}{B_2} \quad a_2 \leq \bar{x}_2 \leq b_2$$

$$= 0 \text{ elsewhere}$$

Figure 7

Find c:

$$f_2(c) \cdot \int_c^{b_2} f_1(\bar{x}_1)\, d\bar{x}_1 = f_1(c) \cdot \int_{a_1}^{c} f_2(\bar{x}_2)\, d\bar{x}_2$$

$$\frac{1}{B_2} \cdot \frac{1}{B_1}(b_2 - c) = \frac{1}{B_1} \cdot \frac{1}{B_2}(c - a_1)$$

$$c = \frac{b_2 + a_1}{2}$$

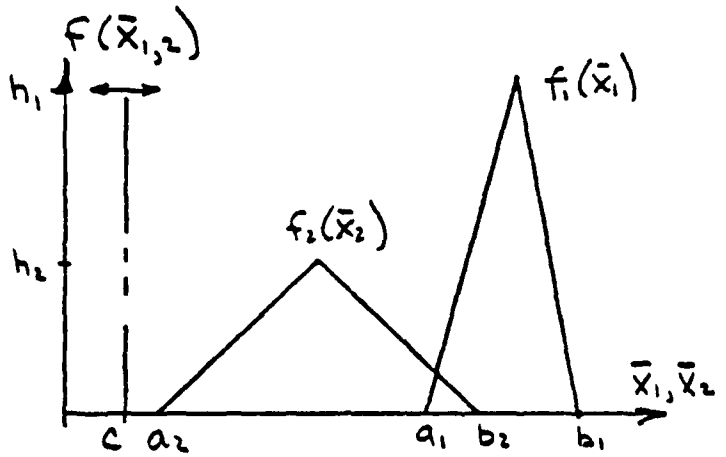
$$\frac{d^2p}{dc^2} = -\frac{2}{B_1 B_2} < 0$$

This value of c maximizes

$$P\{\bar{x}_1 > c,\ \bar{x}_2 < c\} = P\{\bar{x}_1 > c\} \cdot P\{\bar{x}_2 < c\}$$

$$= \frac{(b_1 - c)(c - a_2)}{(b_1 - a_1)(b_2 - a_2)}.$$

9. EXAMPLE 2 FOR COMPARING TWO SAMPLE DISTRIBUTIONS OF SAMPLE MEANS



$$h_1 = \frac{2}{B_1} = \frac{2}{b_1 - a_1}$$

$$h_2 = \frac{2}{B_2} = \frac{2}{b_2 - a_2}$$

$f_1(\bar{x}_1)$ and $f_2(\bar{x}_2)$ are isosceles.

Figure 8

For $h_1 = h_2$ and $a_2 < a_1 < b_2$, find c.

$$f_1(c) \cdot \int_{a_1}^{c} f_2(\bar{x}_2)\, d\bar{x}_2 = f_2(c) \cdot \int_{c}^{b_2} f_1(\bar{x}_1)\, d\bar{x}_1$$

$$c = \frac{a_1 + b_2}{2} = \frac{\mu_{\bar{x}_1} + \mu_{\bar{x}_2}}{2}$$

$(d^2p/dc^2) < 0$ from inspection of equation 12. This value of c maximizes

$$P\{\bar{x}_1 > c,\ \bar{x}_2 < c\} = P\{\bar{x}_1 > c\} \cdot P\{\bar{x}_2 < c\}$$

$$= \left[1 - \frac{f_1(c)\ (c - a_1)}{2}\right]^2 = \left[1 - \frac{h\ (c - a_1)^2}{2\ (\mu_{\bar{x}_1} - a_1)}\right]^2$$

$$= \left[1 - \frac{f_2(c)\ (b_2 - c)}{2}\right]^2 = \left[1 - \frac{h\ (b_2 - c)^2}{2\ (b_2 - \mu_{\bar{x}_2})}\right]^2$$

since $f_1(c) = f_2(c)$ and $(c - a_1) = (b_2 - c)$.

10. EXAMPLE 3 FOR COMPARING TWO SAMPLE DISTRIBUTIONS OF SAMPLE MEANS

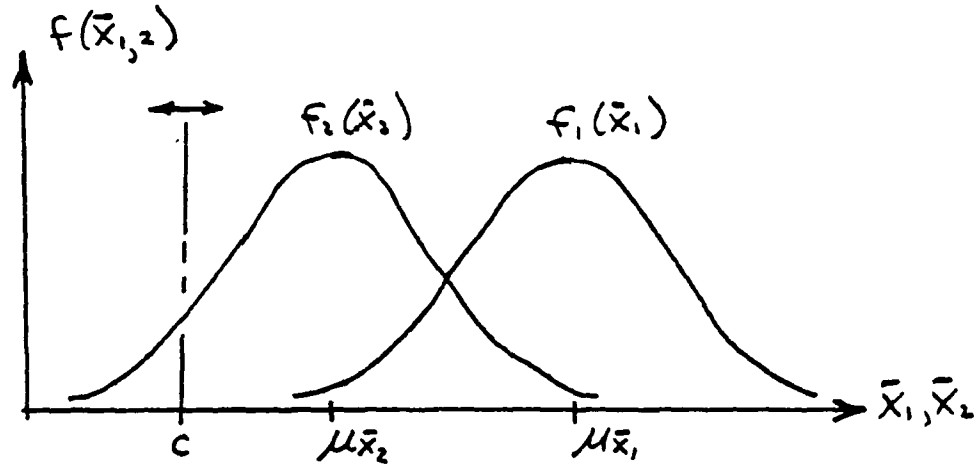


Figure 9

$f_2(\bar{x}_2) = G(\mu_{\bar{x}_2}, \sigma_{\bar{x}_2})$, a Gaussian power density function.

$f_1(\bar{x}_1) = G(\mu_{\bar{x}_1}, \sigma_{\bar{x}_1})$

For $\sigma_{\bar{x}_1} = \sigma_{\bar{x}_2}$, solve for c:

$$f_1(c) \cdot \int_{-\infty}^{c} f_2(\bar{x}_2)\, d\bar{x}_2 = f_2(c) \cdot \int_{c}^{+\infty} f_1(\bar{x}_1)\, d\bar{x}_1$$

$c = \dfrac{\mu_{\bar{x}_1} + \mu_{\bar{x}_2}}{2}$ maximizes $P\{\bar{x}_1 > c, \bar{x}_2 < c\} = P\{\bar{x}_1 > c\} \cdot P\{\bar{x}_2 < c\}$,

since $(d^2p/dc^2) < 0$ from inspection of equation 12.

$$p\{\bar{x}_1 > c, \bar{x}_2 < c\} = \int_{c}^{\infty} f_1(\bar{x}_1)\, d\bar{x}_1 \cdot \int_{-\infty}^{c} f_2(\bar{x}_2)\, d\bar{x}_2$$

For $\sigma_{\bar{x}_1} \neq \sigma_{\bar{x}_2}$, a trial-and-error method may be used to estimate c to the accuracy desired such that equation 11 is satisfied. Once c is known, the values of $P\{\bar{x}_1 > c\}$, $P\{\bar{x}_2 < c\}$, and $P\{\bar{x}_1 > c, \bar{x}_2 < c\}$ may be calculated.

11. SUMMARY

This paper reviews the limitations of hypothesis testing in that the system specification requirements may be compromised and the system user may be required to take an inordinate risk. It provides an analytical procedure that permits direct comparison of data with specification requirements, the probability associated with a conclusion, and the risk in making the conclusion where the risk is more equitably distributed between the contractor and the system user. This technique is developed further to include comparison of two distributions of sample means. The latter includes a maximum likelihood estimate of the joint probability that one random variable is greater than some value c and the second random variable less than c, and the risk associated with the conclusion.

12. ACKNOWLEDGEMENT

I wish to thank the Army Materiel Test and Evaluation Directorate, White Sands Missile Range, New Mexico, and the Hawk Project Office, Redstone Arsenal, Alabama, for the opportunity and encouragement to document this analysis and present it at this conference.

New Sequential and Parallel Methods for Unconstrained Optimization [1]

Robert B. Schnabel
Department of Computer Science
University of Colorado
Boulder, Colorado 80309

Extended Abstract. We give an overview of our recent research on several topics in unconstrained optimization. First we discuss methods for "orthogonal distance regression", data fitting when the measure of the distance from the data points to the fitted curve is the Euclidean distance rather than the standard vertical distance. We describe an efficient algorithm for this problem, and experience with its use. Second we summarize our research into "tensor methods" for nonlinear equations and optimization. These methods use higher order Taylor series information in a way that appears to significantly improve the performance of standard methods without significantly adding to their storage requirements or arithmetic cost per iteration. Finally we describe our research into parallel methods for optimization problems.

One of the most widely used methodologies in scientific and engineering research is the fitting of equations to data by least squares. In cases where significant observation errors exist in all data variables, however, the ordinary least squares approach, where all errors are attributed to the observation variable, is often inappropriate. An alternative approach, suggested by several researchers, involves minimizing the sum of squared orthogonal distances between each data point and the curve described by the model equation. We refer to this as orthogonal distance data fitting. We have developed a method for solving the orthogonal distance fitting problem that is a direct analog of the trust region Levenberg-Marquardt algorithm. The number of unknowns involved is the number of model parameters plus the number of data points, often a very large number. By exploiting sparcity, however, our algorithm has a computational effort per step which is of the same order as required for the Levenberg-Marquardt method for ordinary least squares. The description of this algorithm, an analysis of its mathematical properties, and the results of computational tests on some examples that illustrate some differences between the two approaches are given in Boggs, Byrd, and Schnabel [1987]. A software package that implements this approach is described in Boggs, Byrd, Donaldson, and Schnabel [1987], and is available from the authors.

Tensor methods are a new class of methods for solving systems of nonlinear equations and unconstrained optimization problems. Standard methods for nonlinear equations are related to Newton's method, and use a linear model of the nonlinear functions at each iteration. While they are effective on most problems, they are slow if the first derivative matrix at the root is singular or nearly singular. Tensor methods augment the standard linear model with a simple, low rank second order term, in a way that makes the method require no more function and derivative evaluations per iteration, and hardly more storage or arithmetic operations per iteration, than standard methods. In our tests, tensor methods are significantly more efficient than standard methods on both nonsingular and singular problems. This research is described in Schnabel and Frank [1984, 1987]. More recently we have developed tensor methods for unconstrained optimization. These methods augment the quadratic model, upon which standard optimization methods are based, with low rank third and fourth order terms. Again, the costs per iteration of the tensor method are hardly more than for the standard method, and the method requires substantially fewer total iterations and function evaluations in our tests. This research is described in

[1] Supported by the U.S. Army Research Office

Schnabel and Chow [1988].

Parallel optimization research at the University of Colorado has focused upon designing and implementing parallel algorithms for two optimization problems. One of these is the global optimization problem, which is to find the lowest minimizer of a nonlinear function of multiple variables that has multiple local minimizers. We have developed two types of parallel algorithms for this problem, both based on the stochastic approach of Rinnooy Kan and co-workers. The first is a rather straightforward parallelization of the sequential algorithm, while the second is a new, adaptive, dynamic method that is suggested by considerations of parallelism. Some of this research is described in Byrd, Dert, Rinnooy Kan, and Schnabel [1986]. The second optimization problem we have investigated is the standard local unconstrained optimization problem. We have studied new parallel optimization algorithms that use speculative function evaluations to evaluate part, but not all, of the Hessian matrix at each iteration. We have also analyzed alternatives for parallelizing the matrix updating calculations that constitute the main linear algebra cost of these methods. This research is described in Schnabel [1987] and Byrd, Schnabel, and Schultz [1988].

References

P. T. Boggs, R. H. Byrd, J. R. Donaldson, and R. B. Schnabel [1987], "ODRPACK - Software for weighted orthogonal distance regression," University of Colorado Technical Report CU-CS-360-87, to appear in *ACM Transactions on Mathematical Software.*

P. T. Boggs, R. H. Byrd, and R. B. Schnabel [1987], "A stable and efficient algorithm for nonlinear orthogonal distance regression," *SIAM Journal on Scientific and Statistical Computing*, 8, pp. 1052-1078.

R. H. Byrd, C. Dert, A. H. G. Rinnooy Kan, and R. B. Schnabel [1986], "Concurrent stochastic methods for global optimization", Technical Report CU-CS-338-86, Department of Computer Science, University of Colorado at Boulder, to appear in *Mathematical Programming.*

R. H. Byrd, R. B. Schnabel, and G. A. Shultz, "Parallel quasi-Newton methods for unconstrained optimization," Technical Report CU-CS-396-88, Department of Computer Science, University of Colorado at Boulder, to appear in *Mathematical Programming.*

R. B. Schnabel [1987], "Concurrent function evaluations in local and global optimization," *Computer Methods in Applied Mechanics and Engineering* 64, pp. 537-552.

R. B. Schnabel and T. Chow [1984], "Tensor methods for unconstrained optimization", in preparation.

R. B. Schnabel and P. Frank [1984], "Tensor methods for nonlinear equations", *SIAM Journal on Numerical Analysis* 21, pp. 815-843.

R. B. Schnabel and P. Frank [1987], "Solving systems of nonlinear equations by tensor methods", *The State of the Art in Numerical Analysis*, A. Iserles and M.J.D. Powell, eds., Clarendon Press, Oxford, pp. 245-271.

Two Generalized Fields and Their Governing Equations Applied to Helicopter Acoustics

A. Ünal [1]

C. Tung [2]

July 5, 1988

[1]Research scientist, US Army Aviation Research and Technology Activity, Moffett Field, CA 94035.

[2]Research scientist, US Army Aviation Research and Technology Activity, Moffett Field, CA 94035.

1 Abstract

In studying the refraction of the helicopter sound field from a shear layer, we face the problem of interaction of the sound field with another body (shear layer). In this interaction, we need the induced velocity in addition to the pressure since the boundary condition at the foreign body (shear layer) surface is with respect to the normal velocity. Therefore, a formula in terms of the sound pressure only is not sufficient. We need both pressure and velocity expressions so that we can invoke the interface conditions (continuity of the pressure and continuity of the normal velocity).

We are, therefore, motivated to find two equations in terms of two acoustic fields ; pressure fluctuation and velocity fluctuation.

In this paper, by defining two generalized functions , we develop an approach which yields two field equations. We suggest to use these two equations in any interaction problem of the helicopter sound field and in particular, in studying the refraction from a shear layer for all frequency ranges.

It is also found that the spectral methods seem to be more efficient in refraction problems.

2 Introduction

We present an alternate analytical description of the acoustic field of a moving body in a uniform flow.

Instead of using Ffowcs Williams- Hawkings [1] version of acoustic analogy, we formulate sources on a surface enclosing the moving body and its adjacent nonlinear flow field.

This approach avoids the laborious work of quadrupole terms and can be considered as a generalization of the Kirchhoff-Helmholtz theorem of acoustics.

In helicopter acoustics community, it has become a tradition to take FW-H extension of Lighthill's acoustic analogy concept [2] as the starting point. In this general formulation the acoustic field of a body, moving in a locally nonuniform, unsteady flow field, is expressed in terms of a monopole and a dipole source distribution over the body surface and a quadrupole source distribution over the volume containing the non-uniform, unsteady field in which the body moves.

Here the quadrupole source terms correspond to the nonlinearities in the flow equations.

At large distance the medium is at rest, apart from perturbations of acoustic order. Thus, in order to evaluate the quadrupole source terms, it is in principle necessary to know the complete flow field external to the body in advance.

3 An Alternate Formulation

Our alternate formulation is as follows :

The acoustic field is described in terms of the flow variables at the outer boundary of the volume containing the quadrupole distribution. In this sense

we can call it a generalisation of the Kirchoff-Helmholtz theorem [3]. Some of the salient features of our formulation are :

- a surface integral has to be evaluated instead of a volume integral.
- laborious calculation of the complicated quadrupole terms are avoided.
- we have expressions for both pressure and velocity fluctuations to be used in solving interaction problems of the sound field with other bodies.

We consider a uniform flow in the fluid since we are interested in uniform forward motions of the body. But for an arbitrary motion of the body, an arbitrary flow can be taken and the method allows this. In a uniform flow, we show that not only an acoustic field is generated at the boundary but also a hydrodynamic, vortical velocity field naturally emerges.

4 What is the relation to FW-H ?

The two methods do coincide when the induced velocity perturbations are small which in turn implies that quadrupole field is relatively weak and therefore can be neglected. The methods converge to each other since volume sources in FW-H vanish and in our formulation, the source surface shrinks to the actual body (blade) surface.

For conditions with a non-negligible quadrupole source field, it is straightforward to apply the present method with source surfaces ($S = 0$) at some distance from the blade surface provided that the aerodynamic field is given.

Contrary to this situation, the inclusion of the quadrupole source srength in a FW-H implies considerable analytical and numerical efforts.

5 Governing Equations

Let $S(\mathbf{x},t)=0$ describe a surface enclosing a body moving in a flow such that outside $S=0$, the field is, to leading order, governed by the linearised flow equations for the pressure and velocity fluctuations induced by the body $(p,\mathbf{v})$. We denote by $S<0$ the inside of the surface and by $S>0$ the outside of the surface.

We can make the linearised flow equations formally valid throughout the space by multiplying them with $H(S)$, the Heaviside function of S[4,5].

$$H(S)[\frac{D}{Dt}p+\nabla\circ\mathbf{v}]=0 \tag{1}$$

$$H(S)[\frac{D}{Dt}\mathbf{v}+\nabla p]=0 \tag{2}$$

where

$$\frac{D}{Dt}=\frac{\partial}{\partial t}+M\frac{\partial}{\partial x} \tag{3}$$

We nondimensionalize the equations by using a characteristic length, the mass density and the speed of sound in the fluid at infinity.

We have the following equations for the generalized pressure and generalized velocity outside $S=0$.

$$\frac{D}{Dt}[pH(S)]+\nabla\circ[\mathbf{v}H(S)]=Q\delta(S) \tag{4}$$

$$\frac{D}{Dt}[\mathbf{v}H(S)] + \nabla[pH(S)] = \mathbf{F}\delta(S) \tag{5}$$

where

$$Q = p\frac{DS}{Dt} + \mathbf{v} \circ \nabla S \tag{6}$$

$$\mathbf{F} = \mathbf{v}\frac{DS}{Dt} + p\nabla S \tag{7}$$

Elimination of $\mathbf{v}H(S)$ results in convected wave equation for the sound pressure field outside $S = 0$, driven by a surface source distribution at $S = 0$.

$$(\nabla^2 - \frac{D^2}{Dt^2})[pH(S)] = \nabla \circ [\mathbf{F}\delta(S)] - \frac{D}{Dt}[Q\delta(S)] \tag{8}$$

$$(\nabla^2 - \frac{D^2}{Dt^2} + \nabla \times \nabla\times)[\mathbf{v}H(S)] = \nabla Q\delta(S) - \frac{D}{Dt}\mathbf{F}\delta(S) \tag{9}$$

Green's function, which is defined as the acoustic field of an impulsive point source, has to satisfy

$$(\nabla^2 - \frac{D^2}{Dt^2})\mathcal{G} = -\delta(\mathbf{x} - \xi)\delta(t - \tau) \tag{10}$$

hence the generalized pressure fluctuation becomes

$$pH(S) = \int_\tau \int \int \int_\xi [\nabla_0 \mathcal{G} \circ \mathbf{F} - \frac{D\mathcal{G}}{D\tau}Q]\delta(S)d\xi d\tau \tag{11}$$

what we have here is the expression for the acoustic pressure of a source region in a uniform mean flow.

Note that we did not assume an irrotational velocity fluctuation field.

6 Green's Function Formalism

Green's functions are very instrumental in bringing in other boundaries in the fluid (such as wind tunnel walls, shear layers or other foreign bodies). There are two ways to introduce the Green's function: the traditional way and the less traditional spectral way. We think the spectral way is more advantageous as it will be shown later.

- The traditional Green's function :

$$\mathcal{G} = \delta \frac{(T-R)}{4\pi R} \tag{12}$$

where

$$T = \beta(t-\tau) + M\frac{(x-\xi)}{\beta} \tag{13}$$

$$\beta = (1-M^2)^{1/2} \tag{14}$$

and R is the distance between the source and the observer. This form of the Green's function is the most appealing as it clearly describes the spherical propagation of an acoustic pulse modified by mean flow convection. In the literature, the use of this form resulted in time-domain methods [6]. Farassat's method [5] is a modern version of time-domain method.

- The spectral Green's function : If we take a Fourier-Hankel transform, we obtain a frequency domain result. In the literature, the use of this form resulted in frequency methods [6].

$$\int_{\phi-\pi}^{\phi+\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} e^{-i(n\theta+\omega t+\alpha x)} r J_n(\gamma r)\, [\,]\, dr dx dt d\theta \tag{15}$$

In wave number space, our alternate form of Green's function becomes

$$\hat{G}(\alpha,\gamma,\omega \mid \xi,\rho,\phi,\tau) = e^{-i(n\phi+\omega\tau+\alpha\xi)}\frac{J_n(\gamma\rho)}{\gamma^2+\alpha^2-(\omega+M\alpha)^2} \qquad (16)$$

where ξ is the axial source coordinate, ρ is the radial source coordinate, τ is the source time, ϕ is the angular source coordinate.

The advantage of spectral Green's function is that the final solution of the acoustic field automatically yields an expansion in time-space harmonics.

It is also important to notice that since our approach includes the possibility of a non-uniform incident field (and since sound field is significantly affected by assymmetries in the flow) we should be able to describe the acoustic field well.

7 An Application : Refraction from a Shear Layer

To compute the interaction of the sound field with another body, we need the induced velocity in addition to the pressure since the boundary condition at the foreign body surface is w.r.t. the normal velocity.

We shall use spectral analysis to obtain an expression for the generalized velocity fluctuations.

$$\hat{\mathbf{v}} = \int\int_{-\infty}^{\infty} e^{-i(\omega t+\alpha x)}[\mathbf{v}H(S)]dxdt \qquad (17)$$

and similarly

$$\hat{p} = \int \int_{-\infty}^{\infty} e^{-i(\omega t + \alpha x)} [pH(S)] dx dt \tag{18}$$

$$\hat{\mathbf{F}} = \int \int_{-\infty}^{\infty} e^{-i(\omega t + \alpha x)} \mathbf{F} \delta(S) \tag{19}$$

Let us take the Fourier transform of the momentum equation in time and axial direction,

$$\frac{D}{Dt}[pH(S)] + \nabla \circ [\mathbf{v}H(S)] = Q\delta(S) \tag{20}$$

$$\frac{D}{Dt}[\mathbf{v}H(S)] + \nabla[pH(S)] = \mathbf{F}\delta(S) \tag{21}$$

$$\hat{\mathbf{v}_F} = \frac{1}{i(\omega + M\alpha)}[\hat{\mathbf{F}} - (i\alpha \mathbf{i}_x + \mathbf{i}_r \frac{\partial}{\partial r} + \mathbf{i}_\theta \frac{1}{r}\frac{\partial}{\partial \theta})\hat{p_F}] \tag{22}$$

$$\hat{\mathbf{v}_Q} = -\frac{1}{i(\omega + M\alpha)}(i\alpha \mathbf{i}_x + \mathbf{i}_r \frac{\partial}{\partial r} + \mathbf{i}_\theta \frac{\partial}{r\partial \theta})\hat{p_Q} \tag{23}$$

We should observe here that the velocity is exclusively in $(S = 0)$ surface quantities.

The velocity in physical space and time can be obtained by inverse transforming :

$$\mathbf{v}_F H(S) = \frac{1}{(2\pi)^2} \int \int_{-\infty}^{\infty} e^{i(\omega t + \alpha x)} \hat{\mathbf{v}_F}(\alpha, r, \theta, \omega) d\alpha d\omega \tag{24}$$

$$\mathbf{v}_Q H(S) = \frac{1}{(2\pi)^2} \int \int_{-\infty}^{\infty} e^{i(\omega t + \alpha x)} \hat{\mathbf{v}_Q}(\alpha, r, \theta, \omega) d\alpha d\omega \tag{25}$$

We can interpret this equation as a generalized theorem for the generalized velocity fluctuations.

Shear layer refraction problem is reduced to a boundary-value problem of the governing equations with additional boundary conditions coming

from the shear layer model. We have introduced four different shear layer model [Ünal, Tung, 7] the simplest one is a vortex sheet. The boundary conditions on the vortex sheet are :

- Continuity of the normal velocity
- Continuity of the pressure

We end our paper by stating that the shear layer refraction of the helicopter sound is a boundary-value problem defined on our two field equations.

We intend to solve this boundary-value problem for different configurations and conditions numerically.

8 References

[1] Ffowcs Williams, J.E. and Hawkings, D.L., Sound generation by turbulence and surfaces in arbitrary motion, Philosophical Transactions of the Royal Society of London, Vol. A264, 1969, pp. 321-342.

[2] Lighthill, M. J., "On the Sound Generated Aerodynamically, I. General Theory." Proceedings of the Royal Society (London) A211, 564-587, 1952.

[3] Pierce A., "Acoustics" , McGraw Hill Pub. Co. 1981.

[4] Kanwal, R. P., "Generalized Functions : Theory and Technique", Academic Press, New York, 1964.

[5] Farassat, F., "Linear Acoustic Formulas for Calculation of Rotating Blade Noise", AIAA Journal, vol. 19, No. 9, pp.1122-1130, 1981.

[6] Hanson, B. and Fink, M.R., "The Importance of Quadrupole Sources in Prediction of Transonic Tip Speed Propeller Noise", Jrnl. of Sound and Vibration, 62(1), 1979, 19-38.

[7] Ünal, A. and Tung, C.,"Inverse Source Modeling in Helicopter Acoustics", Sixth Army Con. on Applied Math. and Computing, 31 May- 3 June 1988, Univ. of Colorado, Boulder, Colorado.

INVERSE SOURCE MODELING IN HELICOPTER ACOUSTICS

A. Ünal[1] and C. Tung[2]

1 Abstract

In our efforts to compute the sound created by a moving body (helicopter), we face the evaluation of the near field either numerically or experimentally which are both difficult tasks and could easily lack precision.

To circumvent these difficulties we recast the governing equations into non-linear integral equations and define the helicopter source sound characterization as an inverse problem using the far field computations or measurements.

The logic of the approach lies in the fact that it is both easier to compute or to measure the far field.

Like any other inverse problem, the helicopter source characterization also faces the problem of multiple solutions.

We claim and later demonstrate that to the benefit of the helicopter research, the form of the kernel of the integral equation eliminates the problem of non-uniqueness.

Hence, we can use the inverse source modeling concept to obtain an equivalent source characterization using the far field data and then propagating the fields according to the governing equations for evaluation of the acoustic pressure fluctuations at an arbitrary observer.

[1]Dept. of Aero. and Astro., Stanford University, Stanford, California, 94305.
[2]Research scientist, US Army Aviation Research and Technology Activity, Moffett Field, CA 94035.

2 Mathematical Analysis

We have formulated the helicopter acoustics problem in terms of two generalized functions and two field equations in [1].

These equations are :

$$\left[\nabla^2 - \frac{D^2}{Dt^2}\right][\tilde{p}H(S)] = \nabla \circ [\tilde{\mathbf{v}}\frac{DS}{Dt} + \tilde{p}\nabla S]\delta(S) - \frac{D}{Dt}[\tilde{p}\frac{DS}{Dt} + \tilde{\mathbf{v}} \circ \nabla S]\delta(S) \quad (1)$$

$$\left[\nabla^2 - \frac{D^2}{Dt^2} + \nabla \times \nabla \times\right][\tilde{\mathbf{v}}H(S)] = \nabla[\tilde{p}\frac{DS}{Dt} + \tilde{\mathbf{v}} \circ \nabla S]\delta(S) - \frac{D}{Dt}[\tilde{\mathbf{v}}\frac{DS}{Dt} + \tilde{p}\nabla S]\delta(S) \quad (2)$$

Once we have the governing field equations in hand, we can proceed via two approaches :

- Approach I : Inject the numerically simulated pressure and velocity fields on a chosen surface (a computational surface or an airfoil) into the right-hand-side source terms and solve the partial differential equations numerically.
- Approach II : Use the far field data and the concept of inverse source modeling to replace the right-hand-side of the field equations by equivalent sources then propagate the fields to an arbitrary observer location through the equations.

The difference between these two approaches lies in whether we are using

near field data (for the first approach) or far field data (for the second approach).

Since it is always difficult to take near-field measurements precisely and it is always more costly to compute the near-field precisely compared to far field measurements or far field computations, there is a good reason to introduce the inverse source modeling notions in helicopter acoustics.

Our approach will consist of two parts :

- Part I : Solve an inverse problem.
- Part II : Use the inverse source modeling of Part I, solve the direct problem.

3 Statement of the Problem

What is an equivalent source for our helicopter in its arbitrary motion and what about its uniqueness?

4 Mathematical Analysis

Let us say that the sources are distributed within a region S of space with intensity $Q(\mathbf{r}, t)$ and let us say that we have the governing equa-

tion (1) for what we named a generalized acoustic pressure fluctuation

$$\left[\nabla^2 - \frac{D^2}{Dt^2}\right][\tilde{p}H(S)] = Q(\mathbf{r}, t) \tag{3}$$

which is an inhomogeneous convected scalar wave equation.

Let us say that we have an unbounded fluid medium with zero initial conditions, namely ;

$$\tilde{p}H(S) = \frac{\partial}{\partial t}[\tilde{p}H(S)] \tag{4}$$

$$= 0 \tag{5}$$

for $t < 0$.

Using the Green's function formalism and generalized function theory [2]., we can write down the solution as

$$\left[\nabla^2 - \frac{D^2}{Dt^2}\right] G(\mathbf{r}, t; \rho, t') = -\delta(\mathbf{r} - \rho)\delta(t - t') \tag{6}$$

thus

$$\tilde{p}H(S) = -\int dt' \int d\rho G(\mathbf{r}, t; \rho, t')Q(\rho, t') \tag{7}$$

Here the integrations are taken over the entire space-time domain. The causality property of G precludes integration over times t' that are later than t, while the spatial integration is extended only through the region S where $Q(\mathbf{r}, t)$ is nonzero.

Let us interpret the integral equation :

- If $Q(\rho, t')$ is known, then we solve the direct problem and determine $\tilde{p}H(S)$.

- If, however, $Q(\rho, t')$ is the unknown quantity, then the integral equation is a linear integral equation for $Q(\rho, t')$ with kernel G.

For any finite source distribution Q, the wave function $\tilde{p}H(S)$ must fall off sufficiently rapidly as r approaches infinity.

$$\underbrace{\tilde{p}H(S)}_{1} = -\int dt' \int d\rho G(\mathbf{r}, t; \rho, t') \underbrace{Q(\rho, t')}_{2} \tag{8}$$

(1) is a known quantity and (2) is an unknown quantity.

The integral equation is a Fredholm equation of the first kind. Usually the solution of this type of equation is not a simple matter but in this case the kernel G is a Green's function. For such kernels, it is well known that a solution can be obtained if the Green's function can be expanded in a series of orthonormal functions.

However, there remains a problem of nonuniqueness which we shall discuss thoroughly under the section of nonuniqueness.

- We can consider the integral equation as a description of $\tilde{p}H(S)$: $\tilde{p}H(S)$ is the integral transform of $Q(\rho, t')$ with kernel G. Then the inverse transform will yield Q in terms of $\tilde{p}H(S)$. This will work in cases where G has the form of known integral transforms.

  Let us pose the problem as follows : If we know the far field values of $\tilde{p}H(S)$ at a finite number of locations, can we solve for Q?

  Answer : Yes, if we discretize the integrations as well as truncate

any resulting infinite sums, and if we overcome the nonuniqueness.

From a physical point of view, it is clear that in order to obtain Q from the inverse transformation, it is necessary to know $\tilde{p}H(S)$ over the entire domain of space and time. This seems to be an impossible problem at first sight but as we shall see it is possible to attack this problem as follows :

1. To invert the integral equation, solving for Q, we must satisfy the Green's function G. The form of G depends on the properties of the unbounded medium that supports the acoustic waves propagating away from the source region $\mathcal{S}$. For scalar waves, the $\tilde{p}H(S)$ profile appearing in the governing equation varies in time and space, in which case, it is generally difficult to solve explicitly for G. But if the velocity of sound (c) is constant, then the Green's function is well-known:

$$G(\mathbf{r}, t; \rho, t') = \frac{1}{4\pi} \frac{\delta[t - t' - \frac{|\mathbf{r}-\rho|}{c}]}{|\mathbf{r} - \rho|} \tag{9}$$

We can change the convected wave operator into Helmholtz operator by using a transformation of moving coordinates. Let us then follow the analysis on this transformed equation.

The Helmholtz equation is given by :

$$\left[\nabla^2 + k^2\right] \tilde{p}H(\mathbf{r}, w) = Q(\mathbf{r}, w) \tag{10}$$

which is a Fourier transform of the original equation with respect to time. Here, $\mathbf{k} = \frac{\mathbf{w}}{\mathbf{c}}$. The Green's function for this case is

$$G(\mathbf{r}, \rho) = (\frac{1}{4\pi}) \frac{e^{ik|\mathbf{r}-\mathbf{r}_0|}}{|\mathbf{r} - \rho|} \tag{11}$$

The equivalent integral equation is then

$$\tilde{p}H(S)(\mathbf{r}, w) = -\frac{1}{4\pi} \int d\rho \frac{e^{ik|\mathbf{r}-\rho|}}{|\mathbf{r} - \rho|} Q(\rho, w) \tag{12}$$

The most crucial observation here is that the relation between $Q(\rho, w)$ and $\tilde{p}H(S)(\mathbf{r}, w)$ is linear. This, in turn means that we can use various approximate methods to solve for Q in terms of $\tilde{p}H(S)(\mathbf{r}, w)$. The position vector r is entirely arbitrary. Hence, if $\tilde{p}H(S)$ is known or measured at a sufficiently large number of arbitrary locations, an attempt can be made to solve for Q. In practice only two possibilities are of interest :

(a) when $\tilde{p}H(S)$ is measured in the proximity of the helicopter,

(b) when $\tilde{p}H(S)$ is measured far away from the helicopter, the far field case.

We shall concentrate on the second case. Far field assumption implies immediately that

$$r = |\mathbf{r}| \gg |\rho_{max}| \tag{13}$$

in the support of Q i.e. in the region where Q is nonsero ; namely

$$r \gg \frac{\rho_{max}^2}{\lambda} \tag{14}$$

If we use the inequality to expand the kernel,

$$[\tilde{p}H(S)]_{FF}(\mathbf{r}, \mathbf{k}) = -\frac{e^{ikr}}{4\pi r}\int d\rho e^{-ik\frac{\mathbf{r}}{r}} Q(\rho, \mathbf{k}) \tag{15}$$

but the integrand is nothing other than the Fourier Transform of the source distribution. The point in the transformed space being given by the values of the vector

$$\mathbf{k} = k\frac{\mathbf{r}}{r} \tag{16}$$

With the foregoing interpretation, it would seem a simple matter of taking the inverse Fourier transform and thus solving explicitly for $Q(\mathbf{r})$. But actually, the situation is somewhat more complicated since the wave vector $k\frac{\mathbf{r}}{r}$, which of necessity becomes the integration variable in the three-dimensional integral defining the inverse Fourier transform, is limited by the condition

$$|\mathbf{k}| = |k\frac{\mathbf{r}}{r}| \tag{17}$$

$$= k \tag{18}$$

$$= \frac{w}{c} \tag{19}$$

Let us denote the Fourier transform of Q by $\hat{Q}(\mathbf{K}, \mathbf{k})$, then

$$\hat{Q}(\mathbf{K}, \mathbf{k}) = \int d\rho e^{i\mathbf{K}\circ\rho} Q(\rho, \mathbf{k}) \tag{20}$$

The restriction on $(\mathbf{k}\frac{\mathbf{r}}{\mathbf{r}})$ vector can be accounted for by defining a generalized source (which is also an effective source) that is given in Fourier space by

$$\hat{Q}_e = \hat{Q}(\mathbf{K}, \mathbf{k})\delta(K - k) \tag{21}$$

Now, $\hat{Q}_e$ is defined for all possible K. This is indeed the definition of our third generalized function which in turn defines the generalized effective source. Note that it might be mathematically more proper if we call

$$\hat{Q}_e = \hat{Q}_g \tag{22}$$

the Fourier transform of the generalized source function or the generalized equivalent source function in the K space. This multiplication in K space will result in a convolution in real space. Note also that the physical dimensions of the generalized effective source is different from those of Q because the delta function $\delta(\mathbf{K} - \mathbf{k})$ has dimensions of length.

Inverse Fourier transform yields ;

$$\hat{Q}_e(\mathbf{K},\mathbf{k}) = \hat{Q}(\mathbf{K}-\mathbf{k}) \tag{23}$$

$$= \left[\int d\rho e^{i\mathbf{K}\circ\rho} Q(\rho,\mathbf{k})\right] \delta(\mathbf{K}-\mathbf{k}) \tag{24}$$

$$\hat{Q}_e(\mathbf{K}-\mathbf{k}) = \int d\rho e^{i\mathbf{K}\circ\rho} Q(\rho,\mathbf{k})\delta(\mathbf{K}-\mathbf{k}) \tag{25}$$

$$Q_e(\mathbf{r},\mathbf{k}) = \frac{1}{(2\pi)^3}\int d\mathbf{K} e^{\mathbf{K}\circ\mathbf{r}} \hat{Q}(\mathbf{K},\mathbf{k})\delta(\mathbf{K}-\mathbf{k}) \tag{26}$$

$$= \frac{\mathbf{k}}{2\pi^2}\int d\rho \mathbf{Q}(\rho,k)\frac{sin\mathbf{k}\mid \mathbf{r}-\rho\mid}{\mid \mathbf{r}-\rho\mid} \tag{27}$$

$$\mathbf{k} = \mathbf{F}(k,\theta,\phi) \tag{28}$$

$$Q_e(\mathbf{r},\mathbf{k}) = -k^2\mathcal{A}\frac{e^{-ik\mathcal{A}}}{2\pi^2}$$

$$\int_0^{2\pi} d\phi \int_0^{\pi} d\theta sin\theta[\tilde{p}H(S)]_{FF}(\mathcal{A},\theta,\phi)e^{\mathbf{k}\circ\mathbf{r}} \tag{29}$$

The right-hand-side integrals represent the angular Fourier transform of the values measured.

$$\mathcal{A} = cT \tag{30}$$

$\mathcal{A}$ is the point in the far field where measurements are taken. c is the speed of propagation of sound waves and T is the travel time of the wave from the actual source to the point.

$$\mathbf{k}\circ\mathbf{r} = kr cos\theta \tag{31}$$

where

$$\mathbf{r} = r[sin\theta cos\phi, sin\theta sin\phi, cos\theta] \tag{32}$$

Equation (29) can be considered as an algorithm that yields the effective source distribution at any point r in terms of the measured values of the pressure in the far field.

5 Equivalent Source Distribution at a Point

Let

$$Q(\mathbf{r}) = pf_0\delta(\mathbf{r}) \tag{33}$$

Delta function will be imaged as distributions that are smeared out in space. The extent of smearing depends on the resolution of the imaging process.

Let us write the expansion for the point equivalent source in terms of spherical harmonics as :

$$f = f_0 P_0 cos\theta \tag{34}$$

P_0 is zero-order Legendre polynomial.

$$P_0 = 1 \tag{35}$$

$$f = f_0 cos\theta \tag{36}$$

f_0 is the value of the form function.

$$Qe^N(\mathbf{r}) = 2\sum_{n=0}^{\infty} i^n f_n(k) j_n(kr) P_n cos\theta \tag{37}$$

$$Qe^N = 2f_0 j_0(kr) \tag{38}$$

where j_0 is the zero-order spherical Bessel function which is the sinc function. Its first zero occurs at $kr_0 = \pi$ or in terms of wave lengths of

$$r_0 = \frac{\lambda}{2} \tag{39}$$

6 Non-Uniqueness

It is now a well-established fact that the Fredholm integral equations of the first kind do not always possess unique solutions. Solutions depend on arbitrary constants, which must ultimately be determined from criteria not given in the original problem.

For certain kernels, such as Fourier, Laplace, Mellin, and others, it is well known that the appropriate integral equations have unique solutions with only mild restrictions imposed on the allowed class of functions.

Effective source distribution in Fourier space is

$$\hat{Q}_e(\mathbf{K}, \mathbf{k}) = \hat{Q}(\mathbf{K}, \mathbf{k}) \triangle (\mathbf{K}, \mathbf{k}) \tag{40}$$

$$\triangle (\mathbf{K}, \mathbf{k}) = \delta(\mid \mathbf{K} - \mathbf{k}) \tag{41}$$

although $\bar{Q}(\mathbf{K}, \mathbf{k})$ was known only on the spherical shell $\mathbf{K} = \mathbf{k}$, the choice of $\triangle$, albeit arbitrary, extends the definition

of the function $\tilde{Q}$ to all points of K space. This extension modifies the integral equation into one with an unrestricted Fourier kernel and thus into one with a unique solution.

