

DTIC FILE COPY

ARO Report 90-1

(S)

TRANSACTIONS OF THE SEVENTH ARMY  
CONFERENCE ON APPLIED MATHEMATICS  
AND COMPUTING

AD-A219 220

DTIC  
ELECTE  
MAR 15 1990  
S D CS D



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

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE ASSISTANT SECRETARY OF THE ARMY FOR  
RESEARCH, DEVELOPMENT, AND ACQUISITION

00 03 15 035

U.S. ARMY RESEARCH OFFICE

Report No. 90-1

February 1990

TRANSACTIONS OF THE SEVENTH ARMY CONFERENCE  
ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Host

U.S. Military Academy  
West Point, New York

6-9 June 1989

Accession For	
NTIS CRAMI	✓
DTIC TAB	✓
Unannounced	✓
Justification	
By	
Date	
Authoring Org	
Availability	
Notes	
A-1	

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

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



FOREWORD

The Seventh Army Conference on Applied Mathematics and Computing was held at the U.S. Military Academy, West Point, New York, on 6-9 June 1989. This is the second time the Military Academy has served as the host for this series of Army conferences. For each of these meetings we were fortunate to have the heads of the Department of Mathematics as Chairpersons on Local Arrangements. This year Colonel Frank Giordano served in this capacity. He was assisted in this task by Lieutenant Colonel David Arney and Captain Suzanne Swann. These individuals are to be commended for their efforts in coordinating all the details required to conduct this large successful scientific meeting.

This 1989 conference was attended by more than 80 scientists and engineers representing academia and various Army agencies. The meeting featured seven invited speakers. These general talks covered several topics of current interest, including multi-scale methods and wavelet transforms, high performance computing, phase transformations, multivariate splines, and stochastic control. The names of these speakers, together with the titles of their addresses, are listed below. The second part of the program consisted of special sessions on topics such as stochastic methods for image analysis, mathematical issues in computer science, computational methods for multibody dynamics, and mechanics of large deformations. In addition, about 40 contributed papers were presented by both Army and academic participants.

SPEAKER AND AFFILIATION

TITLE OF ADDRESS

Professor Alan S. Willsky  
Massachusetts Institute  
of Technology

Estimation of Spatially-Distributed  
Processes

Professor Richard D. James  
University of Minnesota

Microstructure of Crystals Undergoing  
Phase Transformation

Professor Robert V. Pohn  
New York University

Modelling Microstructure by Energy  
Minimization

Professor S. Lennart Johnsson  
Yale University

High Performance Computing

Professor Mark H.A. Davis  
Imperial College of Science

Theory and Application of Piecewise-  
Deterministic Processes

Professor A. Cohen  
Universite of Paris-  
Dauphine

Wavelet Transforms

Professor Carl de Boor  
University of Wisconsin-  
Madison

What's New in Multivariate Splines?

One of the sessions at this conference was called "Mathematics at West Point." In it, members of the Department of Mathematics outlined a new program for the cadets entitled "USMA's Mathematics Program in 1990 and Beyond." The first article in these proceedings is devoted to this curriculum.

This conference is part of a continuing program of Army-wide symposia held under the auspices of the Army Mathematics Steering Committee (AMSC) to promote better communication between Army scientists and the Army Research Office investigators. In order that this mission be accomplished, a large number of scientists had to expend a great deal of effort. The members of the AMSC would like to thank all these individuals for their excellent presentations and their valuable contributions to the field of science.

Part 2  
TABLE OF CONTENTS:

<u>Title</u>	<u>Page</u>
Foreword.....	iii
Table of Contents.....	v
Agenda.....	ix
USMA'S MATHEMATICS PROGRAM FOR 1990 AND BEYOND David C. Arney, Lee S. Dewald, Sr., and John R. Edwards.....	1
MODELING AND ESTIMATION FOR MULTIREOLUTION STOCHASTIC PROCESSES A. S. Willsky.....	13
ROBUST IMAGE MODELS FOR IMAGE RESTORATION AND TEXTURE EDGE DETECTION R. L. Kashyap and Kie Bum Eom.....	21
ON THE STROH FORMALISM FOR ANISOTROPIC ELASTICITY AND ITS APPLICATIONS TO COMPOSITES T. C. T. Ting.....	69
TOTAL ABSORPTION IN ELASTIC MEDIA William W. Hager and Rouben Rostamian.....	85
TRANSIENT SHEAR RESPONSE OF A RIGID BLOCK AND FLEXIBLE SUPPORT ASSEMBLY SUBJECTED TO A LATERAL IMPACT Aaron Das Gupta.....	105
ON THE CONTINUUM MECHANICS OF THE MOTION OF A PHASE INTERFACE Morton E. Gurtin.....	113
NONLINEARITY OF INVERSE PROBLEMS T. Mura and Z. Gao.....	117
QUADRATIC DYNAMICAL SYSTEMS DESCRIBING SHEAR FLOW OF NON-NEWTONIAN FLUIDS D. S. Malkus, J. A. Nohel, and B. J. Plohr.....	131
SMART ALGORITHMS FOR COMPLEX PROBLEMS IN FLUID DYNAMICS J. Tinsley Oden.....	149
EFFECT OF CONSTITUTIVE MODELLING ON THE DYNAMIC DEVELOPMENT OF SHEAR BANDS IN VISCOPLASTIC MATERIALS R. C. Batra and C. H. Kim.....	183
A PROPERTY OF LINEAR FEEDBACK SHIFT REGISTER SEQUENCES Harold Fredricksen and Gary Krahn.....	193

DIGITAL REDESIGN OF PSEUDO-CONTINUOUS-TIME SUBOPTIMAL REGULATORS FOR LARGE-SCALE DISCRETE SYSTEMS Jason S.H. Tsai, Leang S. Shieh, Jian L. Zhang, and Norman P. Coleman.....	205
NEW METHODOLOGIES IN RENEWAL THEORY B. D. Sivazlian.....	239
STEPWISE CLOSED FORM TECHNIQUES FOR COMPUTER SIMULATION OF GUIDED PROJECTILES M. J. Amoruso, R. Campbell, and H. Cohen.....	261
GROBNER BASIS OPTIMIZATION Moss Sweedler and Lee Taylor.....	271
THE RELATIONSHIP BETWEEN LINEAR AND NONLINEAR VARIATIONAL MODELS OF COHERENT PHASE TRANSITIONS Robert V. Kohn.....	279
RELATION BETWEEN MICROSCOPIC AND MACROSCOPIC PROPERTIES IN CRYSTALS UNDERGOING PHASE TRANSFORMATION R. D. James.....	305
CONCURRENT SPECIFICATIONS AND THEIR GUREVICH-HARRINGTON GAMES AND REPRESENTATION OF PROGRAMS AS STRATEGIES Alexander Yakhnis.....	319
EXTRACTION OF CONCURRENT PROGRAMS FROM GUREVICH-HARRINGTON GAMES Vladimir Yakhnis.....	333
UNSTABLE INTERFACES AND ANOMALOUS WAVES IN COMPRESSIBLE FLUIDS John W. Grove, Ralph Menikoff, and Qiang Zhang.....	345
CHARACTERISTICS AND STABILITY IMPLICATIONS OF A STREAMWISE VORTEX IN BOUNDED SHEAR FLOW Joseph D. Myers and Frederick H. Abernathy.....	371
GENERALIZED REWRITING IN TYPE THEORY David A. Basin.....	429
SOLVING THE EULER EQUATIONS USING ADAPTIVE MESH MOTION AND REFINEMENT David C. Arney, Rupak Biswas, and Joseph E. Flaherty.....	441
LINE ITERATIVE METHODS FOR CYCLICALLY REDUCED NON-SELF-ADJOINT ELLIPTIC PROBLEMS Howard C. Elman and Gene H. Golub.....	457
A COMPUTER SIMULATION OF THE FREELY-ASSOCIATING NEOCORTEX M. Johnson, R. Scanlon, and M. Cipollo.....	467

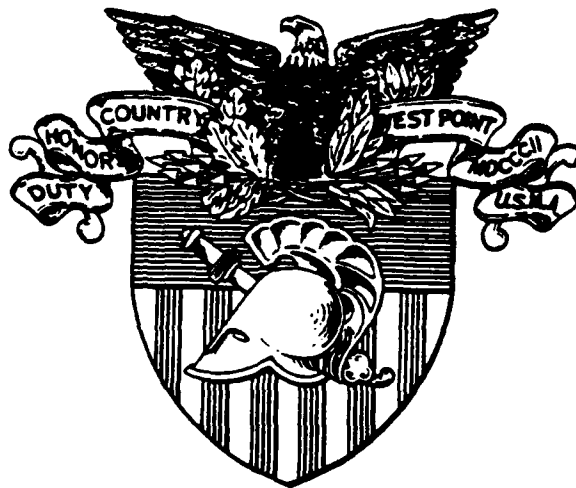
ON TOPOLOGICAL COMPLEXITY OF SOLVING POLYNOMIAL EQUATIONS OF SPECIAL TYPE A. Libgober.....	475
DISCRETE APPROXIMATION OF DECONVOLUTION OPERATORS Carlos A. Berenstein and B. A. Taylor.....	479
SYMBOLIC UNCOUPLING AND EFFICIENT SOLUTION OF TREE-STRUCTURED LINEAR EQUATION SYSTEMS Roger A. Wehage.....	491
APPLICATION OF GENERALIZED NEWTON-EULER EQUATIONS AND RECURSIVE PROJECTION METHODS TO DYNAMICS OF DEFORMABLE MULTIBODY SYSTEMS Roger A. Wehage and Ahmed A. Shabana.....	527
AUTOSIM: A COMPUTER LANGUAGE FOR REPRESENTING MULTIBODY SYSTEMS IN SYMBOLIC FORM TO AUTOMATICALLY FORMULATE EFFICIENT SIMULATION CODES Michael W. Sayers.....	547
EFFECT OF THERMAL SOFTENING ON THE RESPONSE OF SHEARING MOTIONS Athanasios E. Tzavaras.....	581
ELASTIC-PLASTIC ANALYSIS OF A THICK-WALLED COMPOSITE TUBE SUBJECTED TO INTERNAL PRESSURE Peter C.T. Chen.....	589
ULTRAFAST THERMODYNAMIC PROCESSES Richard A. Weiss.....	599
THE INTERNAL PHASE STRUCTURE OF ATOMS Richard A. Weiss.....	609
NEWTONIAN GRAVITY IN MATTER WITH BROKEN INTERNAL SYMMETRY Richard A. Weiss.....	637
WAVE PROPAGATION IN ASYMMETRIC MEDIA Richard A. Weiss.....	691
ELASTIC DEFORMATION AND SLUG FLOW AS APPLICATIONS OF FRONT TRACKING X. Garaizar, J. Glimm, and W. Guo.....	705
NUMERICAL SOLUTION OF AN APERTURE ANTENNA INTEGRAL EQUATION M. A. Hussain, Ben Noble, Wen-Tai Lin, and B. Becker.....	719
COMPUTATION OF LEADING EIGENSPACES FOR GENERALIZED EIGENVALUE PROBLEMS Abraham Kribus.....	735
APPLICATIONS OF FIBONACCI SEQUENCES AND TILING Joseph Arkin, David C. Arney, Lee S. Dewald, and Charles Kennedy.....	743

APPROXIMATION AND INTERPOLATION FORMULAS FOR REAL-TIME APPLICATIONS Charles K. Chui and Harvey Diamond.....	765
AN ENHANCED KNOT SELECTION ALGORITHM FOR LEAST SQUARES APPROXIMATION USING THIN PLATE SPLINES John R. McMahon and Richard Franke.....	773
A MULTIVARIATE EXTENSION OF THE CRAMER-VON MISES TEST FOR GAUSSIANTITY Kevin M. Beam and Albert S. Paulson.....	797
A GENERALIZED HARMONIC BALANCE METHOD FOR FORCE NONLINEAR OSCILLATIONS - NUMERICAL SOLUTION FORMULATION AND RESULTS B. Noble, M. A. Hussain, and J. J. Wu.....	837
OPTIMIZED ANNULUS-BASED POINT-IN-REGION INCLUSION TESTING FOR d DIMENSIONS T. M. Cronin.....	863
List of Attendees.....	881



Agenda for the  
Seventh Army Conference on  
Applied Mathematics and  
Computing

6-9 June 1989



U.S. Military Academy  
West Point, New York

SEVENTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

U.S. Military Academy, West Point, New York

6-9 June 1989

AGENDA

Tuesday, June 6, 1989

- 0800 - 1600      Registration\* - Bartlett Hall, Room 220
- 0800 - 0830      Opening Remarks - Bartlett Hall, Room 220  
                    Frank Giordano, U.S. Military Academy, West Point, NY
- 0830 - 0930      General Session I - Bartlett Hall, Room 220  
  
                    Chairperson: Frank A. Giordano, U.S. Military Academy, West  
  Point, New York
- Estimation for Spatially-Distributed Processes  
                                    Alan S. Willsky, Massachusetts Institute of Technology,  
  Cambridge, Massachusetts
- 0930 - 1000      Break
- 1000 - 1200      Special Session A - Stochastic Methods for Image Analysis -  
                                    Bartlett Hall, Room 220  
  
                    Chairperson: Gerald Andersen, U.S. Army Research Office,  
  Research Triangle Park, North Carolina
- Robust Image Models for Image Restoration and Texture Edge  
                                    Detection  
                                    R. L. Kashyap, Purdue University, West Lafayette, Indiana  
  and Kie-Bum Eom, Syracuse University, Syracuse, New York
- Computational Problems Motivated by Laser Radar Target  
                                    Recognition  
                                    J. Michael Steele, Princeton University, Princeton, NJ

\*\*\*\*\*

\* On Monday, 5 June 1989, conference attendees can register at the Hotel Thayer during the time period 1900-2200.

Tuesday (Continued)

1000 - 1200

Technical Session 1 - Solid Mechanics I - Bartlett Hall,  
Room 420

Chairperson: John D. Vasilakis, Benet Weapons Laboratory,  
Watervliet, New York

On the Stroh Formalism for Anisotropic Elasticity and Its  
Applications to Composites

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

Phase Transitions and Maximally Dissipative Dynamic Solutions  
in the Riemann Problem for Impact

Thomas J. Pence, Michigan State University, East Lansing, MI

Ellipticity and Deformations with Discontinuous Gradients in  
Elastostatics

Phoebus Rosakis, Cornell University, Ithaca, New York

Optimization of Elastic Materials

William W. Hager and Rouben Rostamian, University of  
Florida, Gainesville, Florida

Note on Modeling Stress Relaxation of Elastomer Cylinders

Arthur Johnson, U.S. Army Materials Technology Laboratory,  
Watertown, Massachusetts

Transient Shear Response of a Rigid Block-Support System  
Subjected to Lateral Impact

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

1200 - 1300

Lunch

1300 - 1530

Special Session B - Large Deformation and Computational Issues  
Bartlett Hall, Room 220

Chairpersons: Arthur Johnson, U.S. Army Materials Technology  
Laboratory, Watertown, Massachusetts and John  
Walter, U.S. Army Ballistic Research Laboratory,  
Aberdeen Proving Ground, Maryland

On the Continuum Mechanics of the Motion of a Phase Interface  
Morton E. Gurtin, Carnegie-Mellon University, Pittsburgh, PA

Nonlinear Phenomena in the Inverse Problem

Toshio Mura, Northwestern University, Evanston, Illinois

Dynamics of Viscoelastic Materials with Non-monotone  
Constitutive Relations

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

Tuesday (Continued)

Smart Algorithms for Complex Problems in Fluid Dynamics  
J. Tinsley Oden, University of Texas at Austin, Austin, TX

Effect of Constitutive Modelling on the Dynamic Development of  
Shear Bands in Viscoplastic Materials  
Romesh C. Batra and C. H. Kim, University of Missouri-Rolla,  
Rolla, Missouri

\*\*\*\*\*

1300 - 1530

Technical Session 2 - Control, Systems and Robotics - Bartlett  
Hall, Room 420

Chairperson: Rickey Kolb, U.S. Military Academy, West Point,  
New York

Kinodynamic Planning in Robotics  
Bruce R. Donald, Cornell University, Ithaca, New York

Restructurable Control Inputs I: The Linear Case  
Charles E. Hall, Jr., U.S. Army Missile Command, Redstone  
Arsenal, Alabama

Identifying the Unknown Shift of the Sum of Two Shifted  
Versions of a Linear Feedback Shift Register Sequence  
Harold Fredricksen, Naval Postgraduate School, Monterey, CA  
and Gary W. Krahn, U.S. Military Academy, West Point, NY

Digital Redesign of Pseudo-Continuous-Time Suboptimal  
Regulators for Large-Scale Discrete Systems  
L. S. Shieh, University of Houston, Houston, Texas and  
Norman Coleman, U.S. Army Armament R&D Center, Picatinny  
Arsenal, New Jersey

New Methodologies in Renewal Theory  
B. D. Sivazlian, University of Florida, Gainesville, Florida

Stepwise Closed Form Techniques for Computer Simulation of  
Guided Projectiles  
M. J. Amoruso and R. Campbell, U.S. Armament R&D Center,  
Picatinny Arsenal, NJ, and Herbert Cohen, U.S. Army Materiel  
Systems Analysis Activity, Aberdeen Proving Ground, Maryland

Variations in Buchberger's Algorithm  
Lee Taylor, Cornell University, Ithaca, New York

Tuesday (Continued)

1530 - 1600 Break

1600 - 1700 General Session II - Bartlett Hall, Room 220

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

Microstructure and Properties of Crystals Undergoing Phase  
Transformation

Richard D. James, University of Minnesota, Minneapolis, MN

\*-----\*

Wednesday, June 7, 1989

0800 - 1600 Registration

0830 - 0930 General Session III - Bartlett Hall, Room 220

Chairperson: Julian J. Wu, U.S. Army Research Office, Research  
Triangle Park, North Carolina

Modelling Microstructure by Energy Minimization

Robert V. Kohn, New York University, New York, NY

0930 - 1000 Break

1000 - 1200 Special Session C1 - Mathematics in Computer Science -  
Bartlett Hall, Room 220

Chairperson: Anil Nerode, Cornell University, Ithaca, New York

Nontraditional Logic in Computer Science

Anil Nerode, Cornell University, Ithaca, New York

IZF and Kripke Models and Program Extraction

James Lipton, Cornell University, Ithaca, New York

Concurrent Specifications and Their Gurevich-Harrington  
Strategies

A. Yakhnis, Cornell University, Ithaca, New York

Extraction of Concurrent Programs from Gurevich-Harrington  
Strategies

V. Yakhnis, Cornell University, Ithaca, New York

\*\*\*\*\*

Wednesday (Continued)

1000 - 1200      Technical Session 3 - Computational Fluid Mechanics - Bartlett  
Hall, Room 420

Chairperson: David C. Arney, U.S. Military Academy, West  
Point, New York

Block Multigrid Implicit Solution of the Euler Equations of  
Compressible Fluid Flow

David A. Caughey and Yoram Yadlin, Cornell University,  
Ithaca, New York

Effect of Particle Velocity Fluctuations on the Inertia  
Coupling in Two-Phase Flow

Donald A. Drew, Rensselaer Polytechnic Institute, Troy, NY

Elastic Deformation and Slug Flow as Applications of Front  
Tracking

James Glimm, State University of New York at Stony Brook,  
Stony Brook, New York, X. Garaizar and W. Guo, Courant  
Institute, New York University, New York, NY

Unstable Interfaces and Anomalous Waves in Compressible Fluids

John W. Grove, State University of New York at Stony Brook,  
Stony Brook, New York

Characteristics and Stability Implications of a Streamwise  
Vortex in Bounded Shear Flow

Joseph D. Myers, U.S. Military Academy, West Point, New York  
and Frederick H. Abernathy, Harvard University, Cambridge,  
Massachusetts

Weakly Nonlinear Expansions for Viscous Rotating Pipe Flow

Alex Mahalov and Sidney Leibovich, Cornell University,  
Ithaca, New York

1200 - 1330      Lunch

1330 - 1430      Special Session C2 - Mathematics in Computer Science - Bartlett  
Hall, Room 220

Chairperson: Anil Nerode, Cornell University, Ithaca, New York

Intuitionistic Modal Logics

D. Wijesekera, Cornell University, Ithaca, New York

Generalized Rewriting in Type Theory

David A. Basin, Cornell University, Ithaca, New York

\*\*\*\*\*

Wednesday (Continued)

- 1330 - 1430      **Technical Session 4 - Adaptive Methods and Other Analytical Techniques - Bartlett Hall, Room 420**
- Chairperson: Louis J. Piscitelle, U.S. Army Natick Laboratories, Natick, Massachusetts
- Numerical Experiments in Adaptive Mesh Methods  
David C. Arney, U.S. Military Academy, West Point, New York,  
Rupak Biswas and Joseph E. Flaherty, Rensselaer Polytechnic Institute, Troy, New York
- Asymptotic Analysis of the np-Junction  
Donald A. Drew, Rensselaer Polytechnic Institute, Troy, NY
- Entropy, Directed Orthogonality, and Magic Distances  
Lee K. Jones and Victor Trutzer, University of Lowell,  
Lowell, Massachusetts
- 1430 - 1445      **Break**
- 1445 - 1530      **U.S. Military Academy Session - Bartlett Hall, Room 220**  
**Mathematics at West Point**
- 1530 - 1600      **Break**
- 1600 - 1700      **General Session IV - Bartlett Hall, Room 220**
- Chairperson: Paul Broome, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD
- High Performance Computing  
S. Lennart Johnsson, Yale University, New Haven, Connecticut

---

A banquet is being planned for Wednesday evening at the West Point Officers' Club.

\*-----\*

Thursday, June 8, 1989

0800 - 1600 Registration - Bartlett Hall, Room 220

0830 - 0930 General Session V - Bartlett Hall, Room 220

Chairperson: Norman Coleman, U.S. Army Armament R&D Center,  
Picatinny Arsenal, New Jersey

Theory and Application of Piecewise-Deterministic Processes  
Mark H.A. Davis, Imperial College of Science and Technology,  
London, England

0930 - 1000 Break

1000 - 1200 Special Session D1 - Computational Methods for Multibody  
Dynamics - Bartlett Hall, Room 220

Chairperson: Roger Wehage, U.S. Army Tank-Automotive Command,  
Warren, Michigan

Implementation of Efficient Simulation Codes for  
Tree-Structured Mechanical Systems, such as Robots  
Martin Otter, German Aerospace Research Establishment  
(DFVLR), West-Germany

Symbolic Factors of Linear System Coefficient Matrices for  
Tree-Structured Systems and Their Efficient Solution  
Roger A. Wehage, U.S. Army Tank-Automotive Command, Warren,  
Michigan

Optimal Cut-Set Selection for Tree-Structured Systems  
Containing Closed Loops  
James L. Overholt and Roger A. Wehage, U.S. Army  
Tank-Automotive Command, Warren, Michigan

\*\*\*\*\*

1000 - 1200 Technical Session 5 - Computer Science - Bartlett Hall,  
Room 420

Chairperson: Terry Cronin, U.S. Army Signal Warfare  
Laboratory, Warrenton, Virginia

A Penalty Function Method for the Standard Equality Constrained  
Minimization Problem  
Thomas F. Coleman and Christian G. Hempel, Cornell  
University, Ithaca, New York



Thursday (Continued)

Iterative Methods for Cyclically Reduced Non-Self-Adjoint  
Linear Systems

Howard Elman, University of Maryland, College Park, MD and  
Gene H. Golub, Stanford University, Stanford, California

A Computer Simulation of the Freely-Associating Neocortex

Mark Johnson, Raymond Scanl and Michael Cipollo, Benet  
Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

On Calculations of the Topological Complexity of Algorithms

Anatoly S. Libgober, University of Illinois at Chicago,  
Chicago, Illinois

Deconvolution of Overdetermined Systems of Convolution  
Equations

B. A. Taylor, University of Michigan, Ann Arbor, Michigan  
and Carlos A. Berenstein, University of Maryland, College  
Park, Maryland

1200 - 1330

Lunch

1330 - 1530

Special Session D2 - Computational Methods for Multibody  
Dynamics - Bartlett Hall, Room 220

Chairperson: Gary Anderson, U.S. Army Research Office,  
Research Triangle Park, North Carolina

Application of Recursive Projection Methods to Dynamics of  
Deformable Multibody Systems

Roger A. Wehage, U.S. Army Tank-Automotive Command, Warren,  
Michigan and A. A. Shabana, University of Illinois at  
Chicago, Chicago, Illinois

Automated Symbolic Formulation of Efficient Vehicle Simulation  
Codes

Michael Sayers, University of Michigan, Ann Arbor, Michigan

A Comparative Study of Different Dynamical Formulations in  
Multibody Dynamics

S. Hanagud, Georgia Institute of Technology, Atlanta, GA

A Generalized Harmonic Balance Method for Forced Nonlinear  
Oscillations

B. Noble, Brunell University, Uxbridge, United Kingdom,  
M. A. Hussain, General Electric Corporate R&D Center,  
Schenectady, New York, and Julian J. Wu, U.S. Army Research  
Office, Research Triangle Park, North Carolina

\*\*\*\*\*

Thursday (Continued)

1330 - 1530      Technical Session 6 - Solid Mechanics II - Bartlett Hall,  
Room 420

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

Effect of Thermal Softening on the Response of Shearing Motions  
Athanasios E. Tzavaras, University of Wisconsin-Madison,  
Madison, Wisconsin

Particle Interactions in Ductile Metals  
Dennis M. Tracey and Paul J. Perrone, U.S. Army Materials  
Technology Laboratory, Watertown, Massachusetts

A Mode I Crack Solution for a Mooney-Rivlin Material  
C. J. Quigley, U.S. Army Materials Technology Laboratory,  
Watertown, Massachusetts and D. M. Parks, Massachusetts  
Institute of Technology, Cambridge, Massachusetts

Elastic-Plastic Analysis of a Thick-Walled Composite Tube  
Subjected to Internal Pressure  
Peter C.T. Chen, Benet Weapons Laboratory, Watervliet  
Arsenal, Watervliet, New York

Distortion of Frequency Spectra of Solids Due to Discretization  
Joseph M. Santiago, U.S. Army Ballistic Research Laboratory,  
Aberdeen Proving Ground, Maryland

Ultrafast Thermodynamic Processes  
Richard A. Weiss, U.S. Army Engineer Waterways Experiment  
Station, Vicksburg, Mississippi

The Internal Phase Structure of Atoms  
Richard A. Weiss, U.S. Army Engineer Waterways Experiment  
Station, Vicksburg, Mississippi

1530 - 1600      Break

1600 - 1700      General Session VI - Bartlett Hall, Room 220

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

Wavelet Transforms  
Jean-Michel Morel\*, Université of Paris-Dauphine, Paris,  
France

\*Replaced by A. Cohen

\*-----\*

Friday, June 9, 1989

0800 - 1100 Registration - Bartlett Hall, Room 220

0830 - 1030 Technical Session 7 - Approximation Theory and Applications -  
Bartlett Hall, Room 220

Chairperson: Royce Soanes, Benet Weapons Laboratory,  
Watervliet, New York

Numerical Solution of One-Dimensional Aperture Integral  
Equation

M. A. Hussain and Wen-Tai Lin, General Electric Corporate  
R&D Center, Schenectady, New York and Ben Noble, Brunell  
University, Uxbridge, United Kingdom

Computation of Leading Eigenspaces for Generalized Eigenvalue  
Problems

Abraham Kribus, Cornell University, Ithaca, New York

Applications of Fibonacci Sequences and Tiling

Joseph Arkin, David C. Arney, Lee S. Dewald and Charles  
Kennedy, U.S. Military Academy, West Point, New York

Approximation and Interpolation Formulas for Real-Time  
Applications

C. K. Chui, Texas A&M University, College Station, Texas

An Enhanced Knot Selection Algorithm for Least Squares  
Approximation Using Thin Plate Splines

John R. McMahon, U.S. Military Academy, West Point, New York  
and Richard Franke, Naval Postgraduate School, Monterey, CA

Selection of Step Sizes in Numerical Modeling and Simulation  
Rao Yalamanchili, U.S. Army Armament R&D Center, Picatinny  
Arsenal, New Jersey

\*\*\*\*\*

0830 - 1030 Technical Session 8 - Mathematical Physics, Statistics, Etc. -  
Bartlett Hall, Room 420

Chairperson: John Robertson, U.S. Military Academy, West  
Point, New York

An Extension of the Cramer-vonMises Goodness of Fit Statistic  
for Multivariate Gaussianity

Kevin M. Beam, U.S. Military Academy, West Point, New York  
and Albert S. Paulson, Rensselaer Polytechnic Institute,  
Troy, New York

The Number of Solutions of Certain Equations Occurring in  
Statistics

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

Friday (Continued)

Numerical Modelling of Large Fires

K. C. Heaton, Defense Research Establishment Valcartier,  
Courcellette, Quebec

Application of Composite Regression Analysis to Satellite Data  
Processing

Roger H. Multer, U.S. Army Engineer Waterways Experiment  
Station, Vicksburg, Mississippi

Optimized Annulus-based Point-in-Region Inclusion Testing for  
d Dimensions

Terry M. Cronin, U.S. Army Signal Warfare Laboratory, Vint  
Hill Farms Station, Warrenton, Virginia

1030 - 1100

Break

1100 - 1200

General Session VII - Bartlett Hall, Room 220

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

What's New in Multivariate Splines?

Carl de Boor, University of Wisconsin-Madison, Madison, WI

1200 - 1215

Adjournment

## USMA'S MATHEMATICS PROGRAM FOR 1990 and BEYOND

David C. Arney  
Lee S. Dewald, Sr.  
John R. Edwards

Department of Mathematics  
United States Military Academy  
West Point, New York 10996-1786

**ABSTRACT:** The following paper stemmed from a special presentation by the authors at the Seventh Annual Army Conference on Applied Mathematics and Computing. The presentation described the background of, motivation for, and broad content of the new core mathematics program for all cadets starting in the Fall of 1990 -- discrete dynamical systems, calculus, and probability and statistics. The United States Military Academy (USMA) is the single largest source of officers to the Army with mathematics, science, and engineering backgrounds. It is necessary to inform Army mathematicians and scientists of these curricular developments and the department's research program.

### 1. Introduction.

Mathematics is the language of science. Continuous mathematics, especially calculus, has been the cornerstone of undergraduate education in the sciences. First models of a behavior are often continuous. We need to continue to teach continuous mathematics since we all, students especially, gain great insights from the closed-form solutions to continuous models that calculus affords, even when these models oversimplify reality. These first models assume the world is linear, continuous, and deterministic. More often, it is nonlinear, ultimately discrete, and usually stochastic. Discrete mathematics is not only the language of the discrete world but also is the language of the computer. Probabilistic mathematics is the language of uncertainty. The study of all these fundamental areas of mathematics would provide a much better basis to view and model our world.

The order of presentation, discrete dynamical systems, calculus and probability and statistics, is important. Discrete mathematics progressing from algebra to matrix algebra to discrete dynamical systems is a better transition from high school mathematics and can be used to preview the more difficult concepts (for example, the limit) that underly continuous mathematics. Finally, probability is based on both continuous and discrete mathematics. With

recent advances in textbooks and software, we at USMA are in an unprecedented position with each cadet possessing a computer (portable in 1990) to present an integrated four-course curriculum treating the fundamental ideas of discrete, continuous and stochastic mathematics.

We feel that an integrated curriculum will permit us to develop the following attitudes in cadets that will carry over into their careers as officers:

- Mathematics is deductive in character. A few principles must be internalized but most notions are derived.
- Mathematics is a medium of communications in which ideas are formalized and through which theories are synthesized.
- Curiosity and experimental disposition are essential characteristics of mathematics education. Through observation one seeks universal truths and establishes them by proof.
- Learning mathematics is an individual responsibility. Textbooks, instructors, and members of study groups only facilitate the process.
- Mathematics is useful.

In the next three sections we describe in more detail the plans for each of the three courses in the four semesters of mathematics - discrete dynamical systems, the calculus, and probability and statistics. In designing our curriculum, we have taken into account several national reports on the current status of mathematics, calculus in particular, and the changes needed to improve the status of our educational program [1,2].

In the final section we briefly describe the research program to which the Department of Mathematics ascribes for tenured and non-tenured faculty as well as the cadets. We see this program as the ultimate capstone of the new program. Problem solving is aggressively encouraged by providing ample opportunities to solve meaningful practical problems requiring the integration of fundamental ideas encompassing one or more lesson blocks from one or more core mathematics courses.

## 2. Discrete Dynamical Systems.

Under USMA's proposed curriculum, the first core mathematics course is MA 103: Discrete Dynamical Systems with Matrix Algebra. This is a 3 credit hour course. It provides introductions to elementary matrix operations and matrix methods to solve systems of linear equations. Several applications of these subjects are also studied. Most of the course is devoted to topics and problems in the mathematics of discrete dynamical systems. Introductory material on modeling problems using difference equations motivates the study of solution techniques for these equations and the eventual study of calculus and differential equations. Concepts and techniques are discussed for first-order linear and nonlinear equations and higher order linear equations, and systems of equations. Computer software is used to demonstrate and solve problems in both the matrix algebra and the discrete dynamical systems sections of the course.

While the placement and scope of our course in the curriculum may be unique, we feel that this will be the ultimate role of a discrete mathematics course. As stated by Maurer in [3], "there may yet be a move toward more discrete math in the first year." We intend to lead the way in designing, testing, and teaching a discrete course for the first semester of college mathematics. In order to start this course in the 1990-1991 academic year, we will have to piece together textual material and write some ourselves. [4,5]

There are several reasons to begin our curriculum with such a course. It provides a logical transition from high school to college mathematics and provides an intuitive motivation for the limiting concepts of the calculus. Discrete mathematics also is the language of the computer, and difference equations provide an intuitive introduction to recursion. Discrete models, many in the form of difference equations, are popular models of dynamic behavior and are worthy of increased study.

Some of the goals for the students in the course are: ability to formulate discrete mathematical models; ability to solve algebraic and discrete models; motivation for the calculus; and internalization of a few principles of mathematics. We hope to use this course to develop the following attitudes early in the curriculum: curiosity and experimental disposition; a desire to structure and communicate quantitative ideas; an appreciation for mathematics as a useful tool to solve real problems; and an appreciation for the power of deductive reasoning.

The first block of 12-15 lessons covers matrix algebra. The major topics in this block are the basic concept of linearity, matrix operations, determinants, inverses, Markov processes, and linear programming. The second block on difference equations covers first order theory and applications, second order theory and applications, first order systems, Markov chains, and nonlinear difference equations. The final block establishes a foundation for calculus by introducing sequences and difference quotients.

One unifying feature of all the blocks in this course is the use of the computer. The computer will be used to demonstrate concepts in class as well as a tool to solve problems. USMA and the Department of Mathematics, in particular, have had an aggressive program of computer assisted instruction for several years [6]. This course intends to establish the foundation for computer use by cadets in solving problems of a mathematical or scientific nature and to establish the computer as a tool in mathematical experimentation.

### 3. Lean and Lively Calculus.

#### 3.1. Background.

From the 1950's through 1974 the core mathematics program at USMA was a strong and stable program in undergraduate mathematics both in content and credit. In four semesters each cadet received the equivalent of six courses in mathematics -- single-variable (integral and differential) calculus, multivariable calculus, linear algebra, ordinary differential equations, and elementary probability and statistics. Cadets attended class six days a week for 17 weeks a semester at 80 minutes per day. All of the textual materials were written at USMA either directly or under the supervision of the Chairman of the Department of Mathematics, Charles P. Nicholas.

Since the mid-70's there has been a constant and steady erosion of the depth and breadth of coverage in the core mathematics at USMA. Part of this was the result of offering academic majors in non-science and non-engineering fields. Regardless of the rationale for the reduced emphasis on mathematics, the effects were the same. By the end of the 1980's the core mathematics program was reduced by 30%. Unlike many other schools, USMA still has maintained an emphasis on mathematics by keeping four mathematics courses in its core curriculum. [7]

The resulting programs never reached a steady-state. Topics would appear, disappear, and reappear from semester to semester. Conceptual development was replaced entirely by the learning of algorithmic skills. There was no real plan.



In 1984 the Chairman of the Department of Mathematics received a report from the tenured faculty which recognized a need to change the core mathematics program at USMA and to take advantage of the technological advances in computers and symbolic manipulation. However, textbooks were not available (essentially are still not) and there was no authority or time to write these textual materials at USMA. Therefore, little was changed.

However, since the beginning of the National Reform Movement in Calculus in 1987 there has been increasing interest in mathematics education in many sectors [1], [2] -- publishers, authors, professors, computer scientists, and students. It is against this backdrop that the West Point version of the "Lean and Lively" Calculus is being developed.

### 3.2. Course Description - Calculus I and II.

These are the second and third courses of the mathematics core curriculum and are each 4.5 credit hours. These standard courses provide study of mathematics as an intellectual discipline and as a foundation for continued study of mathematics and for the subsequent study of physical sciences, social sciences, and engineering. Beginning with functions and the sequential development of the limit, the calculus is covered through the development and evaluation of multiple integrals. No vector calculus is included. Ordinary differential equations are integrated into the course as soon as higher order derivatives are covered. Computers and symbolic manipulation are integrated throughout the program to foster both discovery and intellectual curiosity and to enhance problem-solving.

### 3.3. Objectives of the Calculus Sequence.

There are four basic objectives to the study of calculus which support the overall objectives of the mathematics curriculum at USMA:

a. Students learn the three basic limit ideas of calculus: The limit of a convergent sequence is related to the concept of a continuous function; the limit of a quotient is related to a derivative; the limit of a sum is related to the definite integral.

b. Students be able to prove some of the basic results in the calculus.

c. Students be able to formulate ideas in the mathematics of the calculus.

d. Students be able to solve problems using calculus by formulating the models and applying the appropriate techniques and algorithms.

### 3.4. A "Lean" Calculus.

Two problems contribute to a need for a new "lean" calculus. Both problems lie in the size of the calculus text. Calculus books are too large! Even though there are only three principle ideas there are typically 16-19 chapters of material. What once were applications or examples have been elevated to the status of independent topics. Thus problem one is the growth of "important" topics.

The second problem is the reluctance to remove outdated or irrelevant material from the textbooks. Many topics which are purely algorithmic by nature and easily implemented with a computer, are still being drilled and memorized in calculus classrooms.

There are two approaches to be taken in deriving this new lean calculus -- the butcher's approach or that of the sculptor.

The butcher's approach is relatively easy to implement and requires no new textbooks. Essentially the topics of the textbook are divided between baseline and enhancement. Every student of calculus does the baseline and some percentage of the enhancement depending on background, instructor preference, etc. This idea has essentially been implemented by Scott Foresman Publishers for the Calculus and Analytic Geometry by Al Shenk. On the surface this approach sounds like little improvement. Some agreement across colleges over what is baseline and what is enhancement would be required.

There are however Computer Algebra Systems (CAS) that can support a butcher's approach independent of the choice of textbooks. CAS is the new technology that would make this approach a major improvement over the existing programs. CAS performs symbolic manipulation to include symbolic integration and differentiation in either a hand-held calculator or computer software.

CAS is not a crutch to do for students what they should be able to do for themselves. CAS is a force multiplier that makes for a more efficient use of study time for the student and allows professors to change course priorities. Much of the time that is spent on drill and memorization is eliminated. Topics and problems that were not accessible before can now be explored using CAS and other computer support.

The sculptor's approach to a "lean" calculus will be a work of art and is probably still a couple of years in the making. Central to this approach are new textual materials which incorporate several major differences from the same "ole brewski."

Emphasis must shift to conceptual understanding and problem solving and away from memorization and drilling on formulas. More writing requirements and interpretation of results should be emphasized instead of the production of results. Differential equations should be integrated throughout the calculus textbook instead of being treated in isolation. Finally, CAS and other computer capabilities should be integrated into the text to capitalize even more on the new technology.

### 3.5. A "Lively" Calculus.

Many ideas for implementing a "lean" calculus exist. What seems to be more difficult is the question of how to "liven-up" the calculus program. We look to relevance and experimentation. We intend to emphasize the relevance of calculus to the solving of problems -- motivational and carry-over. We also intend to emphasize experimentation and the discovery of new techniques to solve interesting but previously unsolvable problems.

Several special problems have already been developed for use in the calculus program that emphasize integration and modeling and solution of differential equations. New carry-over problems are being developed in probability and statistics as well as optimization and economics.

Computers, CAS, and specialty software will play two major roles in the lively calculus. The use of computational software opens up a wider variety of problems that are more realistic and interesting for students to solve. The student is also much more inclined to explore the nature of functions and discover their properties with the use of computers.

Cadets at USMA currently own The Calculus Toolkit, the Midshipman's Plotting Package, and DERIVE, CAS for IBM compatible PC's.

#### 4. Probability and Statistics

The USMA probability and statistics course is the capstone of the mathematics required of all cadets and is 3 credit hours. In negotiating this course we expect students to show sophistication and technical maturity. When students come to us from their precollege experience their learning is essentially skill based; and the pedagogy and material at the beginning of our core curriculum reflect this as a point of departure. Our curriculum is designed to gradually wean students from this learning approach culminating in this final course, probability and statistics, in which the learning is wholly cognitive; very little time and class reward is devoted to skill learning. The issues confronted by the student are not closed form and require him to interpret his mathematical manipulations.

Because the course is conceptual in character, we emphasize the unified structure of the study of uncertainty. Computation is pushed off to software (currently MINITAB). Learning is socratic in character; students are directed in such a way that they "discover" the two distributions that form the center of the course; one discrete and one continuous. It is our goal that students internalize the idea that once an issue can be modeled by a random variable and its distribution, one has a complete gauge of the inherent uncertainties. While only two distributions are formally developed in class, students are expected to lift the essence of a distribution to other functions; to generalize the concepts and apply them to problems other than the two 'learning examples.'

Our transition to statistics appeals to the intuitive notion that the character of a population can be forecast from a suitable subset of the population. The notion of "sample space", first discussed in probability, is replaced with the space of all subsets of a fixed size ("the sample space of samples of size  $n$ "). The student observes that measures taken on these samples meet the definition of random variables on this new sample space. At this point the structure of the course quickly narrows his consideration to two such measures: mean for central tendency and variance for spread. This approach causes students to take the perspective that the most important need for interpreting a sample outcome is to characterize the distribution of these measures.

This perspective leads to the study of the Central Limit Theorem. The result of the theorem is motivated experimentally; first by mechanical means (e.g. drawing numbered slips out of a container) and then through computer simulation. It is beyond the scope of the course to provide an analytic proof of the theorem, but students have a strong intuitive appreciation of the result.

Establishing the distribution of the variance is less elegant. It is our assessment that the knowledge required to logically develop the relationship between the distribution of the variance and the corresponding Chi-squared distribution would demand an effort all out of proportion to the gain in a one semester course. Thus the transformation of variable is simply given as an analogue of the Z-transform with which they have become familiar.

Central to the course is a case study that, in an actual example, reviews and reinforces all the ideas discussed. In addition to accomplishing the technical analysis of the issues, students are required to interpret their "numbers." Furthermore, the minimum course standard requires that their case study be in a professionally acceptable format. This includes embedding files from their statistical software into their word processing files and integrating mathematical exhibits with text. To reinforce the Case Study's importance, the grade for the effort is one third of their final exam.

A remark on the choice of case study is in order. We have found the learning is far greater if the topic is taken from actual student experience, something that effects their lives. For example, we selected as a population the grades of a preceding class and asked them to draw conclusions about the types of career success these students enjoyed and how that was correlated to various academy successes: academic grades, military leadership grades, and physical fitness grades. This case study was far more successful as a learning tool than an earlier one that investigated a very important weapons systems (the Bradely fighting vehicle) but a subject that was only vicarious to sophomore level cadet. We concluded that having as an object of study something that the students actually experience and see as real imposes and sense of urgency in their study; students want to understand those issues that influence their lives now.

Completion of the course poises students to address problems:

- That require interval estimates of parameters.
- That establish rational decision values for experimental variables.
- That require simple design of an experiment.

The course does not leave them as skilled statisticians. However, it does blend together the key elements of all preceding mathematics courses. It prepares them to use quantitative methods to solve significant and unstructured

problems that require sophisticated interpretation. And it prepares them to communicate their findings in a clear and professional manner.

#### 5. Research Program.

This is a brief description of the faculty and student research program of the Department of Mathematics. With regard to faculty basic and applied research, the departmental program supports the philosophy of the USMA Superintendent, LTG Palmer. In a position paper, he stated:

"...The faculty at USMA constitutes a valuable resource to the Army [and the nation at large] in that nowhere else is there as large a concentration of highly educated personnel as at West Point. The potential of this resource to solve Army problems should be fully exploited ...

"These officers will take this valuable experience back to the Army with them, and many will put it to good use in positions in the acquisition system such as project manager. Thus USMA will provide the Army with officers that understand the research process and who will not be technically at the mercy of government contractors ..." [8]

There is no question that the Department has committed its faculty to the furtherance of knowledge in the areas of prime concern to the Army and nation. Over 20 of the 60 officers in the Department were directly involved in significant research or consulting projects during the last academic year. These projects are described in [9] and include applications in many areas of mathematics and science, i.e., numerical computing, fluid dynamics, number theory, underwater and atmospheric acoustics, probability distributions, statistical analysis, time series, computer aided design, computer aided instruction, air defense methodology, signal processing, financial modeling, and combat modeling. In addition, 18 officers spent time during the summer of 1989 at an Army laboratory or government agency performing research or consulting. Many other instructors were involved in smaller part-time efforts. The tenured faculty particularly were involved in this effort through consulting with Army laboratories, schools, and agencies, attending conferences, presenting results, and publishing in technical journals.

Under the direction of MAJ John S. Robertson, the Department of Mathematics research program was particularly successful in 1989. Four members of the Department served in Dean's Research positions and were funded by the USMA Science Research Laboratory with money from the Army Research Office. Several other researchers received this same funding. For the first time ever, substantial funding was obtained from external agencies for travel, supplies, computer hardware, software, and library services. With this funding and direction, the future looks bright for the Department's research program.

One enhancement for the program may come from the Department's use of Foundation Schools for instructor education. Starting this year, all the non-tenured faculty (85% of the Department's strength) will receive their masters-level education at one of three schools, Georgia Tech, Rensselaer Polytechnic Institute, or the Naval Postgraduate School with degrees in either applied mathematics or operations research. This program will enable the tenured faculty to interface with the officers at an earlier stage for better control of professional development with emphasis on finding research opportunities that can continue while the officer is assigned to the Department.

The student-research program is focused in two areas: Volunteer Summer Training (VST) and a 3-credit Research Seminar (MA 491). Over the last two years, over 25 cadets have participated in a 4-6 week VST research program at many agencies including TRADOC Analysis Center-Monterey, Ballistic Research Laboratory, Natick Laboratory, Concepts Analysis Agency, and Los Alamos National Laboratory. Several cadets have completed the MA 491 course through their undergraduate research in topics such as numerical computing, chaos and fractals, combat modeling, and financial modeling.

As we head into the 1990's, research has taken an important place in the Department of Mathematics. Student and faculty involvement in research activities has been beneficial and rewarding and most likely will continue to grow in the future.

## REFERENCES

- [1] Steen, Lynn Arthur, "Mathematics for a New Century," Notices of the American Mathematical Society, Vol. 36, No. 2, Feb 1989, pp. 133-138.
- [2] "Everybody Counts: A Report to the Nation on the Future of Mathematics Education." Washington: National Academy Press, 1989.
- [3] Maurer, Stephen B., "Is Discrete Mathematics Dead?" UME Trends, Vol. 1, No. 2, May 1989.
- [4] Sandefur, James T., "Using Discrete Applications to Inspire Calculus", UMAP Journal, Vol. 9, No. 3, 1988, pp. 191-194.
- [5] Sandefur, James T., "Discrete Dynamical Systems - An Alternative to Calculus?" SIAM News, Vol. 20, No. 3, May 1987, p. 3.
- [6] Arney, David C. and Robinson, Bruce T., "Weaving a Computer Thread in Mathematics Using Microcomputers," Coed Journal, Vol. 8, No. 3, 1988, pp. 38-42.
- [7] Sandefur, James T., "Core Curriculum: Where Does Math Fit In?", SIAM News, 1988.
- [8] Palmer, Dave R. "Basic and Applied Research at USMA", 1988 Report on the Conference of Academy Superintendents, 1988.
- [9] "In the Lead", Department Technical Report, Number 89-1, Department of Mathematics, USMA, October 1988.



# Modeling and Estimation for Multiresolution Stochastic Processes

A.S. Willsky<sup>1</sup>

## 1 Multiscale Representations and Homogeneous Trees

The recently-introduced theory of multiscale representations and wavelet transforms [4] provides a sequence of approximations of signals at finer and finer scales. In 1-D a signal  $f(x)$  is represented at the  $m$ th scale by a sequence  $f(m, n)$  which provides the amplitudes of time-scaled pulses located at the points  $n2^{-m}$ . The progression from one scale to the next thus introduces twice as many points and indeed provides a tree structure with the pair  $(2^{-m}, n)$  at one scale associated with  $(2^{-(m+1)}, 2n)$  and  $(2^{-(m+1)}, 2n + 1)$  at the next. This provides the motivation for the development of a system and stochastic process theory when the index set is taken to be a homogeneous dyadic tree. In this paper we outline some of the basic ideas behind our work.

Let  $\mathcal{T}$  denote the index set of the tree and we use the single symbol  $t$  for nodes on the tree. The scale associated with  $t$  is denoted by  $m(t)$ , and we write  $s \preceq t$  ( $s \prec t$ ) if  $m(s) \leq m(t)$  ( $m(s) < m(t)$ ). We also let  $d(s, t)$  denote the distance between  $s$  and  $t$ , and  $s \wedge t$  the common "parent" node of  $s$  and  $t$  (e.g.  $(2^{-m}, n)$  is the parent of  $(2^{-(m+1)}, 2n)$  and  $(2^{-(m+1)}, 2n + 1)$ ). In analogy with the shift operator  $z^{-1}$  used as the basis for describing discrete-time dynamics we also define several shift operators on the tree:  $\cdot 0$ , the identity operator (no move);  $\gamma^{-1}$ , the fine-to-coarse shift (e.g. from  $(2^{-(m+1)}, 2n$  or  $2n + 1)$  to  $(2^{-m}, n)$ );  $\alpha$ , the left coarse-to-fine shift ( $(2^{-m}, n)$  to  $(2^{-(m+1)}, 2n)$ );  $\beta$ , the right coarse-to-fine shift ( $(2^{-m}, n)$  to  $(2^{-(m+1)}, 2n + 1)$ ); and  $\delta$ , the exchange operator ( $(2^{-(m+1)}, 2n) \longleftrightarrow (2^{-(m+1)}, 2n + 1)$ ). Note that  $\cdot 0$  and  $\delta$  are

---

<sup>1</sup>This research was supported in part by the Army Research Office under grant DAAL03-86-K-0171 (Center for Intelligent Control Systems), AFOSR grant AFOSR-88-0032 and the NSF under grant ECS-8700903.

**isometries** in that they are one-to-one, onto maps of  $\mathcal{T}$  that preserve distances. Also we have the relations

$$\delta^2 = \gamma^{-1}\alpha = \gamma^{-1}\beta = 0, \quad \gamma^{-1}\delta = \gamma^{-1}, \quad \delta\beta = \alpha \quad (1.1)$$

It is possible to code all points on the tree via shifts from an arbitrary origin node, i.e. as  $wt_0$ ,  $w \in \mathcal{L}$ , where

$$\mathcal{L} = (\gamma^{-1})^* \cup \{\alpha, \beta\}^* \delta (\gamma^{-1})^* \cup \{\alpha, \beta\}^* \quad (1.2)$$

The **length** of a word  $w$  is denoted  $|w|$  and equals  $d(wt, t)$  (e.g.  $|\gamma^{-1}| = 1$ ,  $|\delta| = 2$ ). Also, since we will be interested in coarse-to-fine dynamic models, we define some notation for **causal** moves:

$$w \preceq 0 \text{ (} w \prec 0 \text{) if } wt \preceq t \text{ (} wt \prec t \text{)} \quad (1.3)$$

## 2 Modeling of Isotropic Processes on Trees

A zero-mean process  $Y_t$ ,  $t \in \mathcal{T}$  is **isotropic** if

$$E[Y_t Y_s] = r_{d(t,s)} \quad (2.1)$$

i.e. if its second-order statistics are invariant under any isometry of  $\mathcal{T}$ . These processes have been the subject of some study, and a Bochner-like spectral theorem has been developed [1,2]. However, many questions remain including an explicit criterion for a sequence  $r_n$  to be the covariance of such a process and the representations of isotropic processes as outputs of systems driven by white noise. Note first that the sequence  $\{Y_{\gamma^{-n}t}\}$  is an ordinary time series so that  $r_n$  must be positive semidefinite; however, the constraints of isotropy require even more. To uncover this structure we have developed in [2] a complete characterization of the class of isotropic autoregressive (AR) models where an AR model of order  $p$  has the form

$$Y_t = \sum_{\substack{w \prec 0 \\ |w| \leq p}} a_w Y_{wt} + \sigma W_t \quad (2.2)$$

where  $W_t$  is a white noise with unit variance. Note that this model is "causal"—i.e. it has a coarse-to-fine direction of propagation—since  $w \preceq 0$ . Also, a first thought might be to examine models with strict past dependence, i.e.  $Y_t$  a function of  $W_{\gamma^{-n}t}$ ; however as shown in [2], the constraints of isotropy allow us to show that only AR(1) has such dependence. Thus we have that AR( $p$ ) involves a full set of  $2^{p-1}$   $a_w$ 's and one  $\sigma$  so that the number of parameters doubles as  $p$  increases by one. In addition as shown in [2], isotropy places numerous polynomial constraints on these parameters. As we develop in [2] a better representation is provided by the generalization of lattice structures which involves only one new parameter as  $p$  increases by one.

Let  $\mathcal{H}\{\dots\}$  denote the Gaussian linear space spanned by the variables in braces and define the ( $n$ th order) past of the node  $t$ :

$$\mathcal{Y}_{t,n} \triangleq \mathcal{H}\{Y_{wt} : w \preceq 0, |w| \leq n\} \quad (2.3)$$

As for time series, the development of models of increasing order involves recursions for the forward and backward prediction errors. Specifically, define the **backward residual space**:

$$\mathcal{Y}_{t,n} = \mathcal{Y}_{t,n-1} \oplus \mathcal{F}_{t,n} \quad (2.4)$$

where  $\mathcal{F}_{t,n}$  is spanned by the backward prediction errors

$$F_{t,n}(w) \triangleq Y_{wt} - E(Y_{wt} | \mathcal{Y}_{t,n-1}) \quad (2.5)$$

where  $w \preceq 0$ ,  $|w| = n$ . These variables are collected into a  $2^{\lfloor \frac{n}{2} \rfloor}$ -dimensional vector (see [2] for the order),  $F_{t,n}$ . For  $|w| < n$  and  $w \succ 0$  (i.e.  $m(wt) = m(t)$ ) define the **forward prediction errors**:

$$E_{t,n}(w) \triangleq Y_{wt} - E(Y_{wt} | \mathcal{Y}_{\gamma^{-1}t, n-1}) \quad (2.6)$$

and let  $\mathcal{E}_{t,n}$  denote the span of these residuals and  $E_{t,n}$  the  $2^{\lfloor \frac{n-1}{2} \rfloor}$ -dimensional vector of these variables (see [2]).

The key to the development of our models is the recursive computation of  $F_{t,n}$  and  $E_{t,n}$  as  $n$  increases. The general idea is the same as for time series but we must deal with the more complex geometry of the tree and the changing dimensions of

$F_{t,n}$  and  $E_{t,n}$ . In particular, as shown in [2], it is necessary to distinguish between  $n$  even and odd and between different groups of the components of  $F_{t,n}$  and  $E_{t,n}$ . For example,  $F_{t,n}$  consists of  $F_{t,n}(w)$  in eq.(2.5) with  $|w| = n$ ,  $w \leq 0$ . Suppose that  $n$  is even and consider elements of  $F_{t,n}$  for which  $|w| = n$ ,  $w < 0$ . In this case  $w = \tilde{w}\gamma^{-1}$  for some  $\tilde{w} \leq 0$ , with  $|\tilde{w}| = n - 1$ , and by an argument exactly analogous to the time series case we obtain the recursion:

$$F_{t,n}(w) = F_{\gamma^{-1}t,n-1}(\tilde{w}) - E[F_{\gamma^{-1}t,n-1}(\tilde{w})|E_{t,n-1}] \quad (2.7)$$

This procedure identifies several projections, as in eq.(2.7), to be calculated. A key result is that these projection operators can in fact be reduced to **scalar** projections involving a single new reflection coefficient and the local averages or **barycenters** of the residuals:

$$e_{t,n} = 2^{-\lfloor \frac{n-1}{2} \rfloor} \sum_{|w| < n, w \geq 0} E_{t,n}(w) \quad (2.8)$$

$$f_{t,n} = 2^{-\lfloor \frac{n}{2} \rfloor} \sum_{|w|=n, w \leq 0} F_{t,n}(w) \quad (2.9)$$

For example, the projection in eq.(2.7) is the same for all such  $\tilde{w}$  and in fact equals  $E[F_{\gamma^{-1}t,n-1}(\tilde{w})|e_{t,n-1}]$ . This and related expressions follow from the properties of isotropy and from a very important fact: any local isometry, i.e. a map  $f$  from one subset of  $A$  onto another that preserves distances, can be extended to a full isometry on  $\mathcal{T}$ .

As a consequence of this result, we can obtain scalar Levinson recursions for the barycenters themselves [2]. These recursions introduce a sequence of reflection coefficients,  $k_n$ , and lead to a generalization of the Schur recursions for time series. In [2] we also show how these same  $k_n$  can be used to construct whitening and modeling filters for  $Y_t$  and we present a stability result analogous to the time series case. In this case, however, the condition is somewhat more complex: for  $n$  odd we have the same condition as for time series, namely  $|k_n| < 1$ ; for  $n$  even, however, we must have  $-\frac{1}{2} < k_n < 1$ . In addition we demonstrate in [2] that the class of AR( $p$ ) processes are completely equivalent to reflection coefficient sequences with  $k_n = 0$ ,

$n > p$  and we show that these processes are exactly the isotropic processes with impulse responses with support on a cylinder of radius  $[\frac{p}{2}]$  about the strict past  $\gamma^{-n}$ .

### 3 State Models and Multigrid Estimation

A second class of models displaying coarse-to-fine structure is specified by state models of the form

$$x(t) = A(m(t))x(\gamma^{-1}t) + B(m(t))w(t) \quad (3.1)$$

where  $w(t)$  is a vector white noise process with covariance  $I$ . The model eq.(3.1) describes a process that is Markov scale-to-scale and, because of this, we can readily calculate its second order statistics. For example in the case in which  $A$  and  $B$  are constant and  $A$  is stable, eq.(3.1) can describe **stationary** processes, where the covariance of  $x$  satisfies the Lyapunov equation

$$P_x = AP_xA^T + BB^T \quad (3.2)$$

and the correlation function is

$$K_{xx}(t, s) = A^{d(t,s\wedge t)}P_x(A^T)^{d(t,s\wedge t)} \quad (3.3)$$

In the scalar case, or if  $AP_x = P_xA^T$ , eq.(3.1) describes an isotropic process, but in general eq.(3.1) describes a somewhat larger set of processes.

Consider now the estimation of  $x(t)$  based on measurements

$$y(t) = C(m(t))x(t) + v(t) \quad (3.4)$$

where  $v(t)$  is white noise of covariance  $R(m(t))$ , independent of  $x$ . In many problems we may only have data at the finest level; however in some applications such as geophysical signal processing or the fusion of multispectral data, data at multiple scales is collected and must be combined. In [3] we describe three different algorithmic structures for estimating  $x(t)$  based on the measurements in eq.(3.4). One of these involves processing from one scale to the next. This structure resembles the

Laplacian pyramid processing structure [4] and can be performed extremely quickly using discrete Haar transforms.

A second structure is based on the following equality which can be derived from the Markovian structure of eq.(3.1):

$$\hat{x}(t) = L_1 \hat{x}(\gamma^{-1}t) + L_2(\hat{x}(\alpha t) + \hat{x}(\beta t)) + L_3 y(t) \quad (3.5)$$

where  $L_1$ ,  $L_2$ , and  $L_3$  are gains (depending upon scale in general). Eq.(3.5) describes a set of coupled equations from scale to scale which can be solved by Gauss-Seidel relaxation that can be structured exactly as in multigrid algorithms for the solution of partial differential equations.

A third algorithm involves a single fine-to-coarse sweep followed by a coarse-to-fine correction. In the first step we recursively calculate the best estimate of  $x(t)$  based on observations in its descendent subtree. This recursion involves three steps, which together define a new Riccati equation: a **backward prediction step** to predict from  $\alpha t$  and  $\beta t$  to  $t$ ; a **merge step**, merging these two estimates; and an **update step** incorporating the measurement at  $t$ . The merge step is the new feature that has no counterpart for standard temporal models. Once we have reached the top node of the tree, the downward sweep has the same form as the Rauch-Tung-Striebel form of the optimal smoother for temporal models (allowing of course for the proliferation of parallel calculations as the algorithm passes from coarser to finer scales): the best smoothed estimate at  $t$  is calculated in terms of the best smoothed estimate at  $\gamma^{-1}t$  and the filtered estimate at that node calculated during the upward sweep.

## References

- [1] J.P. Arnaus, G. Letac, "La formule de representation spectrale d'un processus gaussien stationnaire sur un arbre homogene," Publ. Lab. Stat. and Prob. UA 745, Toulouse.
- [2] M. Basseville, A. Benveniste, and A.S. Willsky, "Multi-Scale Autoregressive

Processes," Center for Intelligent Control Systems Report CICS-P-111, MIT, March, 1989.

- [3] K.C. Chou, A.S. Willsky, A. Benveniste, and M. Basseville, "Recursive and Iterative Estimation Algorithms for Multi-Resolution Stochastic Processes," submitted to 1989 IEEE Conf. on Decision and Control.
- [4] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Comm. on Pure and Appl. Math.*, Vol. 91, 1988, pp. 909-996.

# ROBUST IMAGE MODELS FOR IMAGE RESTORATION AND TEXTURE EDGE DETECTION<sup>†</sup>

R.L. Kashyap

School of Electrical Engineering, Purdue University  
West Lafayette, IN 47907

Kie Bum Eom

Electrical Engineering, George Washington University  
Washington, D.C. 20052

**ABSTRACT.** We present stochastic model-based methods both for restoring images corrupted by impulse noise and for detecting edges in an image which may be caused either by changes in intensity or texture.

The image is represented by a nonsymmetric half plane autoregressive model driven by impulse contaminated Gaussian noise. This type of noise is more commonly encountered in real images, unlike the pure Gaussian noise treated in earlier papers. We develop an image restoration algorithm given only the image corrupted by additive noise, the original clean image or its model being unknown. We show that this method gives much better results than currently available methods based on median filters or alpha-trimmed filters.

Next we develop methods which can detect both intensity edges and texture edges. It is well known that traditional edge detection methods have difficulty in detecting texture boundaries. We first generate edge hypotheses. We use two different procedures for confirming whether it is an intensity edge or a texture edge. We give several examples to illustrate the efficacy of the proposed approach.

**I. INTRODUCTION AND OVERVIEW.** In the past decade, there has been remarkable progress in the research on statistical image models and their applications. Statistical image models (often called random field models or spatial interaction models) represent the image intensity of a given picture by a small number of parameters. There are many applications of image models in image processing and analysis. For instance, they can be used for image synthesis (Kashyap, 1984; Cross and Jain, 1983), image restoration (Chellappa and Kashyap, 1982; Geman and Geman, 1984), image coding (Delp et al., 1979), texture boundary detection (Kashyap and Eom, 1985a), and texture analysis (Kashyap and Khotanzad, 1984).

For the application of image models to such image processing tasks, we need to estimate the parameters in the image models. There are many different estimation algorithms for different image models, but most of these methods are based on the assumption of Gaussian image intensity distribution. However, the actual distribution of image intensity

---

<sup>†</sup>Supported by U.S. Army Research Office



deviates from the Gaussian assumption, and traditional estimation methods are very sensitive to minor deviations from the Gaussian assumption. During the past few decades, many estimators which are robust to the deviations from the Gaussian assumption have been proposed (Huber, 1981), but they are rarely applied to image modelling.

Robust estimation procedures for several different image models are developed and applied to some important image processing problems such as image segmentation and image restoration in this study.

### A. Robust Statistical Procedures

There has been considerable interest in robust methods in statistics in recent years. This is because most statistical inference methods are based on rather restrictive assumptions about the observations and models, such as independence of observations, distribution of observations, etc. However, these assumptions do not always hold, and many statistical procedures are very sensitive to minor deviations from the given assumptions. For example, it is well known that least squares methods are excessively sensitive to a small number of outliers.

The term *robust* was introduced by G.E.P. Box in 1953, and a procedure is called robust if it is reasonably good (optimal or near optimal) if the assumption holds, and it is not sensitive to small deviations from the assumption. Primarily robustness implies distribution robustness, i.e., the robustness about the small deviations from the assumed distribution (usually Gaussian). The resistance to outliers is considered equivalent to the distribution robustness (Huber, 1981).

There are several types of robust procedures: M-estimators, L-estimators, and R-estimators. Among these, M-estimators have an advantage over other procedures because they can be extended to the parameter estimation problems in image models. In contrast, either L-estimators or R-estimators are difficult to generalize well beyond one parameter location or scale problems. The robust M-estimators are applied to the parameter estimation problem of causal autoregressive models. Two different outlier processes are considered, and iterative robust estimation algorithms for both of the outlier processes are developed. Theoretical properties of the proposed robust estimators are investigated.

### B. Image Models

Image models characterize the image intensity surface with a small number of parameters. Image models can be divided into two groups, namely, descriptive and generative models. A descriptive model for an image summarizes the intensity distribution into a finite number of statistics. An example is the cooccurrence matrix (Haralick, 1973) used in texture analysis. The generative model, on the other hand, allows one to synthesize an image obeying the given model by using the model description and a set of random numbers. We will restrict ourselves to generative models since they can be used for many varieties of

applications.

We can further divide the generative models into two large classes. In the first class, the observed intensity function  $y(i,j)$  is assumed to be the sum of a deterministic function - usually polynomial or sinusoid - and an additive noise. In the second class, the image intensity function is generated as the output of a transfer function whose input is a sequence of independent random variables. The transfer function represents the known structural information on the image surface; the independent random sequence accounts for the unknown part. Note that the neighboring pixels are highly correlated, unlike in the earlier case, and the transfer function accounts for the covariance.

### C. Applications

Image restoration and image segmentation are two important branches of image processing. Image restoration is needed to recover the original image from the image corrupted by noise (including impulse noise), and image segmentation procedure, especially edge detection or boundary detection, is involved in most high level image processing problems. Robust image models are developed and applied to the above image processing problems in this study.

#### 1. Image Restoration

An image may be subject to noise and interference from many different sources, and image restoration is used to remove noise from the given image. Traditionally, noise distribution is assumed as a Gaussian distribution, and many different restoration algorithms based on Gaussian assumption have been introduced (Pratt, 1978; Rosenfeld and Kak, 1982).

Recently, image models have been used in image restoration applications. For example, Chellappa and Kashyap (1982) used a simultaneous autoregressive model and conditional Markov model, Wu (1985) used a nonsymmetric half plane autoregressive model and two-dimensional Kalman filtering approach, and Geman and Geman (1984) used a family of Markov models. Even though the above examples show some successful applications of image models in the image restoration problem, all of the above methods are designed to remove Gaussian noise, and are not very effective to remove impulse noise (Pratt, 1978).

Traditionally, median filter and its generalizations (Kassam and Poor, 1985) are used to remove impulse noise (also called salt-and-pepper noise) from the noisy image. These methods are simple applications of robust location parameter estimators, such as median or  $\alpha$ -trimmed mean, where image intensity is assumed constant over a small size window. However, the restored images by these methods are blurred (Pratt, 1978).

Robust image model approaches are applied to the image restoration problem in our study. The original image intensity is assumed to follow an image model, and parameters are estimated by a robust estimation algorithm. The image is restored by applying a data cleaning algorithm with the robustly estimated parameters. The robust model-based method

performs better than any other traditional method in the experiment.

## 2. Edge and Segment Boundary Detection

Edge detection or boundary detection is a fundamental step in scene analysis. Traditionally, an edge is defined as a boundary between two uniform regions, where the intensity of each region is uniform and the intensity difference between two regions is large. Most edge detection algorithms are based on the gradient operator or the Laplacian operator (Robinson, 1977), which is sensitive to a change of intensity. Recently, some model-based edge detection approaches are proposed (Haralick, 1984; Zhou and Chellappa, 1986), but they are also based on the derivatives methods using decision rules with estimated model parameters.

For the higher level processing, the edges should be able to distinguish the shape of each object from the background of an image. However, intensity edges are sometimes not satisfactory to represent an object and distinguish it from the background, because the intensity of an object or a background is not uniform. For instance, a grass lawn in an outdoor scene is homogeneous by its texture property, but it has many intensity edges within the region. The above example suggests the necessity of detecting boundaries (or edges) by its texture property.

Image models are already used in synthesizing textures which are very similar to real textures, and the estimated parameters which are obtained by fitting an image model to the given image can be used as texture features. The texture features derived from image model or from other methods can be used to segment an image by a statistical classification method, if the number and types of textures in the given image are known in advance. However, the above prior information is generally not available.

A composite edge detection algorithm is developed in this study. The composite edge detection algorithm combines the model-based texture boundary detection method and a conventional intensity edge detection method. This algorithm detects all potential edges by a directional derivatives method, and final edges are confirmed whether they are texture edges or intensity edges. This algorithm is also compared with other conventional edge detection methods in the experiment. The composite edge detection algorithm performs better than other conventional methods which detect only intensity edges in the experiment.

## II. AR AND ARMA MODELS.

### A. Introduction

It is claimed traditionally that a complete stochastic description of an  $M \times M$  array of pixel intensities  $y(s)$  is given by the joint probability density of the  $M^2$  intensity variables  $y(\cdot)$ . Even writing down the expression is horrendous considering that the typical value of  $M$  is 128 or 256 or 512. As a consequence, it was often conjectured that probabilistic models may not be of much use in solving interesting problems in image processing. The purpose of

this paper is to draw attention to the existence of a large class of image models which can be characterized completely in terms of the second order properties of the image sequence, i.e., the correlations  $E[y(s)y(s+r)]$  or the corresponding spectral density. Consequently these models are relatively easy to analyze. It must be emphasized that the joint probability density of all the intensities is not assumed to be Gaussian.

In the beginning, we will focus our attention on the two-dimensional generalization of the autoregressive (AR) models and autoregressive moving average (ARMA) models popular in the time series analysis. Basically all these two dimensional models can handle rational spectral densities, i.e., the ratio of two linear combinations of sinusoids in the two frequency variables in the direction, just as in the one dimensional case. However, there are many differences between the 1D and 2D cases which will be highlighted in this section. For example, in the 1D case, the correlation function is an exponentially decaying function of the lag variable. But in the 2D case, one rarely encounters the exponential correlation function. Similarly in the 1D case, the driving input random sequence is both statistically independent and uncorrelated with the dependent variables in the past. In the general 2D case, the input sequence cannot possess both these properties simultaneously.

Secondly, we will consider the various possible ways of defining the weak Markov property in the 2D case. By weak, we mean that the corresponding Markov property can be described completely in terms of the second order properties like correlation or spectral density. The traditional Markov property defined in terms of the probability densities is termed as the strong Markov property. A sequence cannot be strong Markov without being weak Markov. We will characterize the various subclasses of 2D AR and ARMA models which possess various types of weak Markov property.

We recall that the general AR or ARMA models mentioned above are not recursive, in general. Still these models are generative in principle, i.e., it is possible to give an algorithm which generates a sequence which obeys a prespecified model. However, the amount of computation involved may be considerable. We will consider modifications or approximations of the AR or ARMA models so that it is relatively easy to synthesize an image obeying a given model.

#### *Preliminaries:*

We will consider a covariance stationary array of the real numbers  $\{y(i,j), -\infty < i,j < \infty\}$ ,  $i,j$  being integers.  $i,j,k$  stand for integers.  $s,t,r$  stand for two dimensional vectors specifying the grid points. Often we are given a finite  $M \times M$  image  $\{y(i,j), (i,j) \in \Omega\}$ ,  $\Omega = \{(i,j): 0 \leq i,j \leq M-1\}$ .  $y(s)$  is the intensity at the grid point  $s$ . Typically if  $s = (i,j)$ ,  $i$  stands for the row number, numbered increasingly from top to bottom, and  $j$  is the column number, numbered from left to right. The corresponding vector of real frequencies is denoted by  $\lambda = (\lambda_1, \lambda_2)$ ,  $\lambda_1$  being the row frequency, and  $\lambda_2$  being the column frequency. Similarly  $z_1$  and  $z_2$  are the unit lead operators in the row and column directions, respectively. Specifically,  $z_1 y(i,j) = y(i+1,j)$ ,  $z_2 y(i,j) = y(i,j+1)$ . We will also interpret  $z_i$  as complex

variables by the relation  $z_i = \exp[\sqrt{-1} \lambda_i]$ ,  $i = 1, 2$ .  $\lambda, z, r, s, t$ , etc., will be considered to be row vectors. The vectors composed of image intensities  $y(\cdot)$  or the input primitive random variables  $v(\cdot)$ ,  $w(\cdot)$  will usually be column vectors. An image is said to have a trend if  $E[y(i,j)]$  is a deterministic function of  $i$  and  $j$ . An image is said to be covariance stationary if the covariance function defined below is a function of  $i$  and  $j$  alone and not a function of  $s$ , and hence is denoted by  $R(i,j)$

$$E[(y(s) - \bar{y})(y(s + (i,j)) - \bar{y}))] = R(i,j)$$

where  $\bar{y} = E[y(s)]$ .

A covariance stationary random field in which  $\bar{y}$  is a constant is called as weak stationary. A random field  $\{y(s)\}$  is said to be isotropic if  $R(i,j) = R(|i|,|j|) = R(j,i)$ . For a covariance stationary RF, we can define a spectral density.

$$\begin{aligned} S(\lambda) &= \sum_{s \in I^2} R(s) \exp[\sqrt{-1} s \cdot \lambda] \\ &= \sum_{s_1=-\infty}^{\infty} \sum_{s_2=-\infty}^{\infty} R(s_1, s_2) \exp[\sqrt{-1} (s_1 \lambda_1 + s_2 \lambda_2)] \end{aligned}$$

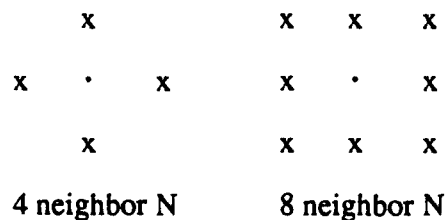
Another important second order measure of an RF model is the variogram  $V_1(s) = E[(y(s) - y(s+r))^2]$  = function of  $r$  only if  $y(\cdot)$  is weak stationary. The covariance function  $R(\cdot)$  can be recovered from  $S(\lambda)$  by the usual Fourier integral

$$R(r) = \int S(\lambda) \exp[\sqrt{-1} \lambda \cdot r] |d\lambda|$$

$$\lambda = (\lambda_1, \lambda_2), r = (i, j) \quad |d\lambda| = |d\lambda_1| |d\lambda_2|$$

Another important concept is the neighbor set. A neighbor set is a set of grid points whose coordinates are near  $\mathbf{0}$ , but  $\mathbf{0}$  itself is not a member of a neighbor set.  $N$  is said to be symmetric if  $r \in N \rightarrow -r \in N$ .

Popular neighbor sets are the ones having 4 nearest neighbors and 8 nearest neighbors.



A neighbor set  $N$  is said to be semicausal if the row coordinates (or column) of all members are the same sign. Some examples of the semicausal neighbor sets are given below.

$$\begin{matrix} \times & \times & \times & \times & \times & \times & \times & \cdot & \times & \times \\ \times & \times & \cdot & \times & \times & \times & \times & \times & \times & \times \end{matrix}$$

$$\begin{matrix} \times & \times & & & \times & \times \\ \times & \cdot & & & \cdot & \times \\ \times & \times & & & \times & \times \end{matrix}$$

where  $\cdot$  stands for origin.

### B. 2D AR Processes

Consider a real valued stationary process possessing a spectral density of the form

$$S(\lambda) = 1/[\text{positive linear combination of sinusoids in } \lambda_1, \lambda_2]$$

Our first step is to enquire whether  $y(\cdot)$  can be expressed as the output of a system characterized by a two dimensional rational transfer function of finite order, the input being some elementary stochastic process, say  $v(\cdot)$ . Toward this end, consider the system described by the difference equation where  $v(\cdot)$  is the elementary input

$$y(s) = \sum_{r \in N} \theta_r y(s+r) + \sqrt{\rho} v(s), \theta_r = \theta_{-r}, \quad (1)$$

where  $N$ , a so-called neighbor set is a set of grid points possessing symmetry, i.e., if  $s \in N$ , then  $-s \in N$ . All  $N_0$  neighbor sets can have the origin  $0$  for its member. However, not all neighbor sets may be symmetric. Define the two dimensional polynomial  $A(z_1, z_2)$  in terms of the coefficients  $\theta_r$

$$A(z_1, z_2) = A(z) = 1 - \sum_{\substack{i \ j \\ (i,j) \in N}} \theta_{i,j} z_1^i z_2^j$$

The coefficients  $\{\theta_r\}$  in (1) obey the following condition defined in terms of the polynomial  $A$ :

$$A(z_1, z_2) > 0 \quad \forall |z_1| = 1 \quad \text{and} \quad |z_2| = 1. \quad (2)$$

In addition, the input  $v(\cdot)$  in (1) is assumed to have zero mean and be orthogonal to all  $y(\cdot)$ , i.e.,

$$E[v(s)y(s+r)] = 0 \quad \forall r \neq 0, \quad (3)$$

We also assume  $E[v^2(s)] = 1$ . The parameter  $\rho$  in (1) can specify the relative power of the input term.

We can also rewrite Eq. (1) compactly in terms of the polynomial A:

$$A(z)y(s) = \sqrt{\rho} v(s). \quad (4)$$

In defining (4),  $z_i$  are interpreted as the unit lead operators in the two directions.

Equation (3) defines the process  $v(\cdot)$  only indirectly. The precise structure of the process  $v(\cdot)$  is not obvious. We will derive later an expression for the spectral density of  $v(\cdot)$  using (1)-(3).

Equation (3) can be thought of as defining a  $v(\cdot)$  process given a  $y(\cdot)$  process. It is not obvious here how to generate a  $y(\cdot)$  and a  $v(\cdot)$  sequence obeying simultaneously (1)-(3). We will later show constructively that there do exist infinite sequences  $y(\cdot)$  and  $v(\cdot)$  obeying (1)-(3).

#### *Structure of $v(\cdot)$ process*

The following theorem gives the spectral densities of the processes  $y(\cdot)$  and  $v(\cdot)$  which obey (1)-(3).

**Theorem 1:** The spectral density of  $y$  and  $v$  obeying (1)-(3) are given below:

$$S_{yy}(\lambda) = \frac{\rho}{A_1(\lambda)}, \quad (5)$$

$$S_{vv}(\lambda) = A_1(\lambda), \quad (6)$$

where  $A_1(\lambda) = A(z_1, z_2)$ ,  $z_i = \exp[\sqrt{-1} \lambda_i]$ .

*Proof:*

We will obtain a difference equation for the covariance function of  $y$ . Note  $E(y(\cdot)) = 0$ . Let  $R(t) = E[y(s)y(s+t)]$ . Multiply (1) by  $v(s)$ , take expectation on both sides, and use (3).

$$\begin{aligned} E[y(s)v(s)] &= \sqrt{\rho} E[v^2(s)] \\ &= \sqrt{\rho}. \end{aligned} \quad (7)$$

Next multiply (1) by  $y(s+t)$  on both sides and take expectation

$$\begin{aligned}
R(t) &= \sum_{r \in \mathbb{N}} \theta_r R(r-t) + \sqrt{\rho} E(v(s)y(s+t)) \\
&= \sum_{r \in \mathbb{N}} \theta_r R(r-t) + \sqrt{\rho} \delta_{t,0},
\end{aligned} \tag{8}$$

by using (3) and (7), where

$$\begin{aligned}
\delta_{t,0} &= 1 \text{ if } t = 0 \\
&= 0 \text{ otherwise}
\end{aligned}$$

Take Fourier transform of (8)

$$(1 - \sum_{r \in \mathbb{N}} \theta_r \exp[\sqrt{-1} \lambda \cdot r]) S_{yy}(\lambda) = \rho,$$

i.e., or  $S_{yy}(\lambda) = \frac{\rho}{A_1(\lambda)}$ .