In other words, the introduction of the generalized function (also named filter function) in the integral equation is a mathematical device that removes the "defect" in the kernel. In the present case, this defect is the restriction of the kernel to the shell $\mathbf{K} = \mathbf{k}$ which is responsible of non-uniqueness. The price paid for regaining uniqueness is that instead of the original quantity of interest, a related quantity is obtained. The latter may be viewed a filter version of the original quantity in the physical space.

$$\int d\rho e^{\mathbf{K}\circ\rho} Q^0(\rho) \tag{42}$$

for $|\mathbf{K}| = \mathbf{k}$. Choose a $Q^0(\rho)$ such that its Fourier transform vanishes for $|\mathbf{K}| = \mathbf{k}$.

$$Q^0(\mathbf{K}) = (\mathbf{K} - \mathbf{k})e^{-\alpha\mathbf{K}} \tag{43}$$

$$= (\mathbf{K} - \mathbf{k})^2 e^{-\alpha\mathbf{K}} \tag{44}$$

These are spherically symmetric distributions in K space which, in turn, yield spherically symmetric distributions in physical space given by

$$Q^0(\mathbf{r}) = \frac{1}{2\pi^2}$$

$$(\frac{\delta}{\delta\alpha}+\mathbf{k})\frac{\delta}{\delta\alpha}\frac{1}{\alpha^2+r^2}-\frac{1}{2\pi^2}(\frac{\delta}{\delta\alpha}+\mathbf{k})^2\frac{\delta}{\delta\alpha}\frac{1}{\alpha^2+r^2} \tag{45}$$

7 An Example

If for example we know a priori that the source is impulsive i.e.

$$F(\mathbf{r},t)=F_0(\mathbf{r})\delta(t) \tag{46}$$

and we wish to determine F_0, we can proceed as follows ;

$$f(\mathbf{r},w)=\int_{-\infty}^{\infty}e^{iwt}F_0(\mathbf{r})\delta(t)dt \tag{47}$$

$$f(\mathbf{r},w)=F_0(\mathbf{r}) \tag{48}$$

$$\tilde{F}_0(\mathbf{k})=u_0(\hat{k},ck) \tag{49}$$

$$F_0(\mathbf{r})=\frac{1}{(2\pi)^3}\int\int\int e^{i\mathbf{k}_0\mathbf{r}}u_0(\hat{k},ck)d^3k \tag{50}$$

If the actual source were

$$F(\mathbf{r},t)=\delta(\mathbf{r}-\mathbf{r}_0)\delta(t) \tag{51}$$

$$u(\mathbf{r},w)=g(\mid\mathbf{r}-\mathbf{r}_0\mid,w) \tag{52}$$

$$\simeq\frac{1}{4\pi r}e^{iw\cdot\frac{r}{c}-iw\frac{\mathbf{r}_0\mathbf{r}_0}{c}} \tag{53}$$

Thus ideally, on δD, we would observe,

$$u_0(\hat{\mathbf{k}},ck)=e^{-i\mathbf{k}_0\mathbf{r}_0} \tag{54}$$

inserting back,

$$F_0(\mathbf{r}) = \frac{1}{(2\pi)^3} \int\int\int e^{i\mathbf{k}_0(\mathbf{r}-\mathbf{r}_0)} d^3\mathbf{k} \tag{55}$$

$$= \delta(\mathbf{r} - \mathbf{r}_0) \tag{56}$$

which is indeed the correct source.

8 References

[1] Ünal, Aynur and Tung, Chee, "Two Generalized Fields and Their Governing Equations Applied to Helicopter Acoustics" Sixth Army Conference on Applied Mathematics and Computing, 31 May - 3 June 1988, University of Colorado, Boulder. Colorado.

[2] Kanwal, R. P., "Generalized Functions : Theory and Technique, New York, Academic Press, 1983.

Diagonal Implicit Multigrid Solution of the Three-Dimensional Euler Equations

Yoram Yadlin and D. A. Caughey

Sibley School of Mechanical and Aerospace Engineering
Cornell University
Ithaca, New York 14853

I. Abstract

A Multigrid Alternating Direction Implicit Scheme has been developed to solve the Euler equations of inviscid, compressible flow in three dimensions. The scheme is an extension of the two-dimensional scheme developed by Caughey [1] to treat three-dimensional problems. The multigrid method is an efficient technique for accelerating the convergence of iterative methods; the Alternating Direction Implicit scheme holds the promise of rapid convergence characteristics, especially on the highly-stretched meshes required to resolve the thin shear layers appearing in high Reynolds number flows; and the diagonalization procedure results in a computationally-efficient implementation of the ADI scheme. The scheme is applied to compute the transonic flow past a swept wing, and results are presented to confirm the accuracy of the method and illustrate the efficiency of the iterative algorithm.

II. Analysis

The Euler equations of inviscid, compressible flow can be written in three space dimensions as

$$\partial\vec{w}/\partial t + \partial\vec{f}_1/\partial x + \partial\vec{f}_2/\partial y + \partial\vec{f}_3/\partial z = 0 \tag{1}$$

where

$$\vec{w} = \{\rho, \rho u_1, \rho u_2, \rho u_3, e\}^T \tag{2a}$$

is the vector of conserved dependent variables, and

$$\vec{f}_1 = \{\rho u_1, \rho u_1^2 + p, \rho u_1 u_2, \rho u_1 u_3, u_1(e+p)\}^T \tag{2b}$$

$$\vec{f}_2 = \{\rho u_2, \rho u_2 u_1, \rho u_2^2 + p, \rho u_2 u_3, u_2(e+p)\}^T \tag{2c}$$

$$\vec{f}_3 = \{\rho u_3, \rho u_3 u_1, \rho u_3 u_2, \rho u_3^2 + p, u_3(e+p)\}^T \tag{2d}$$

are the flux vectors in the x, y, and z coordinates respectively. The pressure p is related to the total energy e by the equation of state

$$e = p/(\gamma - 1) + \rho(u_1^2 + u_2^2 + u_3^2)/2 - \rho H_0, \tag{3}$$

where γ is the ratio of specific heats and H_0 is the total enthalpy.

In order to allow for the treatment of arbitrary geometries, the algorithm is implemented within the framework of a finite volume approximation [2]. The equations under an arbitrary transformation of independent variables to a new curvilinear coordinate system can be written as

$$\partial\vec{W}/\partial t + \partial\vec{F}_1/\partial\xi + \partial\vec{F}_2/\partial\eta + \partial\vec{F}_3/\partial\zeta = 0, \tag{4}$$

where $\vec{W} = h\vec{w}$ is the vector of transformed dependent variables, and

$$\vec{F}_1 = h\{\rho U_1, \rho u_1 U_1 + \xi_x p, \rho u_2 U_1 + \xi_y p, \rho u_3 U_1 + \xi_z p, U_1(e+p)\}^T \tag{5a}$$

$$\vec{F}_2 = h\{\rho U_2, \rho u_1 U_2 + \eta_x p, \rho u_2 U_2 + \eta_y p, \rho u_3 U_2 + \eta_z p, U_2(e+p)\}^T \tag{5b}$$

$$\vec{F}_3 = h\{\rho U_3, \rho u_1 U_3 + \zeta_x p, \rho u_2 U_3 + \zeta_y p, \rho u_3 U_3 + \zeta_z p, U_3(e+p)\}^T \tag{5c}$$

are the transformed flux vectors. Here, h is the determinant of the Jacobian of the transformation (which corresponds to the cell volume), and U_1, U_2, and U_3 are the contravariant components of the velocity given by

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} = \begin{pmatrix} \xi_x & \xi_y & \xi_z \\ \eta_x & \eta_y & \eta_z \\ \zeta_x & \zeta_y & \zeta_z \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}. \tag{6}$$

Diagonal Implicit Multigrid Solution

Artificial dissipation is added as a blend of second and fourth differences of the solution. The fourth differences are necessary to insure convergence to a steady state, while the second differences are necessary to prevent excessive oscillations of the solution in the vicinity of shock waves. Following the two-dimensional implementation of Caughey [1], these dissipation terms are scaled to stabilize the one-dimensional problems in each coordinate direction.

Following Briley and McDonald [3], and Beam and Warming [4], the time linearized implicit operator is approximated as the product of three one-dimensional factors, resulting in a scheme of the form

$$\begin{aligned}
&\{I+\theta\Delta t[A^n_{1i,j,k}\delta_\xi-\epsilon^{(2)}_{i,j,k}\delta^2_\xi(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\xi(1/h)]\}\\
&\{I+\theta\Delta t[A^n_{2i,j,k}\delta_\eta-\epsilon^{(2)}_{i,j,k}\delta^2_\eta(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\eta(1/h)]\}\\
&\{I+\theta\Delta t[A^n_{3i,j,k}\delta_\zeta-\epsilon^{(2)}_{i,j,k}\delta^2_\zeta(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\zeta(1/h)]\}\Delta\vec{W}^n_{i,j,k}=\\
&\quad-\Delta t\{\delta_\xi\vec{F}_{1i,j,k}+\delta_\eta\vec{F}_{2i,j,k}+\delta_\zeta\vec{F}_{3i,j,k}\\
&\quad-\epsilon^{(2)}(\delta^2_\xi+\delta^2_\eta+\delta^2_\zeta)\vec{w}_{i,j,k}+\epsilon^{(4)}(\delta^4_\xi+\delta^4_\eta+\delta^4_\zeta)\vec{w}_{i,j,k}\}^n, \qquad (7)
\end{aligned}$$

where $A^n_{l,i,j,k}=\{\partial\vec{F}_l/\partial\vec{W}\}^n_{l,i,j,k}$ are the Jacobians of the transformed flux vectors with respect to the solution and $\epsilon^{(2)}$and $\epsilon^{(4)}$ are the dissipation coefficients. For computational efficiency, each factor in Eq. (7) is diagonalized by a local similarity transformation, yielding a decoupled set of equations which can be solved using a scalar pentadiagonal solver. The diagonalization is performed using the modal matrices of the Jacobians A_l ($l=1,2,3$). Thus, if Q_l is the modal matrix of A_l, then $Q_l^{-1}A_lQ_l=\Lambda_l$ is a diagonal matrix whose non-zero elements are the eigenvalues of A_l. Applying this transformation at each mesh point, the resulting equations are:

$$\begin{aligned}
&\{I+\theta\Delta t[\Lambda^n_{1i,j,k}\delta_\xi-\epsilon^{(2)}_{i,j,k}\delta^2_\xi(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\xi(1/h)]\}{Q^n_{1,i,j,k}}^{-1}\\
&Q^n_{2i,j,k}\{I+\theta\Delta t[\Lambda^n_{2i,j,k}\delta_\eta-\epsilon^{(2)}_{i,j,k}\delta^2_\eta(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\eta(1/h)]\}{Q^n_{2,i,j,k}}^{-1}\\
&Q^n_{3i,j,k}\{I+\theta\Delta t[\Lambda^n_{3i,j,k}\delta_\zeta-\epsilon^{(2)}_{i,j,k}\delta^2_\zeta(1/h)+\epsilon^{(4)}_{i,j,k}\delta^4_\zeta(1/h)]\}\Delta\vec{V}^n_{i,j,k}=\\
&\quad-\Delta tQ^n_{1,i,j,k}{}^{-1}\{\delta_\xi\vec{F}_{1i,j,k}+\delta_\eta\vec{F}_{2i,j,k}+\delta_\zeta\vec{F}_{3i,j,k}\\
&\quad-\epsilon^{(2)}(\delta^2_\xi+\delta^2_\eta+\delta^2_\zeta)\vec{w}_{i,j,k}+\epsilon^{(4)}(\delta^4_\xi+\delta^4_\eta+\delta^4_\zeta)\vec{w}_{i,j,k}\}^n, \qquad (8)
\end{aligned}$$

where $\Delta\vec{V}^n_{i,j,k}={Q^n_{3,i,j,k}}^{-1}\Delta\vec{W}^n_{i,j,k}$. The elements of Q_l, Q_l^{-1} and Λ_l can be expressed explicitly in terms of $\vec{w}$ and the elements of the Jacobian matrix of the coordinate transformations, and are given by Chaussee and Pulliam [5]. The solution of Eqs. (8) is performed sequentially by solving five scalar-pentadiagonal systems along each line in each of the three mesh directions for each time step. The scheme is incorporated within the multigrid algorithm following the procedure developed by Jameson [6].

The treatment of the explicit boundary conditions in the far field is based on the Riemann invariants of the one-dimensional problem normal to the boundary. On the wing surface and on the symmetry plane, the pressure is interpolated from the interior of the field using the normal momentum equation. The implicit boundary conditions are treated in a manner consistent with the characteristic theory; this is relatively easy to implement since the corrections (or intermediate corrections) determined in each step are approximations to the changes in the characteristic variables in the coordinate direction along which the equations are being solved [1].

III. Results

The algorithm described above has been applied to the problem of transonic flow past a swept wing mounted on a vertical wall, or symmetry plane. The results presented here have been calculated for the ONERA wing M-6 [7] on a C-grid containing $192 \times 32 \times 32$ mesh cells, in the wraparound, normal and spanwise directions, respectively. The grid has been generated by a weak shearing of a square root transformation about a point just inside the leading edge of the wing surface in each plane of constant z.

Figures 1(a) and 1(b) illustrate the general nature of the grid system. Figure 1(a) presents the distribution of cells on the wing planform ($x - z$ plane), while Figure 1(b) presents a perspective view of the wing mounted on the wall; the C-grid shown in the symmetry plane is typical of the mesh in each $x - y$ plane. The far field boundaries are located approximately 10 chords upstream and downstream of the wing, approximately 19 chords laterally from the wing, and at approximately 3.5 semispans from the plane of symmetry in the z direction.

Results have been calculated for a free stream Mach number of 0.839 and 3.06 degrees angle of attack in order to allow comparison with existing wind-tunnel test data [7]. Figure 2 presents the streamwise pressure distribution at each of the computational stations on the upper and lower surfaces of the wing, and Figures 3(a) and 3(b) present contours of constant p/p_∞ on the upper and lower surfaces of the wing. The development of the "Lambda" shock pattern on the wing upper surface, characteristic of supercritical flows past swept wings, is clearly visible. Figures 4(a) and 4(b) present a comparison with wind-tunnel data [7] at two spanwise stations. The calculated results predict quite accurately the strengths and the locations of the shocks, in spite of the neglect of viscous effects in the calculation.

Figure 5 presents contours maps of entropy in several $y - z$ planes, starting at the leading edge and moving downstream. Since the entropy is constant along streamlines for a steady, inviscid flow, these plots can be viewed as representing cuts through stream surfaces, and reveal the generation and evolution of the wingtip vortex. It is clear that the vortex center moves up and inboard as it develops downstream, as is observed in experiment.

Diagonal Implicit Multigrid Solution

Figure 6 presents convergence histories for the Diagonal ADI scheme on a single grid, and when 5 levels of multigrid are used. The logarithm of the average residual, the drag coefficient, and the number of cells in which the local Mach number is supersonic are plotted as a function of computational work, measured in *Work Units*. One *Work Unit* corresponds to the computational labor required for a single time step on the fine grid. For both calculations, local time stepping is used at a Courant Number of $C = 16$. The figure illustrates two points: (1) the Diagonal ADI scheme itself is a reasonably efficient time-stepping algorithm; and (2) an appreciable increase in the convergence rate is achieved when multigrid is used. The aerodynamic force coefficients have converged to within plottable accuracy in about 200 time steps for the single-grid calculation, and in the equivalent of about 30 time steps for the 5-level multigrid calculation.

The calculations were performed on an IBM 3090-600E, and required about 24 minutes of CPU time for 30 Work Units, which is equivalent to approximately 2.4×10^{-4} *s* per mesh cell per *Work Unit*. This is comparable to the time required for the explicit multi-stage (Runge-Kutta) scheme of Jameson *et al* [2], when applied to three-dimensional problems (see, e.g., Jameson & Baker [8]). The efficiency gained by the diagonalization procedure is compounded by the fact that the additional work required to compute the elements of the modal matrices is vectorizable, and thus is efficiently performed on computers with vector capabilities.

IV. Acknowledgements

This work has been supported, in part, by Grant NAG 2-373 from the NASA Ames Research Center and by the U. S. Army Research Office, through the Mathematical Sciences Institute of Cornell University. The computations have been performed at the Cornell National Supercomputer Facility of the Center for Theory and Simulation in Science and Engineering, which is funded, in part, by the National Science Foundation, New York State, and IBM Corporation.

V. References

1. Caughey, D. A., *"A Diagonal Implicit Multigrid Algorithm for the Euler Equations,"* **AIAA Paper 87-0354**, 25th Aerospace Sciences Meeting, Reno, Nevada, January 1987. (Also, AIAA Journal, in press).

2. Jameson, A., Schmidt, W., and Turkel, E., *"Numerical Solutions of the Euler Equations by Finite Volume Methods using Runge-Kutta Time-stepping Schemes,"* **AIAA Paper 81-1259**, Palo Alto, California, June 1981.

3. Briley, W. R. and McDonald, H., *"Solution of the Three-Dimensional Compressible Navier-Stokes Equations by an Implicit Technique,"* **Proc. of Fourth International Conference on Numerical Methods in Fluid Dynamics,** vol. 35, pp. 105-110, Springer Verlag, New York, 1974.

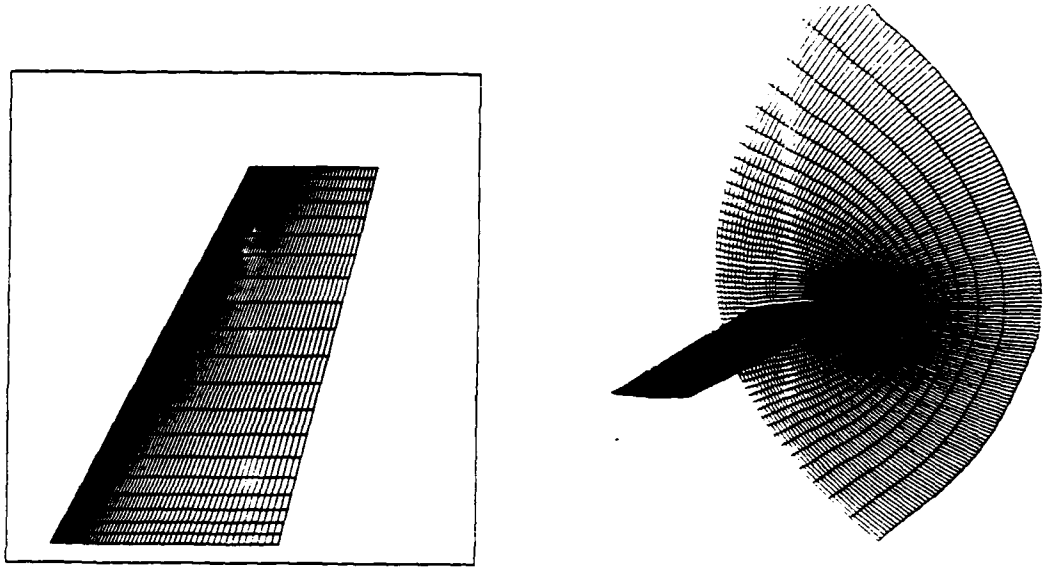
4. Beam, R.M. and Warming, R. F., *"An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation Law Form,"* **Journal of Computational Physics,** vol. 22, pp. 87-110, 1976.

5. Pulliam, T. H. and Chaussee, D. S., *"A Diagonal Form of an Implicit Approximate-Factorization Algorithm,"* **J. Comp. Phys.**, vol. 39, pp. 347-363, 1981.

6. Jameson, A., *"Solution of the Euler Equations by a Multigrid Method,"* **MAE Report 1613,** Princeton University, June 1983.

7. Anonymous, *"Experimental Data Base for Computer Program Assessment,"* **AGARD AR-138,** 1979.

8. Jameson, Antony and Baker, T.J., *"Solution of the Euler Equations for Complex Configurations,"* **Proc. AIAA Computational Fluid Dynamics Conference**, pp. 293-302, Danvers, Mass., July 13-15, 1983.



(a) (b)

Figure 1. Mesh cell distribution for calculation of flow past swept wing.

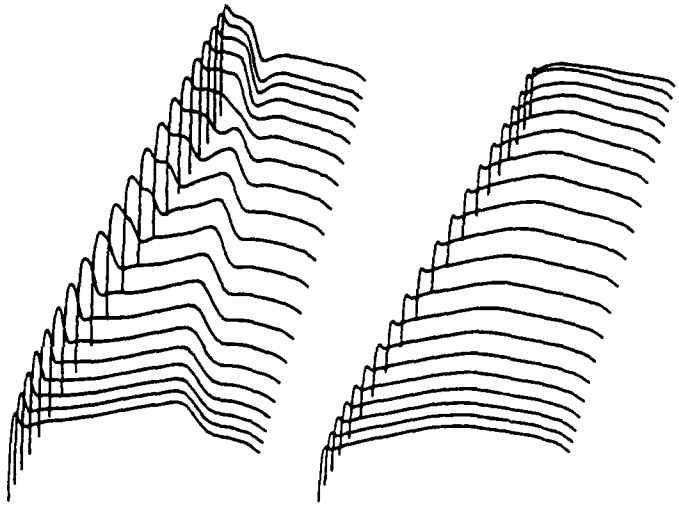
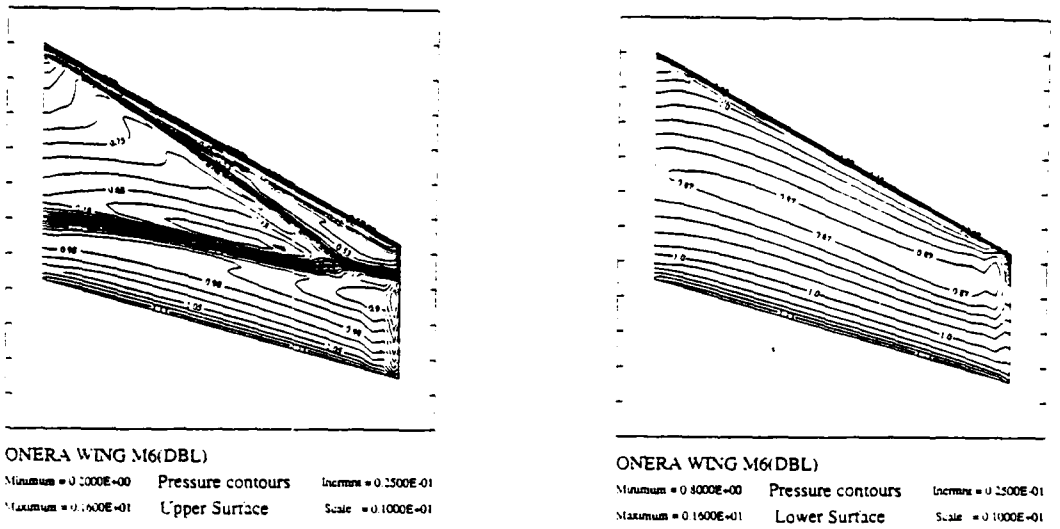


Figure 2. Streamwise distributions of pressure coefficient at computational stations on upper and lower wing surfaces.



(a) (b)

Figure 3. Plan views of constant pressure contours on upper and lower wing surfaces.

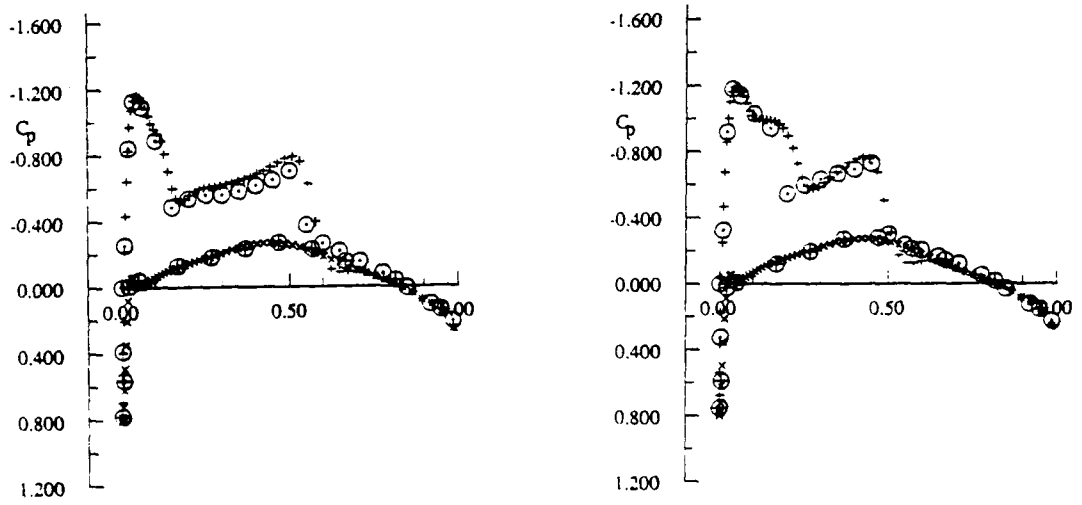


Figure 4. Comparisons of measured and computed pressure coefficients at selected span locations of ONERA Wing M6. Free stream Mach number is $M_\infty = 0.839$ and angle of attack is $\alpha = 3.06°$.

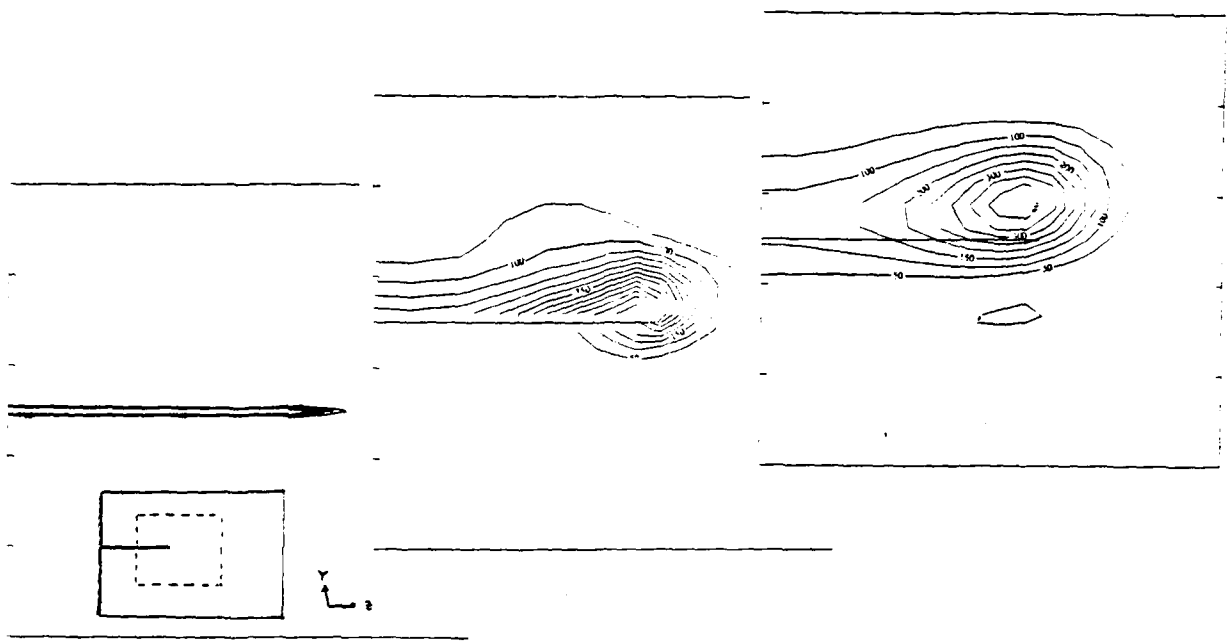


Figure 5. Contours of constant entropy contours in selected cross planes.

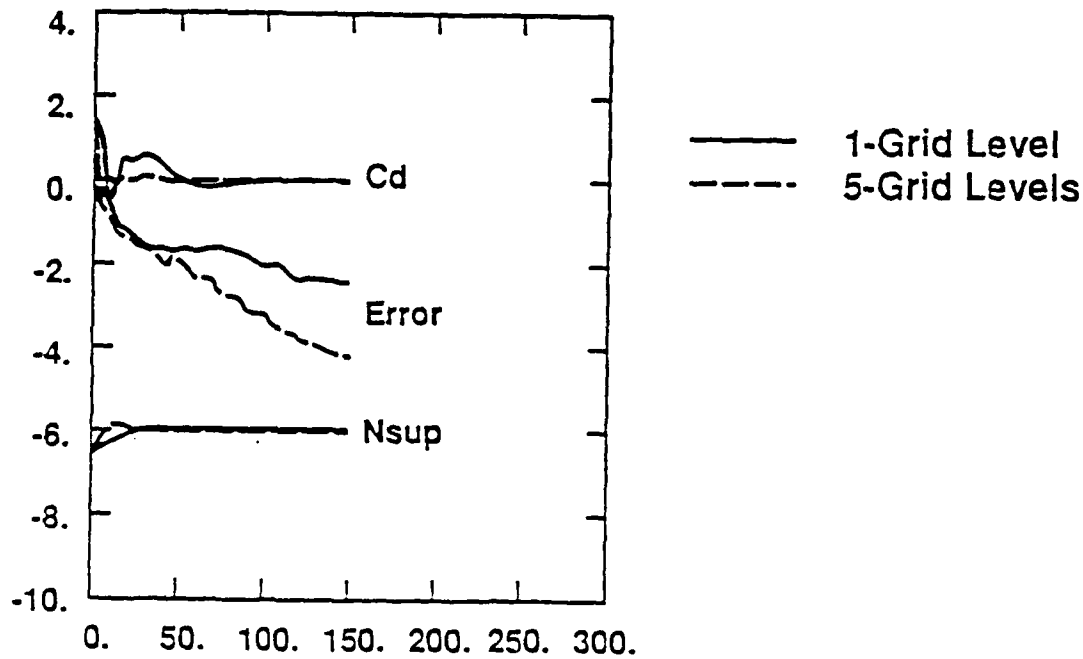


Figure 6. Convergence histories for single-grid and 5-level multigrid schemes.

ADAPTIVE MESH EXPERIMENTS FOR HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS

David C. Arney

Department of Mathematics
United States Military Academy
West Point, NY 10996-1786

Rupak Biswas

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Joseph E. Flaherty

U.S. Army Armament, Munition, and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Laboratory
Watervliet, NY 12189-4050
and
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

ABSTRACT. We discuss experiments conducted on mesh moving and local mesh refinement algorithms that are used with a finite difference scheme to solve initial-boundary value problems for vector systems of hyperbolic partial differential equations in one dimension. The mesh moving algorithms move a coarse base mesh by a mesh movement function so as to follow and isolate spatially distinct phenomena. The local mesh refinement method recursively divides the time step and spatial cells in regions where error indicators are high until a prescribed error tolerance is satisfied.

The adaptive mesh algorithms are implemented in a code with an initial mesh generator, a MacCormack finite difference scheme, and an error estimator. Experiments are conducted for several different problems to determine the efficiency of the adaptive methods and their combinations and to gauge their effectiveness in solving one-dimensional problems.

1. INTRODUCTION. Our goal is to develop expert systems software for solving

time-dependent partial differential equations. The software should allow users to describe problems in a natural language, have a convenient geometric description interface, and not require knowledge of sophisticated numerical analysts. The systems should be intelligent, efficient, reliable, robust and able to solve a large class of problems to prescribed error tolerances.

The power of adaptive techniques is that they are capable of making decisions that change the computational environment. This significantly minimizes the number of a priori decisions demanded of the user and provides dramatic savings in the cost of the computation. This capability is performed by procedures that monitor intermediate results and feed back this data to a control mechanism that modifies the solution strategy. Three popular adaptive techniques for solving partial differential equations are mesh moving or rezoning (r-refinement), mesh refinement (h-refinement), and order enrichment (p-refinement). In r-refinement, the mesh is moved either continuously or statically at discrete times in order to resolve nonuniformities and reduce errors. H-refinement involves the addition or deletion of computational cells to the mesh and p-refinement involves increasing or decreasing the order of a method in different portions of the domain. All strategies attempt to organize the computation so that little effort is expended in regions where the solution is smooth and a much greater effort is devoted to regions where the solution is more difficult to compute.

The different refinement strategies are being combined to yield remarkable results. Babuska and Szabo [8] showed that an hp-refinement scheme produced an exponential rate of convergence on a singular elasticity problem. Arney and Flaherty [6] developed an hr-refinement scheme that moved a 'base' coarse mesh so as to follow important dynamic structures of the solution and recursively refined the base mesh to improve resolution. They found that mesh motion was inexpensive relative to mesh refinement and reduced dispersive errors associated with wave motion but did not always accurately follow structures, especially when interactions occurred, and could not dependably satisfy prescribed tolerances. Recursive mesh refinement can satisfy prescribed tolerances but involves more complicated data structures and greater care at coarse-fine mesh interfaces than r-refinement.

There are numerous other variations of the three adaptive strategies for time-dependent problems. For example, temporal refinement can be done globally to produce an adaptive method of lines strategy [1,13] or locally in combination with the spatial refinement strategy [7,16].

Accurate a posteriori error estimation is essential for codes that strive to satisfy user-prescribed error tolerances. Error estimation is often the most expensive part of an adaptive algorithm. Arney and Flaherty [6] calculated the local discretization error at nodes of the mesh using an algorithm based on Richardson [22] extrapolation. This pointwise estimate can then be used to construct several global measures of the discretization error. The advantage of this method is that it can be used to find error

estimates for any numerical scheme without explicitly knowing the exact form of the error. Details of this error estimate and its implementation on a moving mesh are discussed in Arney [3] and Arney et al. [4].

In this paper, we apply Arney and Flaherty's [6] adaptive mesh moving and refinement technique to one-dimensional hyperbolic systems. As described in Section 2, their approach consists of moving a base mesh of quadrilateral cells so as to isolate important spatial structures of the solution. Refinement, when needed, is performed within cells of coarser meshes. Solutions are generated by a MacCormack [19] finite difference scheme and local error estimates, that are used to control mesh motion and refinement, are computed by Richardson [22] extrapolation. Our goal is to quantify the relative costs and benefits of mesh motion and local mesh refinement. In Section 3, we report the results of computational experiments performed on three one-dimensional problems using several conventional and adaptive numerical procedures. The results obtained demonstrate both the potential and limitations of the adaptive algorithm. We have mixed results showing that the effects of mesh moving can be problem-dependent. Generally, mesh motion is effective for following an isolated structure, but much less so when structures interact. In Section 4, we discuss the utility of our methods, the computational results, and future work.

2. ALGORITHM. We consider an application of Arney and Flaherty's [6] adaptive procedure to one-dimensional vector systems of hyperbolic conservation laws having the form

$$\vec{u}_t + \vec{f}_x(x, \vec{u}, t) = 0, \quad x \in D, \quad t > 0 \tag{1}$$

$$\vec{u}(x, 0) = \vec{u}_0(x), \quad x \in D \cup \partial D, \tag{2}$$

with appropriate well-posed conditions on the boundary ∂D of a domain D. Like them, we discretize Eqs. 1 and 2 using a MacCormack [19] finite difference scheme because of its general applicability [20]. Although this scheme suffers a reduction in order on a moving nonuniform grid, our computations show that proper mesh moving can provide enough efficiency and accuracy to compensate for this order reduction.

The MacCormack scheme produces spurious oscillations near discontinuities because it is a centered scheme with second order accuracy on a uniform mesh. The use of artificial viscosity to make this scheme total variation diminishing (TVD) makes it attractive as a general solver for problems with discontinuities and we use a model due to Davis [12]. The artificial viscosity terms are calculated from the solution data at the beginning of each time step and are added to the solution after the MacCormack solution has been calculated.

Arney and Flaherty's [5] mesh moving procedure is based on an intuitive approach that allows nodes to follow local nonuniformities rather than the more analytic

approaches of equidistribution of error [13] or the solving of variational problems to minimize some given functional [14], which can be expensive and problem-dependent. They derive equations for the nodal velocities so that the mesh moves to follow the geometric propagation of some local nonuniformity. This generally reduces dispersive errors and allows the use of larger time steps while maintaining accuracy and stability. Important factors for mesh moving are to maintain mesh smoothness by controlling adjacent cell ratios, to keep nodes within the domain boundaries, and to move nodes with a velocity that reduces discretization error. In order to prevent mesh distortion that can lead to increased discretization error of the solver, mesh points cannot move independently but must be coupled to at least some of their neighbors.

Some schemes do this coupling by attraction and repulsion of nodes (cf., Rai and Anderson [21]). In these algorithms, the coupling is done globally, where each node influences the velocity of all other nodes in the mesh. Attempting to equidistribute errors can lead to problems where nodes move incorrectly in some regions. This occurs, for example, when a mesh that is following one structure must react to another nonuniformity that arises in another part of the domain. An abrupt grid adjustment can be eliminated if the influence is more local and the movement algorithm is combined with a mesh refinement scheme to add the necessary nodes in the region of the new structure.

At each time step, the selection mechanism of Arney and Flaherty's [6] mesh moving algorithm uses as feedback the current node locations and the nodal values of a mesh movement indicator at the independent moving nodes of a coarse mesh. The local error estimates are used as the mesh movement indicators. Nodes with 'significant error' are grouped into error clusters. This clustering separates the important spatially distinct phenomena of the solution. As time evolves, the clusters can move, change size, collide, or separate. At each time step, new clusters can be created and old ones can vanish.

Mesh movement is then determined by each node's relationship to its nearest error cluster and the propagation velocity of the center of error mass of the cluster. Therefore, the nodal influence is regional. The amount of movement 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. [11]

$$\ddot{r} + \lambda \dot{r} = 0, \tag{3}$$

where r(t) is the position of the center of error mass of a cluster and $(\dot{}) := d(\,)/dt$. Additionally, this movement function smoothes the mesh motion and prevents nodes from moving outside the domain boundary. The distance a node moves is reduced near boundaries in order to prevent it from leaving the domain. Nodes on the domain boundaries are not allowed to move.

Arney and Flaherty [6] perform static rezoning whenever computation with the current mesh would be counterproductive or when the current mesh suffers from poor mesh ratios. There are sophisticated algorithms to check mesh condition and to verify the validity of the mesh (cf., Babuska and Szabo [8] and Simpson [23]). These algorithms can check for gaps between cells and overlapping cells. Since moving meshes can only develop such severe problems over time, mesh degradation can be discovered before it develops complete invalidity. This mesh degradation or ill-conditioning occurs when the mesh angles are severe, the mesh contains poor mesh ratios or poor aspect ratios, or the time step is too restrictive because of the crowding of nodes. Nodes of the coarse mesh can become too crowded when error clusters pass through boundaries or when two or more error clusters converge and trap nodes between them. Static rezoning is performed only when absolutely necessary due to the high cost in accurately interpolating the solution from the existing mesh onto a new one. It was not done in any of the examples presented in this paper.

Arney and Flaherty [6] used local mesh refinement to insure that the user-prescribed error tolerance were satisfied. This was done by recursively introducing finer meshes by binary refinement of space-time cells in regions where nodes with unacceptable error have formed clusters (cf., Berger [9], Flaherty and Moore [16] and Gropp [17]). The clustering algorithm used for refinement is the same as the one used for mesh movement. The clusters are buffered so that high error nodes are in the interior of the refined region. The problem is recursively solved on these fine meshes until the error is within the specified tolerance. The refined subgrids that are adaptively created by the local refinement algorithm, overlay the coarser grids. Each of these subgrids is independently defined. Figure 1 shows a coarse grid with portions overlayed by two fine grids and three finer grids.

Arney and Flaherty's [7] mesh refinement strategy suggests the use of a tree data structure for its description and implementation. In this tree structure, the coarsest grid is the root node and defined as level 0 in the tree. The subgrids of the coarse grid are its offspring in the tree and are defined as level 1. A grid at level ℓ is properly nested in the tree between its parent at level ℓ - 1 and its offspring (if any) at level ℓ + 1. Grids at the same level are given an arbitrary ordering. Due to the clustering and buffering of error regions, grids at the same level of a two-dimensional problem can intersect and overlap. Figure 1 depicts an example of a sequence of meshes that might be produced by our refinement procedure for a coarse grid refinement step. The numbers next to the grids indicate the order in which the solution is computed on each grid.

Such tree data structures are commonly used in adaptive mesh refinement procedures (cf., Berger and Oliger [10] and Flaherty and Moore [16]). Additionally, we use a stack to implement the recursive algorithm (cf., Aho et al. [2] and Horowitz and Sahni [18]).