To prove (6), take spectral density of both sides of (4).

$$\begin{aligned}
\rho S_{vv}(\lambda) &= \|A(z_1 = \exp(\sqrt{-1} \lambda_1), z_2 = \exp(\sqrt{-1} \lambda_2))\|^2 S_{yy}(\lambda) \\
&= \|A_1(\lambda)\|^2 S_{yy}(\lambda),
\end{aligned}$$

Using (5) for  $S_{yy}(\lambda)$ , the above equation yields the required expression for  $S_{vv}(\lambda)$  in (6).

The proof is given in some detail because it gives the difference equation for  $R_y(t)$ . In addition, the above proof indicates the existence of a process  $y(\cdot)$  obeying (1)-(3) by demonstrating its spectral density.

The  $v(\cdot)$  process is an *analog* of a one-dimensional moving average process. Its covariance function is

$$\left. \begin{aligned}
E[v(s)v(s+r)] &= -\theta_r \text{ if } r \in \mathbb{N} \\
&= 1 \text{ if } r = 0 \\
&= 0, \text{ elsewhere}
\end{aligned} \right\} \tag{9}$$

However, one important distinction between 1D and 2D cases lies in the fact that it cannot have a 2D version of moving average representation, i.e., it cannot be represented as a finite



linear combination of independent random variables. The reason is that the symmetric polynomial  $A(z_1, z_2)$  cannot be factored, i.e., it cannot be expressed, in general, as a product of 2 finite polynomials.

*Converse of Theorem 1:*

This section started with the assumption (3) on  $v(\cdot)$ . What would be the structure of the process  $y(\cdot)$  if  $v(\cdot)$  is assumed to be white? We will prove the converse of Theorem 1 and show that a process with inverse sinusoidal spectral density does not in general have any representation other than (1). The exceptions will be handled later.

**Theorem 2:** Consider a zero mean stationary process  $y(\cdot)$  having a spectral density as shown below

$$S_{yy}(\lambda) = \rho / [\text{a positive linear combination of sinusoids in } \lambda_1, \lambda_2]$$

i.e.,  $S_{yy}(\lambda) = \rho / A(z_1, z_2)$ ,  $z_i = \exp(\sqrt{-1} \lambda_i)$ , (10)

$$\text{and } A(z_1, z_2) = 1 - \sum_{r \in N} \theta_r z^r$$

where  $N$  is symmetric,  $\theta_r = \theta_{-r}$  and  $A(\cdot)$  obeys (3). Then define  $v(\cdot)$  as:

$$v(s) \triangleq (y(s) - \sum_{r \in N} y(s+r)) / \sqrt{\rho}$$

Then

$$E[v(s)y(s+r)] = 0, \quad \forall r \neq 0.$$

*Proof:* By definition

$$v(s) = A(z)y(s) / \sqrt{\rho}$$

Multiply both sides by  $y(s+t)$  and take expectation

$$R_{vy}(-t) = A(z)R_{yy}(t) / \sqrt{\rho}$$

Take Fourier transform of both sides

$$S_{vy}(\lambda) \triangleq A(z_1, z_2) S_{yy}(\lambda) / \sqrt{\rho}, \quad z_i = \exp(\sqrt{-1} \lambda_i)$$

$$= \sqrt{\rho} \text{ by (10),}$$

Hence  $E[v(s)y(s+r)] = 0$  if  $r \neq 0$ .

### *Expression for the Correlation*

In the one dimensional case, the correlation function is a linear combination of the exponentially decaying function of the lag term given that the spectral density is a ratio of linear combinations of sinusoids. Such a result is not true in the 2D case. Exponential correlation functions are rare. We can evaluate the correlations from the spectral density by numerical integration. We will give one example below.

*Example:* Consider the 4 member symmetric neighbor set.

$$\text{Let } y(s) = \theta \sum_{r \in N} y(s+r) + \sqrt{\rho} v(s)$$

$$N = [(i,j), |i| = 1 \text{ or } |j| = 1, \text{ not both}]$$

The spectral density is

$$S(\lambda) = \frac{\rho}{1 - 2\theta(\cos\lambda_1 + \cos\lambda_2)}$$

Here  $y(\cdot)$  is isotropic.

For discussion of other models, see (Kashyap, Eom, 1988).

## III. ROBUST ESTIMATION IN CAUSAL AUTOREGRESSIVE MODELS.

### A. Introduction

The importance of model-based techniques for image processing tasks such as edge detection, image synthesis, image coding, image restoration, etc., has been well documented. However, in all of these models, the image intensity array is assumed to be a multivariate Gaussian distribution. The Gaussian assumption is used primarily in estimating the parameters of the image model fitted to the image. The corresponding estimation procedure is relatively easy; for example, for the causal autoregressive model, the maximum likelihood method is the same as the least squares method. However in many applications, it is well known that the Gaussian assumption is not appropriate.

A more realistic assumption is a contaminated Gaussian noise,

$$\zeta(i,j) = \begin{cases} w(i,j), & \text{with probability } 1-\beta \\ v(i,j), & \text{with probability } \beta \end{cases} \quad (11)$$

where  $w(i,j)$  is a regular white Gaussian noise and  $v(i,j)$  is an outlier process and the ratio of outlier  $\beta$  is assumed small (less than 5%).

Unfortunately, least squares estimators or maximum likelihood estimators under the Gaussian assumption are very sensitive to minor deviations from the Gaussian noise assumption. Even a single bad data (outlier) among 1000 observations can cause a large error in the estimator. Because of this excessive sensitivity of least squares estimators, a robust estimator is needed in image models. A robust estimator should possess the following properties:

- (1) It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed noise distribution.
- (2) It should be robust in the sense that a small number of outliers impair the performance only slightly.
- (3) Somewhat larger deviations from the assumed distribution should not cause a catastrophe.

The resistance to outliers (e.g., impulse noise) is equivalent to the distribution robustness by Hampel's theorem (Huber, 1981). Many different robust estimation algorithms have been developed in the last twenty years, mostly on the location parameter estimation. These robust estimation algorithms can be classified into three large types of estimators: M-estimator, L-estimator, and R-estimator. M-estimator is a maximum likelihood type estimator and it is obtained by solving a minimization problem. L-estimator is a linear combination of ordered statistics. R-estimator is derived from the rank tests. We are mostly interested in M-estimator for the application on the image models. M-estimator is easy to extend to the problems of image models, but other types of estimators are difficult to use in problems other than simple location parameter estimation.

M-estimator is defined by the following minimization problem:

$$\text{Minimize } \sum \rho(x_i; \theta) \quad (12)$$

or solve the following implicit function:

$$\sum \psi(x_i; \theta) = 0 \quad (13)$$

where  $\rho$  is a continuous and differentiable convex function possessing bounded and continuous derivative  $\psi(x) = \frac{\partial \rho(x)}{\partial x}$ , and  $\rho$  is symmetric about the origin with  $\rho(0) = 0$ . The convexity of the  $\rho$  function ensures the equivalence of (12) and (13). The boundedness and continuity of the  $\psi$  function is essential in obtaining robustness of the M-estimator. If  $\psi$  is not bounded, then a single gross outlier can completely upset the estimator. If  $\psi$  is not

continuous, then small changes in the observation  $x_i$  may produce a large change in the estimator.

There are several different definitions of robustness of an estimator (Huber, 1981). *Qualitative robustness* is defined by weak continuity of the estimator. M-estimator is qualitatively robust if and only if the corresponding  $\psi$  is bounded and continuous. *Minimax robust estimator* minimizes the maximum degradation over  $\epsilon$  deviations. The M-estimator of location is optimal in the sense of minimax robustness. *Quantitative robustness* is defined by the property of small change in asymptotic bias and asymptotic variance in the contaminated neighborhood.

Even though a robust procedure is necessary in most image processing applications, very little research has been done on the use of a robust procedure in image processing. In this section, we develop estimation algorithms for the causal autoregressive image model.

### B. Causal Autoregressive Model

It is well known that a large class of images can be effectively represented by various types of image models involving a small number of parameters (Kashyap, 1981). Image models are already used in image coding (Delp et al., 1979), image synthesis, texture analysis, and edge detection (Kashyap and Eom, 1985a). Of course, there are many different types of image models and these can be classified into two large classes of image models by their second order statistical structures: classical short correlation models and long correlation models. These different image models and their general properties are discussed by Kashyap (1981).

The causal autoregressive model is a generalization of the one dimensional autoregressive model. This model is simple but has good modelling performance as shown in previous studies. Consider the following  $m \times n$  image (Figure 1).

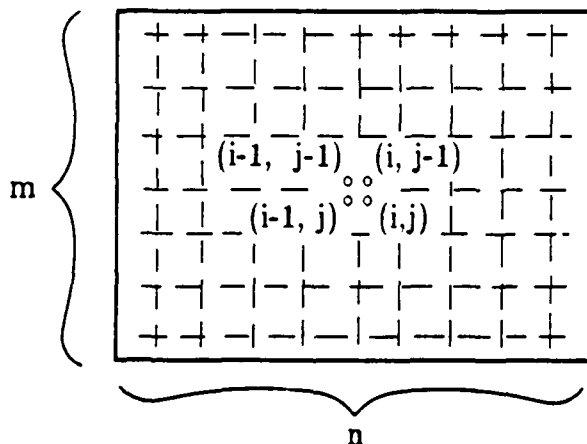


Figure 1. An  $m \times n$  image and three causal neighbors

Assume that the image intensity in this image follows the three neighbor causal autoregressive model. Let  $(i,j)$  be an index for the coordinate location and  $y(i,j)$  be the intensity at the coordinate  $(i,j)$ . Then the causal three neighbors of this pixel are  $\{y(i-1,j), y(i,j-1), y(i-1,j-1)\}$ . This causality is from the convention of raster scanning, and because of the causality, the resulting two dimensional model has all the convenience of the one dimensional model.

Suppose that  $\{\zeta(i,j)\}$  is a two dimensional white noise sequence with outliers as assumed in (11). The variance of the regular part of noise is  $\sigma^2$ . Then the three neighbor causal autoregressive model is represented by the following equation:

$$y(i,j) = \theta^T z(i,j) + \zeta(i,j) \quad (14)$$

where  $\theta$  is a parameter vector and  $z(i,j)$  is a vector consisting of intensities of three causal neighbors and unity. The last element of the vector  $z(i,j)$  is used to represent the constant grey level in the image.

$$z(i,j) = \begin{bmatrix} y(i,j-1) \\ y(i-1,j) \\ y(i-1,j-1) \\ 1 \end{bmatrix} \quad (15)$$

It is assumed that every pixel has all of its neighbors, i.e., for each pixel at  $(i,j)$ , pixels at  $(i,j-1)$ ,  $(i-1,j)$  and  $(i-1,j-1)$  are available.

We consider the robust parameter estimation of the causal autoregressive model for two cases of outliers. First case, we assume that the process  $y(i,j)$  given in (14) can be perfectly observed. In this case, the outlier process is involved only in the noise process  $\zeta(i,j)$  to generate  $y(i,j)$ . Second case, we assume that the observation  $x(i,j)$  of the process  $y(i,j)$  is corrupted by noise  $\xi(i,j)$ . It is given by the following equation:

$$x(i,j) = y(i,j) + \xi(i,j) . \quad (16)$$

The noise process  $\xi$  is assumed to contain outliers. In this case, the outliers are not only involved in generating  $y(i,j)$  but are also involved in observation. In the next section, robust parameter estimation will be discussed for these two different cases of outliers.

### C. Robust Parameter Estimation with Perfect Observations

The parameters of the image model given in (14) can be estimated by robust M-estimator. The M-estimator of the parameters in (14) is a generalization of location M-estimator. Define the following function  $Q(\theta, \sigma)$ .

$$Q(\theta, \sigma) = \frac{1}{mn} \sum \left[ \rho \left( \frac{y(i,j) - \theta^T z(i,j)}{\sigma} \right) + \frac{1}{2} \right] \sigma \quad (17)$$

where  $\rho$  is a continuous, differentiable and convex function possessing a bounded derivative, and it is symmetric about the origin with  $\rho(0) = 0$ . Then M-estimator of the causal autoregressive model is defined by the following minimization problem:

$$\text{Minimize } Q(\theta, \sigma). \quad (18)$$

The M-estimator can also be obtained by solving the following two equations simultaneously.

$$\nabla_{\theta} Q(\theta, \sigma) = \frac{-1}{mn} \sum \psi \left( \frac{y(i,j) - \theta^T z(i,j)}{\sigma} \right) z^T(i,j) = 0 \quad (19)$$

$$\frac{\partial Q(\theta, \sigma)}{\partial \sigma} = \frac{1}{2} - \frac{1}{mn} \sum \chi \left( \frac{y(i,j) - \theta^T z(i,j)}{\sigma} \right) = 0 \quad (20)$$

where  $\psi(x) = \frac{\partial \rho(x)}{\partial x}$  and  $\chi(x) = x\psi(x) - \rho(x)$ , function  $\psi$  is continuous and bounded.

The following  $\rho$ ,  $\psi$ , and  $\chi$  functions satisfy the above conditions on these functions. In this section, it is assumed that the following functions are used in our robust estimation algorithm.

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq c \\ c|x| - \frac{1}{2}c^2, & |x| \geq c \end{cases} \quad (21)$$

$$\psi(x) = \frac{d\rho(x)}{dx} = \begin{cases} c, & x \geq c \\ x, & -c \leq x \leq c \\ -c, & x \leq -c \end{cases} \quad (22)$$

$$\chi(x) = x\psi(x) - \rho(x) = \frac{1}{2}[\psi(x)]^2 \quad (23)$$

### Asymptotic Property

The asymptotic property of the robust M-estimator for autoregression is investigated by Nasburg and Kashyap (1975). The asymptotic property of one dimensional autoregression is also applicable to two dimensional causal autoregressive model. First the following conditions are assumed:

- (i)  $\{y(i,j)\}$  is a weakly stationary random sequence.
- (ii)  $\psi(\cdot)$  is an odd, monotone increasing function satisfying a Lipschitz condition.
- (iii) The noise process  $\zeta$  has finite moments up to third order.
- (iv)  $E[\psi(\zeta(i,j)+c)] = \psi(c)$  for all  $c$ .

Now define  $\hat{\theta}_N$  as an M-estimator which satisfies (18) and is computed with sample size  $N$ . The following Theorem 6 and Theorem 7 are from Nasburg and Kashyap (1975).

**Theorem 6 (consistency):** Under the above assumptions,

$$\hat{\theta}_N \rightarrow \theta \text{ as } N \rightarrow \infty \text{ w.p.1}$$

**Theorem 7 (Asymptotic Normality):** Under the above assumptions,  $\sqrt{N}(\hat{\theta}_N - \theta)$  converges in distribution to a normal distribution with zero mean and variance  $\frac{V_1}{V_2^2}$ , where

$$V_1 = \frac{1}{1-\theta^2} E[\psi^2(\zeta(i,j))]$$

and

$$V_2 = \frac{1}{1-\theta^2} E[\psi'(\zeta(i,j))].$$

### Choice of $\psi$ function

A good choice of  $\psi$  function is not only important for the robustness of the estimator but is also important for the fast convergence of the iterative procedure. The theoretical results in Section III.C are developed with the following monotone  $\psi$  function  $\psi_{HL}$ .

$$\psi_{HL}(x) = \begin{cases} c, & x > c \\ x, & -c \leq x \leq c \\ -c, & x < -c \end{cases} \quad (24)$$

Typical values for  $c$  are between 1.5 and 2.

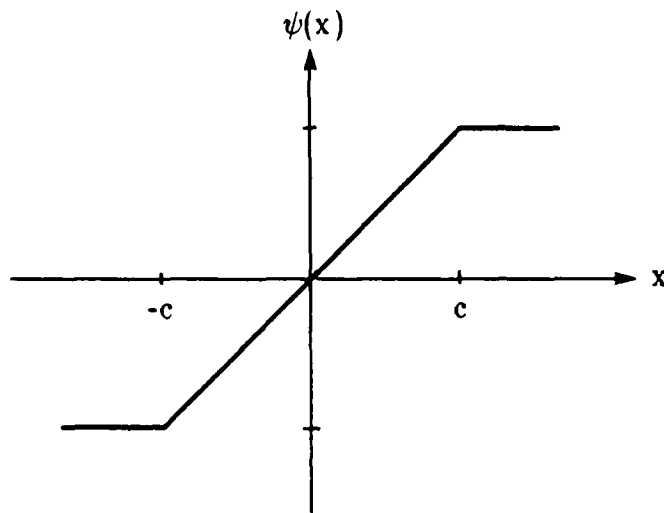


Figure 2. Hard limiter type  $\psi$  function

Even before the theoretical work on robust estimation, the  $3\text{-}\sigma$  edit rule was used for data cleaning for many years. The  $3\text{-}\sigma$  rule is a simple implementation of hard rejection rule and corresponds to the following choice of  $\psi$  function.

$$\Psi_{3\sigma}(x) = \begin{cases} x, & |x| < 3 \\ 0, & |x| \geq 3 \end{cases} \quad (25)$$

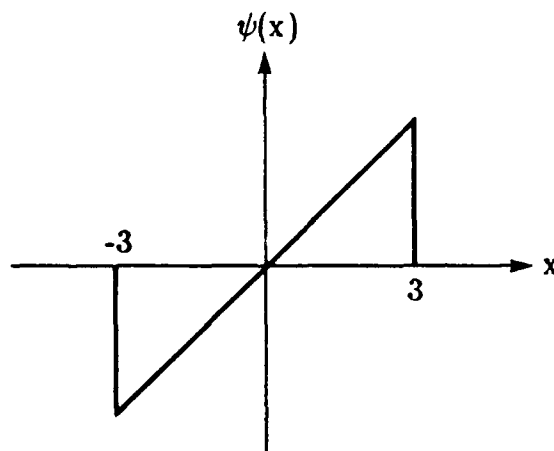


Figure 3.  $\psi$  function for  $3\text{-}\sigma$  rule

The above  $\psi$  function is obviously not continuous. The discontinuity of the  $\psi$  function is not desirable for robust estimation as discussed before.



Another interesting  $\psi$ -function is the following Hampel's  $\psi$ -function. The Hampel's function is also continuous but returns to zero outside of some interval. It is known that the redescending  $\psi$  function yields higher efficiencies than monotone  $\psi$  function for extremely heavy tailed distributions (Huber, 1981; Rey, 1983). This advantage of the redescending  $\psi$  function is also confirmed in our experiment: The procedure converges much faster with Hampel's redescending  $\psi$  function (26) than with Huber's monotone  $\psi$  function (24). This function performed best with parameters  $a=2$ ,  $b=2.5$ ,  $c=4.5$  in our experiment.

$$\Psi_{HA}(x) = \begin{cases} x, & |x| \leq a \\ \frac{a}{(b-a)}(b-x), & a < x \leq b \\ \frac{a}{(b-a)}(b+x), & -b \leq x < -a \\ 0, & |x| > b \end{cases} \quad (26)$$

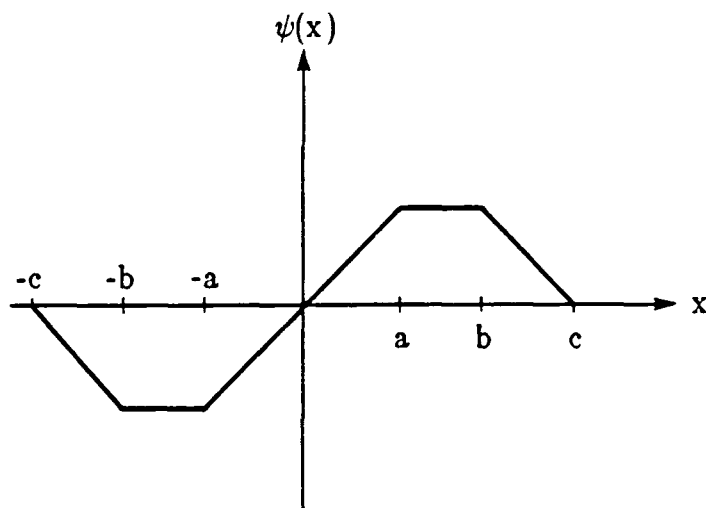


Figure 4. Hampel's  $\psi$  function

These three different  $\psi$  functions are compared in the experiment, and the best performing function is chosen in our algorithm. The Hampel's function performed better than other functions in our experiment with the parameter values given above.

#### IV. IMAGE RESTORATION WITH ROBUST IMAGE MODELLING TECHNIQUES.

##### A. Introduction

Restoration of an image in the presence of noise is one of the fundamental problems in image processing. Let  $x(i,j)$  be the observed image intensity of the original (uncorrupted) image intensity  $y(i,j)$  at the location  $(i,j)$  and is assumed corrupted by additive white noise

$\zeta(i,j)$ .

$$x(i,j) = y(i,j) + \zeta(i,j) \quad (27)$$

To restore image intensity  $\{y(i,j)\}$  from the observation  $\{x(i,j)\}$ , we generally make assumptions on the noise process  $\{\zeta(i,j)\}$  and the original image intensity  $y(i,j)$ . A common assumption on the noise process is that the noise distribution is Gaussian. However, the assumption of Gaussian noise has been seriously questioned, as we discussed in the previous section. A more realistic assumption is that the noise is a mixture of Gaussian and impulse noise.

$$\zeta(i,j) = \begin{cases} w(i,j), & \text{with probability } 1-\beta \\ v(i,j), & \text{with probability } \beta \end{cases} \quad (28)$$

where  $w(i,j)$  is regular Gaussian noise and  $v(i,j)$  is an outlier,  $\beta$  is the fraction of outliers and it is usually less than 5%.

There are many image restoration methods based on the Gaussian noise assumption. Chellappa and Kashyap (1982) used a spatial interaction model to represent image intensity array and restored images with minimum mean square error criterion. Geman and Geman (1984) used the equivalence of Markov random field and Gibbs distribution and restored images by a stochastic relaxation method with maximum a posteriori criterion. Bovick et al. (1985) used an order constrained least squares method. Wu (1985) used a multidimensional Kalman filtering approach and nonsymmetric half plane autoregressive model. Chan and Lim (1985) used a cascade of four 1D adaptive filters in four different directions.

Unfortunately, most image restoration methods based on the Gaussian noise assumption are not effective for impulse noise (Rosenfeld and Kak, 1982). The impulsive component of the noise, which is also called as salt-and-pepper noise, is only a small portion (usually less than 5%) of the total image but difficult to remove by the methods based on the Gaussian noise assumption, because its amplitude is much higher than the signal amplitude. The importance of this problem has been recognized for a long period of time. Traditionally, nonlinear filtering methods such as median filter (Pratt, 1978) or  $\alpha$ -trimmed mean filter (Bovick, et al., 1983) are used to remove impulse noise from the image. These methods use a sliding window and the grey level of the center pixel of the window is estimated by the median or  $\alpha$ -trimmed mean of the samples in the window. The grey level of the center pixel is replaced by this estimate.

These traditional nonlinear filtering methods such as median filter or  $\alpha$ -trimmed mean filters are based on the robust location estimator which uses a linear combination of ordered statistics (robust L-estimator) (Huber, 1981). These methods based on the ordered statistics are used in robust estimation of the location parameter from the 18th century (Rey, 1983). The median or generalized median (linear combination of ordered statistics) are resistant to the contamination of outliers. However, it is based on the assumption of constant grey level

in the window applied to the image. Obviously, this constant intensity assumption is inaccurate. The image intensity in a window is continuously changing, especially near the edges or corners. Because of this constant grey level assumption, the methods based on the linear combination of ordered statistics, such as median filter or  $\alpha$ -trimmed mean filter, have the disadvantage of blurry results. The blurring effect is more severe on the  $\alpha$ -trimmed mean filter than median filter, because of its averaging effect even if the mean square error of the  $\alpha$ -trimmed mean filter is smaller than that of median filter. Median filter generally does better in preserving edges and corners, but the well known examples (Pratt, 1978) show that it also blurs the image.

There are two difficulties in solving the blurring problem in the traditional methods such as median filter or  $\alpha$ -trimmed mean filter. First, the intensity function in the window applied to the image is unknown and difficult to be represented by a simple function. Second, the linear combination or ordered statistics method used in traditional methods have difficulty in accommodating the effect of changing intensity. Even though there has been a facet model-based approach (Yasuoka and Haralick, 1983) to reduce the blurring effect after removing impulse noise, it is based on the least squares estimator which is not robust to impulse noise. We propose a restoration method which uses a statistical image model for the representation of changing intensity and which uses a type of robust method, the so-called M-estimator.

We can use one of the image models mentioned earlier to represent intensity change in a window of the original image. The parameters of the image model can be estimated by robust M-estimator as shown in Section III. The robust M-estimator of the causal autoregressive model can be obtained by the iterative algorithm given in Section III. This estimation algorithm includes a data cleaning procedure at each iteration, and it reduces the outliers in the observed data. The convergence property of the robust parameter estimation algorithm is also discussed in Section III. The image data become noise free as the number of iterations increases, because the parameter estimates converge as the number of iterations increases by the convergence of M-estimator of the causal autoregressive model. By this data cleaning procedure, we can obtain the image from which most of the impulse noise has been removed, and the original sharpness of the edges is preserved. The iterative data cleaning procedure converges relatively fast in our experiment. In most of our experiments, the data cleaning procedure converges only after three iterations with almost noise free results. The restoration algorithm based on the robust estimation algorithm has many advantages over the traditional methods such as median filter or  $\alpha$ -trimmed mean filter. The comparison with other methods will be discussed later.

### B. Intensity Representation for Restoration

The objective of the restoration problem is to estimate the original image intensity  $y(i,j)$  from the given sequence of  $x(i,j)$ . We will fit a causal autoregressive model for the original (noise free) image  $y(\cdot)$ .

Let  $(i,j)$  be an index for the coordinate location and  $y(i,j)$  be the intensity at the location  $(i,j)$ . Then the three neighbor causal autoregressive model is represented by the following equation:

$$y(i,j) = \theta^T z(i,j) + \zeta(i,j) \quad (29)$$

where  $\theta$  is a parameter vector,  $\{\zeta(i,j)\}$  is a two dimensional white noise sequence with outliers as in (28), and  $z(i,j)$  is a vector consisting of intensities of three causal neighbors and unity. The last element of the vector  $z(i,j)$  is used to represent constant grey level in the image.

$$z(i,j) = \begin{bmatrix} y(i,j-1) \\ y(i-1,j) \\ y(i-1,j-1) \\ 1 \end{bmatrix} \quad (30)$$

It is assumed that every pixel has all of its neighbors, i.e., for each pixel at  $(i,j)$ , pixels at  $(i,j-1)$ ,  $(i-1,j)$  and  $(i-1,j-1)$  are available.

We assume that the observation  $x(i,j)$  of the process  $y(i,j)$  is corrupted by noise  $\xi(i,j)$ . It is given by the following equation:

$$x(i,j) = y(i,j) + \xi(i,j) . \quad (31)$$

The noise process  $\xi$  is assumed to contain outliers.

### C. Image Restoration Algorithm

The purpose of image restoration is to remove noise, including impulse noise, from the image. The image degradation process can be represented by the following equation:

$$x(i,j) = y(i,j) + \xi(i,j)$$

where  $x$  is the observation,  $y$  is the original image intensity, and  $\xi$  is the noise process with outlier. Image restoration involves estimation of the original intensity  $y$  from the observation  $x$ . For a small sized image, original image intensity can be modelled by a causal autoregressive model. If the original image intensity indeed obeys a causal autoregressive model, then the original image intensity can be recovered by the robust estimation algorithm for the

noisy observation case (Eom, Kashyap, 1988). The data cleaning procedure removes outliers at each iteration without degrading the original signal.

The restoration method based on the robust image model has an advantage over conventional methods such as median filter or  $\alpha$ -trimmed mean filter. The robust image model-based method does not blur images after restoration. Conventional methods, such as median filter or  $\alpha$ -trimmed mean filter, replace every pixel by its location estimates. Because these methods are based on the constant intensity assumption, the details of the original image are significantly blurred.

This procedure at each iteration is described in the following block diagram (Figure 5) and the algorithm is also summarized below.

### Image Restoration Algorithm

1. Divide the image into small sized (8x8) windows. The following procedures in steps 2-6 are applied for each window.
2. Let  $\{x(i,j)\}$  represent the given noisy data in the window and  $\{y^{(k)}(i,j)\}$  represent the cleaned data at the k-th iteration. Initially,  $y^{(0)}(i,j) = x(i,j)$  for all  $(i,j)$ . Compute initial estimators  $\theta^{(0)}$  and  $\sigma^{(0)}$  by the least squares method.

$$\theta^{(0)} = \left[ \sum_{i,j} z^{(0)}(i,j) z^{(0)T}(i,j) \right]^{-1} \left[ \sum_{i,j} z^{(0)}(i,j) y^{(0)}(i,j) \right] \quad (32)$$

and

$$\sigma^{(0)2} = \frac{1}{mn} \sum_{i,j} [y^{(0)}(i,j) - \theta^{(0)T} z^{(0)}(i,j)]^2 \quad (33)$$

where m and n are row and column dimensions of the image and  $z^{(k)}(i,j)$  is the following state vector.

$$z^{(k)}(i,j) = \begin{bmatrix} x^{(k)}(i,j-1) \\ x^{(k)}(i-1,j) \\ x^{(k)}(i-1,j-1) \\ 1 \end{bmatrix} \quad (34)$$

3. Consider k-th iteration,  $k > 0$ . Compute residuals  $r^{(k)}(i,j)$  and modified residuals  $\hat{r}^{(k)}(i,j)$  by the following formula with the estimated parameters computed in step 2 for all pixels in the window.

$$r^{(k)}(i,j) = y^{(k)}(i,j) - \theta^{(k)T} z^{(k)}(i,j) \quad (35)$$

$$\hat{r}^{(k)}(i,j) = \psi \left[ \frac{r^{(k)}(i,j)}{\sigma^{(k)}} \right] \sigma^{(k)} \quad (36)$$

where  $\psi$  is a bounded and continuous functions as discussed in Section III (e.g., Hampel's redescending  $\psi$ -function).

4. Restore image by the following rule (data cleaning)

$$y^{(k+1)}(i,j) = \theta^{(k)T} z^{(k)}(i,j) + \hat{r}^{(k)}(i,j) \quad (37)$$

5. Update estimators of parameter  $\theta$  and scale parameter  $\sigma^2$  by the following formula.

$$\theta^{(k+1)} = \theta^{(k)} + \left[ \sum_{i,j} z^{(k)}(i,j) z^{(k)T}(i,j) \right]^{-1} \left[ \sum_{i,j} z^{(k)}(i,j) \hat{r}^{(k)}(i,j) \right] \quad (38)$$

and

$$\sigma^{(k+1)^2} = \frac{1}{mn} \sum_{i,j} [\hat{r}^{(k)}(i,j)]^2 \quad (39)$$

6. Repeat steps 3-5 until the difference between estimates in successive iterations becomes small.

The properties of the algorithm are discussed in (Eom and Kashyap, 1988).

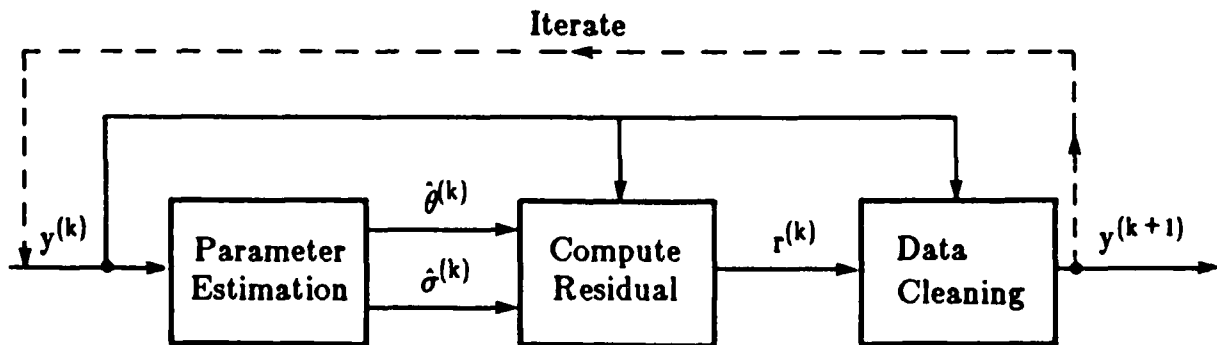


Figure 5. Block diagram of image restoration method at each iteration.  $y^{(k)}$  and  $y^{(k+1)}$  are cleaned data at  $k$ -th and  $(k+1)$ -th iterations, respectively,  $\hat{\theta}^{(k)}$  and  $\hat{\sigma}^{(k)}$  are parameter estimates obtained by algorithm 1, and  $r^{(k)}$  is the residual.

## D. Experimental Results

The restoration algorithm based on the robust modelling approach is applied to five different pictures as shown in Figure 6. Figure 6.a is a 256x256 picture of a bridge. Figure 6.b is a 256x256 picture of the face of a monkey. Figure 6.c is a 256x256 picture of a girl. Figure 6.d is a 256x256 picture of an outdoor scene. Figure 6.e is a 512x512 aerial picture of Purdue University campus. All of these pictures are digitized into 256 grey levels. To measure the performance of different algorithms on the noisy pictures, contaminated images are constructed by adding both Gaussian (0,100) noise and 5% of impulse noise to the originals given in Figure 6. The generated impulse noise has only 2 grey levels, 0 (black) and 255 (white), both with the same probability. In the robust model-based algorithm, Hampel's  $\psi$ -function is used in all experiments. Experiments are designed to clarify three different aspects of the restoration process. First, the convergence of the restoration algorithm is shown with these noisy pictures and the rate of convergence is measured experimentally. Second, the mean square error of three different restoration algorithms, namely, model-based algorithm, median filter, and  $\alpha$ -trimmed mean filter, are compared for different window sizes and different images. Third, the overall performance of three different restoration algorithms are compared qualitatively for different noisy images.

### **Convergence of Image Restoration Algorithm**

The robust model-based restoration algorithm is applied to the contaminated images. Mean square error of the cleaned image is computed at each iteration.

Figures 7.a, 7.b, and 7.c are plots of mean square errors versus the number of iterations for the outdoor scene (Figure 6.d), the girl's image (Figure 6.c) and the bridge scene (Figure 6.a), respectively. Contaminated pictures are made by adding Gaussian (0,100) noise and 5% of impulse noise to the images in Figure 6. Initial mean square errors in all cases are very large because of the additive noise, but they decrease considerably fast in the first two iterations. The mean square error stabilizes in less than three iterations. The convergence of the data cleaning method is also fast (less than three iterations).

### **Mean Square Error Comparison of Image Restoration Methods**

Four different types of image restoration methods with different sizes of windows, 3x3, 5x5, and 7x7, are used in this experiment. These are the mean filter, median filter,  $\alpha$ -trimmed mean filter with the trimming ratio  $\alpha=0.15$ , and the robust model-based method. Note that the popular choice of  $\alpha$  is in the range from 0.1 to 0.15, and the method performed best with choice  $\alpha=0.15$  in our experiment. In the case of robust model-based method, the fixed window size of 8x8 is used. The choice of 8x8 is from convenience and a small change of window size would not adversely affect the performance, because the fitted image model will not change significantly.

Four contaminated images are obtained from the originals in Figure 6 by the same procedure explained in the above section. Different restoration methods which we discussed in

the above are applied to these contaminated images, and mean square error of restored images are computed. In the case of median filter, mean filter, and  $\alpha$ -trimmed mean filter, the mean square error is computed for different window sizes, but in the case of robust model-based method, the plotted mean square error is for the fixed window size 8x8. The computed mean square error is plotted with respect to window size.

Figures 8.a, 8.b, 8.c, 8.d are plots of mean square error computed by different methods for the originals of the outdoor scene (Figure 6.d), girl's images (Figures 6.c), bridge scene (Figure 6.a), and aerial picture of Purdue University campus (Figure 6.e), respectively. The results are consistent for all different types of images. All traditional methods result in relatively large values of mean square error on most of images, especially on the images having many edges. For example, in the outdoor scene, minimum values of mean square error of mean filter, median filter, and  $\alpha$ -trimmed mean filter are 690.1441, 651.1638, and 220.2222, respectively. In contrast, the mean square error of the robust model-based method is 103.9669. The difference, which is significant, corresponds to the fact that the intensity in a window cannot be approximated by a constant because of the edges and corners. Traditional methods have small values of mean square error at window sizes 3x3 or 5x5 depending on the types of images. Mean filter performs worst on all images tested, as expected, and median filter has slightly lower mean square error than that of mean filter.  $\alpha$ -trimmed mean filter performs better than median filter or mean filter but its mean square error is always larger than that of robust model-based method on all images tested. The mean square error comparison shows that the robust model-based method performs better than any other conventional methods on tested images. The minimum values of mean square error in conventional methods are 220.2222 for outdoor scene, 80.6720 for girl, 92.1115 for bridge, and 253.7658 for Purdue campus, respectively. Mean square errors of our approach are 103.9669 for outdoor scene, 52.5648 for girl, 47.3367 for bridge, and 189.1443 for Purdue campus. The level of mean square error of conventional methods are always higher than that of robust model-based method. The detailed comparison is summarized in Table III.

Table III. Mean square error comparison of different restoration methods on four different types of images.

Image	MSE of robust model method	MSE of mean filter	MSE of median filter	MSE of $\alpha$ -TM filter
Outdoor	103.9669	690.1441	651.1638	220.2222
Girl	52.5648	318.9122	300.3172	80.6720
Bridge	47.3367	264.6290	216.3370	92.1115
Campus	189.1433	453.9291	401.6255	253.7658



## Qualitative Comparison of Image Restoration Methods

The noisy images and images restored by different restoration algorithms are shown in Figures 9-10. Figures 9-10 are results on the originals of Figure 6.a-b in the same order. The upper left corner of the each picture of Figures 9-10 is the noisy picture contaminated by noise and is generated by adding white Gaussian (0,100) noise and 5% of impulse noise to the original. This image shows a typical salt-and-pepper noise pattern as well as Gaussian noise degradation. This noisy picture is used to obtain restored images by different methods.

The upper right corner of each picture in Figures 9-10 is the restored image by robust model-based method. This image is obtained after three iterations of data cleaning process. The impulsive noise is almost completely absent and residual Gaussian noise is hardly noticeable. The fine details of the restored image are well preserved. As a matter of fact, almost all details of the original in Figure 6 are still well shown in this picture. For example, guy wire of the bridge (Figure 9), hair of the monkey's face (Figure 10), etc., have sharp edges as in the original. This result shows the important ability of the image model-based approach: it can preserve the edges and corners even with superior performance of noise removal.

The lower left corner of each picture of Figures 9-10 is the image restored by median filter with a 5x5 window. Note that the 5x5 window gives lowest mean square error as well as the 3x3 window in the experiment of the former section. Most of the impulse noise are removed in this picture, but it is much more blurred than the result of robust model-based method. This blurring effect can be more easily observed in the images with many edges and corners than in the images with large areas with constant grey levels. Guy wire and details of the bridge frame (Figure 9) and hairs and eyes of monkey's face (Figure 10) are blurred and cannot be observed in these median filtered images. The regions with small intensity variations are replaced by constant grey level and the transitions between different regions are rather abrupt. This effect is typical in the median filter, and it is because the median filter fails in smoothing images. These effects can be observed in the tower region of the bridge (Figure 9).

The lower right corner of each picture of Figures 9-10 is the image restored by  $\alpha$ -trimmed mean filter with a 5x5 window and  $\alpha=0.15$ . Note that the choice of  $\alpha=0.15$  is considered a good choice in previous studies (Rey, 1983; Bickel, 1977). Even though the  $\alpha$ -trimmed mean filter has lower mean square error than the median filter, the image restored by the  $\alpha$ -trimmed mean filter is more blurred than the median filter. Edges and corners of the image convey more information to human perception and because of this, the image restored by  $\alpha$ -trimmed mean filter is worse than median filter in the visual comparison even though it has smaller mean square error. For example, tower and guy wire in the bridge (Figure 9) and hairs and eyes of monkey's face (Figure 10) are blurred. It is also not successful in removing impulse noise and has considerable residual noise caused by impulse noise. These residual noise can be observed in all images (Figures 9-10).



(a)



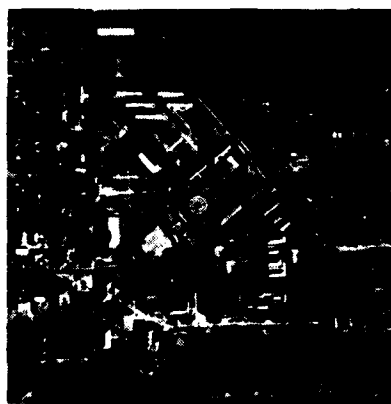
(b)



(c)



(d)



(e)

Figure 6. Originals

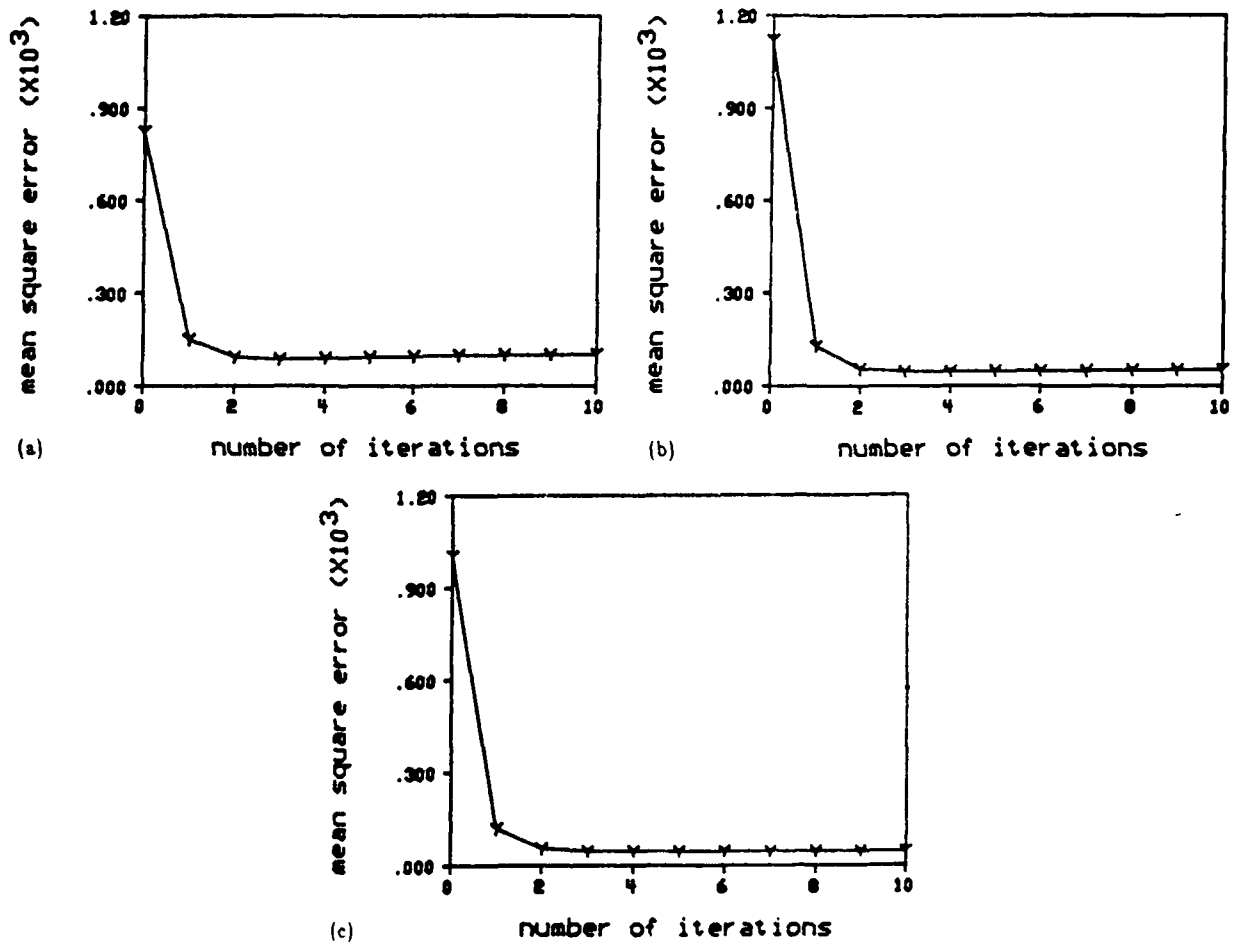


Figure 7. Convergence of mean square errors. (a) The outdoor scene. (b) The image of a Girl. (c) The Bridge scene.

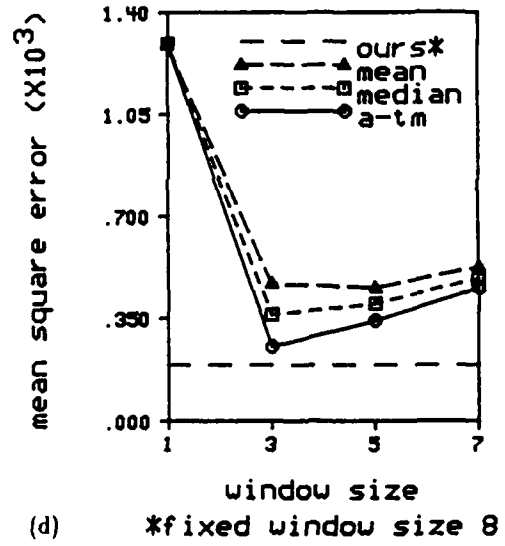
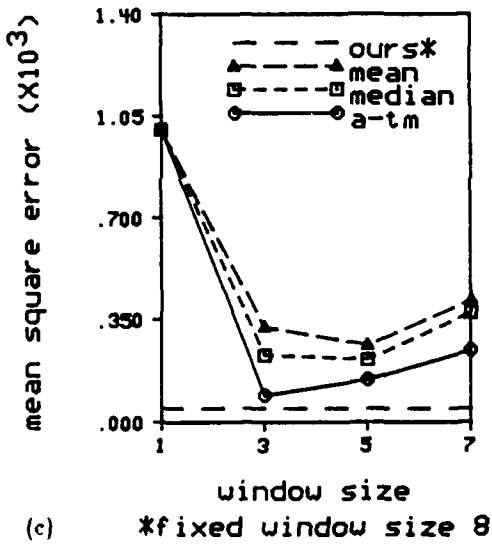
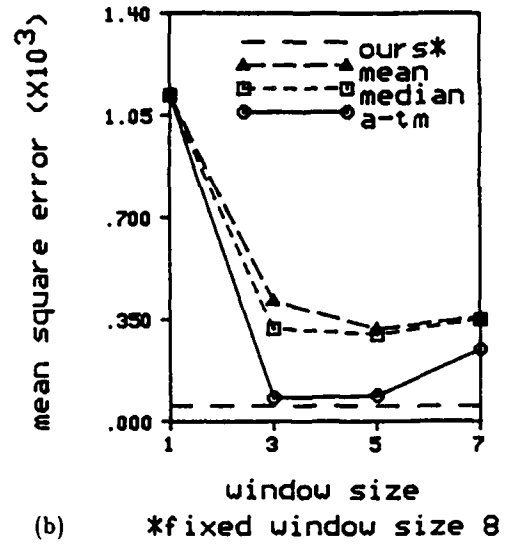
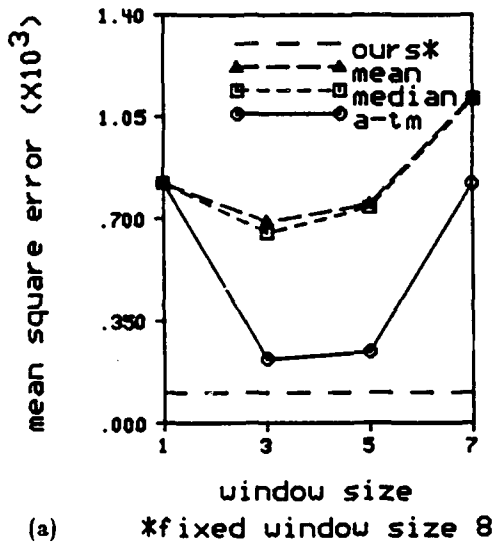


Figure 8. Mean square error comparisons of different methods. (a) The outdoor scene. (b) The image of a Girl. (c) The Bridge scene. (d) Purdue campus scene.

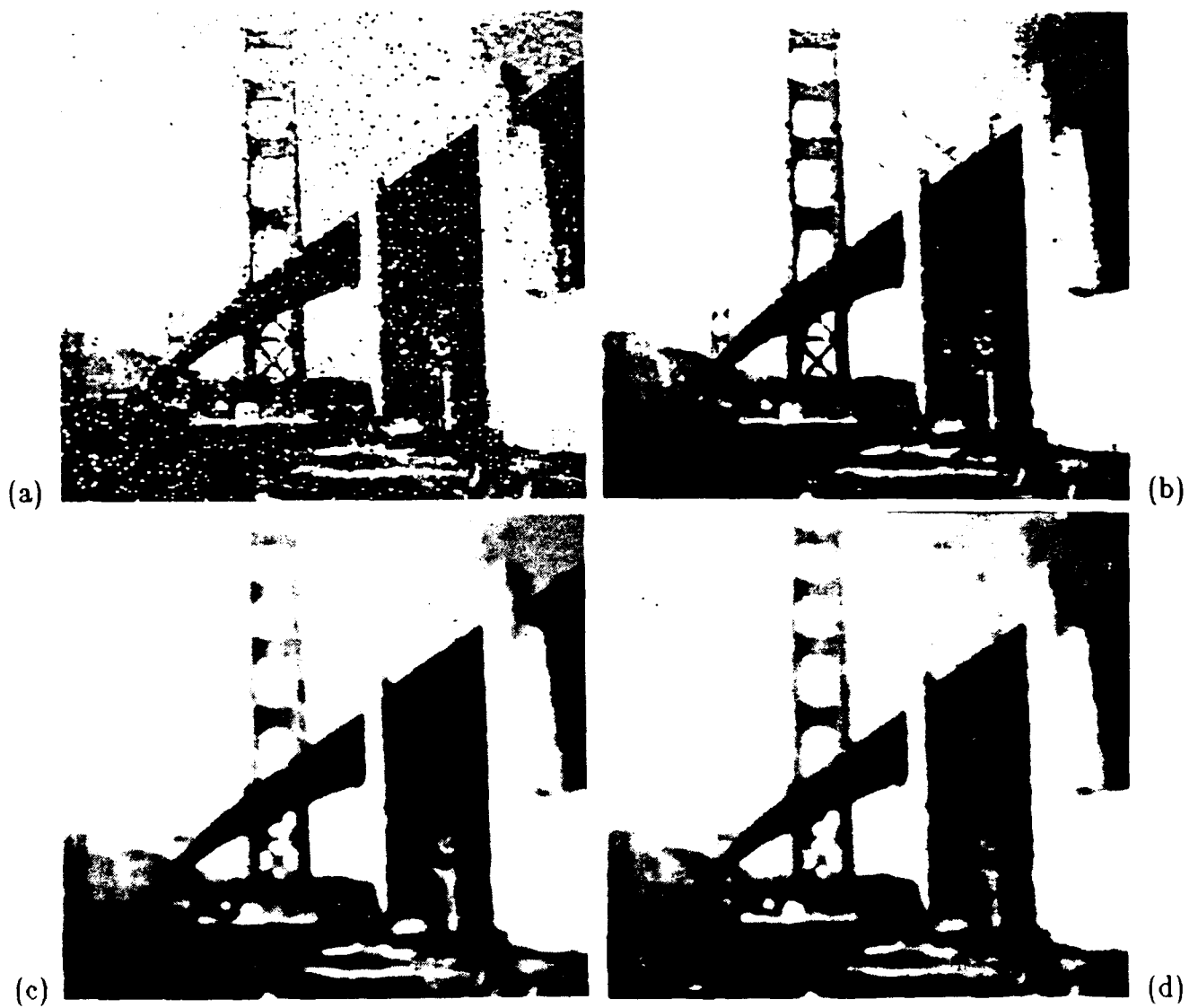


Figure 9. Qualitative comparison for Bridge picture. Most of details, such as guy wire, are clearly shown in the result of model-based approach, but are not clear in others. (a) Contaminated image. (b) Robust model approach. (c) Median filter. (d)  $\alpha$ -trimmed mean filter.

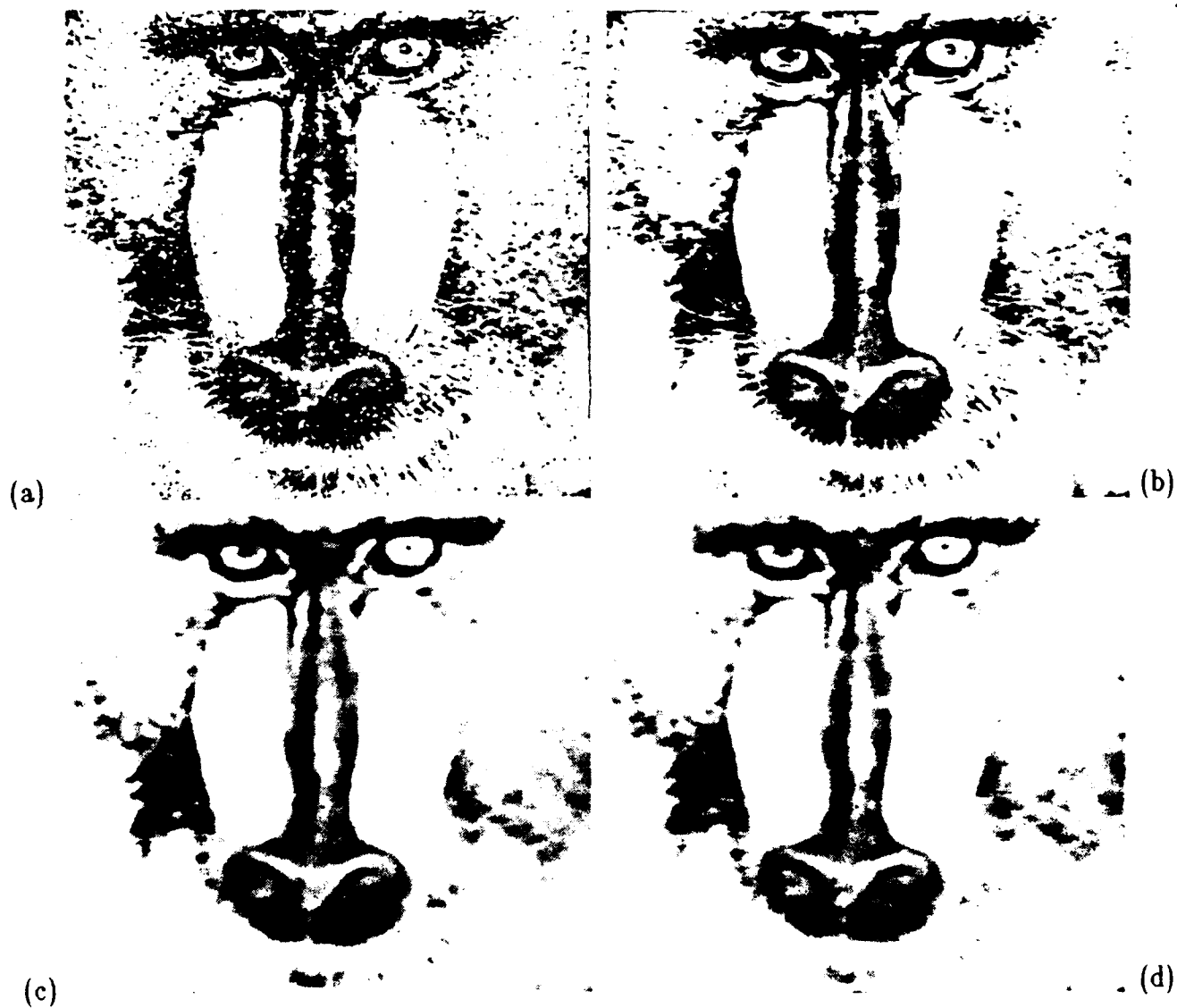


Figure 10. Qualitative comparison for Monkey picture. Most of details, such as hair, eyes, etc., are clearly shown in the result of model-based approach, but are not clear in others. (a) Contaminated image. (b) Robust model approach. (c) Median filter. (d)  $\alpha$ -trimmed mean filter.

## V. COMPOSITE EDGE DETECTION.

### A. Introduction

Edge detection is not only an important topic in image processing in its own right, but also as a tool for the important problem of image segmentation. The traditional methods of edge detection based on the windows of Robert, Prewitt or Sobel (Rosenfeld and Kak, 1982) are based on the fact that there is a sharp change in the intensity on either side of an edge pixel. We can call these types of edges as step edges. Instead of using the step function, we can employ other types of functions like the roof function (Brady, 1982) to characterize the local intensity behavior near the edge. In recent times there have been attempts at characterizing and detecting edges by considering the intensity density over a broad area around the edge pixels. Examples of these methods are the Laplacian on Gaussian operator (Marr and Hildreth, 1980), or difference of Gaussians (DOG) (Wilson and Bergen, 1979), the facet model-based methods (Haralick, 1984), and the causal autoregressive model-based methods (Zhou and Chellappa, 1986).

However, there is another mechanism of creation of an edge which has recently received some attention. Consider the pixels which are at the boundary of two textures, say cotton canvas and raffia. There is no sharp intensity change at the boundary, yet everyone will perceive the existence of a sharp edge at the boundary of the two textures. We can characterize these edge pixels as texture edges. Recently there has been considerable interest in developing methods which can detect all the texture boundaries in a scene involving several textures (Kashyap and Eom, 1985a). These algorithms effectively locate most of the boundaries between the textures which are perceived by a human observer. Of course, any real life images such as an outdoor scene or airport scene will have both intensity edges and texture edges.

When we apply the methods mentioned earlier for detecting edges on outdoor scenes, the final result is not satisfactory for several reasons. For instance, the result given by the Laplacian on Gaussian approach or the facet model approach yields a lot of micro edges corresponding to the leaves of a tree or the inside of a shrub in the house image. These micro edges do not convey much information and only add to the confusion. Even the edges due to runways or highways are often smeared. The texture boundaries are never sharply delineated. These methods cannot distinguish between the edges within a texture like the wood texture and the boundary between the two textures, say wood and cork.

The texture based algorithms also have their limitations. Since the size of the windows or masks needed to detect or discriminate between textures is much bigger than that used in the other methods, sharp edges like highways or runways in the airport are missed by these images.

The purpose of this section is to develop a composite edge detection approach which can detect all types of edges including intensity edges and texture edges. We employ a two stage approach. In the first stage, we use an algorithm which determines all the possible

pixels in an image which are *potential edges* (either intensity edge or texture edge). In addition, the algorithm gives the direction of the potential edge. In the second stage, we submit each candidate edge pixel to two procedures, one of which is designed to test whether the candidate edge pixel is a texture edge or not, and the other is designed to test whether the candidate edge is an intensity edge. We accept only those edges which pass at least one of the two tests. The procedure for testing for the texture edge is a likelihood approach based on a causal autoregressive model. The procedure for testing for a step edge is fairly conventional.

The comprehensive algorithm (Eom, Kashyap, 1987, 1989a, 1989b) presented here was applied to several images, both synthetic as well as real life images. The synthetic images are checkerboard images involving two different textures alternately. Each texture has its own internal structure. The other two images are the outdoor scene and the airport image. We give the results of our algorithm. To bring out the highlights of our approach, we also give the results of the two popular edge detection approaches in recent literature, namely the Laplacian on Gaussian method and the facet model approach, for all four images. The overall approach is given in Figure 11.

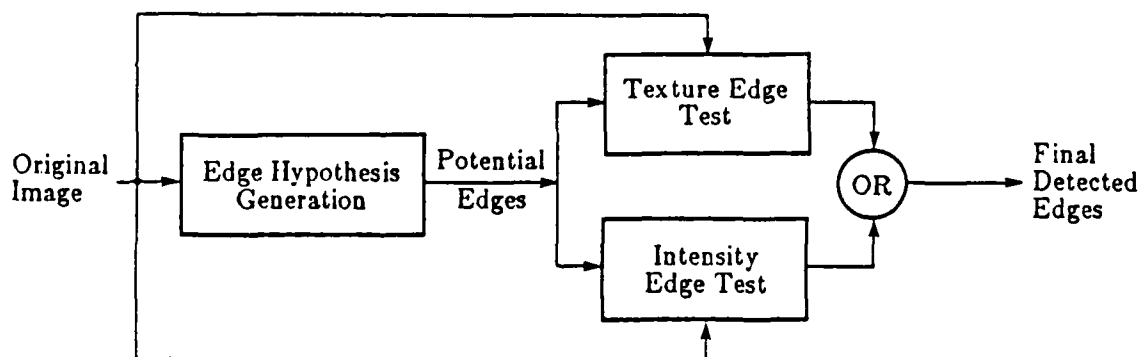


Figure 11. Block diagram of the composite edge detection algorithm

### B. Edge Hypothesis Generation (Algorithm 1)

As indicated in Figure 11, the first step in the composite edge detection algorithm is identifying all pixels which are potential edge pixels. In this process, all potential edge pixels should be detected whether they are step edges, roof edges, or texture edges. Intensity edges, such as step edges or roof edges have abrupt changes of intensity at the edge pixels and these can be detected by a derivative operator. Intensity transition is also involved at the



texture boundary as well as at microedges inside of each texture and it can be detected by a derivative operator. The algorithm used here is based on directional derivatives. We use 3x3 masks so that the edge pixels detected here are relatively sharp. Large mask operators are not adequate because they yield potential edge pixels which are situated away from the actual or true edge pixels.

Let  $g(x,y)$  be the image intensity at position  $(x,y)$ . The first order directional derivative is given by the following equation.

$$\frac{\partial g}{\partial \alpha} = \frac{\partial g}{\partial x} \cos \alpha + \frac{\partial g}{\partial y} \sin \alpha \quad (40)$$

where  $\frac{\partial g}{\partial x}$  and  $\frac{\partial g}{\partial y}$  are partial derivatives of  $g$  in  $x$  and  $y$  directions and can be obtained by convolving with the following differencing operators  $D_x$  and  $D_y$ .

$$D_x = \frac{1}{3} \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad D_y = \frac{1}{3} \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad (41)$$

i.e.,

$$\frac{\partial g}{\partial x}(i,j) = \sum_{k,l=-1}^1 g(i+k,j+l) D_x(k,l) \quad (42.a)$$

$$\frac{\partial g}{\partial y}(i,j) = \sum_{k,l=-1}^1 g(i+k,j+l) D_y(k,l) \quad (42.b)$$

The angle of gradient direction is

$$\alpha = \tan^{-1} \left( \frac{\partial g / \partial y}{\partial g / \partial x} \right). \quad (43)$$

Likewise, the second order directional derivative is given by the following.

$$\frac{\partial^2 g}{\partial \alpha^2} = \frac{\partial^2 g}{\partial x^2} \cos^2 \alpha + 2 \frac{\partial^2 g}{\partial x \partial y} \cos \alpha \sin \alpha + \frac{\partial^2 g}{\partial y^2} \sin^2 \alpha, \quad (44)$$

where second order partial derivatives  $\frac{\partial^2 g}{\partial x^2}$ ,  $\frac{\partial^2 g}{\partial x \partial y}$  and  $\frac{\partial^2 g}{\partial y^2}$  are obtained by convolving  $g$  with the following second order differencing operators  $D_{xx}$ ,  $D_{xy}$  and  $D_{yy}$ .

$$\begin{aligned}
D_{xx} &= \frac{1}{3} \begin{bmatrix} 1 & -2 & 1 \\ 1 & -2 & 1 \\ 1 & -2 & 1 \end{bmatrix} & D_{yy} &= \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ -2 & -2 & -2 \\ 1 & 1 & 1 \end{bmatrix} \\
D_{xy} &= \frac{1}{3} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} & & (45)
\end{aligned}$$

i.e.,

$$\frac{\partial^2 g}{\partial x^2}(i,j) = \sum_{k,l=-1}^1 g(i+k,j+l)D_{xx}(k,l) \quad (46.a)$$

$$\frac{\partial^2 g}{\partial x \partial y}(i,j) = \sum_{k,l=-1}^1 g(i+k,j+l)D_{xy}(k,l) \quad (46.b)$$

$$\frac{\partial^2 g}{\partial y^2}(i,j) = \sum_{k,l=-1}^1 g(i+k,j+l)D_{yy}(k,l) \quad (46.c)$$

An edge hypothesis is made at the pixel whose first directional derivative has a magnitude larger than a threshold  $t_1$  and the corresponding second order directional derivative is negative, i.e.,

$$\left| \frac{\partial g}{\partial \alpha} \right| > t_1, \quad \text{and} \quad \frac{\partial^2 g}{\partial \alpha^2} < 0 \quad (47)$$

Note that the Prewitt operator is a special case of this directional derivatives method and the Prewitt operator does not involve second order directional derivatives.

The angle of the first derivative is given by  $\alpha = \tan^{-1}\left(\frac{\partial g/\partial y}{\partial g/\partial x}\right)$ , and it can be any value between 0 to 360 degrees. The angle of edge direction is quantized into 4 directions as defined in Table IV so that a horizontal or vertical directional strips can be applied. Around each potential edge pixel, a  $m \times 2n$  strip ( $5 \times 16$  is used in this experiment) is constructed so that the strip is perpendicular to the approximated edge direction (Figure 12). For each potential edge pixel at the center of the strip, the following null hypothesis  $H_0$  is assigned.

$H_0$  = An edge exists in the given direction

The above hypothesis is tested by applying decision rules to the image strip. The details of the tests are given later.

Table IV. Quantization rule for estimated edge direction

gradient angle(degrees)	approximated direction	type of strip
315 - 45	0	horizontal
45 - 135	2	vertical
135 - 225	4	horizontal
225 - 315	6	vertical

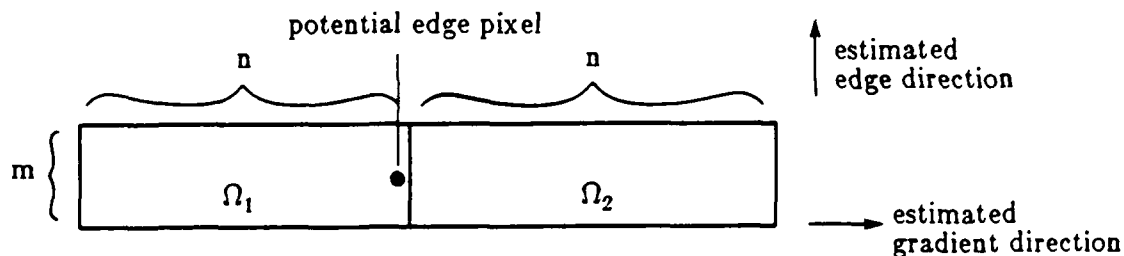


Figure 12.  $m \times 2n$  strip of image with potential edge pixel at center

### C. Confirming the Presence of Edges (Algorithm 2)

The potential edge pixels selected by the edge hypothesis generation process given in Section V.B are not the final edge pixels. Each potential edge can be either an intensity edge, a texture edge, or a spurious edge (micro-edge) caused by intensity changes inside of a texture. We want to detect only valid edges such as intensity edges or texture edges, but microedges (spurious edges) need be deleted from the potential edge map. We need to confirm valid edges at each potential edge pixel. This confirmation process involves two different types of confirmation processes. Intensity edges and texture edges have different generation mechanisms, and these need to be confirmed by separate decision processes.

Therefore, two different types of decision rules are needed to detect both texture edges and intensity edges. The first decision rule tests the existence of a texture edge at the given position and edge direction, and it is based on the likelihood ratio test with statistical texture modelling method. The second decision rule tests the existence of an intensity edge, and it is based on the differencing operator with weighted differencing. The pixels which fail in both of these tests will be deleted from the final edge map.

### 1. Confirming a Texture Edge

A texture edge in an image can be modelled as a boundary between two different texture regions. This is analogous to the intensity edge which is modelled by a boundary between two different grey levels. Detection of a texture edge is much more difficult than detection of intensity edges, because each texture region contains many microedges. The texture edges cannot be detected by the strength of gradient or Laplacian operators, and we need a method to characterize textures before detecting texture edges. Textures can be characterized by a small number of parameters after fitting the image by an image model such as causal autoregressive model.

Consider a horizontal strip of an image intensity array which is sufficiently small, so that the strip can have at most two different textures. If it has two textures, the boundary between textures can be assumed as vertical. In this strip, a texture edge is defined as the boundary between two different textures.

Consider the strip around the candidate pixel defined earlier. Let the null and alternative hypothesis be

$H_0$  = texture edge exists at the given pixel and direction

$H_1$  = no texture edge exists at the given pixel and direction.

Under the hypothesis  $H_0$ , texture in the left of the potential edge (this region will be called  $\Omega_1$ ) and texture in the right of the potential edge (this region will be called  $\Omega_2$ ) are different from each other. These two different textures are modeled by causal autoregressive models. The models in the regions  $\Omega_1$  and  $\Omega_2$  are defined below.

$$g(i,j) = \theta_1^T z(i,j) + \sqrt{\rho_1} \omega(i,j), \text{ if } 0 \leq i \leq m, 0 \leq j \leq n \text{ (region } \Omega_1) \quad (48)$$

$$g(i,j) = \theta_2^T z(i,j) + \sqrt{\rho_2} \omega(i,j), \text{ if } 0 \leq i \leq m, n < j \leq 2n \text{ (region } \Omega_2) \quad (49)$$

where  $\{\omega(i,j)\}$  is a standard 2D white noise sequence,  $\theta_1$  and  $\theta_2$  are parameter vectors for the regions  $\Omega_1$  and  $\Omega_2$ , respectively, and  $z(i,j)$  is a 4-vector.

$$z(i,j) = \begin{bmatrix} g(i,j-1) \\ g(i-1,j) \\ g(i-1,j-1) \\ 1 \end{bmatrix}$$

The parameters of the autoregressive model in the 2 regions  $\Omega_1$  and  $\Omega_2$  will be different under the null hypothesis  $H_0$ .

On the other hand, under the hypothesis  $H_1$ , the strip has only one type of texture, which is also assumed to follow a causal autoregressive model.

$$g(i,j) = \theta_0^T z(i,j) + \sqrt{\rho_0} \omega(i,j), \text{ if } 0 \leq i \leq m, 0 \leq j \leq 2n \text{ (region } \Omega_1 \cup \Omega_2) \quad (50)$$

where  $\theta_0$  is the parameter vector and  $z(i,j)$  is previously defined.

The decision rule based on the likelihood ratio test has the following form:

$$\begin{cases} \text{accept } H_0 & \text{if } \log p(g|H_0) - \log p(g|H_1) \geq K \\ \text{reject } H_0 & \text{if } \log p(g|H_0) - \log p(g|H_1) < K \end{cases} \quad (*)$$

where  $K$  is a constant. The likelihood functions  $\log p(g|H_0)$  and  $\log p(g|H_1)$  for autoregressive model are given in (Kashyap, Eom, 1988; Eom, Kashyap, 1989b). The proof can be found in the reference (Kashyap, 1982).

The texture edge detection by applying the decision rule (\*) on the pixels with edge hypothesis has several advantages over the texture boundary detection algorithms given in (Kashyap, 1985a). First, the texture edge direction is estimated in new method, and this gives more accuracy in detecting edges than applying both horizontal and vertical strips. Second, the new method tests only the existence of a texture edge, and it provides much faster processing.

## 2. Confirming an Intensity Edge

This decision rule tests the existence of an intensity edge at the pixel having edge hypothesis. The intensity edge is modeled by a step edge and the decision is made on the output of the differencing operator with weighted averaging. Briefly speaking, with the strip applied at the given pixel, the difference of the weighted average of grey levels in both sides from the potential edge pixel is computed. If this difference exceeds a threshold, the pixel is accepted. This decision rule also can be extended to detect the local maximum instead of detecting the strength of the weighted differencing operator output. Let  $W(i,j)$  be a weight function. This weight function should be asymmetric with respect to the hypothetical edge pixel and direction. Then the output of the weighted differencing operator is given by the following equation.

$$\bar{g}' = \sum_{i=0}^m \sum_{j=0}^{2n} W(i,j) g(i,j) \quad (51)$$

All edge detection window operators can be considered as a member of this weighted differencing operators. For example, Prewitt, Robert and Sobel operators (Rosenfeld and Kak, 1982) are weighted differencing operators with appropriate weight functions detecting large output as edges, Laplacian on Gaussian (Marr and Hildreth, 1980) operator is also a weighted differencing operator with a derivative of Gaussian weight function detecting local maximum of output. Many variations of weight functions are possible, but we will restrict our attention to the simple operator which can detect the step edges. Probably the simplest weighted differencing operator is the one with uniform weight function. This operator is

defined for the given strip as follows.

$$W(i,j) = \begin{cases} 1 & \text{if } 0 \leq j \leq n \\ -1 & \text{if } n < j \leq 2n \end{cases} \quad (52)$$

The above operator is used to decide the existence of an intensity edge at the potential edge pixel in our experiment. The decision is based on the strength of the operator output, i.e.,

$$\begin{cases} \text{accept edge} & \text{if } \bar{g}' > t \\ \text{reject edge} & \text{otherwise} \end{cases} \quad (53)$$

where  $t$  is a constant.

Experimental results (Figures 13-15) show good performance with this simple decision rule.

#### D. Experimental Results

The composite edge detection algorithm is tested with the following four different images (Figure 13). Figure 13.a is a 128x128 image generated from two textures chosen from Brodatz's photo album (Brodatz, 1966), grass and wood grain textures. This image has only major edges at the boundary of two textures but each square has many weak edges caused by textures.

Figure 13.b is a 128x128 original test image generated by rotating a checker board image generated similarly as Figure 13.a. Textures in this image are the same as in Figure 13.a. The major edges of this image are sloped in a 45-degree direction and each diamond pattern has many weak edges caused by intensity changes within a texture. This image is given to demonstrate that our method can detect edges which are neither horizontal nor vertical. Figure 13.c is a 256x256 monkey image.

#### **Experiment 1: Checker Board Image**

Figure 14.a is the final result of composite edge detection algorithm with a low threshold in the decision rule for the intensity edge. It shows the detected major edges at the boundary of two different textures as well as weak edges inside of each texture. The edges detected inside of textures are close to the actual edge locations. Figure 14.b is the result of composite edge detection algorithm with a high threshold in the decision rule for the intensity edge. It shows only major edges between two different textures and most of the weak edges inside the textures are eliminated. Thus an investigator can get an idea of the texture edges (corresponding to the boundaries between textures) and the intensity edges separately.

Figure 14.c is the result of Laplacian on Gaussian approach with  $\sigma = 0.5$ . Even if we alter the parameters, the final edge map is still similar to the one before. Thus, if we use this approach, we cannot distinguish the edges which are caused by the boundaries of textures

and the microedges within each texture.

Figure 14.d is a result of facet model approach. It shows detected major edges and weak edges. Even if the parameters are changed, the final edge map is similar to the Figure 14.d. Thus if we use this approach, the texture edges and intensity edges are not distinguished. Another noticeable distortion is at the corner of the square. The detected edges around the corner are distorted.

### Experiment 2: Rotated Checker Board Image

Figure 15.a is the final result of composite edge detection algorithm with a low threshold in the decision rule for the intensity edge. It shows all major edges between texture regions and weak edges inside of each texture region. The location of detected edges correspond to actual edge locations of the original image. Figure 15.b is the final result of composite edge detection algorithm with a high threshold in the decision rule for the intensity edge. It shows all major edges between different texture regions, but most of the weak edges inside of a texture region are removed without weakening major edges.

Figure 15.c is the result of Laplacian on Gaussian operator. It shows both major edges and weak edges. The final edge map does not change even if the parameters are changed. Therefore if we use this approach, texture edges and intensity edges are not distinguished. Figure 15.d is the result of facet model approach. It shows severely distorted detected edges. It contains major edges between texture regions and weak edges inside of each texture region, but weak edges cannot be separated from the major edges by changing parameters.

### Experiment 3: Monkey Image

Figure 16.a is the image of pixels having edge hypothesis which is obtained by the edge hypothesis generation process which is described in Section V.B. It shows sharp edges, and the location of these potential edge pixels are very close to actual edge location. For example, eyes of the monkey, lines in the center of the image, etc., are well detected and show good performance of this algorithm as an edge detection method. The performance as an edge detection method is superior than other edge detection methods.

Figure 16.b is the final result of the composite edge detection algorithm. Notice that most of microedges in the texture region in the cheeks of the monkey's face are removed, but most of important edges, such as eyes and nose of the monkey, lines in the center of the picture, are well preserved.

Figure 16.c is the result of Laplacian on Gaussian operator. It shows distorted major edges, and many unwanted edges caused by textures in the cheeks of the monkey's face. This picture not only includes many unwanted microedges but also shows distorted major edges. The edges in the eyes and nose region are distorted and barely distinguishable.

Figure 16.d is the result of facet model approach. Detected edges are distorted and contain many false (spurious) edges. The location of detected edges are relatively far from the

actual edge location.

### E. Discussions and Conclusions

Edges are generated in at least two different ways, namely by the difference in intensity (intensity edge) and by the difference in textures (texture edge). The importance of the texture edge is demonstrated by the examples. Conventional edge detection algorithms cannot distinguish between texture edges and intensity edges. A new edge detection algorithm which can detect both intensity and texture edges is developed.

The performance of the composite edge detection algorithm shown in this experiment can be summarized into the following two points.

1. Edge hypothesis generation procedure developed in this research can be used as an edge detection method, and the performance as an edge detection algorithm is better than other edge detection methods.
2. Our composite edge detection algorithm is flexible enough to detect both major and weak edges by changing threshold. In other words, it can detect only major edges without detecting microedges which are caused by texture for high threshold and can detect both major edges and microedges for lower threshold.