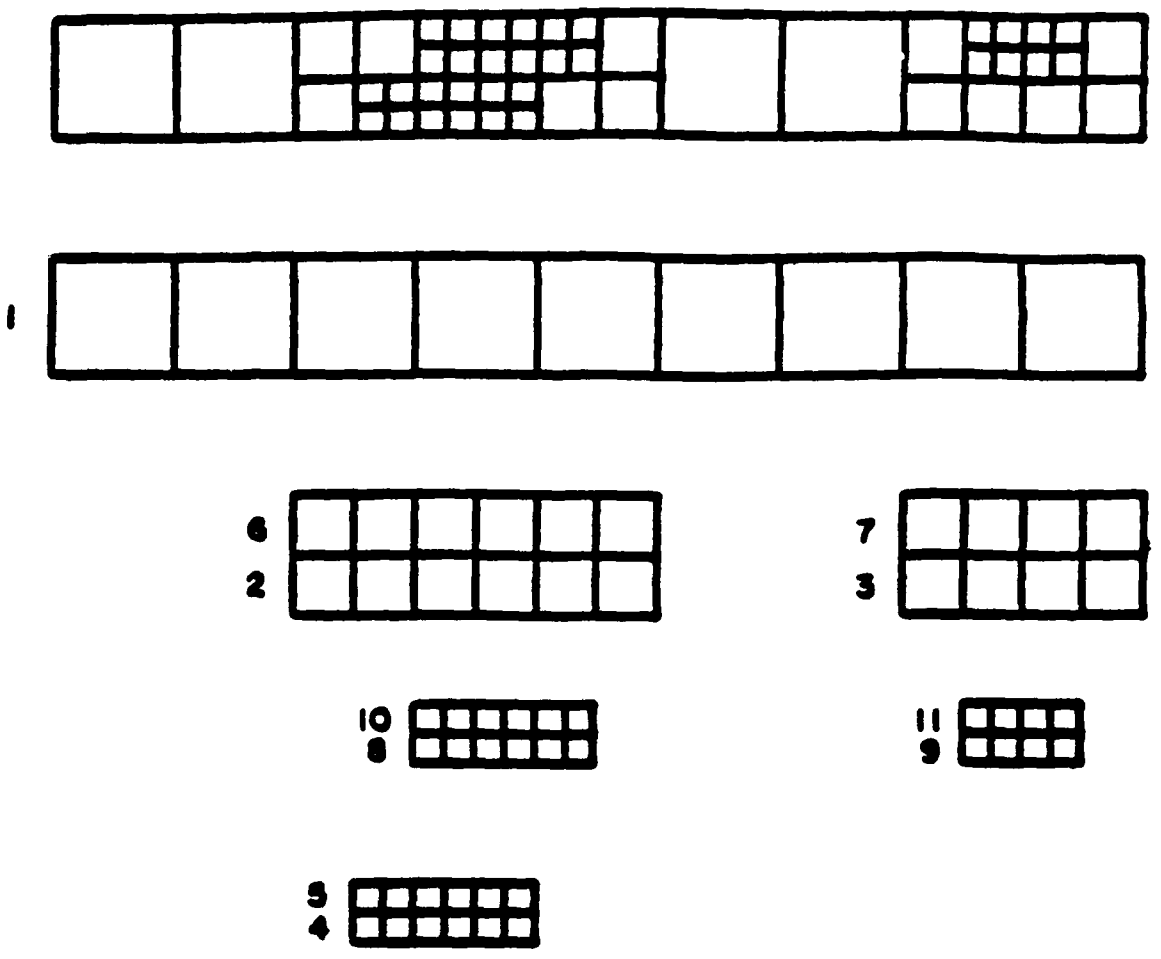


Figure 1. Typical set of local refinement grids for one coarse time step. The numbers indicate the order in which the solution is computed on each grid.

The solution vectors, error estimates, and nodal information are all stored in a dynamic storage area with pointers from the tree to this storage area for each mesh in the tree. For each grid, we store its level in the tree and the number of nodes it contains. The dynamic storage area contains the solution vector and the error estimate at each node, and nodal information for use in the solver and grid interface procedures. Since old mesh data is saved to obtain initial data for the newly refined grids and nodes of the parent grids are updated from the fine grid solutions, nodal relationships between meshes are stored directly in the nodal vector.

3. COMPUTATIONAL RESULTS. We conducted experiments of Arney and Flaherty's [6] adaptive mesh strategy using three problems and our results follow. In each case, errors are measured in the L_1 norm and the CPU times are normalized to unity. All calculations were performed on an IBM 3081D computer.

Example 1. Consider the scalar hyperbolic differential equation

$$u_t + (\cos\ \pi t)\, u_x = 0, \qquad t > 0, \qquad -0.4 \le x \le 1.4, \tag{4}$$

with initial conditions

$$u(x, 0) = \begin{cases} 1, & \text{if} \quad 0.4 \le x \le 0.6 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

and boundary conditions

$$u(-0.4, t) = u(1.4, t) = 0. \tag{6}$$

The exact solution to this problem is

$$u(x, t) = \begin{cases} 1, & \text{if} \quad 0.4 \le x - (\sin\ \pi t) / \pi \le 0.6 \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

which is a square pulse of unit amplitude that oscillates sinusoidally about the center of the domain. Artificial viscosity was added to eliminate oscillations in the solution; however, this resulted in an attenuation and spreading of the square pulse.

Four different adaptive strategies were used to solve this problem for $0 \le t \le 2.5$. The solutions at several times, the mesh trajectories, and the time step profile for the various strategies are shown in Figures 2, 3, 4, and 5. Table 1 summarizes the computational cost and accuracy of the four strategies.

With a stationary uniform mesh, we find that the square pulse is rapidly attenuated and diffused. The time step profile shows how the Courant number is utilized to maintain maximum step sizes without loss of stability. From Figure 3, it is apparent that the results improve when the mesh is allowed to move. The pulse is attenuated less, the error is reduced, but more time steps are needed to complete the computation. The mesh trajectories in Figure 3 demonstrate how well the nodes track the square pulse as it oscillates. Figures 4 and 5, depicting the results of Strategies 3 and 4, respectively, show remarkable improvement when adaptive mesh refinement is used. In both cases, the local error tolerance was specified as 0.001. Errors are reduced and the attenuation of the pulse is almost negligible but shape distortion is still

Adaptive Strategy	$\|\| e \|\|_1$	Number of Space-time Cells	Normalized CPU Time	Attenuation
1. Stationary uniform mesh	0.1090	774	1.000	0.545
2. Moving mesh	0.0903	1134	1.452	0.730
3. Stationary uniform mesh with refinement	0.0614	15718	8.761	0.969
4. Moving mesh with refinement	0.0395	16554	10.069	0.994

Table 1. Comparison of the different adaptive strategies for Example 1.

significant. Notice how well the refinement procedure tracks the pulse; however, the cost of computation increases by almost an order of magnitude. When moving and refinement are combined, the results are even more remarkable. The pulse is attenuated by a factor of only 0.6 percent.

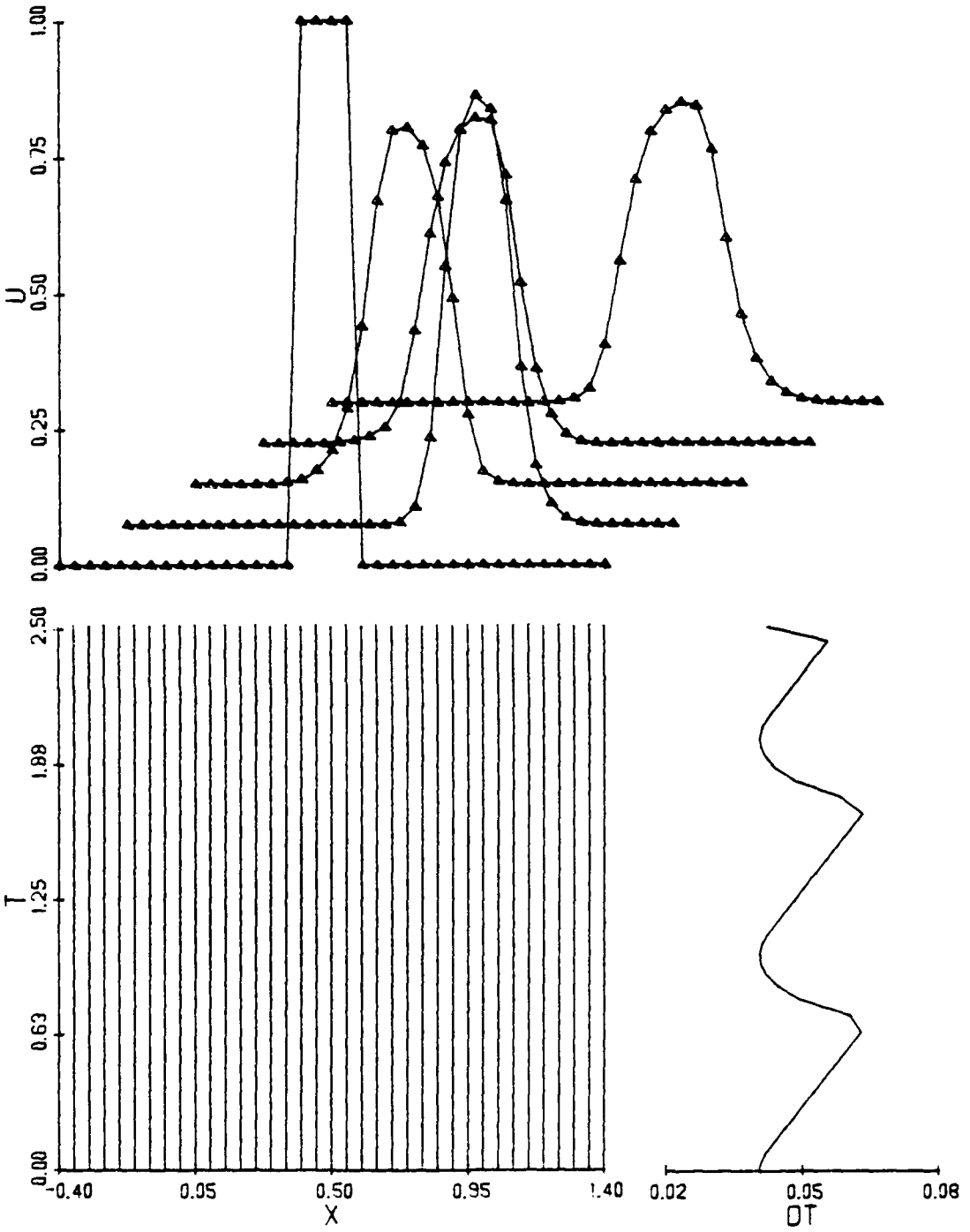


Figure 2. Solution at t = 0, 0.57, 1.15, 1.80, and 2.33, mesh trajectories, and time step profile for Strategy 1 of Example 1.

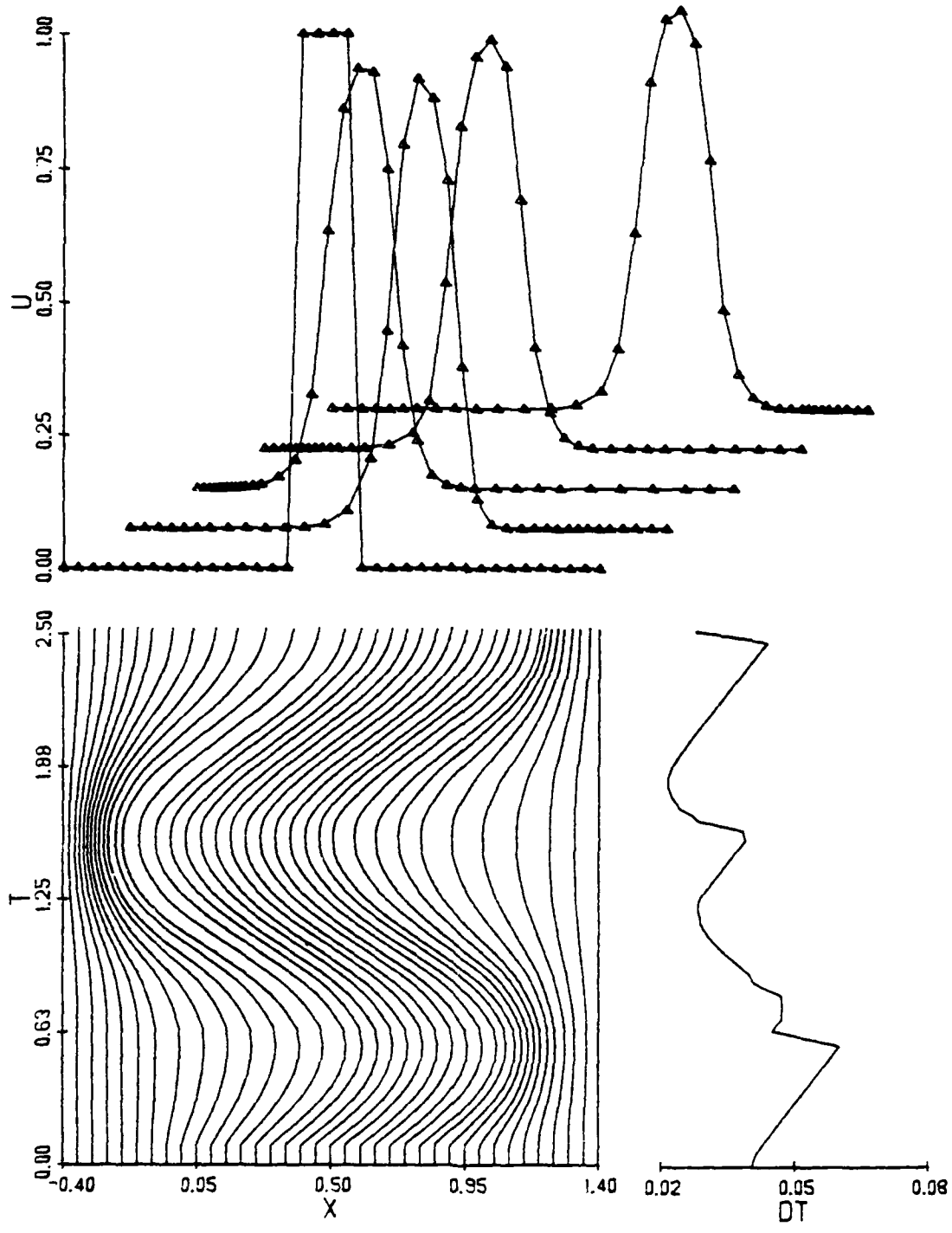


Figure 3. Solution at t = 0, 0.87, 1.42, 1.88, and 2.37, mesh trajectories, and time step profile for Strategy 2 of Example 1.

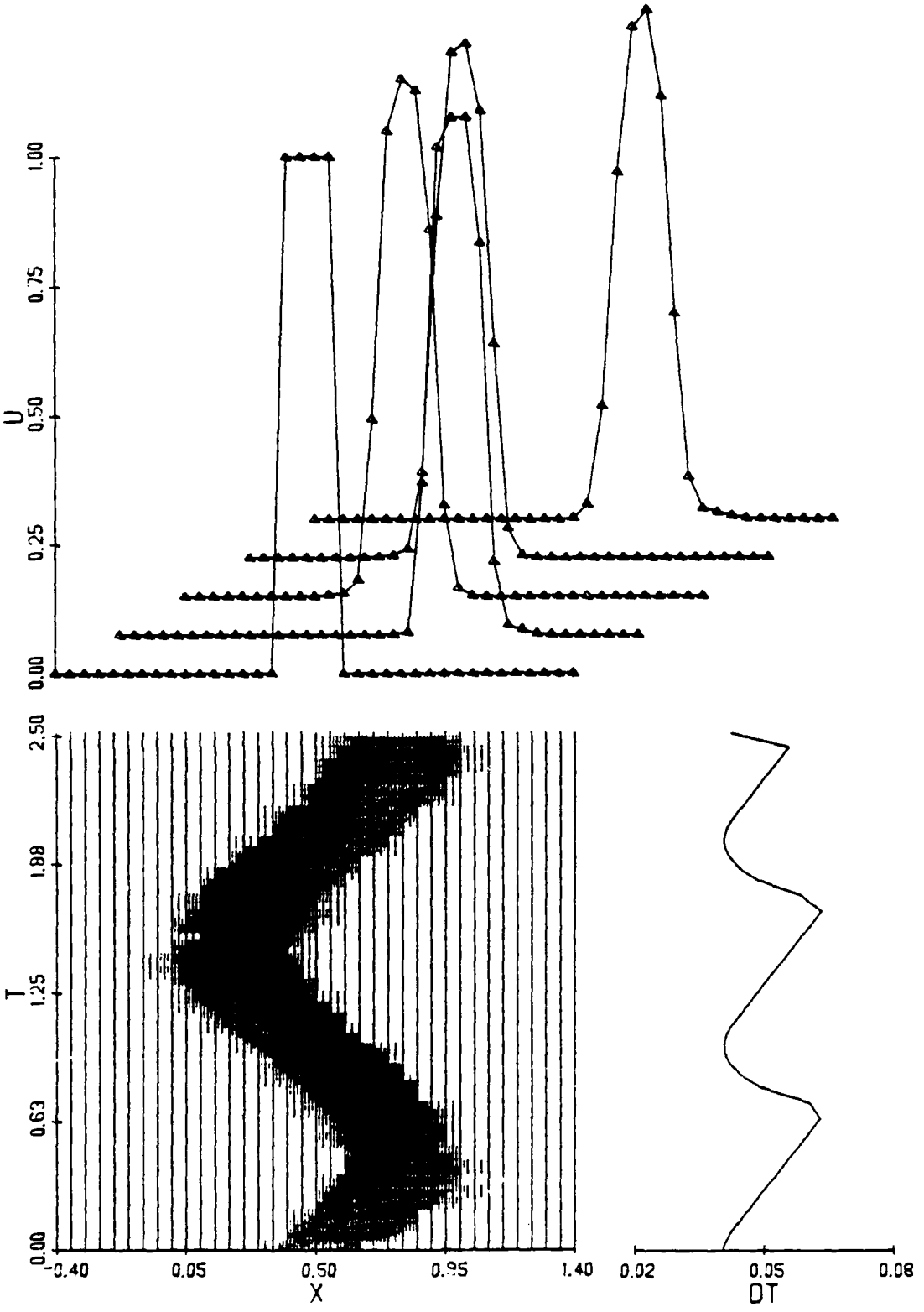


Figure 4. Solution at t = 0, 0.57, 1.15, 1.80, and 2.33, mesh trajectories, and time step profile for Strategy 3 of Example 1.

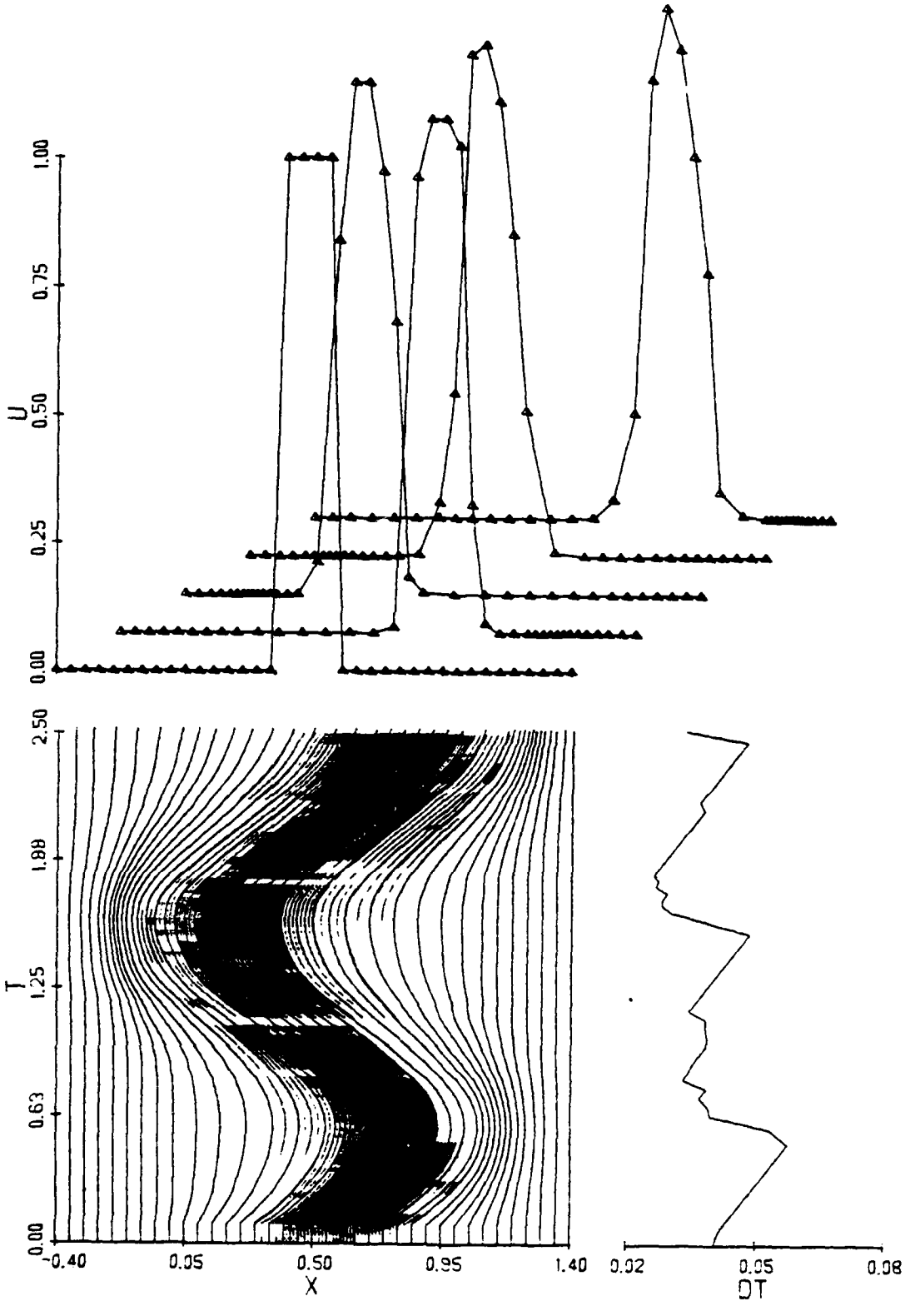


Figure 5. Solutions at t = 0, 0.77, 1.37, 1.91, and 2.50, mesh trajectories, and time step profile for Strategy 4 of Example 1.

Example 2. Consider the linear uncoupled system

$$\begin{aligned} &u_t + (\cos\ \pi t)\, u_x = 0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad t > 0, \quad -0.4 \le x \le 1.4, \\ &v_t - (\cos\ \pi t)\, v_x = 0, \end{aligned} \tag{8}$$

with initial conditions

$$u(x, 0) = v(x, 0) = \begin{cases} 1, & \text{if} \quad 0.4 \le x \le 0.6 \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

and boundary conditions

$$u(-0.4, t) = u(1.4, t) = v(-0.4, t) = v(1.4, t) = 0. \tag{10}$$

The exact solution to this problem is

$$u(x, t) = \begin{cases} 1, & \text{if} \quad 0.4 \le x - (\sin\ \pi t) / \pi \le 0.6 \\ 0, & \text{otherwise,} \end{cases} \tag{11}$$

$$v(x, t) = \begin{cases} 1, & \text{if} \quad 0.4 \le x + (\sin\ \pi t) / \pi \le 0.6 \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The first component u is the same as Example 1, and the second component v moves symmetrically with u. Four different adaptive strategies were used to solve this problem for $0 \le t \le 1.5$. Table 2 summarizes the computational cost and accuracy of the four strategies. The solutions at several times, the mesh trajectories, and the time step profile for mesh strategies 3 and 4 are shown in Figures 6 and 7, respectively.

It is clear that mesh moving does not provide the expected improvement in the results for this problem. In fact, we can see from Table 2 that each time the mesh is moved, the error in the computed solution increases. This is because two identical error regions moving symmetrically about the center of the domain do not contribute equally to the mesh motion due to asymmetries in their error estimates. As a result, the mesh moves incorrectly and the solution deteriorates. This, in turn, leads to further imbalance of the error clusters and subsequently causes catastrophic effects. Comparing Figures 6 and 7, we see how bad the solution is attenuated due to incorrect mesh motion. Improper mesh motion has also lead to refinement in some regions of the mesh where it should not have been necessary. In both cases, the local error tolerance was specified to be 0.005.

Adaptive Strategy	$\|\| e \|\|_1$	Number of Space-time Cells	Normalized CPU Time
1. Stationary uniform mesh	0.1145	1650	1.000
2. Moving mesh	0.1221	5640	3.386
3. Stationary uniform mesh with refinement	0.0541	20828	6.926
4. Moving mesh with refinement	0.0583	48954	18.667

Table 2. Comparison of the different adaptive strategies for Example 2.

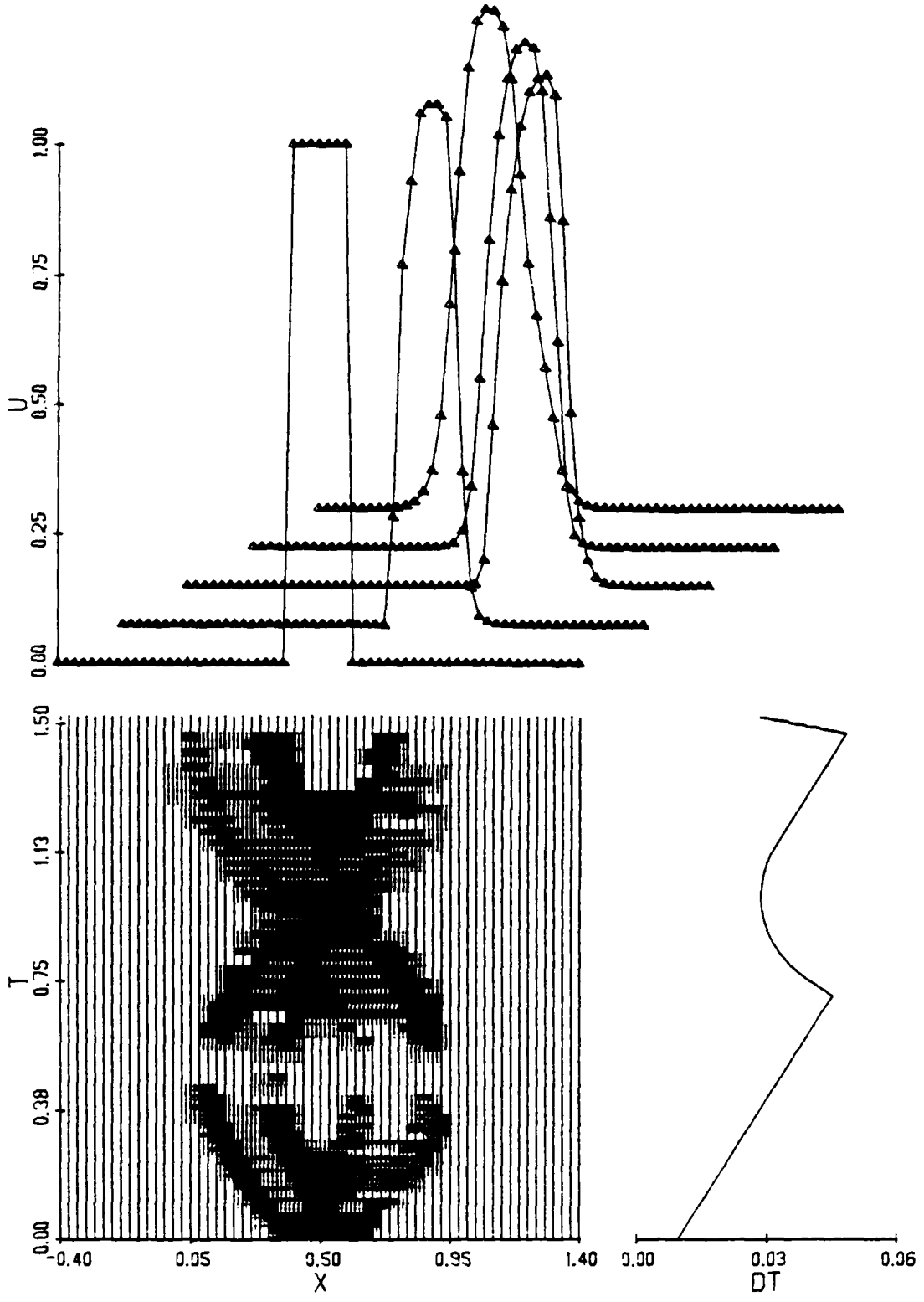


Figure 6. Solutions for u at t = 0, 0.17, 0.48, 0.96, and 1.38, mesh trajectories, and time step profile for Strategy 3 of Example 2.

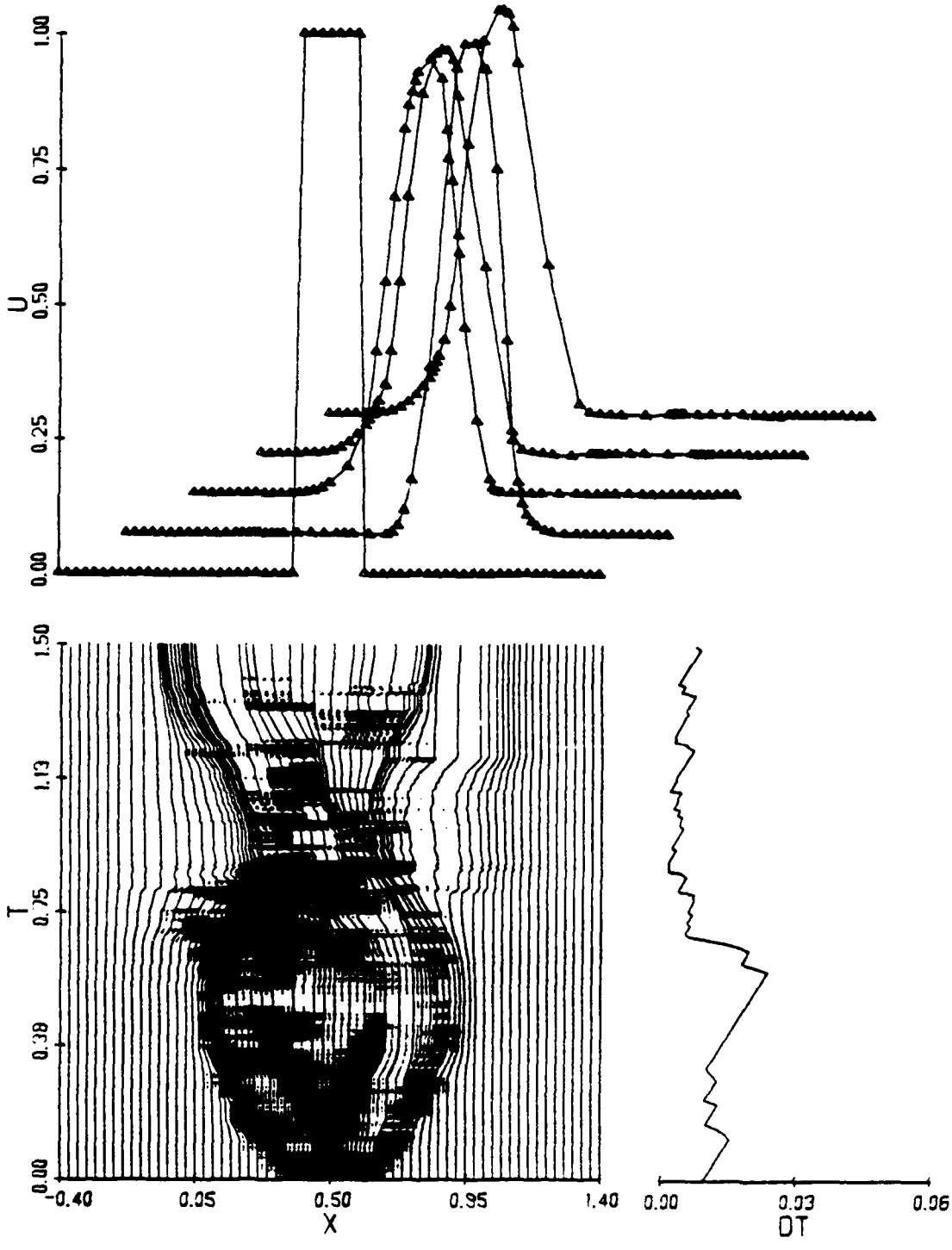


Figure 7. Solutions for u at t = 0, 0.76, 0.95, 1.17, and 1.49, mesh trajectories, and time step profile for Strategy 4 of Example 2.

Example 3. Consider the coupled hyperbolic system from the wave equation

$$\begin{aligned} u_t - v_x &= 0, \\ v_t - u_x &= 0, \end{aligned} \qquad t > 0, \quad -0.3 \le x \le 1.4, \tag{13}$$

with initial conditions

$$u(x, 0) = \begin{cases} 1, & \text{if} \quad 0.4 \le x \le 0.6 \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

$$v(x, 0) = \begin{cases} 1, & \text{if} \quad 0.5 \le x \le 0.7 \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

and boundary conditions satisfying the exact solution

$$u(x, t) = (p(x + t) + q(x - t)) / 2.0 \tag{16}$$

$$v(x, t) = (p(x + t) - q(x - t)) / 2.0, \tag{17}$$

where

$$p(\xi) = \begin{cases} 2, & \text{if} \quad 0.5 \le \xi \le 0.6 \\ 1, & \text{if} \quad 0.4 \le \xi < 0.5 \quad \text{or} \quad 0.6 < \xi \le 0.7 \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

and

$$q(\xi) = \begin{cases} 1, & \text{if} \quad 0.4 \le \xi \le 0.5 \\ -1, & \text{if} \quad 0.6 \le \xi \le 0.7 \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Four different adaptive strategies were used to solve this problem for $0 \le t \le 0.6$. Table 3 summarizes the computational cost and accuracy of the four strategies. The solutions at several times, the mesh trajectories, and the time step profile for mesh strategies 3 and 4 are shown in Figures 8 and 9, respectively.

Once again, mesh motion does not appear to result in the desired improvement in the solution. In this case, there are two error regions moving away with unit speed in opposite directions from the center of the domain. However, the error regions are not identical as was the case in Example 2. With a moving mesh, the solution is attenuated and consequently, the error measure in the L_1 norm increases.

Adaptive Strategy	$\|\| e \|\|_1$	Number of Space-time Cells	Normalized CPU Time
1. Stationary uniform mesh	0.1141	930	1.000
2. Moving mesh	0.1057	3180	3.454
3. Stationary uniform mesh with refinement	0.0527	23236	11.493
4. Moving mesh with refinement	0.0552	39234	20.232

Table 3. Comparison of the different adaptive strategies for Example 3.

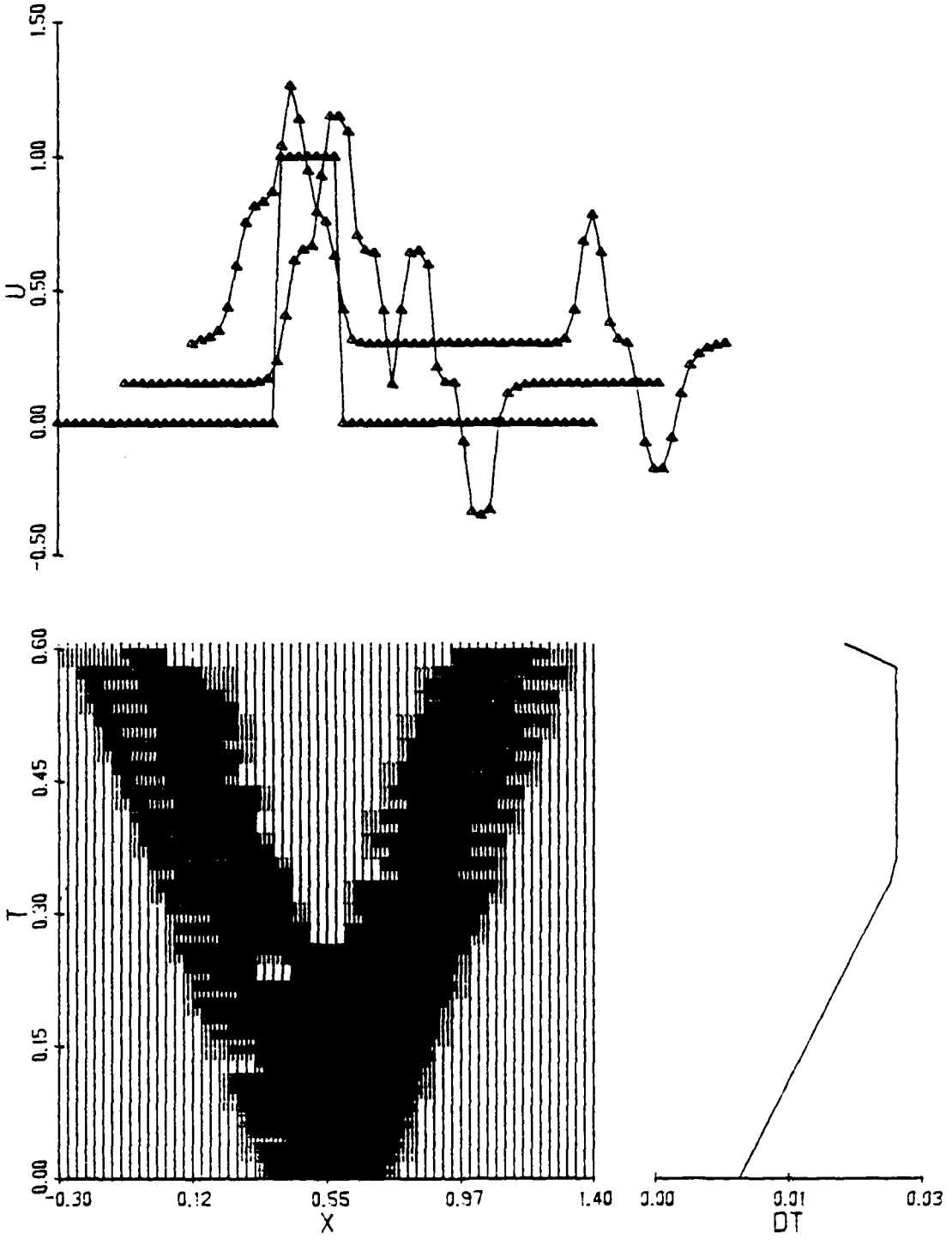


Figure 8. Solutions for u at t = 0, 0.20, and 0.58, mesh trajectories, and time step profile for Strategy 3 in Example 3 with error tolerance = 0.005.

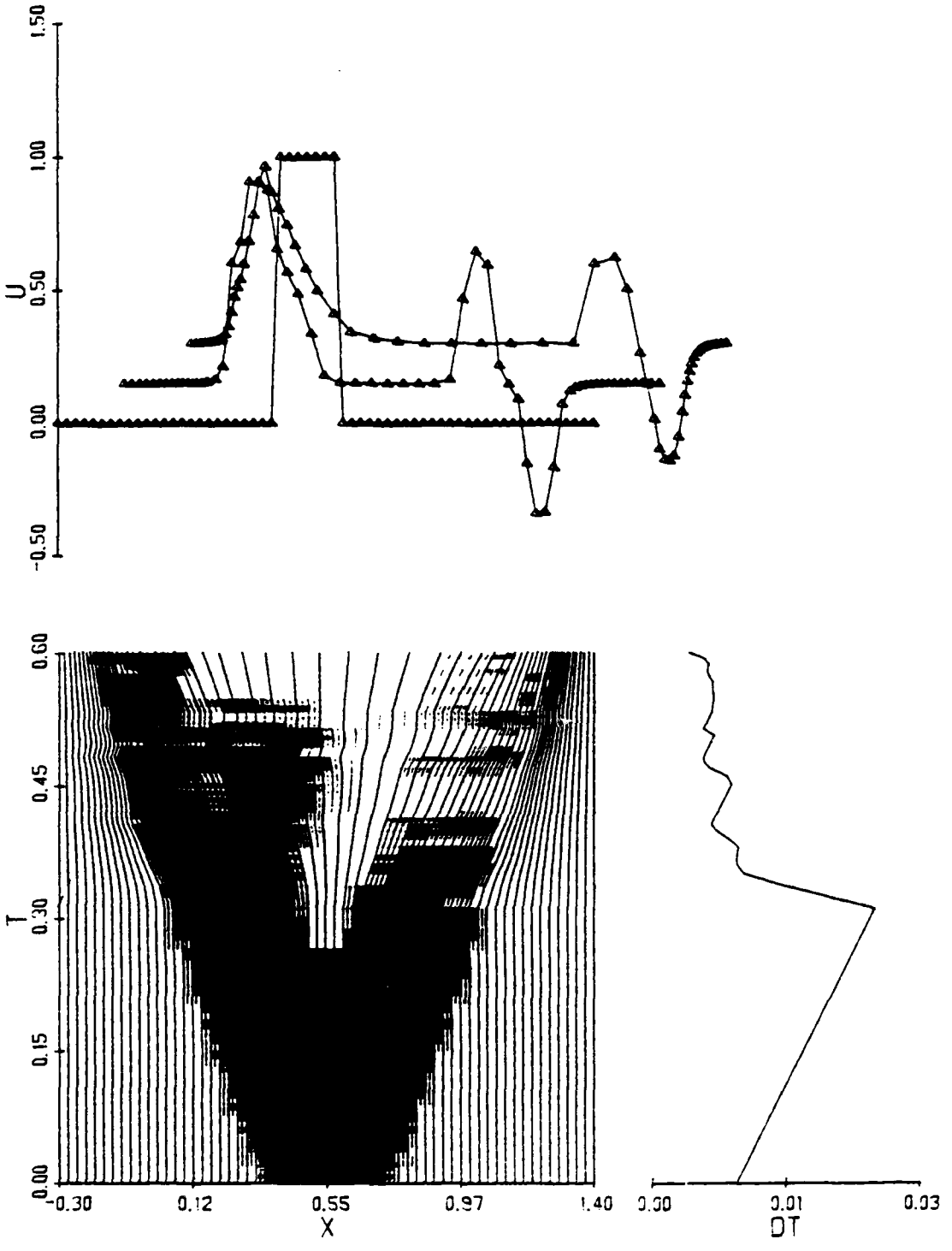


Figure 9. Solutions for u at t = 0, 0.41, and 0.60, mesh trajectories, and time step profile for Strategy 4 in Example 3 with error tolerance = 0.005.

4. CONCLUSION. We have experimented with the adaptive mesh method of Arney and Flaherty [6]. Our results indicate that proper mesh moving can efficiently reduce errors. However, their mesh moving is not effective for problems that have more than one moving structure. We find that whenever there are two error regions, the mesh moving strategy is unable to make an accurate decision. This occurs particularly during the time when the two structures have not completely separated but still form one large error cluster. Results of tests of local refinement show that it can efficiently reduce errors. The most powerful method was the combination of both mesh moving and mesh refinement. Results obtained for Example 1 show that a totally adaptive mesh strategy can be extremely effective. The overhead associated with the clustering and dynamic data structures is only about 5 percent of the time needed to calculate a comparable solution on a uniform mesh.

Additional computation is needed to verify the generality of these conclusions. It is also not clear how much of the difficulties were due to MacCormack's [19] finite difference scheme or the Richardson's [22] extrapolation-based error estimate. A TVD scheme would greatly improve performance near discontinuities.