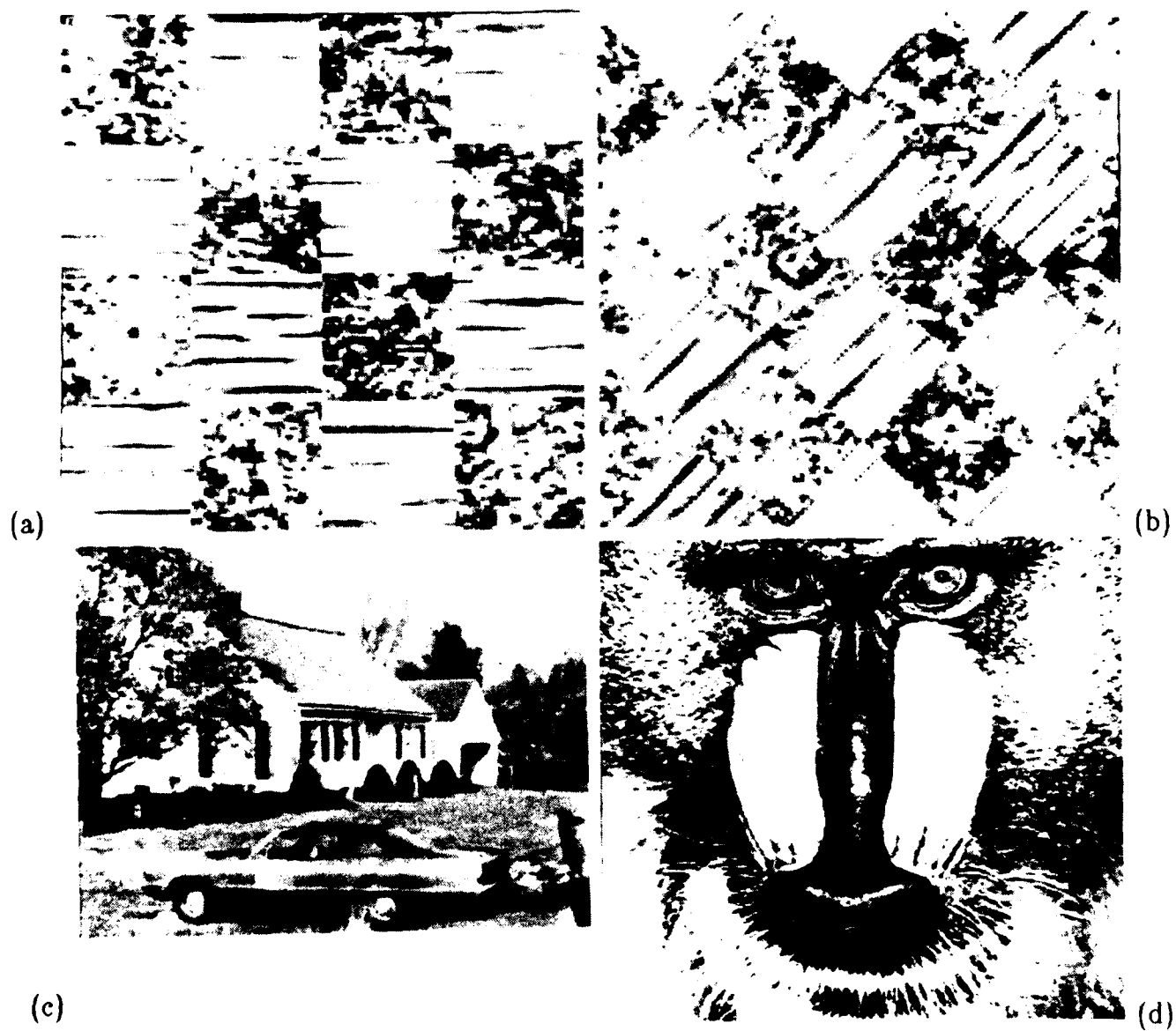


Figure 13. Original images, (a) checker board, (b) rotated checker board, (c) outdoor scene, (d) monkey

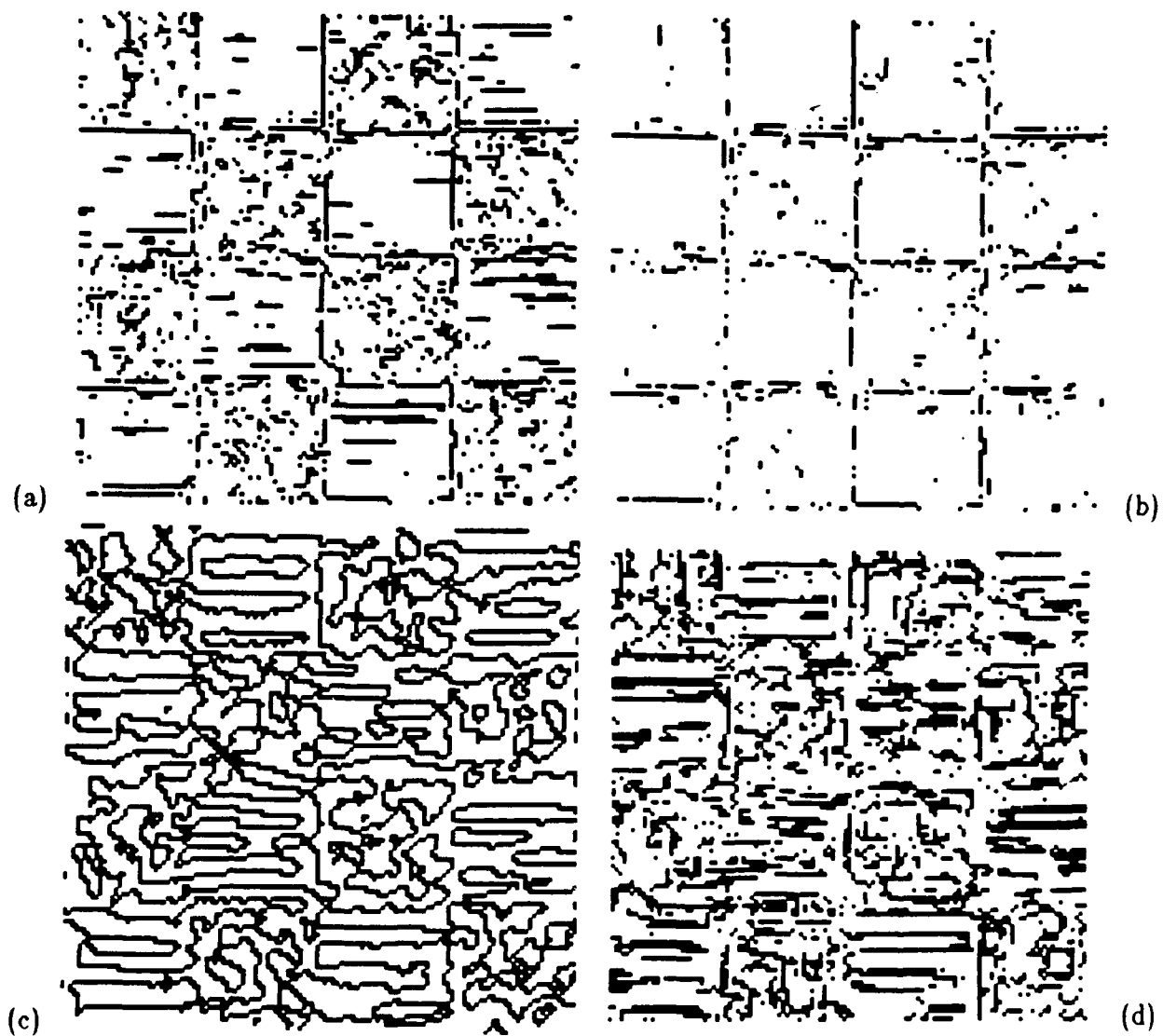


Figure 14. Comparison with checker board image, (a) edges detected by composite edge detection algorithm with low threshold, (b) edges detected by composite edge detection algorithm with high threshold, (c) result of Laplacian on Gaussian method, (d) result of facet model method

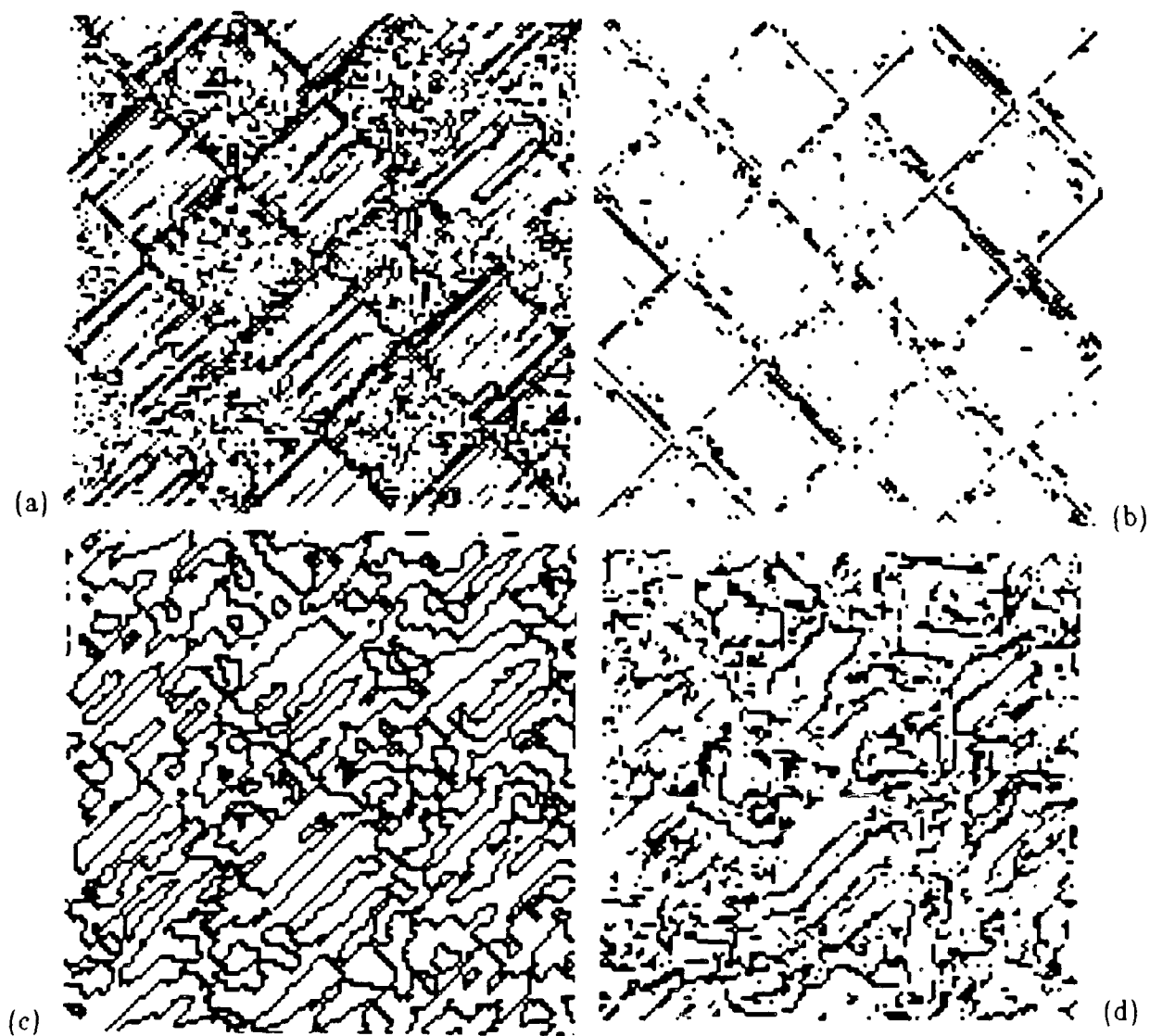


Figure 15. Comparison with rotated checker board image, (a) edges detected by composite edge detection algorithm with low threshold, (b) edges detected by composite edge detection algorithm with high threshold, (c) result of Laplacian on Gaussian method, (d) result of facet model method

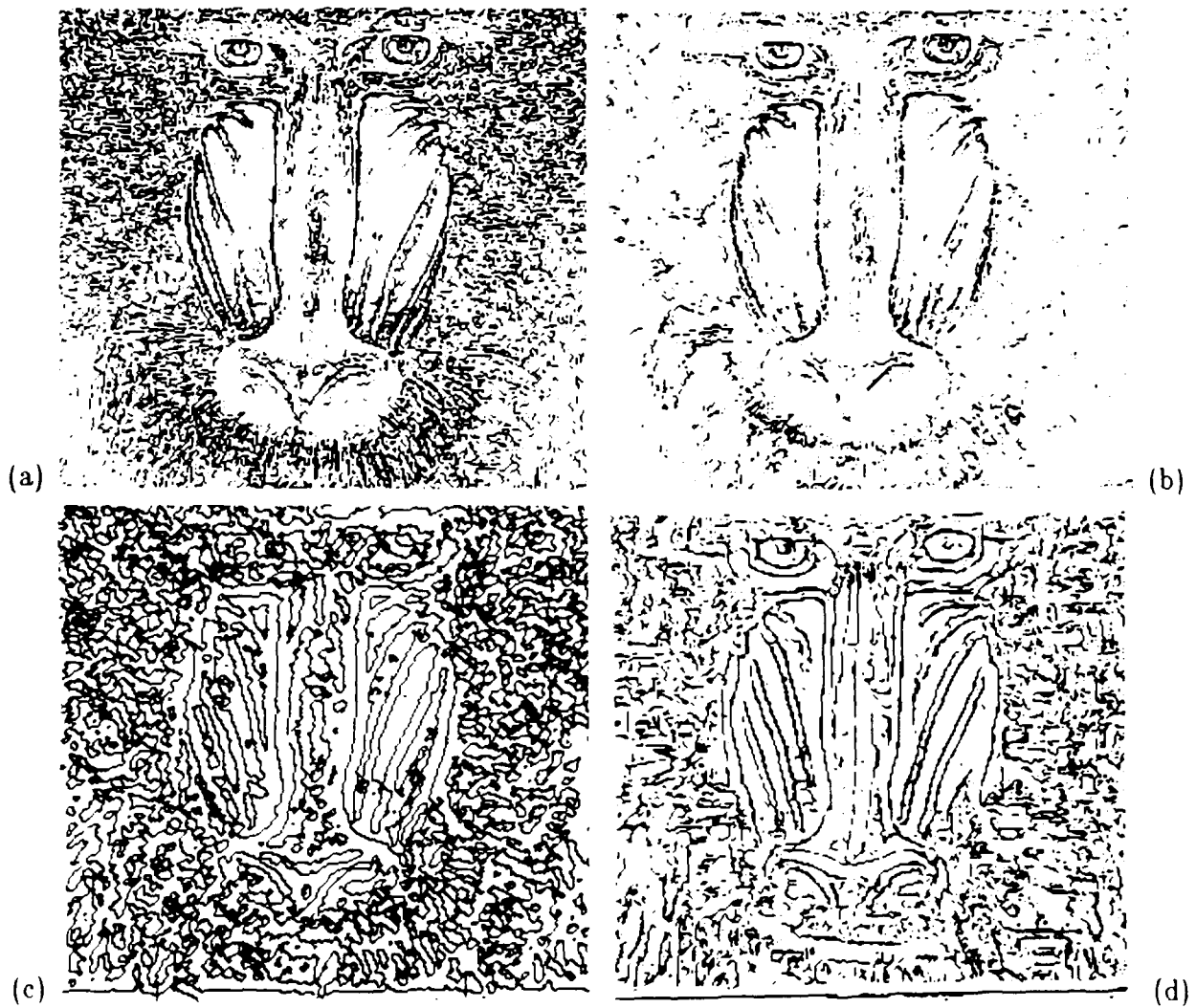


Figure 16. Comparison with monkey image, (a) potential edges (pixels having edge hypothesis) detected by algorithm 1, (b) composite edges detected by algorithm 2 (composite edge detection algorithm), (c) result of Laplacian on Gaussian method, (d) result of facet model method

VI. SUMMARY AND CONCLUSIONS. Robust image models are investigated, and applied to several important image processing problems in this study. Robust image models have potential applications in many problems arising in image processing and computer vision. Image models are already used in image synthesis, texture analysis, image coding, and image segmentation, but they are generally nonrobust to outliers. We applied the robust image models to two important problems in image processing, namely image restoration and edge detection. The robust model-based methods are compared experimentally with conventional methods. The advantage of robust model-based methods over conventional methods in some of image processing problems has been shown in Sections IV and V.

## VII. REFERENCES

- Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics*, Holden-Day Inc.
- Bovick, A.C., Huang, T.S., and Munson, D.C., Jr. (1983). "Generalization of Median Filtering Using Linear Combinations of Order Statistics," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, pp. 1342-1350.
- Bovick, A.C., Huang, T.S., and Munson, D.C., Jr. (1985). "Edge-Sensitive Image Restoration Using Order-Constrained Least Square Methods," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, pp. 1253-1263.
- Brady, M. (1982). "Computational Approaches to Image Understanding," *ACM Computing Surveys*, Vol. 14, pp. 3-71.
- Brodatz, P. (1966). *Textures: A Photographic Album for Artist and Designers*, Dover Publications, New York.
- Chan, P. and Lim, J.S. (1985). "One-Dimensional Processing for Adaptive Image Restoration," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, pp. 117-126.
- Chellappa, R. and Kashyap, R.L. (1982). "Digital Image Restoration Using Spatial Interaction Models," *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-30, pp. 461-472.
- Cross, G.R. and Jain, A.K. (1983). "Markov Random Field Texture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, pp. 25-39.
- Delp, E.J., Kashyap, R.L., and Mitchell, O.R. (1979). "Image Data Compression Using Autoregressive Time Series Models," *Pattern Recognition*, Vol. 11, pp. 313-323.
- Eom, K-B. and Kashyap, R.L. (1987). "Composite Edge Detection With Random Field Models," *Proc. Intl. Conf. Computer Vision (ICCV '87)*, London, England.
- Eom, K-B. and Kashyap, R.L. (1988). "Robust Image Modelling Techniques With an Image Restoration Application," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-36, pp. 1313-1325.
- Eom, K-B. and Kashyap, R.L. (1989a). "Texture Boundary Detection Based On Long Correlation Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-11, pp. 58-67.

- Eom, K-B. and Kashyap, R.L. (1989b). "Composite Edge Detection with Random Field Models," *IEEE Trans. Systems, Man and Cybernetics*, to appear in 1989.
- Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp. 721-741.
- Haralick, R.M., Shanmugam, K. and Dinstein, I. (1973). "Textural Features for Image Classification," *IEEE Trans. Systems, Man and Cybernetics*, Vol. SMC-3, pp. 610-621.
- Haralick, R.M. (1984). "Digital Step Edges from Zero Crossing of Second Directional Derivatives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp. 58-68.
- Huber, P.J. (1981). *Robust Statistics*, John Wiley & Sons, Inc.
- Kashyap, R.L. (1981). "Analysis and Synthesis of Image Patterns by Spatial Interaction Models," *Progress in Pattern Recognition*, Edited by L. N. Kanal & A. Rosenfeld, North-Holland, pp. 149-186.
- Kashyap, R.L. (1982). "Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-4, pp. 99-104.
- Kashyap, R.L. and Khotanzad, A. (1984). "A Stochastic Model Based Technique for Texture Segmentation," *Proc. 7th Intl. Conf. Pattern Recognition*, IEEE Computer Society Publication, Montreal, Canada.
- Kashyap, R.L. and Lapsa, P.M. (1984). "Synthesis and Estimation of Random Fields Using Long-Correlation Model," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp. 800-809.
- Kashyap, R.L. and Eom, K-B. (1985a). "Texture Boundary Detection Based on Long Correlation Model," *Proc. IEEE Intl. Geoscience and Remote Sensing Symposium*, Amhurst, MA, p. 255.
- Kashyap, R.L. and Eom, K-B. (1985b). "Texture Boundary Detection Based on Long Correlation Model," *Proc. The 23rd Annual Allerton Conf. on Control, Communication, and Computers*, pp. 314-323.
- Kashyap, R.L. and Eom, K-B. (1988). "Robust Image Models and Their Applications," *Advances in Electronics and Electron Physics*, Vol. 70, Edited by P. W. Hawkes, Academic Press, pp. 79-157.
- Kassam, S.A. and Poor, H.V. (1985). "Robust Techniques for Signal Processing: Survey," *Proceedings of IEEE*, Vol. 73, pp. 433-481.
- Marr, D. and Hildreth, E. (1980). "Theory of Edge Detection," *Proc. Royal Society, London, Series B*, Vol. 207, pp. 187-217.
- Nasburg, R.E. and Kashyap, R.L. (1975). "Robust Parameter Estimation in Dynamic System," *Proc. of Information Science and Systems Conf.*, Johns Hopkins University.
- Pratt, W.K. (1978). *Digital Image Processing*, John Wiley & Sons, Inc.
- Rey, W.J.J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*, Springer-Verlag, New York.

- Robinson, G.S. (1977). "Edge Detection By Compass Gradient Masks," *Computer Vision, Graphics, and Image Processing*, Vol. 6, pp. 492-501.
- Rosenfeld, A. and Kak, A. (1982). *Digital Picture Processing*, Second Edition, Vol. 1 & 2, Academic Press.
- Wilson, H.R. and Bergen, J.R. (1979). "A Four Mechanism Model for Threshold Spatial Vision," *Vision Research*, Vol. 19, pp. 19-32.
- Wu, Z. (1985). "Multidimensional State-Space Model Kalman Filtering with Application to Image Restoration," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, pp. 1576-1592.
- Yasuoka, Y. and Haralick, R.M. (1983). "Peak Noise Removal by a Facet Model," *Pattern Recognition*, Vol. 16, pp. 3-29.
- Zhou, Y., Chellappa, R. and Venkateswar, V. (1986). "Edge Detection Using Zero-Crossings of Directional Derivatives of a Random Field Model," *Proc. Intl. Conf. ASSP*, pp. 1465-1468.

# ON THE STROH FORMALISM FOR ANISOTROPIC ELASTICITY AND ITS APPLICATIONS TO COMPOSITES

T. C. T. Ting  
Department of Civil Engineering,  
Mechanics and Metallurgy  
University of Illinois at Chicago  
Chicago, IL 60680

**ABSTRACT.** The Stroh formalism for anisotropic elastic materials has contributed much to the determination of solutions of anisotropic elasticity problems. In most cases, however, the solutions are in a complex form and it is desirable to have the solutions in a real form for practical applications. This requires new identities or sum rules which relate eigenvalues and eigenvectors of anisotropic elastic constants to real quantities. The identities serve two important purposes. Firstly, with the identities the problem of repeated eigenvalues disappears. Secondly, the identities enable us to express the final solutions to anisotropic elasticity problems in a real form. The identities and the structural property of certain real matrices in the solution are the keys in solving heretofore unsolved problems and in simplifying existing complex solutions to a real form solution. As a result, some interesting phenomena unnoticed before have been revealed. For instance, it was discovered only recently that the surface traction on any radial plane in the anisotropic elastic material due to a concentrated force and a line dislocation applied at the origin is independent of the choice of the radial plane.

**INTRODUCTION.** Following the work of Eshelby, et al. [1], Stroh in 1958 [2] and 1962 [3] developed a powerful and elegant formalism for treating a certain class of two-dimensional problems involving dislocations, line forces and steady state waves in anisotropic elastic solids. The formalism is well-known in the physics and materials science community (see [4-10], for example). Unlike the two-dimensional anisotropic solutions developed by Green and Zerna [11] which are restricted to plane strain deformations and hence to monoclinic materials, the Stroh formalism applies to general anisotropic elastic materials for which all three displacement components are necessarily coupled. Also, unlike the Lekhnitskii's formalism [12] which breaks down for orthotropic materials [13] and requires a special treatment [14], the Stroh formalism has no restrictions. An excellent review on the Stroh formalism can be found in [8].

The basic elements of Stroh formalism are the eigenvalue  $p$  and the eigenvectors  $\xi$  of anisotropic elastic constants. The solution to an anisotropic elasticity problem is, in general, expressed in terms of  $p$ 's and  $\xi$ 's which are complex. There are identities or sum rules which express certain combinations of  $p$ 's and  $\xi$ 's in terms of real matrices  $N_i$ ,  $i=1,2,3$ ,  $S$ ,  $H$  and  $L$  to be defined later. The identities enable us to rewrite the complex solutions into a real form. The structures of  $N_i$ ,  $S$ ,  $H$ ,  $L$  tell us in depth information on the physical property of the final solution.

We outline in Section 2 Stroh formalism. The eigenvalues  $p$  and eigenvectors  $\xi$  are defined. Problems arise when  $p$  has a repeated root and the modifications required are given in Section 3. The orthogonality relations between the eigenvectors are presented in Section 4. Basic identities between  $p$ ,  $\xi$  and the real matrices  $S$ ,  $H$ ,  $L$  are derived. In



Section 5, an alternative expression for  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  due to Barnett and Lothe is presented. Also presented are the structure of  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  and  $\underline{N}_1$ ,  $\underline{N}_3$ . In the last section, we show new identities which are useful in solving certain problems in anisotropic elastic materials and composites.

**STROH FORMALISM.** In a fixed rectangular coordinate system  $x_i$ ,  $i=1,2,3$ , let  $u_i$  and  $\sigma_{ij}$  be the displacement of a particle and the stress, respectively. The equations of equilibrium and stress-strain laws can be written as

$$(1) \quad \sigma_{ij,j} = 0 ,$$

$$(2) \quad \sigma_{ij} = C_{ijks} u_{k,s} ,$$

in which repeated indices imply summation, a comma stands for partial differentiation and  $C_{ijks}$  are the elastic constants which possess the normal symmetry property

$$C_{ijks} = C_{jiks} = C_{ksij} .$$

Consider a two-dimensional deformation in which  $u_k$ ,  $k=1,2,3$ , depend on  $x_1$  and  $x_2$  only. The general solution has the form

$$(3) \quad u_k = a_k f(z) ,$$

$$(4) \quad z = x_1 + px_2 ,$$

where  $f$  is an arbitrary function of  $z$  and  $p$ ,  $a_k$  are constants. In matrix notation, they are determined by

$$(5) \quad \left\{ \underline{Q} + p(\underline{R} + \underline{R}^T) + p^2 \underline{T} \right\} \underline{a} = \underline{0} ,$$

in which the superscript  $T$  stands for the transpose and

$$(6) \quad \begin{cases} \underline{Q}_{ik} = C_{i1k1} , \\ \underline{R}_{ik} = C_{i1k2} , \\ \underline{T}_{ik} = C_{i2k2} . \end{cases}$$

Equation (5) is obtained by substituting (3) into (2) and (1). We note that  $\underline{Q}$  and  $\underline{T}$  are symmetric and, subject to positiveness of strain energy, positive definite.

Introducing the new vector

$$(7) \quad \underline{b} = (\underline{R}^T + p\underline{T})\underline{a} = -\frac{1}{p}(\underline{Q} + p\underline{R})\underline{a},$$

in which the second equality comes from (5), the stress obtained by substituting (3) into (2) can be written as

$$(8) \quad \sigma_{i1} = -\phi_{i,2}, \quad \sigma_{i2} = \phi_{i,1}, \quad \phi = \underline{b}f(z).$$

Thus  $\phi$  is the stress function.

The two equations in (7) can be rewritten in the following standard eigenrelation

$$(9) \quad \underline{N}\underline{\xi} = p\underline{\xi},$$

$$(10) \quad \begin{cases} \underline{N} = \begin{bmatrix} \underline{N}_1 & \underline{N}_2 \\ \underline{N}_3 & \underline{N}_1 \end{bmatrix}, & \underline{\xi} = \begin{bmatrix} \underline{a} \\ \underline{b} \end{bmatrix}, \\ \underline{N}_1 = -\underline{T}^{-1}\underline{R}^T, & \underline{N}_2 = \underline{T}^{-1} = \underline{N}_2^T, \\ \underline{N}_3 = \underline{R}\underline{T}^{-1}\underline{R}^T - \underline{Q} = \underline{N}_3^T. \end{cases}$$

We see that  $\underline{N}_2$  and  $\underline{N}_3$  are symmetric and  $\underline{N}_2$  is positive definite. It is shown in [15] that  $-\underline{N}_3$  is positive semi-definite. Equation (9) provides six eigenvalues  $p_a$ ,  $a=1,2,\dots,6$ , and six associated eigenvectors  $\underline{\xi}_a$ . Since  $p_a$  cannot be real if the strain energy is positive [1], we let

$$p_{a+3} = \bar{p}_a, \quad \text{Im } p_a > 0,$$

$$\underline{\xi}_{a+3} = \bar{\underline{\xi}}_a, \quad a = 1,2,3,$$

where an overbar denotes the complex conjugate and  $\text{Im}$  stands for the imaginary part. The general solution for the displacement and stress function given by (3) and (8) can be written as

$$\underline{u} = \sum_{a=1}^3 \underline{a}_a f_a(z_a) + \bar{\underline{a}}_a f_{a+3}(\bar{z}_a),$$

$$\phi = \sum_{a=1}^3 \underline{b}_a f_a(z_a) + \bar{\underline{b}}_a f_{a+3}(\bar{z}_a)$$

$$z_a = x_1 + p_a x_2,$$

where  $f_1, f_2, \dots, f_6$  are arbitrary complex functions of their arguments.

For  $\underline{u}$  and  $\underline{\phi}$  to be real, we let

$$f_{a+3} = \bar{f}_a, \quad a = 1, 2, 3,$$

and write the general solution as

$$(11) \quad \begin{cases} \underline{u} = 2 \operatorname{Re} \left\{ \sum_{a=1}^3 \underline{a}_a f_a(z_a) \right\}, \\ \underline{\phi} = 2 \operatorname{Re} \left\{ \sum_{a=1}^3 \underline{b}_a f_a(z_a) \right\}, \end{cases}$$

or

$$(12) \quad \underline{w} = 2 \operatorname{Re} \left\{ \sum_{a=1}^3 \underline{\xi}_a f_a(z_a) \right\},$$

$$\underline{w} = \begin{bmatrix} \underline{u} \\ \underline{\phi} \end{bmatrix}, \quad \underline{\xi}_a = \begin{bmatrix} \underline{a}_a \\ \underline{b}_a \end{bmatrix},$$

where  $\operatorname{Re}$  stands for the real part. We observe that  $\underline{w}$  satisfies the differential equation [8]

$$\underline{w}_{,2} = \underline{N} \underline{w}_{,1}.$$

**DEGENERATE MATERIALS.** Equations (11) or (12) are complete when the  $6 \times 6$  matrix  $\underline{N}$  in (9) is simple, i.e., when the eigenvalues  $p_a$  of  $\underline{N}$  are distinct. It remains complete when  $\underline{N}$  is semisimple, i.e., when  $p_a$  have a repeated eigenvalue and the associated eigenvectors are independent. If  $\underline{N}$  is non-semisimple, i.e., if  $p_a$  have a repeated eigenvalue but the eigenvectors  $\underline{\xi}_a$  are not all independent, the solution given by (12) is not complete. Anisotropic elastic materials for which  $\underline{N}$  is non-semisimple are called degenerate materials. Isotropic materials are a special class of degenerate materials. For isotropic materials, we have  $p_1 = p_2 = i$  and  $\underline{\xi}_1 = \underline{\xi}_2$ . In fact,  $p_3 = i$  also but  $\underline{\xi}_3 \neq \underline{\xi}_1$ . For degenerate materials for which  $\underline{\xi}_1 = \underline{\xi}_2$ , (12) is replaced by [16,17]

$$\underline{w} = 2 \operatorname{Re} \left\{ \underline{\xi}_1 f_1(z_1) + \underline{\xi}_1 \bar{z}_1 f_2(z_2) + \underline{\xi}_3 f_3(z_3) \right\},$$

in which  $\underline{\xi}_1$  satisfies the following equation which is obtained by differentiating (9) with respect to  $p$  and setting  $p = p_1$ :

$$\underline{N}\underline{\xi}'_1 = p_1\underline{\xi}'_1 + \underline{\xi}_1.$$

We see that the solution for degenerate materials destroys the regular expression of the solution given by (12) for general anisotropic materials. This happens not only for the general solution, it also occurs in applications in which the final solution has a nice simple form for general anisotropic materials but has a complicated expression for degenerate materials.

It is desirable, therefore, to have an expression which holds regardless of whether the material is degenerate or not. This means that we need the solution in a form which does not contain the eigenvalues  $p$  and the eigenvectors  $\underline{\xi}$  of the 6x6 matrix  $\underline{N}$ . We could achieve this if we have identities which relate  $p$  and  $\underline{\xi}$  to real quantities represented by  $\underline{N}$ , or by quantities derivable from  $\underline{N}$ . This is the main subject in the following sections.

**THE ORTHOGONALITY RELATIONS.** The left eigenvectors  $\underline{\eta}$  of  $\underline{N}$  is

$$\underline{\eta}^T \underline{N} = p \underline{\eta}^T,$$

or

$$(13) \quad \underline{N}^T \underline{\eta} = p \underline{\eta}.$$

The left eigenvectors  $\underline{\eta}$  and the right eigenvectors  $\underline{\xi}$  associated with different eigenvalues  $p$  are biorthogonal to each other [18]. If  $\underline{N}$  is simple, we can normalize the eigenvectors such that

$$(14) \quad \eta_a^T \xi_\beta = \delta_{a\beta},$$

where  $\delta_{a\beta}$  is the Kronecker delta. If  $\underline{N}$  is semisimple, (14) remains valid because it is possible to choose the eigenvectors associated with the repeated eigenvalue in such a way that (14) holds. If  $\underline{N}$  is non-semisimple, (14) is not valid for the repeated eigenvalue. A modified relation can be found in [8]. If we introduce the 6x6 matrices  $\underline{U}$  and  $\underline{V}$  by

$$\underline{U} = [\underline{\xi}_1, \underline{\xi}_2, \dots, \underline{\xi}_6],$$

$$\underline{V} = [\underline{\eta}_1, \underline{\eta}_2, \dots, \underline{\eta}_6],$$

(14) can be written as

$$\underline{V}^T \underline{U} = \underline{I},$$

where  $\underline{I}$  is the 6x6 identity matrix. This implies  $\underline{V}^T$  and  $\underline{U}$  are inverse of each other and hence the product commutes, i.e.,

$$(15) \quad \underline{U}\underline{V}^T = \underline{I}.$$

Denoting the 6x6 matrix  $\underline{J}$  by

$$\underline{J} = \begin{bmatrix} \underline{0} & \underline{I} \\ \underline{I} & \underline{0} \end{bmatrix},$$

$\underline{I}$  in this context being the 3x3 identity matrix, it can be shown that

$$\underline{J}\underline{N} = (\underline{J}\underline{N})^T = \underline{N}^T\underline{J}.$$

It follows from (9) and (13) that we may set without loss of generality,

$$(16) \quad \underline{\eta} = \underline{J}\underline{\xi}.$$

If we define the 3x3 matrices  $\underline{A}$  and  $\underline{B}$  by

$$\underline{A} = [a_1, a_2, a_3],$$

$$\underline{B} = [b_1, b_2, b_3],$$

we have

$$\underline{U} = \begin{bmatrix} \underline{A} & \underline{\bar{A}} \\ \underline{B} & \underline{\bar{B}} \end{bmatrix}, \quad \underline{V} = \underline{J}\underline{U}.$$

Equation (15) leads to, after carrying out the matrix multiplications,

$$(17) \quad \underline{A}\underline{A}^T + \underline{\bar{A}}\underline{\bar{A}}^T = \underline{0} = \underline{B}\underline{B}^T + \underline{\bar{B}}\underline{\bar{B}}^T,$$

$$(18) \quad \underline{A}\underline{B}^T + \underline{\bar{A}}\underline{\bar{B}}^T = \underline{I} = \underline{B}\underline{A}^T + \underline{\bar{B}}\underline{\bar{A}}^T.$$

Equations (17) imply that  $\underline{A}\underline{A}^T$  and  $\underline{B}\underline{B}^T$  are purely imaginary while (18) tells us that the real part of  $\underline{A}\underline{B}^T$  is  $\underline{I}/2$ . Hence we let

$$(19) \quad \begin{cases} \underline{S} = i(2\underline{A}\underline{B}^T - \underline{I}), \\ \underline{H} = 2i\underline{A}\underline{A}^T = \underline{H}^T, \\ \underline{L} = -2i\underline{B}\underline{B}^T = \underline{L}^T, \end{cases}$$

where  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  are 3x3 real matrices first introduced by Barnett and Lothe [6].  $\underline{H}$  and  $\underline{L}$  are symmetric and can be shown to be positive definite [8,15,19]. Noting that (19) can be written as

$$(20) \quad \begin{bmatrix} \underline{S} & \underline{H} \\ -\underline{L} & \underline{S}^T \end{bmatrix} = 2i \begin{bmatrix} \underline{A} \\ \underline{B} \end{bmatrix} \begin{bmatrix} \underline{B}^T & \underline{A}^T \end{bmatrix} - i\underline{I},$$

and using the following relation which is deduced from (14) and (16),

$$\begin{bmatrix} \underline{B}^T & \underline{A}^T \end{bmatrix} \begin{bmatrix} \underline{A} \\ \underline{B} \end{bmatrix} = \underline{I},$$

we have [8]

$$\begin{bmatrix} \underline{S} & \underline{H} \\ -\underline{L} & \underline{S}^T \end{bmatrix} \begin{bmatrix} \underline{S} & \underline{H} \\ -\underline{L} & \underline{S}^T \end{bmatrix} = -\underline{I}.$$

This leads to

$$(21) \quad \begin{cases} \underline{H}\underline{L} - \underline{S}\underline{S} = \underline{I}, \\ \underline{S}\underline{H} + \underline{H}\underline{S}^T = \underline{0}, \\ \underline{L}\underline{S} + \underline{S}^T\underline{L} = \underline{0}. \end{cases}$$

We see that  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  are not independent of each other. We also see from (21)<sub>2,3</sub> and their counterparts

$$\underline{H}^{-1}\underline{S} + \underline{S}^T\underline{H}^{-1} = \underline{0},$$

$$\underline{S}\underline{L}^{-1} + \underline{L}^{-1}\underline{S}^T = \underline{0},$$

that  $\underline{S}\underline{H}$ ,  $\underline{L}\underline{S}$ ,  $\underline{H}^{-1}\underline{S}$  and  $\underline{S}\underline{L}^{-1}$  are antisymmetric.

**THE STRUCTURE OF  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  AND  $\underline{N}_1$ ,  $\underline{N}_3$ .** The three real matrices  $\underline{S}$ ,  $\underline{H}$ ,

$\underline{L}$ , which are the Barnett-Lothe tensors, appear very often in the solutions to anisotropic elasticity problems (see [7,8,20,21], for example). The expressions given by (19) are based on the assumption that the eigenvectors  $\underline{\xi}$  span a six-dimensional space. When  $\underline{N}$  is non-semisimple, we do not have six independent eigenvectors and (19) are not valid. In

fact, one encounters problems also when  $\underline{N}$  is almost non-semisimple. A modified expression in place of (19) when  $\underline{N}$  is non-semisimple or almost non-semisimple was presented in [22]. The modified expression applies to simple  $\underline{N}$  as well.

An alternate approach which avoids the determination of eigenvectors is the integral formalism introduced by Barnett and Lothe. We generalize the matrices  $\underline{Q}$ ,  $\underline{R}$ ,  $\underline{T}$  defined in (6) to

$$(22) \quad \begin{cases} Q_{ik}(\theta) = C_{ijks} n_j n_s, \\ R_{ik}(\theta) = C_{ijks} n_j m_s, \\ T_{ik}(\theta) = C_{ijks} m_j m_s, \end{cases}$$

in which  $\theta$  is a real parameter and

$$\begin{aligned} n_i &= (\cos\theta, \sin\theta, 0), \\ m_i &= (-\sin\theta, \cos\theta, 0). \end{aligned}$$

When  $\theta = 0$ , (22) reduce to (6). With  $\underline{Q}$ ,  $\underline{R}$ ,  $\underline{T}$ , defined by (22), the three 3x3 matrices  $\underline{N}_i$  and the 6x6 matrix  $\underline{N}$  of (10) also depend on  $\theta$ . Equation (9) now becomes

$$(23) \quad \underline{N}(\theta)\underline{\xi} = p(\theta)\underline{\xi}.$$

It can be shown that when  $p_a(\theta)$  are distinct,  $\underline{\xi}_a$  are independent of  $\theta$ . It can also be shown that  $p(\theta)$  is related to  $p(0) = p$  of (9) by [8,17]

$$(24) \quad \begin{aligned} p(\theta) &= \frac{p \cos\theta - \sin\theta}{p \sin\theta + \cos\theta} \\ &= \frac{d}{d\theta} \{ \ln (\cos\theta + p \sin\theta) \} \end{aligned}$$

We now consider the integrals

$$(25) \quad \begin{cases} \hat{S}(\theta) = \frac{1}{r} \int_0^\theta \underline{N}_1(\omega) d\omega, & \hat{H}(\theta) = \frac{1}{r} \int_0^\theta \underline{N}_2(\omega) d\omega, \\ \hat{L}(\theta) = -\frac{1}{r} \int_0^\theta \underline{N}_3(\omega) d\omega. \end{cases}$$

When  $\theta = \pi$ , the integrals in (25) are called complete integrals. Barnett and Lothe [6] proved that  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  of (19) are identical to the complete integrals

$$(26) \quad \underline{S} = \hat{\underline{S}}(\pi), \quad \underline{H} = \hat{\underline{H}}(\pi), \quad \underline{L} = \hat{\underline{L}}(\pi).$$

This provides an alternate to the determination of the three real matrices  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$ . In (26) the need of determining the eigenvectors are circumvented and hence the problem of repeated eigenvalue disappears.

Equations (25) can be integrated explicitly for isotropic materials. For  $\theta = \pi$ , we have

$$(27) \quad \left\{ \begin{array}{l} \underline{S} = \begin{bmatrix} 0 & -s & 0 \\ s & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \underline{H} = \frac{1}{\mu} \begin{bmatrix} (1-s^2)/\gamma & 0 & 0 \\ 0 & (1-s^2)/\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \underline{L} = \mu \begin{bmatrix} \gamma & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}, \end{array} \right.$$

$$s = \frac{1-2\nu}{2(1-\nu)}, \quad \gamma = \frac{1}{1-\nu},$$

in which  $\mu$  and  $\nu$  are the shear modulus and Poisson's ratio, respectively. Complete integrals of (25) for transversely isotropic materials can be found in [8] but that for more general anisotropic materials have not been available.

For general anisotropic materials, Chadwick and Ting [23] have shown that  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  have the same structure as (27) for isotropic materials if a proper basis and proper tensor components are chosen for the tensors  $\underline{S}$ ,  $\underline{H}$  and  $\underline{L}$ . They showed that the eigenvalues of  $\underline{S}$  are 0,  $\pm is$  where  $s$  is real and positive. Let the associated eigenvectors be  $\underline{e}_3$ ,  $\underline{e}_1 \mp i\underline{e}_2$  where  $\underline{e}_i$ ,  $i=1,2,3$  are real vectors and let the reciprocal eigenvectors  $\underline{e}^i$  be defined by

$$\underline{e}_i \cdot \underline{e}^j = \delta_{ij}.$$

If we choose the following tensor components for  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$ ,

$$\underline{S} = S_{ij}^i \underline{e}_i \underline{e}^j,$$

$$\underline{H} = H^{ij} \underline{e}_i \underline{e}_j,$$

$$\underline{L} = L_{ij}^i \underline{e}^i \underline{e}^j,$$



it can be shown that

$$(28) \quad \begin{cases} S_{ij} = \begin{bmatrix} 0 & -s & 0 \\ s & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & H^{ij} = \frac{1}{\mu} = \begin{bmatrix} (1-s^2)/\gamma & 0 & 0 \\ 0 & (1-s^2)/\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ L_{ij} = \mu \begin{bmatrix} \gamma & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}, \end{cases}$$

where  $\mu, \gamma, s$  are constants. The identical structure between this and (27) is striking. It should be noted that (28) for general anisotropic materials has only three constants  $0 < s < 1, \mu > 0$  and  $\gamma > 0$ .

The three matrices  $\underline{N}_i$  defined in (10) depend on the elastic constants in a complicated way, particularly so for  $\underline{N}_1$  and  $\underline{N}_3$ . It is shown in [15] that  $-\underline{N}_3$  is positive semi-definite and  $\underline{N}_1, \underline{N}_3$  have the structure

$$\underline{N}_1 = \begin{bmatrix} * & -1 & * \\ * & 0 & * \\ * & 0 & * \end{bmatrix}, \quad \underline{N}_3 = \begin{bmatrix} * & 0 & * \\ 0 & 0 & 0 \\ * & 0 & * \end{bmatrix},$$

in which the \* elements can be expressed in terms of the elastic compliances which are the reciprocal of  $C_{ijks}$ . The fact that  $\underline{N}_3$  has the property shown above was crucial in solving the problem of the elastic wedge subject to uniform tractions on the sides of the wedge [24]. Clearly, the property of  $\underline{N}_1$  and  $\underline{N}_3$  will be useful also in solving other anisotropic elasticity problems.

While  $\underline{N}_1(\theta), \underline{N}_3(\theta)$  do not have the same structure as  $\underline{N}_1, \underline{N}_3$  except at  $\theta = 0$ , if we write

$$\begin{aligned} \underline{N}_i^*(\theta) &= \underline{\Omega}(\theta) \underline{N}_i(\theta) \underline{\Omega}^T(\theta), \\ \underline{\Omega}^T(\theta) &= [\underline{n}(\theta), \underline{m}(\theta), \underline{e}_3], \quad \underline{e}_3^T = (0, 0, 1), \end{aligned}$$

$\underline{N}_1^*(\theta)$  and  $\underline{N}_3^*(\theta)$  have the same structure as  $\underline{N}_1$  and  $\underline{N}_3$ . This is not surprising because  $\underline{N}_i^*(\theta)$  are  $\underline{N}_i$  referred to the rotated coordinate system  $\underline{x}^* = \underline{\Omega} \underline{x}$  [25]. It is clear that  $-\underline{N}_3^*(\theta)$  as well as  $-\underline{N}_3(\theta)$  are also positive semi-definite.

**NEW IDENTITIES.** In many applications, the arbitrary functions  $f_a(z_a)$  in (11) or (12) assume the same function form for all  $a$ . The simplest ones are power of  $z_a$ , i.e.,

$$f_a(z_a) = q_a z_a^\lambda, \quad (a \text{ not summed}),$$

where  $\lambda$  and  $q_a$ ,  $a=1,2,3$ , are arbitrary complex constants. If we define the diagonal matrix

$$\underline{Z}^\lambda = \text{diag}(z_1^\lambda, z_2^\lambda, z_3^\lambda),$$

we may write (11) as

$$\underline{u} = 2 \text{Re} \left\{ \underline{AZ}^\lambda \underline{q} \right\},$$

$$\underline{\phi} = 2 \text{Re} \left\{ \underline{BZ}^\lambda \underline{q} \right\},$$

in which the elements of  $\underline{q}$  are  $q_1, q_2, q_3$ . Replacing the complex constant  $\underline{q}$  by two real constants  $\underline{g}$  and  $\underline{h}$  through

$$\underline{q} = \underline{A}^T \underline{g} + \underline{B}^T \underline{h},$$

we have

$$(29) \quad \begin{cases} \underline{u} = 2 \text{Re} \left\{ \underline{AZ}^\lambda \underline{B}^T \right\} \underline{h} + 2 \text{Re} \left\{ \underline{AZ}^\lambda \underline{A}^T \right\} \underline{g}, \\ \underline{\phi} = 2 \text{Re} \left\{ \underline{BZ}^\lambda \underline{B}^T \right\} \underline{h} + 2 \text{Re} \left\{ \underline{BZ}^\lambda \underline{A}^T \right\} \underline{g}. \end{cases}$$

This form of solution can be used for analyzing stress singularities in a composite. In [21], the order of stress singularities  $\lambda$  at an interface crack was obtained in closed form for general anisotropic elastic materials. With  $\lambda$  obtained explicitly in closed form, one can look at the imaginary part of  $\lambda$  and study under what combination of materials the oscillations in displacement near the crack tip disappears.

When  $\lambda$  is an integer, positive or negative, the quantities in the brackets in (29) can be expressed explicitly in real form. Using (4) and (9), we have

$$z_\pm^\lambda = (x_1 + px_2)\xi_\pm = (x_1 \underline{I} + x_2 \underline{N})\xi_\pm,$$

or

$$\begin{bmatrix} \underline{AZ}^\lambda \\ \underline{BZ}^\lambda \end{bmatrix} = (x_1 \underline{I} + x_2 \underline{N})^\lambda \begin{bmatrix} \underline{A} \\ \underline{B} \end{bmatrix}.$$

If we post-multiply both sides by  $[\underline{B}^T \ \underline{A}^T]$  and use (20), we have the identity

$$(30) \quad 2 \text{Re} \begin{bmatrix} \underline{AZ}^\lambda \underline{B}^T & \underline{AZ}^\lambda \underline{A}^T \\ \underline{BZ}^\lambda \underline{B}^T & \underline{BZ}^\lambda \underline{A}^T \end{bmatrix} = (x_1 \underline{I} + x_2 \underline{N})^\lambda,$$

which provides a real expression for the quantities in (29) without determining the eigenvalues  $p$  and the eigenvectors  $\xi$ .

Although (30) applies also for negative integer  $\lambda$ , the right-hand side of (30) is not a very useful form since it requires an inverse of 6x6 matrix. However, if we use the polar coordinate system

$$x_1 = r \cos \theta, \quad x_2 = r \sin \theta$$

and employ the following identity proved in [25],

$$\{\cos \theta \underline{I} + \sin \theta \underline{N}\}^{-1} = \{\cos \theta - \sin \theta \underline{N}(\theta)\}.$$

the right-hand side of (30) becomes

$$r^\lambda \{\cos \theta - \sin \theta \underline{N}(\theta)\}^{-\lambda},$$

which is a useful form for a negative integer  $\lambda$ .

For the wedge problems [24,25],  $\lambda = 1$  is used for uniform tractions applied on the sides of the wedge while  $\lambda = -1$  is used for a concentrated couple applied at the wedge apex.

Another function form for  $f(z)$  which appears in the problem of a concentrated force and a line dislocation in anisotropic elastic materials is

$$f_a(z_a) = q_a \ln z_a, \quad (a \text{ not summed}).$$

Defining the diagonal matrix

$$\ln \underline{Z} = \text{diag} [\ln z_1, \ln z_2, \ln z_3],$$

we have

$$(31) \quad \begin{cases} \underline{u} = 2 \text{Re} \left\{ \underline{A}(\ln \underline{Z}) \underline{B}^T \right\} \underline{h} + 2 \text{Re} \left\{ \underline{A}(\ln \underline{Z}) \underline{A}^T \right\} \underline{g}, \\ \underline{\phi} = 2 \text{Re} \left\{ \underline{B}(\ln \underline{Z}) \underline{B}^T \right\} \underline{h} + 2 \text{Re} \left\{ \underline{B}(\ln \underline{Z}) \underline{A}^T \right\} \underline{g}. \end{cases}$$

To find the real expression for the quantities in brackets, we first notice that

$$z = r(\cos \theta + p \sin \theta),$$

and, using (24),

$$\begin{aligned} \ln z &= \ln r + \ln (\cos \theta + p \sin \theta) \\ &= \ln r + \int_0^\theta p(\omega) d\omega. \end{aligned}$$

Next, from (23) we have

$$(\ln z)\underline{\xi} = \left\{ (\ln r)\underline{I} + \tau\tilde{N}(\theta) \right\} \underline{\xi},$$

or

$$(32) \quad \begin{bmatrix} \underline{A}(\ln Z) \\ \underline{B}(\ln Z) \end{bmatrix} = \left\{ (\ln r)\underline{I} + \tau\tilde{N}(\theta) \right\} \begin{bmatrix} \underline{A} \\ \underline{B} \end{bmatrix},$$

where, following (25),

$$(33) \quad \tilde{N}(\theta) = \frac{1}{\tau} \int_0^\theta \underline{N}(\omega) d\omega = \begin{bmatrix} \underline{\hat{S}}(\theta) & \underline{\hat{H}}(\theta) \\ -\underline{\hat{L}}(\theta) & \underline{\hat{S}}^T(\theta) \end{bmatrix}.$$

Finally, we post-multiply both sides of (32) by  $[\underline{B}^T \ \underline{A}^T]$  and use (20) to obtain the identity

$$(34) \quad 2 \operatorname{Re} \begin{bmatrix} \underline{A}(\ln Z)\underline{B}^T & \underline{A}(\ln Z)\underline{A}^T \\ \underline{B}(\ln Z)\underline{B}^T & \underline{B}(\ln Z)\underline{A}^T \end{bmatrix} = \left\{ (\ln r)\underline{I} + \tau\tilde{N}(\theta) \right\}.$$

With (34), (31) can be written as

$$(35) \quad \begin{cases} \underline{u} = \left\{ (\ln r)\underline{I} + \tau\tilde{S}(\theta) \right\} \underline{h} + \tau\tilde{H}(\theta)\underline{g}, \\ \underline{\phi} = -\tau\tilde{L}(\theta)\underline{h} + \left\{ (\ln r)\underline{I} + \tau\tilde{S}^T(\theta) \right\} \underline{g}. \end{cases}$$

The surface traction  $\underline{t}_\theta$  on any radial plane  $\theta = \text{constant}$  is determined by differentiating  $\underline{\phi}$  with respect to  $r$ . We have

$$(36) \quad \underline{t}_\theta = r^{-1}\underline{g},$$

which is independent of  $\theta$ . In [26], (36) is derived from equations of equilibrium without employing the stress-strain laws. Therefore, (35) applies also to composite spaces [27] and to angularly inhomogeneous anisotropic materials [26,28]. It should be pointed out that, although  $\underline{\phi}$  in (35) is not valid for angularly inhomogeneous materials,  $\underline{u}$  in (35) remains valid in such materials. The only modification required is in (33) where the integrand  $\underline{N}(\theta)$  contains  $C_{ijks}$  which depend on  $\theta$ .

**CONCLUDING REMARKS.** The Stroh formalism is elegant and powerful. The formalism is also very effective in treating the surface waves [3,8,29], Stoneley waves [30,31] and waves in layered composites [32]. The real matrices  $\underline{N}_i(\theta)$ , the incomplete

integrals  $\hat{S}(\theta)$ ,  $\hat{H}(\theta)$ ,  $\hat{L}(\theta)$  and the complete integrals  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$ , which are the Barnett-Lothe tensors, appear often in the solutions to anisotropic elasticity problems. The striking simplicity in the structure of  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  for general anisotropic elastic materials as shown in (28) is puzzling. It is believed that  $\hat{S}(\theta)$ ,  $\hat{H}(\theta)$ ,  $\hat{L}(\theta)$ , as well as  $\underline{S}$ ,  $\underline{H}$ ,  $\underline{L}$  have physical interpretations. For  $\hat{L}(\theta)$ , it is shown in [24] that if the stress in the anisotropic elastic material depends on  $\theta$  only, the stress tensor is, with the exception of the  $\sigma_{33}$  component, proportional to  $\hat{L}(\theta)$ .

**ACKNOWLEDGEMENTS.** The work presented here is supported by the U.S. Army Research Office through grant DAAL 03-88-K-0079 with the University of Illinois at Chicago.

#### **REFERENCES.**

- [1] Eshelby, J. D., Read, W. T. and Shockley, W., "Anisotropic Elasticity with Applications to Dislocation Theory," *Acta Metal.*, 1, 251-259, 1953.
- [2] Stroh, A. N., "Dislocations and Cracks in Anisotropic Elasticity," *Phil. Mag.*, 7, 625-646, 1958.
- [3] Stroh, A. N., "Steady State Problems in Anisotropic Elasticity," *J. Math. Phys.*, 41, 77-103, 1962.
- [4] Ingebrigtsen, K. A. and Tonning, A., "Elastic Surface Waves in Crystals," *Phys. Rev.*, 184, No. 3, 942-951, 1969.
- [5] Malen, K., "A Unified Six-Dimensional Treatment of Elastic Green's Functions and Dislocations," *Phys. Status Solidi B*, 44, 661-672, 1971.
- [6] Barnett, D. M. and Lothe, J., "Synthesis of the Sextic and the Integral Formalism for Dislocations, Greens Functions and Surface Waves in Anisotropic Elastic Solids," *Phys. Norv.*, 7, 13-19, 1973.
- [7] Asaro, R. J., Hirth, J. P., Barnett, D. M. and Lothe, J., "A Further Synthesis of Sextic and Integral Theories for Dislocations and Line Forces in Anisotropic Media," *Phys. Status Solidi B*, 60, 261-271, 1973.
- [8] Chadwick, P. and Smith, G. D., "Foundations of the Theory of Surface Waves in Anisotropic Elastic Materials," *Adv. Appl. Mech.*, 17, 303-376, 1977.
- [9] Bacon, D. J., Barnett, D. M. and Scattergood, R. O., "The Anisotropic Continuum Theory of Lattice Defects," *Progress in Materials Science*, 23, 51-262, 1978.
- [10] Kirchner, H. O. K. and Lothe, J., "On the Redundancy of the N Matrix of Anisotropic Elasticity," *Phil. Mag., A.*, 53, L7-L10, 1986.
- [11] Green, A. E. and Zerna, W., *Theoretical Elasticity*, The Clarendon Press, Oxford, Chap. 6, 1954.

- [12] Lekhnitskii, S. G., Theory of Elasticity of an Anisotropic Body, MIR Publishers, Moscow, 1981.
- [13] Ting, T. C. T. and Chou, S. C., "Stress Singularities in Laminated Composites," Proc. Second USA-USSR Symposium on Fracture of Composite Materials, G. Sih and V. Tamuzs, Editors, Noordhoff, 265-277, 1982.
- [14] Wang, S. S. and Choi, I., "Boundary-Layer Effects in Composite Laminates: Part I - Free-Edge Stress Singularities," J. Appl. Mech., 49, 541-548, 1982.
- [15] Ting, T. C. T., "Some Identities and Structure of  $N_1$  in the Stroh Formalism of Anisotropic Elasticity," Q. Appl. Math., 46, 109-120, 1988.
- [16] Ting, T. C. T. and Chou, S. C., "Edge Singularities in Anisotropic Composites," Int. J. Solids Structures, 17, 1057-1068, 1981.
- [17] Ting, T. C. T., "Effects of Change of Reference Coordinates on the Stress Analyses of Anisotropic Elastic Materials," Int. J. Solids Structures, 18, 139-152, 1982.
- [18] Hilderbrand, F. B., Method of Applied Mathematics, Prentice-Hall, 1954.
- [19] Gundersen, S. A., Barnett, D. M. and Lothe, J., "Rayleigh Wave Existence Theory. A Supplementary Remark," Wave Motion, , 319-321, 1987.
- [20] Barnett, D. M. and Lothe, J., "Line Force Loadings on Anisotropic Half-Spaces and Wedges," Phys. Norv., 8, 13-22, 1975.
- [21] Ting, T. C. T., "Explicit Solution and Invariance of the Singularities at an Interface Crack in Anisotropic Composites," Int. J. Solids Structures, 22, 965-983, 1986.
- [22] Ting, T. C. T. and Hwu, C., "Sextic Formalism in Anisotropic Elasticity for Almost Non-Semisimple Matrix  $N$ ," Int. J. Solids Structures, 24, 65-76, 1988.
- [23] Chadwick, P. and Ting, T. C. T., "On the Structure and Invariance of the Barnett-Lothe Tensors," Q. Appl. Math., 45, 419-427, 1987.
- [24] Ting, T. C. T., "The Critical Angle of the Anisotropic Elastic Wedge Subject to Uniform Traction," J. Elasticity, 20, 113-130, 1988.
- [25] Ting, T. C. T., "The Anisotropic Elastic Wedge Under a Concentrated Couple," Q. J. Mech. Appl. Math., 41, 563-578, 1988.
- [26] Ting, T. C. T., "Line Forces and Dislocations in Angularly Inhomogeneous Anisotropic Elastic Wedges and Spaces," Q. Appl. Math., 47, 123-128, 1989.
- [27] Ting, T. C. T., "Line Forces and Dislocations in Anisotropic Elastic Composite Wedges and Spaces," Physica Status Solidi B, 146, 81-90, 1988.
- [28] Kirchner, H. O. K., "Line Defects Along the Axis of Rotationally Inhomogeneous Media," Phil. Mag. A, 55, 537-542, 1987.

- [29] Barnett, D. M. and Lothe, J., "Free Surface (Rayleigh) Waves in Anisotropic Elastic Half-Spaces: The Surface Impedance Method," Proc. Royal Soc. Lond., A, 402, 135-152, 1985.
- [30] Chadwick, P. and Currie, P. K., "Stoneley Waves at an Interface Between Elastic Crystals," Q. J. Mech. Appl. Math., 27, 497-503, 1974.
- [31] Barnett, D. M., Lothe, J., Gavazza, S. D. and Musgrave, M. J. P., "Considerations of the Existence of Interfacial (Stoneley) Waves in Bonded Anisotropic Elastic Half-Spaces," Proc. Roy. Soc. Lond., A, 402, 153-166, 1985.
- [32] Ting, T. C. T. and Chadwick, P., "Harmonic Waves in Periodically Layered Anisotropic Elastic Composites," ASME Symposium on Wave Propagation in Structural Composites, AMD-Vol. 90, G00426, 69-79, 1988.

# TOTAL ABSORPTION IN ELASTIC MEDIA\*

William W. Hager  
Department of Mathematics  
University of Florida  
Gainesville, FL 32611 USA

and

Rouben Rostamian  
Department of Mathematics  
University of Maryland Baltimore County  
Catonsville, MD 21228 USA

**ABSTRACT.** We examine the problem of designing a homogeneous, isotropic elastic slab that totally absorbs an incident plane wave.

**INTRODUCTION.** In [1] we present a systematic method to analyze the interaction of steady-state, harmonic plane waves with a stratified elastic media. In this paper we apply our analysis to design a homogeneous, isotropic elastic slab that totally absorbs an incident plane wave propagating through an adjacent fluid half-space. To begin, let us consider a homogeneous, isotropic elastic solid half-space in contact with a fluid half-space. When a harmonic plane wave travels through the fluid and strikes the solid-fluid interface, the propagation directions of the reflected and refracted waves are determined by Snell's Laws, while the amplitudes of the reflected and refracted waves are determined from the continuity of displacements and tractions at the interface. For a solid-fluid interface, we have no control over the reflected and refracted waves; the outcomes are governed by the fundamental laws of physics. But when an elastic slab is inserted between the fluid and the solid half-spaces, we show that the mechanical properties of the slab can be chosen so that the amplitude of the reflected wave vanishes. The choice of the mechanical properties depends on the frequency and the angle of incidence for the incoming wave.

---

\*This research was supported by U.S. Army Research Office Contract DAAL03-89-G-0082.



**NORMAL INCIDENCE.** In this section, we consider the case where the incident wave is normal to the fluid-slab interface. This particular case can be treated directly, without reference to our earlier analysis. On the other hand, the results provide insight into the analysis of oblique incidence. Suppose that the  $x$ -axis is oriented perpendicular to the slab and the slab occupies the region  $0 \leq x \leq T$ , where  $x = T$  is the slab-fluid interface. The equation of motion is

$$\rho(x) \frac{\partial^2 v}{\partial t^2} = \frac{\partial}{\partial x} \left( \kappa(x) \frac{\partial v}{\partial x} \right),$$

where  $v = v(x, t)$  is the displacement at position  $x$  and at time  $t$ . We assume that

$$\begin{aligned} \kappa(x) &= \kappa_1 \quad \text{and} \quad \rho(x) = \rho_1 \quad \text{for} \quad x > T, \\ \kappa(x) &= \kappa \quad \text{and} \quad \rho(x) = \rho \quad \text{for} \quad 0 \leq x \leq T, \\ \kappa(x) &= \kappa_0 \quad \text{and} \quad \rho(x) = \rho_0 \quad \text{for} \quad x < 0. \end{aligned}$$

Assuming harmonic time dependence and a unit amplitude for the incident wave, the general solution of the equation of motion has the form  $v(x, t) = u(x)e^{i\omega t}$  where  $\omega$  is the wave frequency,

$$\begin{aligned} u(x) &= e^{i\omega s_1(x-T)} + r e^{-i\omega s_1(x-T)} \quad \text{for} \quad x > T, \\ u(x) &= \tau_+ e^{i\omega s x} + \tau_- e^{-i\omega s x} \quad \text{for} \quad 0 \leq x \leq T, \\ u(x) &= \tau e^{i\omega s_0 x} \quad \text{for} \quad x < 0. \end{aligned}$$

Here the slowness parameters  $s_1$ ,  $s$ , and  $s_0$  are defined by

$$s_1 = \sqrt{\rho_1/\kappa_1}, \quad s = \sqrt{\rho/\kappa}, \quad \text{and} \quad s_0 = \sqrt{\rho_0/\kappa_0}.$$

(The slowness is the reciprocal of the wave speed.) The amplitudes  $r$ ,  $\tau$ ,  $\tau_+$ , and  $\tau_-$  can be determined from the continuity of displacement  $v$  and stress  $\kappa \partial v / \partial x$  at the interfaces  $x = 0$  and  $x = T$ . Altogether, there are four equations of continuity:

$$\begin{aligned}
\tau &= \tau_+ + \tau_- \\
1 + r &= \tau_+ e^{i\omega s T} + \tau_- e^{-i\omega s T} \\
s_0 \kappa_0 \tau &= \kappa s (\tau_+ - \tau_-) \\
s_1 \kappa_1 (1 - r) &= \kappa s (\tau_+ e^{i\omega s T} - \tau_- e^{-i\omega s T})
\end{aligned}$$

Solving these equations for  $r$  and setting  $r = 0$  yields the relation

$$(1) \quad \frac{\sigma_1 - \sigma}{\sigma_1 + \sigma} = \frac{\sigma_0 - \sigma}{\sigma_0 + \sigma} e^{-2i\omega s T},$$

where

$$\sigma_1 = \sqrt{\kappa_1 \rho_1}, \quad \sigma = \sqrt{\kappa \rho}, \quad \text{and} \quad \sigma_0 = \sqrt{\kappa_0 \rho_0}.$$

Since the mechanical parameters are all real, equation (1) only holds when the exponential term is +1 or -1. Hence, there are two cases to consider:

Case 1.  $e^{-2i\omega s T} = -1$ .

In this case, the exponent  $2\omega s T$  is an odd multiple of  $\pi$ . In other words,  $\omega s T = (m + 1/2)\pi$  for some integer  $m$ , or, equivalently,

$$(2) \quad \omega T \sqrt{\rho/\kappa} = (m + 1/2)\pi.$$

Substituting -1 for the exponential term in (1) gives  $\sigma^2 = \sigma_0 \sigma_1$ , or, equivalently,

$$(3) \quad \kappa \rho = \sqrt{\kappa_0 \rho_0} \sqrt{\kappa_1 \rho_1}.$$

Thus the impedance  $\sqrt{\kappa \rho}$  of the slab is the geometric mean of the impedances of the half-spaces it separates. Together, equations (2) and (3) determine the ratio  $\rho/\kappa$  and the product  $\rho \kappa$ . Therefore, they determine a unique  $\rho$  and  $\kappa$  for each choice of the integer  $m$  in (2).

Case 2.  $e^{-2i\omega sT} = +1$ .

In this case, the exponent  $2\omega sT$  is an even multiple of  $\pi$ , which implies that

$$(4) \quad \omega sT = \omega T \sqrt{\rho/\kappa} = m\pi$$

for some integer  $m$ . Again, this equation restricts the slowness to a countable set of discrete values. However, when we substitute  $+1$  for the exponential term in (1), we see that  $\sigma_0 = \sigma_1$ . That is, this case occurs only when the materials in the two half-spaces have the same impedance. On the other hand, if the impedances match, then for each integer  $m$ , there is a 1-parameter family of slab materials, with slowness given by (4), that totally absorbs the incoming wave.

OBLIQUE INCIDENCE. Now let us consider a plane wave that strikes the solid-fluid interface at an oblique angle, generating reflected and transmitted (refracted) waves. Again, we will show that the material in the slab can be chosen to annihilate the reflected wave.

To begin, we briefly review wave propagation in homogeneous, isotropic materials. Let  $\rho$  denote the density, and  $\mu$  and  $\lambda$  denote the Lamé moduli of a homogeneous, isotropic linearly elastic material. If  $\mu > 0$  and  $2\mu + \lambda > 0$ , then exactly two types of waves propagate in the elastic media: *dilatational waves*, in which the directions of displacement and propagation coincide, and *shear waves*, in which the directions of displacement and propagation are orthogonal to each other. Let  $c_d$  and  $c_s$  denote the dilatational and shear wave speeds defined by

$$c_d = \sqrt{\frac{2\mu + \lambda}{\rho}} \quad \text{and} \quad c_s = \sqrt{\frac{\mu}{\rho}}.$$

and let  $D$  and  $S$  denote the dilatational and shear *slowness* given by

$$D = 1/c_d \quad \text{and} \quad S = 1/c_s.$$

For any unit vector  $\mathbf{d}$ , the expression  $\mathbf{v}(\mathbf{x}, t) = \mathbf{d}f(t - D\mathbf{x} \cdot \mathbf{d})$  defines a plane dilatational wave which formally satisfies the equation of motion. Similarly, given two unit vectors  $\mathbf{s}$  and  $\mathbf{p}$  where  $\mathbf{s} \cdot \mathbf{p} = 0$ , the expression  $\mathbf{v}(\mathbf{x}, t) = \mathbf{p}g(t - S\mathbf{x} \cdot \mathbf{s})$  defines a plane shear wave which formally satisfies the equation of motion. The functions  $f$  and  $g$  are called the *wave profiles*, the vectors  $\mathbf{d}$  and  $\mathbf{s}$  are the *propagation vectors*, and the shear wave is said to be *polarized* in the direction  $\mathbf{p}$ . Throughout this paper, we consider harmonic waves; in principle, waves of more general form can be synthesized by the superposition of harmonic waves. The motion of harmonic waves is described by the real or the imaginary parts of the expressions

$$\mathbf{v}(\mathbf{x}, t) = \delta \mathbf{d} e^{i\omega(t - D\mathbf{d} \cdot \mathbf{x})} \quad \text{and} \quad \mathbf{v}(\mathbf{x}, t) = \sigma \mathbf{p} e^{i\omega(t - S\mathbf{s} \cdot \mathbf{x})}.$$

Consider a plane interface  $I$  separating two distinct half-spaces of homogeneous, isotropic elastic materials. A dilatational wave striking the interface typically generates a reflected dilatational wave, a reflected shear wave, a refracted dilatational wave, and a refracted shear wave. Similarly, a shear wave striking the interface typically generates waves of all four types. Therefore, when a combination of dilatational and shear waves impinges upon the interface, eight different waves are generated altogether. The plane formed by the propagation vector of an incident wave and the normal to the interface  $I$  is called the *plane of incidence* for the wave. The propagation vectors of the outgoing waves are determined by a set of equations known as *Snell's Laws* which we state as follows:

The propagation vectors  $\mathbf{d}_r$  and  $\mathbf{s}_r$  for a reflected wave and the propagation vectors  $\mathbf{d}_t$  and  $\mathbf{s}_t$  for the transmitted wave lie in the plane of incidence for the incoming wave. Moreover, if  $\mathbf{m}$  is a unit vector in the intersection of the interface and the plane of incidence, then for an incident dilatational wave with propagation vector  $\mathbf{d}$ , we have

$$(5) \quad D\mathbf{d} \cdot \mathbf{m} = D\mathbf{d}_r \cdot \mathbf{m} = S\mathbf{s}_r \cdot \mathbf{m} = D_t\mathbf{d}_t \cdot \mathbf{m} = S_t\mathbf{s}_t \cdot \mathbf{m}.$$

and, for an incident shear wave with propagation vector  $s$ , we have

$$(6) \quad Ss \cdot m = Dd_r \cdot m = Ss_r \cdot m = D_t d_t \cdot m = S_t s_t \cdot m.$$

Given the unit propagation vectors  $d$  and  $s$  of the incident waves, equations (5) and (6) determine the propagation vectors of the corresponding scattered waves. Note that if a pair of incident dilatational and shear waves share a common plane of incidence and if they satisfy the relation  $Dd \cdot m = Ss \cdot m$ , then the two reflected waves have the same direction as do the two refracted waves. In other words, there are four rather than eight outgoing waves. A pair  $(d, s)$  of incident waves which lie in the same plane of incidence and which satisfy the relation  $Dd \cdot m = Ss \cdot m$  will be called a *conjugate* pair of waves. Note that, when a wave strikes an interface between two homogeneous materials, both the reflected and the transmitted waves form conjugate pairs.

Let us now consider a homogeneous elastic slab of thickness  $T$  separating an isotropic fluid to the right of the slab from a homogeneous, isotropic elastic material to the left of the slab. The ratio of the amplitudes of the reflected and incident waves is the *reflectivity* of the slab. In the paper [1], we obtain a formula for the reflectivity in terms of a *local impedance tensor*. Suppose that a conjugate pair of waves have propagation directions  $d$  and  $s$  and polarization direction  $p$  contained in the plane of incidence for an elastic material. Viewing  $d$  and  $p$  as 2-dimensional vectors in the plane of incidence, we define  $2 \times 2$  matrices

$$A = [d \mid p] \quad \text{and} \quad B = [D\{2\mu(d \cdot n)d + \lambda n\} \mid S\mu\{(s \cdot n)p + (p \cdot n)s\}].$$

Then the local impedance tensor  $H$  is given by  $H = BA^{-1}$ .

Let  $H$  denote the local impedance tensor of the slab,  $H_0$  the local impedance tensor of the left half-space,  $n$  the normal to the slab-fluid interface (pointing into the slab) and  $d_1$  and  $D_1$  the propagation direction and slowness of the incident wave. We regard the fluid as a degenerate elastic solid with Lamé moduli  $\mu_1 = 0$  and  $\lambda_1 > 0$ . By Lemmas 4.1 and 5.1 in [1],

the reflectivity  $r$  of the slab can be expressed as

$$r = \frac{\mathbf{n}^T \mathbf{d}_1 - \lambda_1 D_1 \mathbf{n}^T \Gamma \mathbf{n}}{\mathbf{n}^T \mathbf{d}_1 + \lambda_1 D_1 \mathbf{n}^T \Gamma \mathbf{n}}$$

where

$$(7) \quad \Gamma = [\mathbf{I} + \mathbf{L}][\mathbf{H} - \mathbf{P}\mathbf{H}\mathbf{P}\mathbf{L}]^{-1}, \quad \mathbf{L} = \mathbf{P}\mathbf{A}\mathbf{A}\mathbf{A}^{-1}\mathbf{P}[\mathbf{P}\mathbf{H}\mathbf{P} + \mathbf{H}_0]^{-1}[\mathbf{H} - \mathbf{H}_0]\mathbf{A}\mathbf{A}\mathbf{A}^{-1},$$

$$\mathbf{P} = \mathbf{I} - 2\mathbf{n}\mathbf{n}^T, \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} e^{-i\omega D T \mathbf{d} \cdot \mathbf{n}} & 0 \\ 0 & e^{-i\omega S T \mathbf{s} \cdot \mathbf{n}} \end{bmatrix}.$$

Again,  $T$  denotes the thickness of the slab,  $D$  and  $S$  are the dilatational slowness and shear slowness of the slab, and  $\omega$  is the frequency of the incident wave.

Now consider the problem of choosing the slab material in order to annihilate the reflected wave. Observe that the reflectivity is zero if and only if

$$(8) \quad \mathbf{n}^T \Gamma \mathbf{n} = \frac{\mathbf{n}^T \mathbf{d}_1}{\lambda_1 D_1}.$$

Since the right side of this equation is real, the left side must be real, also. The only way that complex numbers enter  $\Gamma$  is through the diagonal matrix  $\mathbf{A}$ , which appears as two factors of  $\mathbf{L}$ . Let  $a = \omega D T \mathbf{d} \cdot \mathbf{n}$  and  $b = \omega S T \mathbf{s} \cdot \mathbf{n}$  be the parameters that appear in the exponents on the diagonal of  $\mathbf{A}$ . In order to ensure that  $\mathbf{L}$  is real, we must choose  $a$  and  $b$  such that

$$e^{-2ia} = \pm 1, \quad e^{-2ib} = \pm 1, \quad \text{and} \quad e^{-i(a+b)} = \pm 1.$$

Hence, either  $a = m\pi$  and  $b = n\pi$ , or  $a = (m + 1/2)\pi$  and  $b = (n + 1/2)\pi$  for integers  $m$  and  $n$ .

Defining the matrix  $\mathbf{J}$  by

$$\mathbf{J} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

there are essentially 4 distinct  $\Lambda$ , corresponding to different choices of  $m$  and  $n$ , that we need to consider:

$$\begin{aligned}\Lambda = \text{I}, & \quad a = m\pi, \quad b = n\pi, \quad m \text{ and } n \text{ even,} \\ \Lambda = \text{J}, & \quad a = m\pi, \quad b = n\pi, \quad m \text{ odd and } n \text{ even, or } m \text{ even and } n \text{ odd,} \\ \Lambda = \text{iI}, & \quad a = (m + \frac{1}{2})\pi, \quad b = (n + \frac{1}{2})\pi, \quad m \text{ and } n \text{ even,} \\ \Lambda = \text{iJ}, & \quad a = (m + \frac{1}{2})\pi, \quad b = (n + \frac{1}{2})\pi, \quad m \text{ odd and } n \text{ even, or } m \text{ even and } n \text{ odd.}\end{aligned}$$

Each of these cases will be analyzed in the following sections. When studying normal incidence, we saw that there was one "degenerate" case in which total absorption was only possible if the impedances of the left and right half-spaces were identical. For oblique incidence, the degenerate case is  $\Lambda = \text{I}$ .

#### THE CASE $\Lambda = \text{I}$ .

LEMMA 1. *If  $\Lambda = \text{I}$ , then  $r = 0$  if and only if*

$$\mathbf{n}^T \mathbf{H}_0^{-1} \mathbf{n} = \frac{\mathbf{n}^T \mathbf{d}_1}{\lambda_1 D_1}.$$

*Proof.* Observe that  $\mathbf{L} = [\mathbf{PHP} + \mathbf{H}_0]^{-1}[\mathbf{H} - \mathbf{H}_0]$  when  $\Lambda = \text{I}$ . Hence, the second factor in the definition of  $\Gamma$  can be written

$$\begin{aligned}\mathbf{H} - \mathbf{PHPL} &= \mathbf{H} - \mathbf{PHP}[\mathbf{PHP} + \mathbf{H}_0]^{-1}[\mathbf{H} - \mathbf{H}_0] \\ &= \mathbf{H} - (\mathbf{PHP} + \mathbf{H}_0 - \mathbf{H}_0)[\mathbf{PHP} + \mathbf{H}_0]^{-1}[\mathbf{H} - \mathbf{H}_0] \\ &= \mathbf{H}_0(\mathbf{I} + [\mathbf{PHP} + \mathbf{H}_0]^{-1}[\mathbf{H} - \mathbf{H}_0]) \\ &= \mathbf{H}_0(\mathbf{I} + \mathbf{L}).\end{aligned}$$

Referring to the definition of  $\Gamma$ , we see that  $\Gamma = \mathbf{H}_0^{-1}$ . Equation (8) completes the proof.  $\square$

When  $\Lambda = \text{I}$ , the incoming wave is absorbed only if the elastic material in the left half-space satisfies the special condition given in Lemma 1. On the other hand, if the condition of Lemma 1 is satisfied, there is a 1-parameter family of slab materials that annihilates the

reflected wave. In particular, any material that satisfies the conditions

$$\omega T \mathbf{d} \cdot \mathbf{n} = m\pi \quad \text{and} \quad \omega T \mathbf{s} \cdot \mathbf{n} = n\pi$$

where  $m$  and  $n$  are even integers annihilates the reflected wave. Given the angle of the incident wave, the expressions  $\mathbf{d} \cdot \mathbf{n}$  and  $\mathbf{s} \cdot \mathbf{n}$  can be evaluated using Snell's Laws. Omitting the algebra, it follows that the incident wave is absorbed totally when

$$(9) \quad \mu = \rho g_n, \quad \lambda = \rho(g_m - 2g_n), \quad g_n = \frac{\omega^2 T^2}{n^2 \pi^2 + (\omega T D_1 \sin \alpha_1)^2},$$

where  $\alpha_1$  is the angle of the incident wave relative to the normal to the slab interface and  $m$  and  $n$  are even integers. Treating the wave frequency  $\omega$ , the slab thickness  $T$ , and the angle of incidence  $\alpha_1$  as constants, there is a 1-parameter family of perfect absorbers, with Lamé moduli  $\mu$  and  $\lambda$  given by (9) in terms of the parameter  $\rho$ , corresponding to each pair of even integers  $m$  and  $n$ .

**THE STRUCTURE OF  $\Gamma$ .** In order to analyze the other choices of  $\Lambda$ , it helps to see how  $\mathbf{H}$  depends on  $\rho$ . From the definition of  $a$  and  $b$ ,

$$c_d = \frac{\omega T \mathbf{d} \cdot \mathbf{n}}{a} \quad \text{and} \quad c_s = \frac{\omega T \mathbf{s} \cdot \mathbf{n}}{b}.$$

Also, by Snell's Laws, we have

$$c_d = \frac{c_{d_1} \mathbf{d} \cdot \mathbf{m}}{\mathbf{d}_1 \cdot \mathbf{m}} \quad \text{and} \quad c_s = \frac{c_{d_1} \mathbf{s} \cdot \mathbf{m}}{\mathbf{d}_1 \cdot \mathbf{m}}.$$

Hence, the tangent of the angles  $\alpha$  and  $\beta$  (relative to the interface normal  $\mathbf{n}$ ) of the dilatational and shear waves transmitted in the slab are determined:



$$(10) \quad \tan \alpha = \frac{\mathbf{d} \cdot \mathbf{m}}{\mathbf{d} \cdot \mathbf{n}} = \frac{\omega T d_1 \cdot \mathbf{m}}{a c_{d_1}} \quad \text{and} \quad \tan \beta = \frac{\mathbf{s} \cdot \mathbf{m}}{\mathbf{s} \cdot \mathbf{n}} = \frac{\omega T d_1 \cdot \mathbf{m}}{b c_{d_1}}$$

Finally, the wave speeds are given by

$$c_d = \frac{c_{d_1} \sin \alpha}{\mathbf{d}_1 \cdot \mathbf{m}} \quad \text{and} \quad c_s = \frac{c_{d_1} \sin \beta}{\mathbf{d}_1 \cdot \mathbf{m}}$$

In [1] we present an explicit formula for  $\mathbf{H}$  relative to a rectangular coordinate system with  $\mathbf{n}$  pointing along the positive  $x_1$  axis and with  $x_2$  in the interface between the fluid and the slab. Relative to the geometry of Figure 1, we have  $\mathbf{H} = \rho \bar{\mathbf{H}}$  where

$$\bar{\mathbf{H}} = \frac{c_s}{\cos(\alpha - \beta)} \begin{bmatrix} \phi \cos \beta & -\sin(\alpha - 2\beta) \\ \sin(\alpha - 2\beta) & \cos \alpha \end{bmatrix}, \quad \phi = c_d/c_s,$$

and  $\alpha$  and  $\beta$  are given by (10). In summary, for fixed  $a$  and  $b$ ,  $\mathbf{H}$  is a linear function of the

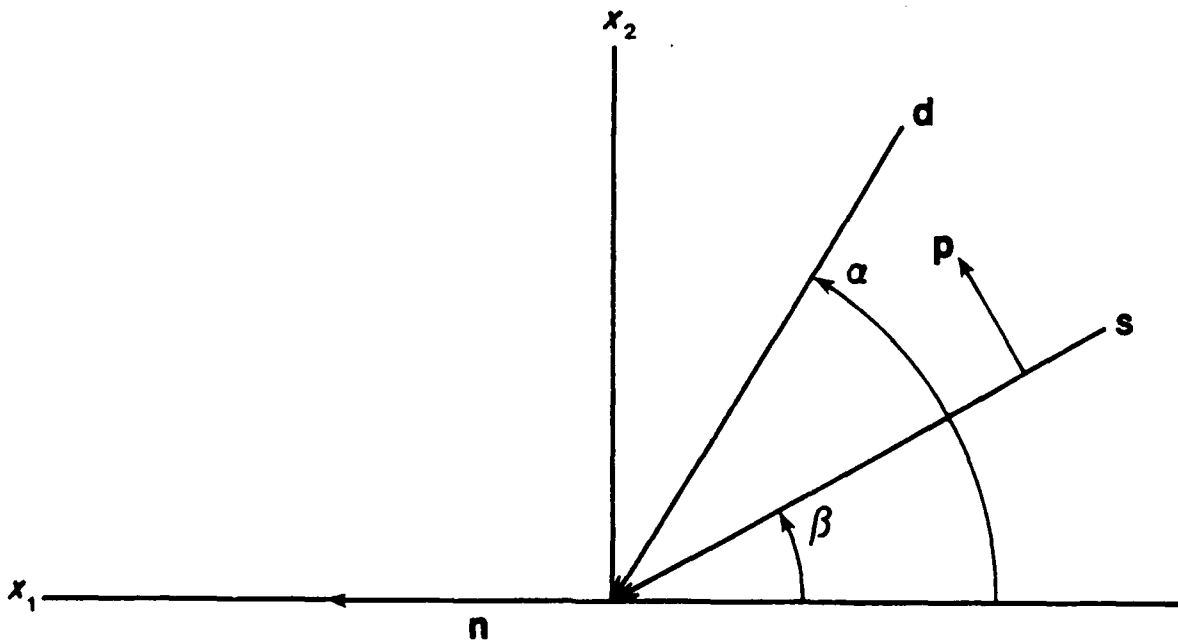


Figure 1. Propagation and polarization vectors.

density  $\rho$  of the slab. Observe that  $\bar{H}$  is invertible whenever  $c_d > 0$  and  $c_s > 0$ . Moreover,  $H^{-1} = \rho^{-1}\bar{H}^{-1}$  where

$$\bar{H}^{-1} = \frac{\cos(\alpha - \beta)}{c_d \cos \beta \cos \alpha + c_s \sin^2(\alpha - 2\beta)} \begin{bmatrix} \cos \alpha & \sin(\alpha - 2\beta) \\ -\sin(\alpha - 2\beta) & \phi \cos \beta \end{bmatrix}$$

and  $\phi = c_d/c_s$ .