We are re-examining the entire process in order to determine an effective mesh motion procedure. Future computations will be performed using more advanced shock capturing difference schemes (e.g., Engquist and Osher [15]).

LITERATURE CITED

1. Adjerid, S., and Flaherty, J. E., "A Moving Finite Element Method for Time Dependent Partial Differential Equations with Error Estimation and Refinement," SIAM J. Numer. Anal., 23, 1986, pp. 778-795.

2. Aho, A., Hopcroft, J., and Ullman, J., The Design and Analysis of Computer Algorithms, Addison-Wesley, Reading, Massachusetts, 1974.

3. Arney, D. C., "An Adaptive Mesh Algorithm for Solving Systems of Time Dependent Partial Differential Equations," Ph.D. Thesis, Department of Mathematics, Rensselaer Polytechnic Institute, 1985.

4. Arney, D. C., Biswas, R., and Flaherty, J. E., "A Posteriori Error Estimation of Adaptive Finite Difference Schemes for Hyperbolic Systems," Trans. Fifth Army Conf. on Appl. Math. and Comput., 1988, pp. 437-457.

5. Arney, D. C., and Flaherty, J. E., "A Two-Dimensional Mesh Moving Technique for Time Dependent Partial Differential Equations," J. Comput. Phys., 67, 1986, pp. 124-144.

6. Arney, D. C., and Flaherty, J. E., "An Adaptive Method with Mesh Moving and Local Mesh Refinement for Time-Dependent Partial Differential Equations," Trans. Fourth Army Conf. on Appl. Math. and Comput., 1987, pp. 1115-1141.

7. Arney, D. C., and Flaherty, J. E., "An Adaptive Local Mesh Refinement Method for Time-Dependent Partial Differential Equations," to appear in Appl. Num. Math.

8. Babuska, I., and Szabo, B., "On the Rates of Convergence of the Finite Element Method," Intl. J. Numer. Methods in Engrg., 18, 1982, pp. 323-341.

9. Berger, M., "Data Structures for Adaptive Mesh Refinement," Adap. Comp. Methods for Part. Diff. Eqns., I. Babuska, J. Chandra, and J. E. Flaherty (Eds.), SIAM, Philadelphia, 1983, pp. 237-251.

10. Berger, M., and Oliger, J., "Adaptive Mesh Refinement for Hyperbolic Partial Differential Equations," J. Comput. Phys., 53, 1984, pp. 484-512.

11. Coyle, M., Flaherty, J. E., and Ludwig, R., "On the Stability of Mesh Equidistributing Strategies for Time Dependent Partial Differential Equations," J. Comput. Phys., 62, 1986, pp. 26-39.

12. Davis, S. F., "A Simplified TVD Finite Difference Scheme via Artificial Viscosity," SIAM J. Sci. Stat. Comput., 8, 1987, pp. 1-18.

13. 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., 3, 1982, pp. 6-27.

14. Djomehri, J., and Miller, K., "A Moving Finite Element Code for General Systems of PDE's in 2-D," PAM-57, Center for Pure and Appl. Math., University of California, Berkeley, 1981, pp. 1-49.

15. Engquist, B., and Osher, S., "One-sided Difference Approximations for Nonlinear Conservation Laws," Math. Comp., 36, 1981, pp. 321-351.

16. Flaherty, J. E., and Moore, P. K., "A Local Refinement Finite Element Method for Time Dependent Partial Differential Equations," Trans. Second Army Conf. on Appl. Math. and Comput., 1985, pp. 585-596.

17. Gropp, W. D., "A Test of Moving Mesh Refinement for 2-D Scalar Hyperbolic Problems," SIAM J. Sci. Stat. Comput., 1, 1980, pp. 191-197.

18. Horowitz, E., and Sahni, S., Fundamentals of Data Structures. Computer Science Press, Potomac, Maryland, 1976.

19. MacCormack, R. W., "Numerical Solution of the Interaction of a Shock Wave with a Laminar Boundary Layer," Proc. Second Intl. Conf. on Numer. Methods in Fluid Dynamics., M. Holt (Ed.), Lecture Notes in Physics, 8, Springer-Verlag, New York, 1971.

20. Mitchell, A. R., and Griffiths, D. F., The Finite Difference Method in Partial Differential Equations, Wiley, New York, 1980.

21. Rai, M. M., and Anderson, D. A., "Grid Evolution in Time Asymptotic Problems," J. Comput. Phys., 43, 1981, pp. 327-344.

22. Richardson, L. F., "The Deferred Approach to the Limit, I. Single Lattice," Trans. Roy. Soc. London, 226, 1927, pp. 299-349.

23. Simpson, R. B., "A Two-Dimensional Mesh Verification Algorithm," SIAM J. Sci. Stat. Comput., 2, 1981, pp. 455-473.

COMPUTATIONS OF TRANSONIC FLOW OVER PROJECTILES AT ANGLE OF ATTACK

Jubaraj Sahu
Computational Aerodynamics Branch
Launch and Flight Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5066

ABSTRACT

The determination of aerodynamic coefficients by shell designers is a critical step in the development of any new projectile design. Of particular interest is the determination of the aerodynamic coefficients at transonic speeds. It is in this speed regime that the critical aerodynamic behavior occurs and a rapid increase in the aerodynamic coefficients is observed. The three-dimensional transonic flowfield computations over projectiles have been made using an implicit, approximately factored, partially flux-split algorithm. Use of a composite grid scheme has been made to provide the increased grid resolution needed for accurate numerical simulation of three-dimensional transonic flows. Details of the asymmetrically located shockwaves on the projectiles have been determined. Computed surface pressures have been compared with experimental data and are found to be in good agreement. The pitching moment coefficient, determined from the computed flowfields, shows the critical aerodynamic behavior observed in free flights.

I. INTRODUCTION

The flight of projectiles covers a wide range of speeds. The accurate prediction of projectile aerodynamic at these speeds is of significant importance in the early design stage of a projectile. The critical aerodynamic behavior occurs in the transonic speed regime, $0.9 < M < 1.1$ where the aerodynamic coefficients have been found to increase by as much as 100%. Of particular interest is the determination of the pitching moment coefficient since it determines the static stability of the flight of the projectile. The critical behavior in this case is usually characterized by a rapid increase in the coefficient followed by a sharp drop. This rapid change in the pitching moment coefficient can be attributed in part to the complex flow structure and in particular, to the asymmetrically located shock waves that exist on the projectiles flying at transonic speeds at angle of attack. Computations of three-dimensional flowfields at transonic speeds are thus needed to predict the critical aerodynamic behavior.

In recent years a considerable research effort has been focused on the development of modern predictive capabilities for determining projectile aerodynamics. Numerical capabilities have been developed primarily using Navier-Stokes[1-4] computational technique and used to compute flow over slender bodies of revolution at transonic speeds. Flowfield computations included both axisymmetric[3] and three-dimensional situations.[1,2,4] References 1 and 2

did not include the computations in the wake or base region of a projectile and thus, ignored the upstream effect of the base region flow on the afterbody flowfield and the asymmetrical locations of the shock wave. An axisymmetric base flow code[3] was developed to compute the entire projectile flowfield including the base region using a flowfield segmentation procedure. This technique was later extended[4] into three dimensions to calculate the pitch plane aerodynamics at transonic speeds. Due to lack of computer resources, only one solution was obtained and reported in Reference 4. In addition, the calculations in References 1, 2, and 4 generally did not have sufficient grid resolution due to lack of adequate computer resources. Due to the availability of supercomputers such as Cray X-MP/48 and Cray 2, it is now possible to provide the increased grid resolution needed for accurate computations of three-dimensional transonic flows.[5]

The numerical scheme plays an equally important role for accurate predictions of transonic flows. All the calculations in Reference 1-4 were made using the compressible, thin-layer Navier-Stokes equations which were solved using the implicit Beam and Warming central finite difference scheme.[6-8] Such schemes require artificial dissipation to be added to control numerical oscillations. Upwind schemes can have several advantages over central difference schemes including natural numerical dissipation and better stability properties. The numerical scheme used here is an implicit scheme based on flux-splitting[9] and upwind spatial differencing in the streamwise direction.

Other factors that have direct impact on the 3-D numerical simulation are the geometric complexity and efficient management of large 3-D data sets. These factors make it necessary to develop zonal or patched methods where a large 3-D problem is divided into a number of smaller problems. Each smaller piece is then solved separately. The break-up of the large data base can be achieved in various ways.[10-13] Reference 10 and 11 are earlier applications where the data base structure follows a pencil format. These numerical calculations, although promising, were based on limited computer resources. Reference 12 shows the development of a chimera grid scheme. This scheme provides multiple regions where communications between grids are done by interpolating in regions of overlap. The blocked grid approach[13] does not require interpolations at the interfaces. The schemes in References 12 and 13 are generally complicated since they allow to embed a block or zone into another. Recently, a simple composite grid scheme[14] has been developed where a large single grid was partitioned into smaller grids so that each of the smaller problem could be solved separately with simple data transfers at the interfaces. The initial results obtained were very promising. The present effort extends the use of this composite grid scheme to include the correct modeling of the base region of a projectile. Three-dimensional flowfields have been computed for two different projectiles at various transonic speeds $0.8 < M < 1.2$ in order to determine the critical aerodynamic behavior.

II. NUMERICAL METHOD

1. GOVERNING EQUATIONS

The three-dimensional Navier-Stokes conservation equations of mass, momentum, and energy can be represented in flux vector form as:

$$\partial_\tau \hat{Q} + \partial_\xi(\hat{F} + \hat{F}_v) + \partial_\eta(\hat{G} + \hat{G}_v) + \partial_\zeta(\hat{H} + \hat{H}_v) = 0 \qquad (1)$$

where the independent variable τ is the time and the spatial variables ξ, η, ζ are chosen to map a curvilinear body conforming discretization into a uniform computational space. Here $\hat{Q}$ contains all the dependent variables and $\hat{F}$, $\hat{G}$ and $\hat{H}$ are the inviscid fluxes. The flux terms $\hat{F}_v$, $\hat{G}_v$ and $\hat{H}_v$ contain viscous derivatives and throughout a nondimensional form of the equations is used. The conservative form of the equations is maintained mainly to capture the Rankine Hugoniot shock jump relations as accurately as possible.

For body conforming coordinates and high Reynolds number flow where ζ is the coordinate away from the surface, the thin layer approximation can be made in the ζ direction and the governing equations can be written as:

$$\partial_\tau \hat{Q} + \partial_\xi \hat{F} + \partial_\eta \hat{G} + \partial_\zeta \hat{H} = Re^{-1} \partial_\zeta \hat{S} \quad . \qquad (2)$$

Here the viscous terms in ζ have been collected into the vector $\hat{S}$ and the nondimensional reciprocal Reynolds number is extracted to indicate a viscous flux term.

In differencing these equations it is often advantageous to difference about a known base solution denoted by subscript o as:

$$\delta_\tau(\hat{Q} - \hat{Q}_o) + \delta_\xi(\hat{F} - \hat{F}_o) + \delta_\eta(\hat{G} - \hat{G}_o) + \delta_\zeta(\hat{H} - \hat{H}_o)$$

$$- Re^{-1}\delta_\zeta(\hat{S} - \hat{S}_o) = -\partial_\tau \hat{Q}_o - \partial_\xi \hat{F}_o - \partial_\eta \hat{G}_o - \partial_\zeta \hat{H}_o + Re^{-1}\partial_\zeta \hat{S}_o \qquad (3)$$

where δ indicates a general difference operator, and ∂ is the differential operator. If the base state can be properly chosen, the differenced quantitatives can have smaller and smoother variation and therefore less differencing error. The freestream is used as a base solution in the present formulation.

2. IMPLICIT FINITE DIFFERENCE ALGORITHM

The implicit approximately factored scheme for the thin layer Navier-Stokes equations that uses central differencing in the η and ζ directions and upwinding in ξ is written in the form:

$$[I + h\delta_\xi^b(\hat{A}^+)^n + h\delta_\zeta\hat{C}^n - hRe^{-1}\bar{\delta}_\zeta J^{-1}\hat{M}^n J - D_i|_\eta]$$

$$\times [I + h\delta_\xi^f(\hat{A}^-)^n + h\delta_\eta\hat{B}^n - D_i|_\eta]\Delta\hat{Q}^n = -\Delta t\{\delta_\xi^b|(\hat{F}^+)^n - \hat{F}_\infty^+] + \delta_\xi^f|(\hat{F}^-)^n - \hat{F}_\infty^-] \quad (4)$$

$$+ \delta_\eta(\hat{G}^n - \hat{G}_\infty) + \delta_\zeta(\hat{H}^n - \hat{H}_\infty) - Re^{-1}\bar{\delta}_\zeta(\hat{S}^n - \hat{S}_\infty)\} - D_e(\hat{Q}^n - \hat{Q}_\infty)$$

where $h = \Delta t$ and the freestream base solution is used. Here δ is typically a three point second order accurate central difference operator, $\bar{\delta}$ is the mid-point operator used with the viscous terms, and the operators δ_ξ^b and δ_ξ^f are backward and forward three-point difference operators. The flux $\hat{F}$ has been split into $\hat{F}^+$ and $\hat{F}^-$, according to its positive and negative eigenvalues and the matrices $\hat{A}$, $\hat{B}$, $\hat{C}$ and $\hat{M}$ result from local linearization of the fluxes about the previous time level. Here J denotes the Jacobian of the coordinate transformation. Dissipation operators, D_e and D_i are used in the central space differencing directions.[5] The factored left hand side operators can be readily inverted by sweeping and inversion of tridiagonal matrices with 5 × 5 blocks. This two factor implicit scheme is readily vectorized or multi-tasked in planes of ξ = constant.

3. COMPOSITE GRID SCHEME

In the present work, a composite grid scheme[14] has been used where a large single grid is split into a number of smaller grids so that computations can be performed on each of these grids separately. Each of tnese grids use the available core memory in turn, while the rest are stored on an external disk storage device such as the SSD of the Cray X-MP/48 computer. On Cray 2 super-computer, large in core memory is available to fit the large single grid. However, for accurate geometric modeling of complex projectile configurations that include blunt nose, sharp base corner and base cavities etc., it is also desirable to split the large data base into few smaller zones on Cray 2 as well.

A code developed for a single grid can be made to work for a block grid structure by: 1) mapping and storing the information for each grid onto a large memory; and 2) supplying interface boundary arrays, pointers and updating procedures. Consider the situation in Figure 1 in which the single grid from J = 1, J_{max} is partitioned into four grids, G1 through G4. The base region of the projectile is included by adding another zone G5. This procedure preserves the actual base corner and no approximation is made. This zonal scheme has been modified to allow more than one zone in the wake for accurate modeling of other complicated base configurations including cavities, etc.

The use of a composite or blocked grid scheme requires special care in storing and fetching the interface boundary data, i.e., the communication between the various zones. For the simple partitioning shown in Figure 1, all subgrid points are members of the original grid. There is no mismatch of the

grid points at the interface boundaries and no interpolations are required. This procedure thus, has the advantage over-patched or overset grid schemes which do need interpolations. The partitioned grid has six interface boundaries, $J_1 = J_{1_{max}}$, $J_2 = 1$, $J_2 = J_{2_{max}}$, $J_3 = 1$, $J_3 = J_{3_{max}}$ and $J_4 = 1$ in the streamwise direction and two interface boundaries in the normal direction between grids G4 and G5. Data for these planes are to be supplied from the other grids by injecting interior values of the other grid onto the interface boundaries. The details of the data storage, transfer and other pertinent informations such as metric and differencing accuracy can be found in Reference 14.

The differencing accuracy near the interfaces is quite important. Three point backward and forward difference operators are used at the interior points. Near the interface, for example, at $J_2 = J_{2_{max}} - 1$ three point forward difference operator cannot be used with one grid point overlap as shown in Figure 1. The differencing accuracy can be dropped from second order to first order; however, this leads to inaccuracies in the flowfield solution near the interfaces.[14] To maintain second order accuracy near the interfaces, we difference, for example, $\frac{\partial F}{\partial \xi}$ at $J_2 = J_{2_{max}} - 1$ as,

$$\frac{\partial F}{\partial \xi} = \partial_\xi^b(F^+) + \partial_\xi^f(F^-)$$

where ∂_ξ^b is the usual three point backward difference operator and δ_ξ^f is now a central difference operator, i.e.,

$$\frac{\partial F}{\partial \xi} = \frac{3F_j^+ - 4F_{j-1}^+ + F_{j-2}^+}{2\Delta\xi} + \frac{F_{j+1}^- - F_{j-1}^-}{2\Delta\xi} \quad .$$

Near the other interface of grid 2 ($J_2 = 2$), the ∂_ξ^b operator is correspondingly replaced by a central difference operator while ∂_ξ^f is a usual three point forward difference operator. The planes $J_2 = 1$ and $J_2 = J_{2_{max}}$ are, of course, boundaries for grid 2 and get their data from interior flowfield solutions from neighboring grids. Second order accuracy at and near the interfaces is thus maintained. Partial use of central differencing near the interfaces has not adversely affected the stability of the scheme.

III. MODEL AND COMPUTATIONAL GRIDS

The first model used for the experiment and computational study presented here is an idealization of a realistic artillery projectile geometry. The experimental model shown in Figure 2 is a secant-ogive-cylinder-boattail (SOCBT) projectile. It consists of a three-caliber (one-caliber = maximum body diameter), sharp, secant-ogive nose, a two-caliber cylindrical mid-

section and a one-caliber 7° conical afterbody or boattail. A similar model was used for the computational studies with the only difference being a five percent rounding of the nose tip. The nose tip rounding was done for computational efficiency and is considered to have little impact on the final integrated forces. Experimental pressure data[15] are available for this shape and were obtained in the NASA Langley eight foot Pressure Tunnel using a sting mounted model. The test conditions of 1 atm supply pressure and 320 K supply temperature resulted in a Reynolds number of 4.5×10^6 based on model length.

The computational grid used for this computation is shown in Figure 3. Figure 3a shows the longitudinal cross section of the 3D grid while Figure 3b shows an expanded view of the three-dimensional base region grid. As shown in Figure 3a, the clustering of grid points near the body surface is done to resolve the viscous boundary layer near the body surface. Grid clustering has also been used in the longitudinal direction near the boattail and the base corners where large gradients in the flow variables are expected. In addition, the composite grid scheme preserves the sharp base corner. The grid consists of 202 points in the streamwise direction, 36 points in the circumferential direction and 50 points in the normal direction. This amounts to about 16 million words of storage for the code on the Cray X-MP/48. Only up to 4 Mw of central core memory was easily accessible; therefore, the full grid was partitioned into five smaller grids (including a base region grid) each of which would use the core memory in turn while the rest is stored on the SSD device. These computations were performed on the Cray X-MP/48 at the US Army Ballistic Research Laboratory (BRL). Each numerical simulation (includes all partitioned grids) took over 20 hours of computer time.

Another grid shown in Figure 4 was obtained to simulate the model with the sting in the base region again for the SOCBT projectile. This is again a longitudinal cross-section of the 3-D grid. The grid is wrapped around the base corner in this case. It consists of 238 points in the axial direction, 39 points in the circumferential direction and 50 points in the normal direction. Computations on this grid were performed on the Cray 2 computer at BRL using the same code. The computing time for these simulations was comparable to that on the Cray X-MP/48.

The second projectile under consideration is the M549 projectile shown in Figure 5. This projectile has a short boattail of about 1/2 a caliber in length. For simplicity, the flat nose was again modeled with nose tip rounding and the rotating band was eliminated. Experimental aerodynamic coefficient data are available for this configuration which is a compilation of the wind tunnel and free flight range data.[16] Computations for this projectile have been made for atmospheric conditions. Figure 6 shows an expanded view of the grid around this projectile and shows both the wind side and lee side planes. The full grid consists of 298 points in the axial direction, 39 points in the circumferential direction and 50 points in the normal direction. Calculations for this projectile were performed on the Cray 2 computer at BRL. Each of these calculations took over 30 hours of computer time.

IV. RESULTS

The implicit time marching procedure was used to obtain the desired steady state result starting from initial freestream conditions everywhere. Boundary

conditions were updated explicitly at each time step. The solution residual dropped at least three orders of magnitude before converged solutions were obtained. In addition, the surface pressure distribution was checked for time invariance. For the computation of turbulent flow, a turbulence model must be supplied. In the present calculation, a two layer algebraic eddy viscosity model due to Baldwin and Lomax[17] was used. Results are now presented for two cases: (1) SOCBT projectile with and without sting; and (2) M549 projectile.

1. SOCBT PROJECTILE, $0.9 < M_\infty < 1.2$, $\alpha = 4$

Results have been obtained at various transonic speeds for both cases with and without modeling of the sting. Figures 7-10 show the Mach contours for the projectile in the windward and leeward planes. These figures show the expansions at the ogive-cylinder and cylinder-boattail corners. These figures indicate the presence of shock waves on the cylinder and also on the boattail which typically occur on the projectile at transonic speeds. Sharp shocks are clearly observed on the boattail flowfield which are asymmetrically located (the one on the wind side being closer to the base than its counterpart on the lee side). The asymmetry can also be seen in the wake flow behind the bluff base. As the Mach number is increased from 0.94 to 0.96 and then to 0.98, the shocks become stronger and move towards the base of the projectile. At higher transonic speeds past the speed of sound (see Figure 10), these shocks become weak; however, a bow shock forms in front of the nose of the projectile.

Computations have also been made to investigate the effect of the sting on transonic projectile aerodynamics. A typical plot of Mach contours for this simulation is shown in Figure 11a for $M_\infty = 0.96$ and $\alpha = 4°$. As expected, the sting has a large effect on the qualitative features of the wakefield which in turn has moved the boattail shocks further upstream from the base corner. Experimentally obtained shadowgraph picture at the same Mach number and flow conditions is shown in Figure 11b. As shown in Figures 11a and 11b, the agreement of the shock wave positions between the computation and experiment is very good. Figures 12a and 12b shows the velocity vectors in the base region for both windside and leeside. Figure 12a is for the case with no sting whereas Figure 12b includes the sting in the base region. In both cases, asymmetry in the flowfield can be observed between the windside and leeside. Three pairs of separated flow bubbles can be seen in the near wake for the case of no sting (Figure 12a). For the case with sting (Figure 12b), one can see the large primary bubble along with a counter rotating small bubble near the junction of sting and the base. The primary bubble is more elongated on the windside and the flow reattaches further downstream of the base.

Figures 13-15 show the surface pressure distributions as a function of the longitudinal position and are compared with experimental data.[15] Figures 13a and 13b show the comparison at $M_\infty = 0.96$ for windside and leeside, respectively. Computed results are shown for two grids, one which wrapped around the base corner and the other which did not. As shown in these figures, the computed results are virtually the same for both computations except near the base corner where a small difference can be noticed. The agreement with experimental data however, is very good for both windside and leeside. The expansions and recompressions near the ogive-cylinder and cylinder-boattail junctions are captured very accurately. Figures 14a and 14b shows the surface pressure distribution for $M_\infty = 0.98$. Computed results are shown for both

cases with and without sting for turbulent flow. In the experiment,[15] the model was sting mounted and no boundary layer trip was used. Therefore, it is not clear if the flow was laminar or turbulent. Computed results were obtained for the sting mounted case for laminar flow condition and is also included in the comparison with experimental data. Comparison of pressure on the windside (Figure 14a) shows generally good agreement of the computed pressures with the experimental data. The largest differences between the computed results are seen on the rear part of the boattail where no experimental results are available. The comparison on the leeside (Figure 14b) again shows good agreement of the computed results with experimental data for most of the projectile except on the second half of the boattail. As expected, the computed result with no sting has the largest discrepancy. Computed results with sting simulation compare well with experimental data especially for laminar flow conditions. A typical result at a high transonic speed $M_\infty = 1.1$ is shown in Figure 15. The agreement of the computed surface pressures with experiment is very good. At this high transonic Mach number, the shocks on the cylinder as well as on the boattail are very weak as evidenced by the absence of sharp rise in pressure in those areas. The expansions and recompressions near the ogive-cylinder and cylinder-boattail junctions can be clearly observed in Figure 15.

The computed surface pressures have been integrated to obtain the aerodynamic forces and moments. The slope of the pitching moment coefficient (C_{m_α}) is generally of greater concern in projectile aerodynamics since it is the parameter that determines the static stability of the projectile. Figure 16 shows the variation of the slope of the pitching moment coefficient with Mach number. It clearly shows the critical aerodynamic behavior in the transonic speed regime, i.e., the sharp rise in the coefficient between M = 0.92 and 0.96 and its subsequent sharp drop. This is followed by a smooth decrease in the coefficient as Mach number is increased further. The increase in C_{m_α} between M = 0.92 and 0.96 is of the order of 20% which is a typical value obtained from a number of range tests for similar projectiles.

2. M549 PROJECTILE, $0.7 < M_\infty < 1.5$, $\alpha = 2$

Numerical computations were made for the M549 projectile at various transonic speeds $0.7 < M_\infty < 1.5$ and at angle of attack, $\alpha = 2°$. Qualitative features of the flowfield obtained from some of these calculations are shown in Figures 17-21 where Mach number contours have been plotted for M = 0.85, 0.90, 0.92, 0.94, and 0.98 for both windward and leeward planes. The asymmetry in the wake region flow is obvious from these figures. These figures indicate the development and asymmetric locations of shock waves on the projectile at transonic speeds. At low transonic speeds, for example, at M = 0.85 (Figure 17) the shock waves are just beginning to form especially near the boattail junction. As Mach number is increased to M = 0.90, the shocks are already formed on the projectile both near the cylinder as well as boattail junctions. The flow expansions at these junctions can also be clearly seen in this figure. The small asymmetry in shockwave locations can be observed particularly with the boattail shocks. The shockwave on the boattail in the windside is a little closer to the base than its counterpart in the leeside. In addition, these shocks have moved little downstream from the boattail junction. As shown in Figures 19-21, with further increase in Mach

number to 0.92, 0.94, and 0.98, the shockwaves (both on the cylinder and the boattail) become stronger and gradually move downstream. The asymmetry in the location of the shock waves can be seen more and more clearly. As seen in Figure 21 for M = 0.98, the shockwaves pattern is complicated and the boattail shocks are located very close to the base corners.

The static aerodynamic coefficients have been obtained from the computed flowfields. As pointed out earlier, the slope of moment coefficient (C_{m_α}) is of greater concern. Figure 22 shows the development of C_m over the projectile for various transonic speeds. Actually, it is the accumulative moment coefficient referenced to the nose and thus, the value at the end (X/D = 5.645) is the final result. The difference in this coefficient over the nose portion is practically negligible for all transonic Mach numbers. The largest effect is seen on the cylinder and boattail sections. The boattail has a dramatic effect as evidenced by the sharp rise in all the curves. Figure 23 shows the C_{m_α} comparison for the computation and the experimental data.[16] Here C_{m_α} is referenced to center of gravity (C.G.) of the projectile. One can clearly see the sharp rise in C_{m_α} between M = 0.8 to 0.94 which is followed by the sharp drop with further increase in Mach number in both the computation and the experimental data. This critical aerodynamic behavior observed in the experimental data is clearly predicted in the computations and good agreement has been obtained between the computed result and the data.

V. CONCLUDING REMARKS

In conjunction with a new Navier-Stokes code, a simple composite grid scheme has been developed which allows fine computational grids needed for accurate transonic flow computations to be obtained on CRAY X-MP/48 or Cray 2 computers. The numerical method uses an implicit, approximately factored, partially upwind (flux-split) algorithm.

The three dimensional transonic flowfield computations have been made over two projectiles for different flow conditions and angle of attack. The computed flowfields show the development of the asymmetrically located shockwaves on the projectile at various transonic speeds. For the SOCBT projectile, computed surface pressures have been compared with experimental data and are found to be in good agreement. The slope of the pitching moment coefficient (C_{m_α}), determined from the computed flowfields, shows the critical aerodynamic behavior. For M549 projectile, computed C_{m_α} has been compared with experimental data. It shows the same critical behavior in the data and the agreement between the computed result and experimental data is good.

The results of this research provide the basis for a new capability to compute three dimensional transonic flowfields over projectiles. This capability in conjunction with the supercomputers at the US Army Ballistic Research Laboratory has led to the first successful prediction of the critical aerodynamic behavior in C_{m_α} of artillery shell at transonic speeds. The next

step is the numerical prediction of magnus force and moment for spinning projectiles at angle of attack which involves calculations of the full three dimensional transonic flowfields.

REFERENCES

1. Deiwert, G.S., "Numerical Simulation of Three Dimensional Boattail Afterbody Flowfield," AIAA Journal, Vol. 19, May 1981.

2. Nietubicz, C.J., et al, "Computations of Projectiles Magnus Effect at Transonic Velocities, AIAA Journal, Vol. 23, No. 7, July 1985.

3. Sahu, J., Nietubicz, C.J. and Steger, J.L., "Navier-Stokes Computations of Projectile Base Flow With and Without Base Injection," ARBRL-TR-02532, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, November 1983. (AD A135738) (See also AIAA Journal, Vol. 23, No. 9, September 1985, pp. 1348-1355.)

4. Sahu, J., "Three Dimensional Base Flow Calculation for a Projectile at Transonic Velocity," AIAA Paper No. 86-1051, May 1986.

5. Ying, S.X., Steger, J.L., Schiff, L.B. and Baganoff, D., "Numerical Simulation of Unsteady, Viscous, High-Angle-of-Attack Flows Using a Partially Flux-Split Algorithm," AIAA Paper No. 86-2179, August 1986.

6. Beam, R. and Warming, R.F., "An Implicit Factored Scheme for the Compressible Navier-Stokes Equations," AIAA Paper No. 85-1815-CP, August 1985.

7. Steger, J.L., "Implicit Finite Difference Simulation of Flow About Arbitrary Geometries with Application to Airfoils," AIAA Journal, Vol. 16, No. 4, July 1978, pp. 679-686.

8. Pulliam, T.H. and Steger, J.L., "On Implicit Finite-Difference Simulations of Three-Dimensional Flow, AIAA Journal, Vol. 18, No. 2, February 1980, pp. 159-167.

9. Steger, J.L. and Warming, R.F., "Flux Vector Splitting of the Inviscid Gasdynamic Equations with Application to Finite-Difference Methods," Journal of Computational Physics, Vol. 40, No. 2, 1981, pp. 263-293.

10. Lomax, H. and Pulliam, T.H., "A Fully Implicit Factored Code for Computing Three Dimensional Flows on the ILLIAC IV, Parallel Computations," G. Rodrigue Ed., Academic Press, New York, 1982, pp. 217-250.

11. Deiwert, G.S. and Rothmund, H., "Three-Dimensional Flow Over a Conical Afterbody Containing a Centred Propulsive Jet: A Numerical Simulation," AIAA Paper No. 83-1709, 1983.

12. Benek, J.A., et al, "A 3-D Chimera Grid Embedding Technique," AIAA Paper No. 85-1523, July 1985.

13. Belk, D.M. and Whitfield, D.L., "Three-Dimensional Euler Solutions on Blocked Grids Using an Implicit Two-Pass Algorithm," AIAA Paper No. 87-0450, January 1987.

14. Sahu, J. and Steger, J.L., "Numerical Simulation of Three Dimensional Transonic Flows," AIAA Paper No. 87-2293, August 1987.

15 Kayser, L.D. and Whiton, F., "Surface Pressure Measurements on a Boat-tailed Projectile Shape at Transonic Speeds," ARBRL-MR-03161, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, March 1982. (AD A113520)

16. Kline, R., Herrman, W.R., and Oskay, V., "A Determination of the Aerodynamic Coefficients of the 155mm, M549 Projectile," Technical Report 4764, Picatinny Arsenal, Dover, New Jersey, November 1974.

17. Baldwin, B.S. and Lomax, H., "Thin Layer Approximation and Algebraic Model for Separated Turbulent Flows," AIAA Paper No. 78-257, 1978.

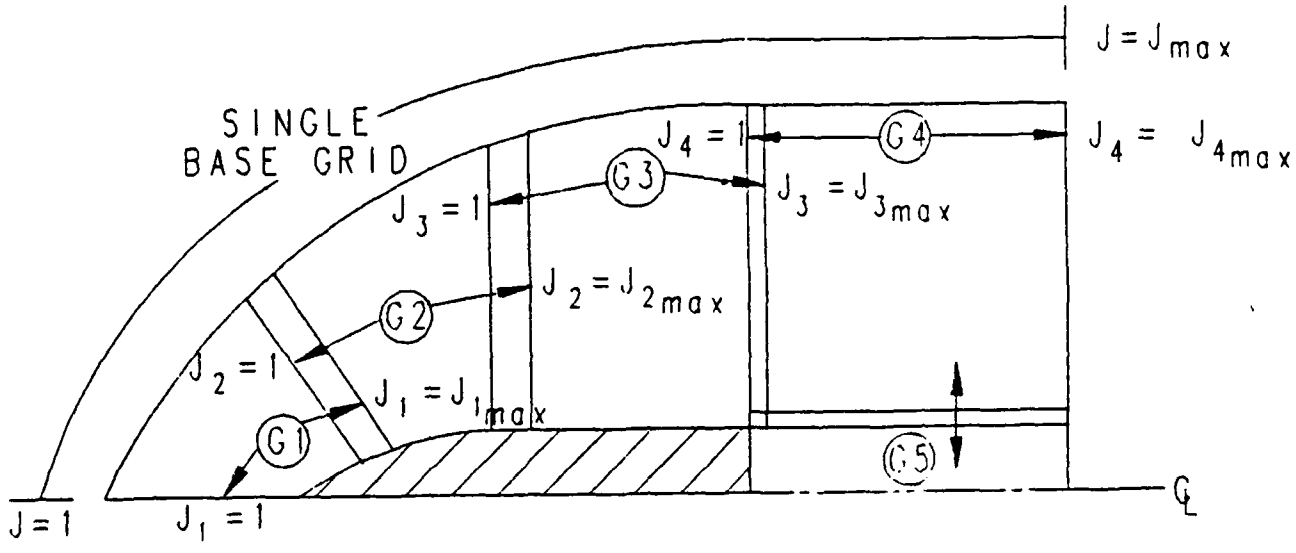
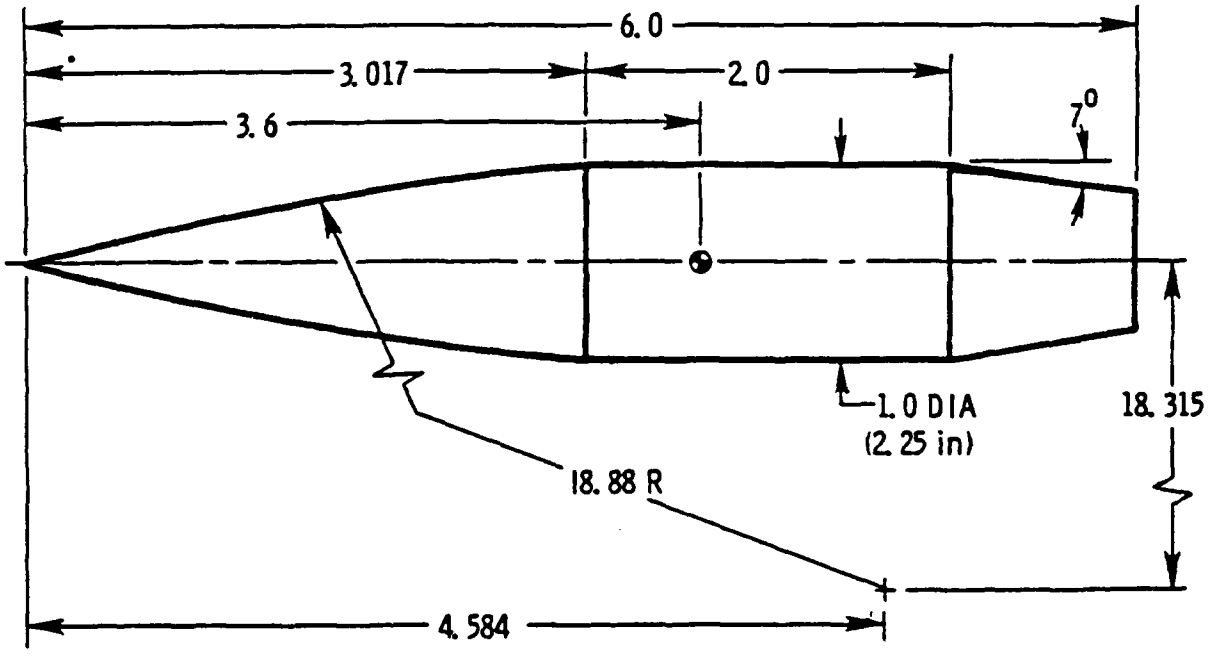


Figure 1. Schematics of grid partitioning.



ALL DIMENSIONS IN CALIBERS

Figure 2. Model geometry of the SOCBT projectile.

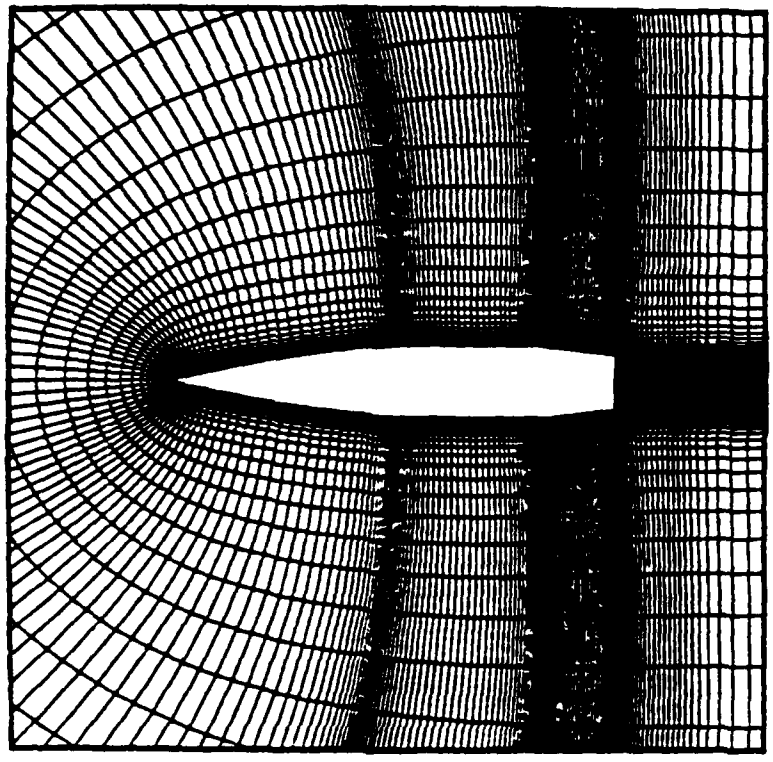


Figure 3a. Longitudinal cross-section of the 3D grid.

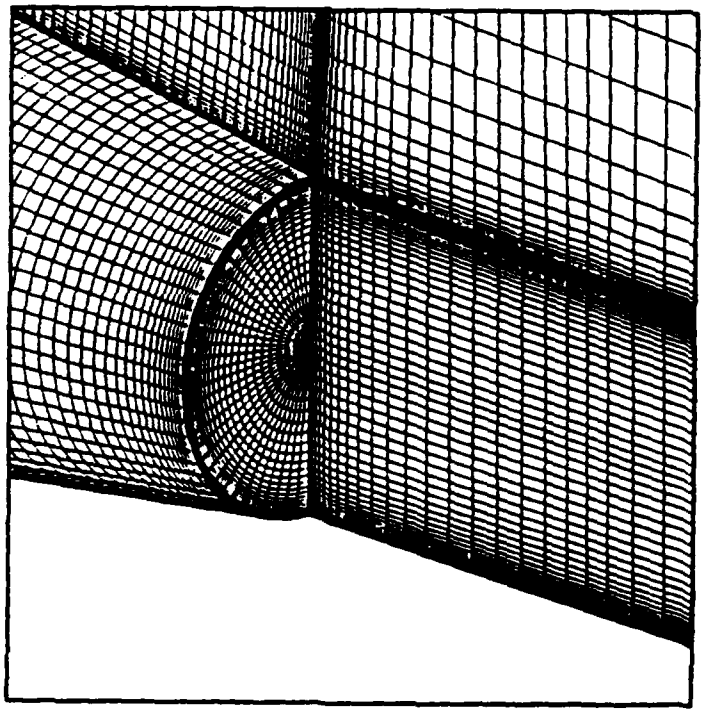


Figure 3b. Expanded view of the base region grid.

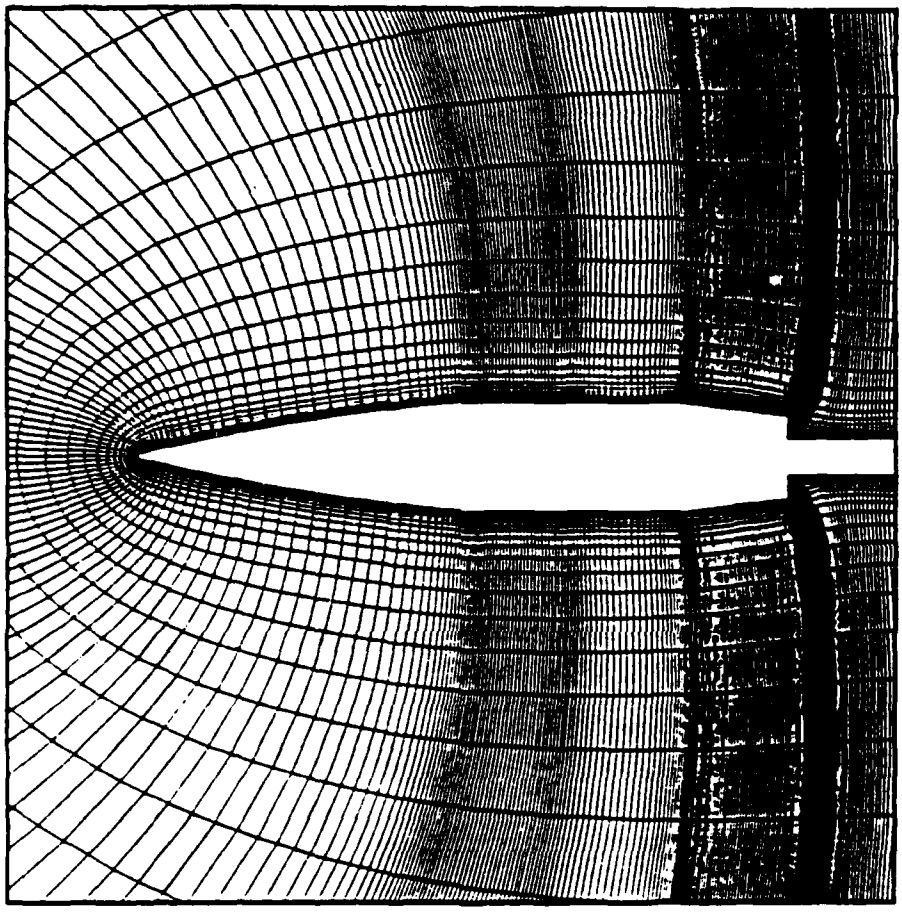


Figure 4. Grid for the sting mounted SOCBT projectile.

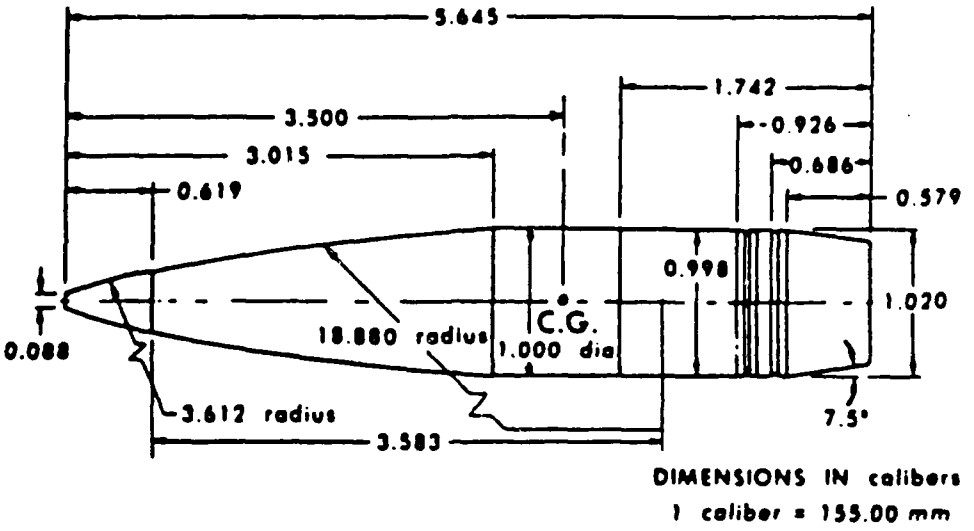


Figure 5. M549 projectile.

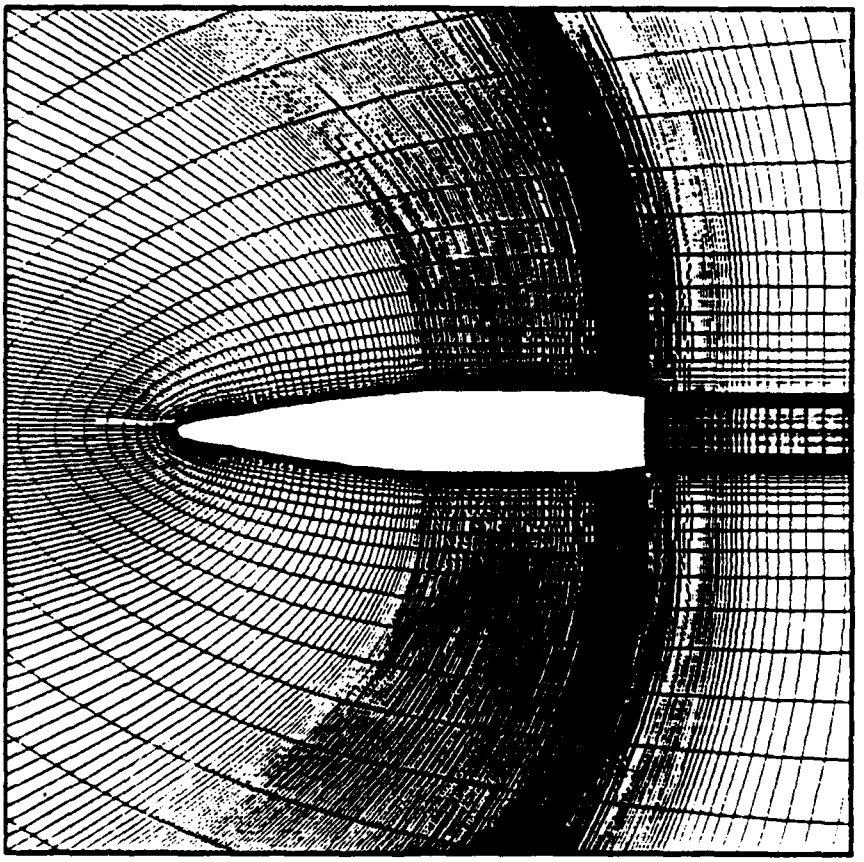


Figure 6. Longitudinal cross-section of the grid for M549 projectile.

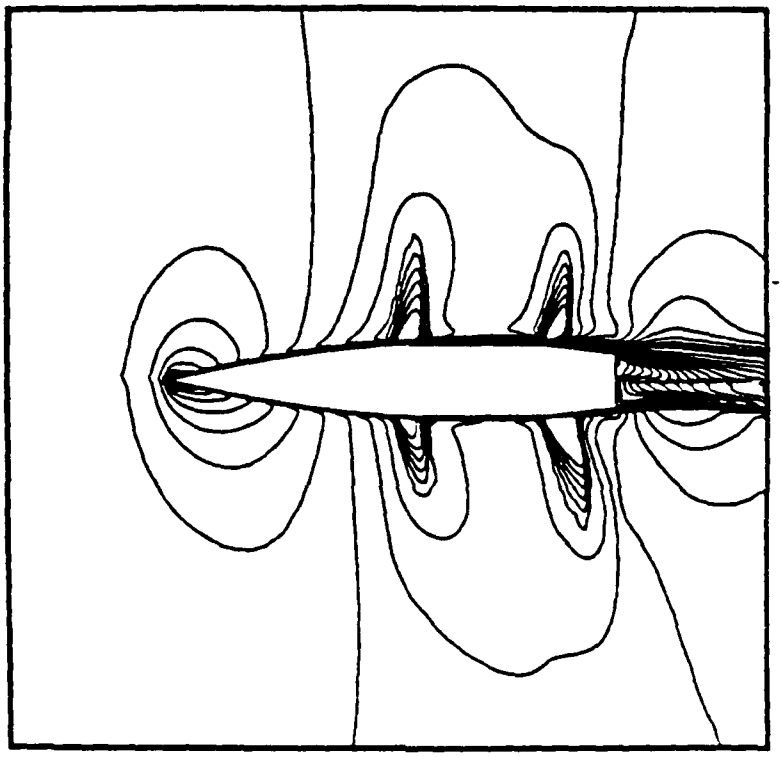


Figure 7. Mach contours, SOCBT projectile, $M_\infty = 0.94$, $\alpha = 4°$.

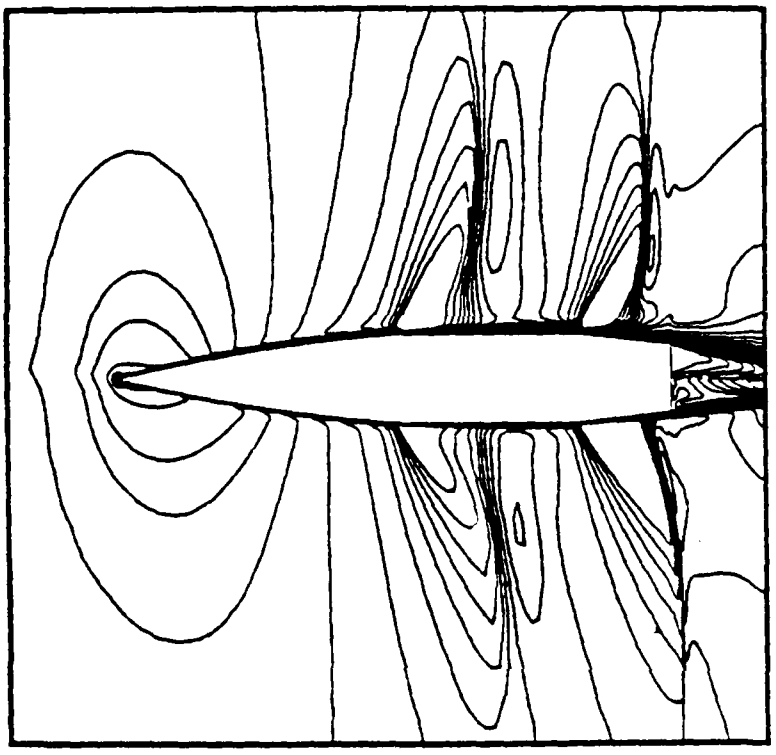


Figure 8. Mach contours, SOCBT projectile, $M_\infty = 0.96$, $\alpha = 4°$.

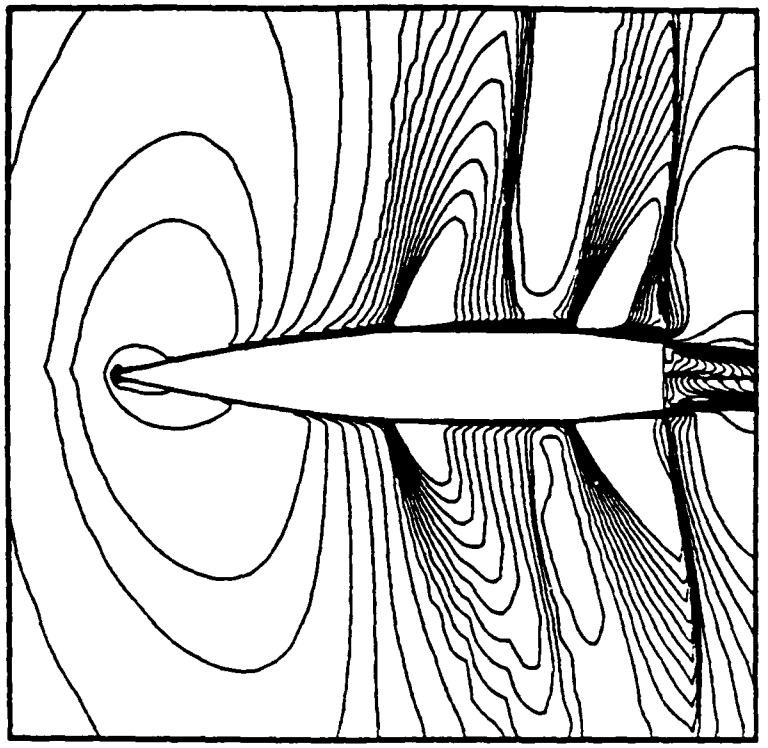


Figure 9. Mach contours, SOCBT projectile, $M_\infty = 0.98$, $\alpha = 4°$.

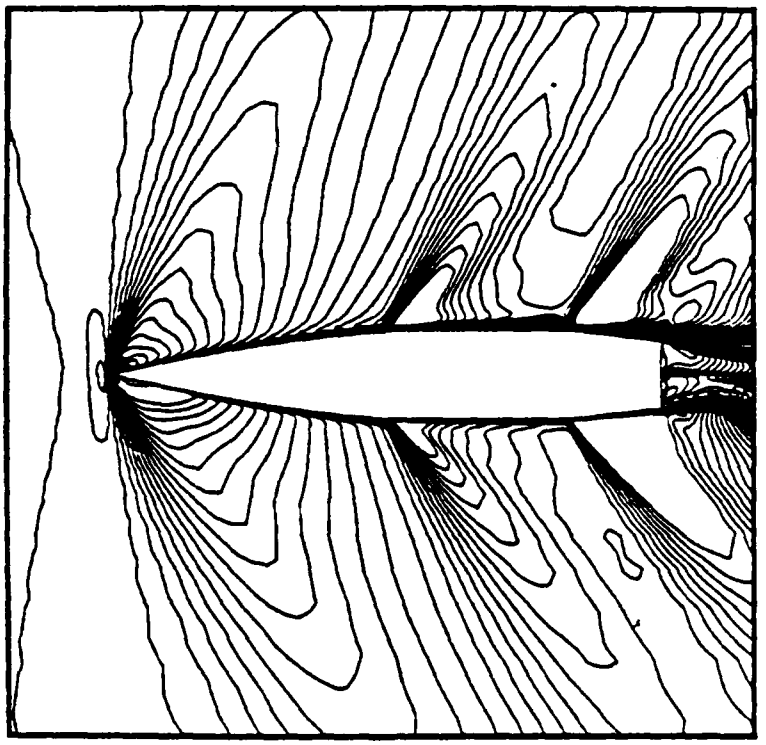


Figure 10. Mach contours, SOCBT projectile, $M_\infty = 1.1$, $\alpha = 4°$.

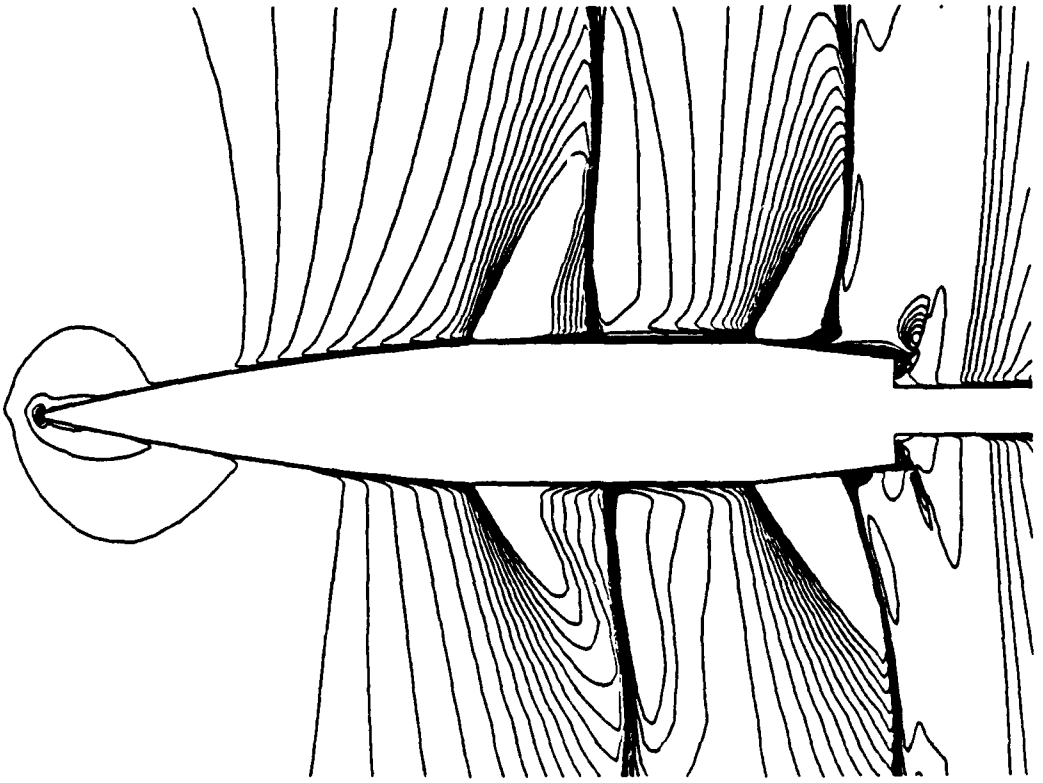


Figure 11a. Computed Mach contours, $M_\infty = 0.96$, $\alpha = 4°$, SOCBT projectile (with sting).

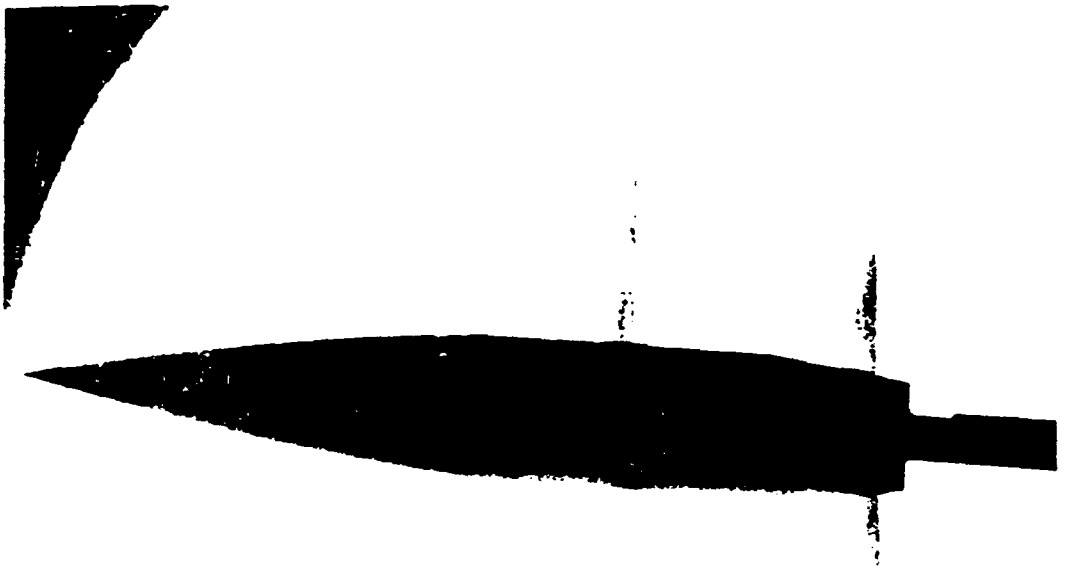


Figure 11b. Experimental shadowgraph, $M_\infty = 0.96$, $\alpha = 4°$, SOCBT projectile (with sting).

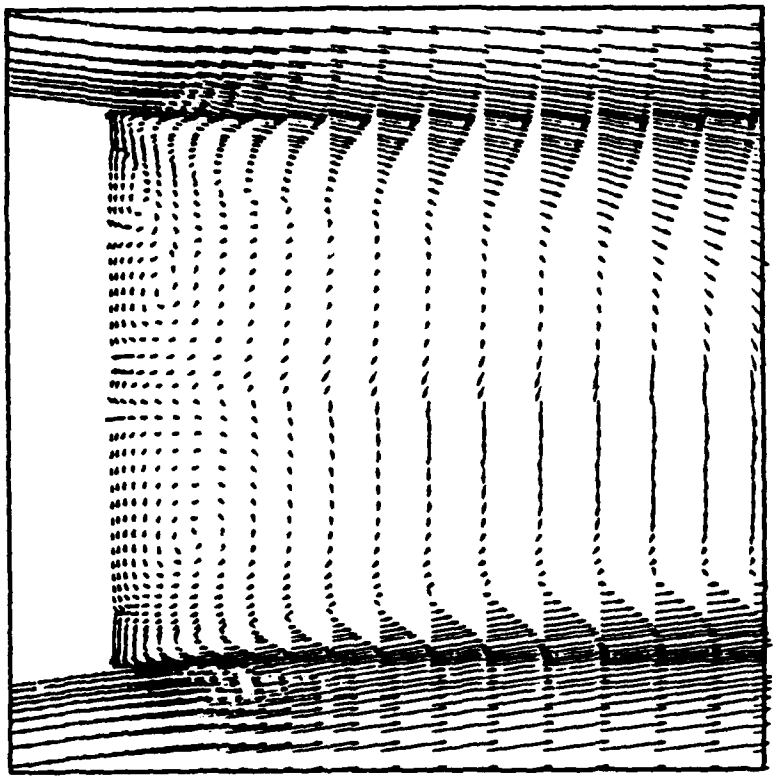


Figure 12a. Velocity vectors in the base region, $M_\infty = 0.96$, $\alpha = 4°$, SOCBT projectile (without sting).

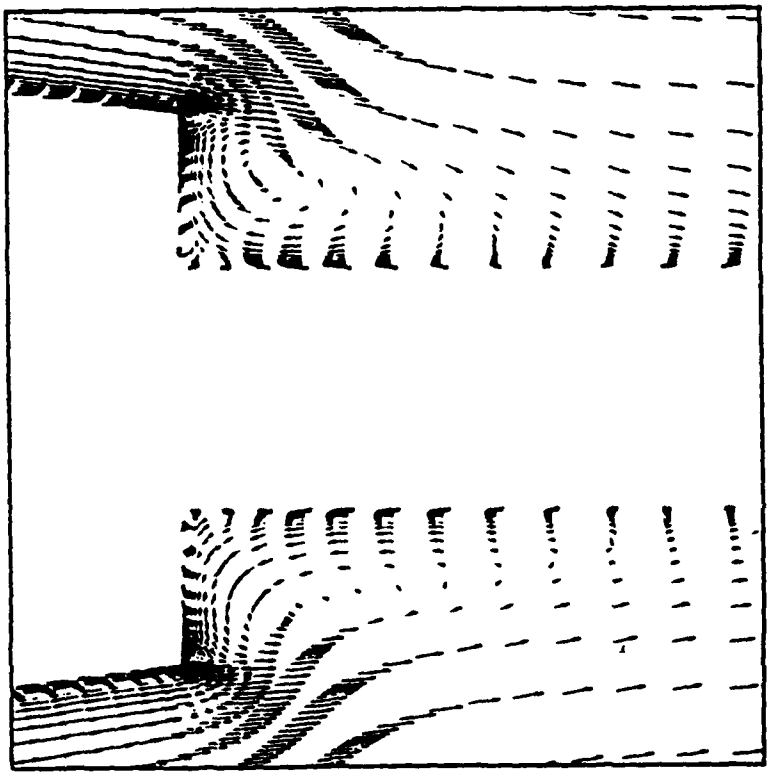


Figure 12b. Velocity vectors in the base region, $M_\infty = 0.96$, $\alpha = 4°$, SOCBT projectile (with sting).

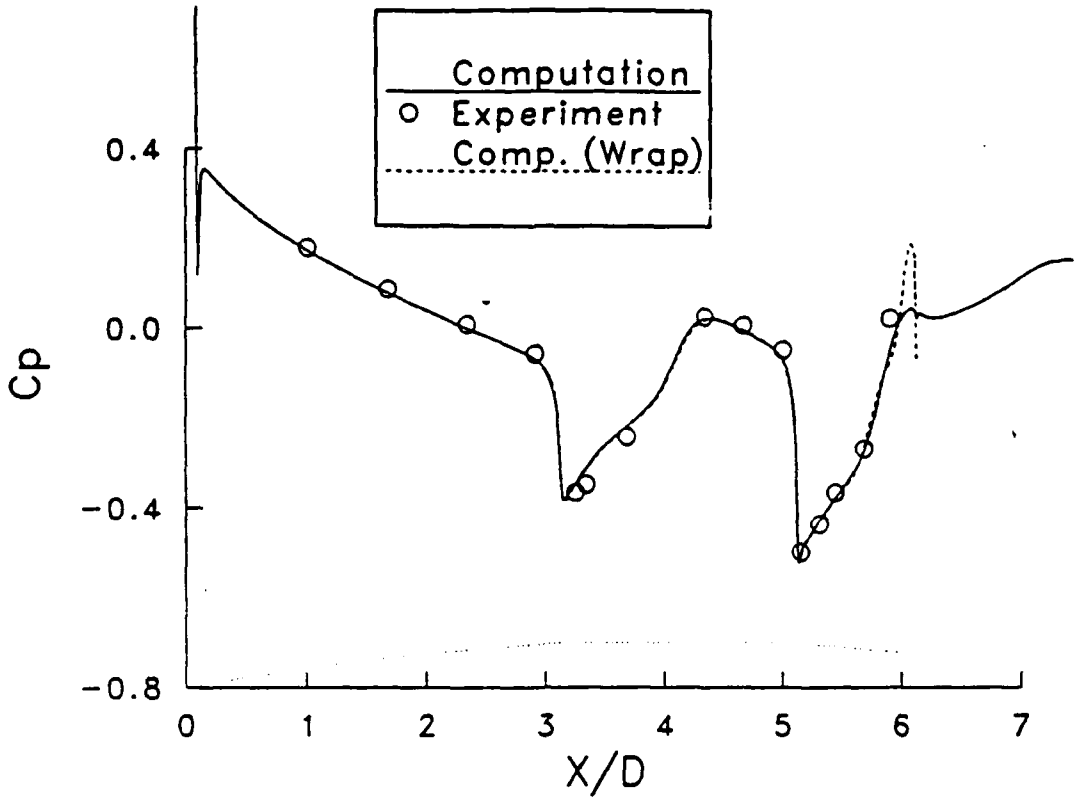


Figure 13a. Longitudinal surface pressure distribution, SOCBT projectile, $M_\infty = 0.96$, $\alpha = 4°$, windside.

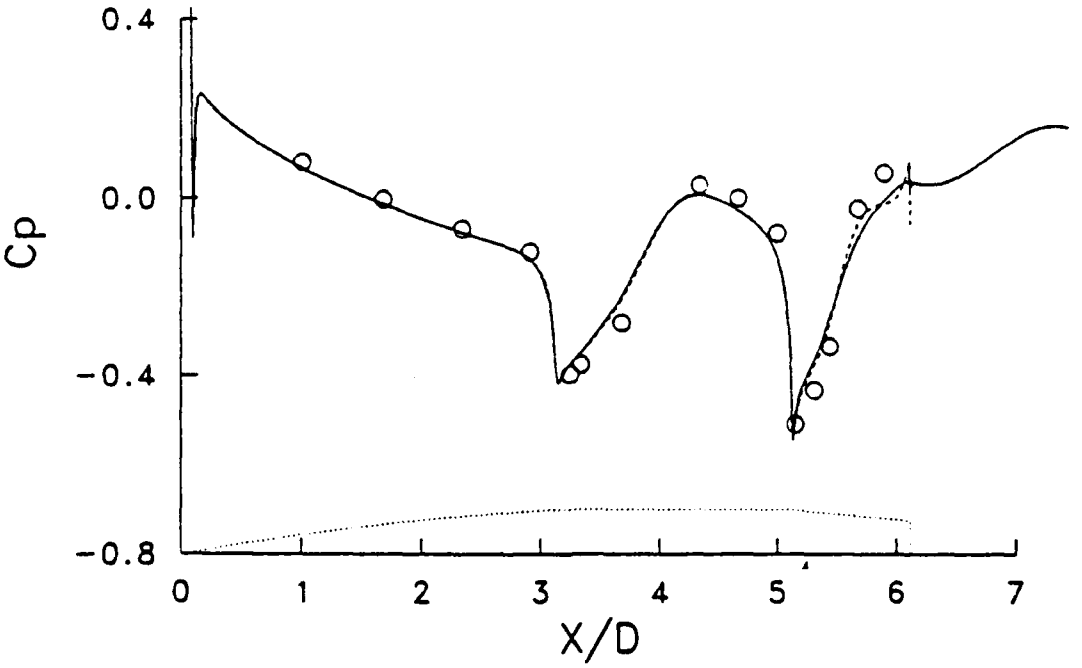


Figure 13b. Longitudinal surface pressure distribution, SOCBT projectile, $M_\infty = 0.96$, $\alpha = 4°$, leeside.

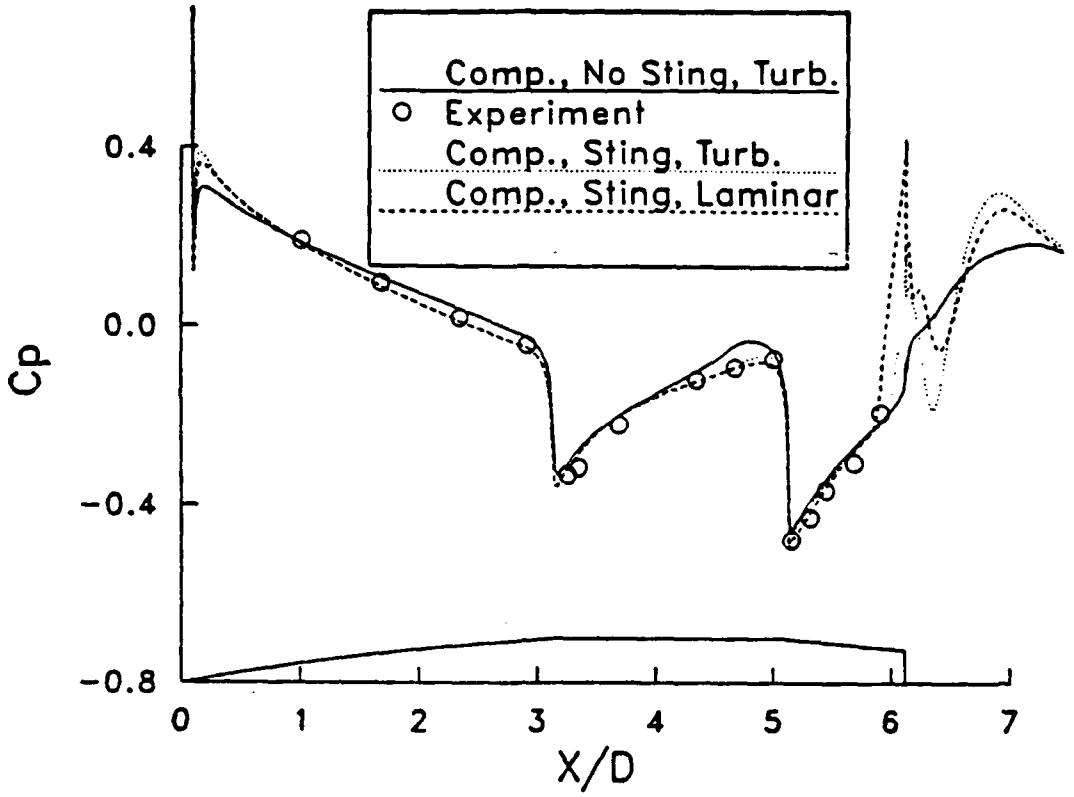


Figure 14a. Longitudinal surface pressure distribution, SOCBT projectile, $M_\infty = 0.98$, $\alpha = 4°$, windside.

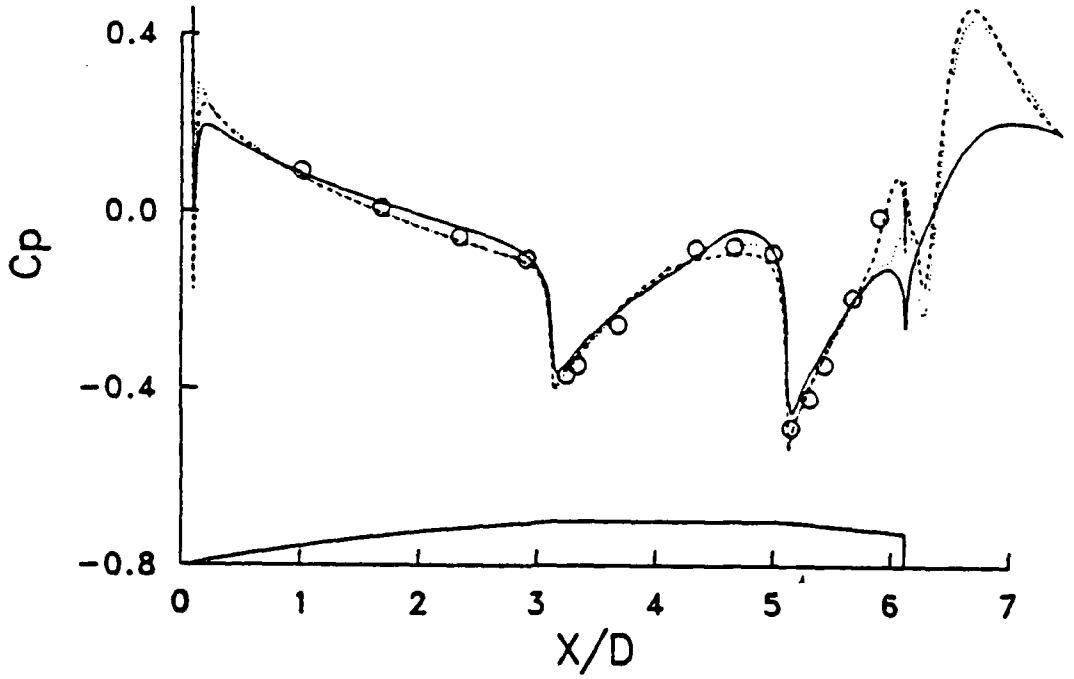


Figure 14b. Longitudinal surface pressure distribution, SOCBT projectile, $M_\infty = 0.98$, $\alpha = 4°$, leeside.

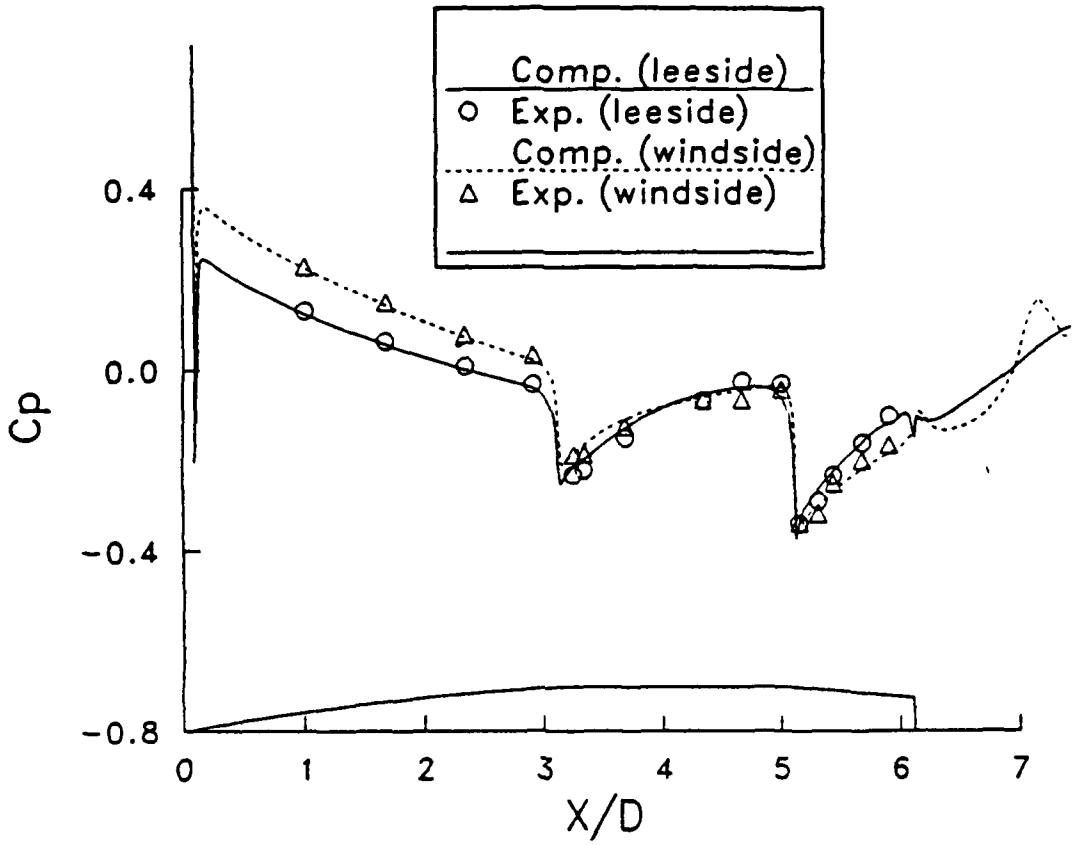


Figure 15. Longitudinal surface pressure distribution, SOCBT projectile, M_∞ - 1.1, α = 4°.

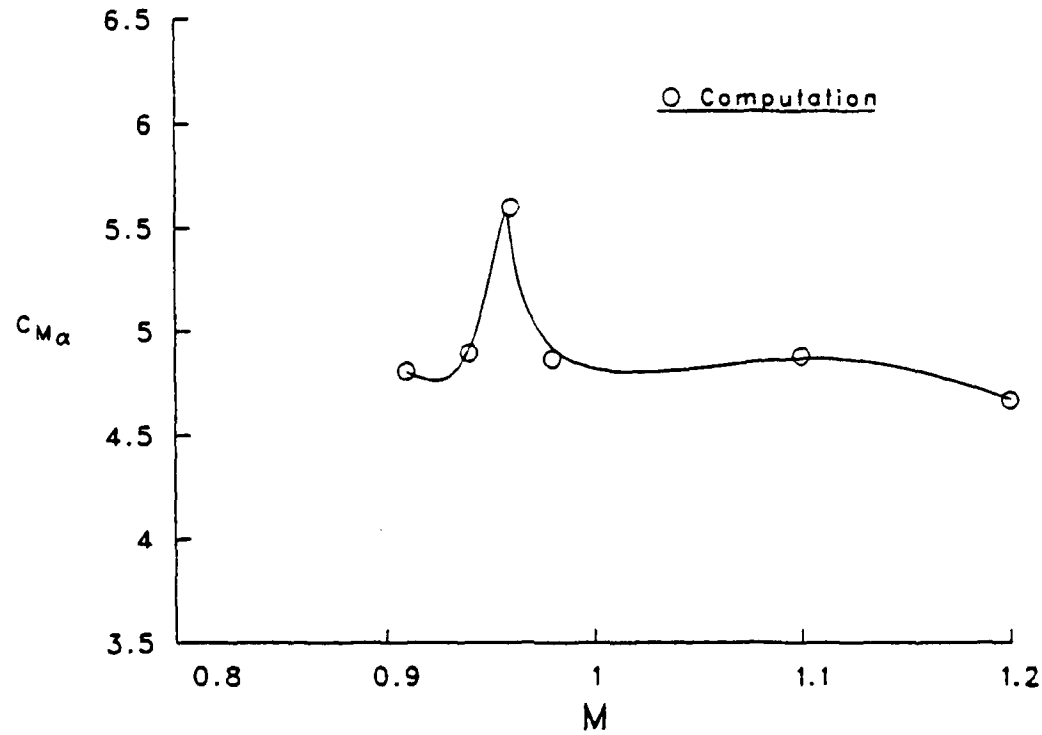


Figure 16. Slope of pitching moment coefficient, C_{m_α}, vs Mach number, SOCBT projectile.

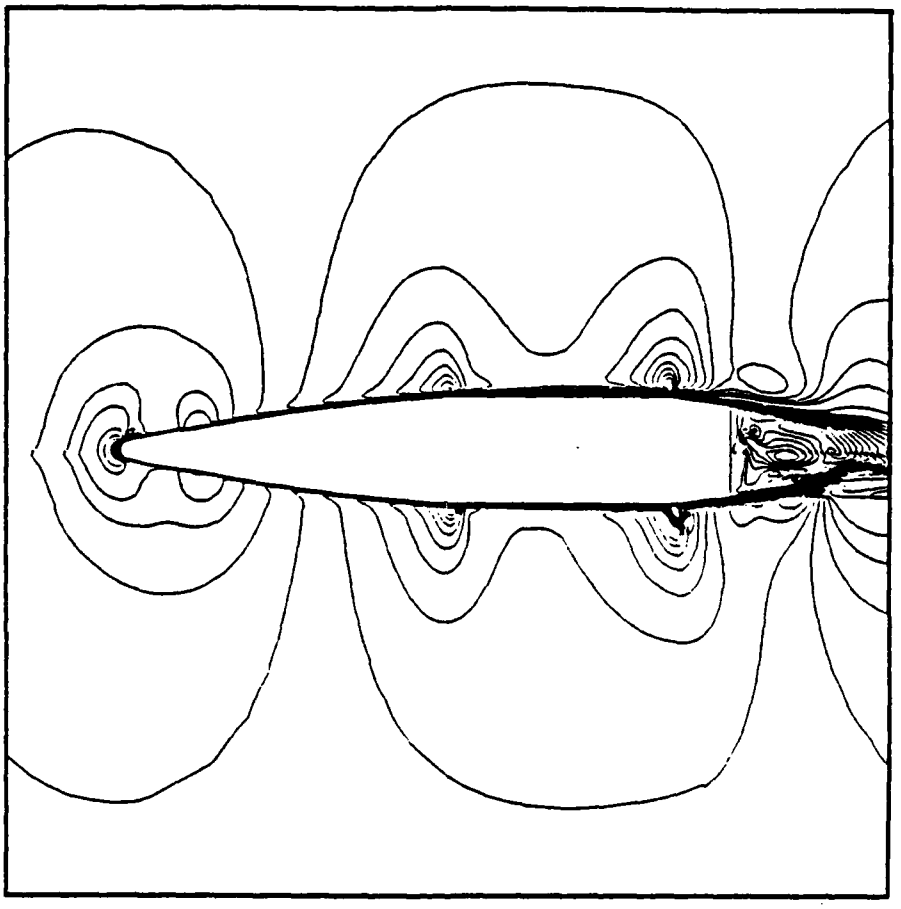


Figure 17. Mach contours, M549 projectile, $M_\infty = 0.85$, $\alpha = 2°$.