**LARGE  $\rho$ .** Let us now determine the limiting behavior of  $\Gamma$  as  $\rho$  tends to infinity. Since  $H$  depends linearly on  $\rho$ , we see from (7) that each element of  $\Gamma$  is a rational function of  $\rho$ ; that is, each element of  $\Gamma$  is the ratio of two polynomials. Recall that a rational function is either constant (independent of  $\rho$ ), or it has a finite number of zeros and poles. We already discovered one case where this rational function is completely constant -- when  $\Lambda = I$ ,  $\Gamma = H_0^{-1}$  independent of  $\rho$ . However, in general  $\Gamma$  depends on  $\rho$ . Since  $H^{-1}H_0$  approaches zero as  $\rho$  tends to infinity, the limit of  $L$  as  $\rho$  tends to infinity is easily evaluated:

$$\lim_{\rho \rightarrow \infty} L = \bar{L} \quad \text{where} \quad \bar{L} = P\Lambda\Lambda^{-1}\bar{H}^{-1}P\bar{H}\Lambda\Lambda^{-1}.$$

Referring to (7),  $\Gamma$  has the following asymptotic form as  $\rho$  tends to infinity:

$$(11) \quad \Gamma = \rho^{-1}\bar{\Gamma} + O(\rho^{-2}) \quad \text{where} \quad \bar{\Gamma} = (I + \bar{L})(\bar{H} - P\bar{H}P\bar{L})^{-1}.$$

Note that the formula (11) only makes sense when the quantity  $\bar{H} - P\bar{H}P\bar{L}$  is nonsingular. In particular, for the special case  $\Lambda = I$ , we have

$$\bar{H} - P\bar{H}P\bar{L} = 0,$$

which explains why the asymptotic limit (11) is incorrect when  $\Lambda = I$ . However, for the other choices of  $\Lambda$ , the quantity  $\bar{H} - P\bar{H}P\bar{L}$  is generally nonsingular. In particular, in the case  $\Lambda = iI$ , it is readily verified that

$$\bar{H} - \bar{P}\bar{H}\bar{P}\bar{L} = 2\bar{H},$$

which is nonsingular since  $\bar{H}$  is nonsingular. For the case  $\Lambda = J$ , let us employ the coordinate system depicted in Figure 1. In this coordinate system,  $P$  is equal to  $J$  and the matrix  $\bar{H} - \bar{P}\bar{H}\bar{P}\bar{L}$  is a nonsingular multiple of the expression

$$B J B^{-1} - J B J B^{-1} J,$$

where  $B$  is the matrix that appears in the definition of the impedance tensor:  $H = B A^{-1}$ . Similarly, taking  $\Lambda = iJ$ , it follows that  $\bar{H} - \bar{P}\bar{H}\bar{P}\bar{L}$  is a nonsingular multiple of the expression

$$B J B^{-1} + J B J B^{-1} J.$$

Focusing on  $B J B^{-1} \pm J B J B^{-1} J$ , we have

LEMMA 2. *Given a nonsingular matrix*

$$B = \begin{bmatrix} a & c \\ -b & d \end{bmatrix},$$

the expression  $B J B^{-1} - J B J B^{-1} J$  is nonsingular if and only if  $a, b, c,$  and  $d$  are nonzero. The expression  $B J B^{-1} + J B J B^{-1} J$  is nonsingular if and only if  $bc \neq ad$ .

*Proof.* This is verified by evaluating the determinants:

$$\det(B J B^{-1} - J B J B^{-1} J) = \frac{16abcd}{\det B}$$

and

$$\det(B J B^{-1} + J B J B^{-1} J) = \frac{-4(ad - bc)^2}{\det B} \quad \square$$

In [1] we provide the following representation of  $B$  for the geometry depicted in Figure 1:

$$\mathbf{B} = \frac{1}{c_d} \begin{bmatrix} (2\mu + \lambda) \cos 2\beta & \phi\mu \sin 2\beta \\ -\mu \sin 2\alpha & \phi\mu \cos 2\beta \end{bmatrix}$$

where  $\phi = c_d/c_s$ . First let us consider the case  $\Lambda = \mathbf{J}$  so that  $\bar{\mathbf{H}} - \mathbf{P}\bar{\mathbf{H}}\mathbf{P}\bar{\mathbf{L}}$  is nonsingular if and only if every element of  $\mathbf{B}$  is nonzero. Assuming the angle of incidence is not normal to the slab (normal incidence was studied earlier), the (1,2) and (2,1) elements of  $\mathbf{B}$  are nonzero. Although the (1,1) and (2,2) elements of  $\mathbf{B}$  are zero if  $\beta = \pi/4$ , an infinitesimal perturbation in the thickness or the frequency yields  $\beta \neq \pi/4$  and  $\bar{\mathbf{H}} - \mathbf{P}\bar{\mathbf{H}}\mathbf{P}\bar{\mathbf{L}}$  invertible. In the case  $\Lambda = i\mathbf{J}$ , it follows from Lemma 2 that  $\bar{\mathbf{H}} - \mathbf{P}\bar{\mathbf{H}}\mathbf{P}\bar{\mathbf{L}}$  is singular if and only if

$$(2\mu + \lambda) \cos^2 2\beta = \mu \sin 2\alpha \sin 2\beta.$$

Utilizing (10), this relation is equivalent to

$$(12) \quad \beta = \arctan \sqrt{\psi + 1 \pm \sqrt{\psi^2 + 2\psi}} \quad \text{where} \quad \psi = 2a/b.$$

Referring to (10), we see that there are special values for the frequency and thickness that lead to singularity; but again, an infinitesimal perturbation of  $T$  or  $\omega$  restores invertibility. We say that the slab is *singular* if either  $\Lambda = \mathbf{I}$ ,  $\Lambda = \mathbf{J}$  and  $\beta = \pi/4$ , or  $\Lambda = i\mathbf{J}$  and  $\beta$  satisfies (12).

In summary, when the slab is nonsingular,  $\Gamma$  approaches (asymptotically)  $\bar{\Gamma}/\rho$  as  $\rho$  increases. In particular,  $\Gamma$  tends to *zero* as  $\rho$  increases.

**SMALL  $\rho$ .** Let us consider a nonsingular slab and the geometry depicted in Figure 1. In this case,  $\mathbf{n}^T \Gamma \mathbf{n}$  equals the (1,1) element of  $\Gamma$  which we denote  $\gamma$ . Since  $\Gamma$  is a rational function of  $\rho$ ,  $\gamma$  is a rational function of  $\rho$  that tends to zero as  $\rho$  increases. In a separate paper, we will show that  $\gamma \geq 0$  for every choice of the density. Consequently,  $\gamma$  has no poles along the positive real axis, and we can satisfy (8) for some  $\rho$  whenever

$$0 \leq \frac{\mathbf{n}^T \mathbf{d}_1}{\lambda_1 D_1} \leq \gamma_0,$$

where  $\gamma_0$  denotes the limit of  $\gamma$  as  $\rho$  tends to zero ( $\gamma$  approaches zero as  $\rho$  becomes large,  $\gamma$  approaches  $\gamma_0$  as  $\rho$  tends to zero, and  $\gamma$  depends continuously on  $\rho$ ). Thus the value of  $\gamma_0$  provides insight concerning the incident waves that can be absorbed.

We have evaluated  $\gamma_0$  for each of the choices  $\Lambda = \mathbf{J}$ ,  $\Lambda = i\mathbf{I}$ , and  $\Lambda = i\mathbf{J}$ . It turns out that the evaluation of  $\gamma_0$  is quite difficult since very complicated trigonometric matrices must be multiplied together and simplified. With the assistance of a symbolic manipulation package, we found that for  $\rho$  near 0 and for each choice of  $\Lambda$ ,  $\Gamma$  has an expansion of the form:

$$\Gamma = T\rho^{-1} + S^T H_0^{-1} S + O(\rho)$$

where the  $T$  and  $S$  corresponding to the various choices of  $\Lambda$  appear in Table 1.

Observe that in each case, the (1, 1) element of  $T$  is zero. Thus for each choice of  $\Lambda$ ,  $\gamma_0$  is the (1, 1) element of  $S^T H_0^{-1} S$ , which is easily evaluated:

LEMMA 3.

$$\begin{aligned} \gamma_0 &= \frac{(H_0^{-1})_{11}}{(1 - 2 \cos 2\beta)^2} && \text{when } \Lambda = \mathbf{J}, \\ \gamma_0 &= \frac{(H_0^{-1})_{22} \cos^2 \alpha}{\sin^2(\alpha - 2\beta)} && \text{when } \Lambda = i\mathbf{I}, \\ \gamma_0 &= \frac{(H_0^{-1})_{22} \cos^2 \alpha \cos^2(\alpha - \beta)}{(\sin \alpha \cos(\alpha + \beta) + \cos \beta \sin 2\beta)^2} && \text{when } \Lambda = i\mathbf{J}. \end{aligned}$$

*For a nonsingular slab and for each choice of  $\Lambda$ , there exists a value of  $\rho$  that absorbs the incident wave whenever*

$$(13) \quad 0 \leq \frac{\mathbf{n}^T \mathbf{d}_1}{\lambda_1 D_1} \leq \gamma_0$$

Finally, let us verify the claim made at the beginning of this paper concerning the existence of a material that totally absorbs any given incident wave.

$$\Lambda = J$$


---

$$T = \frac{2 \sin \beta}{c_r (1 - 2 \cos 2\beta)} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$S = \frac{1}{1 - 2 \cos 2\beta} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$


---

$$\Lambda = iI$$


---

$$T = \frac{\cos(\alpha - \beta)}{c_r \sin(\alpha - 2\beta)} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad S = \frac{1}{\sin(\alpha - 2\beta)} \begin{bmatrix} 0 & \phi \cos \beta \\ -\cos \alpha & 0 \end{bmatrix}$$


---

$$\Lambda = iJ$$


---

$$T = \frac{\cos 2\alpha + \cos 2\beta}{2c_r (\sin \alpha \cos(\alpha + \beta) + \cos \beta \sin 2\beta)} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$S = \frac{\cos(\alpha - \beta)}{\sin \alpha \cos(\alpha + \beta) + \cos \beta \sin 2\beta} \begin{bmatrix} 0 & \phi \cos \beta \\ \cos \alpha & 0 \end{bmatrix}$$


---

Table 1. T and S for various choices of  $\Lambda$ .

**THEOREM 1.** *For any given incident wave, the mechanical properties of the slab can be chosen so that the amplitude of the reflected wave is zero.*

*Proof.* By equation (10), the angles  $\alpha$  and  $\beta$  can be made arbitrarily close to zero by taking  $m$  and  $n$  sufficiently large. By Lemma 3,  $\gamma_0$  tends to infinity as  $\alpha$  and  $\beta$  tend to zero if  $\Lambda = iI$ . Also, by (11)  $\gamma$  tends to zero as  $\rho$  increases when  $\Lambda = iI$ . Since the inequality (13) holds for  $m$  and  $n$  sufficiently large, there exists a density for which the slab totally absorbs the incident wave.  $\square$

**NUMERICAL EXPERIMENTS.** The inequality (13) provides a lower bound on the range of incident angles and frequencies that can be absorbed. Although the range of  $\gamma$  as a function of  $\rho$  contains the interval  $[0, \gamma_0]$ , potentially the range extends outside the interval. Experimentally, we find that  $\gamma$  is nearly a monotone function of  $\rho$  so that the interval  $[0, \gamma_0]$  accurately predicts the incident waves that can be absorbed. Figures 2, 3, and 4 show typical plots of  $\gamma$  as a function of  $\rho$  for various choices of  $\Lambda$ . These graphs correspond to a material like steel in the left half-space and the fluid water in the right half-space. In particular, the following mechanical parameters were employed:

Right half-space:  $\rho_1 = 1 \text{ gm/cm}^3$ ,  $c_{d_1} = 140000 \text{ cm/sec}$ ,  $\lambda_1 = \rho_1 c_{d_1}^2$ .  
 Left half-space:  $\rho_0 = 7 \text{ gm/cm}^3$ ,  $\lambda_0 = \lambda_1$ ,  $\mu_0 = \lambda_0/3$ .  
 Slab:  $T = 5 \text{ cm}$ .  
 Incident wave:  $\alpha_1 = 30 \text{ degrees}$ ,  $\omega = 600\pi \text{ rad/sec}$ .

Since the graphs in Figures 2, 3, and 4 appear monotone, the range of  $\gamma$  is accurately estimated by the interval  $[0, \gamma_0]$ . (Note though that the numerical values of  $\gamma$  in Figure 4 deviate from monotonicity in the fourth significant digit near  $\rho = 0$ , a deviation that is imperceptible to the eye, but which is large enough to undermine any proof of monotonicity for  $\gamma$ .) In Figure 5 we plot the density of the material that totally absorbs the incoming wave versus the angle of incidence. Observe that as the angle of incidence approaches 90 degrees (with the wave speed fixed), the density tends to infinity. In Figure 6 we plot the density of the material that totally absorbs the incoming wave versus the wave frequency. Observe that

as frequency tends to zero (with the wave speed fixed), density tends to infinity, and as frequency tends to infinity, density tends to zero.

Numerically, we investigated singular slabs associated with  $\Lambda = J$  and  $\Lambda = iJ$ . We found that  $\gamma$  was equal to  $\gamma_0$ , independent of  $\rho$  (for fixed wave speed).

#### REFERENCES

- [1] W. W. Hager and R. Rostamian, Reflection and refraction of elastic waves for stratified materials, *Wave Motion*, 10(1988), pp. 333-348.

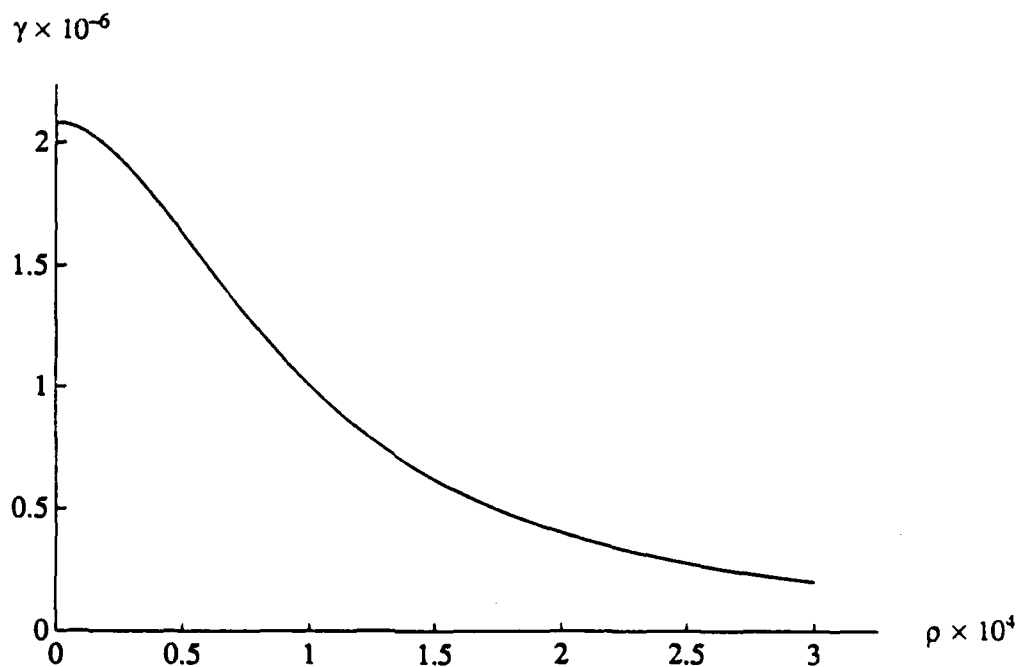


Figure 2.  $\gamma$  versus density for  $\Lambda = J$ ,  $m = 1$ , and  $n = 2$ .



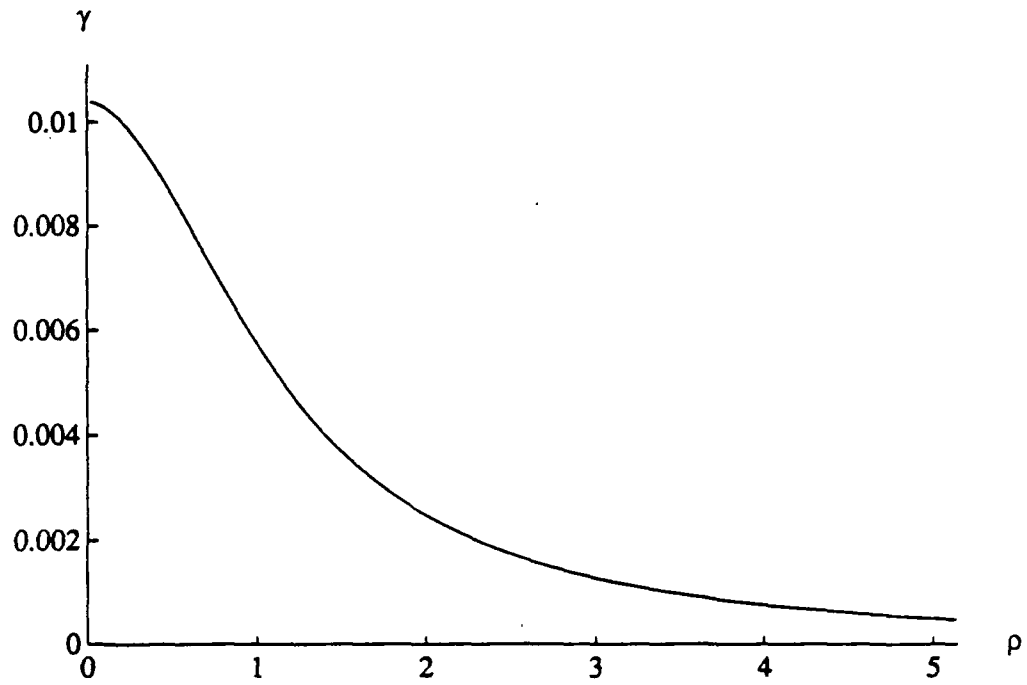


Figure 3.  $\gamma$  versus density for  $\Lambda = iI$ ,  $m = 0$ , and  $n = 0$ .

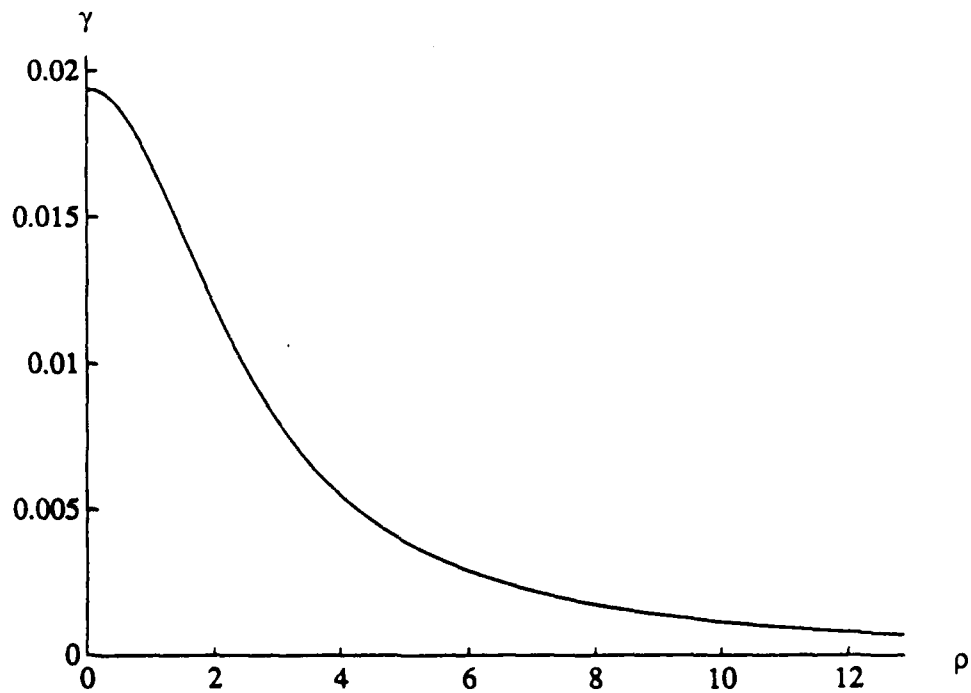


Figure 4.  $\gamma$  versus density for  $\Lambda = iJ$ ,  $m = 1$ , and  $n = 2$ .

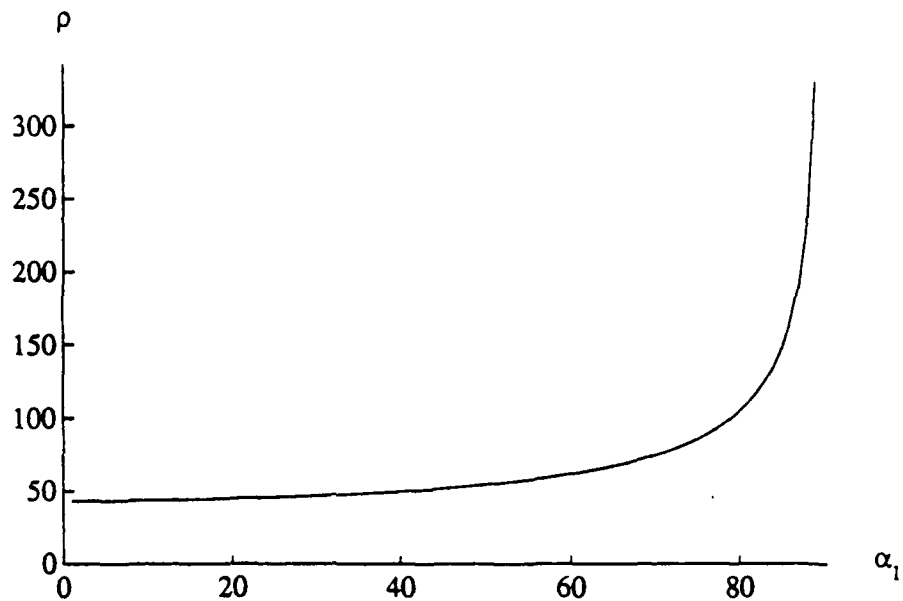


Figure 5. Density of totally absorbing slab versus angle of incidence (degrees) for  $\Lambda = iJ$ ,  $m = 0$ , and  $n = 1$ .

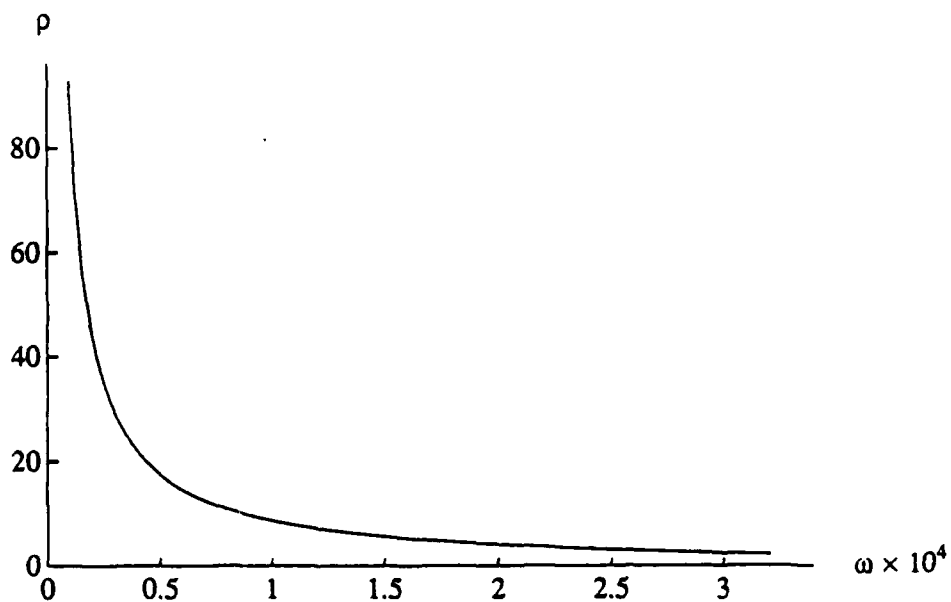


Figure 6. Density of totally absorbing slab versus frequency (radians/second) of incident wave for  $\Lambda = iJ$ ,  $m = 0$ , and  $n = 1$ .

# TRANSIENT SHEAR RESPONSE OF A RIGID BLOCK AND FLEXIBLE SUPPORT ASSEMBLY SUBJECTED TO A LATERAL IMPACT

Aaron Das Gupta  
Research Mechanical Engineer  
U.S. Army Ballistic Research Laboratory  
Aberdeen Proving Ground, MD 21005-5066

## ABSTRACT

A shear block approach has been used to model the transient shear response of a rigid block-flexible support system subjected to a nonpenetrating side-on hypervelocity projectile impact. The initial velocity imparted to the shear block due to impact has been calculated using a momentum balance between the projectile and the rigid block and has been imposed as an initial condition for the transverse dynamic equation of motion of the block-support assembly. The spring constant of the support has been evaluated based on the support height and the shear area. The forcing function has been computed assuming that the entire projectile length is consumed at a constant rate in a finite length of time and a triangular force-time relationship is imposed on the system. The nonhomogeneous transverse equation of motion for the assembly is solved for transverse displacement, velocity and acceleration and the constants for the complimentary and the particular part of the solution are evaluated using a set of initial and boundary conditions. The displacement solution is optimized by setting the velocity equal to zero and obtaining a peak response time at which the displacement is an optimum. The acceleration at this time is found to be negative ensuring that the solution for displacement is a global maximum. Once the peak transverse displacement of the block-support system is known, peak shear stress and strain can be easily calculated and compared to the shear yield strength of the parent material in order to ensure the structural integrity of the system from a shear strength standpoint or predict the occurrence of dynamic shear failure of the assembly at the interface between the block and the support.

## INTRODUCTION

The capability to predict the effect of hypervelocity impact of a missile upon a rigid or deformable structure is a necessity as a first step towards the design and safe operation of nuclear reactors (1,2) as well as defense systems subjected to extreme environments. This problem is also of considerable interest to the Ballistic Research Laboratory (BRL) due to possibility of sustaining severe damage at a vulnerable location of the target structure when impacted by a projectile at a specific angle of obliquity.

A number of studies have been performed and damage data gathered (3-7) 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 terms of dynamic stress-strain response of multibody systems consisting of interconnected rigid and deformable bodies subjected to impact and sudden change in the structure.

Recently, computation using hydrodynamic codes (8-10) has been reported. Unfortunately, setting up an accurate computational model, code computation and assimilation as well as correct interpretation of the results are expensive and time consuming and require considerable expertise for the project leader. Because of limited time and cost constraints it was decided early on to resort to a feasible analytical approach in lieu of a numerical approach which will eliminate undue complexities of the real problem while retaining the essential features of the loading process and giving an insight into the impact phenomena, the dominating stress and failure mechanisms.

## IMPACT CONDITION

Let us assume that a large projectile of mass  $M_1$  travelling with an initial velocity,  $V_1$ , in a horizontal direction collides with a stationary massive object of mass  $M_2$  supported underneath by a series of plates which in turn are connected to an even larger mass by means of continuous double seam welding. Because of the nature of these masses and type of construction, shear phenomena appears to dominate stresses and failure in such structures rather than bending which is the governing mechanism in mass-beam coupled systems.

Assuming the target to be rigid and a constant average deceleration rate upon impact based upon a linear decay of velocity from  $V_1$  to a zero velocity as well as an average duration time to consume the total length of the impactor, it is possible to calculate a linearly decaying forcing function with a triangular equivalent impulse which can be imposed upon the rigid mass in a side-on horizontal direction such that

$$F_p = M_1(V_1 - 0)/(T_1 - T_2) = M_1 V_1 / T \quad (1)$$

where  $T$  is the duration time and  $F_p$  is the decelerating force.

Invoking a momentum balance between the impactor and the target mass which is now allowed to move in a horizontal direction, it is possible to compute the imparted final velocity of the target as follows:

$$\begin{aligned} M_1 V_1 &= V_2 (M_1 + M_2) \\ \text{or, } V_2 &= M_1 V_1 / (M_1 + M_2) \end{aligned} \quad (2)$$

where,  $M_1$  is the impacting mass with an initial velocity of  $V_1$  and  $M_2$  is the target mass. Once the imparted velocity of the target mass is known it can be imposed as a constraint condition for the equation of motion to solve the boundary value problem.

## PROBLEM FORMULATION

Prior to shear stress computation it is necessary to obtain the dynamic equation of motion of the target-support assembly in the form:

$$M_2 \ddot{x} + Kx = F(t) \quad (3)$$

where  $F(t)$  is the externally applied force upon the target,  $K$  is the support stiffness and

$x$  is the horizontal displacement of the target as shown in Figure 1. From this figure shear strain,  $y$ , at the block interface and shear stress,  $T$ , can be given as

$$\begin{aligned} y &= x/h \\ T &= Gy \end{aligned} \quad (4)$$

where  $h$  is the height of the shear block support and  $G$  is the shear modulus.

The spring constant for the support is evaluated by referring to the free body diagram as shown in Figure 2 where  $F_{s,h}$  is the shear force at the interface given as

$$\begin{aligned} F_{s,h} &= 2TA = 2GyA = 2GA(x/h) = kx \\ k &= 2GA/h \end{aligned} \quad (5)$$

where  $A$  is the shear area at the interface between the block and the plate. The equation of motion of the block support system could be rewritten as

$$M_2 \ddot{x} + 2GAx/h = F(t) \quad (6)$$

#### METHOD OF SOLUTION

The dynamic equation of motion of the block support assembly subjected to a horizontal side-on impact load as given in the previous section, needs to be solved for the time dependent displacement,  $x$ , subjected to the constraint conditions that initial displacement is zero at time  $t = 0$  when initial velocity is  $V_2$  which is the initial target velocity obtained earlier from invoking the momentum balance.

The forcing function,  $F$ , could now be assumed to be a triangular force-time curve with linear decay in the form

$$F = F_p[1 - t/t_p] \quad (7)$$

where  $F_p$  is the peak impact force and  $t_p$  is the positive phase duration. At  $t = 0$ ,  $F$  reduces to  $F_p$  and at  $t = t_p$ ,  $F$  vanishes which satisfies the initial constraint conditions.

The equation of motion could now be rewritten as

$$M_2 \ddot{x} + 2GAx/h = F_p[1 - t/t_p] \quad (8)$$

The solution of above equation of motion can be expressed as a sum

$$x(t) = x_c(t) + x_p(t) \quad (9)$$

where  $x_c(t)$  is the complimentary solution satisfying the homogeneous equation

$$\ddot{x} + 2GAx/(M_2h) = 0 \quad (10)$$

and  $x_p(t)$  is the particular solution satisfying the nonhomogeneous equation

$$\ddot{x} + 2GAx/(M_2h) = F_p[1 - t/t_p]/M_2 \quad (11)$$

A complimentary solution for the standard homogeneous equation above can be given as

$$x_c(t) = A\cos(\omega t) + B\sin(\omega t) \quad (12)$$

where A and B are constants to be evaluated from the initial and boundary conditions and  $\omega$  is calculated as

$$\omega = 2GA/(M_2h) \quad (13)$$

Similarly, following the procedures outlined above with some modification for the non-homogeneous part of the equation of motion, a particular solution could be obtained as a function of the peak load, target mass, plate stiffness, positive phase duration and the elapsed time as shown below :

$$x_p(t) = [F_p/(\omega^2 M_2)][1 - t/t_p] \quad (14)$$

Hence the total solution for displacement of the shear block in a horizontal direction is given as

$$x(t) = A\cos(\omega t) + B\sin(\omega t) + [F_p(1 - (t/t_p))]/(\omega^2 M_2) \quad (15)$$

Once displacement-time history of the impacted structure is known it is possible to obtain velocity of the block in a horizontal direction by differentiating the above equation with respect to time which results in

$$\dot{x}(t) = B\omega\cos(\omega t) - [A\omega\sin(\omega t) + F_p/(\omega^2 M_2 t_p)] \quad (16)$$

where A and B are constants evaluated from initial and boundary conditions for the problem.

Acceleration-time relationship for the target-support assembly can be easily obtained by differentiating the velocity in equation above with respect to time which yields

$$\ddot{x}(t) = -(B \omega^2 \sin(\omega t) + A\omega^2 \cos(\omega t)) \quad (17)$$

The minus sign on the right hand side of the equation indicates negative acceleration or deceleration of the block with time which is to be expected due to the restraining action of the welded supporting plates underneath the block.

#### OPTIMIZATION PROCEDURE

In order to predict the magnitude of peak displacement and peak shear stresses as well as strains realized by the shear block at the interface between the block and the beam, it is necessary to determine the specific time of occurrence of the peak response. Optimization of the peak response by some means is essential to arrive at an optimum occurrence time.

A standard mathematical approach to optimization is adopted in lieu of a trial and error minimization scheme. In order to maximize the displacement the derivative of the displacement with respect to time or the velocity can be set to zero such that

$$Bw\cos(\omega t_{op}) - [Aw\sin(\omega t_{op}) + F_p/(\omega^2 M_2 t_p)] = 0 \quad (18)$$

where  $t_{op}$  is the optimum time at which the peak displacement response occurs. The above equation is required to be solved for the unknown optimum time. In the particular case where  $\omega t_{op}$  is small the above equation simplifies to a form :

$$Bw - Aw^2 t_{op} + F_p/(\omega^2 M_2 t_p) = 0 \quad (19)$$

$$\text{or, } t_{op} = (1/Aw)[(B + F_p/(M_2 t_p \omega^3))] \quad (20)$$

Peak displacement could now be easily computed by substituting the expression for the optimum time given above in the equation for displacement time relationship obtained earlier which can be reduced to a simpler form

$$x_p' = A + B\omega t_{op} + F_p[1 - t_{op}/t_p]/M_2 \omega^2 \quad (21)$$

where  $x_p'$  is the peak displacement of the block assembly at time  $t_{op}$ . Now substituting the value of the optimum occurrence time in the above equation one can arrive at an algebraic expression for the peak displacement of the block in the form

$$x_p' = ((A^2 + B^2)/A) + [F_p/(\omega^2 M_2)][1 - F_p/(AM_2 t_p^2 \omega^4)] \quad (22)$$

where  $F_p, \omega, M_2$  and  $t_p$  are previously defined known quantities with specific values for a particular problem and A, B are constants evaluated from initial and boundary conditions.

The displacement is guaranteed to be a global maximum provided the double derivative of the displacement with respect to time or the acceleration at the optimum time of occurrence is negative such that

$$-(B\omega^2 \sin(\omega t_{op}) + A\omega^2 \cos(\omega t_{op})) < 0 \quad (23)$$

$$\text{or, } \tan(\omega t_{op}) > -(A/B) \quad (24)$$

For each optimum time,  $t_{op}$ , the above inequality must be checked out for the specific problem in order to ensure that the peak displacement is indeed a global maximum. Similarly the optimum response time at which the velocity of the shear block attains a peak could be determined by setting the right hand expression of the acceleration equation equal to zero and verifying that the derivative of the acceleration with respect to time at this time of occurrence is negative which ensures that the peak velocity is a global maximum.

## RESULTS AND DISCUSSION

Although it may be possible to solve exactly for the optimum time from the equations resulting from the optimization procedure described in the previous section, it is sufficient for most problems to adopt a trial and error approach where various suitable values of the optimum time are substituted in the left hand side expression of the velocity equation. The difference between the calculated velocity and zero which is the right hand side of the equation is treated as an error which is minimized by adjusting the optimum time until it nearly vanishes.

Once the peak shear displacement is obtained as outlined, it is fairly easy to calculate the peak shear strain as a ratio of the peak transverse displacement and the height of the support plates. The peak shear stress at the interface between the block and the beam can be easily obtained by multiplying the shear strain with the shear modulus of the material for the block-support assembly. The shear stress could be compared with the ultimate or yield shear strength of the parent material in order to determine the structural integrity of the assembly. A factor of safety can be worked out by taking a ratio of the ultimate or yield strength of the material for the support plates to the actual shear stress developed at the interface. If the factor of safety is less than or equal to 1.0, structural failure in shear is indicated at the interface requiring redesign of the block-support assembly. However, if the factor of safety is greater than 1.0 a margin of safety can be given as a measure of structural integrity.

Although the analysis resorts to several simplifying assumptions regarding the loading function and the details of the assembly, it gives a valuable insight into the dynamic shear response behavior of a class of structures subjected to side-on impact loading. The analysis could be extended to side-on overpressure loading due to a blast by modifying the forcing function and reformulating the equation of motion resulting in a somewhat different type of solution appropriate for explosive loading. The procedure outlined above is a quick and inexpensive method of solution of response of structures dominated by shear phenomena occurring at interfaces.

## ACKNOWLEDGEMENTS

Valuable assistance of Drs Andrew V. Mark and 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 and 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, "Impact of an Elastic Perfectly-Plastic Plate on a Rigid Target," Volume 3, Bk. No. G0431C, *Proceedings of the 1988 ASME International Computers in Engineering Conference*, San Francisco, CA, August 1-3, 1988.



4. J.T. Dehn, "Models of Explosively Driven Metals," U.S. Army Ballistic Research Laboratory Technical Report No. BRL-TR-2626, Aberdeen Proving Ground, MD, 1984.
5. R.M. Norman, "Deformation in Flat Plates Exposed to HE Mine Blast," AMSAA-TM-74, U.S. 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. Greczuk 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," Air Force Weapons Laboratory Report No. AFWL-TR-70-144, Kirtland Air Force Base, Albuquerque, NM, 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," U.S. Army Ballistic Research Laboratory Report No. ARBRL-MR-03249, Aberdeen Proving Ground, MD, 1978.

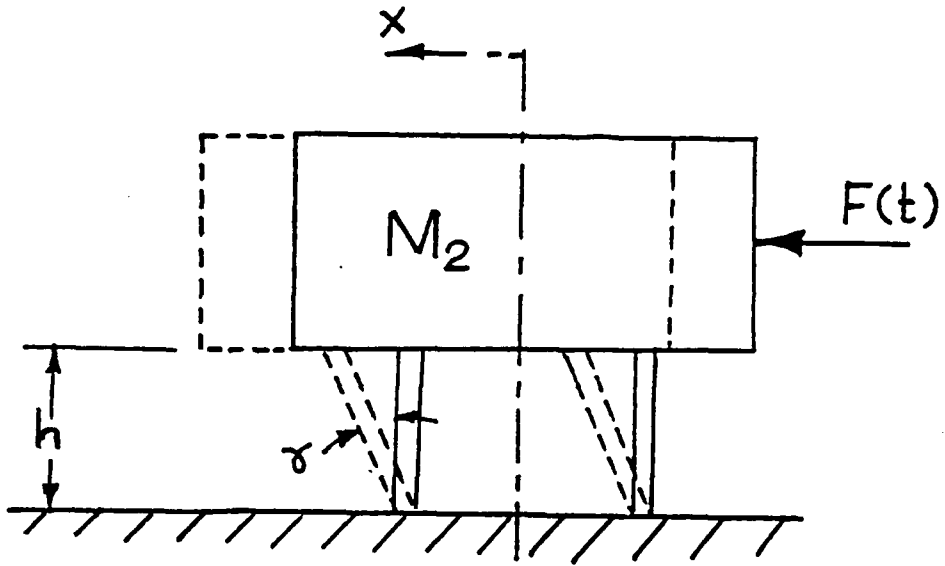


Figure 1: Shear Block Model Subjected to an Impact Load.

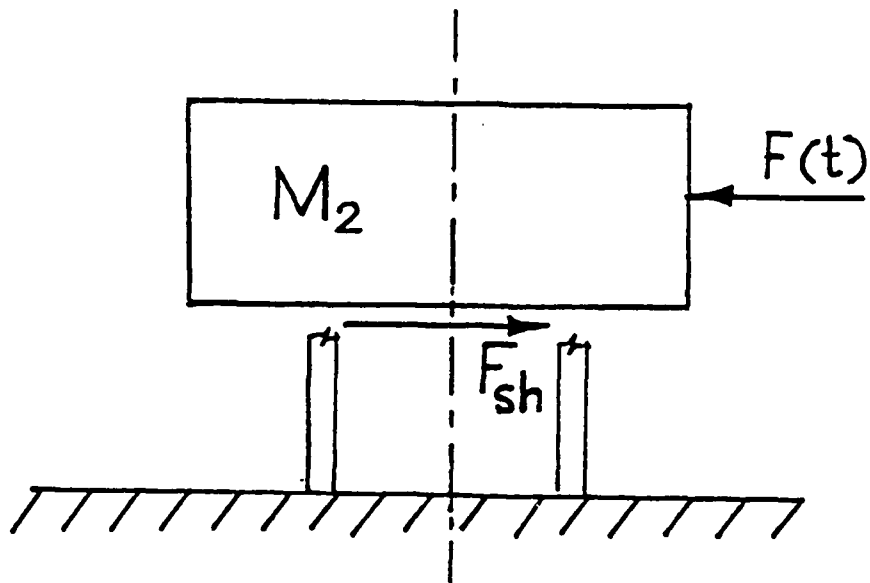


Figure 2: Free-Body Diagram for the Shear Block Model.

# ON THE CONTINUUM MECHANICS OF THE MOTION OF A PHASE INTERFACE<sup>1</sup>

Morton E. Gurtin  
Department of Mathematics  
Carnegie-Mellon University  
Pittsburgh, PA 15213

ABSTRACT. A recent series of papers [G,AG,GS] began an investigation whose goal is a thermomechanics of two-phase continua based on Gibbs's notion of a sharp phase-interface endowed with thermomechanical structure. In [G] a new balance law, balance of capillary forces, was introduced and then applied in conjunction with suitable statements of the first two laws of thermodynamics; the chief results are thermodynamic restrictions on constitutive equations, exact and approximate free-boundary conditions at the interface, and a hierarchy of free-boundary problems. [AG] applied this theory to perfect conductors, in which the underlying equations reduce to a single evolution equation for the interface. [G] and [AG] were limited to rigid systems; [GS] extends the theory to include bodies that deform as they solidify or melt. These theories involve several new concepts, examples being: the creation of new material points; work intrinsic to a moving interface; the formulation of conservation laws for a moving interface. Here I shall discuss some of the new ideas involved in [GS].

MECHANICS AND ENERGETICS OF DEFORMING, ACCRETING  
CRYSTALS. In [GS],<sup>2</sup> the body, ostensibly a crystal, is allowed:

<sup>1</sup>Supported by the U. S. Army Research Office.

<sup>2</sup>[GS] was motivated by studies of Leo and Sekerka [LS], Alexander and Johnson [AJ,JA], and Larche and Cahn [LC], which derive *equilibrium* relations for the crystal surface as Euler-Lagrange equations corresponding to a stationary global Gibbs function. Such derivations are appropriate to statics but tend to obscure the fundamental nature of balance laws as *basic axioms* in any dynamical framework which includes inertia and dissipation.

- (i) to crystallize through the addition or deletion of material points at the crystal surface, a process termed *accretion*;
- (ii) to *deform*.

In conjunction with these kinematical processes, *two distinct force systems are introduced*:

- (i) a system of *accretive forces* which acts within the crystal lattice to drive the crystallization process;
- (ii) a system of *deformational forces* to be identified with the more or less standard forces that act in response to the local motion of material points.

Because of the nonclassical nature of accretive forces, it is not at all clear that there should be an accompanying balance law, let alone what it should be and how it should relate to the deformational system. For that reason the underlying mechanical balance laws are derived from the requirement that the *mechanical production* - the rate of kinetic energy minus the rate of working - be independent of the observer. Here it is necessary to introduce a new idea, that of a *lattice observer*: in addition to the standard observer who measures the gross velocities of the continuum, *there is a second observer,<sup>3</sup> who studies the lattice and measures the velocity of the accreting crystal surface*. This procedure leads, not only to the "standard" balance laws for linear and angular momentum, but to new laws expressing balance of (micro)forces and (micro)moments within the crystal lattice at the crystal surface.

One of the chief differences between theories involving phase transitions and the more classical theories of continuum mechanics is the creation and deletion of material points as the phase interface moves relative to the underlying material. We associate with this process internal forces whose working provides an outflow of "mechanical energy" associated with the attachment and release of atoms as they are exchanged between phases. We write an energy balance relating these internal forces, the forces

---

<sup>3</sup>The use of more than one observer might be useful in other continuum theories, such as theories of liquid crystals, of structured continua, or of mixtures, in which "force"-balance laws over and above the standard laws arise.

described previously, and the bulk energy of the two phases at the crystal surface.

COHERENT CRYSTAL-CRYSTAL INTERACTIONS. To illustrate the results of the general theory,<sup>4</sup> consider an isothermal crystal-crystal interaction,<sup>5</sup> in which the environment consists of a second solid phase of the crystal material, and in which the reference lattices can be chosen to match exactly at the interface, even though the states of stress and deformation will generally differ across the interface. For such an interface balance of linear momentum has the form

$$\operatorname{div}_s \mathbf{S} + (\mathbf{S}_\beta - \mathbf{S}_\alpha) \mathbf{n} = \rho \mathbf{v} (\mathbf{v}_\alpha - \mathbf{v}_\beta), \quad (\text{LM})$$

while the accretive laws for force and energy may be combined to form a single accretive balance law

$$\begin{aligned} \Psi_\beta - \Psi_\alpha &= (\mathbf{S}_\beta \mathbf{n}) \cdot (\mathbf{F}_\beta \mathbf{n}) - (\mathbf{S}_\alpha \mathbf{n}) \cdot (\mathbf{F}_\alpha \mathbf{n}) + \\ &\quad \frac{1}{2} \rho v^2 \{ |\mathbf{F}_\alpha \mathbf{n}|^2 - |\mathbf{F}_\beta \mathbf{n}|^2 \} + \\ &\quad \pi - \sigma \kappa - \operatorname{div}_s \mathbf{c} + (\mathbf{F}^T \mathbf{S}) \cdot \mathbf{L}. \end{aligned} \quad (\text{AB})$$

Here  $\alpha$  and  $\beta$  identify the two phases;  $\mathbf{S}$ ,  $\mathbf{v}$ ,  $\Psi$ , and  $\mathbf{F}$  (appropriately labelled) designate the bulk Piola-Kirchhoff stress, the bulk velocity, the bulk free energy, and the bulk deformation gradient;  $\rho$  is the common referential density of the two phases;  $\sigma$ ,  $\mathbf{S}$ ,  $\mathbf{c}$  and  $\pi$  are the surface tension, the interfacial Piola-Kirchhoff stress, the accretive shear, and the normal attachment force;  $\mathbf{n}$  is the outward unit normal to phase  $\alpha$ ;  $v$ ,  $\mathbf{L}$ ,  $\kappa$ , and  $\operatorname{div}_s$  are the normal velocity, the curvature tensor, twice the mean curvature, and the surface divergence for the interface.

The balance laws (LM) and (AB) are general relations, independent of the particular material under consideration. [GS] gives a thermodynamic argument in support of the interfacial

<sup>4</sup>[GS] also derives equations for a solid crystal in a liquid melt.

<sup>5</sup>Cf. Larche and Cahn [LC].

constitutive equations

$$\begin{aligned}\sigma &= \psi^{\wedge}(\mathbf{F}, \mathbf{n}), \\ \mathbf{S} &= \partial_{\mathbf{F}} \psi^{\wedge}(\mathbf{F}, \mathbf{n}), \\ \mathbf{c} &= -D_{\mathbf{n}} \psi^{\wedge}(\mathbf{F}, \mathbf{n}), \\ \pi &= \beta(\mathbf{F}, \mathbf{n}) \nu,\end{aligned}\tag{CE}$$

where  $\psi^{\wedge}(\mathbf{F}, \mathbf{n})$  is a constitutive function for the interfacial free energy,  $\mathbf{F}$  is the tangential deformation gradient,  $D_{\mathbf{n}}$  is the derivative with respect to  $\mathbf{n}$  following the interface, and  $\beta(\mathbf{F}, \mathbf{n}) \geq 0$  is a material function.

#### REFERENCES.

- [AG] Angenent, S. and M. E. Gurtin, Multiphase thermomechanics with interfacial structure. 2. Evolution of an isothermal interface. Arch. Rational Mech. Anal., forthcoming.
- [AJ] Alexander, J. I. D. and W. C. Johnson, Thermomechanical equilibrium in solid-fluid systems with curved interfaces. J. Appl. Phys. 58, 816-824 (1985).
- [G] Gurtin, M. E., Multiphase thermomechanics with interfacial structure. 1. Heat conduction and the capillary balance law. Arch. Rational Mech. Anal. 104, 195-221 (1988).
- [GS] Gurtin, M. E. and A. Struthers, Multiphase thermomechanics with interfacial structure. 2. On the force systems associated with crystal growth, forthcoming.
- [JA] Johnson, W. C. and J. I. D. Alexander, Interfacial conditions for thermomechanical equilibrium in two phase crystals. J. Appl. Phys. 59, 2735-2746 (1986).
- [LC] Larche, F. C. and J. W. Cahn, Thermomechanical equilibrium of multiphase solids under stress, Acta Met. 26, 1579-1589 (1978).
- [LC] Leo, P. H. and R. F. Sekerka, The effect of surface stress on crystal-melt and crystal-crystal equilibrium, forthcoming.

## NONLINEARITY OF INVERSE PROBLEMS

T. Mura and Z. Gao  
Department of Civil Engineering  
Northwestern University  
Evanston, IL 60208

### ABSTRACT

In this paper, we analyze the inverse problem in which residual surface displacements are used to evaluate nonelastic deformation in a domain, which is called the damage domain, of a solid. The problem is taken as an example to elucidate the nonlinearity of a class of inverse problems.

The problem can be formulated as a system of multi-dimensional Fredholm integral equations of the first kind. It is a complicated nonlinear problem since both damage domain (which appears as the domain of integration in the integral equation) and the nonelastic strains are unknown. The surface data are not sufficient to determine the shape of the damage domain and the exact distribution of the nonelastic strains. However, these data can be used to obtain some important characteristic quantities associated with the non-elastic deformation of the solid, such as elastic energy, stresses in certain region of the solid or the fracture toughness enhancement due to localized nonelastic deformation.

The research shows an interesting example of conversions between nonlinear and linear problems. By introducing the concept of equivalent damage domain, the general nonlinear problem is first converted into a linear one which is more tractable, but still ill-posed. A variational problem is then imposed. This leads to a new linear problem with a parameter determined by a nonlinear algebraic function. The payoff of the second conversion is the well-posedness (uniqueness and stability) of the new problem. This new problem is essentially a nonlinear problem again, but a much easier one compared with the original nonlinear problem. A numerical scheme is easily constructed due to the monotonic property of the nonlinear algebraic function.

## 1. Introduction

In recent years, inverse problems are becoming increasingly important in many scientific fields. Inverse scattering problems deal with the determination of the existence, locations and sizes of defects in mechanical structures by measurements of scattered ultrasonic wave. Increasing numbers of results, especially experimental ones, have been reported (e.g., Ogura, 1983). In the inverse problems of vibration, natural frequencies are used to reconstruct mass distribution of the structure (e.g., Gladwell, 1986). Intensive work has been done in this area, particularly for in-line discrete systems and one dimensional continuous systems, in which the corresponding mathematical problems are relatively simple and analytical results can be derived. Backus and Gilbert (1967, 1970, 1980) have studied the problem of determining the density distribution in the earth as well as wave velocities from observed travel time data, together with the known mass and moment of inertia of the earth, and the frequencies of certain normal modes of vibration.

This is a paper dealing with inverse problems in solid mechanics. Our objective is to characterize nonelastic deformation in bulk after a series of loadings by using only the residual surface displacements instead of the entire loading history. The residual surface displacements are relative and are defined as the difference of the initial and final values of the displacements.

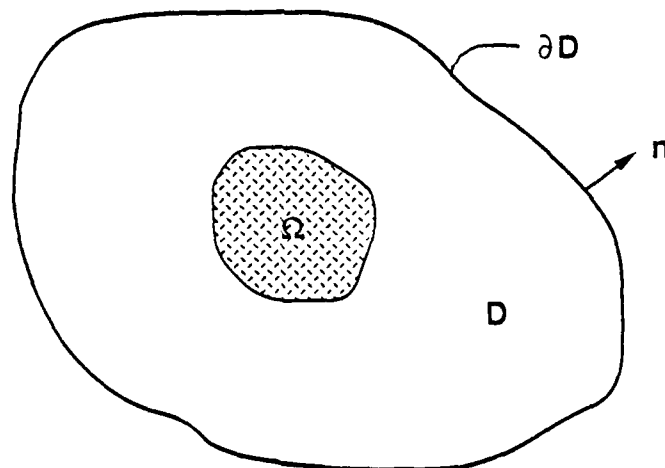


Fig. 1. - A traction free body  $D$  with a sub-region where residual nonelastic strains are accumulated.



Suppose that nonelastic strains  $\epsilon_{ij}^P$  are caused in a subdomain  $\Omega$  (damage domain) of a given body  $D$  after a series of unknown loadings (Fig. 1). The integral equation relating  $\epsilon_{ij}^P$  to the residual displacements  $u_i$  is written as (Gao and Mura, 1989).

$$\int_{\Omega} C_{ijkl} G_{km,\ell}(\underline{x} - \underline{x}') \epsilon_{ij}^P(\underline{x}) d\underline{x} - \int_{\partial D} C_{ijkl} G_{km,\ell}(\underline{x} - \underline{x}') u_i(\underline{x}) n_j ds + \frac{1}{2} u_m(\underline{x}') \quad (1)$$

$\underline{x}' \in \partial D$

where  $\partial D$  is the boundary of  $D$ ;  $n_j$  is the outer normal of  $\partial D$ ;  $C_{ijkl}$  is the elastic modulus tensor of the material and  $G_{km}(\underline{x} - \underline{x}')$  is the Green's function for an infinite elastic medium, i.e.,  $G_{km}(\underline{x} - \underline{x}')$  is the displacement at point  $\underline{x}$  in the  $x_k$  direction due to a unit force at point  $\underline{x}'$  in the  $x'_m$  direction.  $G_{km,\ell}(\underline{x} - \underline{x}')$  represents  $(\partial/\partial x_\ell)G_{km}(\underline{x} - \underline{x}')$ .

Equation (1) can be obtained by using the Betti's reciprocal theorem. We refer readers to Gao and Mura (1989) for detailed derivation.

## 2. Uniqueness

Our objective is to determine nonelastic strains  $\epsilon_{ij}^P$  and the domain  $\Omega$  (a nonlinear problem). However, neither of these two quantities can be obtained from equation (1).

Let  $\Omega^*$  be a domain inside the body  $D$ . When a distribution of nonelastic strains is compatible in  $\Omega^*$ , the remainder  $D - \Omega^*$  is not disturbed. Hence, the displacements and stresses in  $D - \Omega^*$  vanish. This implies that the homogeneous equation of (1) has nonzero solution  $\epsilon_{ij}^P(\underline{x})$  for arbitrarily chosen domain  $\Omega^*$ . It is then clear that  $\epsilon_{ij}^P$  and  $\Omega$  cannot be determined uniquely from equation (1).

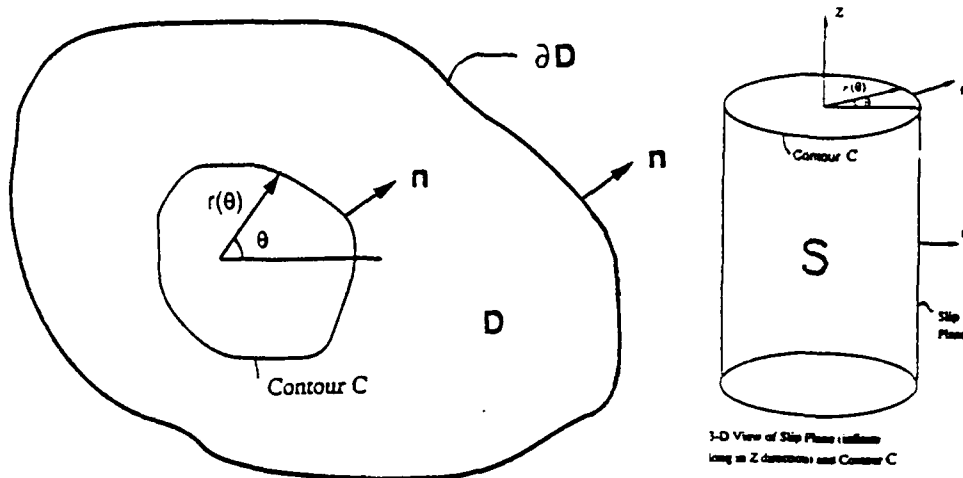


Fig. 2. - A two-dimensional body with a line defect (or dislocation loop).

Consider a two-dimensional case. Equation (1) has the unique solution when  $\Omega$  is a contour C (Fig. 2). One interpretation of this case is that C is a dislocation loop. The Somigliano's dislocation density  $\underline{b}$  yields nonelastic strains, defined on C,

$$\epsilon_{ij}^p = \frac{1}{2} (b_i n_j + b_j n_i)$$

Let the equation of the closed contour C be  $r = r(\theta)$ . Equation (1) is then changed to

$$\int_0^{2\pi} C_{ijkl} G_{km,l}(\underline{x} - \underline{x}') \Big|_{\substack{x_1 = r(\theta)\sin\theta \\ x_2 = r(\theta)\cos\theta}} \epsilon_{ij}^p(\theta) r(\theta) d\theta - \int_{\partial D} C_{ijkl} G_{km,l}(\underline{x} - \underline{x}') u_i(\underline{x}) n_j ds + \frac{1}{2} u_m(\underline{x}') \quad (2)$$

$$\underline{x}' \in \partial D.$$

$\epsilon_{ij}^p(\theta)$ , as well as function  $r(\theta)$  (shape of contour C) are determined uniquely from the surface displacements. The reason for the uniqueness is as follows.

We have shown (Gao and Mura, 1989) that the displacement field, of the points not belonging to  $\Omega$ , are uniquely determined from residual surface displacements. Therefore, if rigid body motion is properly excluded, the displacements inside and outside contour  $C$ , denoted by  $u_i^{(I)}$  and  $u_i^{(O)}$  respectively, are determined by the surface displacements.  $\epsilon_{ij}^P(\theta)$  is then deduced from the mismatch of  $u_i^{(I)}$  and  $u_i^{(O)}$  on the contour  $C$ . Hence, equation (2) has the unique solution for  $\epsilon_{ij}^P(\theta)$  and  $r(\theta)$ .

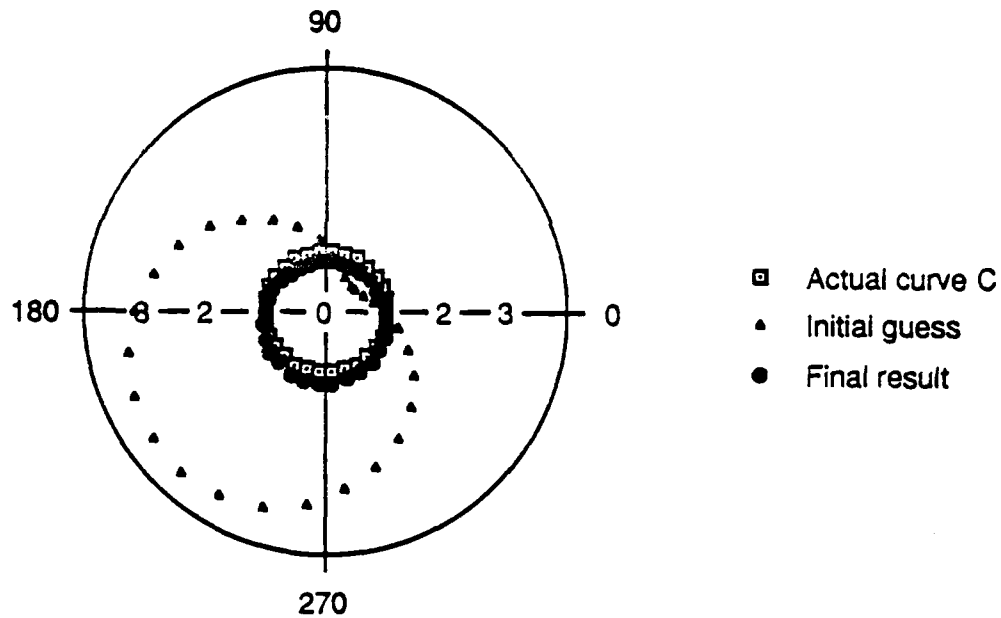


Fig. 3. - The initial and final configurations of iterations compared to the actual shape of contour  $C$ . The actual values of nonelastic strains are  $\epsilon_{ij}^P = 1$  while the computed values are  $\epsilon_{11}^P = 0.97$ ,  $\epsilon_{22}^P = 0.97$ ,  $\epsilon_{12}^P = 0.9$ .

An example is shown in Fig. 3, in which we expand  $r = r(\theta)$  by Fourier series

$$r(\theta) = \sum_{n=0}^{\infty} (a_n \sin n\theta + b_n \cos n\theta)$$

$$= b_0 + a_1 \sin\theta + b_1 \cos\theta. \quad (3)$$

If we assume  $\epsilon_{ij}^p$  are constants, equation (2) becomes a nonlinear equation for  $\epsilon_{11}^p$ ,  $\epsilon_{12}^p$ ,  $\epsilon_{22}^p$ ,  $b_0$ ,  $a_1$  and  $b_1$ . Choosing an initial configuration of  $r(\theta)$ , the nonlinear equation is solved by an optimization algorithm (Subroutine ZXMIN in IMSL Library) which minimizes the difference between the right and left hand sides of (2). The initial and final configurations of  $r = r(\theta)$  are compared to the actual shape of  $r(\theta)$  (Fig. 3). The actual values of nonelastic strains are  $\epsilon_{ij}^p = 1$  while the computed ones are  $\epsilon_{11}^p = 0.97$ ,  $\epsilon_{22}^p = 0.97$  and  $\epsilon_{12}^p = 0.9$ , respectively.

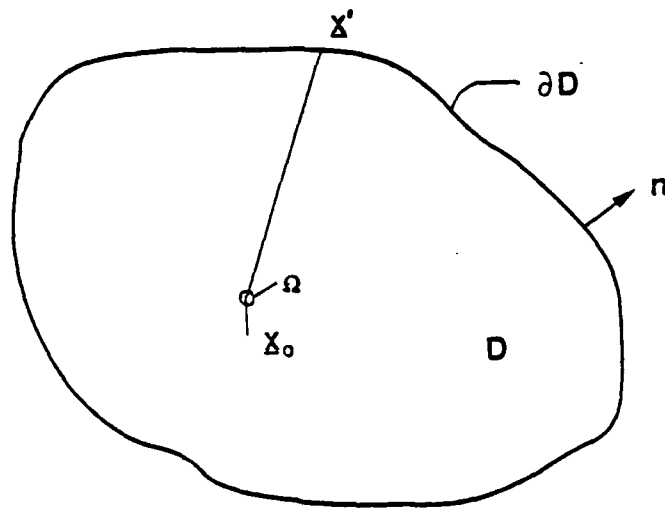


Fig. 4. - A body with a point defect  $\Omega$ .  $\underline{x} - \underline{x}' = \underline{x}_0 - \underline{x}'$   
for  $\underline{x} \in \Omega$  and  $\underline{x}' \in \partial D$ .

Another interesting case is as follows. When domain  $\Omega$  is small and far away from surface  $\partial D$ ,  $\Omega$  can be treated as a point defect (Fig. 4). Note  $\underline{x} - \underline{x}' \approx \underline{x}_0 - \underline{x}'$  for  $\underline{x}' \in \partial D$ ,  $\underline{x} \in \Omega$  and a fixed point  $\underline{x}_0$  inside  $\Omega$ . Equation (1) can be simplified as

$$C_{ijkl} G_{km,l} (\underline{x}_0 - \underline{x}') \int_{\Omega} \epsilon_{ij}^p (\underline{x}) d\underline{x} - \int_{\partial D} C_{ijkl} G_{km,l} (\underline{x} - \underline{x}') u_i (\underline{x}) n_j ds + \frac{1}{2} u_m (\underline{x}') \quad (4)$$

$$\underline{x}' \in \partial D.$$

The location of the defect  $\underline{x}_0$ , as well as the quantity  $\int_{\Omega} \epsilon_{ij}^p (\underline{x}) d\underline{x}$  can be obtained by employing a proper algorithm for the nonlinear problem (4).

### 3. From the Nonlinear Problem to a Linear Problem

As we have mentioned, in general, the nonlinear problem (1) to determine  $\Omega$  and  $\epsilon_{ij}^p$  cannot be solved uniquely. Even for certain specific problems (e.g., equations (2) and (4)), where the uniqueness is guaranteed, the construction of a proper algorithm is still a difficult task. On the other hand, the degree of difficulty will be greatly reduced if the nonlinear problem is converted into a linear one.

Problem (1) becomes a linear problem when domain  $\Omega$  is specified. The question is how to specify  $\Omega$  since we really do not know its shape and location.

Choose a domain  $\Omega^*$  (equivalent damage domain) such that  $\Omega$  is contained inside  $\Omega^*$ . Equation (1) is changed to

$$\int_{\Omega^*} C_{ijkl} G_{km,l} (\underline{x} - \underline{x}') \epsilon_{ij}^p (\underline{x}) d\underline{x} - \int_{\partial D} C_{ijkl} G_{km,l} (\underline{x} - \underline{x}') u_i (\underline{x}) n_j d\underline{x} + \frac{1}{2} u_m (\underline{x}') \quad (5)$$

$$\underline{x}' \in \partial D$$

which is a linear problem to determine  $\epsilon_{ij}^P$ . Now we discuss the relationship between the solutions of nonlinear problem (1) and linear problem (5).

Conclusion 1. The stress field caused by both solutions are identical in the region outside  $\Omega^*$ .

Conclusion 2. The minimum elastic energy (or any other quadratic function of  $\epsilon_{ij}^P$ ) of all the  $\epsilon_{ij}^P$  satisfying (5) is a lower bound of that of actual nonelastic strains satisfying (1).

The above conclusions are based on the fact that if surface displacements and traction forces are zero on a part of the boundary of an elastic body, the displacements and stresses are identically zero in the whole elastic body. The details of the discussion can be found in Gao and Mura (1989). The same idea applies to the problem of calculating the shielding effects due to an unknown distribution of micro-defects in an unknown domain  $\Omega$  by measuring the crack opening displacements (Gao and Mura, 1990; also see Fig. 5).

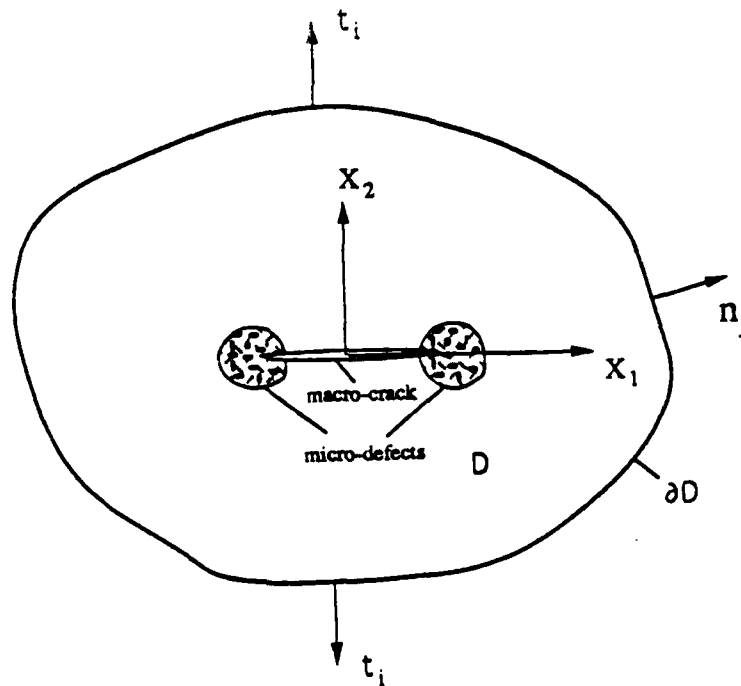


Fig. 5. - An infinite medium with a crack. The shielding effects of the micro-defects can be calculated from measurements of crack opening displacements.

#### 4. From the Linear Problem to a New Nonlinear Problem

By specifying domain  $\Omega$  as  $\Omega^*$ , we have changed the nonlinear problem (1) to a linear equation (5). The solution of the linear problem (5) preserves important characteristics of the actual nonelastic strains. However, equation (5) is still an ill-posed problem (nonunique and unstable), which cannot be solved directly.

Let's write equation (5) as

$$\underline{U}(\underline{x}') = \int_{\Omega^*} \underline{K}(\underline{x}, \underline{x}') \underline{V}(\underline{x}) d\underline{x} \quad \text{for } \underline{x}' \text{ on } \partial D \quad (6)$$

where  $\underline{U}(\underline{x}')$  is known since  $u_i(\underline{x})$  is given on  $\partial D$ .  $\underline{K}(\underline{x}, \underline{x}') = C_{ijkl} G_{km,l}(\underline{x}-\underline{x}')$  and  $\underline{V}(\underline{x})$  is an unknown vector whose components are  $\epsilon_{ij}^p$ .

Now consider a variational problem

$$\text{Min } || \underline{V}(\underline{x}) ||^2 \quad (7)$$

$$\text{subjected to } || \int_{\Omega^*} \underline{K}(\underline{x}, \underline{x}') \underline{V}(\underline{x}) d\underline{x} - \underline{U}(\underline{x}') ||^2 = \epsilon$$

where  $\epsilon$  is a small number chosen from the accuracy of measurement and

$$|| \underline{V}(\underline{x}) ||^2 = \int_{\Omega^*} \underline{V}^T(\underline{x}) \underline{V}(\underline{x}) d\underline{x} \quad (8)$$

$\underline{V}^T$  is the transpose vector of  $\underline{V}$ .

The use of a Lagrange multiplier  $\lambda$  transforms (7) to

$$\text{Min } \{ || \underline{V}(\underline{x}) ||^2 + \lambda ( || \int_{\Omega^*} \underline{K}(\underline{x}, \underline{x}') \underline{V}(\underline{x}) d\underline{x} - \underline{U}(\underline{x}') ||^2 - \epsilon ) \}. \quad (9)$$

The Euler equation of (9) becomes

$$\int_{\Omega^*} \underline{K}^*(\underline{x}, \underline{y}) \underline{V}(\underline{x}) d\underline{x} + \alpha \underline{V}(\underline{y}) = \underline{U}^*(\underline{y}), \quad \text{for } \underline{y} \in \Omega^* \quad (10)$$

where

$$\alpha = 1/\lambda$$

$$\underline{K}^*(\underline{x}, \underline{y}) = \int_{\partial D} \underline{K}^T(\underline{y}, \underline{x}') \underline{K}(\underline{x}, \underline{x}') d\underline{x}' \quad (11)$$

$$\underline{U}^*(\underline{y}) = \int_{\partial D} \underline{K}^T(\underline{y}, \underline{x}') \underline{U}(\underline{x}') d\underline{x}' .$$

The integral equation (10) is solved for  $\underline{V}(\underline{x})$  with parameter  $\alpha$ . The value of  $\alpha$  is determined from

$$f(\alpha) = \left\| \int_{\Omega^*} \underline{K}(\underline{x}, \underline{x}') \underline{V}(\underline{x}) d\underline{x} - \underline{U}(\underline{x}') \right\|^2 - \epsilon = 0. \quad (12)$$

Equation (10) is a well-posed Fredholm integral equation of the second kind with a self-adjoint kernel. For any chosen parameter  $\alpha$ , equation (10) can be solved by employing conventional techniques such as finite element method. However, the parameter  $\alpha$  must satisfy the nonlinear algebraic equation (12). Therefore, the new problem is essentially a nonlinear problem again, but a much easier one compared with the original nonlinear problem (1).

The nonlinear function  $f(\alpha)$  is an increasing function of  $\alpha$  and has only one root (Gao and Mura, 1989). Therefore, the root can be solved by the bisection algorithm. The algorithm converges rather fast since in each iteration, the interval containing the root of  $f(\alpha)$  is reduced by half.

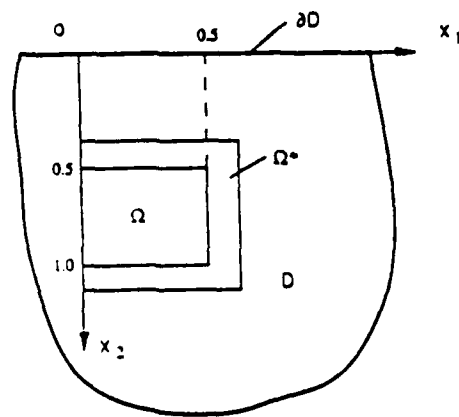


Fig. 6. - Even though  $\Omega$  (where nonelastic strains are distributed) is unknown, it is always possible to cover  $\Omega$  with a chosen domain  $\Omega^*$ . The original nonlinear problem is, therefore, changed into a linear problem.



Let us consider the example shown in Fig. 6. Two dimensional half space is given by  $x_2 \geq 0$  and the nonelastic strains distributed inside  $\Omega$  are

$$\gamma_{12}^P = 2 \epsilon_{12}^P = 28 (x_1 - 0.6) (x_2 - 1.0)$$

$$\epsilon_{22}^P = - \epsilon_{11}^P = 20 (x_2 - 1.7) \cos 2x_1. \quad (13)$$

Table 1

$\Omega^*$	$\ V\ $	$\gamma$	Element Number	Element Length
$0 \leq x_1 \leq 0.5;$				$L[x_1] : 0.1$
$0.5 \leq x_2 \leq 1.$	8.34	1.0	25	$L[x_2] : 0.1$
$0 \leq x_1 \leq 0.54;$				$L[x_1] : 0.09$
$0.48 \leq x_2 \leq 1.02$	8.10	1.17	36	$L[x_2] : 0.09$
$0 \leq x_1 \leq 0.616;$				$L[x_1] : 0.088$
$0.48 \leq x_2 \leq 1.096$	7.98	1.52	49	$L[x_2] : 0.088$
$0 \leq x_1 \leq 0.744;$				$L[x_1] : 0.093$
$0.47 \leq x_2 \leq 1.27$	7.44	2.38	64	$L[x_2] : 0.1$

$\Omega = \{0 \leq x_1 \leq 0.5; 0.5 \leq x_2 \leq 1\}$ . The exact value of  $\|V\|$  is 8.33.

$$\gamma = \text{area of } \Omega^* / \text{area of } \Omega$$

Table 1 lists some characteristic quantities for the calculations of different choices of domain  $\Omega^*$ . For instance, from the second row of Table 1 we know  $\Omega^*$  is taken as  $0 \leq x_1 \leq 0.54$ ,  $0.48 \leq x_2 \leq 1.02$  to cover  $\Omega$ . 36 elements are used and the element lengths in both directions are 0.09. The computed value of  $||V(\underline{x})||$  is 8.1 which is a lower bound of the actual value 8.33.

It should be mentioned that if we replace  $||V(\underline{x})||^2$  in equation (7) by the elastic strain energy and change (10) properly, we can obtain a lower bound of the elastic strain energy.

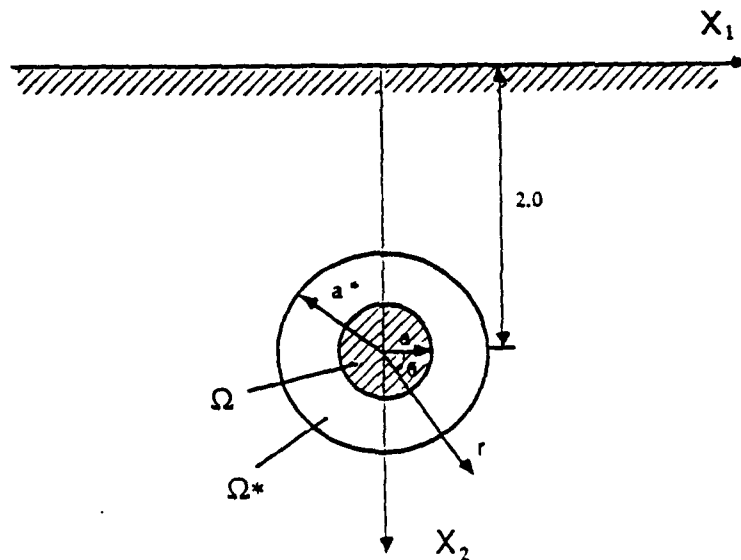


Fig. 7. -  $\Omega$  is the circular domain in a half-space  $x_2 \geq 0$ .  $\Omega^*$  covers  $\Omega$ .

Another example is shown in Fig. 7. The nonelastic strains

$$\epsilon_{12}^p = \frac{1}{2}, \quad \epsilon_{22}^p = -\epsilon_{11}^p = 1 \quad (14)$$

occur in domain  $\Omega$  ( $r \leq 0.2$ ). Domain  $\Omega^*$  is chosen to include  $\Omega$  inside. By solving equations (10) and (12), we obtain a distribution of  $\epsilon_{ij}^P$  in  $\Omega^*$ , which is different from that in (14). However, as indicated by Fig. 8, in the region outside  $\Omega^*$  the stress field induced by the computed  $\epsilon_{ij}^P$  in  $\Omega^*$  is the same as the one by (14).