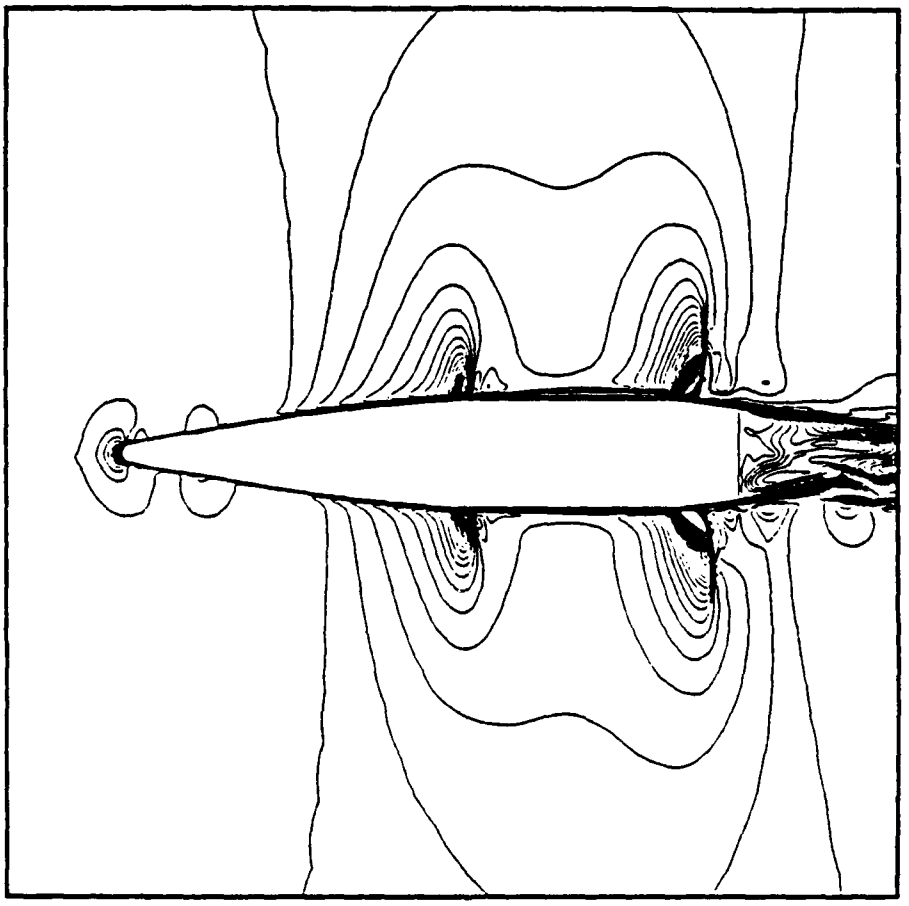


Figure 18. Mach contours, M549 projectile, $M_\infty = 0.90$, $\alpha = 2°$.

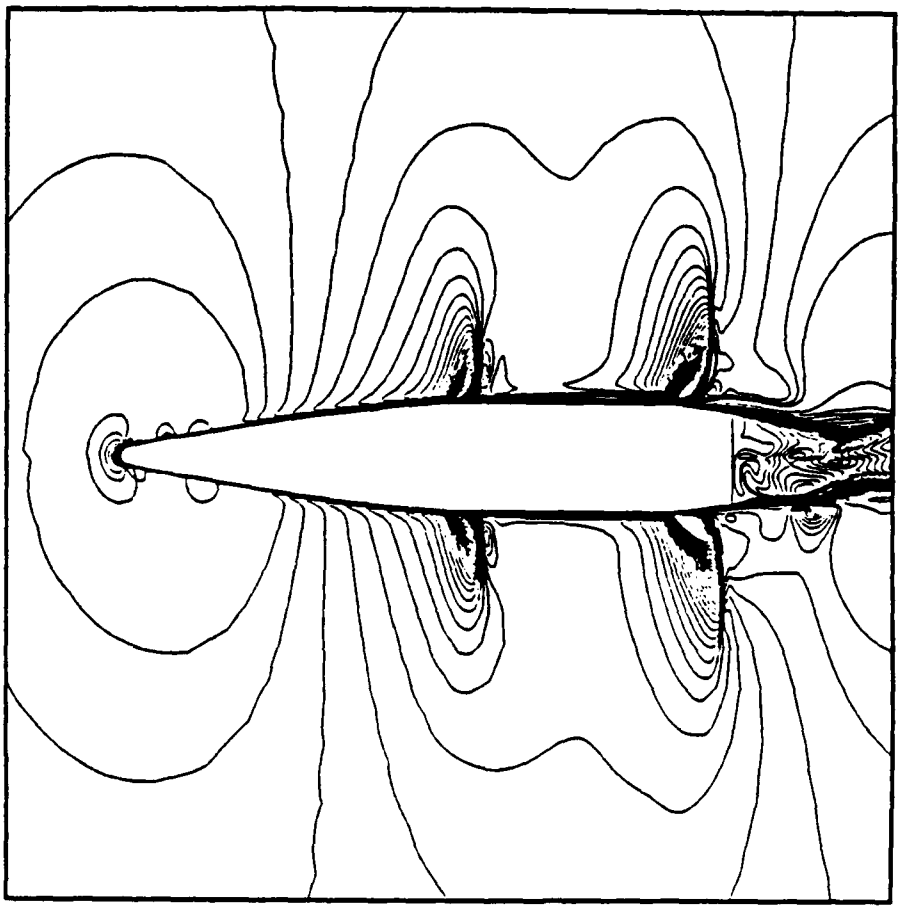


Figure 19. Mach contours, M549 projectile, $M_\infty = 0.92$, $\alpha = 2°$.

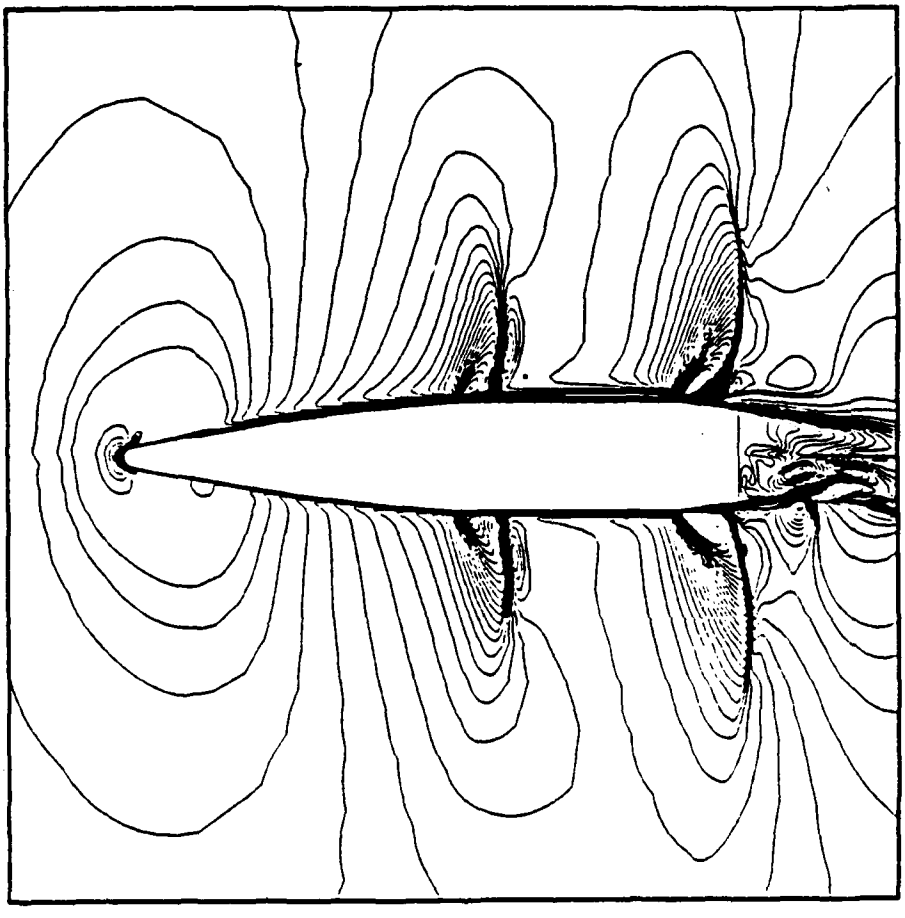


Figure 20. Mach contours, M549 projectile, $M_\infty = 0.94$, $\alpha = 2°$.

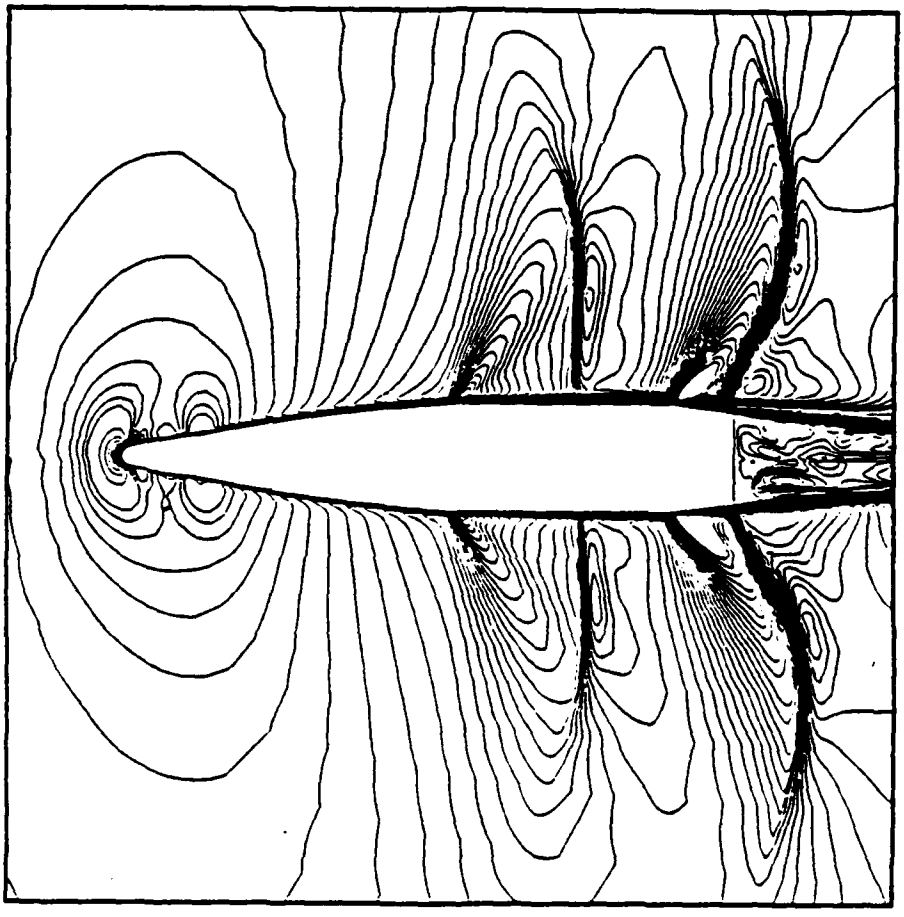


Figure 21. Mach contours, M549 projectile, $M_\infty = 0.98$, $\alpha = 2°$.

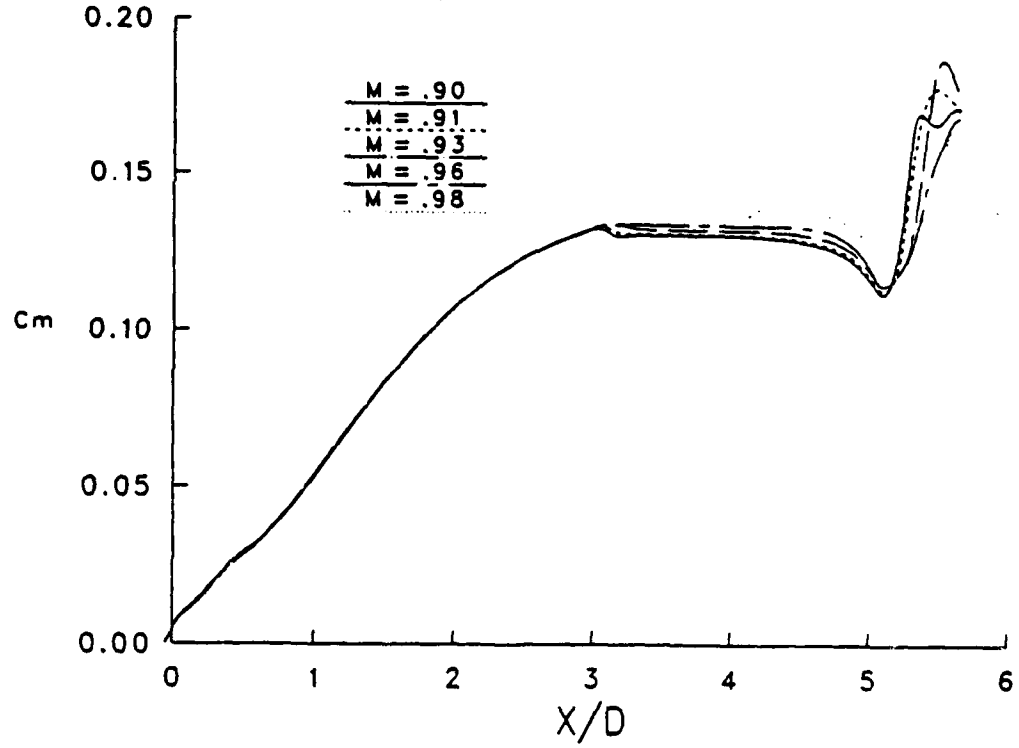


Figure 22. Development of pitching moment coefficient over the M549 projectile, $\alpha = 2°$.

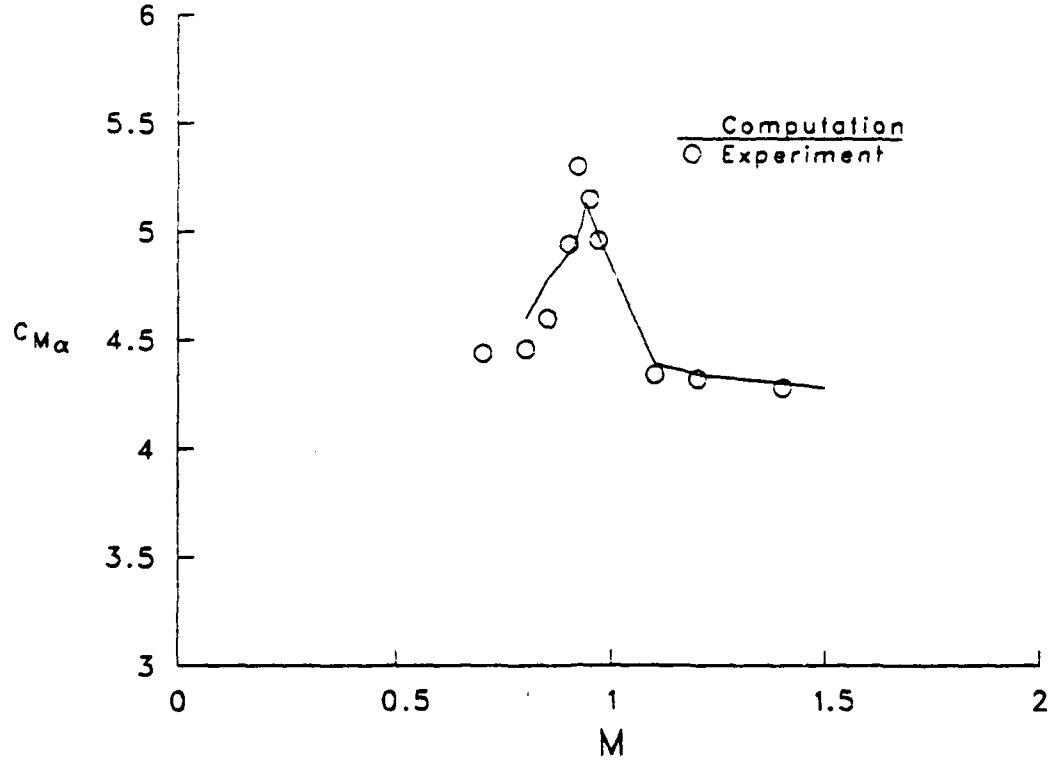


Figure 23. Slope of pitching moment coefficient, C_{m_α} vs Mach number, M549 projectile.

DYNAMIC RESPONSE OF RECTANGULAR STEEL PLATES OBLIQUELY IMPACTING A RIGID TARGET

Aaron Das Gupta
US Army Ballistic Research Laboratory, US Army Laboratory Command
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. Dynamic response of rectangular steel plates of two different thicknesses obliquely impacting a rigid semi-infinite target was modeled using the STEALTH-3D hydrodynamic code based on an explicit Lagrangian finite-difference formulation for solids, structural and thermo-hydraulic analysis. Motivation for this analysis arises from the need to assess transient loads and deformation in the contact zone due to impact of plates of various materials, geometry, initial velocity and angle of obliquity in order to assure structural integrity and avoid premature failure. The STEALTH code was developed by Science Applications Inc. and a DOD solids and structural version without the thermo-hydraulic analysis capability has been employed for this investigation.

The plates were impulsively driven to a high velocity prior to oblique impact upon a rigid wall and were modeled using a three-dimensional 11*11*3 mesh configuration. Only one-half of the plates were modeled by virtue of their lateral symmetry. The Mie-Gruneisen equation of state for steel under hydrostatic compression and elastic perfectly-plastic behavior for deviatoric material strength were included for the material model. Initially a very small time step two orders of magnitude below the Courant stability criteria for the smallest mesh was used to stabilize the explicit integration calculations near the impacted region.

The results indicate large hydrostatic pressure rise at the initial contact zone resulting in severe elasto-plastic deformation and plate bending causing seperation of the leading edge while the trailing zone contacts the rigid surface as the plate continues to slide along the rigid wall for the relatively thick plate impact problem. For the thin plate, occurrence of a plastic hinge, localized bending near the leading edge and sliding along the target surface are observed. Hourglassing instabilty along the boundary did not adversely affect computation near the leading edge and a major portion of the impulse occurred during the initial 40-50 microseconds. The computations indicate that impact loading upon the wall can be accurately estimated using a refined mesh near the leading edge of the plate.

1. INTRODUCTION. The capability to predict the effect of hypervelocity plate impact on a rigid structure is a necessity as a first step towards the design and safe operation of protective enclosures (1,2) commonly used in nuclear power plants. This problem is also of interest to the Ballistic Research Laboratory due to the possibility of fragment induced damage (3) to target enclosures in the terminal ballistics test facilities which might result in catastrophic rupture when the blast loading is applied.

A number of studies have been performed and damage data obtained (4-7) over the years. However, most data available are in the form of impulse correlation curves and crater shapes in plates due to slender rods while

relatively little has been reported in plate impact upon walls involving contact, sliding and seperation.

Recently, computation using hydrodynamic codes (8-10) has been reported. This paper documents numerical studies conducted using the explicit finite-difference computer code, STEALTH (11) to simulate and predict the transient nonlinear behavior resulting from the oblique impact of a plate on a rigid surface. The goals of this numerical simulation were to aid in understanding the process by which fragments impact and deform prior to seperation as well as to demonstrate the applicability of the code in obtaining loading functions in the contact region.

2. IMPACT CONDITION. The flyer plate is assumed to be a 60 cm * 35 cm * 2.54 cm rectangular steel plate impacting a rigid surface at an inclination of 20 degrees to the horizontal surface for the thick plate problem. The plate is assumed to be accelerated to a constant velocity of 200 m/s prior to impact.

For the second case involving the thin plate impact problem the plate which is only 1.6 mm in thickness, impacts a rigid surface at an angle of obliquity of 60 degrees. The plate is assumed to be 31 cm in length and 22 cm in width. The initial contact occurs along the length of the plate at the bottom edge. The plate is assumed to have an initial velocity of 900 m/s prior to impact.

3. MATERIAL MODEL. Computation of stresses and strains in the STEALTH code involves calculation of deviatoric components which depend upon shear strength characteristics as well as hydrostatic components which are governed by high pressure equation of state of the deformable material.

For the thick plate problem quasi-static properties of mild steel and an elastic perfectly-plastic representation of the constitutive relationship was employed for the deviatoric strength of the plate material. This model is available in the standard material library in the computational code. The yield strength was 3.0 KBar and the shear modulus was approximately 0.82 MBar. The shear modulus G was calculated from the Young's modulus E using the relationship

$$G = E / (2 (1 + v))$$

where v is Poisson's ratio which was found to be 0.29.

For the hydrostatic compression, a modified form of the Mie-Gruneisen equation of state for shock propagation in solids was available in the code and could be described as

$$P(\mu,E) = A\mu + B\mu^2 + C\mu^3 + U (D + F\mu + H\mu^2)$$

where A, B, C and D, F, H are material constants determined experimentally from Hugoniot pressure-volume states obtained in shock transitions and μ is a ratio of the specific volume change and the initial volume. U is the internal energy density. Material parameters are available in the standard material library in the STEALTH code for a variety of solids.

The bulk modulus, K, was calculated from the relationship

$$K = E / [3 (1 - 2 v)]$$

The shear modulus, G, is related to the bulk modulus for both loading and unloading phases as

$$G = 3 K (1- 2 v) / [2 (1 + v)]$$

The sound speed, c, in the parent material could be obtained from the deviatoric and hydrostatic compression behavior as

$$c = A + 2B\mu + 3C\mu^2 + U (F + 2H\mu) + (P/V^2)(D + F\mu + H\mu^2) + 1.333 G$$

where V is specific volume and P is the hydrostatic pressure.

Artificial viscosity parameters for both linear and quadratic damping were employed to damp out spurious oscillations. However, zero energy modes due to hourglassing at the boundary could not be damped out due to lack of an appropriate tensorial hourglass viscosity parameter. This was not considered to be a significant problem since the instability did not appear to grow with time and did not seem to affect the contact zone near the leading edge.

In the second case for the thin plate impact problem an AISI 304 grade steel already available in the standard material library was considered. For the deviatoric strength an elastic strain-hardening model representation was used. The yield strength was 20.0 KBar and a hardening exponent of 0.035 was employed. The shear modulus was 0.77 MBar while the spallation threshhold was -0.02. The model also included thermal softening capability.

For the hydrostatic compression part a Gruneisen volume coefficient of 1.4753 and an energy coefficient of 2.17 were employed in the Mie-Gruneisen Equation of State for the 304 Grade steel as opposed to null values used in the earlier case for the low carbon steel. Additionally, a hardening coefficient of 40.0 and a corresponding hardening exponent of 0.35 was used to model the strain hardening part of the deviatoric strength behaviour in contrast with perfectly-plastic assumptions in the thick plate problem. The initial bulk modulus was approximately 1.648 MBar.

4. COMPUTATIONAL ALGORITHM. The STEALTH code was used to simulate the dynamic response due to impact in both cases. The code (11,12) solves the partial differential equations of continuum mechanics using an explicit finite-difference method formulated in a Lagrangian moving coordinate frame.

In the Lagrange system, fixed mass units translate, rotate, compress, expand and distort. Momentum is associated with the motion of the mass and internal energy is fixed to the mass unit. The STEALTH solutions are second-order accurate in space and time. A complete description of the Lagrangian equations solved by the STEALTH code is given in the user's manual (11).

Several rezoning options are available in the program for updating grid point locations and variables in case of large mesh distortion or grid entanglement. Pressure discontinuities are handled by smearing out the discontinuity with a von Neumann quadratic artificial viscosity. Zone to zone oscillations are damped out by means of a linear artificial viscosity.

Stability of the differential equations is automatically regulated by the Courant stability criterion which can be described as

$$\Delta t = \frac{\Delta tc}{n} = \frac{\Delta Lmin}{n\sqrt{Eymax/p}}$$

where Δtc is the minimum Courant stability step size,

$\Delta Lmin$ is the distance of the two closest mesh points in the system,

Eymax is the Young's modulus for the stiffest material,

p is the density of the material, and

n is the number of time steps with which we wish to represent the shock wave in passing through the distance ΔL

5. NUMERICAL MODELING. In both cases only one-half of each plate was modeled by virtue of its lateral symmetry. For the first problem the generic three-dimensional computational model has 11 meshes along the length and an equal number along the width as well as 3 rows of mesh through the thickness of the flyer plate segment. An isometric view of initial configuration of the undeformed mesh is shown in Figure 1.

In the second case the entire plate was modeled using a somewhat refined mesh with 14 rows of mesh along length and width as well as 4 rows through the thickness of 0.16 cm. Because of this narrow thickness an initial time step an order of magnitude lower than in the previous case was required in order to avoid a violation of the stability criteria in the explicit integration scheme at the outset. A refined mesh in the contact region is expected to result in a more accurate description of the impact forces. However, a very small time step approximately 1% of the wave speed transit time to traverse the smallest mesh was needed to stabilze the computation in the impacted region.

The computational procedure used in modeling the angular impact process can be summarized as follows :

a. The rigid surface acts as a fixed boundary for the finite-difference grid.

b. The forces acting on the rigid surface in the impacted region is calculated by STEALTH from the stresses developed in the deformable plate and summed to give the resultant cell averaged contact pressure. Frictionless contact is assumed between the plate and the rigid surface.

c. Additionally STEALTH computes the new grid point positions due to sliding upon impact. These updated locations are then used as input for the next computational cycle. No rezoning is used in this calculation.

6. DYNAMIC RESPONSE COMPUTATION. The entire bottom and front surfaces of the plate was designated as a wall interaction boundary to allow initial contact and subsequent interaction with the rigid surface. Dynamic response calculations were initially performed for 40 microseconds upon impact and was

later extended upto 0.3 ms to monitor the post-impact response which should show sliding and seperation in addition to transverse bending effects.

The dynamic response of the relatively thick plate is shown in Figures 2a-2e which describes the behavior from 10-300 microseconds after impact. After initial contact along the leading edge, surface contact at the bottom frontal element is visible in Figure 2a at an elapsed time of 10 microseconds signalling the onset of bending. Significant amount of bending is visible in Figure 2b at 125 microseconds beyond impact at the first two rows of elements at the forward end of the plate while the contact area has increased beyond the first set of elements at the bottom surface.

At an increased response time of 200 microseconds after impact bending appears to propagate and affect the first three rows of elements while the contact area has progressed to the first two rows of elements at the forward bottom location of the plate as shown in Figure 2c. Initial compression of the leading edge resulting in development of compressive stresses near the forward end are significantly altered by the onset of bending causing compressive stresses in the top fibers and tensile stresses in the bottom fibers.

With further increase in response times bending stress wave propagates towards the rear of the plate. At 250 microseconds lifting of the forward edge and seperation from the rigid surface are indicated as shown in Figure 2d. This is accompanied by a backward shift of the contact region as the third row of elements at the bottom comes in contact with the rigid wall boundary. Bending has now progressed beyond the first three rows of elements at the forward end.

At an extended response time of 300 microseconds the post-impact process continues causing further seperation and upward bending of the leading edge and the forward end while the contact zone shifts backward indicating partial contact of the second and fourth rows of elements at the bottom surface and complete contact with the third rows of elements as shown in Figure 2e. This process of deformation is realistic in the sense that significant bending and seperation of the forward end is expected at a shallow angle of attack.

For the thin plate problem dynamic response studies were conducted upto 700 cycles corresponding to an elapsed time of 40 microseconds only. This is because the automatically adjusted time steps were an order of magnitude lower than those for the previous case due to very small thickness of the plate and the use of a refined mesh scheme for this model.

Typical deformation patterns at 20 and 40 microseconds are shown in in Figures 3a and 3b respectively. Due to lack of sufficient resolution the grid appears as a dark band in the end view. The onset of bending is clearly visible in Figure 3a. Comparison of the two figures indicate sliding along the rigid surface in a direction along the horizontal component of the initial velocity vector. This is expected since a zero friction coefficient has been imposed along the interacting boundary. Additionally, evidence of thin plate buckling and formation of a plastic hinge approximately two mesh points away from the leading edge can be observed in Figure 3b. These phenomena create severe compression of elements near the leading edge along the thickness direction requiring further drop in the allowable computational time step to avoid instability problem. At longer time steps spurious

oscillations throughout the plate and grid instability is observed due to hourglassing and consequent zero energy modes which may be controlled by a tensor-triangle artificial viscosity.

A typical forcing function for the thick plate problem generated by the STEALTH code is shown in Figure 4. The forcing function is computed as a cell averaged pressure at the center of the bottom mesh near the leading edge where initial contact occurs. Since the contact zone propagates along this bottom mesh surface, the cell averaged pressure can be a reasonably acceptable measure of the impact load in the contact zone. The accuracy of the measure in representing the actual contact load can be improved upon by refining the grid particularly near the impacted region of the plate.

As depicted in Figure 4 the impact pressure has a rather steep climb and a sharp peak of 2.5 KBar at an early time of 4.0 microseconds after impact. This is followed by an equally steep drop between 4.0 and 12.0 microseconds unloading to zero. Subequently the plate reloads to approximately 6.0 KBar at 46.0 microseconds and oscillates about the 3.0 KBar level for a rather extended period of time beyond 46.0 microseconds. This wringing behavior of the plate is probably due to reflection of stress waves from the top and bottom surfaces near the forward end of the plate and it appears to gradually decay in amplitude with time. Beyond 40 microseconds seperation of the leading edge and the forward bottom mesh from the rigid surface causes a drop in the impact load to the fully unloaded level.

The contact pressure-time history due to oblique impact of a flyer plate for the thin plate problem shows a similar trend although a peak pressure of 7.5 KBar occurs early with a considerably smaller duration and the subsequent oscillation is rather noisy due to propagation of stress waves through the plate material. Hourglassing instability did not affect the contact zone adversely and the contact pressure could be monitored until step instability or tip seperation takes place. However, a major part of the total impulse is contained in the first 40-50 microseconds and further computation does not contribute significantly to the forcing function.

7. CONCLUSION. Results from numerical simulation of a thick as well as a thin plate on a rigid surface are presented. Realistically these results are conservative in the sense that peak pressures and deformation from impact on a rigid surface should be higher than those due to impact on a deformable surface. If the forcing function from a nonresponding surface is applied to drive a responding model in a structural response code, a small error in terms of somewhat higher displacements and stress levels should be expected. In some cases this may be desirable since the margin of safety from a structural integrity standpoint could be enhanced using this procedure.

In the absence of any experimental data, the deformation patterns ensure increased confidence in the predicted results from the STEALTH code which yield valuable insight into the post-impact response behavior of plates. The complex phenomena of sliding and seperation are demonstrated using the 3-D computational model. Inspite of simplistic assumptions of frictionless sliding, nonresponding surface and quasi-static idealized materials data, useful data for plates of varying thickness, initial velocity and inclination could be generated in a cost-effective and efficient manner using the STEALTH code.

ACKNOWLEDGEMENTS.

Valuable assistance of Dr. Gayman Yeh of Science Applications Inc. and Dr. Joseph M. Santiago of the Terminal Ballistics Division during the course of this investigation is gratefully acknowledged.

REFERENCES.

1. J.T. Gordon Jr. and J.E. Reaugh, "Strain-Rate Effects on Turbine Missile Casing Impact", Computers & Structures, Vol.13, pp. 311-318, 1981.

2. H.R. Yoshimura and J.T. Schauman," Preliminary Results of Turbine Missile Casing Tests" EPRI Research Project Report No. 399, EPRI, Palo Alto, California, 1978.

3. A.D. Gupta and H.L. Wisniewski, "Dynamic Response of the Hemispherical Containment Structure Subjected to Transient Loads at the R-9 Firing Range", Ballistic Research Laboratory Memorandum Report ARBRL-MR-03249, (ADA # 125530), Aberdeen Proving Ground, Maryland, March 1983.

4. J.T. Dehn, "Models of Explosively Driven Metal", US Army Ballistic Research Laboratory Technical Report # BRL-TR-2626, Aberdeen Proving Ground, Maryland, 1984.

5. R.M. Norman, "Deformation in Flat Plates Exposed to HE Mine Blast", AMSAA-TM-74, US Army Material Systems Analysis Agency, APG, MD, 1970.

6. N.E. Hoskin, J.W. Allan, W.A. Bailey, J.W. Lethaby and I. Skidmore, "The Motion of Plates and Cylinders Driven at Tangential Incidence", Fourth International Symposium on Detonation, ONR ACR-126, p.14, 1965.

7. J.A. Zukas, T. Nicholas, H.F. Swift, L.B. Gresczuk and D.R. Curran, "Impact Dynamics",pp 150-165, John Wiley and Sons, 1982.

8. B.D. Lambourn and J.E. Hartley, "The Calculation of the Hydrodynamic Behavior of Plane One-Dimensional Explosive/Metal System", Fourth International Symposium on Detonation, ONR ACR-126, 1965.

9. W.E. Johnson, "Code Correlation Study", AFWL-TR-70-144, Air Force Weapons Laboratory, Kirtland, New Mexico, 1971.

10. R.E. Lottero and K.D. Kimsey, "A Comparison of Computed versus Experimental Loading and Response of a Flat Plate Subjected to Mine Blast", ARBRL-MR-03249, US Army Ballistic Research Laboratory, APG, MD, 1978.

11. R. Hoffman, "STEALTH, A Lagrange Explicit Finite-Difference Code for Solids, Structural and Thermohydraulic Analysis", EPRI NP-200, Volume 1-4, Electric Power Research Institute, Palo Alto, California, 1976.

12. R. Hoffman, "STEALTH, A Lagrange Explicit Finite-Difference Code for Solids, Structural and Thermohydraulic Analysis", EPRI NP-176, Summary, Electric Power Research Institute, Palo Alto, California, 1976.

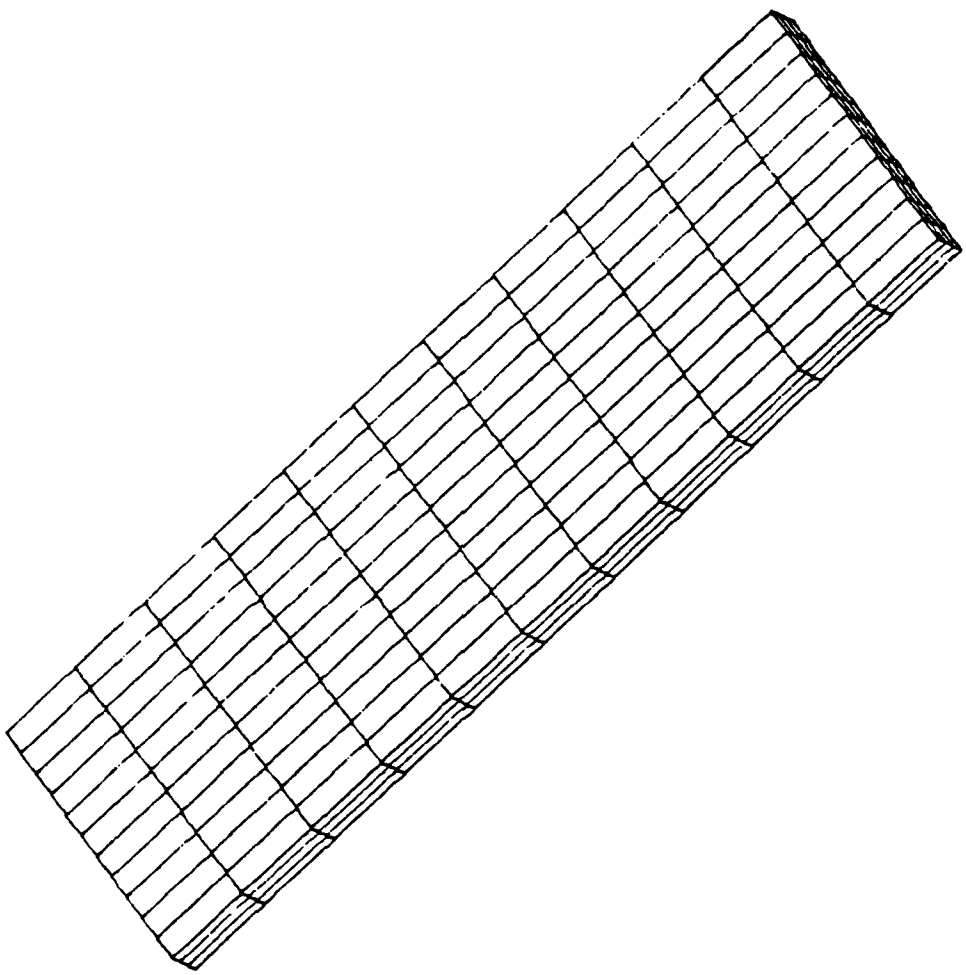
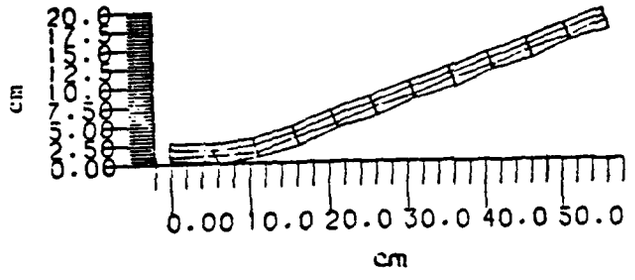
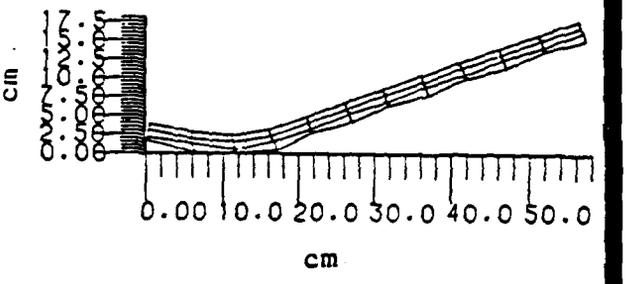


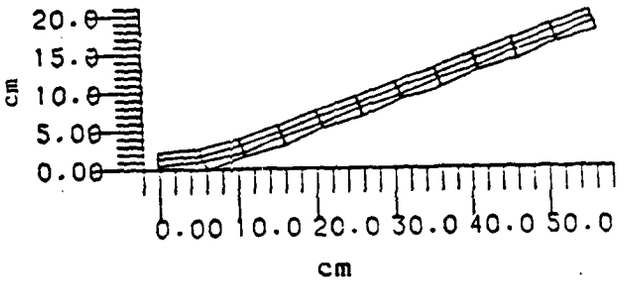
Figure 1: Initial configuration of the undeformed mesh for the plate model.



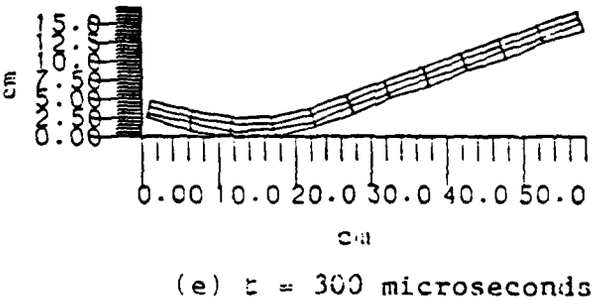
(c) t = 200 microseconds



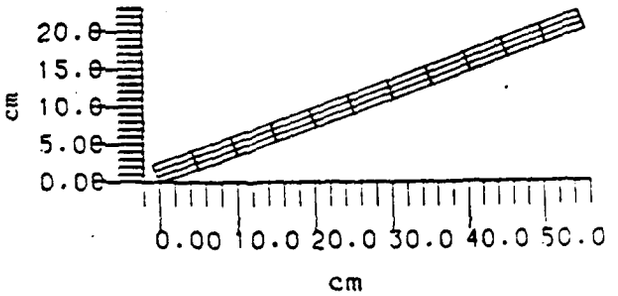
(d) t = 250 microseconds



(b) t = 125 microseconds

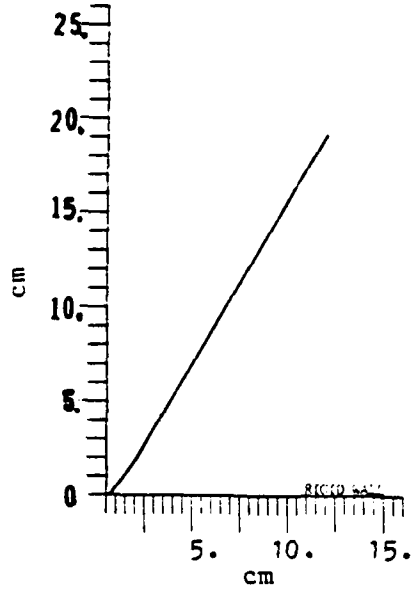


(e) t = 300 microseconds

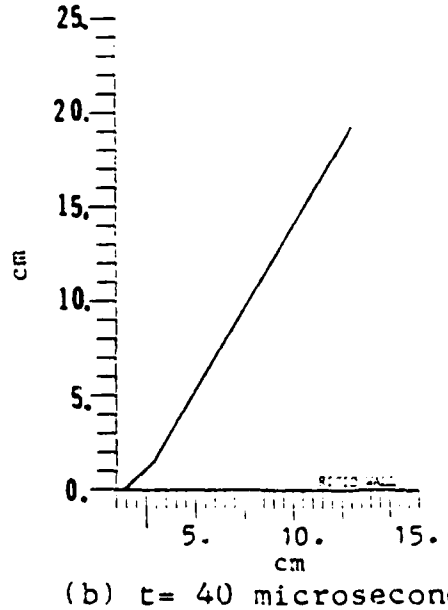


(a) t= 10 microseconds

Figure 2: Dynamic response behavior of the plate impacting a rigid wall.



(a) t = 20 microseconds



(b) t= 40 microseconds

Figure 3: Dynamic response behavior of the thin plate impacting a rigid wall.

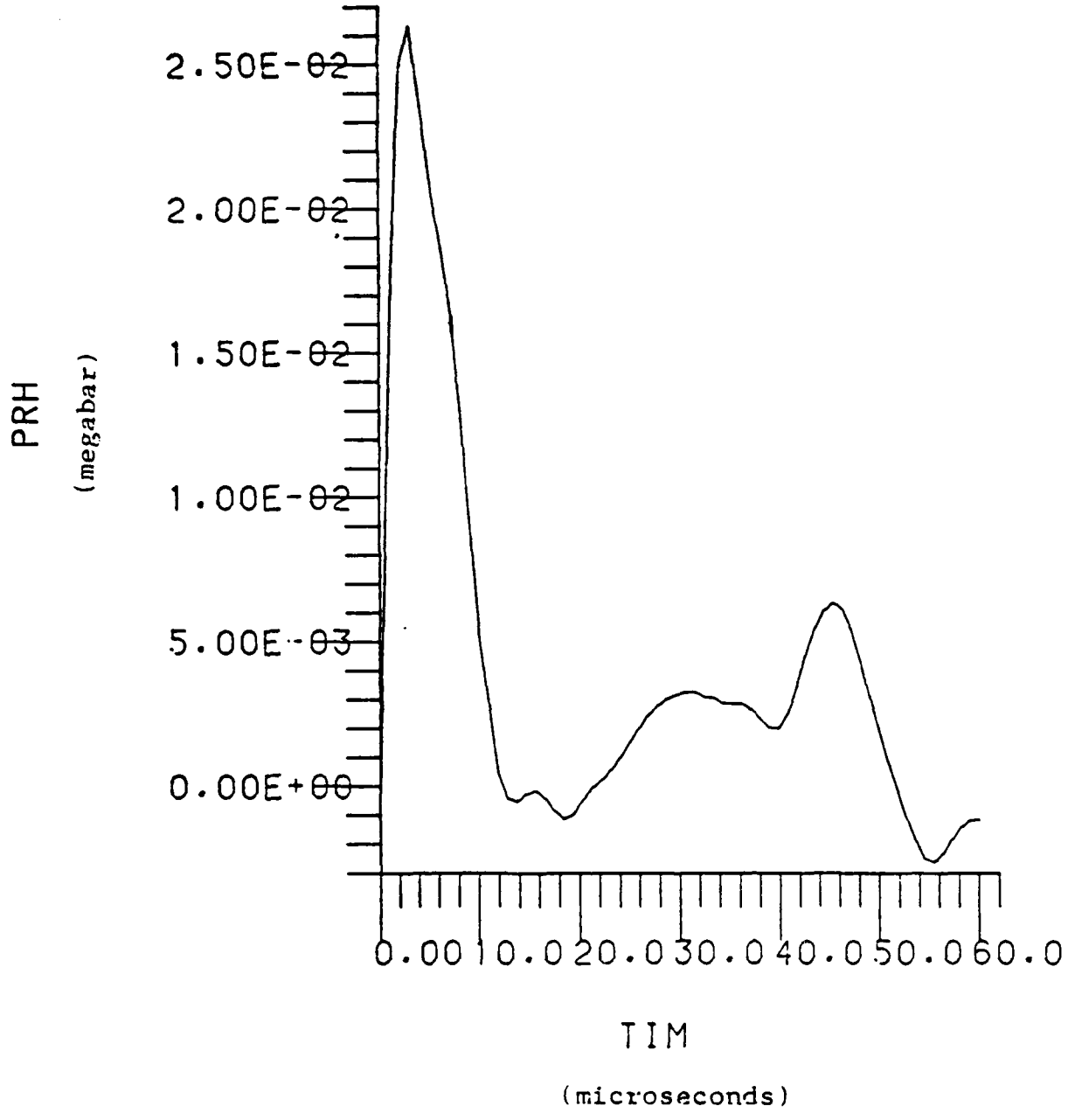


Figure 4: Typical contact pressure generated by STEALTH code for the plate impact problem.

An Adaptive Mesh Method for Solving Blast Problems Using the Euler Equations

David C. Arney

Department of Mathematics
United States Military Academy
West Point, NY 10996-1786

Garry Carofano

U.S. Army Armament Research, Development and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

Erin Misner

U.S. Army Armament Research, Development and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

and

Department of Mathematics
United States Military Academy
West Point, NY 10996-1786

ABSTRACT. We discuss mesh moving and local mesh refinement that are used with MacCormack's scheme to solve the Euler equations for inviscid compressible flow in two space dimensions. A coarse base mesh of quadrilateral cells is moved by an algebraic mesh movement function so as to follow and isolate distinct phenomena. The local mesh refinement algorithm recursively divides the time step and spatial cells of the moving base mesh in regions where the error indicators are high. A mesh generation procedure is used to create the initial base mesh. MacCormack's scheme is given total variation diminishing (TVD) artificial viscosity in order to compute shocks and discontinuities. The time step is adjusted automatically to maintain stable computation by calculating the maximum eigenvalues of the Euler Equations on the computational mesh. Results are presented for computational examples involving planar and cylindrical blasts.

1. INTRODUCTION. The numerical solution of the Euler equations is often difficult because the nature, location, and duration of fine-scale structures is often not known in advance. Thus calculation on a uniform or prescribed mesh can fail to adequately resolve the fine-scale phenomena or have excessive computational costs. Adaptive mesh procedures that evolve with the solution offer a robust, reliable, and efficient alternative. Such techniques for time-dependent problems are either capable of creating finer meshes in regions of excessive error [1,2,3,4,5] or moving meshes to follow isolated dynamic phenomena [1,2,5,6,7,8,9]. The use of these techniques is enhanced when they

are capable of providing an accurate error estimate for the computed solution [1,2,4,10,11,12].

Our procedure solves the two-dimensional Euler equations

$$\vec{u}_t + \vec{f}_x(\vec{u}) + \vec{g}_y(\vec{u}) = 0,$$

$$\vec{u} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix}, \quad \vec{f}(\vec{u}) = \begin{bmatrix} \rho u \\ \rho u^2+p \\ \rho uv \\ u(e+p) \end{bmatrix}, \quad \vec{g}(\vec{u}) = \begin{bmatrix} \rho u \\ \rho uv \\ \rho v^2+p \\ v(e+p) \end{bmatrix}, \tag{1}$$

on a rectangular domain Ω with well-posed initial and boundary conditions. Here, u and v are the velocity components of the fluid in the x and y directions, ρ is the fluid density, e is the total energy of the fluid per unit volume, and p is the fluid pressure. For an ideal gas the equation of state is

$$p = (\gamma - 1)\left[e - \rho(u^2 + v^2)/2\right], \tag{2}$$

where γ is the ratio of the specific heats.

We use MacCormack's [13] explicit finite difference scheme with Davis's [14] artificial viscosity to calculate solutions of (1) at each node of a moving mesh of quadrilateral cells. We use a density switch (gradient) to to estimate error at the nodes of the mesh and a procedure to select the proper time step size.

Our adaptive mesh algorithm was modified from a general-purpose scheme and consists of three main parts (i) movement of a coarse base mesh (cf. Arney and Flaherty [6]), (ii) local refinement of the base mesh in regions where the resolution is inadequate (cf. Arney and Flaherty [3]), and (iii) regeneration of the base mesh when it becomes too distorted and unsuitable for further computation (cf. Arney and Flaherty [15]). Proper mesh motion can reduce errors; however, mesh motion alone cannot produce solutions that satisfy prescribed error tolerances. Therefore, local mesh refinement is added to recursively solve local problems in regions where error tolerances are not satisfied. If the base mesh becomes too distorted a mesh regeneration procedure is used to produce a better base mesh. The combination of our solution method, error estimation, and adaptive mesh techniques creates a powerful algorithm since the solution method provides robustness, the error estimation and mesh refinement provide accuracy, and the mesh moving, mesh regeneration, and time step selection provide efficiency.

We briefly explain the various parts of our algorithm in Sections 2 and 3. The results of their use on an example problem are given in Section 4. The status of our algorithm and future considerations of adaptive methods for the Euler equations are discussed in Section 5.

2. SOLUTION SCHEME AND ERROR ESTIMATION ON A MOVING NONUNIFORM MESH.

MacCormack's scheme has had wide use in solving Euler equations. The use of artificial viscosity to make this scheme total variation diminishing (TVD) makes it more attractive to solve problems with discontinuities. Our Richardson's extrapolation-based error estimation produces a pointwise approximation of the local discretization error which can be used to construct global or local measures of the error.

a. MacCormack's Scheme. In order to discretize (1) on a moving nonuniform mesh, we introduce a transformation

$$\xi = \xi(x,y,t), \quad \eta = \eta(x,y,t), \quad \tau = t, \tag{3}$$

from the physical (x,y,t) domain to the computational (ξ,η,τ) domain where a uniform rectangular grid is used. Under this transformation (1) becomes

$$\vec{u}_\tau + \vec{u}_\xi \xi_t + \vec{u}_\eta \eta_t + \vec{f}_\xi \xi_x + \vec{f}_\eta \eta_x + \vec{g}_\xi \xi_y + \vec{g}_\eta \eta_y = 0. \tag{4}$$

The two-step MacCormack's scheme [13] then uses first-order forward temporal and spatial difference approximations in the predictor step and first-order backward differences in the corrector step to solve (4). Hindman [16] showed that proper differencing of this chain-rule form (4) with its metrics produces a consistent approximation. Arney and Flaherty [10] showed that conservation is also maintained. We have also used the finite-difference method of Harten [17] in this adaptive mesh algorithm. However, because of flux splitting the mesh must be constrained to remain rectangular. This constraint limits the benefits of mesh moving but does not affect the mesh refinement.

The explicit MacCormack's scheme has a stability restriction that limits the time step allowed for a given spatial mesh. For efficient computation, we choose the next time step adaptively to be close to the maximum allowed by the Courant, Friedrichs, Lewy theorem.

b. Davis's Artificial Viscosity. MacCormack's scheme, being a second-order accurate centered scheme, produces spurious oscillations near discontinuities. In order to eliminate or reduce these oscillations, artificial viscosity is added to the solution to diffuse the discontinuity. We use an artificial viscosity model due to Davis [14] which is not problem dependent and only requires knowledge of the maximum eigenvalues. This artificial viscosity model is designed to convert MacCormack's scheme into a TVD scheme in one dimension. There are other artificial viscosity models for the Euler equations that produce TVD schemes (cf. Pulliam [18]).

Davis's artificial viscosity is based on a flux limiter that does not depend on explicitly determining the upwind direction and, with a modification by Roe [19], does not affect the region of stability of MacCormack's scheme. Because MacCormack's scheme does not determine the upwind direction, the combined use of MacCormack's scheme and Davis's artificial viscosity is computationally simple. The artificial viscosity terms are calculated from the solution data at the beginning of the time step. The maximum absolute

eigenvalues for the Euler equations on the computational mesh are computed from the maximum absolute values of

$$\xi_t + \xi_x u + \xi_y v \pm \sqrt{\left(\frac{\gamma p}{\rho}\right)} \sqrt{\xi_x^2 + \xi_y^2} \tag{5}$$

and

$$\eta_t + \eta_x u + \eta_y v \pm \sqrt{\left(\frac{\gamma p}{\rho}\right)} \sqrt{\eta_x^2 + \eta_y^2} \ ,$$

in the ξ and η directions, respectively [18]. For two-dimensional problems separate dissipative terms are computed in the ξ and η directions. However, this scheme is not TVD in two dimensions.

c. Error Estimation. A posteriori error estimation is an integral part of our adaptive system. The general-purpose scheme we modified estimated the local temporal and spatial portions of the discretization error on a moving mesh using an algorithm based on Richardson's extrapolation (cf. Arney and Flaherty [10]). Flaherty and Moore [20] and Berger and Oliger [4] used a similar form of Richardson's extrapolation to estimate error on uniform meshes.

The Richardson's extrapolation error estimation procedure is expensive, costing up to four times more to compute than the solution, and is based on the assumption of a smooth solution, which is not the case for blast problems (cf. Section 4). Therefore, we use a less expensive error indicator $\left(e^k_{i,j}\right)$ called the density switch which is computed as

$$e^k_{i,j} = 0.5 \left[\frac{\left|\rho^k_{i,j} - \rho^k_{i-1,j}\right|}{\rho^k_{i,j} + \rho^k_{i-1,j}} + \frac{\left|\rho^k_{i,j} - \rho^k_{i,j-1}\right|}{\rho^k_{i,j} + \rho^k_{i,j-1}} \right] , \tag{6}$$

using one-sided differences. We also used a form with centered differences. However, this error indicator may not coverge to zero as the mesh is refined in blast problems and therefore, a maximum level of refinement must be used in connection with the error tolerance to control mesh refinement.

3. ADAPTIVE MESH PROCEDURES. An algorithm of our adaptive procedure is presented in Figure 1. This procedure integrates the Euler equations from time (tint) to (tfinal) while keeping the local error estimates below a tolerance of (tol). The base level time step Δt is initially specified, but is changed during the solution to maintain stability.

```
procedure adaptive_PDE_solver(tinit, Δt, tfinal, tol:  real; M, N:  integer);

begin
  Generate an initial base mesh;
  t := tinit;

  while t < tfinal do
    begin
      mesh_move(M, N, Δt, tol);
      local_refine(0, t, Δt, tol);
      t := t + Δt;
      Select an appropriate Δt;
      if base mesh is too distorted then regenerate a base mesh
    end
end { adaptive_PDE_solver };
```

Figure 1. Description of our adaptive algorithm to solve (1) to within a tolerance of (tol).

The rectangular domain Ω is initially discretized into a coarse moving spatial grid of $M \times N$ quadrilateral cells. The base mesh is moved for each base time step Δt and (1) is solved on this mesh. This is followed by recursive local mesh refinement. The value of Δt is calculated from the eigenvalues (5) and the Courant-Friedrichs-Lewy condition to maintain stability for the next time step. Finally, a new base mesh is generated if necessary. The solution, error estimation, mesh moving, local refinement, and mesh regeneration procedures are explicit and uncoupled from one another reducing their computational cost and providing flexibility. Therefore, the solution and error estimation procedures could be replaced with ones suitable for the Navier-Stokes equations.

a. Mesh Moving. Our mesh moving procedure is based on an intuitive approach. The essential idea is that the mesh moves to follow isolated nonuniformities, such as wave fronts and shocks, which manifest themselves with high error estimates. Proper mesh movement generally reduces dispersive errors and can allow the use of larger time steps if the eigenvalues are reduced in the Courant-Friedrichs-Lewy condition while maintaining accuracy and stability.

The algorithm for our mesh moving procedure mesh_move is presented in Figure 2. At each base time, we scan the base mesh and locate significant-error nodes as those having error indicator greater than twice the mean nodal error estimate and also greater than ten percent of tol. This strategy avoids having the mesh respond to fluctuations with too small an error estimate, yet is sensitive enough to avoid missing significant dynamic phenomena. If there are no significant-error nodes, computation proceeds on a stationary mesh. The nearest neighbor clustering algorithm of Berger and Oliger [4] is used to gather the significant error nodes into rectangular error clusters.

```
procedure mesh_move (M, N:  integer; Δt, tol:  real);
  begin
    for j := 1 to M × N do
        compile list of significant error nodes using tol;
    if no significant error nodes then no mesh movement
       else cluster significant error nodes into k error clusters.
    for m := 1 to k do
       calculate propagation of error cluster from Δt;
    for j := 1 to M × N do
        move nodes based on function of the velocity of
        the nearest error clusters;
    smooth the node movement to reduce deformation;
  end {. mesh_move };
```

Figure 2. Pseudo-PASCAL description of mesh moving algorithm to move mesh for one base time step (Δt).

We determine individual node movement from the velocity of propagation, the orientation, and the size of the error clusters. We assume that nodes in the same cluster have related solution characteristics, so that we determine individual node movement from the propagation of the center of the nearest error cluster.

b. Local Mesh Refinement. The local refinement procedure is invoked after the base mesh has moved. Our refinement strategy consists of first calculating the solution and error estimates on the base mesh. Finer grids are created in untolerable-error regions by locally bisecting the time steps and the sides of the quadrilateral cells of the base grid. The solution and error estimates are computed on the finer grids. The refinement scheme is recursive; thus, fine grids may be refined to create finer grids.

This grid relationship leads to a tree data structure. Information regarding the base grid is stored in the root node or level 0 of the tree. Subgrids of the base grid are stored in level 1 of the tree. The structure continues, with a grid at level ℓ having subgrids at level $\ell+1$. Grids at level ℓ are given arbitrary ordering and we denote them by $G[\ell,j]$.

Our recursive local mesh refinement algorithm local-refine is presented in Figure 3. The procedure integrates (1) from time tinit to tinit + dt, attempting to satisfy the error tolerance tol.

Our technique for introducing finer subgrids consists of four steps: (i) scanning level ℓ grids to locate untolerable-error nodes, (ii) clustering those nodes into rectangular regions [4], (iii) buffering the regions in order to reduce problems associated with prescribing initial and boundary conditions at coarse/fine grid interfaces, and (iv) cellularly refining the level ℓ meshes and time steps inside the buffered clusters. Base mesh motion is maintained on the refined subgrids to insure proper nesting in their parent grid. If there are no untolerable-error nodes, the solution is acceptable and no further refinement is necessary.

```
procedure local_refine(ℓ:  integer; tinit, Δt, tol:  real);
  begin
    for j := 1 to N[ℓ] do
      begin
        Integrate the partial differential system from tinit to tinit + Δt
          on grid G[ℓ,j];
        Calculate error indicators at tinit + Δt at all nodes of grid
          G[ℓ,j];
        if any error indicators > tol then introduce level ℓ + 1 subgrids
           of G[ℓ,j]
      end { for }

    if any error indicators > tol then
       begin
         local_refine(ℓ + 1, tinit, Δt/2, tol/2);
         local_refine(ℓ + 1, tinit + Δt/2, Δt/2, tol/2)
       end
    end { local_refine };
```

Figure 3. Pseudo-PASCAL description of a recursive local refinement procedure to find a solution of the partial differential system (1) on all grids at level ℓ of the tree.

c. Generation and Regeneration of the Base Mesh. The efficiency of our adaptive strategies depends on the ability to generate a suitable initial base mesh and to regenerate a new base mesh should it become distorted. The two essential elements of mesh generation or regeneration are determination of the number of nodes and their optimal location. A base mesh with too few nodes will result in excessive refinement or may completely miss a fine structure while one having too many nodes will reduce efficiency. Our approach is to use the error estimation of a trial solution for one on a $K \times L$ mesh time step to determine the number of nodes ($M \times N$) and their placement in the initial mesh that approximately equidistributes the error estimates.

The node placement algorithm for the base mesh is similar to the mesh moving algorithm except that nodes are moved toward the center of the nearest error cluster. Nodes nearly equidistant from two or more error clusters are moved by a weighted average toward those nearest error clusters to maintain a smooth mesh. Nodes on $\partial\Omega$ remain on $\partial\Omega$, and nodes near the boundary are moved a reduced distance in order to prevent the formation of large aspect ratios. This construction generates a base mesh that depends on the solution of (1) as well as the initial conditions. The mesh generation algorithm is presented in Figure 4.

The base mesh can become distorted by mesh motion for some problems. We regenerate a new base mesh whenever this happens. The mesh regeneration or static rezone procedure consists of three steps: (i) determining the need for a new base mesh, (ii) creating the new mesh, and (iii) interpolating the solution from the old base mesh to the new base mesh.

```
procedure mesh_generation (K,L,M,N: integer; Δt, tol:  real);
  begin
    solve (1) using Δt on K × L uniform test mesh;
    determine new mesh spacing M and N from error estimation;

    for j := 1 to K × L do
          compile list of error nodes exceeding tol;
    if no error nodes then use uniform mesh
       else cluster error nodes into P clusters;

    for j := 1 to M × N do
       begin
         move nodes toward center of nearest error cluster,
         smooth node spacing near boundaries of domain and between clusters;
       end {for}
    end { mesh_generation };
```

Figure 4. Pseudo-PASCAL description of mesh generation algorithm.

The base mesh is regenerated whenever any interior angle of a cell is less than 40 degrees or more than 140 degrees or the aspect ratio for any cell is greater than 15. A new base mesh is generated using the same procedure used to generate the initial one. The error clusters used for regeneration are those already determined in the mesh moving step. Once the new base mesh has been constructed, the solution on the old one is interpolated to the new one using bilinear interpolation.

4. COMPUTATIONAL EXAMPLES.

EXAMPLE 1. Consider (1) where a planar Mach 10 shock in air moves down a channel containing a wedge with a half-angle of thirty degrees. This problem was used as a test problem by Woodward and Collela [21]. Like them, we orient a rectangular computational domain, $-0.3 \le x \le 3.2$, $0 \le y \le 1$, so that the top edge of the wedge is is on the bottom of the domain in the interval $y = 0$. Thus, in the computational domain it appears like a Mach 10 shock is impinging on a flat plate at an angle of sixty degrees. The initial conditions are

$$\rho = 8.0,\ p = 116.5,\ e = 563.5,\ u = 4.125\sqrt{3},\ v = -4.125,$$

$$\text{if } y < \sqrt{3}(x-1/6), \qquad (7)$$

and

$$\rho = 1.4,\ p = 1.0,\ e = 2.5,\ u = 0,\ v = 0,$$

$$\text{if } y \ge \sqrt{3}(x-1/6).$$

Along the left boundary and bottom boundary left of the wedge, we prescribe Dirichlet conditions of (7). Along the top boundary we prescribe the exact motion of a Mach 10 shock. Along the right boundary, all normal derivatives are set to 0. Along the wedge reflecting boundary conditions are used.

The solution of this problem is a self-similar structure called a double-Mach reflection [22]. The geometries of the structures are very fine. The interesting structures are primarily confined to a small region that moves along the wedge behind the incident shock.

We calculate a solution for $0 < t < 0.19$. Refinement was restricted to a maximum of two levels and a tolerance level of 0.6 in the maximum norm was prescribed. This is necessary because our pointwise error estimate based on the assumption of smooth solutions is not appropriate for problems having discontinuities. Without restricting the maximum level of refinement, we could refine excessively in the vicinity of a discontinuity and exhaust the available storage. Our base mesh generation procedure provided a 29×11 discretization of the domain.

The sequence of meshes in Figure 5 shows that the coarse mesh is able to follow the dynamic structures and that refinement is performed in the vicinity of discontinuities. The density obtained from the adaptive mesh calculation, which used a range of nodes from 960 on the first time step to 3450 on the last time step, is shown in Figure 6 (top). The adaptive solution compares favorably with the solution computed on a 120 x 40 uniform mesh shown in Figure 6 (bottom).

Severe distortion of the mesh in the reflected shock region caused a static mesh regeneration to occur at $t = 0.162$. The overhead of mesh moving for this problem is approximately five percent in terms of total computational time. The CPU time for the adaptive solution using 94 coarse mesh time steps was 51 percent of the time for the stationary uniform mesh solution using 200 coarse mesh time steps.

EXAMPLE 2. Consider (1) where an infinite cylindrical piston is expanding radially creating a radially expanding shock. We orient the computational domain, $0 \le x \le 0.05$, $0 \le y \le 0.05$, to solve in one quadrant of the expansion. The initial conditions were computed by solving the following ordinary differential equations from [23,24,25]:

$$\frac{dv}{dr^*} = \frac{-jv}{r^*}\left(1 - \frac{1}{a^2}(U_p r^* - v)^2\right)^{-1}$$
$$\frac{da}{dr^*} = \frac{-jv}{r^*}(\gamma - 1)\left(\frac{U_p r^* - v}{2a}\right)\left(1 - \frac{1}{a^2}(U_p r^* - v)^2\right)^{-1}, \tag{8}$$

where U_p is the piston velocity, v is the fluid velocity, a is the acoustic speed, r^* is a nondimensional variable defined as $r/U_p t$, and j is a dimensional parameter which is one for a cylindrical piston. These same equations apply to an expanding plane ($j = 0$) and an expanding sphere ($j = 2$). r is the radial distance from the center of the cylinder. The expanding sphere problem is an axi-symmetric two-dimensional problem which can be solved with our code by adding the appropriate forcing terms to (1) found in Carofano [26].

The initial conditions and shock Mach number (M) can be determined by a bisection method for a given piston velocity by matching the fluid velocity. Tables of initial values, Mach number, and shock locations have been assembled in Brantley [25] allowing us to use a Runge-Kutta scheme to solve (8) directly. The parameter values chosen for this problem are $M = 1.7752$ and $U_p = 1.6185$. The solution computed with (8) for three different times is shown in Figure 7.

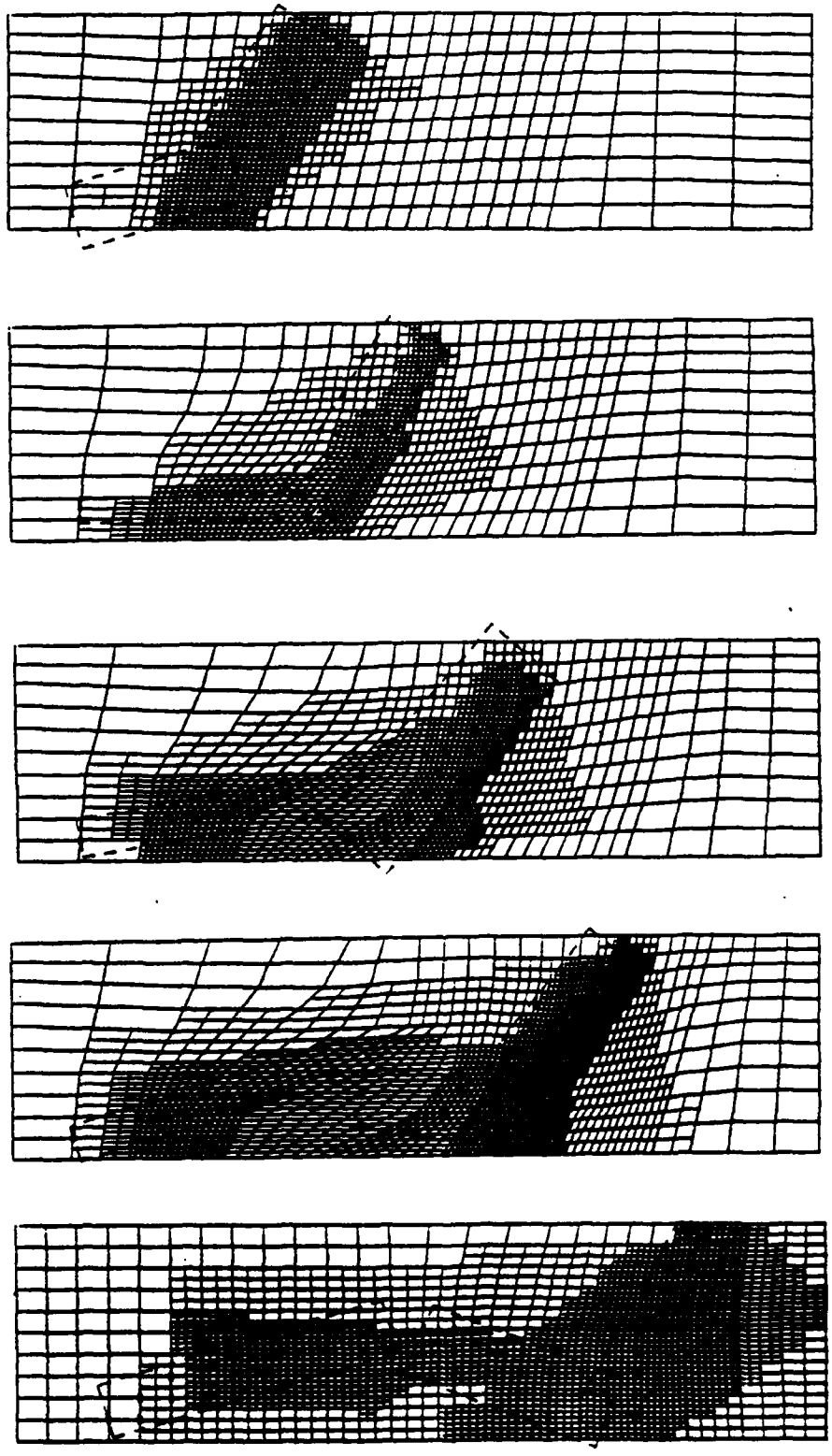


Figure 5. Grids created for the adaptive mesh solution at t = .038, 0.076, 0.114, 0.152, and 0.19 (top to bottom). The rectangular boxes represent the error clusters used to move the mesh for the current time step.

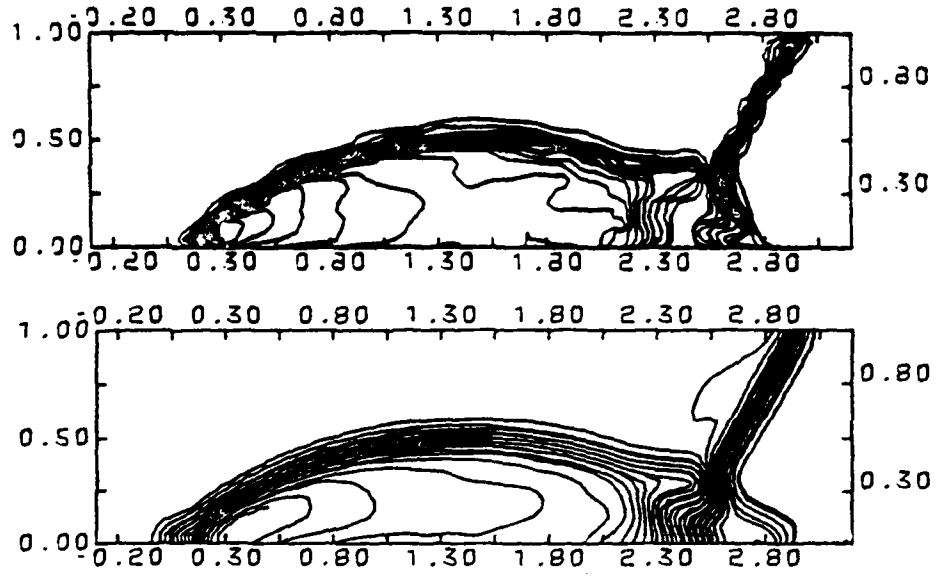


Figure 6. Contours of the density at $t = 0.9$ for an adaptive solution on a 29×11 base mesh (top) and for a uniform stationary 120×40 base mesh (bottom).

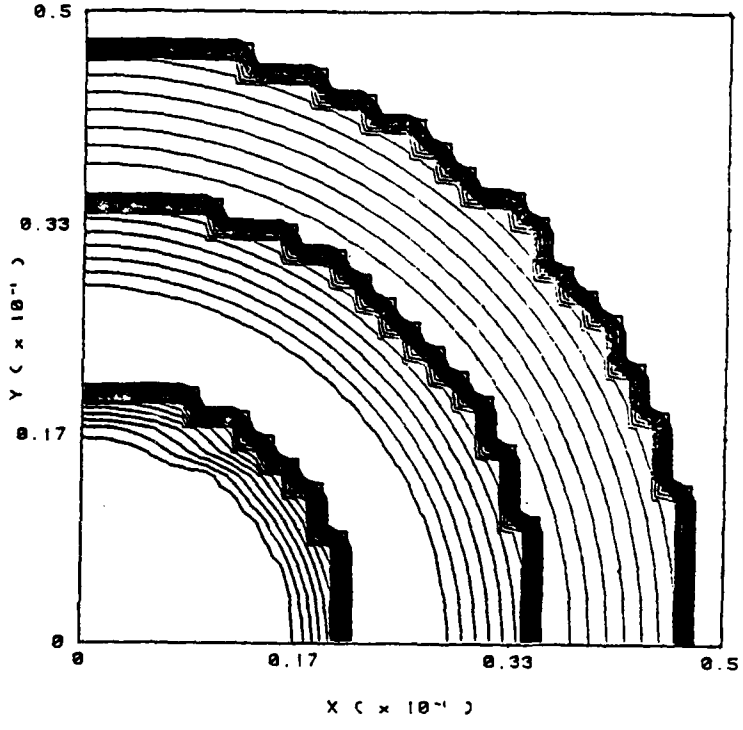


Figure 7. Contours of the density for cylindrical piston using Equation (8). The three separate contour profiles are for $t = 0$, 0.007, and 0.0128.

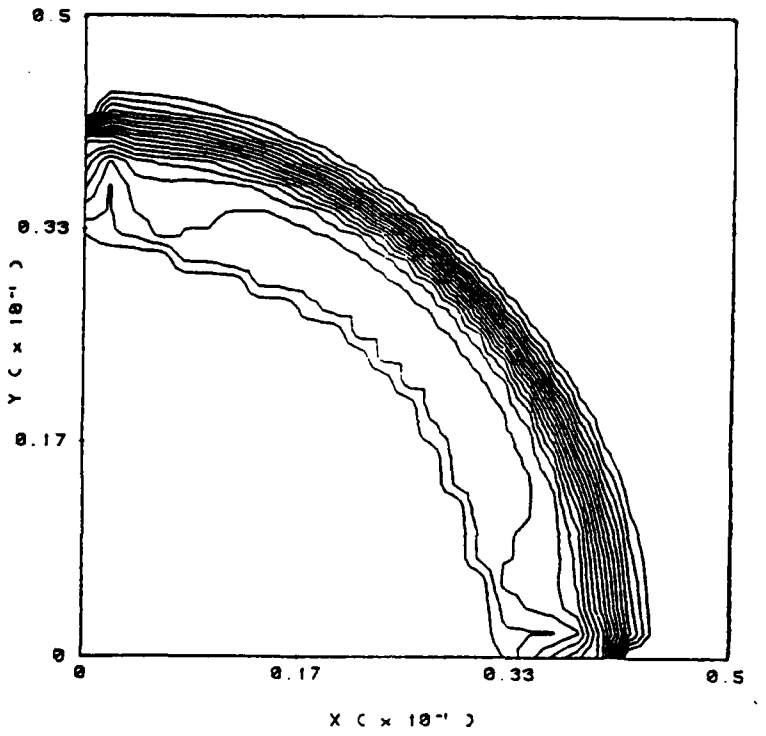


Figure 8. Contours of the density at t = 0.0096 using a uniform, stationary 26 × 26 mesh.

Our adaptive method was used to solve this problem. For comparison, the density computed for t = 0.0096 on a uniform stationary mesh is shown in Figure 8. The meshes at a sequence of four time steps and the density solution for t = 0.0096 using mesh refinement only are shown in Figures 9 and 10, respectively. Four adaptive meshes at different times and the density solution for t = 0.0096 using both mesh moving and refinement are shown in Figures 11 and 12. The algorithm performed a static mesh regeneration at t = 0.0085. Dirichlet boundary conditions were used on the left boundary (x = 0), and a reflective symmetry boundary was implemented on the bottom boundary (y = 0). The results in Figure 12 show the refinement and boundary conditions on the bottom boundary to have performed better than those on the left boundary. Due to memory restrictions on a PRIME 850 minicomputer refinement was restricted to one level and a tolerance of 0.05 was prescribed. A base mesh of 26 × 26 was used for all the computations.

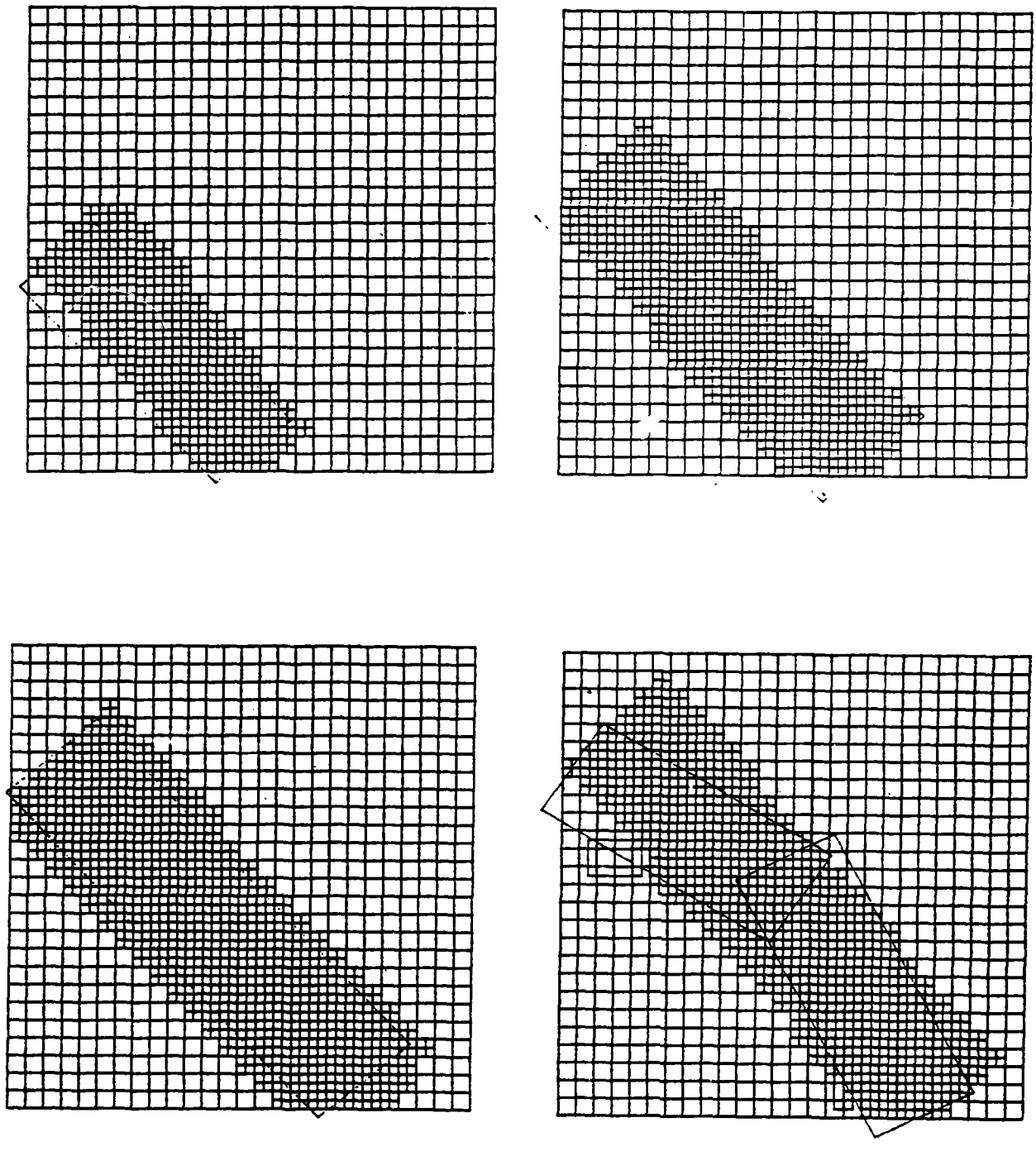


Figure 9. Grids created for the mesh refinement method at t = 0 (upper left), t = 0.0032 (upper right), t = 0.0064 (lower left), and t = 0.0096 (lower right).

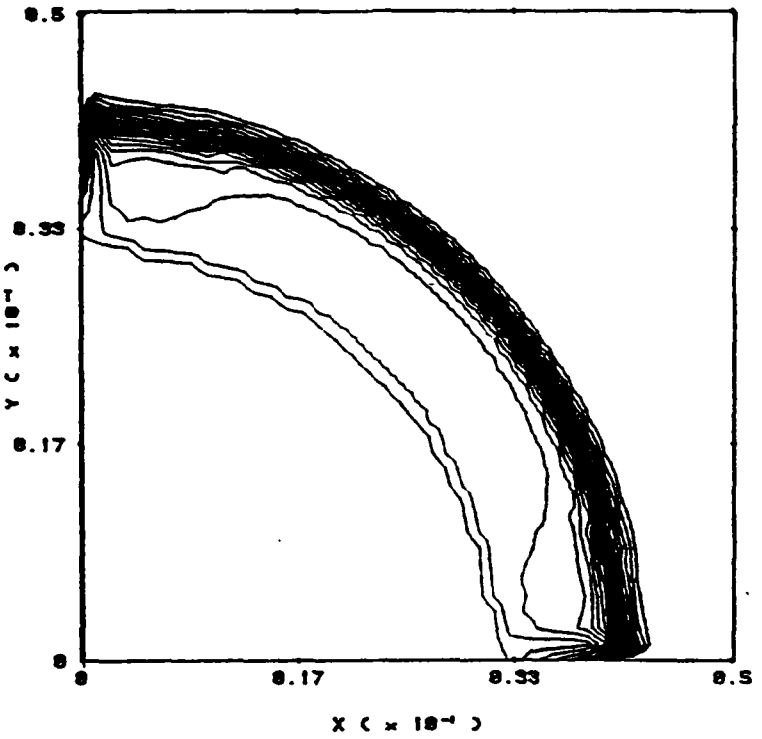


Figure 10. Contours of the density at t = 0.0096 using one-level of mesh refinement on a 26 × 26 base mesh.

5. DISCUSSION. We have described an adaptive procedure for solving the Euler equations in two-space dimensions that combines both mesh moving and local mesh refinement techniques. The algorithm also contains procedures for initial mesh generation and mesh regeneration. We used MacCormack's scheme with Davis's artificial viscosity and a density switch error indicator. This combination of techniques provided good results on example problems while costing less than a comparable uniform mesh calculation or calculation of a comparable solution using mesh refinement only (cf., Arney and Flaherty [3]).