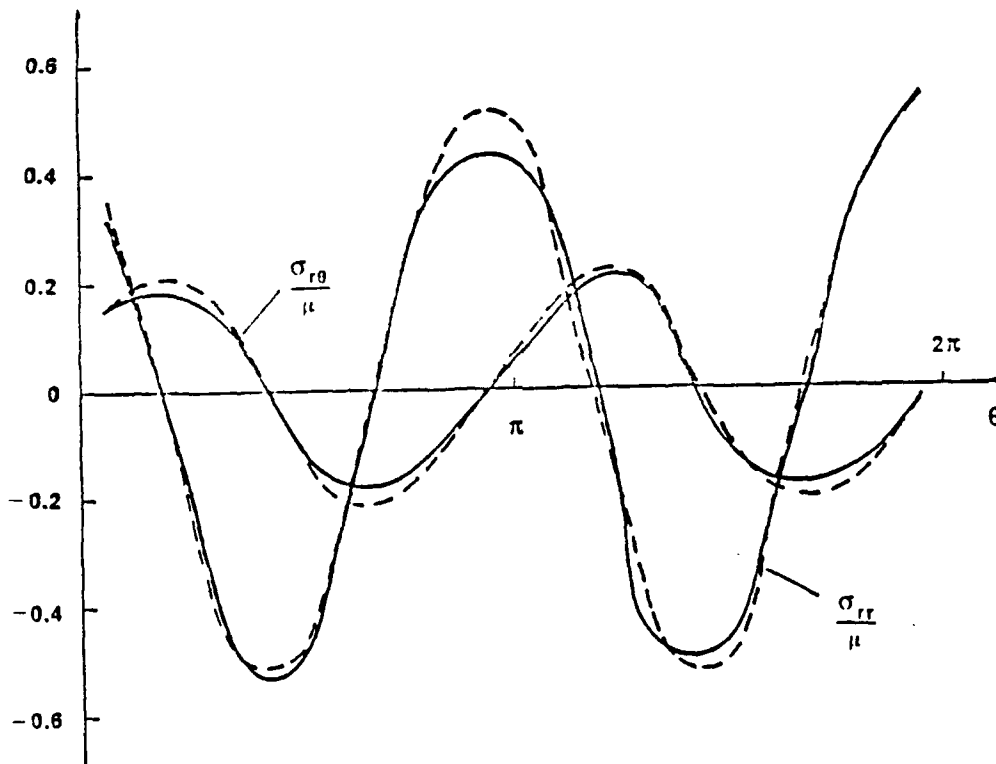


Fig. 8. - The residual stresses  $\sigma_{rr}$  and  $\sigma_{r\theta}$  at  $r/a^* = 1.5$  caused by (14) are identical to those by the computed  $\epsilon_{ij}^P$  in  $\Omega^*$ . The solid lines are the computed results and the dashed lines are induced by  $\epsilon_{ij}^P$  in (14).

### Acknowledgment

This research was supported under U. S. Army Research Office Contract No. DAAL03-88-C-0027 through a subcontract with Rockwell International Science Center.

### References

- Backus, G. E. and Gilbert, J. F. (1967). Numerical applications of a formalism for geophysical inverse problems, *Geophy. J. R. Astr. Soc.* 13.
- Backus, G. E. and Gilbert, J. F. (1970). Uniqueness in the inversion of inaccurate gross earth data, *Phil. Trans, Roy. Soc.* 266.
- Backus, G. E. and Gilbert, J. F. (1986). The resolving power of gross earth data, *Geophy. J. R. Astr. Soc.* 16.
- Gao, Z. and Mura, T. (1989). On the inversion of residual stresses from surface displacements, *J. Appl. Mech*, Vol. 56, No. 3.
- Gao, Z. and Mura, T. (1990). IUTAM Symposium on inelastic deformation of composite materials (submitted).
- Gladwell, G. M. L. (1986). *Inverse problems in vibration*, Martinus Nijhoff, Dordrecht.
- Ogura, Y. (1983). Height determination studies for planar defects by means of ultrasonic testing. *The Non Destructive Testing Journal, Japan.* Vol. 1, No. 1.

# QUADRATIC DYNAMICAL SYSTEMS DESCRIBING SHEAR FLOW OF NON-NEWTONIAN FLUIDS \*

D. S. Malkus<sup>1</sup>, J. A. Nohel<sup>2</sup>, and B. J. Plohr<sup>3</sup>

Center for the Mathematical Sciences  
University of Wisconsin-Madison  
Madison, WI 53705

## Abstract

Phase-plane techniques are used to analyze a quadratic system of ordinary differential equations that approximates a single relaxation-time system of partial differential equations used to model transient behavior of highly elastic non-Newtonian liquids in shear flow through slit dies. The latter one-dimensional model is derived from three-dimensional balance laws coupled with differential constitutive relations well-known by rheologists. The resulting initial-boundary-value problem is globally well-posed and possesses the key feature: the steady shear stress is a non-monotone function of the strain rate. Results of the global analysis of the quadratic system of ode's lead to the same qualitative features as those obtained recently by numerical simulation of the governing pde's for realistic data for polymer melts used in rheological experiments. The analytical results provide an explanation of the experimentally observed phenomenon called spurt; they also predict new phenomena discovered in the numerical simulation; these phenomena should also be observable in experiments.

---

\* 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.

<sup>1</sup> Also Department of Engineering Mechanics.

<sup>2</sup> Also Department of Mathematics.

<sup>3</sup> Also Computer Sciences Department.

## 1. Introduction

The purpose of this paper is to analyze novel phenomena in dynamic shearing flows of non-Newtonian fluids that are important in polymer processing [17]. One striking phenomenon, called "spurt," was apparently first observed by Vinogradov *et al.* [19] in experiments concerning quasi static flow of monodisperse polyisoprenes through capillaries or equivalently through slit dies. They found that the volumetric flow rate increased dramatically at a critical stress that was independent of molecular weight. Until recently, spurt has been associated with the failure of the flowing polymer to adhere to the wall [5]. The focus of our current research is to offer an alternate explanation of spurt and related phenomena.

Understanding these phenomena has proved to be of significant physical, mathematical, and computational interest. In our recent work [12], we found that satisfactory explanation and modeling of the spurt phenomenon requires studying the full dynamics of the equations of motion and constitutive equations. The common and key feature of constitutive models that exhibit spurt and related phenomena is a non-monotonic relation between the steady shear stress and strain rate. This allows jumps in the steady strain rate to form when the driving pressure gradient exceeds a critical value; such jumps correspond to the sudden increase in volumetric flow rate observed in the experiments of Vinogradov *et al.* The governing systems used to model such one-dimensional flows are analyzed in [12] by numerical techniques and simulation, and in the present work by analytical methods. The systems derive from fully three-dimensional differential constitutive relations with relaxation times (based on work of Johnson and Segalman [8] and Oldroyd [16]). They are evolutionary, globally well posed in a sense described below, and they possess discontinuous steady states of the type mentioned above that lead to an explanation of spurt. The governing systems for shear flows through slit-dies are formulated from balance laws in Sec. 2.

Specifically, we model these flows by decomposing the total shear stress into a polymer contribution, evolving in accordance with a differential constitutive relation with a single relaxation time and a Newtonian viscosity contribution (see system ( $JSO$ ) in Sec. 2.). The flows can also be modelled by a system based on a differential constitutive law with two widely spaced relaxation times (see system ( $JSO_2$ ) in [13].) but no Newtonian viscosity contribution. Numerical simulation [9, 12] of transient flows at high Weissenberg (Deborah) number and very low Reynolds number using the model ( $JSO$ ) exhibited spurt, shape memory, and hysteresis; furthermore, it predicted other effects, such as latency, normal stress oscillations, and molecular weight dependence of hysteresis, that should be analysed further and tested in rheological experiment.

In earlier work, Hunter and Slemrod [7] used techniques of conservation laws to study the qualitative behavior of discontinuous steady states in a simple one-dimensional viscoelastic model of rate type with viscous damping. They predicted shape memory and hysteresis effects related to spurt. A salient feature of their model is linear instability and loss of evolutionarity in a certain region of state space.

The objective of the present paper is to develop analytical techniques, the results of which verify these rather dramatic implications of numerical simulation. Based on scaling introduced in [12], appropriate for the highly elastic and very viscous polyisoprenes used in

the spurt-experiment, we are led to study the following pair of quadratic autonomous ordinary differential equations that approximates the governing system (*JSO*) in the relevant range of physical parameters for each fixed position in the channel:

$$\begin{aligned}\dot{\sigma} &= (Z + 1) \left( \frac{\bar{T} - \sigma}{\varepsilon} \right) - \sigma, \\ \dot{Z} &= -\sigma \left( \frac{\bar{T} - \sigma}{\varepsilon} \right) - Z.\end{aligned}\tag{1.1}$$

Here the dot denotes the derivative  $d/dt$ ,  $\bar{T}$  is a parameter that depends on the driving pressure gradient as well as position  $x$  in the channel, and  $\varepsilon > 0$  is a ratio of viscosities. System (1.1) is obtained by setting  $\alpha = 0$  in the momentum equation in system (*JSO*); this approximation is reasonable because  $\alpha$  is at least several orders of magnitude smaller than  $\varepsilon$ . We show that steady states of system (*JSO*), some of which are discontinuous for non-monotone constitutive relations, correspond to critical points of the quadratic system. We deduce the local characters of the critical points, and we prove that system (1.1) has no periodic orbits or closed separatrix cycles. Moreover, this system is endowed with a natural Lyapunov-like function with the aid of which we are able to determine the global dynamics of the approximating quadratic system completely and thus identify its globally asymptotically stable critical points (i.e. steady states) for each position  $x$ . This analysis is carried out in Sec. 3. When  $\alpha$ , the ratio of Reynolds to Deborah numbers, is strictly positive, the stability of discontinuous steady states of system (*JSO*) remains to be settled. Recently, Nohel, Pego and Tzavaras [15] established such a result for simple model in which the polymer contribution to the shear stress satisfies a single differential constitutive relation; for a particular choice, their model and system (*JSO*) with  $\alpha > 0$  have the same behavior in steady shear. Their asymptotic stability result, combined with numerical experiments and research in progress, suggest that the same result holds for the full system (*JSO*), at least when  $\alpha$  is sufficiently small.

In Sec. 4., the analysis of Sec. 3. is applied to each point  $x$  in the channel, allowing us to explain spurt, shape memory, hysteresis, and other effects originally observed in the numerical simulations in terms of a continuum of phase portraits. We discuss asymptotic expansions of solutions of systems (*JSO*) and (*JSO*<sub>2</sub>) of Ref. [13] in powers of  $\varepsilon$  that enable us to explain latency (a pseudo-steady state that precedes spurt). The asymptotic analysis also permits a more quantitative comparison of the dynamics of the two models when  $\varepsilon$  is sufficiently small. In Sec. 5., we discuss physical implications of the analysis, particularly those that suggest new experiments. In Sec. 6., we draw certain conclusions. Although the analysis in this paper applies only to the special constitutive models we have studied, we expect that the qualitative features of our results appear in a broad class of non-Newtonian fluids. Indeed, numerical simulation by Kolkka and Lerley [10] using another model with a single relaxation time and Newtonian viscosity exhibits very similar character.

## 2. A Johnson-Segalman-Oldroyd Model for Shear Flow

The motion of a fluid under incompressible and isothermal conditions is governed by the balance of mass and linear momentum. 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. [2] for a survey. Of particular interest is a class of differential models with  $m$ -relaxation times, derived in a three-dimensional setting in Refs. [12] and [13]; these models can be regarded as a special cases of the Johnson-Segalman model [8] when the memory function is a linear combination of  $m$ -decaying exponentials with positive coefficients or of the Oldroyd differential constitutive equation [16].

Essential properties of constitutive relations are exhibited in simple planar Poiseuille shear flow. We study shear flow of a non-Newtonian fluid between parallel plates, located at  $x = \pm h/2$ , with the flow aligned along the  $y$ -axis, symmetric about the center line, and driven by a constant pressure gradient  $\bar{f}$ . We restrict attention to the simplest model of a single relaxation-time differential model that possesses steady state solutions exhibiting a non-monotone relation between the total steady shear stress and strain rate, and thereby reproduces spurt and related phenomena discussed below. The total shear stress  $T$  is decomposed into a polymer contribution and a Newtonian viscosity contribution. When restricted to one space dimension the initial-boundary value problem, in non-dimensional units with distance scaled by  $h$ , governing the flow can be written in the form (see Refs. [9, 12]):

$$\begin{aligned} \alpha v_t - \sigma_x &= \varepsilon v_{xx} + \bar{f}, \\ \sigma_t - (Z + 1)v_x &= -\sigma, \\ Z_t + \sigma v_x &= -Z \end{aligned} \quad (JSO)$$

on the interval  $[-1/2, 0]$ , with boundary conditions

$$v(-1/2, t) = 0 \quad \text{and} \quad v_x(0, t) = 0 \quad (BC)$$

and initial conditions

$$v(x, 0) = v_0(x), \quad \sigma(x, 0) = \sigma_0(x), \quad \text{and} \quad Z(x, 0) = Z_0(x), \quad \text{on} \quad -1/2 \leq x \leq 0; \quad (IC)$$

symmetry of the flow and compatibility with the boundary conditions requires that  $v_0(-1/2) = 0$ ,  $v_0'(0) = 0$  and  $\sigma_0(0) = 0$ .

The evolution of  $\sigma$ , the polymer contribution to the shear stress, and of  $Z$ , a quantity proportional to the normal stress difference, are governed by the second and third equations in system (JSO). As a result of scaling motivated by numerical simulation and introduced in Ref. [12], there are only three essential parameters:  $\alpha$  is a ratio of Reynolds number to Deborah number,  $\varepsilon$  is a ratio of viscosities, and  $f$  is the constant pressure gradient.

When  $\varepsilon = 0$ , and  $Z + 1 \geq 0$ , system (JSO) is hyperbolic, with characteristics speeds  $\pm[(Z + 1)/\alpha]^{1/2}$  and 0. Moreover, for smooth initial data in the hyperbolic region and compatible with the boundary conditions, techniques in [18] can be used to establish



global well-posedness (in terms of classical solutions) if the data are small, and finite-time blow-up of classical solutions if the data are large. If  $\varepsilon > 0$ , system (*JSO*) for any smooth or piece-wise smooth data; indeed, general theory developed in [15] (see Sec. 3 and particularly Appendix A) yields global existence of classical solutions for smooth initial data of arbitrary size, and also existence of almost classical, strong solutions with discontinuities in the initial velocity gradient and in stress components; the latter result allows one to prescribe discontinuous initial data of the same type as the discontinuous steady states studied in this paper.

The steady-state solutions of system (*JSO*) play an important role in our discussion. Such a solution, denoted by  $\bar{v}$ ,  $\bar{\sigma}$ , and  $\bar{Z}$ , can be described as follows. The stress components  $\bar{\sigma}$  and  $\bar{Z}$  are related to the strain rate  $\bar{v}_x$  through the relations

$$\bar{\sigma} = \frac{\bar{v}_x}{1 + \bar{v}_x^2}, \quad \bar{Z} + 1 = \frac{1}{1 + \bar{v}_x^2}. \quad (2.1)$$

Therefore, the steady total shear stress  $\bar{T} := \bar{\sigma} + \varepsilon \bar{v}_x$  is given by  $\bar{T} = w(\bar{v}_x)$ , where

$$w(s) := \frac{s}{1 + s^2} + \varepsilon s. \quad (2.2)$$

The properties of  $w$ , the steady-state relation between shear stress and shear strain rate, are crucial to the behavior of the flow. By symmetry, it suffices to consider  $s \geq 0$ . For all  $\varepsilon > 0$ , the function  $w$  has inflection points at  $s = 0$  and  $s = \sqrt{3}$ . When  $\varepsilon > 1/8$ , the function  $w$  is strictly increasing, but when  $\varepsilon < 1/8$ , the function  $w$  is not monotone. Lack of monotonicity is the fundamental cause of the non-Newtonian behavior studied in this paper; hereafter we assume that  $\varepsilon < 1/8$ .

The graph of  $w$  is shown in Fig. 1. Specifically,  $w$  has a maximum at  $s = s_M$  and a minimum at  $s = s_m$ , where it takes the values  $\bar{T}_M := w(s_M)$  and  $\bar{T}_m := w(s_m)$  respectively. As  $\varepsilon \rightarrow 1/8$ , the two critical points coalesce at  $s = \sqrt{3}$ .

The momentum equation, together with the boundary condition at the centerline, implies that the steady total shear stress satisfies  $\bar{T} = -\bar{f}x$  for every  $x \in [-\frac{1}{2}, 0]$ . Therefore, the steady velocity gradient can be determined as a function of  $x$  by solving

$$w(\bar{v}_x) = -\bar{f}x. \quad (2.3)$$

Equivalently, a steady state solution  $\bar{v}_x$  satisfies the cubic equation  $P(\bar{v}_x) = 0$ , where

$$P(s) := \varepsilon s^3 - \bar{T}s^2 + (1 + \varepsilon)s - \bar{T}. \quad (2.4)$$

The steady velocity profile in Fig. 2 is obtained by integrating  $\bar{v}_x$  and using the boundary condition at the wall. However, because the function  $w$  is not monotone, there might be up to three distinct values of  $\bar{v}_x$  that satisfy Eq. (2.3) for any particular  $x$  on the interval  $[-1/2, 0]$ . Consequently,  $\bar{v}_x$  can suffer jump discontinuities, resulting in kinks in the velocity profile (as at the point  $x_*$  in Fig. 2). Indeed, a steady solution must contain such a jump if the total stress  $\bar{T}_{\text{wall}} = \bar{f}/2$  at the wall exceeds the total stress  $\bar{T}_M$  at the local maximum  $M$  in Fig. 1.

Finally, we remark that the flow problem discussed here can also be modelled by a system based on a differential constitutive law with two widely spaced relaxation times but no Newtonian viscosity contribution (see system (*JSO*<sub>2</sub>) in Sec. 2. of [13]); with an appropriate choice of relevant parameters, the resulting problem exhibits the same steady states and the same characteristics as (*JSO*).

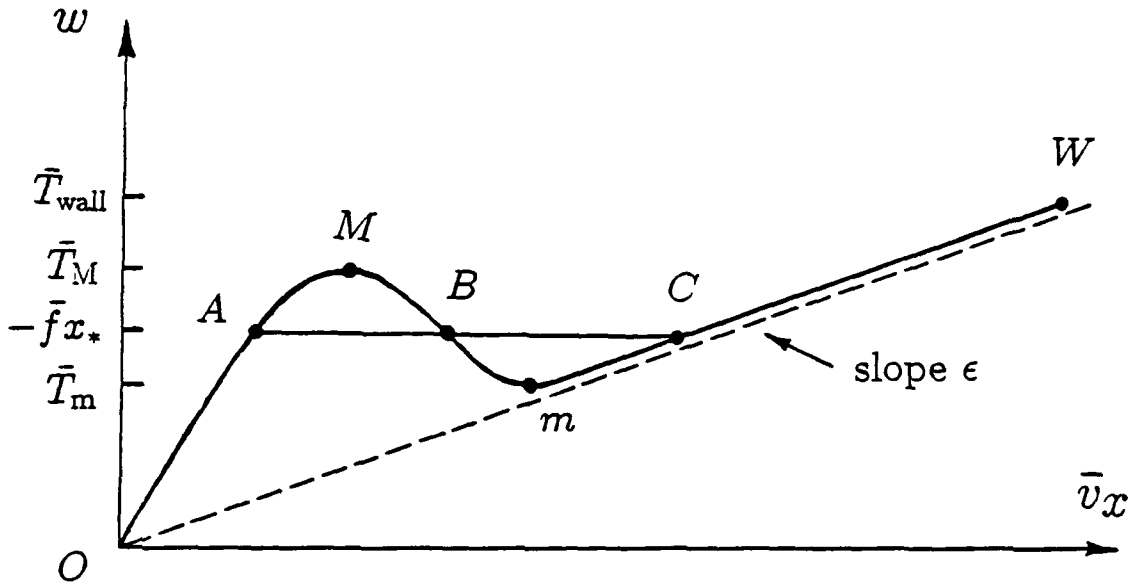


Fig. 1: Total steady shear stress  $\bar{T}$  vs. shear strain rate  $\bar{v}_x$  for steady flow. The case of three critical points is illustrated; other possibilities are discussed in Sec. 3.

### 3. Phase Plane Analysis for System ( $JSO$ ) When $\alpha = 0$

When  $\alpha$  is not zero, numerical simulation developed in [9, 11, 12] discovered striking phenomena in shear flow and suggested the analysis that follows. A great deal of information about the structure of solutions of system ( $JSO$ ) can be garnered by studying a quadratic system of ordinary differential equations that approximates it in a certain parameter range, the dynamics of which is determined completely. Motivation for this approximation comes from the following observation: in experiments of Vinogradov *et al.* [19],  $\alpha$  is of the order  $10^{-12}$ ; thus the term  $\alpha v_t$  in the momentum equation of system ( $JSO$ ) is negligible even when  $v_t$  is moderately large. This led us to the approximation to system ( $JSO$ ) obtained when  $\alpha = 0$ .

When  $\alpha = 0$ , the momentum equation in system ( $JSO$ ) can be integrated to show that the total shear stress  $T := \sigma + \epsilon v_x$  coincides with the steady value  $\bar{T}(x) = -\bar{f}x$ . Thus  $T = \bar{T}(x)$  is a function of  $x$  only, even though  $\sigma$  and  $v_x$  are functions of both  $x$  and  $t$ . The remaining equations of system ( $JSO$ ) yield, for each fixed  $x$ , the autonomous, quadratic,

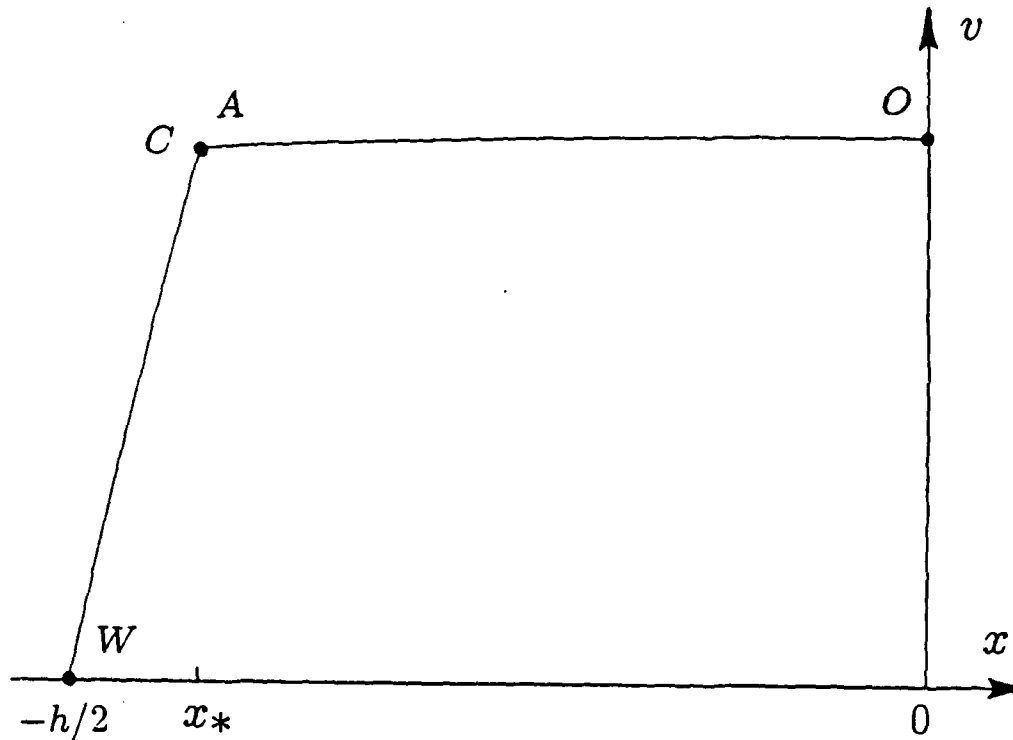


Fig. 2: Velocity profile for steady flow.

planar system of ordinary differential equations

$$\begin{aligned}\dot{\sigma} &= (Z + 1) \left( \frac{\bar{T} - \sigma}{\varepsilon} \right) - \sigma, \\ \dot{Z} &= -\sigma \left( \frac{\bar{T} - \sigma}{\varepsilon} \right) - Z.\end{aligned}\tag{3.1}$$

Here the dot denotes the derivative  $d/dt$ . We emphasize that for each  $\bar{f}$ , a different dynamical system is obtained at each  $x$  on the interval  $[-1/2, 0]$  in the channel because  $\bar{T} = -\bar{f}x$ . By symmetry, we may focus attention on the case  $\bar{T} > 0$ ; also recall from Sec. 2 that  $\varepsilon < 1/8$ ; these are assumed throughout. The dynamical system (3.1) can be analyzed completely by a phase-plane analysis outlined below; the reader is referred to Sec. 3 in [13] for further details. Here we state the main results.

The critical points of system (3.1) satisfy the algebraic system

$$\begin{aligned}(Z + 1 + \varepsilon) \left( \frac{\sigma}{\bar{T}} - 1 \right) + \varepsilon &= 0, \\ \frac{\bar{T}^2}{\varepsilon} \frac{\sigma}{\bar{T}} \left( \frac{\sigma}{\bar{T}} - 1 \right) - Z &= 0.\end{aligned}\tag{3.2}$$

These equations define, respectively, a hyperbola and a parabola in the  $\sigma$ - $Z$  plane; these curves are drawn in Fig. 3, which corresponds to the most comprehensive case of three critical points. The critical points are intersections of these curves. In particular, critical points lie in the strip  $0 < \sigma < \bar{T}$ .

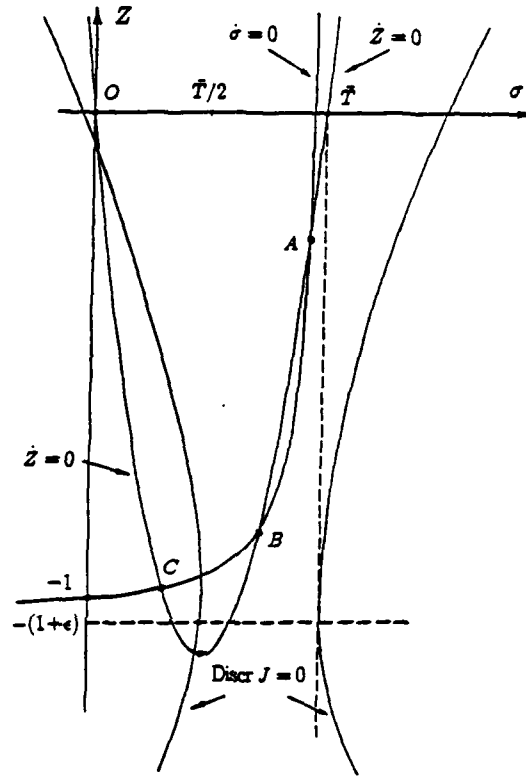


Fig. 3: The phase plane in the case of three critical points.

Eliminating  $Z$  in these equations shows that the  $\sigma$ -coordinates of the critical points satisfy the cubic equation  $Q(\sigma/\bar{T}) = 0$ , where

$$Q(\xi) := \left[ \frac{\bar{T}^2}{\epsilon} \xi(\xi - 1) + 1 + \epsilon \right] (\xi - 1) + \epsilon. \quad (3.3)$$

A straightforward calculation using Eq. (2.4) shows that

$$P(\bar{v}_x) = P\left(\frac{\bar{T} - \sigma}{\epsilon}\right) = -\frac{\bar{T}}{\epsilon} Q(\sigma/\bar{T}). \quad (3.4)$$

Thus each critical point of the system (3.1) defines a steady-state solution of system (JSO): such a solution corresponds to a point on the steady total-stress curve (see Fig. 1) at which the total stress is  $\bar{T}(x)$ . Consequently, we have:

**Proposition 3.1:**

For each position  $x$  in the channel and for each  $\epsilon > 0$ , there are three possibilities:

- (1) there is a single critical point  $A$  when  $\bar{T} < \bar{T}_m$ ;
- (2) there is also a single critical point  $C$  if  $\bar{T} > \bar{T}_M$ ;
- (3) there are three critical points  $A$ ,  $B$ , and  $C$  when  $\bar{T}_m < \bar{T} < \bar{T}_M$ .

For simplicity, we ignore the degenerate cases, where  $\bar{T} = \bar{T}_M$  or  $\bar{T} = \bar{T}_m$ , in which two critical points coalesce.

To determine the qualitative structure of the dynamical system (3.1), we first study the nature of the critical points. The behavior of orbits near a critical point depends on the linearization of system (3.1) at this point, i.e., on the eigenvalues of the Jacobian matrix  $J$  associated with Eq. (3.1), evaluated at the critical point. To avoid solving the cubic equation  $Q(\sigma/\bar{T}) = 0$ , the character of the eigenvalues of  $J$  can be determined from the signs of the trace of  $J$  denoted by  $\text{Tr } J$ , the determinant of  $J$  denoted by  $\text{Det } J$ , and the discriminant of  $J$  denoted by  $\text{Discrm } J$  at the critical points. We omit these tedious calculations, a result of which is a useful fact: at a critical point,  $\epsilon \text{Det } J = Q'(\sigma/\bar{T})$ . This relation is important because  $Q'$  is positive at  $A$  and  $C$  and negative at  $B$ . To assist the reader, Fig. 3 shows the hyperbola on which  $\dot{\sigma} = 0$ , the parabola on which  $\dot{Z} = 0$  [see Eqs. (3.2)], and the hyperbola on which  $\text{Discrm } J$  vanishes. As a result of the analysis above, we draw the following conclusions:

- (1)  $\text{Tr } J < 0$  at all critical points;
- (2)  $\text{Det } J > 0$  at  $A$  and  $C$ , while  $\text{Det } J < 0$  at  $B$ ; and
- (3)  $\text{Discrm } J > 0$  at  $A$  and  $B$ , whereas  $\text{Discrm } J$  can be of either sign at  $C$ . (For typical values of  $\epsilon$  and  $\bar{T}$ ,  $\text{Discrm } J < 0$  at  $C$ ; in particular,  $\text{Discrm } J < 0$  if  $C$  is the only critical point. But it is possible for  $\text{Discrm } J$  to be positive if  $\bar{T}$  is sufficiently close to  $\bar{T}_m$ .)

Standard theory of nonlinear planar dynamical systems (see, e.g., Ref. [3, Chap. 15]) now establishes the local characters of the critical points  $A, B, C$  in Proposition 3.1:

**Proposition 3.2:**

- (1)  $A$  is an attracting node (called the classical attractor);
- (2)  $B$  is a saddle point;
- (3)  $C$  is either an attracting spiral point or an attracting node (called the spurt attractor).

The next task is to determine the global structure of the orbits of system (3.1). In this direction, we modify an argument suggested by A. Coppel [4] and establish the crucial result, the proof of which involves a change in the time scale and an application of Bendixson's theorem:

**Proposition 3.3:**

System (3.1) has neither periodic orbits nor separatrix cycles.

To understand the global qualitative behavior of orbits, we construct suitable invariant sets. In this regard, a crucial tool is that system (3.1) is endowed with the identity (3.5)

$$\frac{d}{dt} \left\{ \sigma^2 + (Z + 1)^2 \right\} = -2 \left[ \sigma^2 + \left( Z + \frac{1}{2} \right)^2 - \frac{1}{4} \right]. \quad (3.5)$$

Thus the function  $V(\sigma, Z) := \sigma^2 + (Z+1)^2$  serves as a Lyapunov function for the dynamical system. Notice that identity (3.5) is independent of  $\bar{T}$  and  $\varepsilon$ .

Let  $\Gamma$  denote the circle on which the right side of Eq. (3.5) vanishes, and let  $C_r$  denote the circle of radius  $r$  centered at  $\sigma = 0$  and  $Z = -1$ , i.e.  $C_r := \{(\sigma, Z) : V(\sigma, Z) = r, r > 0\}$ ; each  $C_r$  is a level set of  $V$ . The circles  $\Gamma$  and  $C_1$  are shown in Fig. 4, which corresponds to the case of a single critical point, the spiral point  $C$ . Eq. (3.5) also implies the critical points of system (3.1) lie on  $\Gamma$ . If  $r > 1$ ,  $\Gamma$  lies strictly inside  $C_r$ . Consequently, Eq. (3.5) shows that the dynamical system (3.1) flows inward at points along  $C_r$ . Thus the interior of  $C_r$  is a positively invariant set for each  $r > 1$ . Furthermore, the closed disk bounded by  $C_1$ , which is the intersection of these sets, is also positively invariant. Therefore the above argument establishes:

**Proposition 3.4:** *Each closed disk bounded by the circle  $C_r, r \geq 1$  is a positively invariant set for the system (3.1).*

The above results combined with identification of suitable invariant sets were used to determine the global structure of the orbits of system (3.1) in the cases of one and three critical points, and to analyze the stable and unstable manifolds of the saddle point at  $B$ . These results are shown in Figs. 5 and 6 and summarized in the following result.

**Proposition 3.5:**

*The basin of attraction of  $A$ , i.e., the set of points that flow toward  $A$  as  $t \rightarrow \infty$ , comprises those points on the same side of the stable manifold of  $B$  as is  $A$ ; points on the other side are in the basin of attraction of  $C$ . Moreover, the arc of the circle  $\Gamma$  through the origin, between  $B$  and its reflection  $B'$  is contained in the basin of attraction of  $A$ . In particular, the stable manifold for  $B$  cannot cross its boundary, so that it cannot cross  $\Gamma$  between  $B$  and  $B'$ .*

*All qualitative features of the dynamics of system (3.1) (except possibly whether  $C$  is a node or a focus) carry over to one that approximates the system ( $JSO_2$ ) in the case of two widely separated relaxation times (see system (4.9) in [13]).*

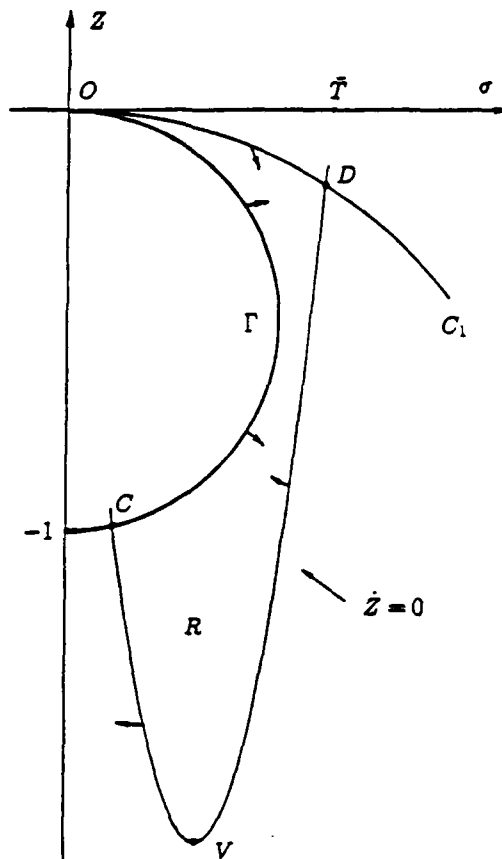


Fig. 4: The phase plane when the spurt attractor  $C$  is the only critical point.

#### 4. Qualitative Features of ( $JSO$ ) Based on Phase Plane Analysis

The discussion that follows sketches an explanation of recent numerical simulations of ( $JSO$ ) described in Refs. [9, 12]. These exhibited several effects related to spurt: latency, shape memory, and hysteresis. Fig. 7 shows the result of simulating a "quasi-static" loading sequence in which the pressure gradient  $\bar{f}$  is increased in small steps, allowing sufficient time between steps to achieve steady flow [9]. The loading sequence is followed by a similar quasi-static unloading sequence, in which the driving pressure gradient is decreased in steps. The initial step used zero initial data, and succeeding steps used the results of the previous step as initial data. The resulting hysteresis loop includes the shape memory predicted by Hunter and Slemrod [7] for a simpler model by a different approach. The width of the hysteresis loop at the bottom can be related directly to the molecular weight of the sample [9].

We explain spurt, shape memory, hysteresis and latency. We consider experiments of the following type: the flow is initially in a steady state corresponding to a forcing  $\bar{f}_0$ , and the forcing is suddenly changed to  $\bar{f} = \bar{f}_0 + \Delta\bar{f}$ . We call this process "loading" (resp. "unloading") if  $\Delta\bar{f}$  has the same (resp. opposite) sign as  $\bar{f}_0$ . The initial flow can be described by specifying, for each channel position  $x$ , whether the flow is at a classical

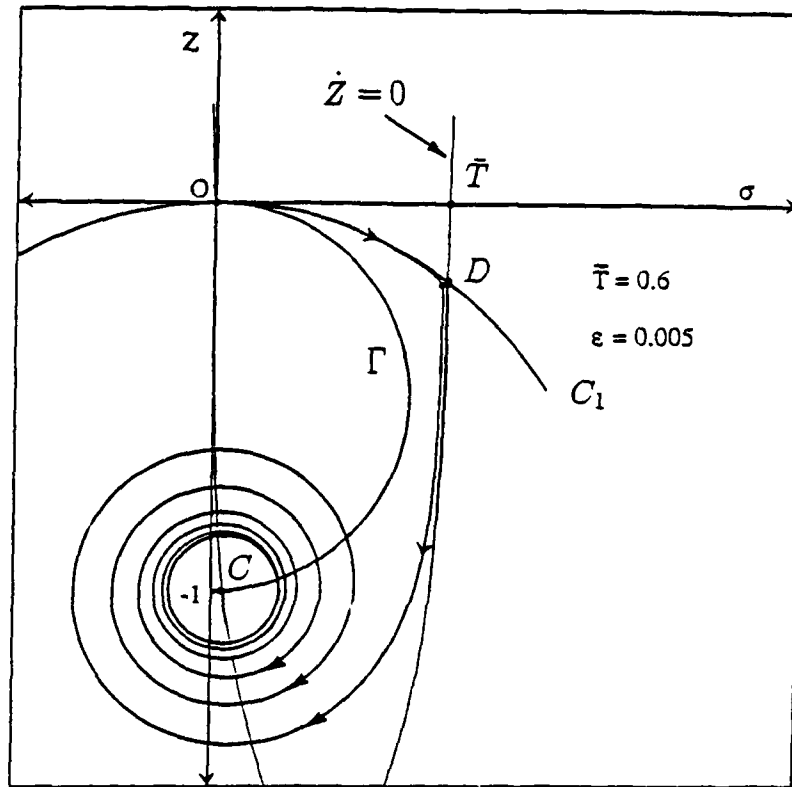


Fig. 5: The orbit through origin when the spurt attractor  $C$  is the only critical point.

attractor  $A$  ( $x$  is a “classical point”) or a spurt attractor  $C$  ( $x$  is a “spurt point”) for the system (3.1) with  $\bar{T} = -\bar{f}_0 x$ . We shall say that any point lying on the same side of the stable manifold of  $B$  as is  $A$  lies on the “classical side”; points lying on the other side are said to be on the “spurt side.” The outcome of the experiment depends on the character of the phase portrait with  $\bar{T} = -\bar{f}x$ . To determine this outcome, we need only decide when a classical point becomes a spurt point or vice versa.

The principle mathematical properties of the dynamical system (3.1) that determine the outcome of loading and unloading experiments are embodied in the following consequence of the phase plane analysis.

**Proposition 4.1:**

- (1) A classical point  $A_0$  for the initial forcing  $\bar{f}_0$  lies in the domain of attraction of the classical attractor  $A$  for  $\bar{f}$ , provided that  $A$  exists (i.e.,  $|\bar{f}x| < \bar{T}_M$ );
- (2) A spurt point  $C_0$  for the initial forcing  $\bar{f}_0$  lies in the domain of attraction of the spurt attractor  $C$  for  $\bar{f}$  unless (a)  $C$  does not exist (i.e.,  $|\bar{f}x| < \bar{T}_m$ ); or (b)  $C$  lies on the classical side of the stable manifold of the saddle point  $B$  for  $\bar{f}$ .

Consider starting with  $\bar{f}_0 = 0$  and loading to  $\bar{f} > 0$ . Thus the initial state for each  $x$



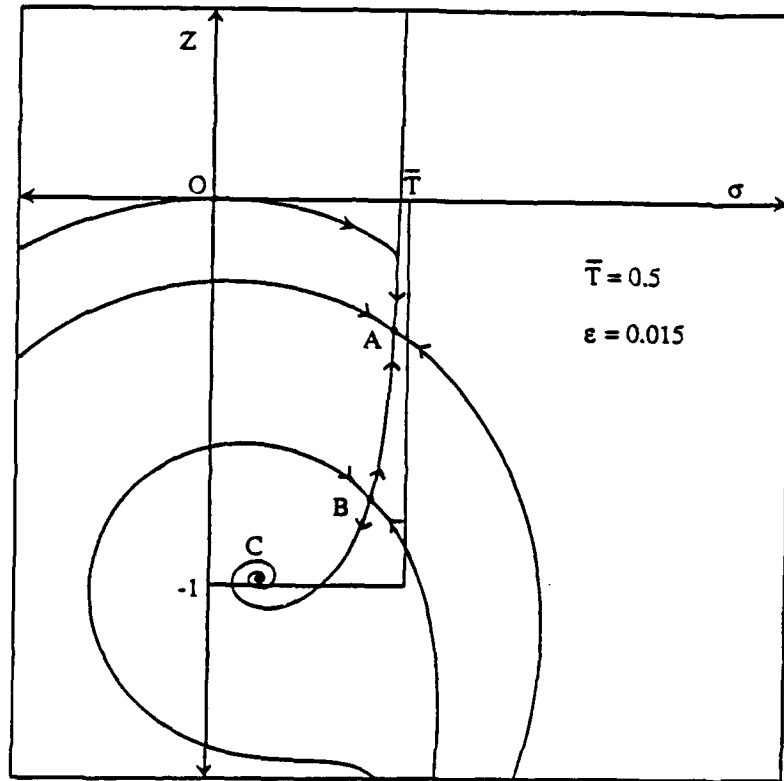


Fig. 6: Phase portrait in the case of three critical points, with  $C$  being a spiral.

lies at the origin  $\sigma = 0$ ,  $Z = 0$ . Then according to 4.1(1) above, each  $x \in [-1/2, 0]$  such that  $\bar{f}|x| < \bar{T}_M$  is a classical point, while the  $x$  for which  $\bar{f}|x| > \bar{T}_M$  are spurt points (because there is no classical attractor). Consequently, we draw two conclusions:

**Proposition 4.2:**

- (a) If the forcing is subcritical (i.e.,  $\bar{f} < \bar{f}_{\text{crit}} := 2\bar{T}_M$ ), the asymptotic steady flow is entirely classical.
- (b) If the forcing is supercritical ( $\bar{f} > \bar{f}_{\text{crit}}$ ), there is a single kink in the velocity profile (see Fig. 2), located at  $x_* = -\bar{T}_M/\bar{f}$ ; those  $x \in [-1/2, x_*)$ , near the wall, are spurt points, whereas  $x \in (x_*, 0]$ , near the centerline, are classical.

The solution in case (b) can be described as “top jumping” because the stress  $\bar{T}_* = \bar{T}_M$  at the kink is as large as possible, and the the kink is located as close as possible to the wall.

Next, consider increasing the load from  $\bar{f}_0 > 0$  to  $\bar{f} > \bar{f}_0$ . A point  $x$  that is classical for  $\bar{f}_0$  remains classical for  $\bar{f}$  unless there is no classical attractor for  $\bar{T} = -\bar{f}x$ , i.e.,  $\bar{f}|x| > \bar{T}_M$ . A spurt point  $x$  for  $\bar{f}_0$ , on the other hand, is always a spurt point for  $\bar{f}$ . As a result, a point in  $x$  in the channel can change only from a classical attractor to a spurt

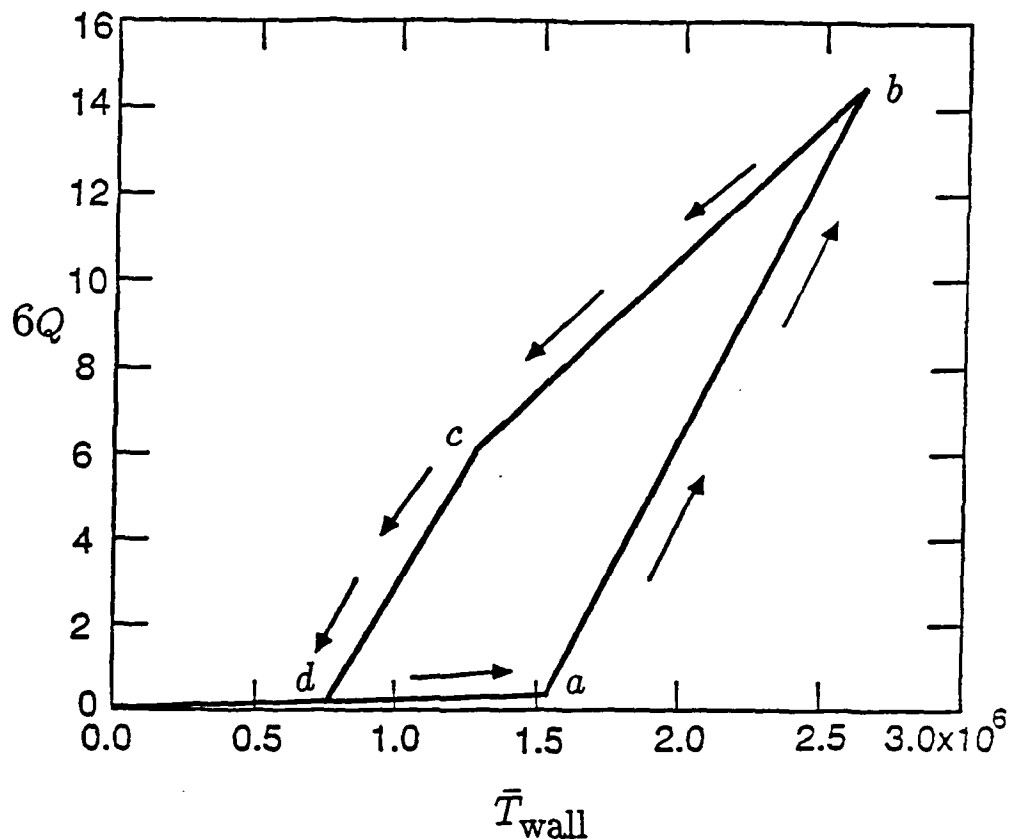


Fig. 7: Hysteresis under cyclic load: normalized throughput  $6Q$  vs. wall shear stress  $\bar{T}_{\text{wall}}$  [9].

attractor, and then only if  $\bar{f}|x|$  exceeds  $\bar{T}_M$ . When  $\bar{f}$  is chosen to be supercritical, loading causes the position  $x_*$  of the kink in Fig. 2 to move away from the wall, but only to the extent that it must: a single jump in strain rate occurs at  $x_* = -\bar{T}_M/\bar{f}$ , where the total stress is  $\bar{T}_* = \bar{T}_M$ . These conclusions are valid, in particular, for a quasi-static process of gradually increasing the load from  $\bar{f}_0 = 0$  to  $\bar{f} > \bar{f}_{\text{crit}}$ .

Now consider unloading from  $\bar{f}_0 > 0$  to  $\bar{f} < \bar{f}_0$ ; assume, for the moment, that  $\bar{f}$  is positive. Here, the initial steady solution need not correspond to top jumping. For this type of unloading, a point  $x$  that is classical for  $\bar{f}_0$  always remains classical for  $\bar{f}$ : the classical attractor for  $\bar{f}$  exists because  $\bar{f}|x| < \bar{f}_0|x|$ . By contrast, a spurt point  $x$  for  $\bar{f}_0$  can become classical at  $\bar{f}$ . This occurs if: (a) the total stress  $\bar{T} = -\bar{f}x$  falls below  $\bar{T}_m$ ; or (b) the spurt attractor  $C_0$  for  $\bar{T} = -\bar{f}_0x$  lies on the classical side of the stable manifold of the saddle point  $B$  for  $\bar{T} = -\bar{f}x$  (see Proposition 4.1(2b)).

Combining the analysis of loading and unloading leads to the following summary of quasi-static cycles and the resulting flow hysteresis.

*Kinks move away from the wall under top jumping loading; they move toward the wall under bottom jumping unloading; otherwise they remain fixed. The hysteresis loop opens from the point at which unloading commences; no part of the unloading path retraces the*

loading path until point  $d$  in Fig. 7.

To explain the latency effect that occurs during loading, assume that  $\epsilon$  is small. It is readily seen that the total stress  $\bar{T}_M$  at the local maximum  $M$  is  $1/2 + O(\epsilon)$ , while the local minimum  $m$  corresponds to a total stress  $\bar{T}_m$  of  $2\sqrt{\epsilon}[1 + O(\epsilon)]$ . Furthermore, for  $x$  such that  $\bar{T}(x) = O(1)$ ,  $\sigma = \bar{T} + O(\epsilon)$  at an attracting node at  $A$ , while  $\sigma = O(\epsilon)$  at a spurt attractor  $C$  (which is a spiral). Consider a point along the channel for which  $\bar{T}(x) > \bar{T}_M$ , so that the only critical point of the system (3.1) is  $C$ , and suppose that that  $\bar{T} < 1$ . Then the evolution of the system exhibits three distinct phases, as indicated in Fig. 6: an initial "Newtonian" phase ( $O$  to  $N$ ); an intermediate "latency" phase ( $N$  to  $S$ ); and a final "spurt" phase ( $S$  to  $C$ ).

The Newtonian phase occurs on a time scale of order  $\epsilon$ , during which the system approximately follows an arc of a circle centered at  $\sigma = 0$  and  $Z = -1$ . Having assumed that  $\bar{T} < 1$ ,  $Z$  approaches

$$Z_N = (1 - \bar{T}^2)^{\frac{1}{2}} - 1 \quad (4.1)$$

as  $\sigma$  rises to the value  $\bar{T}$ . (If, on the other hand,  $\bar{T} \geq 1$ , the circular arc does not extend as far as  $\bar{T}$ , and  $\sigma$  never attains the value  $\bar{T}$ ; rather, the system slowly spirals toward the spurt attractor. Thus the dynamical behavior does not exhibit distinct phases.)

The latency phase is characterized by having  $\sigma = \bar{T} + O(\epsilon)$ , so that  $\sigma$  is nearly constant and  $Z$  evolves approximately according to the differential equation

$$\dot{Z} = -\frac{\bar{T}^2}{Z+1} - Z. \quad (4.2)$$

Therefore, the shear stress and velocity profiles closely resemble those for a steady solution with no spurt, but the solution is not truly steady because the normal stress difference  $Z$  still changes. Integrating Eq. (4.2) from  $Z = Z_N$  to  $Z = -1$  determines the latency period. This period becomes indefinitely long when the forcing decreases to its critical value; thus the persistence of the near-steady solution with no spurt can be very dramatic. The solution remains longest near point  $L$  where  $Z = -1 + \bar{T}$ . This point may be regarded as the remnant of the attracting node  $A$  and the saddle point  $B$ . Eventually the solution enters the spurt phase and tends to the critical point  $C$ . Because  $C$  is an attracting spiral, the stress oscillates between the shear and normal components while it approaches the steady state.

Asymptotic analysis carried out in Sec. 6 of [13] shows that when  $\epsilon$  is sufficiently small, system ( $JSO_2$ ) of [13] has the same asymptotic properties as system ( $JSO$ ). Thus system ( $JSO$ ) approximates ( $JSO_2$ ) quantitatively as well as qualitatively.

## 5. Physical Implications

One of the widely accepted explanations of spurt and similar observations is that the presence of the wall affects the dynamics of the polymer system near the wall. Conceivably, there could be a variety of "wall effects," the most obvious is the loss of chemical bond between wall and fluid, or wall slip [5]. Perhaps the most distinguishing feature of our alternative approach is: it predicts that spurt stems from a material property of the polymer and is not related to any external interaction. The spurt layer forms at the wall in situations such as top jumping because the stresses are higher there; for the same reason, of course, chemical bonds would break at the wall; however, our approach predicts that the layer of spurt points spreads into the interior of the channel on continued loading. Layer thickness is predicted to grow continuously in loading to a thickness that should be observable, provided secondary (two-dimensional) instabilities do not develop.

Our analysis suggests other ways in which experiments might be devised to verify the dependence of spurt on material properties: (i) produce multiple kinks with spurt layer separated from the wall, (ii) produce hysteresis in flow reversal (Fig. 9). Our model predicts circumstances under which a different path can be followed in sudden reversal of the flow than would be followed by a sequence of solutions in which the pressure gradient is reduced to zero and reloaded again (with the opposite sign) to a value of somewhat smaller magnitude. Such behavior does not seem likely to be explainable by a wall effect.

The most important and perhaps the easiest experiment to perform to verify our theory is to produce latency. Our analysis predicts long latency times for data corresponding to realistic material data; no sophisticated timing device would be required, nor would the onset of the instability be hard to identify. The increase in throughput is predicted to be so dramatic that simple visual inspection of the exit flow would probably be sufficient.

## 6. Conclusions

Although our analysis applies only to the special constitutive models we have studied, we expect that the qualitative features of our results appear in a broad class of non-Newtonian fluids. Our analysis has identified certain universal mathematical features in the shear flow of viscoelastic fluids described by differential constitutive relations that give rise to spurt and related phenomena. The key feature is that there are three widely separated time scales, each associated with an important non-dimensional number ( $\alpha$ ,  $\varepsilon$ , and 1, respectively), when scaled by the dominant relaxation time,  $\lambda^{-1}$ . Each of these time scales can be associated with a particular equation in system (JSO) [13]. The key to understanding the dynamics of such systems is fixing the location of the discontinuity in the strain rate induced by the non-monotone character of the steady shear stress vs. strain rate.

## Acknowledgments

We thank Professor A. Coppel for suggesting an elegant argument that rules out the existence of periodic and separatrix cycles for the systems (3.1). We also acknowledge helpful discussions with D. Aronson, M. Denn, G. Sell, M. Slemrod and A. Tzavaras, and M. Yao.

## References

1. A. Andronov and C. Chaikin, *Theory of Oscillations*, Princeton Univ. Press, Princeton, 1949.
2. R. Bird, R. Armstrong, and O. Hassager, *Dynamics of Polymeric Liquids*, John Wiley and Sons, New York, 1987.
3. E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
4. A. Coppel, , 1989. private communication.
5. M. Denn, "Issues in Viscoelastic Fluid Dynamics," *Annual Reviews of Fluid Mechanics*, 1989. to appear.
6. M. Doi and S. Edwards, "Dynamics of Concentrated Polymer Systems," *J. Chem. Soc. Faraday* 74 (1978), pp. 1789-1832.
7. J. Hunter and M. Slemrod, "Viscoelastic Fluid Flow Exhibiting Hysteretic Phase Changes," *Phys. Fluids* 26 (1983), pp. 2345-2351.
8. 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.
9. 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.* 29 (1988), pp. 303-325.
10. R. Kolkka and G. Ierley, "Spurt Phenomena for the Giesekus Viscoelastic Liquid Model," *J. Non-Newtonian Fluid Mech.*, 1989. To appear.
11. D. Malkus, J. Nohel, and B. Plohr, "Time-Dependent Shear Flow Of A Non-Newtonian Fluid," in *Conference on Current Problems in Hyperbolic Problems: Riemann Problems and Computations (Bowdoin, 1988)*, ed. B. Lindquist, Amer. Math. Soc., Providence, 1989. Contemporary Mathematics, to appear.
12. D. Malkus, J. Nohel, and B. Plohr, "Dynamics of Shear Flow of a Non-Newtonian Fluid," *J. Comput. Phys.*, 1989. To appear.
13. D. Malkus, J. Nohel, and B. Plohr, "Analysis of New Phenomena In Shear Flow of Non-Newtonian Fluids," *SIAM J. Appl. Math.*, 1989. Submitted.
14. T. McLeish and R. Ball, "A Molecular Approach to the Spurt Effect in Polymer Melt Flow," *J. Polymer Sci.* 24 (1986), pp. 1735-1745.
15. J. Nohel, R. Pego, and A. Tzavaras. "Stability of Discontinuous Steady States in Shearing Motions of Non-Newtonian Fluids," *Proc. Roy. Soc. Edinburgh, Series A*, 1989. submitted.

16. J. Oldroyd, "Non-Newtonian Effects in Steady Motion of Some Idealized Elastico-Viscous Liquids," *Proc. Roy. Soc. London A* 245 (1958), pp. 278-297.
17. J. Pearson, *Mechanics of Polymer Processing*, Elsevier Applied Science, London, 1985.
18. 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.
19. 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.
20. M. Yao and D. Malkus, "Analytical Solutions of Plane Poiseuille Flow of a Johnson-Segalman Fluid," in preparation, 1989.

# SMART ALGORITHMS FOR COMPLEX PROBLEMS IN FLUID DYNAMICS<sup>1</sup>

J. Tinsley Oden<sup>2</sup>

## Abstract

Some recent results obtained using adaptive finite element methods and so-called smart algorithms in two- and three-dimensional problems in fluid mechanics are discussed. These include applications of *h*- and *h-p*-adaptive methods on unstructured meshes.

---

<sup>1</sup>The present manuscript is extracted from the full paper on this subject that is to appear in *Computers and Structures* in a special volume compiled from the Symposium on Frontiers in Computational Mechanics, held at M.I.T., March 1989, in honor of Professor T.H.H. Pian on the occasion of his seventieth birthday.

<sup>2</sup>Texas Institute for Computational Mechanics, The University of Texas at Austin

# 1 Introduction

The significant challenges of numerical simulation of complex phenomena in fluid dynamics have encouraged the development of new and innovative methods for producing good computational results as efficiently as possible with existing computing capabilities. Chief among these are adaptive methods, which attempt to adapt the mesh size and topology or the spectral order of the approximation so as to yield acceptable results with minimum numbers of grid points or degrees of freedom. In this paper, some recent results are surveyed. Some of the results described are also discussed in the conference paper [1,2]; others are new results obtained using recently completed three-dimensional adaptive  $h$ -codes and  $h$ - $p$  codes.

## 2 Brief Review of Adaptive Methods

Adaptive methods in computational mechanics are generally based on a simple idea: when the error in a computation is too large, change the structure of the approximation (the mesh size, the location of grid points, the order of the approximation, etc.) to reduce it. Interest in such procedures has grown gradually in recent years with the realization that they may embody ways to optimize computations — to deliver the best answers in some sense for the least effort. However, implementation of the adaptive idea constitutes a significant departure from conventional methods in CFD and involves many open problems. For instance, the very notion that one attempts to reduce error implies that the error is known or can be estimated in some sense. Thus, the first step in adaptivity is to develop measures of “goodness” of solutions, and such measures may range from ad hoc checks of solution gradients to rigorous *a posteriori* error estimates. While progress in *a posteriori* error estimation has been made in recent months, this subject remains an area of active research.

Having an estimate of the error in the solution at a grid cell, what can one do to systematically reduce it below some preset level? In general, one can refine the mesh size  $h$  ( $h$ -methods of adaptivity), increase the density of grid points by relocating nodes ( $r$ -methods of adaptivity), increase the local spectral order  $p$  of the approximation ( $p$ -methods of adaptivity), or use combinations of these techniques (e.g.,  $h$ - $p$  techniques). Each of these choices puts new demands on the overall approach to the computational problem. In particular, adaptive techniques (1) must generally function on unstructured meshes, (2) require elaborate and complicated data structures, (3) employ explicit or iterative solution techniques since direct solvers are of limited value on dynamically evolving unstructured meshes, (4) cope with special issues of stability of numerical schemes that must function with continually changing structures and orders, and (5) attempt to minimize the computational overhead of the error estimation and of implementation of the adaptive process itself. These are the



major challenges of adaptive methods in computational fluid dynamics.

In recent months, we have attempted to meet these challenges through a series of studies on each of the above issues. Some of our results have been implemented in a collection of Navier-Stokes solvers that we refer to as the *ADAPT<sup>TM</sup>* code; other results deal with new data structures and adaptive methods and require further work before they represent effective tools for complex flow simulations. Some of our results on special issues of adaptivity and on selected applications are briefly summarized.

### 3 Local Approximation of the Navier Stokes Equations on Unstructured Meshes

Most of the applications discussed here pertain to numerical solutions of the compressible Navier-Stokes equations. In three dimensions, without body forces or external heat sources, these can be written,

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}}{\partial x} + \frac{\partial \mathbf{F}}{\partial y} + \frac{\partial \mathbf{G}}{\partial z} = \text{div } \mathbf{S} \quad (3.1)$$

where  $\mathbf{U}$ ,  $\mathbf{E}$ ,  $\mathbf{F}$ , and  $\mathbf{G}$  are vectors and  $\mathbf{S}$  is a matrix of stresses, power, and heat flux,

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho e \end{bmatrix} \quad (3.2)$$

$$\mathbf{E} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ (\rho e + p)u \end{bmatrix} \quad \mathbf{F} = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ (\rho e + p)v \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ (\rho e + p)w \end{bmatrix} \quad (3.3)$$

$$\text{div } \mathbf{S} = \begin{bmatrix} 0 \\ \tau_{xx,x} + \tau_{xy,y} + \tau_{xz,z} \\ \tau_{xy,x} + \tau_{yy,y} + \tau_{yz,z} \\ \tau_{xz,x} + \tau_{yz,y} + \tau_{zz,z} \\ \sum_{i=1}^3 \sum_{j=1}^3 \tau_{ij} u_{j,i} + \sum_{i=1}^3 q_{i,i} \end{bmatrix} \quad (3.4)$$

Here  $\rho$  is the total mass density,  $u$ ,  $v$  and  $w$  are the velocity components,  $p$  is the fluid pressure,  $\tau_{ij}$  are the components of the viscous stresses,  $e$  is the total energy defined by

$e = \hat{e} + \frac{1}{2}(u^2 + v^2 + w^2)$  where  $\hat{e}$  is the thermodynamic internal energy per unit mass, and  $q$  is the heat flux vector.

The constitutive relation used to evaluate the viscous stress  $\tau_{ij}$  is given by

$$\tau_{ij} = \mu[u_{i,j} + u_{j,i}] + \lambda u_{k,k} \delta_{ij}$$

where  $\mu$  and  $\lambda$  are the first and second coefficients of viscosity,  $u_{i,j}$  are the components of the velocity gradient ( $u_{i,j} = \partial u_i / \partial x_j$ ,  $x_1 = x$ ,  $x_2 = y$ ,  $x_3 = z$ ), and  $\delta_{ij}$  is the Kronecker delta.

In addition to the partial differential equations above, two thermodynamic relations are also needed to close the system of equations. These relations are the ideal-gas state equation

$$p = (\gamma - 1)\rho e \quad (3.5)$$

and an equation which relates the temperature to the internal energy

$$e = c_v T \quad (3.6)$$

Here  $\gamma$  is the specific heat ratio and  $c_v$  is the specific heat at constant volume. With these two additional equations we now have a complete system which can be solved for the vector of unknown quantities ( $\rho$ ,  $u$ ,  $v$ ,  $w$ ,  $e$ ) and for  $p$  and  $T$ .

For the class of problems considered here, a weak formulation is defined in terms of two classes of functions:  $V$ , the class of trial functions, to which the solution  $U$  belongs, and  $W$ , the class of test (or weight) functions which are integrated against the residual of the governing equations. The resulting weak form is:

Find  $U$  in a class  $V$  such that

$$\begin{aligned} & \int_0^T \int_{\Omega} (U_t^T \phi - E^T \phi_x - F^T \phi_y - G^T \phi_z) d\Omega dt \\ & = \int_0^T \int_{\Omega} S(U) : \nabla \phi d\Omega dt + \int_0^T \int_{\partial\Omega} \alpha^T \phi ds dt \end{aligned} \quad (3.7)$$

for all test functions  $\phi = [\phi_1, \phi_2, \dots, \phi_5]$  in  $W$ : where  $[0, T]$  is the time interval of interest,  $\Omega$  is the region through which the fluid moves,  $\partial\Omega$  is the boundary of the flow region  $\Omega$ , and  $\alpha$  is the vector of boundary fluxes. It is understood that the viscous stress terms on the right-hand side of (3.7) may also appear in the integrated form,

$$\int_0^T \left( \int_{\Omega} -\tau_{ij} \phi_{i,j} d\Omega + \int_{\partial\Omega} \tau_{ij} n_i \phi_j ds \right) dt$$

so that differentiability of  $\tau_{ij}$  in  $L^1$  is not necessarily required.

Our numerical approximation of the flow problem will begin with a discrete approximation of the alternate weak form for a time interval  $[t_1, t_2]$ :

Find  $U = U(x, t) \in V$  such that

$$\begin{aligned}
 & - \int_{t_1}^{t_2} \int_{\Omega(t)} U^T \phi_t d\Omega dt \\
 & + \int_{\Omega(t_2)} U^T(\mathbf{x}, t_2) \phi(\mathbf{x}, t_2) d\Omega - \int_{\Omega(t_1)} U^T(\mathbf{x}, t_1) \phi(\mathbf{x}, t_1) d\Omega \\
 & = \int_{t_1}^{t_2} \int_{\Omega(t)} \mathbf{Q}(U) : \nabla \phi d\Omega dt - \int_{t_1}^{t_2} \int_{\partial\Omega(t)} \phi^T (\mathbf{Q}(U)) \mathbf{n} ds dt \\
 & \int_{t_1}^{t_2} \int_{\Omega(t)} \mathbf{S}(U) : \nabla \phi d\Omega dt + \int_{t_1}^{t_2} \int_{\partial\Omega(t)} \alpha^T \phi ds dt
 \end{aligned} \tag{3.8}$$

for all  $\phi \in W$

Here,  $\phi = \{\phi_1, \phi_2, \dots, \phi_5\}^T$ ,  $d\Omega = dx = dx_1 dx_2 dx_3$ ,  $\phi_\alpha = \phi_\alpha(\mathbf{x}, t)$ ,  $\mathbf{x} \in \Omega(t)$ ,  $\alpha = 1, 2, \dots, 5$ ,  $\phi_t = \partial\phi/\partial t$ , and  $\mathbf{n}$  is an outward unit normal vector to the boundary. Also, here  $\mathbf{Q}$  is the convective flux  $\mathbf{Q}(U) = (\mathbf{E}, \mathbf{F}, \mathbf{G})$ . It is easily verified that (3.8) is equivalent to the entire system of Navier-Stokes equations, Rankine-Hugoniot jump conditions (when  $\mathbf{S} = \mathbf{O}$ ), and initial conditions on  $U$  (at  $t = t_1$ ) whenever  $U$  is a  $C^0$ -function everywhere except at surfaces of discontinuity where the jump conditions hold.

In a strictly formal way, the finite element approximation of the flow is obtained from the weak statement of the conservation laws, by interpreting  $\Omega$  as a quadrilateral or brick element, replacing  $U$  by the discrete approximation  $U^h$  and replacing the test functions  $\phi$  by the discrete functions  $\phi^h$ .

## 4 Various Adaptive Methods and Data Structures

**Various Adaptive Strategies.** As is well known, several distinctly different adaptive strategies for CFD problems have emerged over the last several years. We classify them as follows:

**r-methods (or moving finite element methods).** These methods "relocate" grid points in a mesh so that the grid density is large in regions of high error. Here, in general, a fixed number of elements and nodes is used. These classes of methods include those designed to merely enhance orthogonality and smoothness of grids, reduce the  $L^2$ -error in residuals, or to equidistribute the error.

**h-methods.** These methods involve automatic refinement of mesh sizes  $h$ . Data structures for  $h$ -refinement vary in detail; a critique and survey of such data structures was given by Demkowicz and Oden [3].

***p*-methods.** The *p*-methods involve the adaptive enrichment of the spectral order of the approximations over subdomains in a fixed grid. The *p*-methods are closely related to the spectral element methods of Patera (e.g. [4]) and have been extensively developed in the solid mechanics literature by Szabo and his collaborations (e.g. [5]). Adaptive *p*-methods for two dimensional Navier-Stokes equations were first presented in [6].

**Combined methods.** The most effective techniques generally involve a combination of *h* and *r* or *h* and *p* techniques. However, the complexity of data structures for some combined adaptive methods can be substantial [7].

The *h-r* adaptive strategy is regarded here as primarily a preprocessing technique, wherein nodes are positioned in an initial mesh to align gridlines with special flow features such as shocks, boundary-layers, etc. Then, we superimpose on a pre-processed *r*-grid a full *h* or *h-p* adaptive scheme. We describe in the next section a simple *r*-adaptive strategy adequate for preprocessing and a general *h-p* scheme for two-dimensional problems.

#### 4.1 An *h*-Refinement/Unrefinement Strategy.

One *h*-procedure involves the following steps:

1. For a given domain  $\Omega$ , a coarse finite element mesh is constructed which contains only a number of elements sufficient to model basic geometrical features of the flow domain.
2. As our adaptive process will be designed to handle groups of four elements at a time (for the two-dimensional case), we may generate a finer starting grid by a bisection process to obtain an initial set of element groups.
3. We initiate the numerical solution procedures on this initial coarse grid, and compute error indicators  $\theta_e$  over all  $M$  elements in the grid. Let

$$\theta_{\text{MAX}} = \max_{1 \leq e \leq M} \theta_e$$

4. Next, we scan groups of a fixed number  $P$  of elements and compute

$$\theta_{\text{GROUP}}^k = \sum_{k=1}^P \theta_{e_k}$$

where  $e_k$  is the element for group  $k$ . We take  $P = 4$  in our current code.

5. Error tolerances are defined by two real numbers,  $0 < \alpha, \beta < 1$ . If

$$\theta_e \geq \beta\theta_{\text{MAX}}$$

we refine element  $\theta_e$ . This is done by bisecting  $\theta_e$  into four new subelements. If

$$\theta_{\text{GROUP}}^k \leq \alpha\theta_{\text{MAX}}$$

we unrefine the group  $k$  by replacing this group with a single new element with nodes coincident with the corner nodes of the group.

This general process can be followed for any choice of an error indicator. Moreover, it can also be implemented at each time step. Three-dimensional generalizations are straightforward with eight brick elements constituting a group.

One possible adaptive scheme for time-dependent problems is:

1. Advance the solution  $N$  time steps  $\Delta t$  using an appropriate time-marching scheme.
2. Calculate error estimates.
3. Refine the mesh.
4. Redo the  $N$  time-step calculations using the new refined mesh.
5. Redo the error estimation.
6. Unrefine the mesh.
7. Go to 1.

There are several rather obvious alternative versions of this algorithm, but this is the approach used in the sample calculations presented later in this paper.

The  $h$ -methods used in all calculations reported here use 1-irregular refined meshes. Full details of these types of  $h$ -refinement strategies are discussed in [3].

### A $p$ -Method

The idea of increasing the order of an approximation while keeping mesh sizes fixed is a natural one in the case of problems with thin boundary layers or singularities. In results to be outlined later, we employ a hierarchical  $p$ -version of the finite element method. The idea is to choose element shape functions of the form

$$\varphi_{ij}(\xi, \eta) = \sum_{i,j} \chi_i(\xi)\chi_j(\eta)$$

where

$$\chi_i = \text{polynomial of degree } \leq p \text{ in } \xi \in [-1, 1]$$

These polynomials have hierarchical structure, which ensures the property that the element matrices corresponding to an approximation of degree  $p$  contain as proper submatrices all of those element matrices corresponding to approximations of degree less than  $p$ . For a two-dimensional quadrilateral element, the degrees of freedom are the nodal values  $u_i, i = 1, 2, 3, 4$  at the vertices, the tangential derivatives  $\partial^k u / \partial \tau^k, k = 1, 2, \dots, p$  at the midsides, and mixed derivatives  $\partial^m u / \partial \xi^\ell \partial \eta^r, \ell + r = m = 1, 2, \dots, p$  at the centroid.

To fix ideas, consider first the one-dimensional case. In the classical FEM (e.g., Lagrange interpolation), shape functions for various order of approximation are constructed independently. For example, passing from a linear element with two linear shape functions to a quadratic element, we construct the three quadratic shape functions independently of the shape functions for the linear element. An alternative way to construct the *same* second order approximation is to complete the set of two linear shape functions by including a third, quadratic shape function. At the moment the definition of this third shape function and a corresponding degree of freedom is somewhat arbitrary, the only restriction being that the set of shape functions must form a *dual* basis to the set of degrees of freedom, i.e.,

$$\varphi_i(\chi_j) = \delta_{ij} \quad i, j = 0, 1, 2$$

where  $\varphi_i, i = 0, 1, 2$  denote the degrees of freedom and  $\chi_j, j = 0, 1, 2$  the corresponding shape functions. Since the two degrees of freedom associated with the linear shape functions are function values at the endpoints, this implies that the added quadratic shape function must vanish at both endpoints. The remaining hierarchical functions for  $p > 2$  also vanish at the vertices.

### An $r$ -Method

If the mesh size  $h$  and the polynomial degree  $p$  are fixed, one can show that the optimal mesh is that for which the nodes are positioned so that the error is equidistributed over the mesh; i.e., the error over each element is the same. Diaz, Kikuchi, and Taylor [8] have used this fact to produce a simple algorithm for  $r$ -adaptivity:

1. For fixed  $h$  and  $p$  on a mesh of quadrilateral elements, compute error estimators  $\theta_e$  for all elements in a mesh  $\Omega_h \subset \mathbb{R}^2$ .
2. For each 4-element group calculate the error per unit area  $\theta_e/A_e$  of each element in the group.

3. Compute the "centroid of error"

$$\mathbf{y}_k = \frac{\sum_{j=1}^4 \mathbf{r}_j \times \frac{\theta_j}{A_j} \mathbf{e}_3}{\sum_{j=1}^4 \frac{\theta_j}{A_j}}, \quad k = \text{group no.}$$

where  $\mathbf{r}_j$  is the position vector of the centroid of element  $j$  in group  $k$ .

4. Relocate the central node in the group at  $\mathbf{y}_k$  so as to (approximately) equidistribute error over the 4-element group.
5. Continue this process over all groups until the error is equidistributed over the entire mesh.

This simple procedure is easy to implement and is effective in many classes of problems.

### A General $h$ - $p$ Data Structure

Space does not permit the discussion of a general  $h$ - $p$  data structure recently developed by the author and his colleagues [7]; however, some  $h$ - $p$  results will be mentioned later in this paper.

## 5 Some Sample Results

We next cite some representative results obtained recently on adaptive finite element methods in CFD. Additional details are given in earlier papers on this subject (e.g., [9-30]).

*A.  $h$ -Adaptive Schemes for Unsteady Compressible Navier-Stokes Codes for Rotor-Stator Interaction.* Our *ADAPT<sup>TM</sup> 2D* and *ADAPT<sup>TM</sup> 3D* codes were originally built around an  $h$ -adaptive data structure for transient, subsonic, transonic, and supersonic flows in turbines. These flow simulators employ an algebraic turbulence model and a sliding mesh technique to model the motion of rotor blades relative to stator blades in turbomachinery. A Euler code for these problems, which was reported in [14, 19], was a predecessor of these programs.

Typical results of a two-dimensional rotor-stator calculation are shown in Figs. 1-4. There one sees a dynamically changing mesh generated after each of a specified number of time steps in such a way to reduce computed errors below a preset error tolerance. Note the continuity of density (and pressure) contours across mesh interfaces and the interaction of shocks on the moving turbine blades. The code also computes the time history of stresses in the blade due to fluid pressure and shear. It is also interesting to note that flows during

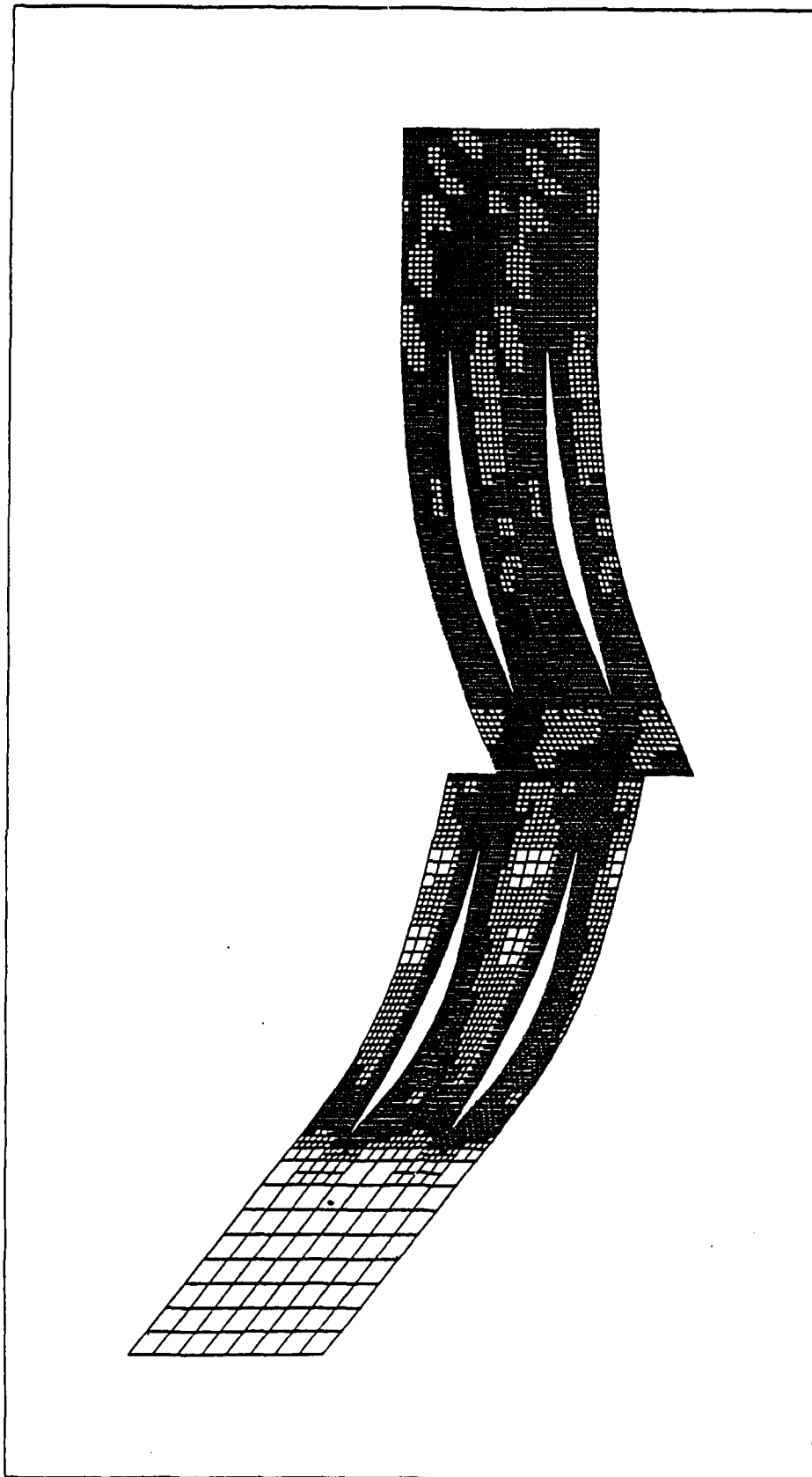


Figure 1: Dynamically evolving adaptive grid at one time instant for Navier-Stokes solution of rotor-stator interaction. Mesh for rotor blades on right is moving relative to fixed stator blade mesh along sliding interface.



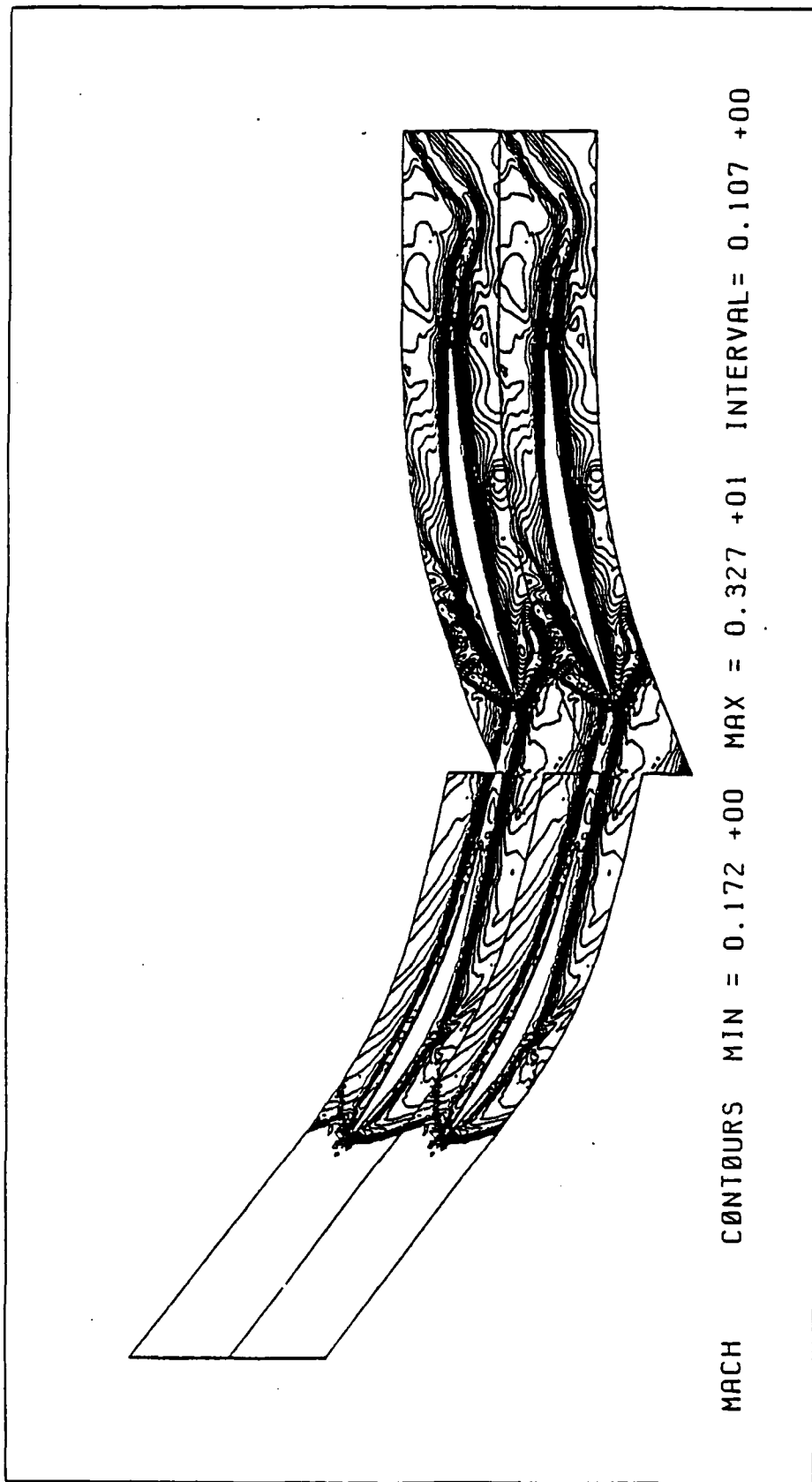


Figure 2: Computed instantaneous Mach number contours at 3.5 cycles for rotor-stator flow interaction simulation.

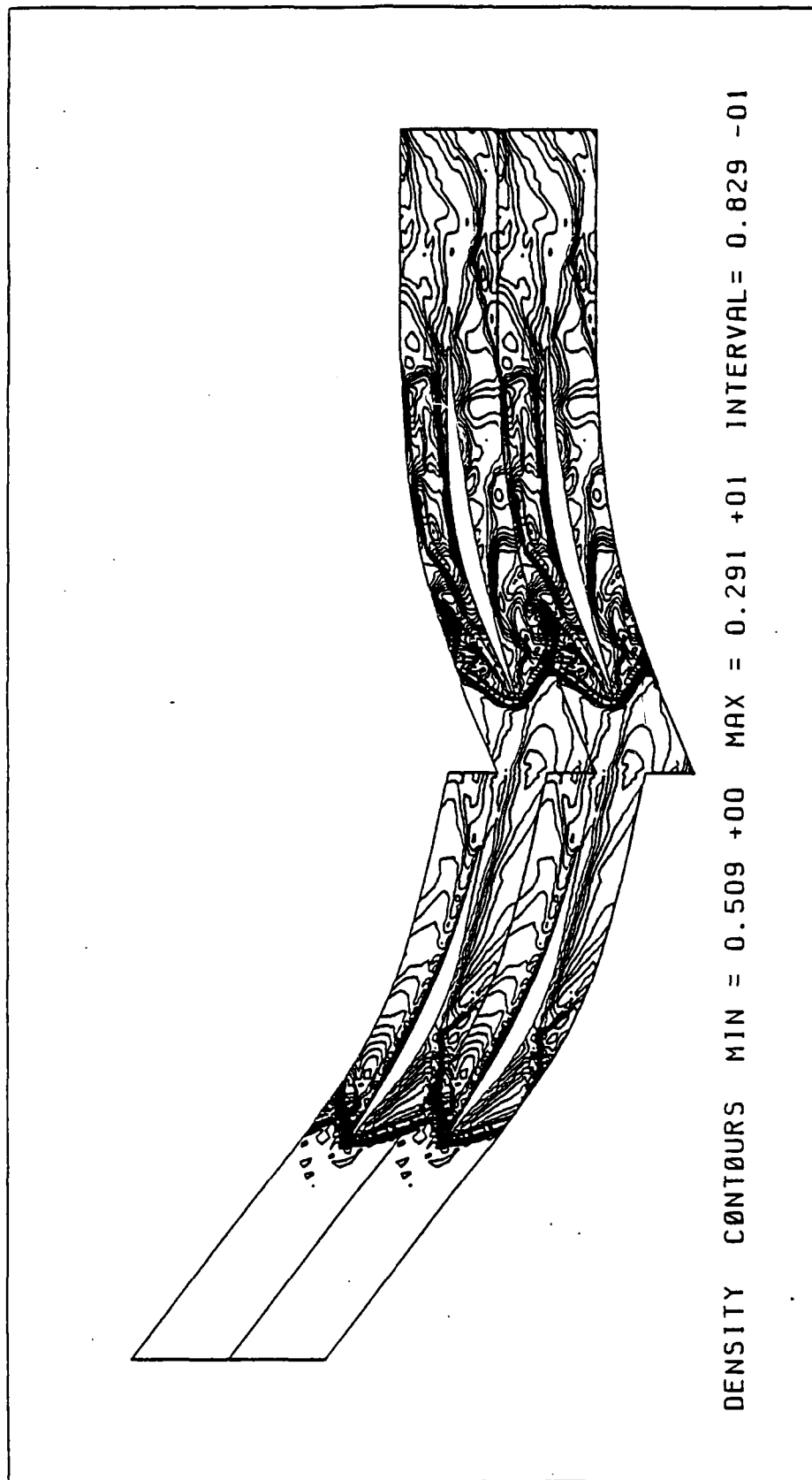


Figure 3: Computed instantaneous density contours at 3.5 cycles of rotor-stator flow interaction simulation.

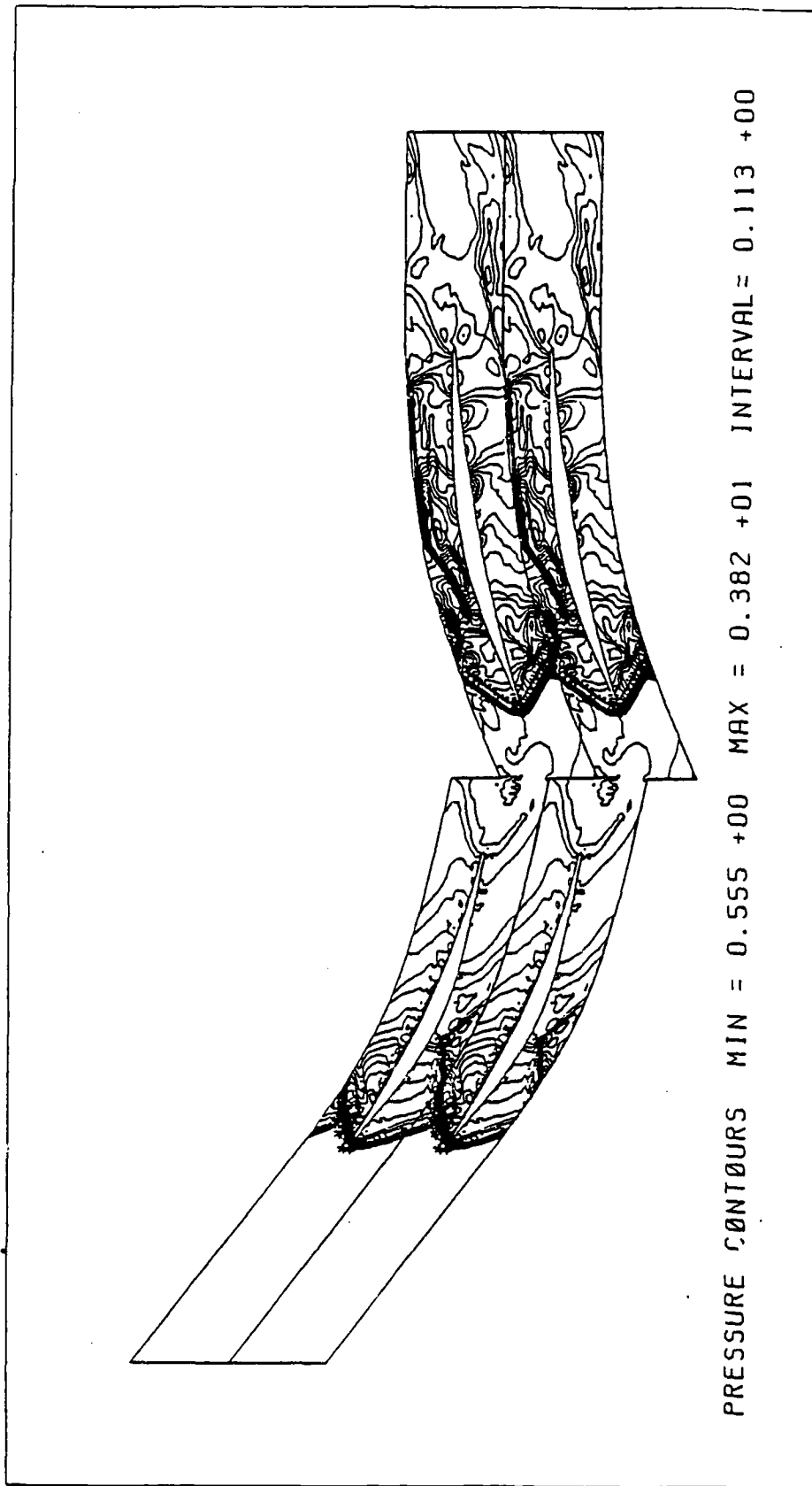


Figure 4: Computed instantaneous pressure contours at 3.5 cycles of viscous rotor-stator flow interaction.

multiple cycles of the blade rows have been computed from a start-up uniform flow to flow with periodic structure which exists after the turbine is in operation. Such calculations simulate as many as 21 complete blade revolutions and 100,000 time steps. Note also that the adaptive process, which consumes only 1 percent of the total calculation, uses around 3,500 elements to deliver the desired accuracy at any particular time during the computation, while a uniform mesh needed for the same accuracy consists of around 14,000 elements. This version of *ADAPT<sup>TM</sup> 2D* employs bilinear quadrilateral elements; the 3D code uses trilinear brick elements.

New results on three-dimensional calculations are shown in Figs. 5-10. A three-dimensional, three-level adapted mesh around a pair of stationary blades is shown in Fig. 5. Pressure contours on planes normal to the blade are shown in Fig. 6 with the adapted grid on the leading blade shown again in Fig. 7. Results for moving blades are in Figs. 8, 9, and 10 with the computed adaptive grid for a rotor blade moving with respect to a fixed stator shown in Figs. 9 and 10 after 1.5 and 6 cycles, respectively.

*B. Low-Mach Number Flow Around a Cylinder.* Figures 11 and 12 show the versatility of the 2D *h*-adaptive strategy for subsonic flow around a cylinder. Note the dynamically changing mesh and the resolution of vortices spinning off the cylinder at  $M = 0.65$ . Fully implicit schemes which function on unstructured meshes are being developed for these problems, but the results shown were obtained with an explicit flow solver, the effectiveness of which was made possible by the use of a near-optimal mesh at the end of each of a designated collection of time steps.

Three-dimensional results for the cylinder are shown in Figs. 13 and 14.

*C. Supersonic Flow Over a Ramp.* Adapted meshes and density contours for flow over a three-dimensional ramp are shown in Figs. 15 and 16.

*D. An *h-r* Adaptive Calculation of Shocks Structure on a Blunt Body.* A moving node technique (an *r*-method) is used to condition mesh structures prior to an *h*-adaptive calculation. Figures 17 and 18 show a typical calculation. There we observe an *h-r* adaptive mesh and density contours of an inviscid gas impinging on a blunt body. Our results indicate that *r*-method preprocessing can be beneficial in aligning the initial mesh with shocks in steady supersonic flow problems with the result that a given level of *h*-refinements produces better solutions than a pure *h*-process which is initiated on an unaligned mesh.

*E. A New *h-p* Scheme for Optimal Computations.* Both two- and three-dimensional *h-p* adaptive codes are operational for the analysis of general linear boundary-value problems. Extensions to steady-state Euler equations are under study. These codes employ an optimization algorithm which chooses the optimal distribution of *h* (mesh size) and *p* (polynomial degree/spectral order) to produce a solution with a given level of local accuracy with a minimum number of unknowns.

## Example 2: 3D Rotor-Stator Interaction?

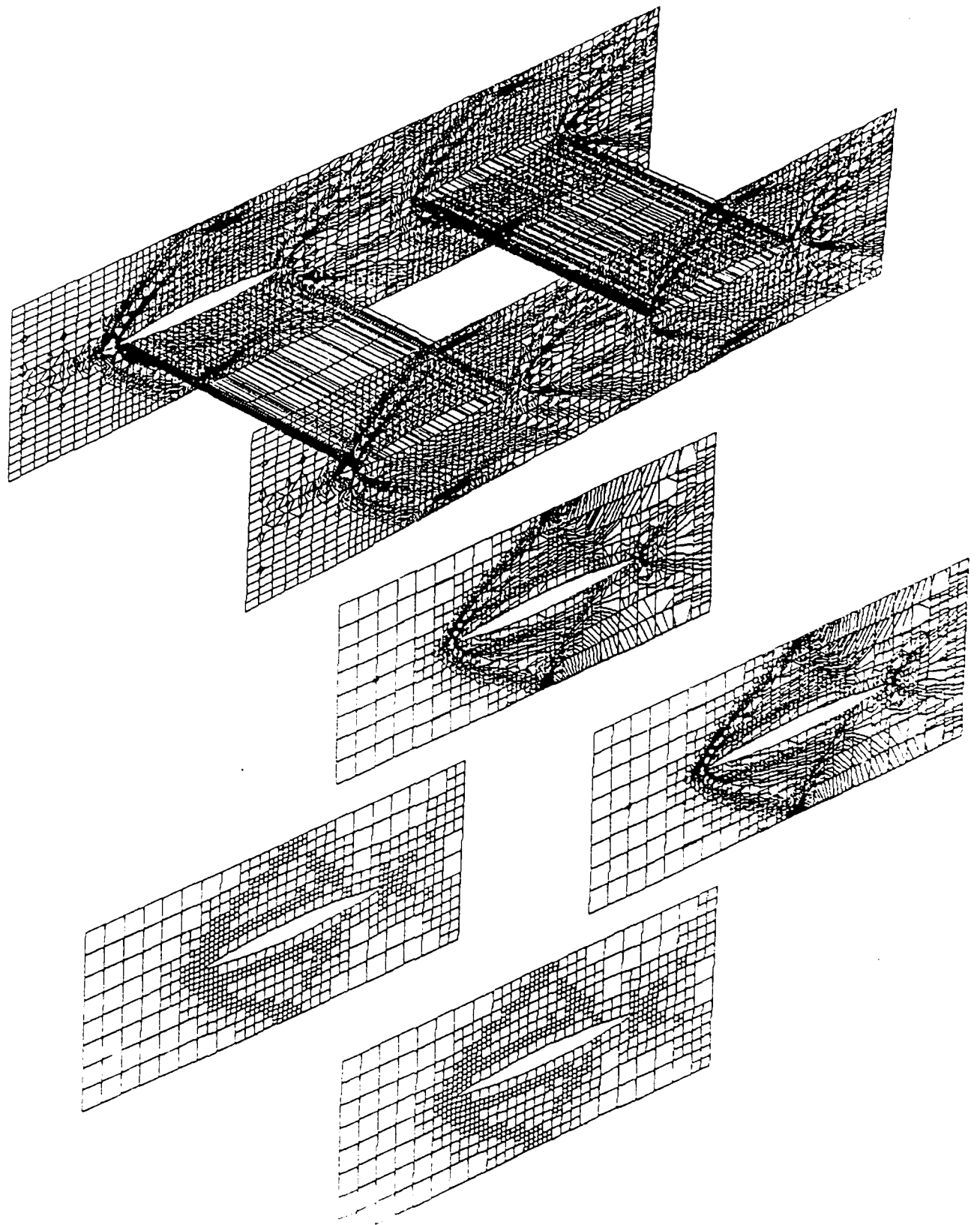


Figure 5: Supersonic flow past two stationary turbine blades: final adapted grid and density contours.

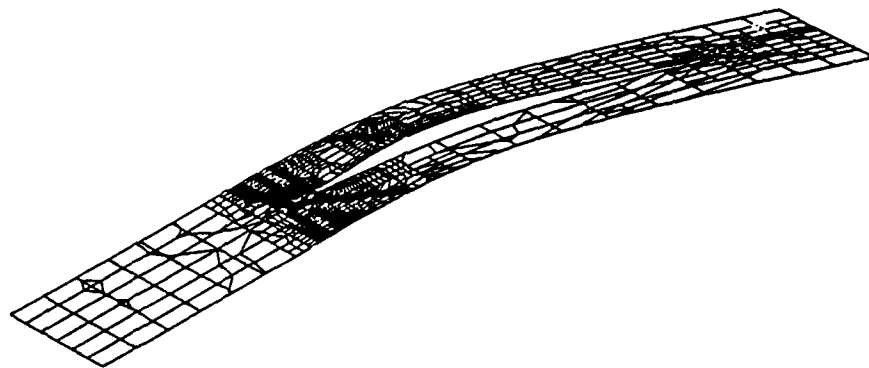
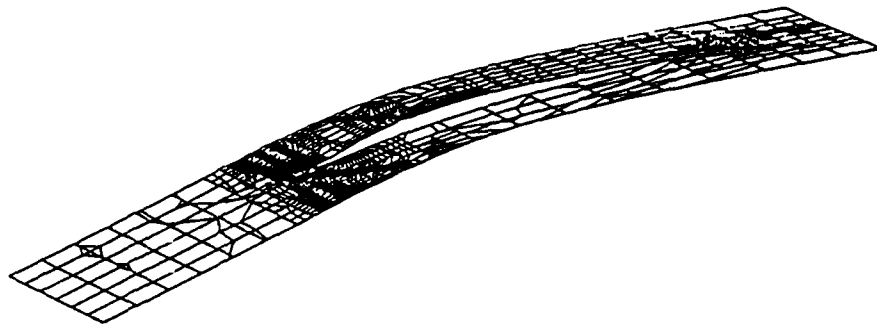


Figure 6: Supersonic flow past a rotor blade. Pressure contours on surfaces through blade cross sections.

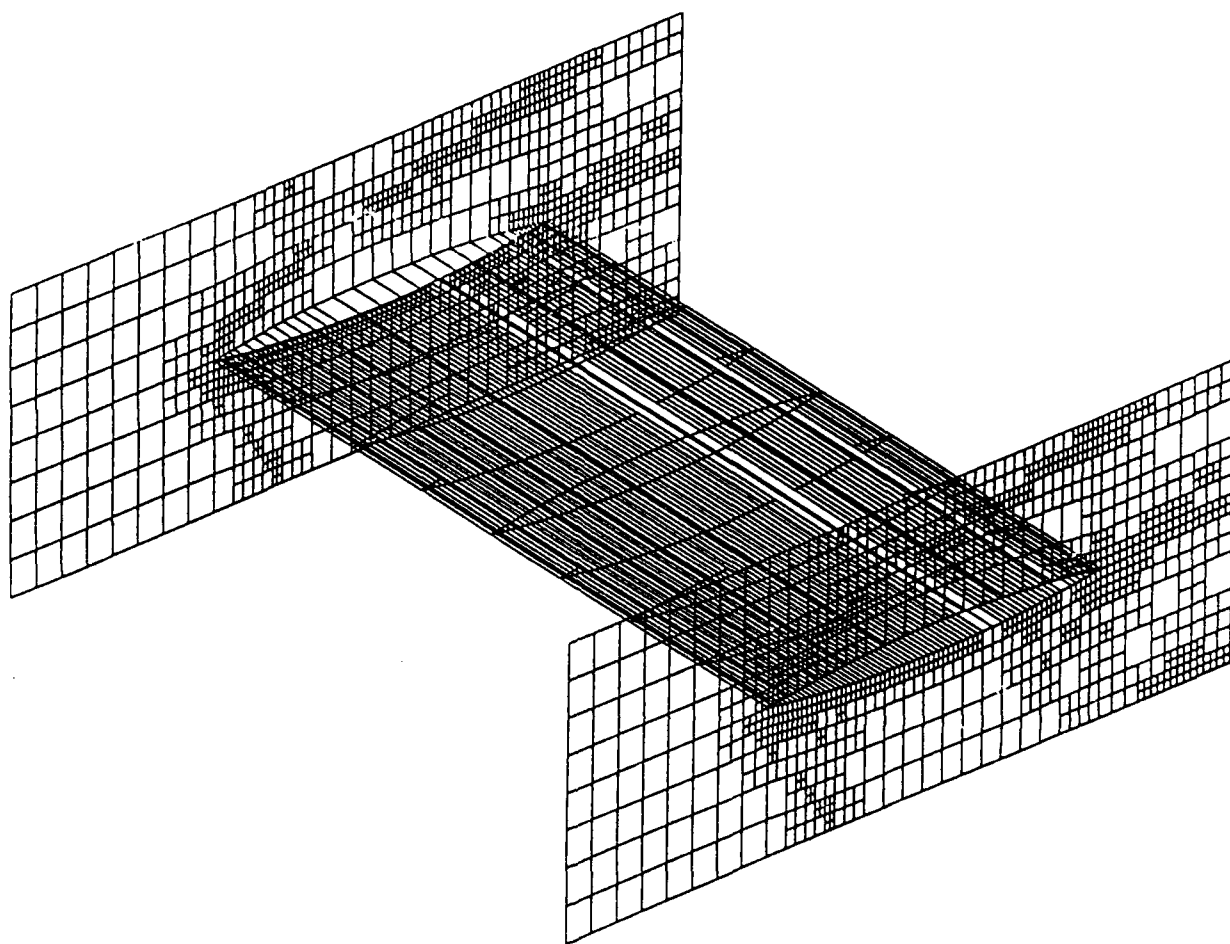


Figure 7: Supersonic flow over a rigid blade in motion, adapted grid at 300 steps.

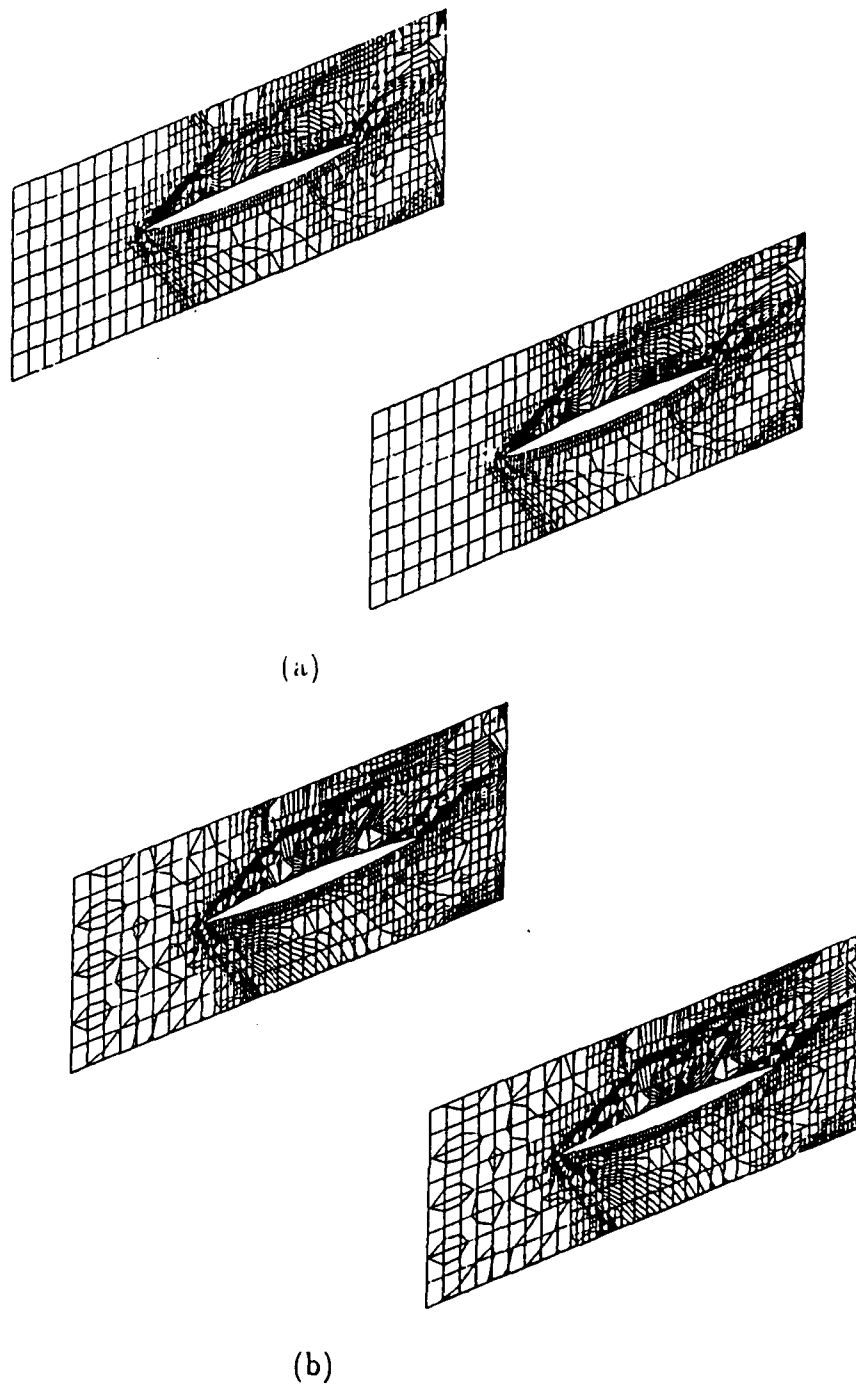
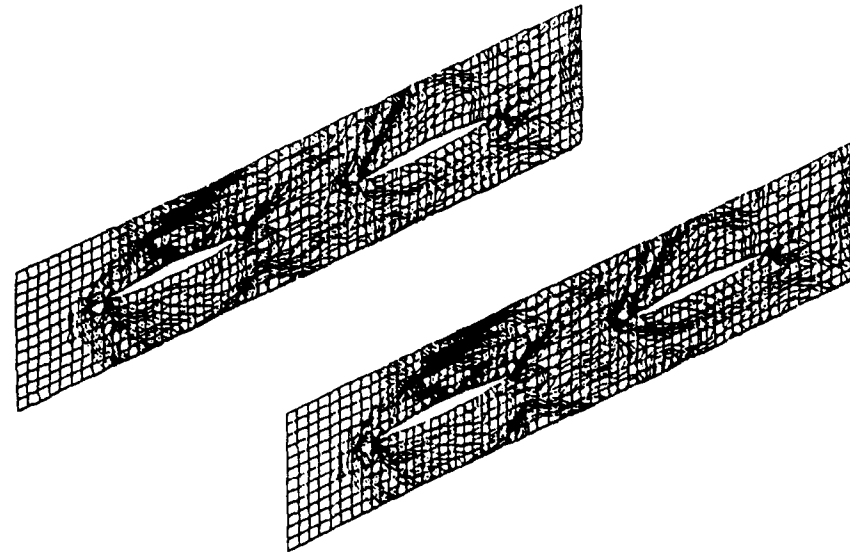
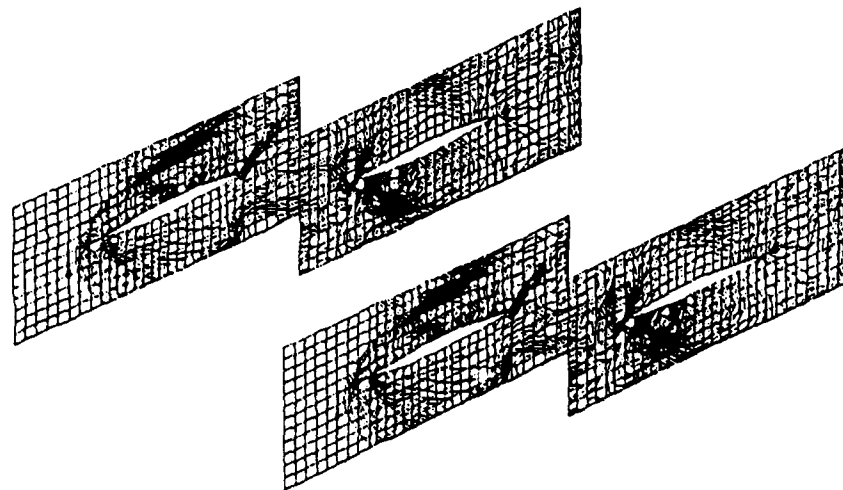


Figure 8: Supersonic flow over a rigid blade in motion. (a) Pressure contours after 300 time steps, (b) density contours after 300 time steps.





(a)



(b)

Figure 9: Rotor-stator interaction simulation. (a) Density contours after 1 cycle, (b) density contours after  $\sim 1.5$  cycles.

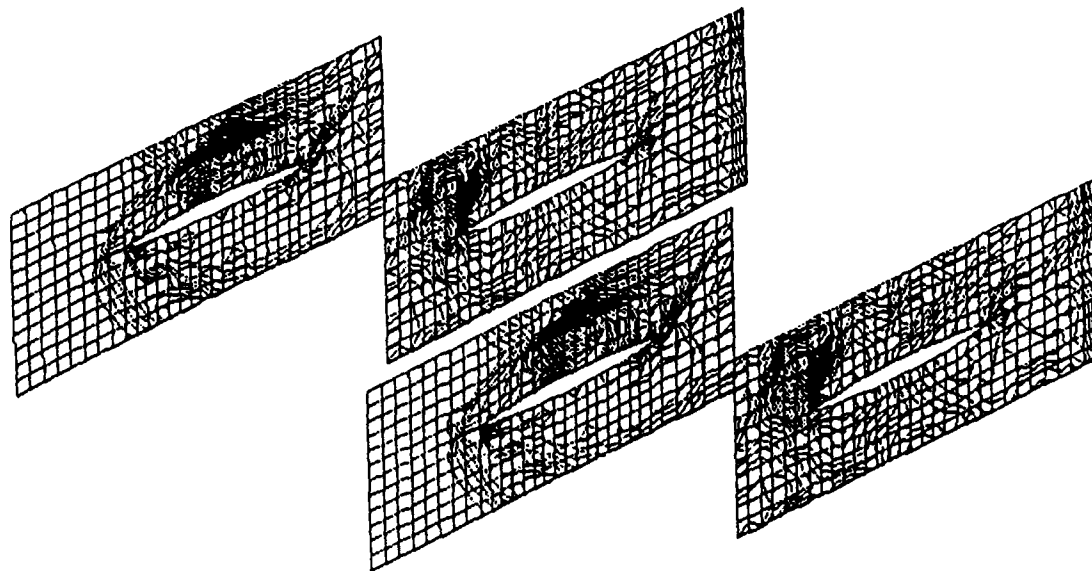


Figure 10: Rotor-stator interaction simulation, density contours for  $\sim 6$  cycles.

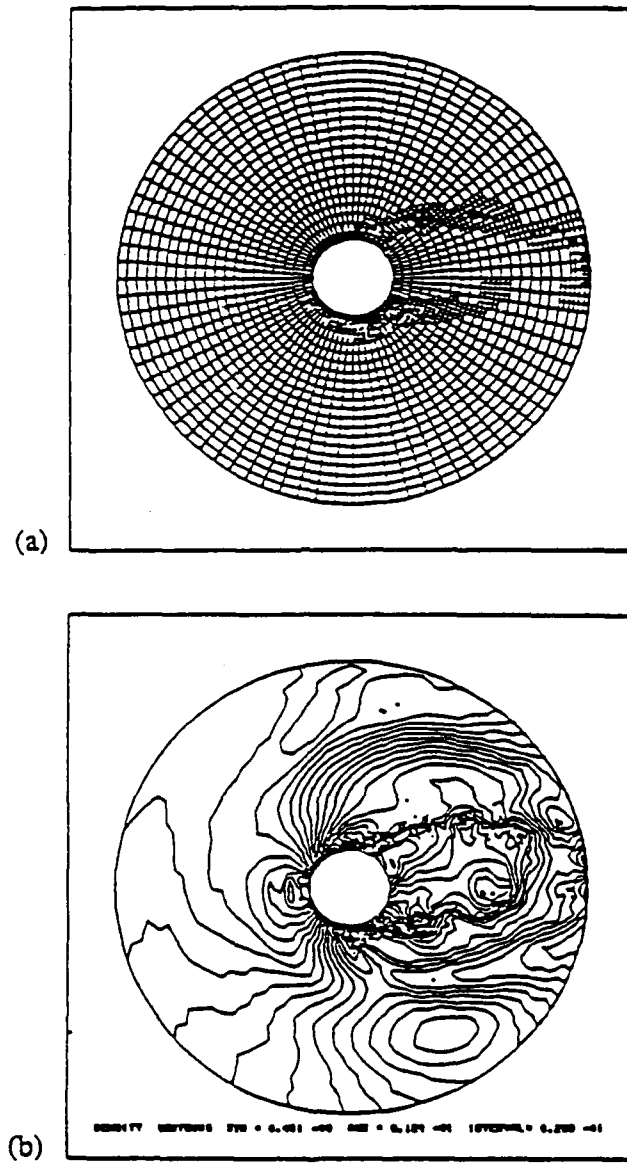


Figure 11: Viscous cylinder problem with  $M = 0.64$ , flow perturbed after 2000 time steps. Vortices are generated and shed; (a) instantaneous grid and (b) density contours.

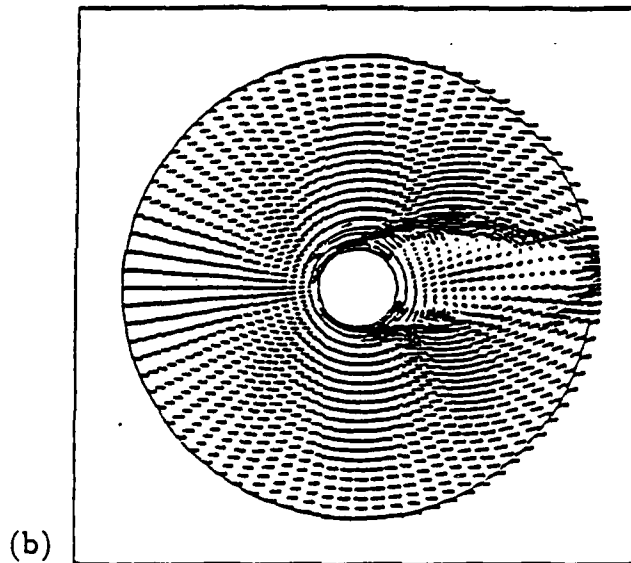
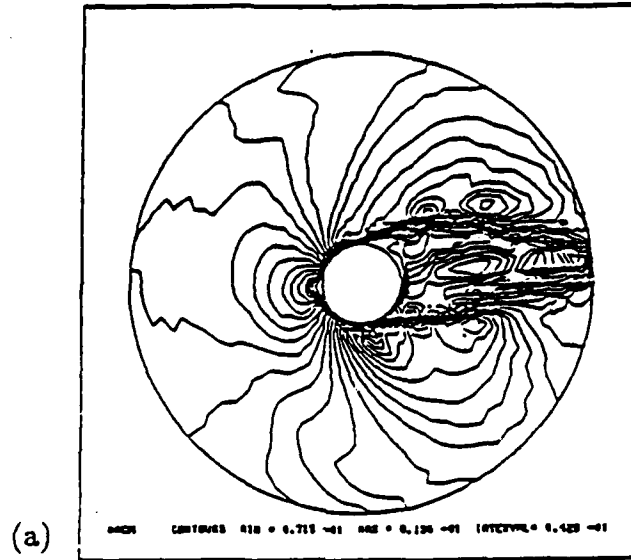
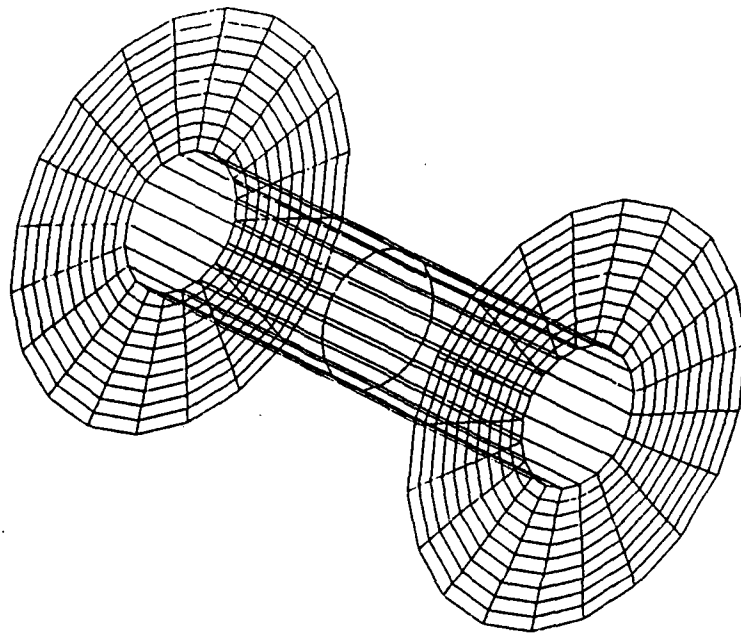
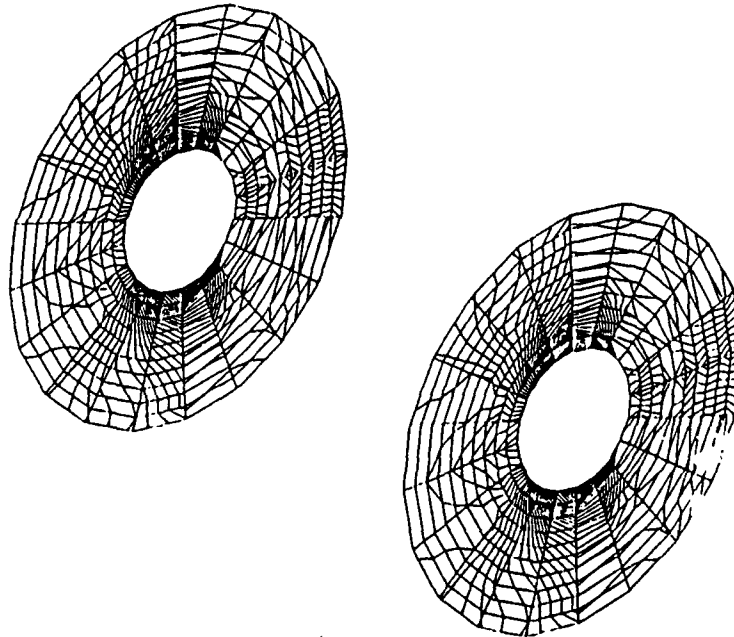


Figure 12: (a) Computed Mach number contours and (b) velocity vectors for viscous cylinder problem.

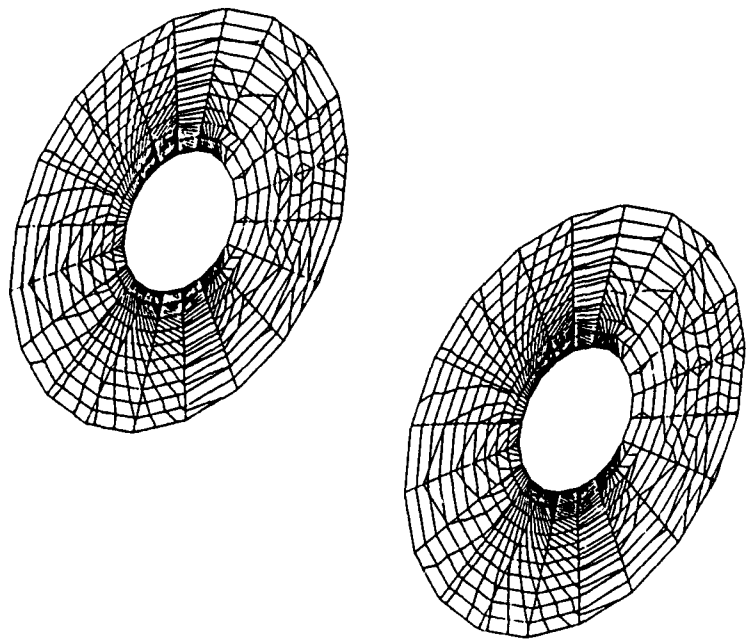


(a)

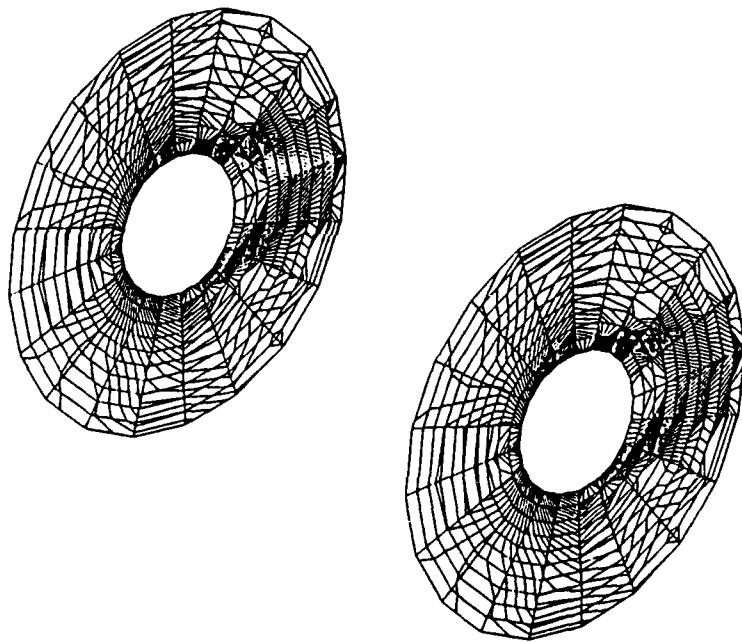


(b)

Figure 13: Subsonic flow past a rigid cylinder, Mach = 0.41. (a) Initial grid, (b) density contours.

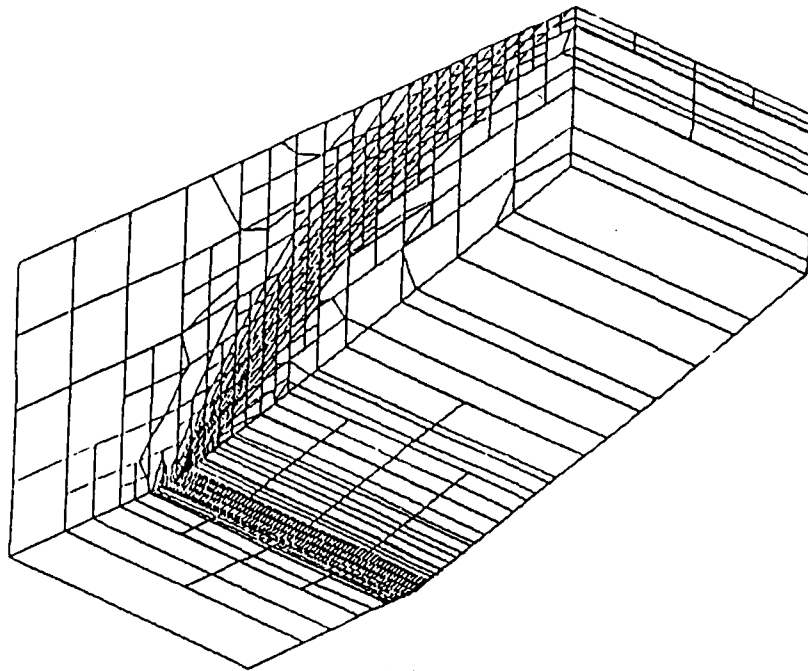


(a)

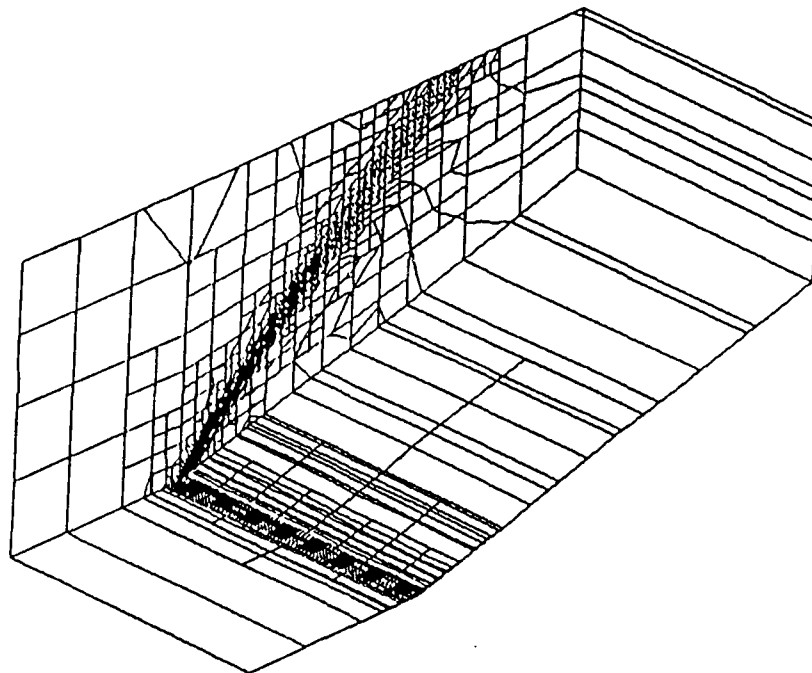


(b)

Figure 14: Subsonic flow past a rigid cylinder. (a) Pressure contours, (b) Mach contours.

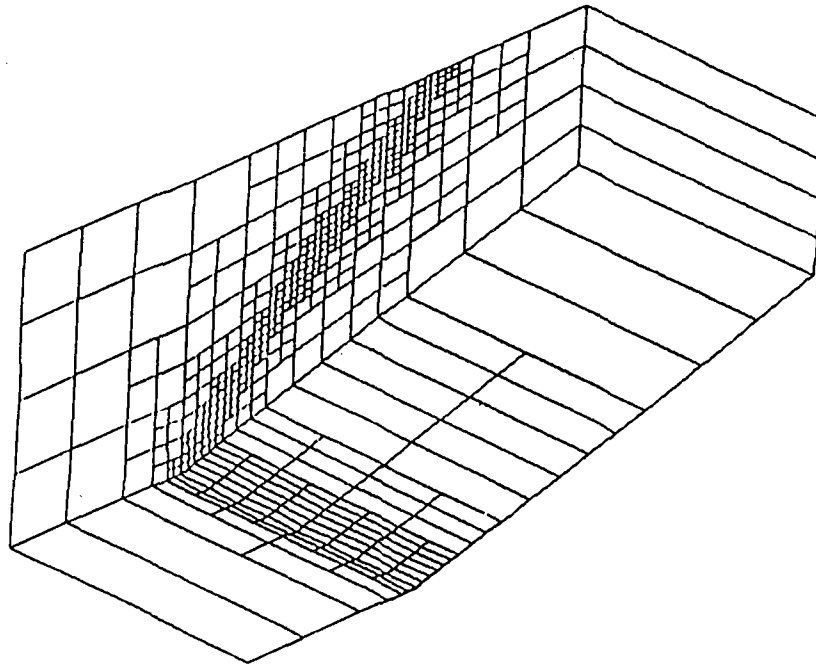


(a)

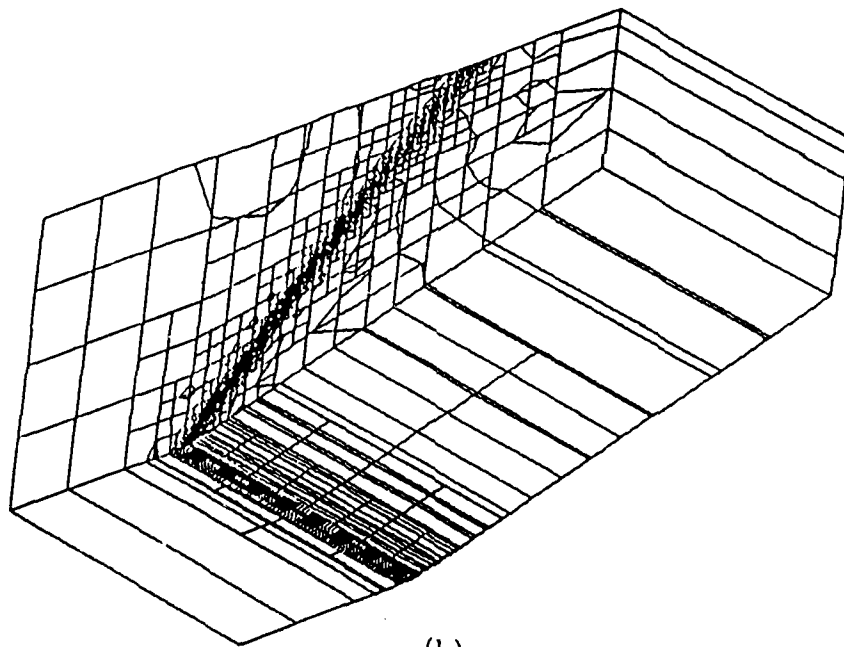


(b)

Figure 15: Supersonic flow over a  $15^\circ$  ramp. (a) Density contours at 60 steps, (b) density contours at 120 steps.



(a)



(b)

Figure 16: Supersonic flow over a  $15^\circ$  ramp. (a) Final adapted grid, (b) density contours at 160 steps.



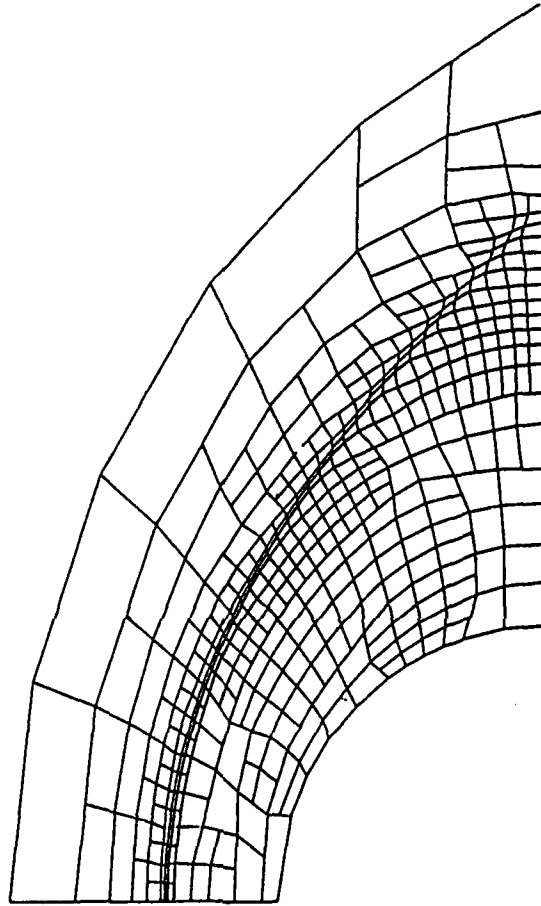


Figure 17: An  $h$ - $r$  adaptive mesh with combined node relocation and mesh refinement.

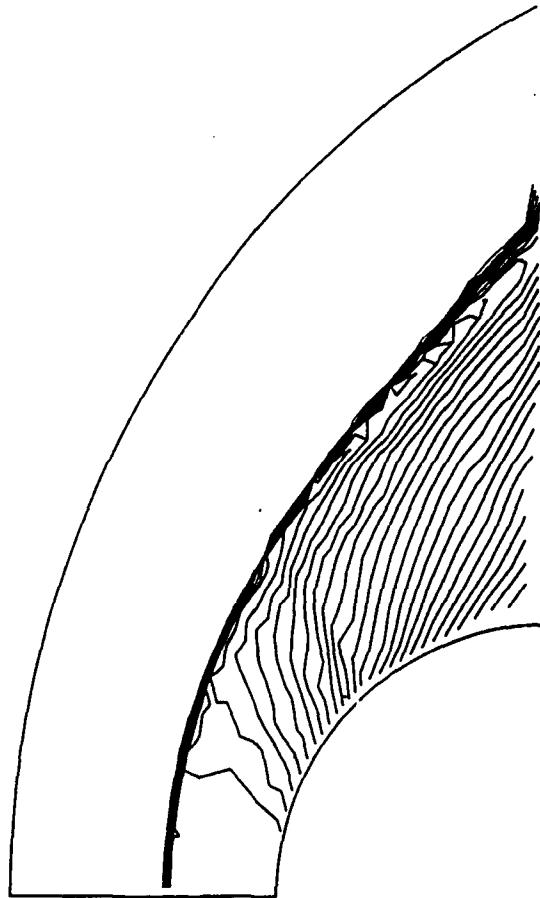


Figure 18: Computed density contours for supersonic flow calculation of bow shocks on a blunt body.

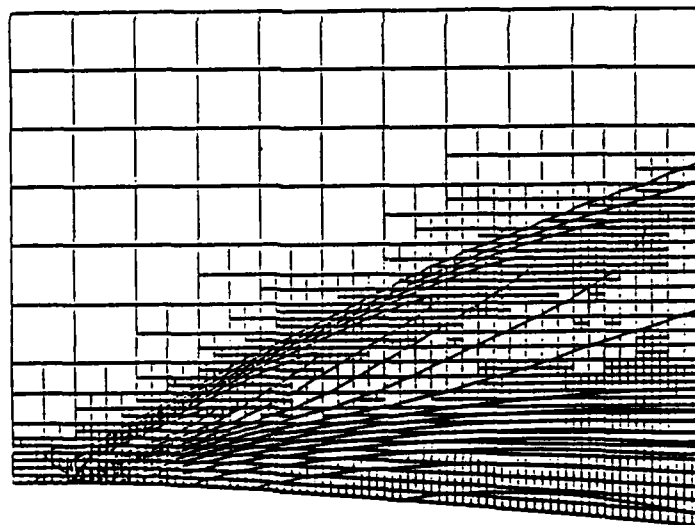
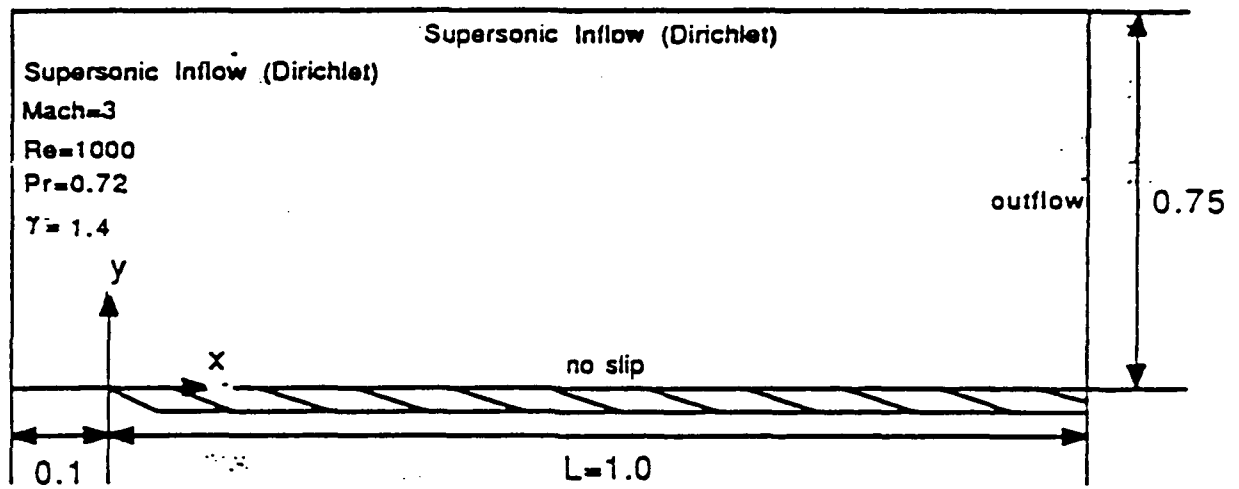


Figure 19: An instantaneous adapted  $h$ - $p$  mesh with overlaid density contours for viscous compressible flow over a deforming elastic plate; high-order spectral elements are used to model the viscous boundary layer while  $h$ -adaptive refinement is used to capture the shock.

Figure 19 contains an optimal  $h$ - $p$  mesh and density contours over a flexible elastic plate deformed by the action of a viscous incompressible fluid. Low-order elements are used to capture the shock while higher-order elements are used to model the viscous boundary layer. The flow is quasi-steady and a fully implicit solver with a multigrid iterative solver is used to compute successive optimal meshes as the plate deforms.

## Acknowledgements

*We wish to acknowledge the help of several colleagues with our work on adaptive CFD methods. Specifically, thanks are due Leszek Demkowicz, Theofanis Strouboulis, Philippe Devloo, and Waldek Rachowicz who have contributed to various components of the research since 1984. The work on  $h$ - $r$  methods was done in collaboration with Michael Edwards.*

## 6 References

1. Oden, J. T., "Adaptive Methods in Computational Fluid Dynamics," *Finite Element Analyses in Fluids*, Edited by T-J Chung and G. R. Karr, UAH Press, Huntsville, pp. 2-10, 1989.
2. Oden, J. T., "Progress in Adaptive Methods in Computational Fluid Dynamics," *Adaptive Methods in Partial Differential Equations*, Edited by J. Flaherty, SIAM Publications, Philadelphia (to appear).
3. Demkowicz, L., and Oden, J. T., "A Review of Local Mesh Refinement Techniques and Corresponding Data Structures in  $h$ -Adaptive Finite Element Methods," **TICOM Report TR-88-02**, The University of Texas at Austin, 1988.
4. Patera, A., "Advances and Future Directions of Research on Spectral Methods," in *Computational Mechanics Advances and Trends*, Edited by A. K. Noor, American Society of Mechanical Engineers Publication, AMD, Vol. 75, pp. 441-428, 1986.
5. Szabo, B. A., "Estimation and Control of Error Based on  $p$ -Convergence," in I. Babuška, O. C. Zienkiewicz, J. P. de S. R. Gago and A. de Oliveira, (Eds.), **Adaptive Methods and Error Refinement in Finite Element Computations**, John Wiley and Sons, Ltd., London, 1986.
6. Demkowicz, L., Oden, J. T., Strouboulis, T., and Devloo, P., "An Adaptive  $p$ -Version Finite Element Method for Transient Flow Problems With Moving Boundaries," in **Finite Elements in Fluids**, Vol. 6, Edited by R. H. Gallagher, G. F. Carey, J. T., Oden, and O. C. Zienkiewicz, John Wiley and Sons Ltd., Chichester, pp. 291-305, 1985.

7. Demkowicz, L., Oden, J. T., Rachowicz, W., and Hardy, O., "Toward a Universal  $h$ - $p$  Finite Element Strategy, Part 1. Data Structure and Constrained Approximation," *Computer Methods in Applied Mechanics and Engineering*, (to appear).
8. Diaz, Kikuchi, N., and Taylor, J., "A method of Grid Optimization for Finite Element Methods," *Computer methods in Applied Mechanics and Engineering*, Vol. 41, pp. 29-45, 1983.
9. Oden, J. T., and Demkowicz, L., "Advances in Adaptive Improvements: A Survey of Adaptive Methods in Computational Fluid Mechanics," **State of the Art Surveys in Computational Mechanics**, Edited by A. K. Noor and J. T. Oden, American Society of Mechanical Engineers, N.Y., 1989.
10. Oden, J. T., Demkowicz, L., and Strouboulis, T., "Adaptive Finite Element Methods for Flow Problems With Moving Boundaries. I: Variational Principles and *A Posteriori* Estimates," *Computer methods in Applied Mechanics and Engineering*, Vol. 46, pp. 217-251, 1984.
11. Oden, J. T., Demkowicz, L., Strouboulis, T., and Devloo, P., "Adaptive Methods for Problems in Solid and Fluid Mechanics," **Accuracy Estimates and Adaptive Refinements in Finite Element Computations**, Edited by I. Babuška, O. C. Zienkiewicz, J. Gago, and E. R. A. de Oliveira, John Wiley and Sons Ltd., London, 1986.
12. Demkowicz, L., and Oden, J. T., "An Adaptive Characteristic Petrov-Galerkin Finite Element Method for Convection-Dominated Linear and Nonlinear Parabolic Problems in One Space Variables," *Journal of Computational Physics*, Vol. 68, No. 1, pp. 188-273, 1986.
13. Oden, J. T., and Bass, J. M., "Adaptive Finite Element Methods for a Class of Evolution Problems in Viscoplasticity," *International Journal of Engineering Science*, Vol. 25, No. 6, pp. 623-653, 1987.
14. Oden, J. T., Strouboulis, T., and Devloo, P., "Adaptive Finite Element Methods for the Analysis of Inviscid Compressible Flow: I. Fast Refinement/Unrefinement and Moving Mesh Methods for Unstructured Meshes," *Computer Methods in Appl. Mech. and Engrg.*, Vol. 59, No. 3, 1986.
15. Oden, J. T., Devloo, P., and Strouboulis, T., "Implementation of an Adaptive Refinement Technique for the SUPG Algorithm," *Computational Methods in Appl. Mech. and Engrg.*, Vol. 61, pp. 339-358, 1987.

16. Oden, J. T., Strouboulis, T., Devloo, P., and Howe, M., "Recent Advances in Error Estimation and Adaptive Improvement of Finite Element Calculations," **Computational Mechanics Advances and Trends**, Edited by A. K. Noor, AMD, Vol. 75, ASME, pp. 369-410, 1986.
17. Oden, J. T., "Adaptive Finite Element Methods for Problems in Solid and Fluid Mechanics," **Finite Element Theory and Application Overview**, Edited by R. Voight, Springer-Verlag, N.Y., 1988.
18. Oden, J. T., Strouboulis, T., and Devloo, P., "Adaptive Finite Element Methods for Compressible Flow Problems," **Numerical Methods for Compressible Flows — Finite Difference, Element and Volume Techniques**, Edited by T. E. Tezduyar and T. J. R. Hughes, AMD, Vol. 78, ASME, New York, pp. 115-126, 1987.
19. Oden, J. T., Strouboulis, T., and Devloo, P., "Adaptive Finite Element Methods for High-Speed Compressible Flows," *International Journal for Numerical Methods in Fluids*, Vol. 7, pp. 1211-1228, 1987.
20. Devloo, P., Oden, J. T., and Pattani, P., "An  $h$ - $p$  Adaptive Finite Element Method for the Numerical Simulation of Compressible Flow," *Computer Methods in Applied Mechanics and Engineering*, Vol. 70, pp. 203-235, 1988.
21. Bass, J. M., and Oden, J. T., "Adaptive Computational Methods for Chemically-Reacting Radiative Flows," *International Journal of Engineering Science*, Vol. 26, No. 9, pp. 959-992, 1988.
22. Oden, J. T., Strouboulis, T., and Bass, J. M., "Paradigmatic Error Calculations for Adaptive Finite Element Approximations of Convection Dominated Flows," **Recent Advances in Computational Fluid Dynamics**, ASME Publication.
23. Demkowicz, L., and Oden, J. T., "An Adaptive Characteristic Petrov-Galerkin Finite Element Method for Convection-Dominated Linear and Nonlinear Parabolic Problems in Two Space Variables," *Computer Methods in Applied Mechanics and Engineering*, Vol. 55, pp. 63-87, 1986.
24. Demkowicz, L., and Oden, J. T., "On a Mesh Optimization Method based on a Minimization of Interpolation Error," *International Journal of Engineering Science*, Vol. 24, No. 1, pp. 55-68, 1986.
25. Oden, J. T., "Adaptive FEM in Complex Flow Problems," **The Mathematics of Finite Elements With Applications**, Edited by J. R. Whiteman, London Academic Press Ltd., Vol. 6, pp. 1-29, 1988.

26. Oden, J. T., Demkowicz, L., Westermann, T., and Rachowicz, W., "Toward a Universal  $h$ - $p$  Adaptive Finite Element Strategy, Part 2. A *Posteriori* Error Estimates," *Computer Methods in Applied Mechanics and Engineering*, (to appear).
27. Rachowicz, W., Oden, J. T., and Demkowicz, L., "Toward a Universal  $h$ - $p$  Adaptive Finite Element Strategy, Part 3. A Study of the Design of  $h$ - $p$  Meshes," *Computer Methods in Applied Mechanics and Engineering*, (to appear).
28. Oden, J. T., "Smart Algorithms and Adaptive Methods in Computational Fluid Dynamics," *Proceedings CANCAM (Canadian Congress on Applied Mechanics)*, Ottawa, Canada, 1989.
29. Edwards, J., Oden, J. T., and Demkowicz, L., "An  $h$ - $r$  Adaptive TVD Scheme With Grid Movements for the Euler Equations in Two Dimensions," (in review).
30. Demkowicz, L., and Oden, J. T., "A Review of Local Mesh Refinement Techniques and Corresponding Data Structures in  $h$ -Type Adaptive Finite Element Methods," **TICOM Report TR 88-2**, The University of Texas at Austin, 1988.

EFFECT OF CONSTITUTIVE MODELLING ON THE  
DYNAMIC DEVELOPMENT OF SHEAR BANDS IN  
VISCOPLASTIC MATERIALS\*

R. C. Batra and C. H. Kim  
Department of Mechanical and Aerospace Engineering  
and Engineering Mechanics  
University of Missouri-Rolla  
Rolla, MO 65401

**ABSTRACT.** We model the viscoplastic response of a HY-100 steel by a Power law, and flow rules proposed by Litonski, Bodner and Partom, and Johnson and Cook. Each of these flow rules is first calibrated by using the torsional test data at a strain-rate of  $3,300 \text{ sec}^{-1}$ . These material models are then used to study the thermomechanical deformations of a block made of the HY-100 steel and undergoing simple shearing deformations at a nominal strain-rate of  $5000 \text{ sec}^{-1}$ . A material defect is simulated by assuming a non-uniform initial temperature distribution within the block. Whereas all of the flow rules used predict a rapid drop of the shear stress as a shear band forms, only for the Litonski Law for nonpolar materials, does an unloading elastic wave emanate outwards from the shear band.

**INTRODUCTION.** Noting that Batra (1987) has briefly reviewed the work done on shear bands till 1986, we discuss below some of the work done since then. For strain-rate hardening but thermally softening materials Wright and Walter (1987) found that the shear stress within a band collapses rapidly as the band grows. Batra and Kim (1989a) accounted also for material elasticity and work hardening effects and found that if the rate of collapse of the shear stress is large, then an unloading elastic wave emanates outwards from the shear band and propagates towards the boundaries of the specimen. The development of shear bands in plane strain problems have been studied, among others, by Anand et al. (1988), Needleman (1989), LeMonds and Needleman (1986a,1986b), Batra and Liu (1989a,1989b). These works have employed different flow rules and have modeled a material defect by introducing either a temperature perturbation or assuming the existence of a weak material at the site of the defect. Batra and Kim (1989b) have recently studied the development of a shear band in a block of HY-100 steel undergoing overall simple shearing adiabatic deformations and compared computed results with the experimental observations of Marchand and Duffy (1988). They found that the dipolar theory due to Wright and Batra (1987) and Batra (1987, 1989) and the Bodner-Partom (1975) law predict most of the features of the shear band.

We note that Molinari and Clifton (1987), Tzavaras (1987) and Wright (1989) have studied the problem analytically. For rigid/perfectly plastic materials, Wright (1989) has developed a criterion that ranks materials according to their tendency to form adiabatic shear bands. Hartley et al. (1987), Giovanola (1987), and Marchand and Duffy (1988) have reported the observed histories of the temperature and strain within a band as it develops.

Here we presume that the torsional experiments on thin-walled steel tubes can be analyzed by studying the thermomechanical deformations of a viscoplas-

\*Supported by the U.S. Army Research Office Contract DAAL 03-88-K-0184 to the University of Missouri-Rolla.



tic block undergoing overall adiabatic simple shearing deformations. We find the values of the material parameters appearing in different flow rules by solving an initial-boundary-value problem and comparing computed results with the experimental stress-strain curve at a nominal strain-rate of  $3,300 \text{ sec}^{-1}$ . These flow rules are then used to compute the initiation and growth of a shear band when the applied nominal strain-rate is  $5,000 \text{ sec}^{-1}$ . It is found that the rate of stress drop during the growth of a shear band as predicted by the Bodner-Partom law and the dipolar theory due to Wright and Batra (1987) is similar to that observed experimentally.

GOVERNING EQUATIONS. In terms of non-dimensional variables, equations governing the thermomechanical deformations of a viscoplastic block undergoing overall adiabatic deformations are (e.g. see Batra and Kim (1989a))

$$\rho \dot{v} = (s - \ell \sigma, y), y \quad 0 < y < 1, \quad (2.1)$$

$$\dot{\theta} = k \theta, yy + s \dot{\gamma}_p + \ell \sigma \dot{d}_p, \quad 0 < y < 1, \quad (2.2)$$

$$\dot{s} = \mu (v, y - \dot{\gamma}_p), \quad (2.3)$$

$$\dot{\sigma} = \mu \ell (v, yy - \dot{d}_p), \quad (2.4)$$

$$\dot{\gamma}_p = g(s, \sigma, \gamma_p, d_p, \theta, \ell), \quad (2.5)$$

$$\dot{d}_p = h(s, \sigma, \gamma_p, d_p, \theta, \ell). \quad (2.6)$$

These equations, written for dipolar materials, reduce to those for non-polar materials when  $\ell$  is set equal to zero. Here  $\rho$  is the mass density,  $v$  the velocity of a material particle in the direction of shearing, a superimposed dot indicates the material time derivative,  $s$  is the shearing stress,  $\ell$  a material characteristic length,  $\sigma$  the dipolar stress, and a comma followed by  $y$  signifies partial differentiation with respect to  $y$ . Furthermore,  $k$  is the thermal conductivity,  $\gamma_p$  the plastic strain-rate,  $d_p$  the dipolar plastic strain-rate,  $\mu$  the shear modulus, and  $\theta$  is the temperature change from that in the reference configuration. Equation (2.1) expresses the balance of linear momentum and (2.2) the balance of internal energy, equations (2.3)-(2.6) are constitutive relations. The different viscoplastic flow rules differ in the functional forms of  $g$  and  $h$  and are given below in the next section.

For the initial conditions we take

$$v(y, 0) = 0, s(y, 0) = 0, \sigma(y, 0) = 0, \theta(y, 0) = \epsilon(1-y^2)^9 e^{-5y^2}. \quad (2.7)$$

That is, in the initial rest state of the block, it is taken to be stress free. The initial temperature distribution simulates the defect or inhomogeneity in the block assumed to be present near the point  $y = 0$  and the value of  $\epsilon$  represents the strength of the defect.

We presume that the overall deformations of the block are adiabatic and the lower surface is at rest while the upper surface is assigned a velocity that increases linearly from 0 to 1 in time  $t_r$  and then stays at the constant value of 1.0. Thus,

$$\theta, y(0, t) = 0, \theta, y(1, t) = 0, v(0, t) = 0, \quad (2.8)$$

$$v(1, t) = t/t_r, \quad 0 \leq t \leq t_r, \quad (2.9)$$

$$-1, t \geq t_r,$$

and for dipolar materials, we also assume that

$$\sigma(0,t) = 0, \sigma(1,t) = 0. \quad (2.10)$$

Computations for the domain  $-1 \leq y \leq 1$  and with boundary conditions  $\sigma(-1,t) = 0, \sigma(1,t) = 0$  have given  $\sigma(0,t) = 0$ .

3. VISCOPLASTIC FLOW RULES. In order to calibrate the various flow rules against the shear stress-shear strain curve given by Marchand and Duffy (1988) for a strain-rate of  $3,300 \text{ sec}^{-1}$ , we solved numerically, the initial-boundary-value problem outlined above with

$$s(y,0) = 1.0, \gamma_p(y,0) = 0.012, v(y,0) = y, \theta(y,0) = 0^\circ \text{ c}, \epsilon = 0,$$

$$t_r = 0.033, \rho = 7,860 \text{ kg/m}^3, c = 473 \text{ J/kg}^\circ\text{c}, k = 49.73 \text{ w/m}^2 \text{ }^\circ\text{c}, H = 2.5 \text{ mm},$$

$$\dot{\gamma}_0 = 3,300 \text{ sec}^{-1}.$$

Here  $H$  is the height of the block and  $\dot{\gamma}_0$  is the average applied strain-rate. With no initial temperature perturbation, the block deforms uniformly and homogeneously and the dipolar effects vanish identically. As far as possible we kept the values of the strain-hardening exponent and the strain-rate-hardening exponent equal to those given by Marchand and Duffy (1988), and adjusted the values of other parameters till the computed stress-strain curve came out close to that given by Marchand and Duffy.

3.1 Litonski's Law for Nonpolar and Dipolar Materials. Wright and Batra (1987) generalized the constitutive relation proposed by Litonski (1977) to be applicable to nonpolar and dipolar materials. Batra and his co-workers (1987, 1988, 1989) have used it to study the initiation and growth of shear bands. It may be written as:

$$\dot{\gamma}_p = \Lambda s, \dot{d}_p = \frac{\Lambda}{l} \sigma, \quad (3.1)$$

$$\Lambda = \max \left[ 0, \left\{ \frac{s_e}{(1-\alpha\theta) \left(1 + \frac{\varphi}{\varphi_0}\right)^n} \right\}^{1/m} - 1 \right] / b s_e, \quad (3.2)$$

$$s_e = (s^2 + \sigma^2)^{1/2}, \quad (3.3)$$

$$\dot{\varphi} = \Lambda s_e^2 / \left(1 + \frac{\varphi}{\varphi_0}\right)^n. \quad (3.4)$$

Here  $\varphi$  can be viewed as an internal variable that describes the work hardening of the material. Its evolution is given by equation (3.4). In equation (3.2),  $(1-\alpha\theta)$  describes the softening of the material due to its heating,  $b$  and  $m$  characterize its strain-rate hardening, and  $\varphi_0$  and  $n$  its work hardening. The following values of material parameters resulted in a stress-strain curve that was close to the one observed experimentally.

$$\alpha = 0.00185/^\circ\text{c}, \varphi_0 = 0.012, n = 0.107, m = 0.0117, b = 10^4 \text{ sec}, \ell = 0.005$$

3.2 Power Law. For nonpolar materials and assuming that there is no loading surface, this flow rule for the HY-100 steel can be written as

$$\dot{\gamma}_p = (10^{-4}) s^{85.47} \left( \frac{\gamma}{0.012} \right)^{9.145} \left( \frac{\theta}{300} \right)^{-64.103} \quad (3.5)$$

Here  $\theta$  is the current temperature in degrees Kelvin and  $\gamma$  is the total strain at a material particle.

3.3 Bodner-Partom Law. For the HY-100 steel, the constitutive relation proposed by Bodner and Partom (1972), can be written as

$$\dot{\gamma}_p = 10^8 \exp \left[ -\frac{1}{2} \left( \frac{K^2}{3s^2} \right)^n \right], \quad n = \frac{1200}{\theta}, \quad K = 1600 - 300 \exp(-5 W_p) \quad (3.6)$$

Here  $\theta$  is the absolute temperature of a material particle and  $W_p$  is the plastic work done.

3.4. Johnson-Cook Law. The constitutive relation proposed by Johnson and Cook (1983) takes the following form for the HY-100 steel.

$$\dot{\gamma}_p = \exp \left[ \left\{ \frac{s}{(0.45 + 1.433 \gamma_p^{0.107}) (1-T)^{0.7}} - 1.0 \right\} / 0.0277 \right], \quad (3.7)$$

$$T = (\theta - \theta_0) / 1200.$$

Here  $\theta_0$  equals the ambient temperature.

4. DETERMINATION OF THE SIZE OF THE PERTURBATION. Here we model the cumulative effect of the change in the thickness of the specimen and possibly the slight variation in the material properties by assuming a nonuniform initial temperature distribution as given by Eqn. (2.7). For different flow laws, the value of  $\epsilon$  was determined so as to initiate a shear band, as signified by a rapid drop in the shear stress, at a value of the average strain close to that found experimentally. The initial-boundary-value problem outlined in Section 2 with  $t_r = 0.033$  was solved by the finite element method. Values of  $\epsilon$  equal to  $1^\circ\text{c}$ ,  $2^\circ\text{c}$ ,  $5^\circ\text{c}$  and  $9^\circ\text{c}$  for the Litonski Law for nonpolar and dipolar materials, Power Law, and the Bodner-Partom Law and the Johnson-Cook Law, respectively, result in stress-strain curves shown in Fig. 1.

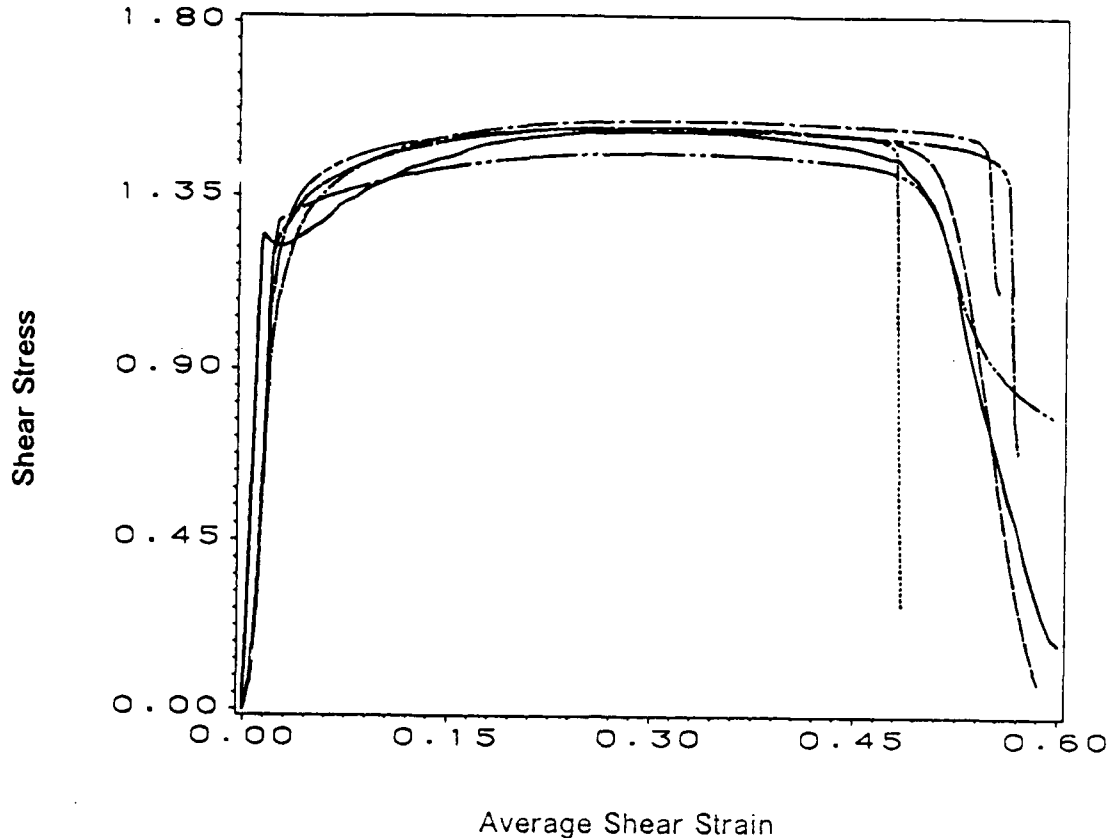


Fig. 1. Shear stress-shear strain curves computed with different flow rules and with different initial temperature perturbations.

\_\_\_\_\_ experimental, ..... Bodner-Partom, ..... Litonski (non-polar), \_\_\_\_\_ Litonski (dipolar) \_\_\_\_\_ Power,  
 \_\_\_\_\_ Johnson-Cook.

These curves vividly reveal that until the time the shear stress begins to drop rapidly, all of the flow rules considered predict material behavior in reasonable agreement with the experimental observations. For nonpolar materials Litonski's Law, the Power Law and the Johnson-Cook Law give essentially a catastrophic drop in the shear stress with virtually no increase in the nominal shear strain. This does not agree with the experimental data since Marchand and Duffy observed that during the drop of the shear stress, the nominal strain increases by approximately 5 percent. The Litonski Law for dipolar materials and the Bodner-Partom Law for nonpolar materials do predict the gradual drop in the shear stress in agreement with the experimental data. However, for the Bodner-Partom Law the shear stress does not drop as much as it does during the tests and it reaches a plateau.

5. RESULTS FOR A NOMINAL STRAIN-RATE OF 5,000 SEC.<sup>-1</sup>. With the values of material parameters and the size of the temperature perturbation found above kept fixed, we increased the prescribed velocity on the upper boundary so as to deform the block at a nominal strain-rate of 5,000 sec<sup>-1</sup>. Note that the values of some of the non-dimensional variables appearing in the governing equations will change. For each one of the flow rules used, the shear stress attained a maximum value when the average shear strain was approximately equal to 0.30. For subsequent deformations, we have plotted in Figs. 2 and 3 the evolution of the shear stress and the particle velocity within the specimen. The value of the nominal shear strain at which the shear stress drops and the shear band initiates depends upon the flow rule used. However, in each case, the value of the nominal shear strain when a band initiates is noticeably more than the value at which the shear stress attains a maximum value.

For nonpolar materials, the rate of drop of the shear stress is highest for the Litonski law as compared to that for the other three flow rules used. For the Bodner-Partom law, the shear stress drops initially, but then seems to reach a plateau. For the Power law, the shear stress oscillates both in space and time and there was no unloading wave observed. With the Johnson-Cook law, the shear stress drops almost as rapidly as with the Litonski law, but seems to stay uniform throughout the specimen. For the Litonski law, as the shear stress drops, an unloading elastic wave emanates out of the shear band and travels towards the other end of the specimen. Batra and Kim (1989a) found this unloading wave and their computed wave speed was very close to the analytical value of  $(\mu/\rho)^{1/2}$ . The propagation of the wave is more clear from the particle velocity plot depicted in Fig. 3. We note that we assumed the existence of a yield surface only for the Litonski law. For other flow rules, plastic deformations are assumed to occur at all times.

For nonpolar materials, only Litonski's law as generalized by Wright and Batra was used. In this case, even though the shear stress drop was larger near the center as compared to that for nonpolar materials, no wave phenomenon was noticed. This becomes transparent from the velocity plot in Fig. 3.

For nonpolar materials, the velocity plots indicate that the particle velocity increases rapidly from zero at  $y = 0$  to as high as 2 at a point close to  $y = 0$  and then decreases to the prescribed value of 1 at  $y = 1.0$ . The overshoot in the particle velocity is highest for the Litonski law. The flow rule used affects the evolution of the particle velocity significantly. With the Johnson-Cook law, no oscillations in the particle velocity are observed. With the Bodner-Partom flow rule, no spatial oscillations in the particle velocity are seen but after a shear band has initiated, the velocity of a material particle oscillates in time. The spatial and temporal variation in the particle velocity with the Power law is noticeably different from that computed with the other three flow rules. A glance at the velocity and the shear stress plot seems to indicate that there is no unloading wave emanating out of the shear band in this case.

6. CONCLUSIONS. For overall adiabatic simple shearing thermomechanical deformations of a viscoplastic block, we first calibrated the four different flow rules so as to give essentially identical shear stress-shear strain curves at a nominal strain-rate of 3,300 sec<sup>-1</sup>. Then, the size of the initial temperature perturbation was adjusted to yield the initiation of the shear band, as indicated by a significant drop in the shear stress for very little change in the nominal shear strain, at almost the same value of the nominal

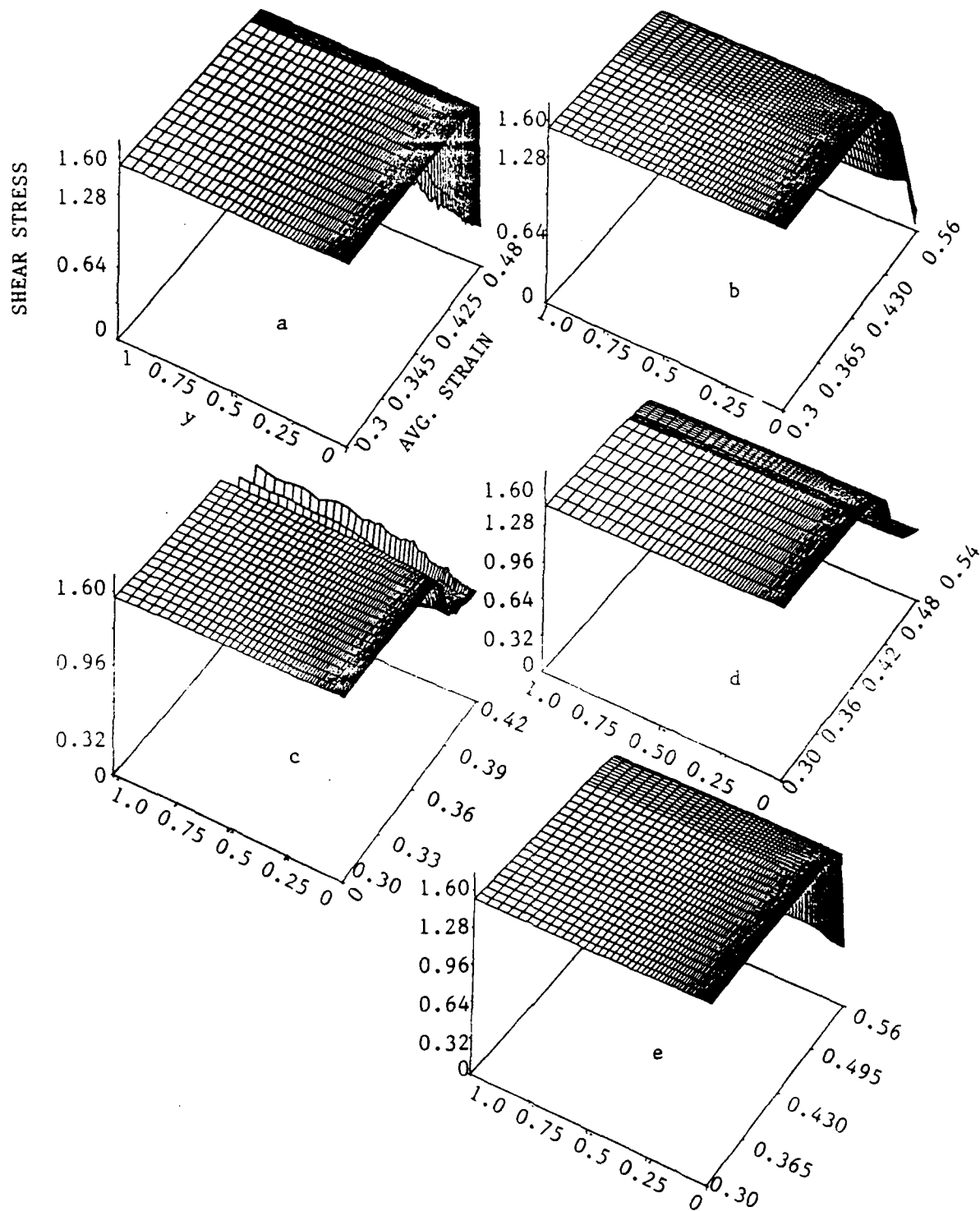


Fig. 2. The Evolution of the Shear Stress Within the Specimen After the Shear Stress has Attained Its Peak Value.

- |                  |  |
|------------------|--|
| (a) Litonski Law | (b) Litonski's Flow Rule for Dipolar Materials |
| (c) Power Law    | (d) Bodner-Partom Law, (e) Johnson-Cook Law    |

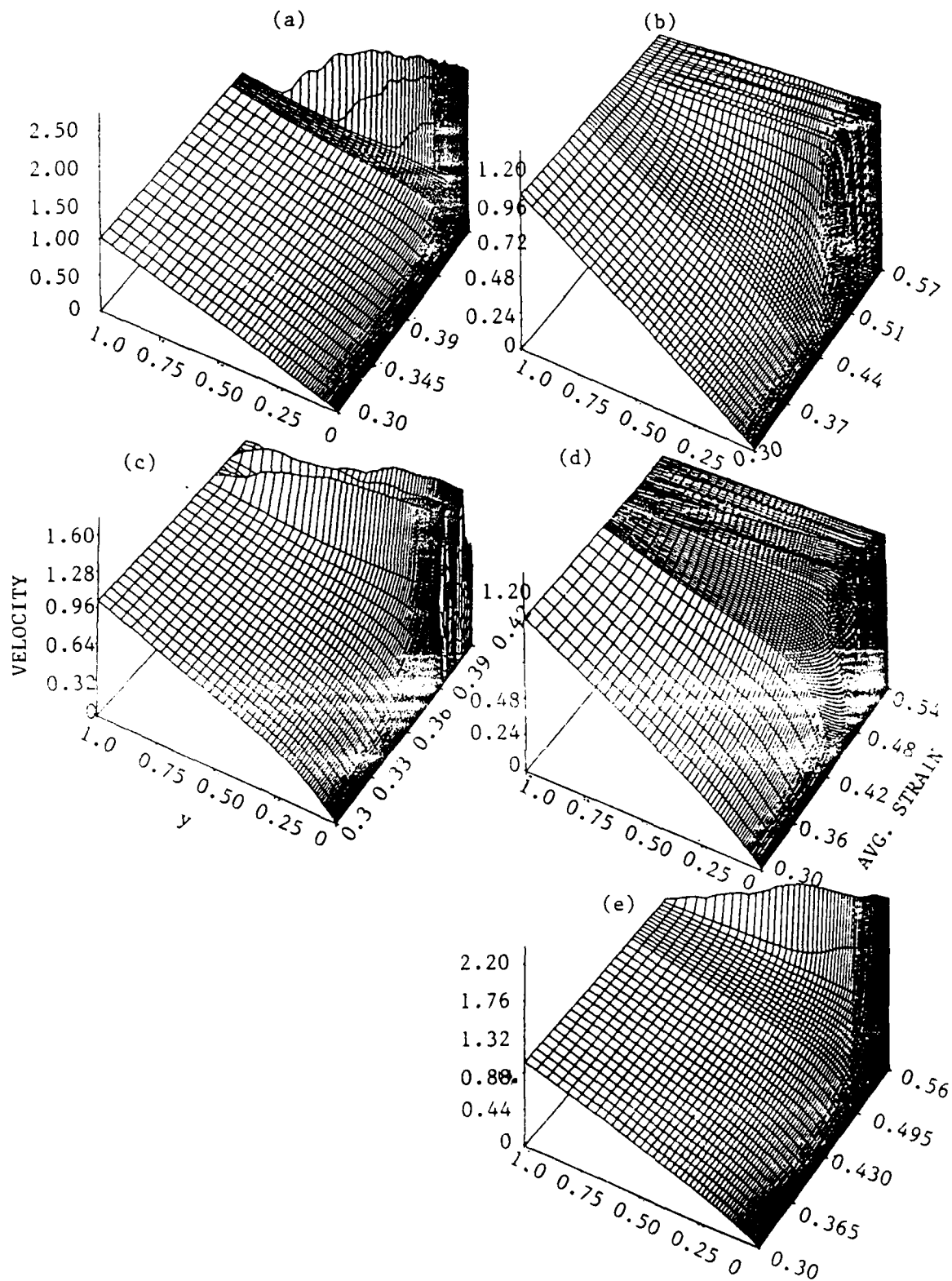


Fig. 3. The Evolution of the Velocity Field Within the Specimen After the Shear Stress has Attained Its Peak Value.

- (a) Litonski's Law, (b) Litonski's Flow Rule for Dipolar Materials,  
(c) Power Law (d) Bodner-Partom Law, (e) Johnson-Cook Law

strain. These flow rules when used to compute the initiation and growth of shear bands at a nominal strain-rate of  $5,000 \text{ sec}^{-1}$  gave noticeably different values of the nominal strain at which a shear band initiates. Also, the rate of drop of the shear stress as predicted by the Bodner-Partom law and the dipolar theory of Wright and Batra was closer to that observed experimentally. For nonpolar materials, the Litonski law predicts the emanation of an unloading elastic wave out of the shear band as it grows. The other three flow rules do give the overshoot in the particle velocity at the edges of the band as also given by the Litonski law, but do not predict the propagation of the unloading elastic wave. This could possibly be due to the use of a yield criterion for the Litonski law and not using any such criterion for the other flow rules.

Acknowledgements: We thank Mr. Ko for his help in plotting Figs. 2 and 3.

#### REFERENCES

- Anand, L., Lush, A. M., and Kim, K. H., 1988, "Thermal Aspects of Shear Localization in Viscoplastic Solids", Thermal Aspects in Manufacturing, M. H. Attia and L. Kops, eds., ASME-PED-Vol. 30, pp. 89-103.
- Batra, R. C., 1987, "The Initiation and Growth of and the Interaction Among, Adiabatic Shear Bands in Simple and Dipolar Materials", Int. J. Plasticity, Vol. 3, pp. 75-89.
- Batra, R. C. and Kim, C.H., 1989a, "Adiabatic Shear Banding in Elastic-Viscoplastic Nonpolar and Dipolar Materials", Int. J. Plasticity, (in press).
- Batra, R. C. and Kim, C. H., 1989b, "Effect of Viscoplastic Flow Rules on the Initiation and Growth of Shear Bands at High Strain Rates", (pending publication).
- Batra, R. C. and Liu, De-Shin, 1989a, "Adiabatic Shear Banding in Plane Strain Problems", J. Appl. Mechs., Vol. 56, (in press).
- Batra, R. C. and Liu, De-Shin, 1989b, "Adiabatic Shear Banding in Dynamic Plane Strain Compression of a Viscoplastic Material", Int. J. Plasticity (in press).
- Batra, R. C., 1989, "Effect of Nominal Strain Rates on Adiabatic Shear Banding in Dipolar Materials", Proc. 1st Pan American Congress of Applied Mechs., Rio de Janeiro, Brazil, pp. 79-82.
- Bodner, S. R. and Partom, Y., 1975, "Constitutive Equations for Elastic-Viscoplastic Strain-Hardening Materials", J. Appl. Mechs., Vol. 42, pp. 385-389.
- Giovanola, J., 1987, Proc. Impact Loading and Dynamic Behavior of Materials, Bremen, W. Germany.
- Hartley, K. A., Duffy, J. and Hawley, R. H., 1987, "Measurement of the Temperature Profile During Shear Band Formation in Steels Deforming at High Strain Rates", J. Mechs. Phys. Solids, Vol. 35, pp. 283-.



- Johnson, G. R. and Cook, W. H., 1983, "A Constitutive Model and Data for Metals Subjected to Large Strains, High Strain Rates and High Temperatures", Proc. 7th Int. Symp. Ballistics, The Hague, The Netherlands, pp. 1-7.
- Kwon, Y. W. and Batra, R. C., 1988, "Effect of Multiple Initial Imperfections on the Initiation and Growth of Adiabatic Shear Bands in Nonpolar and Dipolar Materials", Int. J. Engng. Sci., 26, pp. 1177-1187.
- LeMonds, J. and Needleman, A., 1986a, "Finite Element Analysis of Shear Localization in Rate and Temperature Dependent Solids", Mechs. Materials, Vol. 5, pp. 339-.
- LeMonds, J. and Needleman, A., 1986b, "An Analysis of Shear Band Development Incorporating Heat Conduction", Mechs. Materials, Vol. 5, pp. 363-.
- Litonski J., 1977, "Plastic Flow of a Tube Under Adiabatic Torsion", Bulletin de l'Academie Polonaise des Sciences, Sciences Tech., Vol. 25, pp. 7-14.
- Marchand, A. and Duffy, J., 1988, "An Experimental Study of the Formation of Adiabatic Shear Bands in a Structural Steel", J. Mech. Phys. Solids, Vol. 36, pp. 251-.
- Molinari, A. and Clifton, R. J., 1987, "Analytical Characterization of Shear Localization in Thermoviscoplastic Materials", J. Appl. Mech., Vol. 54, pp. 806-812.
- Needleman, A., 1989, "Dynamic Shear Band Development in Plane Strain", J. Appl. Mechs., Vol. 56, pp. 1-8.
- Tzavaras, A. E., 1987, "Effect of Thermal Softening in Shearing of Strain-Rate Dependent Materials", Arch. Rat. Mech. Anal., Vol. 99, pp. 349-374.
- Wright, T. W., and Batra, R. C., 1987, "Adiabatic Shear Bands in Simple and Dipolar Plastic Materials", Proc. Macro- and Micro-Mechanics of High Velocity Deformation and Fracture, IUTAM Symp., Tokyo, Aug. 1985, pp. 189-201.
- Wright, T. W., 1989, "Approximate Analysis for the Formation of Adiabatic Shear Bands", J. Mechs. Phys. Solids (in press).

# A PROPERTY OF LINEAR FEEDBACK SHIFT REGISTER SEQUENCES

Harold Fredricksen  
Mathematics Department  
Naval Postgraduate School  
Monterey, CA 93940

and

Gary Krahn  
Mathematics Department  
United States Military Academy  
West Point, NY 10996

## I. INTRODUCTION

Modern high-speed communication demands high-speed techniques to generate random-like sequences. One of the simplest and most efficient devices for generating deterministic, random looking binary sequences is the shift register. In a sense, no sequence which depends on a few parameters, such as the feedback connections of a linear feedback shift register, can be considered truly random. Solomon W. Golomb introduced the name "pseudo-random" for periodic binary maximum length linear sequences (m-sequences) because they satisfied the three randomness properties of balance, runs, and correlation. These "pseudo-random" sequences have a tremendous amount of combinatorial and algebraic structure. Their suitability derives chiefly from their efficiency of generation and, more importantly, their randomness properties. The applications for linear feedback shift register sequences of maximum length (LFSRS) include secure data transmission, multiple address coding, error correcting codes, radar range measuring and random number generation. Their ideal correlation property and other randomness properties can be derived from the shift-and-add (SAA) property. The ultimate goal of completely understanding the theoretical behavior of LFSR has not yet been achieved. A better understanding of LFSRS would provide improvements for the design and analysis of communication systems.

In this note we analyze the SAA property of the LFSRS. An algebraic approach is used in the analysis of the "unknown shift" of the sum of two shifted versions of the LFSRS.

## II. SHIFT - AND - ADD PROPERTY

A LFSRS of span  $n$  is a sequence of length  $N = 2^n - 1$  containing all non-zero binary strings of length  $n$ . The LFSRS is specifically generated by an  $n$  stage shift register whose feedback polynomial is a primitive polynomial  $f(x)$  of degree  $n$  over the Galois field of two elements,  $GF(2)$ . The computations in this note are performed modulo  $2^n - 1$  where  $n$  is the degree of  $f(x)$ , and modulo 2 and modulo the generating polynomial  $f(x)$ .

The SAA property states that if two shifted versions of the same LFSRS are added termwise modulo 2, the resulting sequence is also a shifted version of the same sequence.

**DEFINITION:** Let  $n$ ,  $i$  and  $j$  be positive integers with both  $i$  and  $j \leq 2^n - 2$ . The ordered pair  $(i, j)$  is a shift-and-add pair if and only if the modulo 2 sum sequence:

$$\{s^k\} + \{s^{k+i}\} + \{s^{k+j}\} = \{0\}, \quad \text{the zero sequence.}$$

Here,  $\{s^{k+t}\}$  is the cyclic shift by  $t$  of the original sequence  $\{s^k\}$ .

The existence of SAA pairs for  $m$ -sequences follows from the shift-and-add property of  $m$ -sequences. In fact, for every given shift  $i$  of the sequence  $S$ , the shift  $j$  of the sequence  $S$  is a unique function of  $i$ . That is  $j = \gamma(i)$ . The function  $\gamma(i)$  to compute  $j$  from  $i$  exists but is difficult to determine. Four properties of  $\gamma$  are as follows:

1.  $\gamma(i)$  is a unique function of  $i$ ,

*Proof:* This follows directly from the definition of SAA pairs. Otherwise,  $\{s^k\}$  would not be of maximal period and thus would be generated by a polynomial of a smaller degree than  $n$ .

2.  $\gamma(\gamma(i)) = i$ ,

*Proof:*  $x^i = 1 + x^{\gamma(i)} = x^{\gamma(\gamma(i))}$  follows from the definition of  $\gamma$ , i.e.,  $\gamma$  is idempotent.

3.  $\gamma(-i) = \gamma(i) - i$ .

*Proof:*  $1 + x^{-i} = (1 + x^i) x^{-i} = x^{\gamma(i)} x^{-i} = x^{\gamma(i) - i}$ .

NOTE: For small values of  $n$ , the SAA pairs  $(i, \gamma(i))$  for the respective polynomial  $f(x)$  are easily determined by analysis of the multiplicative group of the corresponding finite field. Alternate procedures to determine the SAA pairs, relative to a respective generating polynomial  $f(x)$ , also include long division of  $(1 + x^i)$  by  $f(x)$  to determine the remainder  $x^{\gamma(i)}$ .

Thus we see that, either by our original definition of SAA pairs or by the procedure just described, the sequence  $S$  generated by a minimal polynomial generator  $f(x)$  satisfies the relation

$$S + S^i + S^{\gamma(i)} = \{0\} \quad \text{the zero sequence or,}$$

$$f(x) \mid (1 + x^i + x^{\gamma(i)}) .$$

The symbol " $f \mid g$ " denotes that  $g$  is a polynomial multiple of  $f$  with coefficients in  $GF(2)$ .

4. If  $\gamma(i) = j$ , then  $\gamma(2i) = 2j$ . Thus,  $(2i, 2j)$  is also a SAA pair for  $f(x)$ .

*Proof:* Suppose  $f(x) \mid (1 + x^i + x^{\gamma(i)})$ , then

$$(1 + x^i + x^{\gamma(i)})^2 = 1 + x^{2i} + x^{2\gamma(i)} \quad \text{in } GF(2).$$

Hence, if  $i$  and  $j = \gamma(i)$  belong to a pair of cyclotomic cosets of size  $n$ , then  $(n-1)$  additional SAA pairs can be computed easily through squaring the polynomial relation.

Multiplication of the relation  $(1 + x^i + x^j)$  by  $x^{-i}$  and by  $x^{-j}$  will generate the SAA pairs  $(-i, j-i)$  and  $(-j, i-j)$ , respectively. Therefore, from one SAA pair it is possible to generate a total of  $(3n)$  pairs by only  $O(n)$  work.

### III. GENERATING ADDITIONAL SAA PAIRS:

In the previous section we show how to find  $3n$  SAA pairs when one SAA pair is known. Here we show how to find additional SAA pairs.

**EXAMPLE:** Let  $f(x) = 1 + x + x^7$ . Thus, (1,7) is a SAA pair and we determine that:

SET R - Under squaring

(1,7)      (2,14)      (4,28)      (8,56)      (16,112)      (32,97)      (64,67)

are SAA pairs by squaring. Multiplying  $f(x)$  by  $x^{-i}$  we get:

SET S - Under multiplication by  $x^{-i}$  and squaring

(6, 126)      (12,125)      (24,123)      (48,119)      (96,111)      (65,95)      (3,63).

Multiplying  $f(x)$  by  $x^{-j}$  we get:

SET T - Under multiplication by  $x^{-j}$  and squaring

(120,121)      (113,115)      (99,103)      (71,79)      (15,31)      (30,62)      (60,124).

The operations above are sufficient to determine only 21 of the 63 SAA pairs for the generator  $f(x) = 1 + x + x^7$ . The SAA mate for  $i = 5$  has not yet been determined. By the following algebraic manipulations we can determine another SAA pair:

$$\begin{array}{rcl}
 1 + x^5 & = & x^j \\
 (1 + x^5)x & = & x^{j+1} \\
 x + x^6 & = & x^{j+1} \\
 x+(1+x^{126}) & = & x^{j+1} \\
 (x+1)+x^{126} & = & x^{j+1} \\
 (x^7)+x^{126} & = & x^{j+1} \\
 x^7(1+x^{119}) & = & x^{j+1} \\
 x^7(x^{48}) & = & x^{j+1} \\
 x^{55} & = & x^{j+1} \\
 x^{54} & = & x^j
 \end{array}$$

Hence, (5,54) is a SAA pair. With this SAA pair we determine an additional 21 pairs by following the previous methods.

<u>SET R</u>	<u>SET S</u>	<u>SET T</u>
(5,54)	(49,122)	(73,78)
(10,108)	(98,117)	(19,29)
(20,89)	(69,107)	(38,58)
(40,51)	(11,87)	(76,116)
(80,102)	(22,47)	(25,105)
(33,77)	(44,94)	(50,83)
(66,27)	(88,61)	(100,39)

Finally, we find the last "family" of SAA pairs. The SAA mate for  $i = 9$  can be found in the same manner as  $i = 5$  as follows:

$$\begin{aligned}
 1 + x^9 &= x^j \\
 (1 + x^9)x^7 &= x^{7+j} \\
 x^7 + x^{16} &= x^{7+j} \\
 x^7 + (1 + x^{112}) &= x^{7+j} \\
 (x^7 + 1) + x^{112} &= x^{7+j} \\
 (x) + x^{112} &= x^{7+j} \\
 x(1 + x^{111}) &= x^{7+j} \\
 x(x^{96}) &= x^{7+j} \\
 x^{97} &= x^{7+j} \\
 x^{90} &= x^j
 \end{aligned}$$

<u>SET R</u>	<u>SET S</u>	<u>SET T</u>
(9,90)	(81,118)	(37,46)
(18,53)	(35,109)	(74,92)
(36,106)	(70,91)	(21,57)
(72,85)	(13,55)	(42,114)
(17,43)	(26,110)	(84,101)
(34,86)	(52,93)	(41,75)
(68,45)	(104,59)	(82,23)

Knowing only the initial SAA pair (1,7), all 63 SAA pairs are derived. The above procedure has thus been used to compute all the SAA pairs for  $f(x)$  in a relatively efficient manner.

#### IV. STRUCTURAL PROPERTIES OF SAA PAIRS:

Here we generalize the previous example and state and prove propositions that can be used to efficiently solve for the unknown families of SAA pairs.

**Proposition I:** Let  $j$  and  $\gamma(j)$  be a SAA pair. Then

$$\gamma(\gamma(j)+j) = \gamma(\gamma(j) - 2j) + 2j.$$

$$\begin{aligned} \text{Proof: } 1 + x^{\gamma(j)+j} &= 1 + x^{\gamma(j)} x^j, \\ &= 1 + (1 + x^j) x^j, \\ &= 1 + x^j + x^{2j}, \\ &= x^{\gamma(j)} + x^{2j}, \\ &= (x^{\gamma(j)-2j} + 1) x^{2j}, \\ &= x^{\gamma(\gamma(j)-2j)} x^{2j}, \\ &= x^{\gamma(\gamma(j)-2j) + 2j} \end{aligned}$$

from which the result follows easily.

**EXAMPLE:** For  $n = 7$  assume that the SAA pair  $(1,7)$  and the family of 21 SAA pairs associated with it are known. Let  $j=1$  and  $\gamma(j) = 7$ . Substitution into Eq (1) yields  $\gamma(7 + 1) = \gamma(7 - 2) + 2$ . Then,  $\gamma(8) = \gamma(5) + 2$  and  $\gamma(5) = 56 - 2 = 54$ .

It is only necessary to check one SAA pair from a given family to determine if a previously unknown family can be generated using proposition I. This is shown as follows:

Let  $(i_1, j_1)$  be a SAA pair and let  $I = \{(i_k, j_k) \mid k=1\}^n$  be the set of SAA pairs formed under squaring. The set  $L = \{(i_k + j_k) \mid k=1\}^n$  forms a cyclotomic coset. Similarly,  $M = \{(i_k - 2j_k) \mid k=1\}^n$  and  $Q = \{(j_k - 2i_k) \mid k=1\}^n$  form cyclotomic cosets. As previously shown, if  $(i_1, j_1)$  is a SAA pair then  $(-i_1, j_1 - i_1)$  and  $(-j_1, i_1 - j_1)$  are SAA pairs. It follows that

$$\{(-i_k + (j_k - i_k) |_{k=1}^n\} = Q, \{(j_k - i_k) - 2(-i_k) |_{k=1}^n\} = L, \{(-i_k - 2(j_k - i_k) |_{k=1}^n\} = M,$$

$$\{-j_k + (i_k - j_k) |_{k=1}^n\} = M, \{(i_k - j_k) - 2(-j_k) |_{k=1}^n\} = L, \text{ and } \{(-j_k - 2(i_k - j_k) |_{k=1}^n\} = Q.$$

**Proposition II:** Let  $(i, \gamma(i))$  and  $(j, \gamma(j))$  be SAA pairs. Then

$$\gamma(\gamma(j) + i) = \gamma(j + i - \gamma(i)) + \gamma(i).$$

*Proof:*

$$\begin{aligned} 1 + x^{\gamma(j) + i} &= 1 + (x^j + 1) x^i, \\ &= 1 + x^{j+i} + x^i, \\ &= x^{\gamma(i)} + x^{j+i}, \\ &= (x^{j+i-\gamma(i)} + 1) x^{\gamma(i)}, \\ &= x^{\gamma(j+i - \gamma(i))} x^{\gamma(i)} \end{aligned}$$

from which the results follows easily.

**EXAMPLE:** For  $n = 7$  let the SAA pair  $(1,7)$  and the family of 21 SAA pairs associated with it be given. Hence,  $(1,7)$  and  $(4, 28)$  are known SAA pairs. Let  $i=1$  and  $j = 4$ . Substitution into Eq (2) yields  $\gamma(28+1) = \gamma(5 - 7) + 7$ . Then,  $\gamma(29) = \gamma(125) + 7$  and  $\gamma(29) = 12 + 7 = 19$ .

A family and the associated SAA pairs are defined to be "generators" if the SAA pairs within this family generate at least one SAA pair from another family using proposition II.

**NOTE:** The application of Proposition II requires the input of two ordered SAA pairs where the order of the input is important and repetition of a SAA pair is allowed. If  $(i, j)$  is a SAA pair, then the four inputs of  $(i, j), (j, i); (i, j), (i, j); (j, i), (j, i);$  and  $(j, i), (i, j)$  into proposition II provide unique relationships. Therefore, within a given family of SAA pairs their are  $36n^2$  possible combinations to substitute into proposition II. However, it is only necessary to consider the input of  $18n$  combinations to determine if a previously unknown family can be generated using proposition II. This is shown as follows:



Let  $J = \{(i_p, \gamma(i)_p) \mid p=1\}$  and  $I = \{(j_k, \gamma(j)_k) \mid k=1\}$  be sets of SAA pairs formed under squaring. The sets  $\{(i_p + j_k), (\gamma(i)_p + \gamma(j)_k) \mid p, k=1\}$  and  $\{(i_p + j_1), (\gamma(i)_p + \gamma(j)_1)\}$  are equal. Hence, only  $36n$  of the possible  $36n^2$  combinations need to be considered when applying proposition II. Furthermore, the input of  $(i, \gamma(i))$  and  $(\gamma(i), i)$  generates the same relationships as the input  $(i - \gamma(i), -\gamma(i))$  and  $(i, \gamma(i))$ . It follows that there remains only  $18n$  combinations that are unique.

The effect of  $18n$  combinations as input to proposition II is to produce relations that develop new SAA pairs. We shall see, that on some occasions no new SAA pairs are derived from the original SAA pair by applying proposition II.

To return to the previous example, the initial condition for the application of proposition II was that exactly one SAA pair is known along with its  $3n$  relatives from procedures R, S and T. From the SAA pair  $(1,7)$  and its relatives, we saw it is possible by proposition II to find all 63 SAA pairs for the polynomial  $f(x) = x^7 + x + 1$ . In fact, it is the case that no matter which irreducible polynomial is chosen for degree  $n = 7$ , proposition II is sufficient to find all of the 63 SAA pairs. However, this is not the case for every value of  $n$ . For larger values of the degree  $n$  we sometimes are not able to find all SAA pairs by application of proposition II to a given SAA pair and its relatives. In the sequel we discuss the probability with which it is possible to complete the table of SAA pairs from a given SAA pair.

**Proposition III:** Let  $(i, \gamma(i))$ ,  $(j, \gamma(j))$  and  $(k, \gamma(k))$  be SAA pairs. Then

$$\gamma(\gamma(i) + \gamma(j) + \gamma(k)) = \gamma\{i + \gamma(j) + \gamma(k) - \gamma(\gamma(j) + \gamma(k))\} + \gamma(\gamma(j) + \gamma(k))$$

*Proof:*

$$\begin{aligned} 1 + x^{\gamma(i) + \gamma(j) + \gamma(k)} &= 1 + x^{\gamma(i)} x^{\gamma(j)} x^{\gamma(k)}, \\ &= 1 + (1 + x^i) x^{\gamma(j)} x^{\gamma(k)}, \\ &= 1 + x^{\gamma(j) + \gamma(k)} + x^{i + \gamma(j) + \gamma(k)}, \\ &= x^{\gamma(\gamma(j) + \gamma(k))} + x^{i + \gamma(j) + \gamma(k)}, \\ &= (1 + x^{i + \gamma(j) + \gamma(k) - \gamma(\gamma(j) + \gamma(k))}) x^{\gamma(\gamma(j) + \gamma(k))}. \\ &= x^{\gamma\{i + \gamma(j) + \gamma(k) - \gamma(\gamma(j) + \gamma(k))\}} x^{\gamma(\gamma(j) + \gamma(k))} \end{aligned}$$

from which the result follow directly.

**EXAMPLE:** For  $n = 7$  allow the SAA pair  $(1,7)$  and the family of 21 SAA pairs associated with it to be known. Hence,  $(99,103)$ ,  $(65,95)$  and  $(15,31)$  are known SAA pairs. Let  $j=99$ ,  $i = 65$  and  $k = 15$ . Substitution into Equation (3) yields  $\gamma(103 + 95 + 31) = \gamma(65 + 103 + 31 - \gamma(103+31)) + \gamma(103 + 31)$ . Then,  $\gamma(102) = \gamma(71) + 1$  and  $\gamma(102) = 79 + 1 = 80$ .

If we let  $\gamma(j) + \gamma(k) = K$  where  $K$  is an element in a family, then substituting  $K$  into proposition III transforms it into proposition II. Hence, proposition III and proposition II are equivalent. As additional terms are included in the right hand side of proposition II, equivalent relationships are created which do not provide further information.

## V. RESULTS:

Proposition II has been used to generate SAA pairs for every PN sequences up to  $n=14$ . Not every SAA pair from a PN sequence possesses the necessary structure to generate all the remaining SAA pairs. At this time, it cannot be determined a priori which initial SAA pairs will be complete generators of all SAA pairs. However, for a SAA pair to be a generator the following conditions must be satisfied:

Let  $(i, j)$  be a SAA pair that forms the family  $\{(l_k, h_k) \mid k=1\}$ , where  $q$  is the size of the family and  $(l_k, h_k)$  is a SAA pair. Generate the following set of 2-tuples:

$$\left\{ \begin{array}{l} \left( (l_k+i), (h_k+(i-j)) \right), \left( (l_k+j), (h_k+(j-i)) \right), \left( (l_k+(-i)), (h_k+(-j)) \right), \\ \left( (h_k+i), (l_k+(i-j)) \right), \left( (h_k+j), (l_k+(j-i)) \right), \left( (h_k+(-i)), (l_k+(-j)) \right) \end{array} \right\}.$$

If one and only one element from any 2-tuple is a member of the family then every SAA pair in this family is a generator. If for every 2-tuple, one and only one element is not a member of the family, then every SAA pair in the family is not a generator. This follows from proposition II. Upon closer examination, these 2-tuples are created by adding  $i, j, (i-j), (j-i), -i,$  and  $-j$  to the each member of  $\{(l_k, h_k) \mid k=1\}$ . These transformations map each element of the pair from one cyclotomic coset to another cyclotomic coset. It is this mapping that determines which SAA pairs will be generators. The function to compute this mapping exists. However, it is difficult to determine.

If a SAA pair is a generator, proposition II provides a tremendous reduction in the calculations needed to find the set of SAA pairs for a given LFSR. An analysis and determination of which SAA pairs are generators has been carried out for values of  $n$  up to 14. The number of cyclotomic cosets, SAA pairs, and families (including their sizes) have been tabulated below:

n	Number of Cosets	Number of SAA pairs	Number of Families	Number of Families by size					
7	18	63	3	3-21					
8	34	127	8	4-24	2-12	1-6	1-1		
9	58	255	11	9-27	1-9	1-3			
10	106	511	20	15-30	4-15	1-1			
11	186	1023	31	31-33					
12	350	2047	63	53-36	1-9	6-18	1-3	1-1	1-12
13	630	4095	105	105-39					
14	1180	8191	202	189-42	12-21	1-1			

The number and percentage of SAA pairs that generate SAA pairs of families other than their own are:

n=7: 63 - 100%

n=11: 462 - 45%

n=8: 72 - 61%

n=12: 720 - 35%

n=9: 162 - 64%

n=13: 819 - 20%

n=10: 285 - 56%

n=14: 1071 - 13%

These numbers are identical for every LFSRS of the same degree. However, as  $n \rightarrow \infty$  the probability that a randomly selected SAA pair generates all families of SAA pairs should approach zero.

We note that more can be gleaned from proposition II if two independent SAA pairs are given. If two or more SAA pairs from different families are known, then the new family becomes the union of the families from each of the separate SAA pairs. Thus, when non-generator families are combined, the structure can be sufficient to generate additional SAA pairs when each SAA pair alone is not a generator.

**EXAMPLE:** Let  $f(x) = x^9 + x^4 + 1$ . Assume that the SAA pairs (30, 66) and (105, 248) have been selected at random. Using proposition II, no additional SAA pairs can be generated by using each SAA pair alone. However, when the two individual families are combined into one larger family, the entire set of SAA pairs for  $f(x)$  can be generated using proposition II.

For  $f(x)$  of degree  $n$ , assume that two SAA pairs,  $(i, j)$  and  $(l, k)$  of different families have been randomly selected. Then, the probabilities that these SAA pairs will generate the remaining SAA pairs for the associated LFSRS are:

n=7: 100%	n=11: 100%
n=8: 99%	n=12: 100%
n=9: 99%	n=13: 93%
n=10: 99%	n=14: 70%

The probabilities have not been calculated when three or more SAA pairs of different families are used to generate the remaining SAA pairs for the associated LFSRS.

Additional research is needed to discover the further structure of SAA pairs. Ultimately we would like to determine the entire nature of the function  $\gamma: i \rightarrow j$ .

#### UNANSWERED QUESTION

Can it be determined a priori which cyclotomic coset each member of the set  $\{2^p i + k \mid p=1\}$  will belong to for a given  $i$  or  $k$ ?

# Digital Redesign of Pseudo-continuous-time Suboptimal Regulators for Large-scale Discrete Systems

Jason S.H. Tsai § , Leang S. Shieh, Jian L. Zhang † and Norman P. Coleman ‡

Indexing terms: Digital control, Large-scale system, Optimal regulators,  
Pole placement, State-space methods.

## Abstract

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 explicitly. The designed pseudo-continuous-time regulator places the eigenvalues of the closed-loop discrete system near 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 solutions of small order Riccati equations only at each stage of the design. Based on matching all the states at all the sampling instants, a new digital redesign technique is presented for finding the pseudo-continuous-time quadratic regulator. An illustrative example is presented to demonstrate the effectiveness of the proposed procedures.

---

§ Department of Electrical Engineering, National Cheng-Kung University, Tainan, Taiwan, R.O.C.

† Department of Electrical Engineering, University of Houston, Houston, TX 77204-4793, U.S.A.

‡ U.S. Army Armament Center, Dover, New Jersey 07801, U.S.A.

## 1. Introduction

Physical realizations of engineering systems result, in general, 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 [3], multi-time scales [9] and modal analysis [15]. 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 [11,12,14]. Mahmoud *et al.* [11] 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 always be feasible to satisfy this condition. Shieh *et al.* [18] have developed an algebraic method based on the matrix sign function [16] for separating the slow (dominant) modes from the fast (non-dominant) 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 [17] of a large-scale system, i.e., decomposing the system into parallel and cascaded structures; for solving non-linear Riccati equations, which often appear in feedback design of systems based on linear quadratic theory; and for model conversions of systems via the computation of the principal  $q$ th root of the system matrix [21,24]. Recently, fast and stable algorithms have been developed for the computation of the matrix sign function [21] and for the computation of the principal  $q$ th root of a complex matrix [24] 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 suboptimal discrete controllers with pole assignment near 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 [2]. 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 non-uniqueness of choice of feedback for multivariable systems. Further, it is too restrictive to place the poles in pre-determined locations [1], since for non-linear systems the exact locations 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. [1]. In this paper, we consider the common region of a circle and a logarithmic spiral in the  $z$ -plane (Fig. 2) for pole assignment. This is equivalent to the sector region (hatched) in Fig. 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 [2], 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.* [19,22] 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 [8] proposed 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. 1. In [23], a pseudo-continuous-time method has been developed to place the eigenvalues of a discrete system (having a sufficiently small sampling period) within the hatched region of Fig. 2. However, it involves the solution of full order Riccati equations, which could be computationally difficult for large-scale systems. The Luenberger 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 paper, at each stage of the design, only reduced order Riccati equations need to be solved and also, the transformation to the

general canonical forms is avoided.

For digital implementation of the designed continuous-time controller, the continuous-time controller (analog controller) needs to be converted into an equivalent discrete-time controller (digital controller). This is a digital redesign problem [10]. Based on a bilinear transform method, a new digital redesign technique is presented for finding the equivalent digital controller from the designed analog controller.

The material in this paper is organized as follows: Section 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. 1. In Section 3, a new digital redesign technique is presented for converting the continuous-time control law to an equivalent discrete-time control law. In Section 4, a method, using the matrix sign function, for block-decomposing the equivalent large scale continuous-time system into a multi-time scale structure is introduced. Then, a pseudo-continuous-time multi-stage design procedure is presented for designing large-scale discrete systems with pole placement near the hatched region of Fig. 2. An illustrative example is given in Section 5 to demonstrate the effectiveness of the proposed design procedure and the conclusions are summarized in Section 6. Some computational algorithms are given in an appendix.

## 2. Continuous-time optimal quadratic regulators with pole placement

Consider the linear controllable continuous-time system described by

$$\dot{x}_c(t) = Ax_c(t) + Bu_c(t); \quad x_c(0) \quad (1)$$

where  $x_c(t)$  and  $u_c(t)$  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 (1) be

$$J = \int_0^{\infty} (x_c^T(t)Qx_c(t) + u_c^T(t)Ru_c(t)) dt \quad (2)$$

where the weighting matrices  $Q$  and  $R$  are  $n \times n$  non-negative definite and  $m \times m$  positive definite symmetric matrices, respectively. The feedback control law that minimizes the performance index in (2) is given by [2]

$$u_c(t) = -K_c x_c(t) + E_c r(t) \equiv -R^{-1}B^T P x_c(t) + E_c r(t) \quad (3)$$



where  $K_c$  is an  $m \times n$  feedback gain,  $E_c$  is an  $m \times m$  forward gain,  $r(t)$  is a reference input, and  $P$ , an  $n \times n$  non-negative definite symmetric matrix, is the solution of the Riccati equation,

$$PBR^{-1}B^T P - PA - A^T P - Q = 0_n \quad (4)$$

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

$$\dot{x}_c(t) = (A - BK_c)x_c(t) + BE_c r(t) \quad (5)$$

The eigenvalues of  $A - BK_c$ , denoted by  $\sigma(A - BK_c)$ , lie in the open left-half plane of the complex  $s$ -plane. Our objective is to determine  $Q$ ,  $R$  and  $K_c$  so that the closed-loop system in (5) has its eigenvalues on or within the hatched region of Fig. 1. The important results along with the design procedure to achieve the desired design are presented in the following.

**Lemma 1** [2,23]: Let  $(A, B)$  be the pair of the given open-loop system in (1). Also, let  $h \geq 0$  represent the prescribed degree of relative stability. Then, the eigenvalues of the closed-loop system  $A - BR^{-1}B^T P$  lie to the left of the  $-h$  vertical line with the matrix  $P$  being the solution of the Riccati equation,

$$PBR^{-1}B^T P - P(A + hI_n) - (A + hI_n)^T P = 0_n \quad (6)$$

where the matrix  $I_n$  is an  $n \times n$  identity matrix. ■

**Theorem 1:** [23] Let the given stable system matrix  $A \in \mathcal{R}^{n \times n}$  have eigenvalues  $\hat{\lambda}_i^-$  ( $i = 1, \dots, n^-$ ) lying in the open sector of Fig. 1 and the eigenvalues  $\hat{\lambda}_i^+$  ( $i = 1, \dots, n^+$ ) outside that sector, with  $n = n^- + n^+$ . Now, consider the two Riccati equations,

$$\hat{Q}BR^{-1}B^T \hat{Q} - \hat{Q}(-A^2) - (-A^2)^T \hat{Q} = 0_n \quad (7a)$$

and

$$PBR^{-1}B^T P - PA - A^T P - \hat{Q} = 0_n \quad (7b)$$

Then, the closed-loop system,

$$A_c = A - \gamma BK_c = A - \gamma BR^{-1}B^T P, \quad (8)$$

will enclose the invariant eigenvalues  $\lambda_i^-$  ( $i = 1, \dots, n^-$ ) and at least one additional pair of complex conjugate eigenvalues lying in the open sector of Fig. 1, for the constant gain  $\gamma$  in (8) satisfying

$$\gamma \geq \max\left\{\frac{1}{2}, \frac{b + \sqrt{b^2 + ac}}{a}\right\} \quad (9)$$

where  $a = \text{tr}[(BR^{-1}B^T P)^2]$ ,  $b = \text{tr}[BR^{-1}B^T P A]$  and  $c = \frac{1}{2}\text{tr}[BR^{-1}B^T \hat{Q}]$ . ■

**Remark 1:** The matrix  $\hat{Q}$ , the solution of the Riccati equation in (7a), contains the eigenvectors associated with the eigenvalues  $\lambda_i^-$  ( $i = 1, \dots, n^-$ ) of  $A$  lying in the open sector of Fig. 1. This matrix is used as a state weighting matrix in the Riccati equation in (7b) for solving the matrix  $P$ . As a result, the asymptotically stable closed-loop system matrix  $A_c$  in (8) contains the invariant eigenvectors and associated eigenvalues  $\lambda_i^-$  ( $i = 1, \dots, n^-$ ) of  $A$ . The steady state solutions of the Riccati equations in (6) and (7) can be found using the matrix sign function techniques [4,17] and the eigenvalue-eigenvector approach [7]. A brief review of this is given in the Appendix.

### Continuous-time Design Procedure

**Step 1:** Let the given continuous-time system be as in (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. 1. Also, assign  $A_0 = A$  and the positive definite matrix  $R$ . Set  $i = 1$ . If the system is unstable, then solve (6) to obtain the closed-loop system  $A_1 = A - \gamma_0 BR^{-1}B^T P_0 = A - \gamma_0 BK_0$ , with  $\gamma_0 = 1$ ; else (stable system) go to Step 2 with  $A_1 = A$ ,  $P_0 = 0_n$  and  $\gamma_0 = 0$ .

**Step 2:** Solve (7a) for  $\hat{Q}_i$  with  $A := A_i$ . Check if  $\frac{1}{2}\text{tr}[BR^{-1}B^T \hat{Q}_i]$  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  $\frac{1}{2}\text{tr}[BR^{-1}B^T \hat{Q}_i] = 0$ , all eigenvalues of the matrix  $A_i$  lie on or within the open sector of Fig. 1.

**Step 3:** Solve (7b) for  $P_i$  with  $A := A_i$  and  $\hat{Q} := \hat{Q}_i$ . Then, the constant gain  $\gamma_i$  can be evaluated using (9). The closed-loop system matrix is

$$A_{i+1} = A_i - \gamma_i BR^{-1}B^T P_i = A_i - \gamma_i BK_i \quad (10a)$$

Set  $i := i + 1$  and go to Step 2.

**Step 4:** Check if  $\text{tr}[(A_j + hI_n)]^+$  (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  $\gamma_{j+1} = 0$ ; else, solve (6) for  $P_{j+1}$  with  $A := A_j$  and obtain the closed-loop system  $A_j - \gamma_{j+1}BR^{-1}B^T P_{j+1} = A_j - \gamma_{j+1}BK_{j+1}$ , with  $\gamma_{j+1} = 1$  and  $K_{j+1} = R^{-1}B^T P_{j+1}$ .

**Step 5:** The designed closed-loop system is

$$A_0 - BR^{-1}B^T \sum_{k=0}^{j+1} \gamma_k P_k \quad (10b)$$

and its eigenvalues lie in the hatched region of Fig. 1. Note that the above system matrix in (10b) is equal to the system matrix in (5),  $A - BR^{-1}B^T \hat{P}$ , where  $\hat{P}$  is the solution of the Riccati equation in (4) with

$$Q = 2h(P_0 + P_{j+1}) + \sum_{i=1}^j (\hat{Q}_i + \Delta\gamma_i P_i BR^{-1}B^T P_i) \gamma_i \quad (10c)$$

In the above equation,  $\Delta\gamma_i = \gamma_i - 1$  and the matrix  $R$  is as originally assigned. Also, the optimal continuous-time regulator can be given as

$$u_c(t) = - \left( \sum_{i=0}^{j+1} \gamma_i K_i \right) x_c(t) + E_c r(t) = -K_c x_c(t) + E_c r(t) \quad (10d)$$

where  $r(t)$  is any reference input,  $E_c$  is any forward gain, and  $K_c$  is the desired state feedback gain.

### 3. Model Conversions and Digital Redesign

For digital implementation of the obtained optimal continuous-time regulator in (10d), we need to convert the continuous-time system in (1) and continuous-time control law in (10d) into an equivalent discrete-time model and discrete-time control law, respectively.

#### 3.1 Model Conversions

Let the state equation of the digital system which approximates the continuous-time system in (1) be represented by

$$\dot{x}_d(t) = Ax_d(t) + Bu_d(t); \quad x_d(0) \quad (11a)$$

where  $u_d(t)$  is a piecewise input function,

$$u_d(t) = u_d(kT) \quad \text{for } kT \leq t < (k+1)T \quad (11b)$$

where  $T$  is the sampling period, then we can write the equivalent discrete-time model as

$$\mathcal{X}_d(kT + T) = G\mathcal{X}_d(kT) + H u_d(kT); \quad \mathcal{X}_d(0) \quad (12a)$$

where

$$G = \exp(AT) \quad \text{and} \quad H = [G - I_n]A^{-1}B \quad (12b)$$

In general, the matrices  $G$  and  $H$  can be determined exactly from the matrices  $A$  and  $B$ , and the input function  $u_d(t)$  in (11b) using the eigenvalue and eigenvector approach [13]. 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 [13] to evaluate approximately  $G$  and  $H$  given in (12), the simplest of them is the truncation of the infinite series of  $\exp(AT)$  [13] which results in good approximation when  $T$  is sufficiently small. A popular method for determining  $G$  and  $H$  approximately is the Pade approximation method [13,20]. Some of the approximations obtained using this method are listed below:

$$G \approx [I_n - \frac{1}{2}AT]^{-1}[I_n + \frac{1}{2}AT] \triangleq G_3 \quad (13a)$$

$$\approx [I_n - \frac{1}{2}AT + \frac{1}{12}(AT)^2]^{-1}[I_n + \frac{1}{2}AT + \frac{1}{12}(AT)^2] \triangleq G_5 \quad (13b)$$

and

$$H \approx T[I_n - \frac{1}{2}AT]^{-1}B \triangleq H_3 \quad (14a)$$

$$\approx T[I_n - \frac{1}{2}AT + \frac{1}{12}(AT)^2]^{-1}B \triangleq H_5 \quad (14b)$$

It can be noted that the matrices  $G_3$  in (13a) and  $H_3$  in (14a) correspond to the popular Tustin approximation (bilinear transformation) [6]. The matrices  $G_5$  and  $H_5$ , when used with even large sampling periods, provide good approximations. The use of scaling and squaring method [13] as shown below, along with one of the above approximations, would result in better approximations:

$$G \approx [e^{AT/m}]^m, \quad m \text{ is a power of two} \quad (15)$$

Now, given a discrete-time model as in (12), an equivalent continuous-time model in (11) can be obtained by using

$$A = \frac{1}{T} \ln (G) \quad \text{and} \quad B = A[G - I_n]^{-1}H \quad (16)$$

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 logarithmic function,  $\ln (G)$ , subject to certain convergence conditions. Shieh *et al.* [20] 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

$$A = \frac{1}{T} \ln (G) \approx \frac{2}{T} R \approx \frac{2}{T} R [I_n - \frac{4}{15} R^2] [I_n - \frac{3}{5} R^2]^{-1} \quad (17)$$

where  $R = [G - I_n][G + I_n]^{-1}$ . The matrix series approximations obtained from truncation or continued fractions converge when  $\text{Re}(\sigma(G)) > 0$ , 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  $q$ th root of the matrix  $G$  [20,21,24] can be made use of. Shieh *et al.* [21] and Tsai *et al.* [24] have recently developed a fast and stable algorithm for computing the principal  $q$ th root of a general complex matrix. This is listed in Appendix, A-1. The eigenvalues of  $\sqrt[q]{G}$  lie in the right half of the complex  $z$ -plane, i.e.,  $\text{Re}(\sigma(\sqrt[q]{G})) > 0$ , for  $q \geq 2$ . Therefore, instead of  $G$  the principal  $q$ th root of  $G$  can be used in determining an approximation for  $A$ . In this case, the matrix equation (16) becomes

$$A = \frac{1}{T} \ln (G) = \frac{q}{T} \ln (\sqrt[q]{G}) \quad (18)$$

As a result, the matrix  $R$  in equation (17) 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.

### 3.2 Digital Redesign by Matching of States

Let the digital control law for the discrete-time open-loop system in (12) be

$$u_d(kT) = -K_d x_d(kT) + E_d r(kT) \quad (19)$$

The designed hybrid closed-loop system in (11) becomes

$$\dot{x}_d(t) = Ax_d(t) - BK_d x_d(kT) + BE_d r(kT); \quad x_d(0) \quad (20)$$

for  $kT \leq t < (k+1)T$ , where  $K_d \in R^{m \times n}$  and  $E_d \in R^{m \times m}$  are the digital state-feedback and forward gains, respectively. A zero-order hold is utilized in (20). Now, the digital redesign problem is reduced to finding the digital state-feedback gain  $K_d$  and forward gain  $E_d$  in (19) from the continuous state-feedback gain  $K_c$  and forward gain  $E_c$  in (3) so that the states of the digital system in (20) are approximately equal to the states of the continuous-time system in (5) at the sampling instants, for a given  $r(t)$ .

Assuming  $r(t) \approx r(kT)$  over one sampling period, we have the respective discrete models of (5) and (20) as follows:

$$x_c(kT + T) = \hat{G}x_c(kT) + \hat{H}E_c r(kT); \quad x_c(0) \quad (21)$$

and

$$x_d(kT + T) = (G - HK_d)x_d(kT) + HE_d r(kT); \quad x_d(0) \quad (22)$$

where  $\hat{G} = e^{(A-BK_c)T}$ ,  $\hat{H} = \int_0^T e^{(A-BK_c)\lambda} B d\lambda = [\hat{G} - I_n](A - BK_c)^{-1}B$ ,  $G = e^{AT}$  and  $H = \int_0^T e^{A\lambda} B d\lambda = [G - I_n]A^{-1}B$ . To match all  $n$  states of the digital system,  $x_d(kT)$  in (22), and those of the continuous-time system,  $x_c(kT)$  in (21), at each sampling instant, it is sufficient that the following equations are satisfied:

$$\hat{G} = G - HK_d \quad (23a)$$

and

$$\hat{H}E_c = HE_d \quad (23b)$$

where the  $m \times n$  (non-square) feedback gain  $K_d$  and the  $m \times m$  forward gain  $E_d$  are unknown matrices to be solved.

Alternative representation of (23) is

$$e^{(A-BK_c)T} = e^{AT} - [(e^{AT} - I_n)A^{-1}B]K_d \quad (24a)$$

and

$$[(e^{(A-BK_c)T} - I_n)(A - BK_c)^{-1}B]E_c = [(e^{AT} - I_n)A^{-1}B]E_d \quad (24b)$$

In the existing digital redesign technique [10], the  $K_d$  and  $E_d$  in (24) are considered as the functions of sampling period, i.e.,  $K_d(T)$  and  $E_d(T)$ , respectively, and they are expanded into a Taylor series about  $T = 0$  as

$$K_d(T) = \lim_{k \rightarrow \infty} \sum_{j=0}^{k-1} \frac{1}{j!} K_d^{(j)}(T) T^j \quad (25a)$$

where  $K_d^{(j)}(T) = \frac{\partial^j K_d(T)}{\partial T^j} |_{T=0}$ , and

$$E_d(T) = \lim_{k \rightarrow \infty} \sum_{j=0}^{k-1} \frac{1}{j!} E_d^{(j)}(T) T^j \quad (25b)$$

where  $E_d^{(j)}(T) = \frac{\partial^j E_d(T)}{\partial T^j} |_{T=0}$ .

In a similar manner, the exponential matrix-valued functions in (24) are expanded into a Taylor series about  $T = 0$ . As a result, (24), together with (25), can be written as

$$\sum_{j=0}^{\infty} \frac{[A - BK_c]^j T^j}{j!} = \sum_{j=0}^{\infty} \frac{A^j T^j}{j!} - \sum_{j=0}^{\infty} \frac{A^j T^{j+1}}{(j+1)!} B \sum_{k=0}^{\infty} \frac{K_d^{(k)}(T) T^k}{k!} \quad (26a)$$

and

$$\sum_{j=0}^{\infty} \frac{(A - BK_c)^j T^{j+1}}{(j+1)!} B E_c = \sum_{j=0}^{\infty} \frac{A^j T^{j+1}}{(j+1)!} B \sum_{k=0}^{\infty} \frac{E_d^{(k)}(T) T^k}{k!} \quad (26b)$$

To match all the states of the continuous and the digital systems with a sufficiently small sampling period, the approximated digital feedback gain (defined as  $\hat{K}_d$ ) and the approximated digital forward gain (defined as  $\hat{E}_d$ ) are determined [10] by taking the first two terms of all associated matrix series expansions in (26) as

$$\hat{K}_d = K_c + \frac{1}{2} K_c (A - BK_c) T \quad (27a)$$

and

$$\hat{E}_d = [I_m - \frac{1}{2}K_c BT]E_c \quad (27b)$$

In this paper, we propose a bilinear transform method to solve explicitly for the desired  $m \times n$  (non-square) feedback gain  $K_d$  and  $m \times m$  forward gain  $E_d$  in (24) by taking infinite terms of the modified Taylor series expansions of  $e^{(A-BK_c)T}$  and  $e^{AT}$  in (24) in the following.

The matrix-valued function of  $e^{XT}$  with  $X \in R^{n \times n}$  and a sampling period  $T$  can be represented by an infinite series [6] as

$$e^{XT} = I_n + XT + \frac{1}{2}(XT)^2 + \sum_{i=3}^{\infty} \frac{1}{i!}(XT)^i \quad (28)$$

The infinite series in (28) can be approximated by a geometric series as

$$e^{XT} \approx I_n + XT + \frac{1}{2}(XT)^2 + \sum_{i=3}^{\infty} \frac{1}{2^{i-1}}(XT)^i \quad (29a)$$

$$= I_n + XT[I_n + (\frac{1}{2}XT) + (\frac{1}{2}XT)^2 + (\frac{1}{2}XT)^3 + \dots] \quad (29b)$$

$$= I_n + XT[I_n - \frac{1}{2}XT]^{-1} \quad \text{for } \|I_n - \frac{1}{2}XT\| < 1 \quad (29c)$$

$$= [I_n - \frac{1}{2}XT]^{-1}[I_n + \frac{1}{2}XT] \quad \text{for } \|I_n - \frac{1}{2}XT\| < 1 \quad (29d)$$

Note that the first three dominant terms of (28) are equal to those of (29a), while other terms differ in weighting factors  $1/i!$  in (28) and  $1/2^{i-1}$  in (29a). Also that the sampling period  $T$  in (29) can be chosen to satisfy the sufficient condition,  $\|I_n - \frac{1}{2}XT\| < 1$ , in (29).

When the matrix  $X$  in (28) is a continuous-time system matrix,  $e^{XT}$  becomes the equivalent discrete-time system matrix. Then the model conversion in (29d) corresponds to the popular Tustin approximation (bilinear transformation). The selection of the suitable sampling period  $T$  for the Tustin approximation method and its applications to digital control system design in the frequency domain have been investigated in [5]. Also, complete analysis, design and implementation of pseudo-continuous-time controllers, developed via the frequency-domain bilinear transformation, for discrete-time systems can be found in [6].



With the use of the bilinear transformation as shown in (29d), we express the matrices  $\hat{G}$ ,  $\hat{H}$ ,  $G$  and  $H$  in (23) as follows:

$$\hat{G} = e^{(A-BK_c)T} \approx [I_n - \frac{1}{2}(A-BK_c)T]^{-1} [I_n + \frac{1}{2}(A-BK_c)T] \quad (30a)$$

$$\hat{H} = [e^{(A-BK_c)T} - I_n](A-BK_c)^{-1}B \approx [I_n - \frac{1}{2}(A-BK_c)T]^{-1}BT \quad (30b)$$

$$G = e^{AT} \approx [I_n - \frac{1}{2}AT]^{-1} [I_n + \frac{1}{2}AT] \quad (30c)$$

and

$$H = [e^{AT} - I_n]A^{-1}B \approx [I_n - \frac{1}{2}AT]^{-1}BT \quad (30d)$$

To solve explicitly for the desired  $m \times n$  (non-square) feedback gain  $K_d$  and  $m \times m$  forward gain  $E_d$  in (23), we make use of the following matrix inversion formula [7]:

$$(A + BC^{-1}D)^{-1} = A^{-1} - A^{-1}B(C + DA^{-1}B)^{-1}DA^{-1} \quad (31)$$

Thus, the matrix  $[I_n - \frac{1}{2}(A-BK_c)T]^{-1}$  in (30a) can be represented as

$$\begin{aligned} & [(I_n - \frac{1}{2}AT) + \frac{1}{2}BK_cT]^{-1} = [(I_n - \frac{1}{2}AT) + BT(2I_m)^{-1}K_c]^{-1} \\ & = (I_n - \frac{1}{2}AT)^{-1} - (I_n - \frac{1}{2}AT)^{-1}BT[2I_m + K_c(I_n - \frac{1}{2}AT)^{-1}BT]^{-1}K_c(I_n - \frac{1}{2}AT)^{-1} \end{aligned} \quad (32)$$

Substituting (30) and (32) into (23) and utilizing the inverse bilinear transformation in (30c) and (30d), we have

$$\begin{aligned} \hat{G} &= (I_n - \frac{1}{2}AT)^{-1}(I_n + \frac{1}{2}AT) - (I_n - \frac{1}{2}AT)^{-1}BT[2I_m + K_c(I_n - \frac{1}{2}AT)^{-1}BT]^{-1} \\ &\quad \times K_c(I_n - \frac{1}{2}AT)^{-1}(I_n + \frac{1}{2}AT) - \frac{1}{2}(I_n - \frac{1}{2}AT)^{-1}BTK_c \\ &\quad + \frac{1}{2}(I_n - \frac{1}{2}AT)^{-1}BT[2I_m + K_c(I_n - \frac{1}{2}AT)^{-1}BT]^{-1}K_c(I_n - \frac{1}{2}AT)^{-1}BTK_c \\ &\approx G - H[2I_m + K_cH]^{-1}K_cG - \frac{1}{2}HK_c + \frac{1}{2}H[2I_m + K_cH]^{-1}K_cHK_c \\ &= G - H[\frac{1}{2}(I_m + \frac{1}{2}K_cH)^{-1}K_c(I_n + G)] \end{aligned} \quad (33a)$$

and

$$\hat{H}E_c = \left\{ (I_n - \frac{1}{2}AT)^{-1}BT \right.$$

$$\begin{aligned}
& - (I_n - \frac{1}{2}AT)^{-1}BT[2I_m + K_c(I_n - \frac{1}{2}AT)^{-1}BT]^{-1} \times K_c(I_n - \frac{1}{2}AT)^{-1}BT \} E_c \\
& \approx [H - H(2I_m + K_cH)^{-1}K_cH]E_c \\
& = H(I_m + \frac{1}{2}K_cH)^{-1}E_c
\end{aligned} \tag{33b}$$

Note that the exact system matrix  $G$  and input matrix  $H$  in (33) are approximations of the respective bilinear transformed system matrix,  $(I_n - \frac{1}{2}AT)^{-1}(I_n + \frac{1}{2}AT)$ , and input matrix,  $(I_n - \frac{1}{2}AT)^{-1}BT$ . These inverse approximations can be justified by the same reason shown in (28) and (29). Comparing (33) with (23), we obtain the desired digital state-feedback gain  $K_d$  and forward gain  $E_d$  as

$$K_d = \frac{1}{2}(I_m + \frac{1}{2}K_cH)^{-1}K_c(I_n + G) \tag{34a}$$

and

$$E_d = (I_m + \frac{1}{2}K_cH)^{-1}E_c \tag{34b}$$

If the exact system matrix  $G$  in (34a) and input matrix  $H$  in (34b) are replaced by the bilinear transformed models  $G_3$  and  $H_3$  in (13a) and (14a), respectively, the resulting digital redesign gains in (34a) and (34b) reduce to

$$\tilde{K}_d = K_c \left( I_n - \frac{1}{2}(A - BK_c)T \right)^{-1} \tag{34c}$$

and

$$\tilde{E}_d = (I_m - \frac{1}{2}\tilde{K}_dBT)E_c \tag{34d}$$

Since the matrix exponential formulation  $G_3$  in (13a) is equivalent to a  $w$ -plane matrix representation (bilinear transformation) [6], the obtained controller in (19), which utilizes the digital gains in (34c) and (34d), is equivalent to a  $w$ -plane pseudo-continuous controller [6].

To compare the digital redesign gains [10] in (27), we represent the matrices  $(I_m + \frac{1}{2}K_cH)^{-1}$  and  $G(= e^{AT})$  in (34a) and (34b) by a respective infinite series as

$$(I_m + \frac{1}{2}K_cH)^{-1} = I_m - \frac{1}{2}K_cH + \sum_{j=2}^{\infty} (-\frac{1}{2}K_cH)^j \tag{35a}$$

and

$$G = e^{AT} = I_n + AT + \sum_{j=2}^{\infty} \frac{1}{j!} (AT)^j \quad (35b)$$

When the sampling period  $T$  is sufficiently small, we can approximate the respective infinite series in (35) by taking the first two terms only and then solving for the desired gains, i.e.,

$$\begin{aligned} K_d &\approx \frac{1}{2} [I_m - \frac{1}{2} K_c H] K_c (2I_n + AT) \\ &= K_c + \frac{1}{2} K_c (A - BK_c) T - \frac{1}{4} K_c B K_c A T^2 \\ &\approx K_c + \frac{1}{2} K_c (A - BK_c) T \triangleq \hat{K}_d \end{aligned} \quad (36a)$$

and

$$E_d \approx [I_m - \frac{1}{2} K_c B T] E_c \triangleq \hat{E}_d \quad (36b)$$

The approximate gains ( $\hat{K}_d$  and  $\hat{E}_d$ ) in (36) are those obtained in (27).

Thus, we conclude that the digital redesign gains in (27) are the approximations of the proposed digital redesign gains in (34).

#### 4. Pseudo-continuous-time Suboptimal Quadratic Regulators

##### 4.1 Block-diagonalization via matrix sign function

In the following, the results leading to the decomposition of a continuous system into a multi-time scale structure are presented.

**Definition 1** [18]: Let the eigenvalues of a continuous-time stable system matrix,  $A \in \mathcal{R}^{n \times n}$ , be  $\lambda_i, i = 1, \dots, n$ . The non-dominant modes of this system are the modes with  $Re(\lambda_i) < -h$ , where  $h$  is a positive real number, while the dominant modes are those having  $Re(\lambda_i) > -h$ , where  $Re(\cdot)$  represents the real part of  $(\cdot)$ .

**Theorem 2** [17]: Let  $A \in \mathcal{R}^{n \times n}$  and  $\{Re(\sigma(A))\} \cap \{h_i, i = 0, 1, \dots, k\} = \emptyset$ , where  $\sigma(A)$  represents the eigenspectrum of  $A$ ,  $h_i \in \mathcal{R}, i = 0, 1, \dots, k$ . Let a set of matrix sign functions (see Appendix A-2) be

$$\text{sign}_{(h_i)}(h(A)) \triangleq \text{sign}(A - h_i I_n) \quad \text{for } i = 0, 1, \dots, k \quad (37a)$$

Define

$$S_i \triangleq \text{ind} [\text{sign}_{(h_{i-1}, h_i)}^+(h(A))] \in \mathcal{R}^{n \times n_i}, \quad 1 \leq i \leq k \quad (37b)$$

where  $\text{ind}(\cdot)$  represents the collection of the linearly independent column vectors of  $(\cdot)$  and

$$\text{sign}_{(h_{i-1}, h_i)}^+(h(A)) \triangleq \frac{1}{2} [\text{sign}_{(h_{i-1})}(h(A)) - \text{sign}_{(h_i)}(h(A))] \quad (37c)$$

with  $h_0 = 0$  and  $\text{sign}_{(0)}(h(A)) = I_n$ . Assume that  $n_i \neq 0$  for  $1 \leq i \leq k$ . Then

$$A_R = M_s^{-1} A M_s = \text{block diag} [A_{Rk}, A_{R(k-1)}, \dots, A_{R1}] \quad (38a)$$

where  $M_s$  is the right block modal matrix given by

$$M_s = [S_k, S_{k-1}, \dots, S_1] \quad (38b)$$

and

$$A_{Ri} = S_i^+ A S_i \in \mathcal{R}^{n_i \times n_i} \quad \text{for } 1 \leq i \leq k \quad (38c)$$

where  $S_i^+ \in \mathcal{R}^{n_i \times n}$  is the left inverse of  $S_i$  and is defined as  $S_i^+ \triangleq (S_i^T S_i)^{-1} S_i^T$ . ■

#### 4.2 Pseudo-continuous-time multi-stage design procedure

Let the given large-scale discrete-time system with appropriate sampling period  $T$  be

$$\mathbf{x}_d(kT + T) = \bar{G} \mathbf{x}_d(kT) + \bar{H} u_d(kT); \quad \mathbf{x}_d(0) \quad (39a)$$

Also, let the dimension of the system be  $n$  and the number of inputs be  $m$ . The procedure is to first transform the discrete-time system to an equivalent continuous-time model. Next decompose the continuous-time model into a multi-time scale structure, using techniques based on the matrix sign function, then design each decomposed subsystem via the design technique shown in Section 2, and finally determine the suboptimal digital regulator for the whole large-scale system via the new digital redesign technique shown in Section 3.2.

**Step 1:** Transform the give discrete-time system to an equivalent continuous-time system using the technique shown in Section 3.1 as

$$\dot{\mathbf{x}}_c(t) = \bar{A} \mathbf{x}_c(t) + \bar{B} u_c(t); \quad \mathbf{x}_c(0) \quad (39b)$$

Step 2: Set  $i = 1$ ,  $A := \bar{A}$ ,  $B := \bar{B}$ ,  $\bar{Q} := 0_n$  and the feedback gain  $\bar{K}_c = 0_{m \times n}$ .

Step 3: Now, specify a positive real scalar  $h_i$  (see Definition 1) and find a transformation matrix  $M_1^{(i)}$  such that the matrix  $A$  can be block-diagonalized into the following form:

$$A := (M_1^{(i)})^{-1} A M_1^{(i)} = \text{block diag} [A_c, \hat{A}_i, \bar{A}_i] \quad (40a)$$

where  $A_c \in \mathcal{R}^{(n-n_i) \times (n-n_i)}$  represents a block, which has already been designed or does not need to be designed, and the matrices  $\hat{A}_i \in \mathcal{R}^{\hat{n}_i \times \hat{n}_i}$  and  $\bar{A}_i \in \mathcal{R}^{\bar{n}_i \times \bar{n}_i}$ , with  $n_i = \hat{n}_i + \bar{n}_i$ , contain eigenvalues with real parts less than and greater than  $-h_i$ , respectively. The transformation matrix  $M_1^{(i)}$  is given by

$$M_1^{(i)} = \text{block diag} [I_{n-n_i}, (S_2, S_1)] \quad (40b)$$

where  $S_1 \in \mathcal{R}^{n \times \hat{n}_i}$  and  $S_2 \in \mathcal{R}^{n \times \bar{n}_i}$  are as defined in (37) with respect to the matrix sign function of the matrix  $A_i$ , where  $A_i := \text{block diag} [\hat{A}_i, \bar{A}_i]$ ,  $i > 1$ , and  $A_i := A$ ,  $i = 1$ . Using  $M_1^{(i)}$ , transform  $B$  as

$$B := (M_1^{(i)})^{-1} B = [B_c^T, \hat{B}_i^T, \bar{B}_i^T]^T \quad (40c)$$

The dimensions of the matrices  $B_c$ ,  $\hat{B}_i$  and  $\bar{B}_i$  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 4: The subsystem considered for design at this stage is  $(\bar{A}_i, \bar{B}_i)$ . Let the immediate optimal closed-loop continuous-time system be  $(\bar{A}_c, \bar{B}_i)$ .

Step 5: Update

$$\bar{K}_c := \bar{K}_c + [0_{m \times (n-\hat{n}_i)}, \bar{K}_i] (M_1^{(i)})^{-1} \quad (41)$$

$$A := A - B [0_{m \times (n-\hat{n}_i)}, \bar{K}_i] = \begin{bmatrix} \hat{A}_i & W_i \\ 0_{\hat{n}_i \times (n-\hat{n}_i)} & \bar{A}_c \end{bmatrix} \quad (42)$$

and

$$\bar{Q} := \bar{Q} + [(M_1^{(i)})^{-1}]^T [\text{block diag} [0_{n-\hat{n}_i}, Q_i]] (M_1^{(i)})^{-1} \quad (43)$$

where  $\bar{A}_i = \text{block diag} [A_c, \hat{A}_i]$ ,  $W_i = -[B_c^T, \hat{B}_i^T]^T \bar{K}_i$  and the dimensions of the matrices  $\hat{A}_i$  and  $W_i$ , and  $Q_i$  are  $(n - \hat{n}_i) \times (n - \hat{n}_i)$ ,  $(n - \hat{n}_i) \times \hat{n}_i$  and  $\hat{n}_i \times \hat{n}_i$ , respectively.

Step 6: Block-diagonalize the partially designed system  $A$  and move the last block of  $A$  in (42) (viz.,  $\bar{A}_{c_i}$ ) to the first block, via a transformation matrix  $M_2^{(i)}$  which is given as

$$M_2^{(i)} = \begin{bmatrix} L_i & I_{n-n_i} \\ I_{n_i} & 0_{n_i \times (n-n_i)} \end{bmatrix}, \quad (M_2^{(i)})^{-1} = \begin{bmatrix} 0_{n_i \times (n-n_i)} & I_{n_i} \\ I_{n-n_i} & -L_i \end{bmatrix} \quad (44a)$$

The matrix  $L_i$  ( $\in \mathcal{R}^{(n-n_i) \times n_i}$ ) can be solved from the following Lyapunov equation [11,12,14,17],

$$\bar{A}_i L_i - L_i \bar{A}_{c_i} + W_i = 0_{(n-n_i) \times n_i} \quad (44b)$$

The transformed system is

$$A := (M_2^{(i)})^{-1} A M_2^{(i)} = \begin{bmatrix} \bar{A}_{c_i} & 0_{n_i \times (n-n_i)} \\ 0_{(n-n_i) \times n_i} & \bar{A}_i \end{bmatrix} \quad (45a)$$

$$B := (M_2^{(i)})^{-1} B = [\bar{B}_i^T, (\bar{B}_i - L_i \bar{B}_i)^T]^T \quad (45b)$$

where  $\bar{B}_i = [B_c^T, \hat{B}_i^T]^T$ . Accumulate the transformations in  $M_1^{(i)} := M_1^{(i)} M_2^{(i)}$ .

Step 7: Set  $i := i + 1$ . If  $i > k$  ( $k$  is the number of time-scales), then go to Step 8; else, go to Step 3.

Step 8: Compute the desired digital state-feedback gain  $\bar{K}_d$  and forward gain  $\bar{E}_d$  as

$$\bar{K}_d = \frac{1}{2} (I_m + \frac{1}{2} \bar{K}_c \bar{H})^{-1} \bar{K}_c (I_n + \bar{G}) \quad (46a)$$

and

$$\bar{E}_d = (I_m + \frac{1}{2} \bar{K}_c \bar{H})^{-1} \bar{E}_c \quad (46b)$$

The digital regulator,

$$u_d(kT) = -\bar{K}_d x_d(kT) + \bar{E}_d r(kT) \quad (47)$$

with  $r(k)$  as any reference input, would place the eigenvalues of the system in (39a) near the hatched region of Fig. 2. Also, the digital regulator is 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. Note that, although some numerically stable algorithms have been suggested in the Appendix for computing some special matrices and functions, the proposed multi-stage design process does not guarantee numerical stability.

## 5. Illustrative Example

Consider an unstable discrete-time system in (39a) with

$$\bar{G} = \begin{bmatrix} 0.822 & -0.440 & 0.008 & 0.074 & 0.062 \\ 1.244 & 0.725 & 0.249 & 0.000 & 0.124 \\ -0.157 & 0.098 & 0.752 & -0.371 & -0.014 \\ 0.037 & 0.176 & 0.284 & 0.724 & -0.025 \\ -0.655 & -0.275 & -0.131 & 0.000 & 0.136 \end{bmatrix}$$

$$\bar{H} = \begin{bmatrix} 0.271 & 0.017 \\ -0.175 & -0.305 \\ -0.099 & 0.196 \\ 0.068 & 0.010 \\ 0.092 & 0.213 \end{bmatrix}$$
(48a)

where  $\sigma(\bar{G}) = \{0.7252 \pm j0.7466, 0.7537 \pm j0.3190, 0.2013\}$  and  $T \approx 0.2$ .

The location of the poles of  $\bar{G}$  in the discrete  $z$ -plane is shown in Fig. 2 and it is seen that except for the one at 0.2013, which is to be kept invariant, the rest of the poles lie outside the region of interest. The objective is to find the pseudo-continuous-time suboptimal regulators for the discrete-time system in (48a) with pole assignment near the specified region in the  $z$ -plane. The pseudo-continuous-time design procedure given in Section 4.2 will be used to achieve the desired design.

We utilize the matrix continued fraction approximation in (17) with  $q = 4$  in (18) to obtain the equivalent continuous-time system in (39b) as

$$\bar{A} = \begin{bmatrix} 0.80993 & -2.05956 & 0.32673 & 0.46503 & 0.89827 \\ 6.66468 & 0.19703 & 1.33276 & 0.00065 & 0.66065 \\ -1.29339 & 0.45801 & -1.07402 & -2.32838 & -0.20294 \\ -0.32191 & 0.82377 & 1.67148 & -1.18838 & -0.36133 \\ -3.51498 & -4.31738 & -0.70176 & 0.00000 & -8.36346 \end{bmatrix}$$

$$\bar{B} = \begin{bmatrix} 0.95385 & -0.38170 \\ -1.66643 & -1.66699 \\ -0.21358 & 1.19415 \\ 0.61711 & 0.05240 \\ 0.87785 & 1.40500 \end{bmatrix}$$
(48b)

where  $\sigma(\bar{A}) = \{0.19984 \pm j3.99969, -1.00180 \pm j2.00218, -8.01498\}$ .

Since the given system is unstable, the first step is to block-decompose the continuous-time system into its stable and unstable parts. Assign  $h_1 = 0$ . The transformation

matrix  $M_1^{(1)}$ , found using the matrix sign function technique given in Section 4.1 to block-diagonalize  $\bar{A}$ , is given by

$$M_1^{(1)} = [S_2, S_1] = \left[ \begin{array}{c} \begin{pmatrix} -0.0469 & -0.0550 & -0.2096 \\ -0.0001 & 0.0004 & 0.0008 \\ 0.2331 & 0.0121 & 1.0467 \\ 0.4191 & 0.0217 & 0.0839 \\ 0.0012 & 0.5259 & -0.0002 \end{pmatrix} \\ \begin{pmatrix} 1.0469 & 0.0550 \\ 0.0001 & 0.9996 \\ -0.2331 & -0.0121 \\ -0.4191 & -0.0217 \\ -0.0012 & -0.5259 \end{pmatrix} \end{array} \right] \quad (49a)$$

where  $S_2 \in \mathcal{R}^{5 \times 3}$  and  $S_1 \in \mathcal{R}^{5 \times 2}$  can be found from (37). The transformed matrices are

$$\begin{aligned} A &:= (M_1^{(1)})^{-1} \bar{A} M_1^{(1)} \\ &= \text{block diag} [\hat{A}_1, \bar{A}_1] \\ &= \left[ \begin{array}{ccc} \begin{pmatrix} 0.00005 & -0.00913 & 4.50053 \\ -0.01599 & -8.01497 & -0.00994 \\ -1.11375 & 0.00050 & -2.00366 \end{pmatrix} & & \\ & 0_{3 \times 2} & \\ & & \begin{pmatrix} 0.19949 & -2.39917 \\ 6.66793 & 0.20019 \end{pmatrix} \end{array} \right] \quad (49b) \end{aligned}$$

$$B := (M_1^{(1)})^{-1} \bar{B} = \begin{bmatrix} \hat{B}_1 \\ \bar{B}_1 \end{bmatrix} = \left[ \begin{array}{c} \begin{pmatrix} 2.49670 & -0.24850 \\ -0.00073 & 1.00378 \\ -0.55700 & 1.16477 \end{pmatrix} \\ \begin{pmatrix} 0.99896 & -0.00210 \\ -1.66645 & -1.66889 \end{pmatrix} \end{array} \right] \quad (49c)$$

The eigenspectra of the diagonal blocks in (49b) are  $\sigma(\hat{A}_1) = \{-1.002 \pm j2.002, -8.015\}$  and  $\sigma(\bar{A}_1) = \{0.200 \pm j4.000\}$ . The unstable subsystem  $(\bar{A}_1, \bar{B}_1)$ ,

$$\bar{A}_1 = \begin{bmatrix} 0.19949 & -2.39917 \\ 6.66793 & 0.20019 \end{bmatrix}, \quad \bar{B}_1 = \begin{bmatrix} 0.99896 & -0.00210 \\ -1.66645 & -1.66889 \end{bmatrix} \quad (49d)$$

with  $\sigma(\bar{A}_1) = \{0.200 \pm j4.000\}$ , is to be designed at this stage. Assign  $h = 1.1$  (i.e., the eigenvalues of the closed-loop system should lie to the left of the vertical line at  $-1.1$  on the negative real axis in the  $s$ -plane) and  $R = I_2$ . To achieve the design, we follow the steps of the continuous-time design procedure in Section 2.1. Let  $A = \bar{A}_1$  and  $B = \bar{B}_1$ . Solving the Riccati equation in (6) with  $(A + hI_2, B)$ , we have

$$P_0 = \begin{bmatrix} 2.246 & -0.038 \\ -0.038 & 0.509 \end{bmatrix}, \quad K_0 = R^{-1} \bar{B}_1^T P_0 = \begin{bmatrix} 2.308 & -0.886 \\ 0.059 & -0.849 \end{bmatrix} \quad (50a)$$

The resulting closed-loop system is

$$A_1 = A - BK_0 = \begin{bmatrix} -2.106 & -1.516 \\ 10.612 & -2.694 \end{bmatrix} \quad (50b)$$



where  $\sigma(A_1) = \{-2.400 \pm j4.000\}$ . Note that  $|(\operatorname{Re} \sigma(A_1))| > 1.1$ . Now, solving the Riccati equation in (7a) with  $(-A_1^2, B)$ , we have

$$\hat{Q}_1 = \begin{bmatrix} 31.350 & 1.258 \\ 1.258 & 2.489 \end{bmatrix}, \quad \text{and} \quad \frac{1}{2} \operatorname{tr}[BR^{-1}B^T\hat{Q}_1] = 20.48 \neq 0 \quad (50c)$$

Solving the Riccati equation in (7b) with  $(A_1, B)$  and  $\hat{Q}_1$ , we obtain

$$P_1 = \begin{bmatrix} 4.078 & 0.070 \\ 0.070 & 0.326 \end{bmatrix}, \quad K_1 = R^{-1}B^TP_1 = \begin{bmatrix} 3.957 & -0.473 \\ -0.126 & -0.544 \end{bmatrix} \quad (50d)$$

From (9), the constant gain is chosen as  $\gamma_1 = 0.6382$ . Therefore, the closed-loop system is

$$A_2 = A_1 - \gamma_1 BK_1 = \begin{bmatrix} -4.628 & -1.215 \\ 14.686 & -3.776 \end{bmatrix}, \quad \hat{Q}_2 = 0_2 \quad (50e)$$

where  $\sigma(A_2) = \{-4.2024 \pm j4.2024\}$ . Note that all the eigenvalues lie on the boundary of the hatched region in Fig. 1. Also, note that  $\operatorname{tr}[(A_2 + hI_2)^+] = 0$  and  $\frac{1}{2} \operatorname{tr}[BR^{-1}B^T\hat{Q}_2] = 0$ , where  $\hat{Q}_2$  is solved from (7a) with respect to  $(-A_2^2, B)$ . This verifies that the design goal has been achieved for the subsystem in (49d). Let us denote this closed-loop subsystem by  $\bar{A}_{c1} = \bar{A}_1 - \bar{B}_1(K_0 + \gamma_1 K_1)$ . The continuous-time feedback gain at this stage is

$$\bar{K}_1 = K_0 + \gamma_1 K_1 = \begin{bmatrix} 4.833 & -1.188 \\ -0.021 & -1.197 \end{bmatrix} \quad (51a)$$

and it is optimal with respect to the performance index in (2) having  $R = I_2$  and

$$\begin{aligned} \bar{Q}_1 &= 2hP_0 + \gamma_1[\hat{Q}_1 + (\gamma_1 - 1)P_1\bar{B}_1R^{-1}\bar{B}_1^TP_1] \\ &= \begin{bmatrix} 21.332 & 1.135 \\ 1.135 & 2.588 \end{bmatrix} \end{aligned} \quad (51b)$$

Using this feedback gain, the updated system is given by

$$\begin{aligned} A &:= A - B[0_{2 \times 3}, \bar{K}_1] \\ &= \begin{bmatrix} \bar{A}_1 & W_1 \\ 0_{2 \times 3} & \bar{A}_{c1} \end{bmatrix} \\ &= \begin{bmatrix} \begin{pmatrix} 0.0000 & -0.0091 & 4.5005 \\ -0.0160 & -8.0150 & -0.0099 \\ -1.1137 & 0.0005 & -2.0037 \end{pmatrix} & \begin{pmatrix} -12.0720 & 2.6690 \\ 0.0251 & 1.2002 \\ 2.7169 & 0.7320 \end{pmatrix} \\ & \quad 0_{2 \times 3} & \begin{pmatrix} -4.6284 & -1.2149 \\ 14.6859 & -3.7764 \end{pmatrix} \end{bmatrix} \end{aligned} \quad (51c)$$

The updated feedback gain  $\bar{K}_c$  and weighting matrix  $\bar{Q}$  are

$$\bar{K}_c := \bar{K}_c + [0_{2 \times 3}, \bar{K}_1](M_1^{(1)})^{-1} = \begin{bmatrix} 4.833 & -1.188 & 0.969 & 0.000 & 0.484 \\ -0.021 & -1.137 & -0.003 & -0.001 & -0.001 \end{bmatrix} \quad (51d)$$

$$\begin{aligned} \bar{Q} &:= \bar{Q} + [(M_1^{(1)})^{-1}]^T [\text{block diag}[0_3, \bar{Q}_1](M_1^{(1)})^{-1}] \\ &= \begin{bmatrix} 21.332 & 1.135 & 4.271 & 0.006 & 2.131 \\ 1.135 & 2.588 & 0.225 & 0.002 & 0.112 \\ 4.271 & 0.225 & 0.855 & 0.001 & 0.427 \\ 0.006 & 0.002 & 0.001 & 0.000 & 0.001 \\ 2.131 & 1.112 & 0.427 & 0.001 & 0.213 \end{bmatrix} \end{aligned} \quad (51e)$$

The solution of the Lyapunov equation in (44b) for  $i = 1$  and  $\bar{n}_i = 2$  is

$$L_1 = \begin{bmatrix} 0.998 & -0.891 \\ -0.552 & 0.129 \\ -1.252 & -0.114 \end{bmatrix} \quad (51f)$$

and thus the transformation matrix  $M_2^{(1)}$  that block-diagonalizes  $A$  in (51c) and swaps the blocks  $\bar{A}_{c1}$  and  $\bar{A}_1$  is given by

$$M_2^{(1)} = \begin{bmatrix} L_1 & I_3 \\ I_2 & 0_{2 \times 3} \end{bmatrix} \quad (51g)$$

The transformed matrices are now given by

$$A := (M_2^{(1)})^{-1} A M_2^{(1)} = \begin{bmatrix} \bar{A}_{c1} & 0_{2 \times 3} \\ 0_{3 \times 2} & \bar{A}_1 \end{bmatrix} \quad (52a)$$

and

$$B := (M_2^{(1)})^{-1} B = \begin{bmatrix} \begin{pmatrix} 0.9990 & -0.0021 \\ -1.6665 & -1.6689 \end{pmatrix} \\ \begin{pmatrix} 0.0153 & -1.7332 \\ 0.7648 & 1.2174 \\ 0.5031 & 0.9711 \end{pmatrix} \end{bmatrix} \quad (52b)$$

The accumulated transformation matrix becomes  $M_1^{(1)} := M_1^{(1)} M_2^{(1)}$ .

Now, we proceed to the second stage of design. Choose  $h_2 = 1.1$ . The transformation matrix  $M_1$  which block-diagonalizes the block  $\bar{A}_1$  in (51c) while preserving the block  $\bar{A}_{c1}$  is given by (as in (40b))

$$M_1^{(2)} = \begin{bmatrix} I_2 & 0_{2 \times 3} \\ 0_{3 \times 2} & (S_2 \ S_1) \end{bmatrix} \quad (52c)$$

where

$$[S_2, S_1] = \begin{bmatrix} \begin{pmatrix} 0.0011 \\ 1.0000 \\ 0.0001 \end{pmatrix} & \begin{pmatrix} 1.0000 & 0.0000 \\ -0.0020 & -0.0001 \\ 0.0000 & 1.0000 \end{pmatrix} \end{bmatrix} \quad (52d)$$

The submatrices  $S_2 \in \mathcal{R}^{3 \times 1}$  and  $S_1 \in \mathcal{R}^{3 \times 2}$  can be found from (37) with respect to  $\bar{A}_1$  and  $h_2$ . The transformed matrices  $A$  and  $B$  are

$$\begin{aligned} A &:= (M_1^{(2)})^{-1} A M_1^{(2)} \\ &= \text{block diag} [\bar{A}_{c1}, \hat{A}_2, \bar{A}_2] \\ &= \begin{bmatrix} \begin{pmatrix} -4.6284 & -1.2149 \\ 14.6859 & -3.7764 \end{pmatrix} & 0_{2 \times 1} & 0_{2 \times 2} \\ 0_{1 \times 2} & (-8.0150) & 0_{1 \times 2} \\ 0_{2 \times 2} & 0_{2 \times 1} & \begin{pmatrix} 0.0001 & 4.5005 \\ -1.1137 & -2.0040 \end{pmatrix} \end{bmatrix} \end{aligned} \quad (53a)$$

$$B := (M_1^{(2)})^{-1} B = \begin{bmatrix} \bar{B}_1 \\ \hat{B}_2 \\ \bar{B}_2 \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 0.9990 & -0.0021 \\ -1.6665 & -1.6689 \end{pmatrix} \\ \begin{pmatrix} 0.7649 & 1.2140 \end{pmatrix} \\ \begin{pmatrix} 0.0145 & -1.7346 \\ 0.5030 & 0.9709 \end{pmatrix} \end{bmatrix} \quad (53b)$$

Again, the accumulated transformation matrix becomes  $M_1^{(2)} := M_1^{(1)} M_1^{(2)}$ . The subsystem to be designed at this stage is  $(\bar{A}_2, \bar{B}_2)$ . Following the same procedures as in the first stage, we obtain the designed continuous-time subsystem as

$$\bar{A}_{c2} = \begin{bmatrix} -0.1156 & 7.0817 \\ -1.2639 & -4.2294 \end{bmatrix} \quad (53c)$$

with  $\sigma(\bar{A}_{c2}) = \{-2.1725 \pm j2.1725\}$ . Note that these eigenvalues are within the hatched region of Fig. 1. The continuous-time feedback gain  $\bar{K}_2$  and weighting matrix  $\bar{Q}_2$  at this stage are

$$\bar{K}_2 = \begin{bmatrix} 0.4205 & 1.5278 \\ -0.0632 & 1.5008 \end{bmatrix} \quad (53d)$$

and

$$\bar{Q}_2 = \begin{bmatrix} 2.0109 & 3.3163 \\ 3.3163 & 9.2675 \end{bmatrix} \quad (53e)$$

The updated feedback gain  $\bar{K}_c$  and weighting matrix  $\bar{Q}$  are given below:

$$\begin{aligned} \bar{K}_c &:= \bar{K}_c + [0_{2 \times 3}, \bar{K}_2](M_1^{(2)})^{-1} \\ &= \begin{bmatrix} 6.74659 & -0.63902 & 2.79527 & 0.20152 & 0.63200 \\ 1.85769 & -1.08109 & 1.88639 & -0.99297 & 0.19219 \end{bmatrix} \end{aligned} \quad (53f)$$

$$\bar{Q} := \begin{bmatrix} 35.87594 & 6.16143 & 17.95729 & 3.93400 & 3.16016 \\ 6.16143 & 4.97772 & 4.80710 & 3.19066 & 0.39323 \\ 17.95729 & 4.80710 & 13.76777 & 3.28138 & 1.41166 \\ 3.93400 & 3.19066 & 3.28138 & 6.20283 & 0.07071 \\ 3.16016 & 3.93227 & 1.41166 & 0.07071 & 0.29412 \end{bmatrix} \quad (53g)$$

where  $\sigma(\bar{Q}) = \{48.298, 7.506, 3.368, 0.000, 1.947\}$  and  $\bar{Q} \geq 0$ .

The eigenvalues of  $A - B[0_{2 \times 3}, \bar{K}_2]$  with  $A$  and  $B$  as in (53a) and (53b) are  $\{-4.2024 \pm j4.2024, -2.1725 \pm j2.1725, -8.0150\}$ . Note that all of them are within the hatched region of Fig. 1, and the non-dominant eigenvalue of the open-loop system at  $-8.0150$  is kept invariant. Therefore, the closed-loop continuous-time system is

$$\begin{aligned} \bar{A}_c &= \bar{A} - \bar{B}\bar{K}_c \\ &= \begin{bmatrix} -4.91622 & -1.86269 & -1.61950 & -0.10620 & 0.36880 \\ 21.00416 & -2.67002 & 9.13548 & -1.31880 & 2.03420 \\ -2.07080 & 1.61251 & -2.72963 & -1.09959 & -0.29746 \\ -4.58264 & 1.27476 & -0.15235 & -1.26071 & -0.76142 \\ -12.04750 & -2.23749 & -5.80596 & 1.21822 & -9.18828 \end{bmatrix} \end{aligned} \quad (54a)$$

The continuous-time optimal regulator is given by

$$u_c(t) = -\bar{K}_c x_c(t) + \bar{E}_c r(t) \quad (54b)$$

where  $\bar{K}_c$  is the total feedback gain as in (53f),  $\bar{E}_c = I_2$ , and  $r(t)$  is any reference input.

The digital redesigned closed-loop system will be of the form as shown in (22) with the digital state-feedback gain  $\bar{K}_d$  and forward gain  $\bar{E}_d$  in (47) to be determined. With  $\bar{G}$  and  $\bar{H}$  as in (48a),  $\bar{E}_c = I_2$ , and  $\bar{K}_c$  as in (53f), the gains  $\bar{K}_d$  in (34a) and  $\bar{E}_d$  in (34b) can be evaluated as follows:

$$\begin{aligned} \bar{K}_d &= \frac{1}{2}(I_2 + \frac{1}{2}\bar{K}_c\bar{H})^{-1}\bar{K}_c(I_5 + \bar{G}) \\ &= \begin{bmatrix} 2.82346 & -0.82731 & 1.05826 & 0.17590 & 0.26367 \\ 0.10625 & -0.85000 & 0.81854 & -0.85244 & 0.02810 \end{bmatrix} \end{aligned} \quad (55a)$$

and

$$\bar{E}_d = (I_2 + \frac{1}{2}\bar{K}_c\bar{H})^{-1}\bar{E}_c = \begin{bmatrix} 0.56005 & -0.20157 \\ -0.09247 & 0.75738 \end{bmatrix} \quad (55b)$$

The designed closed-loop digital system matrix in (22) with the gains in (55) is

$$\begin{aligned} \bar{G} &= \bar{G} - \bar{H}\bar{K}_d \\ &= \begin{bmatrix} 0.05504 & -0.20135 & -0.29270 & 0.04082 & -0.00993 \\ 1.77051 & 0.32097 & 0.68385 & -0.22921 & 0.17871 \\ 0.10170 & 0.18270 & 0.69633 & -0.18651 & 0.00660 \\ -0.15606 & 0.24076 & 0.20385 & 0.72056 & -0.04321 \\ -0.93739 & -0.01784 & -0.40271 & 0.16539 & 0.10576 \end{bmatrix} \end{aligned} \quad (56)$$

where  $\sigma(\hat{G}) = \{0.2582 \pm j0.4070, 0.5905 \pm j0.2718, 0.2013\}$ . These eigenvalues close to the digitized continuous-time optimal eigenvalues,  $\{0.2879 \pm j0.3215, 0.5874 \pm j0.2726, 0.2013\}$ , of the system matrix  $\bar{A} - \bar{B}\bar{K}_c$ .

The simulations of the closed-loop systems in (5) and (22) with  $r(t)$  as a unit-step vector are shown in Fig. 3. It can be seen that all the discrete states  $x_d(kT)$  closely match the continuous-time states  $x_c(t)$  at  $t = kT$ . Also, the simulations of the continuous-time quadratic regulator in (3) with  $\bar{E}_c = I_2$  and  $\bar{K}_c$  as in (53f) and the discrete-time control law in (47) with  $\bar{K}_d$  and  $\bar{E}_d$  as in (55) are shown in Fig. 5. The continuous function  $u_c(t)$  in (3) closely matches the discrete function  $u_d(kT)$  in (47). The same simulation results have been obtained by using the approximated digital gains  $\hat{K}_d$  in (34c) and  $\hat{E}_d$  in (34d). The simulations were also carried out with the approximated digital feedback gain  $\hat{K}_d$  and forward gain  $\hat{E}_d$  as given in (36) and shown in Fig. 4. For this case, a rather large discrepancy occurs in the transient region due to the utilization of the roughly approximated digital gains in (36). It might be interesting to note that the direct use of the digitized  $u_c(t)$  in (54b) with  $\bar{K}_c$  in (53f) and  $\bar{E}_c = I_2$  to the system in (48a) results in an unstable response.

Since all the designed digital states closely match the continuous-time optimal states and the designed digital regulator closely matches the continuous-time quadratic regulator, the designed digital regulator in (47) with  $\bar{K}_d$  and  $\bar{E}_d$  as in (55) can be termed as a pseudo-continuous-time quadratic regulator.

## 6. 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 paper. It has been shown that a large-scale pseudo-continuous-time 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 near 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  $q$ th root of the system matrix and the matrix sign functions. A new digital redesign technique based on matching all the states at all the sampling instants has been developed for finding the pseudo-continuous-time regulator with appropriate pole assignment. With an appropriately sampling period  $T$ , 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. 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 [6] can only be applied to design single variable digital control systems for obtaining the cascaded controllers.

## 7. Acknowledgements

The authors wish to express their gratitude for the valuable remarks and suggestions made by the reviewers and by Dr. Jagdish Chandra, Director of the Mathematical Sciences Division, U. S. Army Research Office.

This work was supported in part by the U.S. Army Research Office under contract DAAL-03-87-K-0001, the U.S. Army Research and Development Command under contract DAAH-01-85-C-A111, and the NASA-Johnson Space Center under contract NAG9-380.

## References

- [1.] Ackermann, J. (1985), *Sampled-data Control Systems*, Springer-Verlag, p. 234.
- [2.] Anderson, B.D.O. and J.B. Moore (1971), *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, New Jersey.
- [3.] Aoki, M. (1968), "Control of large-scale dynamic systems by aggregation," *IEEE Trans. Autom. Contr.*, AC-13, 246-253.
- [4.] Bierman, G.J. (1984), "Computational aspects of the matrix sign function solution to the ARE," *Proc. 23rd Conf. Decision Contr.*, 514-519.
- [5.] Houpis, C.H. (1985), "Refined design method for sampled-data control systems: the pseudo-continuous-time control system design," *IEE Proc.*, Vol. 132, Pt. D, 69-74.
- [6.] Houpis, C.H. and G.B. Lamont (1985), *Digital Control Systems*, McGraw-Hill, New York.
- [7.] Kailath, T. (1984), *Linear Systems*, Prentice-Hall, p. 656.

- [8.] Kawasaki, N. and E. Shimemura (1983), "Determination quadratic weighting matrices to locate poles in a specified region," *Automatica*, Vol. 19, 557-560.
- [9.] Kokotovic, P.V., R.E. O'Malley and P. Sannuti (1976), "Singular perturbations and order reduction in control theory — an overview," *Automatica*, Vol. 12, 123-132.
- [10.] Kuo, B.C. (1980), *Digital Control Systems*, Holt, Rinehart and Winston, 321-338.
- [11.] Mahmoud, M.S., Y. Chen and M.G. Singh (1985), "On eigenvalue assignment in discrete systems with fast and slow modes," *Int. J. Syst. Sci.*, Vol. 16, 61-70.
- [12.] Mahmoud, M.S., Y. Chen and M.G. Singh (1986), "Discrete two-time scale systems," *Int. J. Syst. Sci.*, Vol. 17, 1187-1207.
- [13.] Moler, C. and C.V. Loan (1978), "Nineteen dubious ways to compute the exponential of a matrix," *SIAM Rev.*, Vol. 20, 801-836.
- [14.] Naidu, D.S. and D.B. Price (1987), "Time-scale synthesis of a closed-loop discrete optimal control system," *J. Guidance, Contr. Dynam.*, Vol. 32, 417-422.
- [15.] Porter, B. and R. Crossley (1972), *Modal Control - Theory and Applications*, Taylor and Francis, London.
- [16.] Shieh, L.S., Y.T. Tsay and R.E. Yates (1983), "Some properties of matrix sign functions derived from continued fractions," *IEE Proc.*, Vol. 130, Pt. D, 111-118.
- [17.] Shieh, L.S., Y.T. Tsay, S.W. Lin and N.P. Coleman (1984), "Block-diagonalization and block-triangularization of a matrix via the matrix sign function," *Int. J. Syst. Sci.*, Vol. 15, 1203-1220.
- [18.] Shieh, L.S. and Y.T. Tsay (1984), "Algebra-geometric approach for the model reduction of large-scale multivariable systems," *IEE Proc.*, Vol. 131, Pt. D, 23-36.
- [19.] Shieh, L.S., H.M. Dib and B.C. McInnis (1986), "Linear quadratic regulators with eigenvalue placement in a vertical strip," *IEEE Trans. Autom. Contr.*, AC-31, 241-243.
- [20.] Shieh, L.S., J.S.H. Tsai and S.R. Lian (1986), "Determining continuous-time state equations from discrete-time state equations via the principal  $q$ th root method," *IEEE Trans. Autom. Contr.*, AC-31, 454-457.
- [21.] Shieh, L.S., S.R. Lian and B.C. McInnis (1987), "Fast and stable algorithms for computing the principal square root of a complex matrix," *IEEE Trans. Autom. Contr.*, AC-32, 820-822.
- [22.] Shieh, L.S., H.M. Dib and S. Ganesan (1987), "Linear quadratic regulators with eigenvalue placement in a horizontal strip," *Int. J. Syst. Sci.*, Vol. 19, 1279-1290.
- [23.] Shieh, L.S., H.M. Dib and S. Ganesan (1987), "Continuous-time quadratic regulators and pseudo-continuous-time quadratic regulators with pole placement in a specific region," *IEE Proc.*, Vol. 134, Pt. D, 338-346.
- [24.] Tsai, J.S.H., L.S. Shieh and R.E. Yates (1988), "Fast and stable algorithms for computing the principal  $n$ th root of a complex matrix and the matrix sector function,"

Appendix

**A-1 Principal  $n$ th root of a matrix [21,24]**

**Definition A.1:** Let a matrix  $A \in C^{m \times m}$  have an eigenspectrum  $\sigma(A) = \{\lambda_i, i = 1, \dots, m\}$ ,  $\lambda_i \neq 0$  and  $\arg(\lambda_i) \neq \pi$ . Then, the principal  $n$ th root of  $A$  is defined as  $\sqrt[n]{A} \in C^{m \times m}$ , where  $n$  is a positive integer, and

- (a)  $(\sqrt[n]{A})^n = A$
- (b)  $\arg(\sigma(\sqrt[n]{A})) \in (-\pi/n, +\pi/n)$ .

A generalized fast and stable algorithm with  $k$ th order convergence has been derived in [21,24] for computing the principal  $n$ th root of a given complex matrix  $A \in C^{m \times m}$ . The algorithm corresponding to quadratic convergence ( $k = 2$ ) is listed below.

$$G(k+1) = G(k) \left( [2I_m + (n-2)G(k)] [I_m + (n-1)G(k)]^{-1} \right)^n,$$

$$G(0) \triangleq A, \quad \lim_{k \rightarrow \infty} G(k) = I_m \tag{A.1.a}$$

$$R(k+1) = R(k) [2I_m + (n-2)G(k)]^{-1} [I_m + (n-1)G(k)],$$

$$R(0) \triangleq I_m, \quad \lim_{k \rightarrow \infty} R(k) = \sqrt[n]{A} \tag{A.1.b}$$

**A-2 Matrix sign function [16]**

The matrix sign function of a matrix  $A \in C^{m \times m}$  [16,18,21] is defined as

$$\text{sign}(A) = A(\sqrt{A^2})^{-1} = A^{-1}(\sqrt{A^2}) \tag{A.2}$$

where the matrix  $\sqrt{A^2}$  denotes the principal square root of  $A^2$ . A fast and stable algorithm [21] to compute the matrix sign function is listed below. For  $k = 0, 1, \dots$ ,

$$P_j(k) = P_{j-1}(k) + S^{-2}(k)Q_{j-1}(k), \quad P_1(k) = I_m, \quad \text{and}$$

$$Q_j(k) = P_{j-1}(k) + Q_{j-1}(k), \quad Q_1(k) = I_n, \quad \text{with } j = 2, \dots, r \tag{A.3.a}$$

$$S(k+1) = S(k)Q_r^{-1}(k)P_r(k), \quad S(0) = A, \quad \lim_{k \rightarrow \infty} S(k) = \text{sign}(A) \tag{A.3.b}$$

where  $r$  is the order of the desired rate of convergence.



### A-3 Solving 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

$$PBR^{-1}B^T P - A^T P - PA - Q = 0 \quad (A.4.a)$$

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 [4,17], and the eigenvalue-eigenvector approach [7]. Consider the Hamiltonian associated with the given system

$$H = \begin{bmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{bmatrix} \quad (A.4.b)$$

The following algorithm can be utilized to obtain the solution  $P$ .

$$H_{k+1} = \frac{1}{2}[H_k + H_k^{-1}], \quad H_0 = H, \quad \text{and} \\ \lim_{k \rightarrow \infty} H_k = \text{sign}(H) \quad (A.5.a)$$

Let

$$\text{sign}^+(H) \triangleq \frac{1}{2}[I_{2n} + \text{sign}(H)] \quad (A.5.b)$$

Construct a block modal matrix  $X$  as

$$X = [\text{ind}(\text{sign}^+(H)), \text{ind}(I_{2n} - \text{sign}^+(H))] \triangleq \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \quad (A.6.a)$$

where  $\text{ind}(\cdot)$  represents the collection of the linearly independent column vectors of  $(\cdot)$ .

Then, we have

$$P = X_{22}(X_{12})^{-1} \quad (A.6.b)$$

To alleviate the problems of computing  $H_k^{-1}$ , the Hamiltonian can be transformed into a symmetric form as follows [4]

$$\hat{H} = \hat{J}H = \begin{bmatrix} 0_n & -I_n \\ I_n & 0_n \end{bmatrix} H = \begin{bmatrix} Q & A^T \\ A & -BR^{-1}B^T \end{bmatrix} \quad (A.7.a)$$

Then, the algorithm in (A.5) becomes

$$\hat{H}_{k+1} = \frac{1}{2}[\hat{H}_k + \hat{J}\hat{H}_k^{-1}\hat{J}], \quad \hat{H}_0 = \hat{J}H, \quad \text{and} \\ \lim_{k \rightarrow \infty} (-\hat{J}\hat{H}_k) = \text{sign}(H) \quad (A.7.b)$$

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.6).

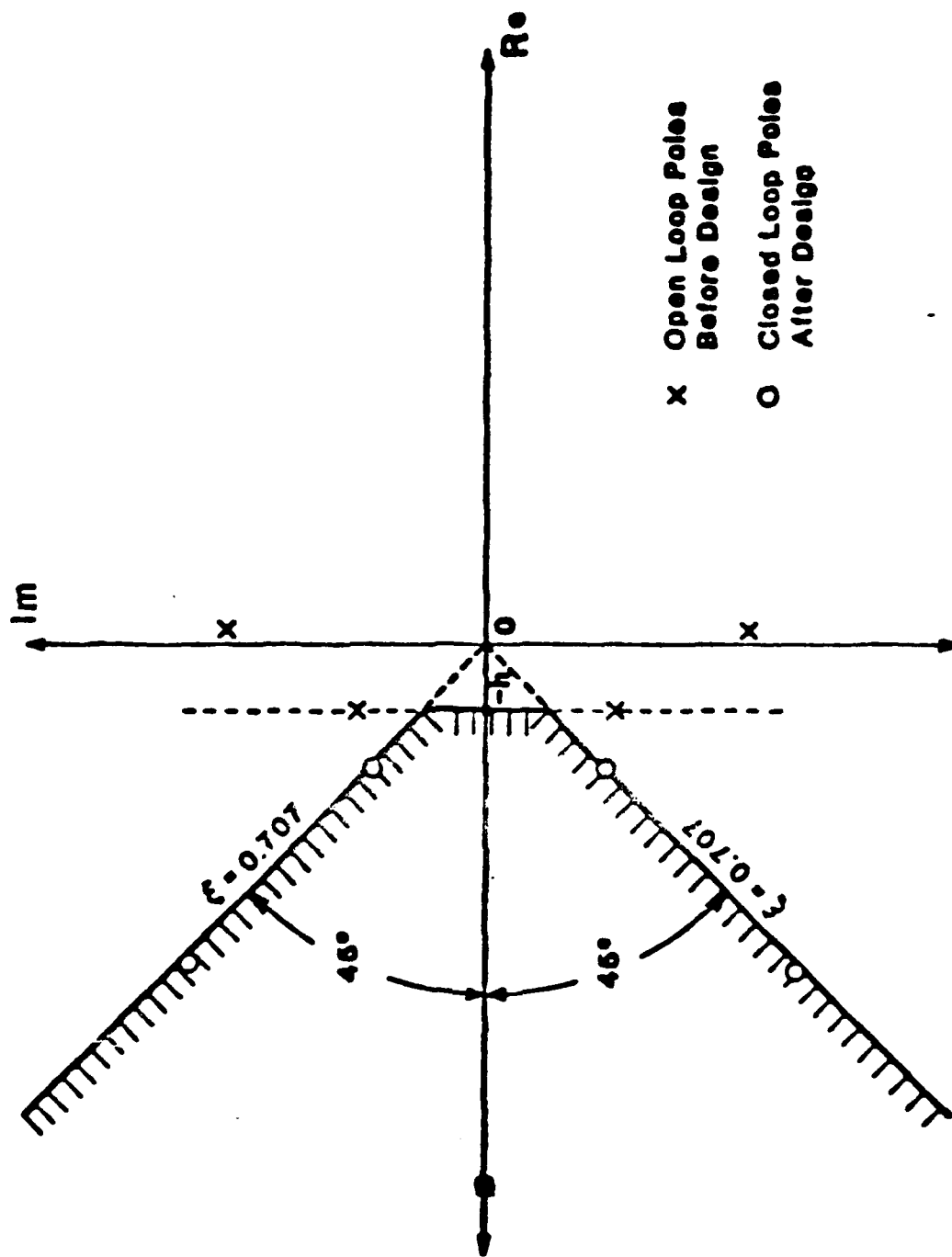


Fig. 1 The region of interest in the continuous-time s-plane.

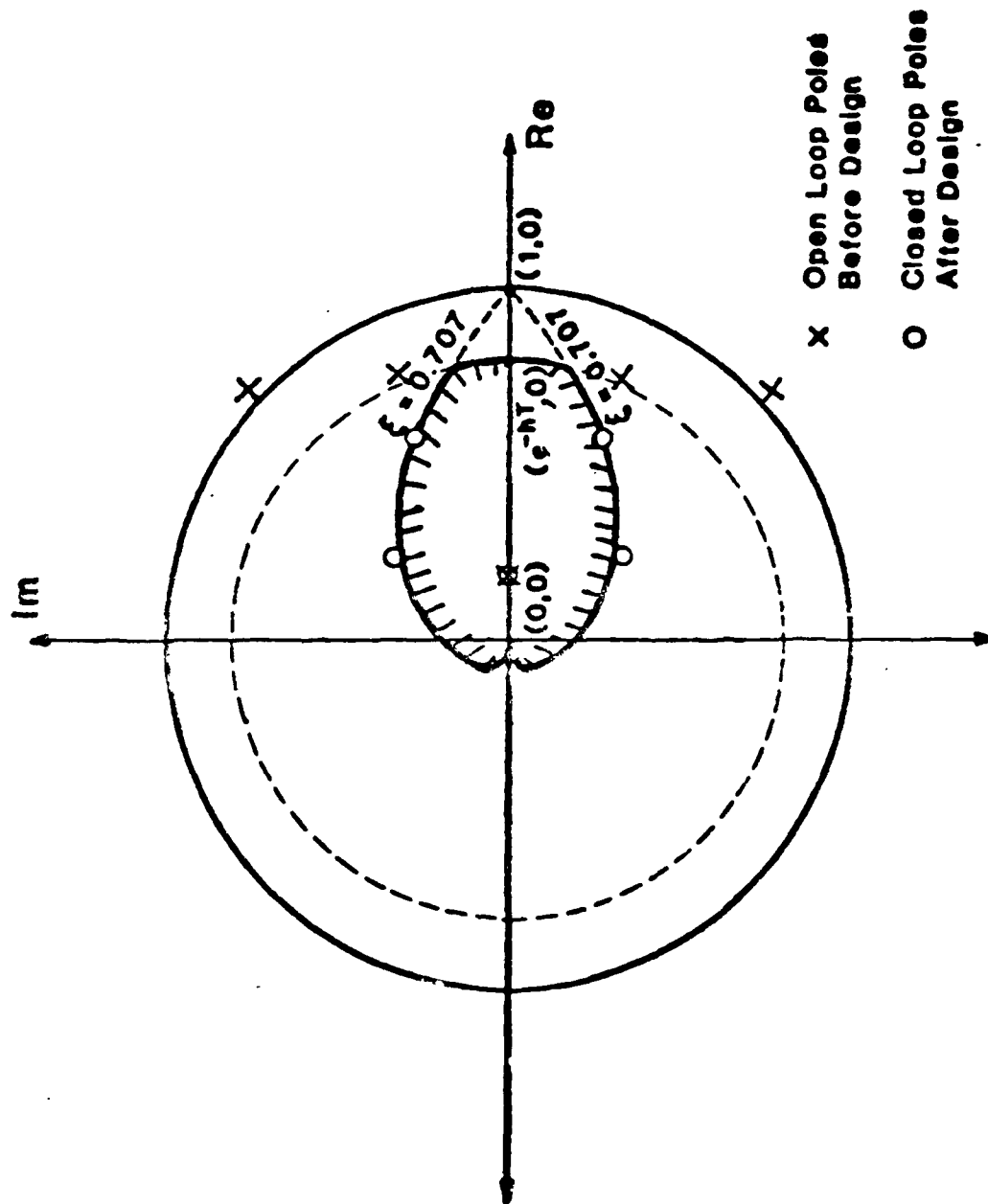


Fig. 2 The region of interest in the discrete-time z-plane.

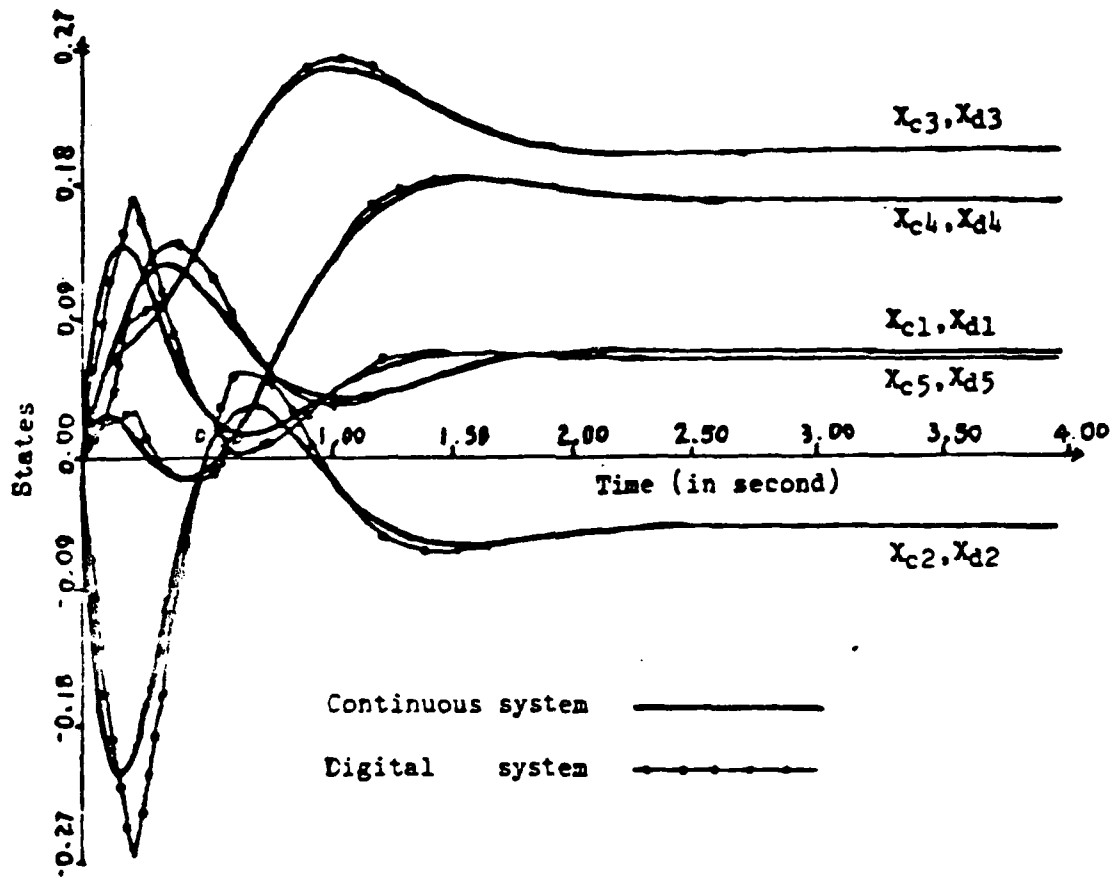


Fig.3. Comparison of the state trajectories of  $X_c(t)$  and  $X_d(t)$  with  $K_d$  and  $E_d$  in equ. (34).

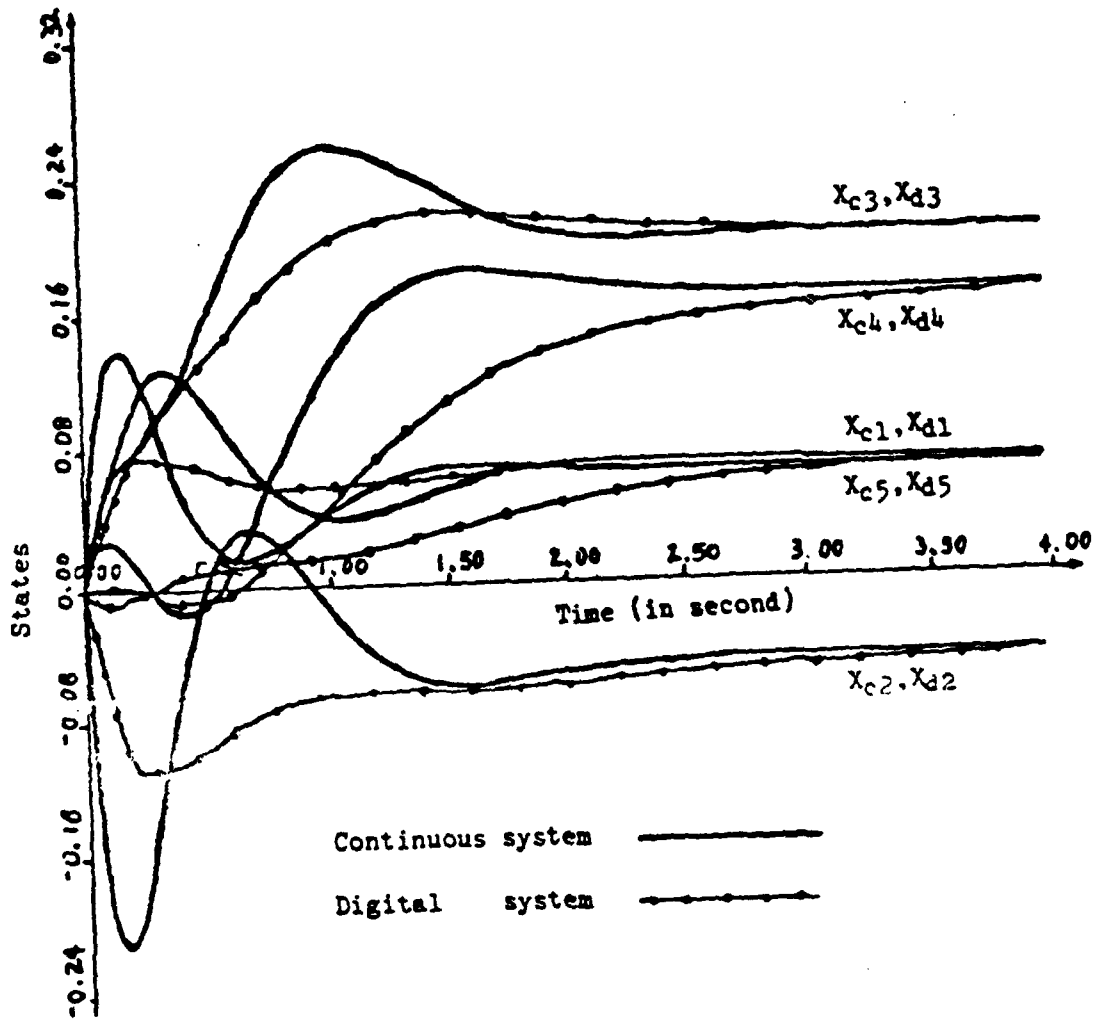


Fig.4. Comparison of the state trajectories of  $X_c(t)$  and  $X_d(t)$  with  $\hat{K}_d$  and  $\hat{E}_d$  in equ. (27).

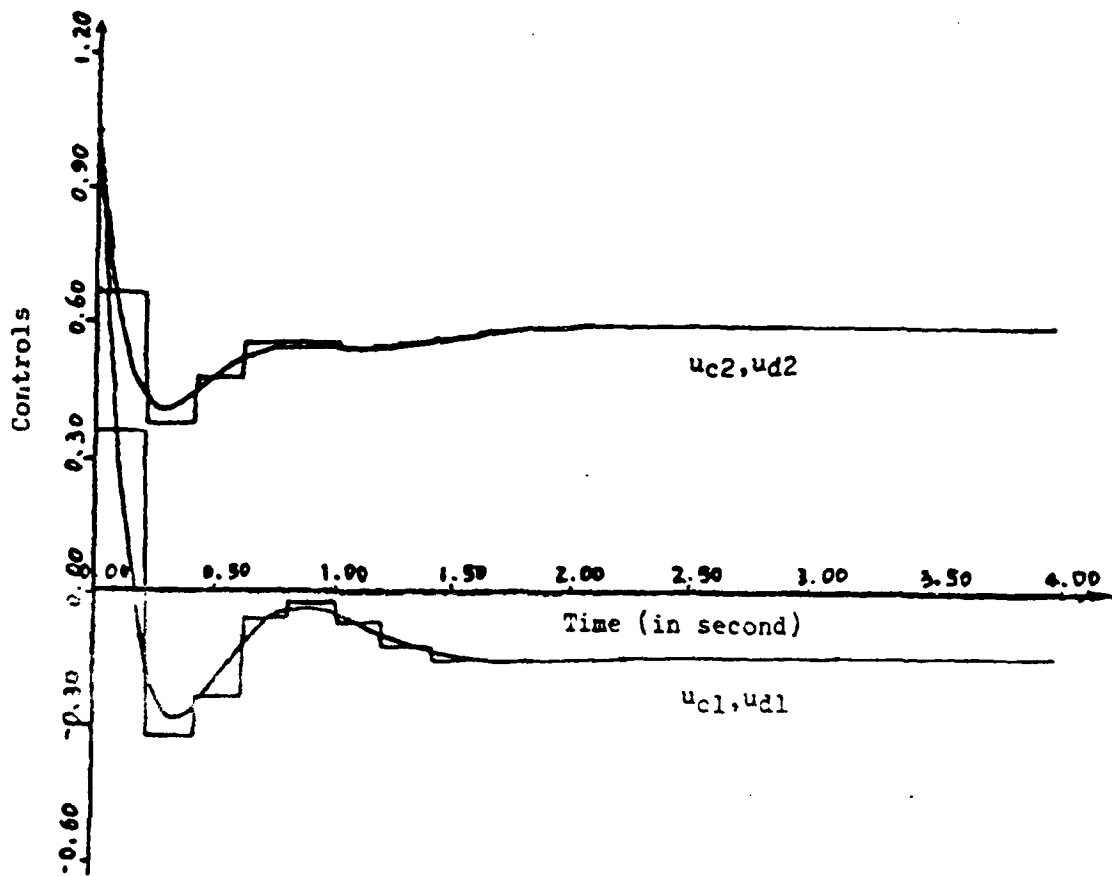


Fig.5. Control signals  $u_c(t)$  in equ. (54b) and  $u_d(kT)$  in equ. (47).

## NEW METHODOLOGIES IN RENEWAL THEORY

B. D. Sivazlian

Department of Industrial and Systems Engineering

The University of Florida

Gainesville, Florida 32611

### ABSTRACT

The derivation of the probability law (joint distribution function) of a renewal counting process, and the analysis of the filtered renewal process, form the theoretical basis 1) to study the prediction problem associated with systems regulated by these processes, and 2) to formulate and solve a number of applied problems arising in reliability, replacement, maintenance, queueing, production and other pertinent areas of interest to engineering, operations research and military systems. The research addresses itself to both theory and applications.

### I. INTRODUCTION

As an important branch of stochastic processes, renewal theory has found many useful applications in statistics and in the mathematical modeling of natural and man-created phenomena, particularly in solving complex problems in operations research such as inventory, queueing, reliability and replacement. As a process generalizing the Poisson process and all its ramifications, renewal theory has found applications in such fields as actuarial sciences, astronomy, astrophysics, ecology, economics, engineering, meteorology and physics.

The literature on renewal process is at least 50 years old. Lotka's paper (1939) on "... self-renewing aggregates ..." contains a list of 74 papers on the subject of renewal equation and its applications dating as far back as 1909 when Herbelot encountered the equation while investigating an actuarial problem. In his fundamental work, Feller (1941) is the first to study formally the integral equation of renewal theory. Smith (1958) provides a thorough review of renewal theory. Cox (1962) discusses many theoretical and applied problems in the area. Daley and Vere-Jones (1988) and Wolff (1988) present a more modern approach to the theory. Without elaborating further on the existing literature, it suffices to say that renewal theory is a standard topic covered by most textbooks on stochastic processes and their applications.

Renewal theory has found a very fruitful area of application in modeling complex systems in reliability, maintainability and availability. Renewal theory has been used for example (see Barlow and Proschan, 1965): 1) to define the operating characteristics of maintenance policies, 2) to solve the age and block replacement problem, 3) to formulate repair problems of single- and multi-units, and 4) to derive optimum inspection and maintenance policies. The theory has also been applied to solve problems in reliability arising from shock processes, cumulative damages and redundancies.

Other areas of applications in operations research where renewal theory has been utilized have been: single and multi-commodity inventory systems, queueing systems, maintenance and replacement systems. More recently, in using diffusion approximation to solve complex queueing systems, such as the machine repair problem with standbys, renewal



theory has been used to generate the infinitesimal means and variances to the diffusion equation.

This research proposes the development of a unifying methodology to bring forth a new perspective to the analysis of renewal processes. The proposed research aims at: 1) obtaining new results in the field such as the characteristics of the probability law of a renewal counting process; 2) studying the theory of filtered renewal process; 3) developing efficient procedures to predict the behavior of systems governed by renewal processes for short-term and medium-term purposes, and 4) using the results obtained in a variety of applied problems arising in engineering, operations research and military systems.

## II. THE THEORY OF RENEWAL PROCESSES

### 1. The Probability Law of a Renewal Counting Process

The study of a stochastic process is not complete until one has characterized its probability law, that is, in our case, the joint distribution function of the number of renewals at distinct time epochs  $t_1, t_2, \dots, t_n$ , where  $0 < t_1 < t_2 < \dots < t_n$  ( $n$  being an arbitrary positive integer). The probability law provides all the necessary information to describe the properties of the process. Unfortunately, despite its early inception and its many usage, the probability law of a renewal counting process has been considered so far to be too difficult a task to tackle, and thus remains an unsolved problem. Only in the special case of the Poisson process defined as a renewal process, has this law been derived, leading to important characterization of the process itself. For example, it may be shown that the Poisson process has stationary independent increments, two properties that form the

basis for many important applications in statistics and operations research (see e.g. Cohen, 1982).

Let  $(T_i)$ ,  $i = 1, 2, \dots$ , be the sequence of interarrival times in an ordinary renewal process, assumed to be i.i.d. random variables with probability density function  $f(x)$ ,  $0 < x < \infty$ , and distribution function  $F(x)$ . Let  $(N(t), t \geq 0)$  be the total number of renewals in  $[0, t]$  where  $N(0) = 0$ . Consider distinct time epochs  $t_1, t_2, \dots, t_m$ , where  $0 < t_1 < t_2 < \dots, t_m$  and  $m$  is any positive integer.

The probability law of the renewal counting process may be defined by

$$P(N(t_1) = n_1, N(t_2) = n_2, \dots, N(t_m) = n_m) \quad (1)$$

where  $0 \leq n_1 \leq n_2 \leq \dots \leq n_m$ . Other representations of the probability law of the process, such as the joint characteristic function, may be appropriate depending on the nature of the intended results.

We propose a new methodology to formulate mathematically the probability law of a renewal counting process and to obtain closed form expressions in terms of the distribution of the interarrival times. This new methodology is based on the properties of certain classes of multiple integrals. In addition, basic properties of the process can be identified and structured such as the distribution of renewal increments  $N(t_2) - N(t_1)$ , the joint distribution of the number of renewals  $N(t)$  and the forward recurrence time  $V(t)$ , etc. The results obtained will be of particular use in developing mathematical models for prediction.

## 2. Methodology

### a. Introduction

The objective of the present section is to show that a class of multiple integrals may be used as a novel mathematical methodology to solve problems arising in renewal counting processes. Multiple integrals provide a natural vehicle to approach these complex problems as one is essentially dealing with sums of independent random variables in the context of interarrival times. The class of multiple integrals typically arising in these problems is of the generalized Liouville type (Sivazlian, 1971). For example, we use this methodology to provide a new derivation to the distribution of the number of renewals in an ordinary renewal process. The primary emphasis is to demonstrate the use of multiple integrals as a method of analysis and solution, rather than to derive the specific intended result in the shortest number of steps. The available method for obtaining this result in the existing literature is much shorter; it relies however, on event arguments relating waiting times to number of renewals, which are restrictive (see e.g. Cox, 1962). The present derivation is based on the joint distribution of the number of renewals and the interarrival times. Moreover, the intent is to suggest a methodology which could be used to solve more complex problems in renewal theory such as 1) deriving the probability law of a renewal counting process 2) characterizing the statistical properties of the filtered renewal process and 3) providing a basis for predicting systems behavior which are of the renewal type. We first state a result in multiple integrals.

b. A Result in Multiple Integral

Theorem

Define for  $x > 0$ , the function  $g(x) \in \mathcal{C}$  (i.e. continuous) and the function  $\phi_i(x) \in \mathcal{K}$  (i.e. with at most a finite number of points of discontinuity in every finite interval and such that the integral

$$\int_0^x |\phi_i(u)| du \text{ has a finite value for every } x > 0, i=1,2,\dots, n \text{ where } n$$

is a positive integer (Mikusinski, 1959). Then

$$\int \int \cdots \int_R g(x_1+x_2 + \cdots + x_n) \phi_1(x_1)\phi_2(x_2) \cdots \phi_n(x_n) dx_1 dx_2 \cdots dx_n$$

$$= \int_0^t g(u) [\phi_1(u)*\phi_2(u) * \cdots * \phi_n(u)] du \quad (2)$$

where  $R = \{x: 0 < x_1+x_2 + \cdots + x_n \leq t, x_i \geq 0\}$  and where the integrand on the right hand single integral is a function of class  $\mathcal{K}$ . Here the notation  $*$  refers to the usual convolution operation. (For a proof see Sivazlian, 1971.)

c. The Distribution of the Number of Renewals

Let  $\{T_i\}$ ,  $i=1,2,\dots$ , be the sequence of interarrival times in an ordinary renewal process, assumed to be independently and identically distributed random variables with probability density function  $f(x)$ ,  $0 < x < \infty$ , and distribution function  $F(x)$ . Let  $\{N(t), t \geq 0\}$  be the total number of renewals in  $[0,t]$  where  $N(0)=0$ . The joint distribution of  $N(t)$  and  $T_1, T_2, \dots, T_{N(t)}$  is:

a. For  $N(t)=0, t \geq 0$ :

$$P(N(t)=0, T_1 > t) = 1 - F(t) \quad (3)$$

b. For  $N(t)=n \geq 1$ ,  $0 < x_1+x_2 + \dots + x_n \leq t$ :

$$\begin{aligned} P(N(t)=n, x_1 < T_1 \leq x_1+dx_1, x_2 < T_2 \leq x_2+dx_2, \dots, \\ x_n < T_n \leq x_n+dx_n, T_{n+1} > t-(x_1+x_2 + \dots + x_n)) \\ = f(x_1) \cdot f(x_2) \dots f(x_n) [1 - F(t-(x_1+x_2 + \dots + x_n))] dx_1 dx_2 \dots dx_n \end{aligned} \quad (4)$$

Thus, the probability mass function of  $N(t)$  is:

For  $N(t)=0$ :

$$P(N(t)=0) = 1 - F(t)$$

For  $N(t)=n \geq 1$ :

$$\begin{aligned} P(N(t)=n) &= \int \int \dots \int_{0 < x_1+x_2+\dots+x_n \leq t} f(x_1) \cdot f(x_2) \dots f(x_n) \\ &\quad (1 - F[t-(x_1+x_2 + \dots + x_n)]) dx_1 dx_2 \dots dx_n \\ &= \int_0^t f^{*(n)}(u) [1 - F(t-u)] du \\ &= F^{(n)}(t) - F^{(n+1)}(t) \end{aligned} \quad (5)$$

which is a well-known result.

#### d. The Joint Distribution Function of the Number of Renewals

Consider two time epochs  $t_1$  and  $t$ , where  $0 < t_1 < t$ , and suppose that it is required to determine  $P(N(t_1)=n_1, N(t)=n)$ . Here, it is necessary to consider several cases depending on the values taken by  $n_1$  and  $n$  ( $0 \leq n_1 \leq n$ ).

For example, if we consider the case  $n_1 \geq 1$ ,  $n \geq n_1 + 2$ , then  $P(N(t)=n_1, N(t)=n)$  will be given by:

$$P(N(t)=n_1, N(t)=n)$$

$$= \int \int \cdots \int_R f(x_1) f(x_2) \cdots f(x_n) (1 - F[t - (x_1 + x_2 + \cdots + x_n)]) dx_1 dx_2 \cdots dx_n \quad (6)$$

where  $R = \{x : 0 < x_1 + x_2 + \cdots + x_{n_1} < t_1 < x_1 + x_2 + \cdots + x_{n_1+1} < x_1 + x_2 + \cdots + x_n < t\}$ . (7)

Clearly the problem reduces to one involving the evaluation of an  $n$ -tuple integral. The development of an appropriate methodology to reduce this multiple integral to a simpler expression which is more amenable to analysis and which can be more useful in characterizing a renewal counting process appears in Sivazlian (1989).

#### e. The Prediction Problem in Renewal Processes

Traditionally, research in renewal theory has delved into either finding solutions to the time-dependent problem, or the derivation of limit theorems. The latter has constituted the majority of work in the more recent years. This is quite understandable since often the solution to a time-dependent problem is not easily obtainable. Although limit theorems may sometimes be used to arrive at desired solutions, nevertheless they are often inadequate for solving certain problems particularly for short term and medium term predictive purposes. It is evident that for this particular area of research many of the answers

would be obtained directly from the formulas expressing the probability law of the process, which provide the solution to the time-dependent problem.

Consider a system whose behavior is regulated by a renewal counting process. Let  $N(t)$  be the number of counts up to time  $t$ . Consider time epochs  $0 < t_1 < t_2 < t$ , and suppose that the system has been operating till time  $t_1$ . For predictive purposes, one may be interested in computing several probability expressions given that some type of information is available about the process at a given time  $t_1$ . These probability expressions provide mathematical models for predicting the statistical behavior of the process at some future time  $t_2 > t_1$ , given some level of knowledge concerning the state of the process at time  $t_1$ . Depending on the circumstances, we consider three cases:

Case 1: At time  $t_1$ , only the number of renewal counts  $N(t_1) = n_1$  is known;

Case 2: At time  $t_1$  the time at which the last renewal count has occurred, is known;

Case 3: At time  $t_1$ , no information is available

Case 1:

When at time  $t_1$  the number of renewal counts  $N(t_1) = n_1$  is known, the expressions of interest for predictive purpose would be the conditional probabilities:

$$P(N(t) = n \mid N(t_1) = n_1) \tag{8}$$

$$P(N(t+dt) = n \mid N(t_1) = n_1) \tag{9}$$

$$P(N(t_2)=n_2, N(t)=n \mid N(t_1)=n_1) \quad (10)$$

$$P(N(t)-N(t_1)=m \mid N(t_1)=n_1) \quad (11)$$

Note in particular that an expression for (9) would yield transition rates for the renewal counting process. Similarly, one may be interested in obtaining the covariance function  $\text{Cov} [N(t_1), N(t_2)]$ , the conditional expectation  $E[N(t)-N(t_1)=m \mid N(t_1)=n_1]$ , the joint distribution function of the backward and forward recurrence times conditional on  $N(t_1)=n_1$ , etc.

Case 2:

Suppose that at time  $t_1$ , it is known that the last renewal count has occurred at time  $r$ ,  $0 < r < t_1$ . Let  $U_1$  be the random variable defining the time elapsed from  $t_1$  until the next renewal occurs. It is evident that the probability density function of  $U_1$ ,  $g_{U_1}(u)$  is

$$g_{U_1}(u) = f(u) \int_{t_1-r}^{\infty} f(u) du, \quad t_1-r < u < \infty \quad (12)$$

Consider now the modified (or delayed) renewal counting process in which the arrival time till the first renewal is  $U_1$ , while the interarrival times for the next renewals is still given by the sequence of i.i.d. random variables  $(T_i)$ ,  $i = 2, 3, \dots$ . One thus is lead to study the probability law of a modified renewal process. Clearly, the same methodology used in Section II-1 would still be applicable here except that instead of using  $f_{X_1}(\cdot)$  in the final expressions obtained, one would substitute  $g_{U_1}(\cdot)$ . Time  $t_1$  would then be considered as the origin



of the process. Expressions for the conditional probability.

$$P(N(t_2 - t_1) = n_1, N(t - t_1) = n \mid \text{last renewal occurred at } \tau), \quad (13)$$

$$0 < \tau < t_1 < t_2 < t$$

or other probabilities could then be straightforwardly obtained.

Case 3:

In this case, no information is available about the state of the process at time  $t_1$ . This case is clearly of the same type as Case 2 except that the time till the first renewal is the forward recurrence time  $V(t_1)$ . Let

$H_V(t_1)$  - distribution function of  $V(t_1)$ ;

$M(y)$  - the renewal function for the original ordinary renewal counting process;

Then  $H_V(t_1)$  is given by the expression:

$$H_V(t_1)(x) = F(t_1 + x) - F(t_1) = \int_0^{t_1} [F(t_1 + x - y) - F(t_1 - y)] dM(y) \quad (14)$$

The method of analysis would be similar to Case 2.

3. A Birth Equation for the Ordinary Renewal Counting Process

Transition rates for the renewal counting process  $\{N(t), t \geq 0\}$  may be derived (Sivazlian, 1989). As a result, the process can be "viewed" as a non-homogeneous state-dependent birth type process. Thus, although  $\{N(t), t \geq 0\}$  does not in general, satisfy the Chapman-Kolmogorov equations, the unconditional distribution of  $N(t)$  satisfies the well known differential - difference equations of the birth type. It is

shown that indeed the solution of these equations yield the well-known results

$$\begin{aligned} P(N(t)=0) &= 1 - F(t) \\ P(N(t)=n) &= F^{(n)}(t) - F^{(n+1)}(t). \end{aligned} \tag{15}$$

Here we define

$$F^{(n)}(t) = \int_0^t f(x) F^{(n-1)}(t-x) dx, \quad n = 1, 2, \dots \tag{16}$$

$$\text{with } F(t) = F^{(0)}(t) = \int_0^t f(x) dx. \tag{17}$$

The results may be summarized as follows:

$$\begin{aligned} P(N(t_1+dt)=n+1 | N(t_1)=n) &= \frac{f^{*(n+1)}(t_1)}{F^{(n)}(t_1) - F^{(n+1)}(t_1)} dt + o(dt) \\ & \quad n=0, 1, 2, \dots \end{aligned} \tag{18}$$

$$\begin{aligned} P(N(t_1+dt)=n | N(t_1)=n) &= 1 - \frac{f^{*(n+1)}(t_1)}{F^{(n)}(t_1) - F^{(n+1)}(t_1)} dt + o(dt) \\ & \quad n=0, 1, 2, \dots \end{aligned} \tag{19}$$

$$P(N(t_1+dt)=n | N(t_1)=n_1) = o(dt) \quad n \geq n_1 + 2 \tag{20}$$

It is then evident that we can write for  $0 \leq m \leq n$ , expressions for:

$$\begin{aligned} P(N(t+dt)=n) &= \sum_m P(N(t+dt)=n, N(t)=m) \\ &= \sum_m P(N(t+dt)=n | N(t)=m) \cdot P(N(t)=m) \end{aligned} \tag{21}$$

Let  $P(n, t) = P(N(t) = n)$ . We find

$$\frac{dP(0, t)}{dt} = - \frac{f(t)}{1 - F(t)} P(0, t) \tag{22}$$

with the initial condition

$$P(0, 0) = 1 \tag{23}$$

Also

$$\frac{dP(n,t)}{dt} + \frac{f^{*(n+1)}(t)}{F^{(n)}(t) - F^{(n+1)}(t)} P(n,t) = \frac{f^{*(n)}(t)}{F^{(n-1)}(t) - F^{(n)}(t)} P(n-1,t) \quad (24)$$

with the initial condition

$$P(n,0) = 0 \quad (25)$$

The solution of these equations may be verified to yield (15).

### III. THE THEORY OF FILTERED RENEWAL PROCESS

#### 1. Introduction

The filtered renewal process is a stochastic process which is a generalization and a natural extension of the concept of the filtered Poisson process (in the sense of Parzen, 1962), in which the underlying process generator is modified to be a renewal counting process rather than a Poisson process. The filtered Poisson process (sometimes loosely called the compound Poisson process) is extensively discussed in Blanc-Lapierre et Fortet (1953), Parzen (1962) and Karlin and Taylor (1981). Filtered renewal processes provide models for a large variety of random phenomena in such areas as queueing theory, physics, economics, astrophysics, and population immigration. They can be regarded as arising by means of linear operations on a renewal process, in which additionally a response function must be specified. Many problems in simple and compound renewal processes, such as the renewal reward process or the cumulative process may be shown to be special cases of the filtered renewal process by judiciously selecting the form of the response function. Note that this response function will remain the same both for the filtered renewal process and the filtered Poisson

process. We now formally define the filtered renewal process following Parzen:

A stochastic process  $\{X(t), t \geq 0\}$  is said to be a filtered renewal process, if it can be represented, for  $t \geq 0$ , by

$$X(t) = \sum_{m=1}^{N(t)} w(t, W_m, Y_m) \quad , \quad 0 < W_1 < W_2 < \dots < W_{N(t)} < t \quad (26)$$

where

- i)  $\{N(t), t \geq 0\}$  is an ordinary renewal counting process with known i.i.d. interarrival time distribution  $\{T_i\}$ , and waiting times  $\{W_i\}$ ,

$$W_i = \sum_{j=1}^i T_j, \quad i = 1, 2, \dots;$$

- ii)  $\{Y_n\}$  is a sequence of i.i.d. random variables, and independent of  $\{N(t), t \geq 0\}$ , with distribution function  $G_Y(\cdot)$ ;  
 $n$

- iii)  $w(t, r, y)$  is a function of three variables called the response function.

For example, if  $W_m = r_m$  is the time at which an event took place, then  $Y_m = y$  represents the magnitude of a signal associated with the event,  $w(t, r_m, y)$  represents the value at time  $t$  of a signal of magnitude  $y$  originating at time  $r_m$ ,  $0 < r_m < t$ , and  $X(t)$  represents the value at time  $t$  of the sum of the signals arising from the events occurring in  $[0, t]$ .

The primary reason for being unable to extend the theory of filtered Poisson process to the filtered renewal process has been the unavailability of mathematical techniques to handle the complex multiple

integral expressions arising in obtaining, for example, the joint characteristic function of  $X(t_1), X(t_2), \dots, X(t_m)$ ,  $0 < t_1 < t_2 < \dots < t_m$  ( $m$  a positive integer).

The theory of filtered renewal processes may be studied by exploiting the properties of multiple integrals developed by the present author (1971), (1983), and extending the methodology to analyze a larger class of integrals, in order to obtain reduction formulas for evaluating these integrals. As a result, expressions for the various statistical characteristics of the process such as the probability law of the process, the joint characteristic function at distinct time epochs, the covariance function,  $E[X(t)]$ ,  $\text{Var}[X(t)]$  and limiting values (as  $t \rightarrow \infty$ ) could conceivably be obtained. One could also establish asymptotic normality of the filtered renewal process for given response functions.

## 2. Methodology

For the filtered renewal process  $(X(t), t \geq 0)$ , the characteristic function is given by  $E[e^{isX(t)}]$ . Using the definition of  $X(t)$ , this is given by

$$\begin{aligned}
 E[e^{isX(t)}] &= E\left[\exp\left(is \sum_{m=1}^{N(t)} w(t, W_m, Y_m)\right)\right] \\
 &= E\left[\exp\left(is \sum_{m=1}^{N(t)} w(t, \sum_{i=1}^m T_i, Y_m)\right)\right] \quad (27)
 \end{aligned}$$

$$= \int \int \dots \int \int_0^\infty \int_0^\infty \dots \int_0^\infty \prod_{m=1}^n \exp\left(is w(t, \sum_{i=1}^m x_i, y_m)\right)$$

$0 < x_1 + x_2 + \dots + x_n \leq t$

$$f(x_1) \cdot f(x_2) \cdot \dots \cdot f(x_n) \cdot F[t - (x_1 + x_2 + \dots + x_n)]$$

$$dx_1 dx_2 \dots dx_n \cdot dG(y_1) dG(y_2) \dots dG(y_n) \quad (28)$$

If we define

$$h(x_1+x_2 + \dots + x_i) = E_Y \left( \exp[isw(t, \sum_{j=1}^i x_j, Y_m)] \right) \\ = \int_0^\infty \exp[isw(t, \sum_{j=1}^i x_j, y)] dG(y) \quad (29)$$

Then clearly, the integral (17) has the form.

$$E[e^{isX(t)}] = \int \int \dots \int_{\substack{0 < x_1+x_2+\dots+x_n \leq t \\ \prod_{i=1}^n h(\sum_{j=1}^i x_j)}} \prod_{i=1}^n f(x_i) \cdot F(t - \sum_{i=1}^n x_i) \prod_{i=1}^n dx_i \quad (30)$$

Although there is a similarity between the multiple integrals (30) and (2), in general the integrand is not the same. It thus becomes necessary to use and extend the methodologies available (Sivazlian, 1971, 1983) to obtain new reduction formulas for evaluating integrals of the type (30). Again, here also, the reduced and simplified formulas thus derived, would be of particular significance in characterizing the filtered renewal process and in obtaining various statistical properties related to the process

#### IV. APPLICATIONS

With a better understanding of the theoretical basis of a process, a better appreciation is gained in characterizing the process as well as

in gaining insight into its various properties. A direct consequence is that the process may become more amenable to a wider variety of application, or may be used to better approximate the behavior of a system. Shortcomings in the theory invariably produce limitations in its applicability. It is hoped that the new results will open new vistas of applications by solving many complex problems in renewal theory and in filtered renewal theory.

Among some of the application areas we consider the following:

i. The Cumulative Damage Problem -

Consider a component subject to wear, where the number of wearout occurrences is regulated by a renewal process, and where in addition, the amount of wear is also regulated by a renewal process. Suppose that the component has been operating for some time. Given the present wearout state of the component, the problem of predicting the future wearout condition, that is the level of degradation of the component, as well as its ultimate failure (first passage time) may be addressed.

ii. The Takacs' Sojourn Problem for an Alternating Renewal Process -

This problem (see Takacs, 1957) has a variety of applications such as: 1) the cumulative damage problem previously described; 2) the problem of a component subject to failure and repair; 3) the traffic delay problem; and 4) other problems in statistics and operations research. It addresses itself to determining the cumulative effect of a certain condition (such as the total amount of repair time since time origin) for a system which can only be in two states rather than the total number of events related to that condition (such as the total number of repairs since time origin). Predictive models associated with this class of problem may be formulated.

iii. The Time-Dependent  $G/G/\infty$  Queueing System -

The variety of applications to the  $G/G/\infty$  queueing system are well established. This problem can be formulated as a filtered renewal process through a judicious choice of the response function, and the time-dependent characteristics of this system may be studied as well as its steady state behavior.

iv. A Production Problem -

Consider a manufacturing system which has unlimited capacity to produce an item. Assume that items arrive for production according to a renewal process and that the production times are i.i.d. random variables with known distribution function. One may be interested in determining for any give time  $t$  since production start the following:

- a) the number of items in production;
- b) the backlog of production time on the items which are in the production process;
- c) the probability of an empty production system.

Clearly, this is a variant of the  $G/G/\infty$  queueing system where the above quantities can be determined by the selection of an appropriate response function.

v. Other Problems -

Some of the other problems that may be formulated are:

- a) Predictive models for reliability systems such as systems with standby components;
- b) Predictive models for replacement and maintenance systems, including individual replacement, group replacement and age replacement;
- c) Predictive models for systems described by the superposition of



renewal processes;

vi. A Military Application -

An important potential area in applying renewal theory is the formulation of generalized stochastic Lanchester equations for models of military combats. It is well-known that these models are of the death or attrition type. The existing models could conceivably be extended based on the results presented in this paper.

For an ordinary renewal counting process  $\{N(t), t \geq 0\}$ , we have shown that a transition rate for the process can be generated which is both non-homogeneous and time-dependent taking the form

$$\lambda_n(t) = \frac{f^{*(n+1)}(t)}{F^{(n)}(t) - F^{(n+1)}(t)} \quad n = 0, 1, 2, \dots$$

As a result, the unconditional distribution of  $N(t)$ , namely,  $P(n, t) = P(N(t)=n)$  satisfies the birth equations

$$\frac{dP(0, t)}{dt} = -\lambda_0(t) P(0, t)$$

$$\frac{dP(n, t)}{dt} = -\lambda_n(t) P(n, t) + \lambda_{n-1}(t) P(n-1, t) \quad n=1, 2, \dots$$

with initial conditions given by  $P(0, 0)=1$  and  $P(n, 0)=0$ , otherwise.

For a death process involving an initial population size  $N$ , we have:

$$P(N, 0) = 1 \text{ and } P(n, 0) = 0 \quad n \neq N$$

We may write for example the following equations involving "linear" death rates  $n \lambda_n(t)$ :

$$\frac{dP(N, t)}{dt} = -N \lambda_N(t) P(N, t), \quad n = N$$

$$\frac{dP(n,t)}{dt} = -n \lambda_n(t)P(n,t) + (n-1) \lambda_{n+1}(t) P(n+1, t), \quad 1 \leq n \leq N-1$$

$$\frac{dP(0,t)}{dt} = \lambda_1(t) P(1,t), \quad n=0$$

This is a natural extension of existing death models. It is evident that this provides a framework of analysis for combat models by generalizing the Markovian models while retaining the structural properties for obtaining closed form solutions. The reader is referred to Sivazlian (1989) for a recent application of combat modeling.

#### V. CONCLUSIONS

The derivation of the probability law (joint distribution function) of a renewal counting process, and the analysis of the filtered renewal process, form the theoretical basis 1) to study the prediction problem associated with systems regulated by these processes, and 2) to apply the results to several useful problems in reliability, replacement, maintenance, queueing, production, combat analysis and other areas of operations research.

The theoretical knowledge acquired by this research should advance the level of knowledge in the statistical characterization of renewal processes and filtered renewal processes. Novel insight into the properties of these processes and a better understanding of their behavior should be gained. The theoretical results derived can be used immediately

1) to solve a number of stochastic problems in operations research and

- other applied areas;
- 2) to assess the performance characteristics of systems which can be modeled as a renewal process or as a filtered renewal process or as a process based on or related to the two previous ones;
  - 3) to develop models for predicting the behavior and effectiveness of systems represented by any of the above processes. Typical measures of effectiveness could include for example, system reliability, system maintainability and system availability.

On a long term basis, it is hoped that these processes may become, through an improved knowledge of their behavior, more amenable to a wider variety of applications.

#### VI. REFERENCES

- Arrow, K. J., S. Karlin and H. Scarf (1958), Studies in the Mathematical Theory of Inventory and Production, Stanford University Press, Stanford, California.
- Barlow, R. E., and F. Proschan (1965), Mathematical Theory of Reliability, J. Wiley, New York.
- Blanc-Lapierre et Fortet (1953), Théorie des Fonctions Aléatoires, Masson, Paris
- Cohen, J. W. (1982), The Single Server Queue, North-Holland, Amsterdam.
- Cox, D. R. (1962), Renewal Theory, Methuen & Co., Ltd., London.
- Daley, D. J. and D. Vere-Jones (1988), "An Introduction to the Theory of Point Processes", Springer-Verlag, New York.
- Feller, W. (1941), "On the Integral Equation of Renewal Theory", Ann. Math. Statist., 12, 243-67.
- Heyman, D. P. (1975), "A Diffusion Model Application for the GI/G/1 Queue in Heavy Traffic", The Bell System Technical Journal, 54, 1637-1646.

- Lotka, A. J. (1939), "A Contribution to the Theory of Self-Renewing Aggregates with Special Reference to Industrial Replacement", Ann. Math. Statist., 10, 1-25.
- Karlin, S. and H. M. Taylor (1981), A Second Course in Stochastic Processes, Academic Press, Inc., New York.
- Parzen, E. (1962), Stochastic Processes, Holden-Day, Inc., San Francisco.
- Mikusinski, J. (1959), Operational Calculus, Pergamon Press, Oxford.
- Ross, S. M. (1983), Stochastic Processes, J. Wiley, New York.
- Sivazlian, B. D. (1971), "A Class of Multiple Integrals", SIAM J. of Math. Anal., 12, 72-75.
- Sivazlian, B. D. (1983), "The Incomplete Liouville Multiple Integral and its Application", Am. Jour. of Math. and Manag. Sci., 3, 297-311.
- Sivazlian, B. D. (1989), "Aircraft Sortie Effectiveness Model" Naval Research Logistics, 36, 127-137.
- Sivazlian, B. D. (1989), "On the Joint Distribution of the Number of Renewals in a Renewal Process" Stochastic Analysis and Applications, (to appear).
- Smith, W. L. (1958), "Renewal Theory and its Ramifications", Jour. Royal Stat. Soc., B, 20, 243-302.
- Takacs, L. (1957), "On Certain Sojourn Time Problems in the Theory of Stochastic Processes" Acta Math. Acad. Sci. Hung., 8, 169-191.
- Wolff, R. W. (1988), Stochastic Modeling and the Theory of Queues, Prentice-Hall, Inc. Englewood Cliffs, NJ.

**STEPWISE CLOSED FORM TECHNIQUES  
FOR COMPUTER SIMULATION OF GUIDED PROJECTILES**

**M. J. Amoruso  
R. Campbell  
ARDEC**

**H. Cohen  
AMSAA**

**Sponsored by: U. S. Army Armament Research Development and Engineering Center  
Picatinny, New Jersey 07806-5000**

**The Army's Armament Research Development and Engineering Center (ARDEC) has been formulating methods for computationally efficient computer simulation for smart munitions. Time constants associated with autopilot components are often small compared with their driving terms. The integration time step is consequently driven to very small values to achieve stable numerical integration, which results increased computer run time. An innovative technique was developed in which exact analytic solutions to differential equations and transfer functions are applied in a piecewise manner within a larger but lower frequency problem that is solved numerically. Closed form analytical solutions were obtained for the following: the first-order lag, the first-order lag with differentiation, the first-order lead/lag, the first-order lag with integrator, the second-order lag/oscillator, a two-axis gimbaled gyro, and an impulse thruster. In addition to formulating piecewise analytic solutions to smart munition components, serial configurations of transfer functions should be replaced by equivalent parallel configurations. This approach avoids difficulties arising from propagation of the signal through a sequential network of widely varying natural frequencies when using a relatively large piecewise integration time step, and produces a decomposition that leads to terms that can be readily integrated analytically. In some cases, considerable time savings were obtained.**

## STEPWISE CLOSED FORM TECHNIQUES FOR COMPUTER SIMULATION OF GUIDED PROJECTILES

When modeling guided projectiles in 6 DOF (six degree of freedom) simulations, differential equations are obtained that describe the various component subsystems. These are then typically integrated numerically within the framework of the 6 DOF simulation. The largest allowable time step to perform the integration is bounded by two constraints. The driving term or input must not vary appreciably during a time step and the time step must be sufficiently small to insure a stable integration.

Since the driving term rates are commensurate with the airframe motion rates, inherently slower processes than those associated with the autopilot, stable integration is a lower bound to integration step size than the requirement for driving terms that remain essentially constant during the integration time step. By using analytic closed-form solutions for the differential equations for the autopilot, the second constraint appears to be eliminated. An innovative technique was developed by which exact analytical solutions to the required transfer functions are applied in a piecewise manner within a larger but lower frequency problem which must be solved numerically. The use of these piecewise analytical solutions to the transfer functions guarantees valid integration of the autopilot transfer functions regardless of the integration time step.

The overall system is usually analyzed into simpler terms in sequential order that are separately solved by numerical integration techniques, assuming that the input or driving term is essentially constant or linear during the integration time step. These factors are concatenated with the output of one factor or block becoming the input to the next block. Since the integration time step is generally quite small to achieve stable numerical integration, negligible errors are introduced as the signal propagates through the usually modest number of transfer function blocks down to the output.

The new approach consists in introducing analytical closed-form solutions for the transfer function factors that were previously treated numerically. Since these are exact closed-form solutions for constant (or linear) driving terms, the solutions bridge the time step perfectly as long as the driving term is essentially constant (or linear) during the time step. See Figure 1.

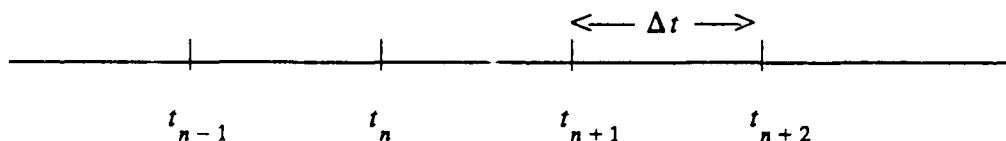


Figure 1. Iterative propagation of the solution

The result of the integration from the previous or  $n-1^{\text{th}}$  time step is used as the initial condition for the current time step. A value for the driving term during the current or  $n^{\text{th}}$  time step along with the initial condition are put into the analytic closed-form solution to propagate the solution from the beginning of the  $n^{\text{th}}$  time step to its end, where the resulting solution becomes the initial condition for the next or  $n+1^{\text{th}}$  time step, and so forth. This approach guarantees stable intergration contingent only upon the input remaining essentially constant or linear during the integration time step.

**Table 1**  
Typical Autopilot and Actuator Transfer Functions

TYPE	LAPLACE OPERATOR	DIFFERENTIAL EQUATION
First Order Lag	$\frac{1}{\tau s + 1}$	$\tau \frac{dy}{dt} + y = T$
First Order Lag with Differentiator	$\frac{s}{\tau s + 1}$	$\tau \frac{dy}{dt} + y = \frac{dT}{dt}$
First Order Lead/Lag	$\frac{\tau_1 s + 1}{\tau_2 s + 1}$	$\tau_2 \frac{dy}{dt} + y = \tau_1 \frac{dT}{dt} + T$
First Order Lag with Integrator	$\frac{1}{s[\tau s + 1]}$	$\tau \frac{d^2 y}{dt^2} + \frac{dy}{dt} = T$
Second order Lag/Oscillator	$\frac{1}{I s^2 + D s + K}$	$I \frac{d^2 y}{dt^2} + D \frac{dy}{dt} + K y = T$
Second order Lag/Oscillator with Differentiator	$\frac{s}{I s^2 + D s + K}$	$I \frac{d^2 y}{dt^2} + D \frac{dy}{dt} + K y = \frac{dT}{dt}$

Closed form solutions have been obtained for the typical transfer functions indicated in Table 1. Note that zero initial conditions are assumed for the

Laplace operators in this table. Non-zero initial conditions will be described below.

Savings in computer time vary from case to case with savings typically up to an order of magnitude. An impulsive thruster that consisted of rapidly burning material in a groove on the side of a spin stabilized projectile was modeled with great savings in execution time. The results in Tables 2 were obtained. Note that three different time steps were used for the numerical integration. With the coarsest time step ( $10^{-6}$  sec), agreement was to only 3 significant digits. For this case the analytical approach ran 23 times