There is still work to be done in order to make use of the power of adaptive methods. Better error estimation is needed so that accurate error can be obtained near discontinuities to avoid excessive mesh refinement. The algorithm must be interfaced with a grid generation package for general domain geometry. The greater efficiency of adaptive techniques will be most beneficial in three dimensions. Therefore, our techniques must be able to take advantage of the latest advances in vector and parallel computing. The tree is a highly parallel structure and we hope to develop a procedure to exploit our mesh refinement data structure on a parallel computing environment. This process will probably include both static and dynamic allocation of multiple processors.

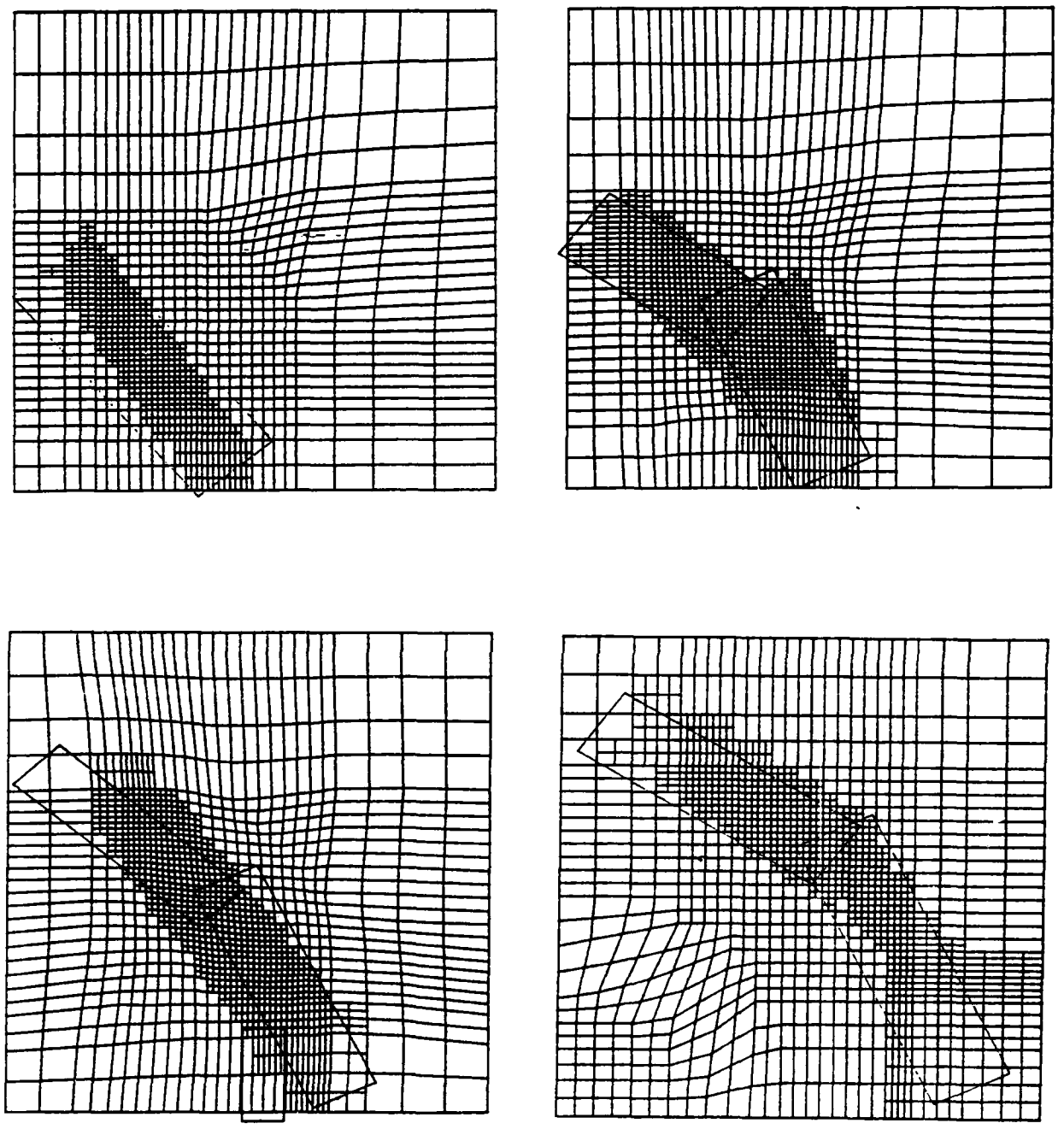


Figure 11. Grids created for the adaptive mesh solution at t = 0 (upper left), t = 0.0032 (upper right), t = 0.00064 (lower left), and t = 0.0096(lower right).

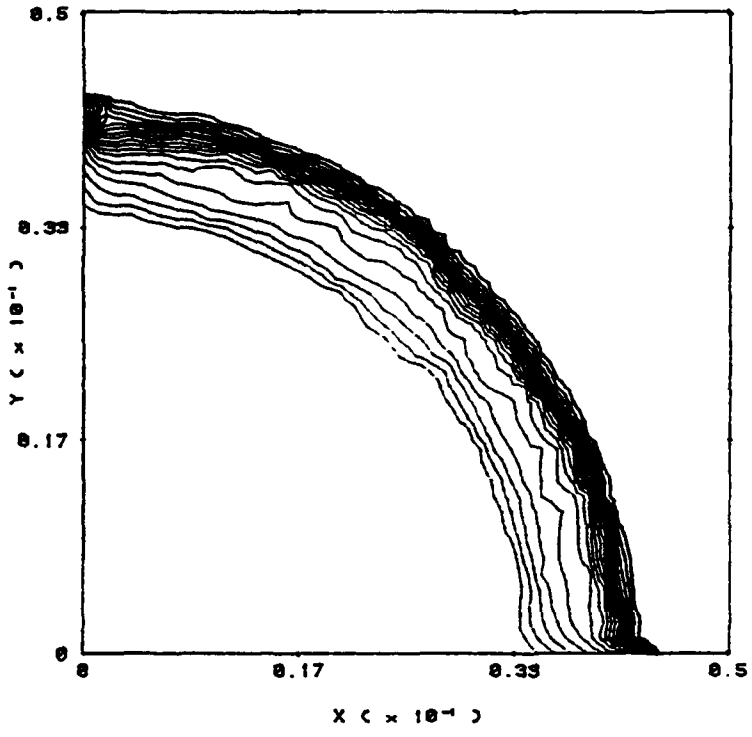


Figure 12. Contours of the density at t = 0.0096 using the adaptive method on a 26 × 26 base mesh.

REFERENCES

1. Adjerid, S. and Flaherty, J.E., A moving finite element method with error estimation and refinement for one-dimensional time dependent partial differential equations, SIAM J. Numer. Anal., 23 (1986), pp. 778-796.

2. Adjerid, S. and Flaherty, J.E., A moving mesh finite element method with local refinement for parabolic partial differential equations, Comp. Meths. Appl. Mech. Engr., 56 (1986), pp. 3-26.

3. Arney, D.C. and Flaherty, J.E., An adaptive local mesh refinement method for time-dependent partial differential equations, Tech. Rep. No. 86-10, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, 1986.

4. Berger, M. and Oliger J., Adaptive mesh refinement for hyperbolic partial differential equations, J. Comput. Phys., 53 (1984), pp. 484-512.

5. Bieterman, M., Flaherty, J.E. and Moore, P.K., Adaptive refinement methods for nonlinear parabolic partial differential equations, Chap. 19 in Accuracy Estimates and Adaptive Refinements in Finite Element Computations, I. Babuska, O.C. Zienkiewicz, J.R. Gago, and E.R. de A. Olivera, Eds., John Wiley and Sons, Chichester, 1986.

6. Arney, D.C. and Flaherty, J.E., A two-dimensional mesh moving technique for time-dependent partial differential equations, J. Comput. Phys., 67 (1986), pp. 124-144.

7. Dwyer, H.A., Grid adaption for problems with separation, cell Reynolds number, shock-boundary layer interaction, and accuracy, AIAA paper No. 83-0449, AIAA 21st Aerospace Sciences Meeting, Reno, 1983.

8. Harten, A. and Hyman, J.M., Self-adjusting grid methods for one-dimensional hyperbolic conservation laws, J. Comput. Phys., 50 (1983), pp. 235-269.

9. Rai, M.M. and Anderson, D., Grid evolution in time asymptotic problems, J. Comput. Phys., 43 (1981), pp. 327-344.

10. Arney, D.C. and Flaherty, J.E., A posteriori error estimation of adaptive finite difference schemes for hyperbolic systems, Trans. Fifth Army Conference on Applied Mathematics and Computing, (1988), pp. 437-457.

11. Babuska, I., Chandra, J., and Flaherty, J.E., Eds., Adaptive Computational Methods for Partial Differential Equations, SIAM, Philadelphia, 1983.

12. Babuska, I., Zienkiewicz, O.C., Gago, J.R. and de A. Olivera, E.R., Eds., Accuracy Estimates and Adaptive Refinements in Finite Element Computations, John Wiley and Sons, Chichester, 1986.

13. MacCormack, R.W., The effect of viscosity in hypervelocity impact cratering, AIAA Paper 69-354, 1969.

14. Davis, S., A Simplified TVD finite difference scheme via artificial viscosity, SIAM J. Sci. Stat. Comput., 8 (1987), pp. 1-18.

15. Arney, D.C. and Flaherty, J.E., An adaptive method with mesh moving and local mesh refinement for time-dependent partial differential equations, Trans. Fourth Army Conf. Appl. Math. Comput., (1987), pp. 1115-1141.

16. Hindman, R., Generalized coordinate forms of governing fluid equations and associated geometrically induced errors, AIAA J, 20 (1982), pp. 1359-1367.

17. Harten, A., High resolution schemes for hyperbolic conservation laws, J. Comput. Phys., 49(1983), pp. 357-393.

18. Pulliam, T.H., Artificial dissipation models for the Euler equations, AIAA J., 24 (1986), pp. 1931-1940.

19. Roe, P.L., Generalized formulation of TVD Lax-Wendroff schemes, ICASE Report No. 84-53, 1984.

20. Flaherty, J.E. and Moore, P.K., An adaptive local refinement finite element method for parabolic partial differential equations, Accuracy Estimates and Adaptive Refinements in Finite Element Computation, (1986) pp. 139-152.

21. Woodward, P. and Collela, P., The numerical simulation of two-dimensional fluid flow with strong shocks, J. Comput. Phys., 54 (1984), pp. 115-173.

22. Ben-Dor, G. and Glass, I.I., Non-stationary oblique shock-wave reflections: Actual isopycnics and numerical experiments, AIAA J., 16 (1978), pp. 1146-1153.

23. Taylor, G.I., The air wave surrounding an expanding sphere, Proc. Roy. Soc., 186 (1946), pp. 273-292.

24. Kimura, T. and Tsutahara, M., Analysis of compressible flows around a uniformly expanding circular cylinder and sphere, J. Fluid Mech., 79 (1977), pp. 625-630.

25. Brantley, M.W., Numerical solution of shock problems in the design of muzzle brakes, United States Military Academy Technical Report 88-2, 1988.

26. Carofano, G., Blast computation using Harten's total variation diminishing scheme, U.S. ARDC Technical Report ARLCB-TR-8409, Benet Weapons Laboratory, Watervliet, NY, October 1984.

HOW TO DESCRIBE OSCILLATIONS OF SOLUTIONS OF NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS

Luc TARTAR
Carnegie Mellon University

I want to describe here some developments in progress concerning the mathematical tools used to describe the relations between microscopic and macroscopic levels. Before describing the new approach, I will first sketch what was my preceding point of view on that question, in order to list the defects of the old classical approach so that one can see which defects will be corrected by the new approach and which one do remain.

A Classical Approach

There are different mathematical models used for the purpose of describing the relations between microscopic and macroscopic levels and the more common one is to use a probabilistic framework : ω denoting the generic point of a probability space Π and E denoting the expectation (i.e. the integration on Π) one can say that if a microscopic variable is denoted by $U(x,\omega)$, then $E(U(x,.))$ will be the associated macroscopic variable.

Another fashionable model is the periodic modulation setting where one deals with functions defined on $\Omega \times R^N$ and periodic in y with unit cell Y. In that model if a microscopic variable is denoted by $U(x,y)$ then the corresponding macroscopic variable is $\int_Y U(x,y)dy/meas(Y)$.

The model that I have been advocating for more than 15 years (which initiated in joint work with F. MURAT) is based on the use of weak convergence : one considers a sequence U_ε of functions defined on an open set Ω of R^N and taking values in R^p; one says that this sequence shows oscillations if it converges weakly but not strongly. Generally if a sequence V_ε converges weakly to V_0 as $\varepsilon \to 0$, then one calls V_ε a microscopic variable and V_0 the corresponding macroscopic variable. This framework extends the periodic setting; indeed if one sets $U_\varepsilon(x) \equiv U(x,x/\varepsilon)$, then as $\varepsilon \to 0$, U_ε converges weakly to U_0 given by $U_0(x) \equiv \int_Y U(x,y)dy/meas(Y)$ (assuming that U is periodic in y and satisfies some regularity hypotheses in (x,y)).

I do not want to go into the technical details of functional analysis involved in the definition of weak and weak* topologies, but it is worth noticing that this point of view is the one used by physicists when they replace a discrete distribution of point masses by a smooth density. The fact that for every continuous function f on [0,L] one has

$$\sum_{i=1}^{m} \frac{L}{m} f(\frac{iL}{m}) \to \int_0^L f(x)dx \text{ as } m \to \infty$$

is equivalent to a weak* convergence of measures, namely

$$\sum_{i=1}^{m} \frac{L}{m} \delta_{\frac{iL}{m}} \to \chi_{[0,L]} \text{ weakly * as } m \to \infty$$

where δ_a denotes the Dirac mass at the point a (it acts on f by evaluating f at a) and $\chi_{[0,L]}$ is the characteristic function of the interval (0,L).

Roughly speaking a sequence of functions U_ε converges weakly to U_0 if $\int_\omega (U_\varepsilon(x)-U_0(x))dx \to 0$ for any measurable set ω while it converges strongly if $\int_\omega |U_\varepsilon(x) - U_0(x)|dx \to 0$.

With these definitions one sees easily that $\sin(x/\varepsilon) \to 0$ weakly while $\sin^2(x/\varepsilon) \to 1/2$ weakly. A similar example, with a simple physical interpretation is the following : if $U_\varepsilon(x) \in R^3$ denotes the microscopic velocity which means the exact velocity for a particle at a point x, then the macroscopic velocity $U_0(x)$ is the average velocity near the point x; the microscopic kinetic energy is $k_\varepsilon(x) = |U_\varepsilon(x)|^2/2$ while the macroscopic total energy is $k_0(x) = |U_0(x)|^2/2 + e(x)$ where $e(x) \geq 0$ denotes the internal energy which is then a macroscopic quantity without analog at the microscopic level.

These examples show that constitutive relations, which are of the form $U_\varepsilon(x) \in$ closed set K of R^p, will not always be satisfied by the macroscopic quantities; indeed one can only say in general that $U_0(x)$ will belong to the closed convex hull of K.

A natural question is then to describe the weak limits of $F(U_\varepsilon)$ for any continuous function F and the answer is given by Young measures : if K is closed and bounded (and after extracting a subsequence) there are probability measures ν_x living on K and depending in a measurable way of x in Ω such that for every continuous function F, $F(U_\varepsilon)$ converges weakly to a limit l_F given by $l_F(x) = \int_K F(k)d\nu_x(k)$. Roughly speaking, making a measurement of the values of $F(U_\varepsilon)$ near x will give a random answer following the probability measure ν_x.

Following this definition one sees that if U(x,y) is periodic in y and we consider $U_\varepsilon(x) \equiv U(x,x/\varepsilon)$ then, under some regularity hypotheses, the Young measures are defined by $\int_K F(k)d\nu_x(k) \equiv \int_Y F(x,y)dy/meas(Y)$.

Another natural question consists in taking into account the fact that the functions that we study (which are related to problems in continuum mechanics or physics) satisfy some differential equations, namely some balance equations

$$\sum_{j=1}^{p}\sum_{k=1}^{N} A_{ijk}\frac{\partial U_j^\varepsilon}{\partial x_k} = f_i^\varepsilon \to f_i^0 \text{ "strongly" for } i = 1,..,q$$

where A_{ijk} are real constants (usually U_ε will converge weakly in $(L^2(\Omega))^p$, and "strongly" will mean $(H^{-1}{}_{loc}(\Omega))^p$ strong, which is implied by $(L^2(\Omega))^p$ weak).

As the balance equations are taken in the sense of distributions the functions U_ε may have jumps accross smooth hypersurfaces; if there is a jump λ in U_ε at a point where the normal is ξ then (λ,ξ) must belong to the following characteristic set V

$$V = \{(\lambda,\xi) \in R^p x S^{N-1} \text{ such that } \sum_{j=1}^{p}\sum_{k=1}^{N} A_{ijk}\lambda_j\xi_k = 0 \text{ for } i = 1,..,q\}$$

The set of possible jumps compatible with the balance equations written in the sense of distributions is then the following characteristic set Λ

$$\Lambda = \{\lambda \in R^p \text{ such that there exists } \xi \in S^{N-1} \text{ with } \sum_{j=1}^{p}\sum_{k=1}^{N} A_{ijk}\lambda_j\xi_k = 0 \text{ for } i = 1,..,q\}$$

Using this characteristic set (which has taken into account some information on the balance equations) one can obtain some new information on limits of quadratic quantities : assume that U_ε converges weakly to U_0 in $(L^2(\Omega))^p$ and $U_{\varepsilon i}U_{\varepsilon j}$ converges weakly to $U_{0i}U_{0j} + R_{ij}$ as measures, then one has the following

Theorem: If U_ε satisfy the balance equations and if $Q(\lambda) \equiv \sum_{ij} q_{ij}\lambda_i\lambda_j$ satisfies $Q(\lambda) \geq 0$ for all $\lambda \in \Lambda$, then $\sum_{ij} q_{ij}R_{ij} \geq 0$.

It means that R, which is always a (measured valued) nonnegative symmetric matrix is constrained by the balance equations through the characteristic set Λ : it must belong to the convex hull of $\{\lambda\otimes\lambda \mid \lambda \in \Lambda\}$.

This theorem, obtained in 1977, extends some results of F. MURAT called compensated compactness, a generalization of a useful remark that we had made in 1974 in our joint work on homogenization, the div-curl lemma.

Although the characteristic set Λ does not contain as much information as the characteristic set V, this theorem led me to a useful ("classical") approach for studying oscillations in the nonlinear partial differential equations of continuum mechanics :

Oscillations are described by the Young measures ν_x constrained by

1. Constitutive relations : the support of ν_x lies on K
2. Balance equations : $\int Q(k)d\nu_x(k) \geq Q(\int k d\nu_x(k))$ for every quadratic Q such that $Q(\lambda) \geq 0$ on Λ.
3. Entropy conditions : one adds any other equations (or inequalities, similar to entropy conditions) implied by constitutive relations and balance equations and apply the two preceding points to the new setting.

One important idea was to show that oscillations were impossible in some situations; indeed if one can show that the only possible Young measures satisfying these constraints were Dirac measures then one would deduce that the subsequence was converging strongly.

I was succesful with this approach for studying a scalar hyperbolic equation and the extension to some hyperbolic systems of conservation laws was made by R. DIPERNA. It covered some cases where other methods were not powerful enough and (although the amount of technical work associated with this method is very important) this approach was followed in subsequent work of D. SERRE, M. RASCLE, C. MORAWETZ, C. DAFERMOS, J. NOHEL among others.

If one could not preclude oscillations, the next question was to study their propagation and interaction, and this required characterizing the structure of the Young measures. I developed such an application to semilinear hyperbolic systems in one space dimension but found that in general one needed to use correlations (which cannot be seen by the Young measures) and did some work in that direction with G. PAPANICOLAOU.

There has been some extensions of these ideas for computing the propagation and interaction of oscillations in linear or semilinear hyperbolic systems by B. ENGQUIST and it has also been extended to some quasilinear hyperbolic systems which are linearly degenerate by D. SERRE and M. BONNEFILLE.

Defects of the Classical Approach

When working on question related to nonlinear partial differential equations of continuum mechanics or physics, it is useful to describe the achievements of a given method but also important to list its limitations. Some of the limitations will be overcome by the new approach, but not all of them.

The real interrelation of the characteristic set V and the constitutive manifold K have not been understood; more differential geometry seems necessary in order to clarify this question, which unfortunately still remains open at the moment.

Oscillations are not the only difficulties encountered in nonlinear partial differential equations and one should also consider the complementary question of concentration, as studied in the work of P.L. LIONS and of R. DIPERNA & A. MAJDA. The new approach will treat oscillations and concentration in a more unified way.

If weak convergence appears to be natural for quantities which can be added (they are usually coefficients of differential forms) there are other quantities for which something different has to be used and for these the adjective averaged should be replaced by effective. This point of view has been developed in the theory of homogenization whose purpose is to understand effective properties of mixtures (periodicity assumptions should be avoided in this context) and it was in connection with homogenization that most of my preceding ideas had been developed. The simple question of computing effective coefficients for layered media (which is well understood) shows the importance of adding to the preceding description at least one geometric parameter for showing the orientation of the layers. Apart from the technical question of finding optimal bounds for effective coefficients (and compensated compactness does play an important role for that purpose) one goal to keep in mind is that one needs to understand the evolution of mixtures. The preceding mathematical tools could not see both the x and the ξ variables and take advantage of the complete characteristic set V; the new one will be able to correct this defect and enable us to address some of the open questions (but obviously not all of them).

Propagation of oscillations and concentration (which is a different matter than the propagation of singularities studied by specialists of linear partial differential equations) cannot be seen by the Young measures ν_x and one idea (which I tried for a long time, without success) is to split the Young measures in directions ξ, so that one could write some kind of transport equation. The new tool will construct a similar object, but not from the Young measures and so will contain a different information, but a compensation for this loss of information will be the useful properties of the new measures, which I have called H-measures in order to remind of their origin in homogenization theory.

A New Mathematical Tool : H-Measures

For a subsequence U_ε converging weakly to 0 in $(L^2(\Omega))^p$ we will define a family of measures $\mu_{x,\xi}$ indexed by $x \in \Omega$ and $\xi \in S^{N-1}$. They will give a better description of the oscillations and their propagation through some kind of microlocal H-calculus enabling us to use in a better way the balance equations. They do not contain all the information of the Young measures based on the constitutive relations, but they will improve the compensated compactness theorem.

For ϕ_1 and ϕ_2 in $D(\Omega)$ and ψ in $D(S^{N-1})$ one defines the H-measure μ with entries μ_{ij} by computing the following limits as $\varepsilon \to 0$ (after extracting some subsequence)

$$\langle \mu_{ij}, \phi_1(x)\phi_2(x)^*\psi(\xi)\rangle = \text{limit} \int_{R^N} F(\phi_1 U_i^\varepsilon) F(\phi_2 U_j^\varepsilon)^* \psi(\xi) d\xi \text{ as } \varepsilon \to 0$$

where F denotes Fourier transform ($Ff(\xi) = \int f(x) e^{2i\pi x.\xi} dx$) and z^* denotes the complex conjugate of z.

Note that there is indeed something to prove here so that the above definition makes sense, namely that if ϕ_1 and ϕ_2 have disjoint support then the above limit is 0; this is the reason why the H-measures are only defined for sequences converging weakly to 0.

Although there is some analogy with the definition of the wave front set of a distribution as it was given by HÖRMANDER, the framework here is entirely different : we deal with sequences and are interested in the difference between weak convergence and strong convergence (which is the case where the H-measure is 0).

Because the function ψ are homogeneous of degree 0, the H-measures cannot distinguish between different frequencies for propagation in some direction, which they will be able to describe.

The H-measures will give us some results in small amplitude homogenization which were traditionnally obtained using 2-point correlations (which can be easily defined both in a periodic or in a random setting, but not in a general case without the use of a characteristic length); the H-measures do not contain the information on 2-point correlations, but can be deduced from them by a singular integral, as was pointed out to me at a later meeting by M. AVELLANEDA who had worked in the random case. At the same time G. MILTON pointed out that one could not expect to find results about scattering unless constructing similar measures using 3-point correlations, but it is not clear if such measures can be constructed that would also retain the other properties that I wanted, namely the use of balance equations to give some information on propagation.

Example 1: Periodic modulation.
If $U_\varepsilon(x) = V(x, x/\varepsilon)$ with $V(x,y)$ having period 1 in each y coordinate, then (under some regularity hypothesis) V admits a Fourier decomposition $V(x,y) = \Sigma'_m v_m(x) e^{2i\pi m.y}$ and the H-measure is equal to $\Sigma'_m |v_m(x)|^2 \delta_{m/|m|}$ where the sum is taken over $m \in Z^N \backslash 0$ ($v_0 = 0$ by hypothesis).

Example 2: Concentration effects.
If $U_\varepsilon(x) = \varepsilon^{-N/2} f(x/\varepsilon)$ for an L^2 function f, then the H-measure is equal to $\delta_0 \otimes \nu$ where ν has a surface density $\nu(\xi) = \int_0^\infty |Ff(t\xi)|^2 t^{N-1} dt$.

Localization Property : Balance equations restrict the H-measures $\mu_{x,\xi}$ to be a combination of hermitian nonnegative matrices of the form $\lambda\otimes\lambda^*$ with $\lambda \in C^p$ such that $(\lambda,\xi) \in$ characteristic set V.

Propagation Property : One assumes that $U_\varepsilon \to 0$ in $(L^2(\Omega))^p$ weak and

$$\sum_{j,k} A_{ijk} \frac{\partial U_{\varepsilon j}}{\partial x_k} = v_{\varepsilon i} \to 0 \text{ in } L^2(\Omega) \text{ weak}$$

then, in some cases, one can obtain some differential equations satisfied by the H-measure $\mu_{x,\xi}$ describing the oscillations of U_ε. It involves a creation term using another H-measure $\nu_{x,\xi}$ describing the joint oscillations of U_ε and v_ε.

I have not elucidated yet what are the algebraic computations to perform in the general case, but in classical examples it does give what the intuition suggests (based on physics or linear theory). If dealing with the wave equation one does obtain the expected transport equation for the H-measures $\mu_{x,\xi}$; however one does not find here any role for caustics because the H-measures describing only the amplitude and the direction of propagation of both oscillations and concentration effects cannot feel the changes of phase that happen when crossing caustics.

In the case of a scalar equation one can see more easily the difference between the static localization property and the dynamic propagation property. It is worth pointing out that our test function ϕ_j or ψ can be chosen to be only continuous and that enable us to derive some kind of pseudo-differential calculus with zero order operators having only a continuous symbol. Let us consider a simple linear equation

$$\sum_{i=1}^{N} a_i(x) \frac{\partial u^\varepsilon}{\partial x_i} + b(x) u^\varepsilon = 0$$

where the coefficients a_i are of class C^1 while b is of class C^0.

On one hand the localization property says that the H-measure $\mu_{x,\xi}$ for u^ε (assuming that u^ε converges weakly to 0) will be supported by the zero set of the function P defined by $P(x,\xi) = \Sigma_j a_j(x)\xi_j$.

On the other hand the propagation property states that for every test function $\Phi(x,\xi)$ of class C^1 on $\Omega x S^{N-1}$ with compact support in x, one has

$$<\mu, \sum_{k=1}^{N} \frac{\partial \Phi}{\partial \xi_k} \frac{\partial P}{\partial x_k} - \frac{\partial \Phi}{\partial x_k} \frac{\partial P}{\partial \xi_k}> + 2<\mu, b\Phi> = 0$$

(Φ is extended as an homogeneous function of degree 0 in ξ); it implies that oscillations and concentration effects do propagate along bicharacteristic rays.

When u^ε is solution of different first order equations then the H-measure

is supported by the intersection of the zero sets of the symbols of the first order operators and satisfies different equations which may happen to be incompatible constraints forcing μ to be 0 and thus precluding oscillations.

Applications to Homogenization

Effective properties of mixtures cannot be obtained from the knowledge of averages or weak limits (except in special situations like layered materials) and it cannot be deduced from H-measures either. In some cases however the introduction of H-measures can help reformulate the problem. For example H-measures are useful in deriving good bounds for effective coefficients, but my results in this direction are too fragmentary at the moment to explain what is the best way to use them.

H-measures are the right mathematical tool for studying small amplitude homogenization; the formula obtained are analog to some that were known to specialists in a periodic or in a random setting. If one considers the diffusion (of electricity or heat) in a mixture of materials of near by conductivities

$$-\mathrm{div}(A^{\varepsilon}(x;\gamma)\mathrm{grad}U^{\varepsilon}(x;\gamma)) = \rho^{\varepsilon} \text{ in } \Omega$$

where

$$A^{\varepsilon}(x;\gamma) = A^{0}(x) + \gamma B^{\varepsilon}(x) + \gamma^{2}C^{\varepsilon}(x) + o(\gamma^{2}) \text{ with } B^{\varepsilon} \rightarrow B^{0} \text{ and } C^{\varepsilon} \rightarrow C^{0} \text{ weakly}$$

then the effective conductivity has the form

$$A_{\mathrm{eff}}(x;\gamma) = A^{0}(x) + \gamma B^{0}(x) + \gamma^{2}\{C^{0}(x) - M(x)\} + o(\gamma^{2})$$

where the correction term M(x) (which is nonnegative) can be computed from the H-measure associated to the sequence $\{B_{\varepsilon} - B_{0}\}$ by a specific integration in ξ on S^{N-1}; for example asssuming that B_{ε} is isotropic so that we have only one scalar H-measure $\mu_{x,\xi}$, the formula is

$$M(x) = \int_{S^{N-1}} \frac{\xi \otimes \xi}{(A_{0}(x)\xi.\xi)} d\mu_{x,\xi}.$$

This procedure extends to other models like linear elasticity and can give relations between different effective properties of a given mixture.

H-measures can also give the exact answer for the effective behaviour in some cases where the oscillating coefficients do not appear in the highest order terms. The following example, which has some similarities with stationary Navier-Stokes equations, is instructive : if u^{ε} is a velocity field solution of

$$-\Delta u^{\varepsilon} + u^{\varepsilon} x [\mathrm{curl} v^{\varepsilon}] + \mathrm{grad} p^{\varepsilon} = f \text{ and } \mathrm{div} u^{\varepsilon} = 0 \text{ in } \Omega$$

where $v_{\varepsilon} \rightarrow v_{0}$ weakly; then the effective equation is

$$-\Delta u^0 + u^0 x [\mathrm{curl} v^0] + M u^0 + \mathrm{grad} p^0 = f \text{ and } \mathrm{div} u^0 = 0 \text{ in } \Omega$$

and the correction term M^0 can be computed from the H-measure associated to the sequence $\{v_\varepsilon - v_0\}$ by a specific integration on S^{N-1}; the formula is

$$M(x) = -\int_{S^{N-1}} \xi \otimes \xi \{ \sum_k d\mu^{kk}_{x,\xi} - \sum_{k,l} \xi_k \xi_l d\mu^{kl}_{x,\xi} \}$$

Such formulas should be useful in order to understand turbulence effects, at least at their onset.

Conclusion :

There are many other areas where I plan to use this simple new approach based on H-measures, and stability questions in continuum mechanics is one of them, but I have a bolder conjecture, i.e. that H-measures may be one of the missing mathematical tools necessary to explain why some rules invented by physicists work so well, in spite of their irrational derivations (it was for understanding situations obviously related to the difficult homogenization problem of propagation of waves in mixtures that physicists have invented quantum mechanics). The computation of the correction terms that have appeared in some of the examples which I have described above present striking analogies with some which are done by following some dogmatic rules of quantum mechanics; those that I have made were entirely deductive and part of a general program of study of nonlinear partial differential equations.

More detailed proofs of the constructions sketched here will appear elsewhere.

REGISTERED ATTENDEES

Myron B. Allen
Department of Mathematics
University of Wyoming
Laramie, Wy 82071

David C. Arney
Department of Mathematics
United States Military Academy
West Point, NY 10996

John B. Bdzil
Group M-9, Mail Stop P952
Los Alamos National Laboratory
Los Alamos, NM 87544

Professor J. Bebernes
Department of Mathematics
Campus Box 426
University of Colorado
Boulder, CO 80309

Professor Ted Belytschko
Department of Civil Engineering
The Technological Institute
Northwestern University
Evanston, IL 60208

Prof. Carlos A. Berenstein
Math. Dept.
Univ. of Maryland
College Park, MD 20742

Vicki Bergmann
T & AM, Kimball Hall
Cornell University
Ithaca, NY 14853

Christian Bischof
Dept. of Computer Science
Upson Hall, Cornell University
Ithaca, NY 14853

Prof. David A. Caughey
218 Upson Hall
Cornell University
Ithaca, NY 14853

B. F. Caviness
National Science Foundation, Rm 304
1800 G Street
Washington, DC 20008

Dr. Jagdish Chandra
US Army Research Office
Mathematical Sciences Division
P. O. Box 12211
Research Triangle Park, NC 27709

T. Y. Chang
Department of Civil Engineering
University of Akron
Akron, OH 44325

Peter C. T. Chen
U.S. Army Research, Development, & Engineering Center
AMCCOM
Watervliet Arsenal
Watervliet, NY 12189-4050

Herbert E. Cohen
AMSAA
U.S. Army Materials Systems Analysis Activity
Aberdeen Proving Ground, MD 21005

Joel Cohen
Dept. of Math & Computer Science
University of Denver
Denver, CO 80208

Vincent T. Coppola
114 Thurston Hall
Cornell University
Ithaca, NY 14853

Dr. Terence M. Cronin
US Army Center for Signals Warfare
Vint Hill Farms Station
Warrenton, VA 22186

Gideon A. Culpepper
Box 14
Mesilla Park, NM 88047

Benjamin Cummings
P. O. Box 96
Churchville, MD 21028-0096

David F. Delchamps
School of Electrical Engineering
Cornell Univ. (Phillips Hall)
Ithaca, NY 14853

W. Donovan
M/S Branch; Interior Ballistics Div.
Ballistic Research Lab.
Aberdeen Proving Grd., MD 21005

Phillip Dykstra
U.S. Army, BRL
Aberdeen Proving Grounds, MD 21005-5066

Walter O. Egerland
SLCBR-SE-C
U.S. Army Ballistic Research Lab.
Aberdeen Proving Ground, MD 21005-5066

Magnus Ewerbring
409 Phillips Hall
Cornell University
Ithaca, NY 14853

Charbel Farhat
Dept. of Aerospace Eng.
Univ. of CO at Boulder
Boulder, CO 80309

Walter T. Federer
337 Warren Hall
Cornell University
Ithaca, NY 14853

Kurt Fickie
Ballistic Research Lab
Aberdeen Proving Ground, MD 21005-5066

Rodrigo Fontecilla
Univ. of MD
Department of Computer Science
College Park, MD 20742

Isaac Fried
Boston University
Mathematics
Boston, MA 02215

Dr. Gopal Gaonkar
Department of Mechanical Engineering
Florida Atlantic Univ.
Boca Raton, FL 33431

John Groff
1590 Theadore Rd.
Rising Sun, MD 21911

Dr. John W. Grove
Courant Inst. of Math. Sci.
New York University
251 Mercer St.
NY, NY 10012

Aaron D. Gupta
Rm 228 B 309 TBD
USA Ballistic Res. Lab.
Aberdeen Proving Grounds, MD 21005-5066

Karl Gustafson
Dept. of Math
Univ. of CO at Boulder
Boulder, CO 80309

Prof. Philip Hall
Exeter Univ.
Exeter, England

Prof. John K. Hunter
Math. Dept
Colorado State Univ.
Ft. Collins, CO 80523

M. A. Hussain
G. E. Research Center
General Electric Company
P. O. Box 8
Schenectady, NY 12301

Dr. William Jackson, Commander
US Army Tank-Automotive Command
Warren, MI 48397-5000

Dr. Arthur R. Johnson
U. S. Army Materials Tech. Lab
Watertown, MA 02172-0001

T. Kailath
Dept. of Elec. Eng.
Stanford Univ.
Stanford, CA 94305-4055

A. K. Kapila
Dept. Math. Sciences
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

D. R. Kassoy
Mech. Eng. B-427
University of CO at Boulder
Boulder, CO 80309

Robert L. Launer
614 New Keat Pl.
Cary, NC 27511

Siegfried H. Lehnigk, Commander
U.S. Army Missile Command
ATTN; AMSMI-RD-RE-OP
Redstone Arsenal, AL 35898-5248

Dening Li
Math. Dept.
Univ. of Colorado at Boulder
Boulder, CO 80309

Zhiming Li
5130 E. Ashbury Ave. 304
Denver, CO 80222

Guopoing Liu
2475 S. York, #103
Denver, CO 80210

Franklin Luk
School of EE
Cornell Univ.
Ithaca, NY 14853

Clarence J. Maday
MAE Dept
North Carolina State Univ.
Raleigh, NC 27695-7910

David S. Malkus
Ctr. for Math. Sci.
610 Walnut Street
Madison, WI 53705

Ralph Menikoff
Los Alamos National Lab
Theoretical Div., MS-B214
Los Alamos, NM 87545

Miles Miller
U.S. Army Chemical Research, Development & Engineering Center
Attn: SMCCR-RSP-A
Aberdeen Proving Grounds, MD 21010-5423

Mike Muuss
U.S. Army Ballistic Research Lab
Aberdeen Proving Grounds, MD 21005-5066

Prof. J. Nagbhushana
Dept. Mech Eng.
Florida Atlantic Univ.
Boca Raton, FL 33431

Yoshisuke Nakano
U.S. Army CRREL
Hanover, NH 03755

Albert Nigrin
2701 Sarah Ave.
Durham, NC 27707

Youngtae Park
Dept. of Electrical Engineering
Univ. of California Irvine
Room 208, IERF
Irvine, CA 92717

Prof. Richard Rand
Dept. TAM, Kimball Hall
Cornell Univ.
Ithaca, NY 14853

Major John S. Robertson
Dept. of Math.
U.S. Military Academy
West Point, NY 10996-1786

Nathanael Roman
P. O. Box 2394
Las Cruces, NM 88004

J. Sahu
Ballistic Research Lab
Aberdeen Proving Grounds, MD

Dr. M. Sambandham
Box 287
Atlanta University
Atlanta, GA 30314

Dr. Joseph M. Santiago
U.S. Army Ballistic RSCH Lab
Aberdeen Proving Gd., MD 21005

Bobby Schnabel
Campus Box 430
Univ. of CO at Boulder
Boulder, CO 80309

Dr. L. S. Shieh
Dept. of Electrical Engineering
Univ. of Houston
Houston, TX 77004

Royce Soanes
Benet Labs
Watervliet, NY 12189-4050

Moss Sweedler
Dept. of Math
White Hall
Cornell University
Ithaca, NY 14853

Prof. Luc Tartar
Dept. of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Alexander Tessler
Army Mater. Tech. Lab
Watertown, MA

Bill Troy
Math. Dept.
U. of Pittsburgh, PA 15260

Dr. Aynur Unal
NASA Ames Res. Lab., 215-1
Moffett Blvd. CA 94035

Xiaodong Zhang
1300 30th St. B4-11
Boulder, CO 80303

John Vasilakis
U.S. Army Research, Development & Eng. Ctr.
AMCCOM
Watervliet Arsenal
Watervliet, NY 12189-4050

Dr. H. Baussus von Luetzow
8021 Garlot Dr.
Annandale, VA 22003

Paul S. Wang
Dept. of Mathematical Sciences
Kent State Univ.
Kent, OH 44242

Richard A. Weiss
U.S. Army Engineer
Waterways Experiment Station
Vicksburg, MS 39180

Michael Wester
1801 Quincy, SE
Albuquerque, NM 87108

Prof. Chelsea C. White, III
Dept. of Systems Engineering
Univ. of VA - Thornton Hall
Charlottesville, VA 22901

Dr. J. R. Whiteman
Dept. of Math.
Texas A & M Univ
College Station, TX 77843-3368

Dr. Arthur Wouk
US Army Research Office
Math. Sciences Division
P. O. Box 12211
Research Triangle Park, NC 27709

Leszek S. Zaremba
Mathematical Reviews
416 Fourth Street
Ann Arbor, MI 48107

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188
Exp. Date Jun 30, 1986

1a. REPORT SECURITY CLASSIFICATION: Unclassified

1b. RESTRICTIVE MARKINGS:

2a. SECURITY CLASSIFICATION AUTHORITY:

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE:

3. DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release: Distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S): ARO Report 89-1

5. MONITORING ORGANIZATION REPORT NUMBER(S):

6a. NAME OF PERFORMING ORGANIZATION: Army Research Office

6b. OFFICE SYMBOL (If applicable): SLCRO-MA

7a. NAME OF MONITORING ORGANIZATION:

6c. ADDRESS (City, State, and ZIP Code): P.O. Box 12211
Research Triangle Park, NC 27709-2211

7b. ADDRESS (City, State, and ZIP Code):

8a. NAME OF FUNDING / SPONSORING ORGANIZATION:

8b. OFFICE SYMBOL (If applicable):

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER:

8c. ADDRESS (City, State, and ZIP Code):

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification): Transactions of the Sixth Army Conference on Applied Mathematics and Computing

12. PERSONAL AUTHOR(S):

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM 5/31/88 TO 6/3/88	1989 February	1150

16. SUPPLEMENTARY NOTATION:

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Fluid and solid mechanics, mathematical physics and numerical methods, symbolic computation, control theory, and stochastic techniques.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

(U) This is a technical report resulting from the Sixth Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat many Army applied mathematical problems.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION:

22a. NAME OF RESPONSIBLE INDIVIDUAL: Dr. Francis G. Dressel

22b. TELEPHONE (Include Area Code): (919) 549-0641

22c. OFFICE SYMBOL: SLCRO-MA

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete.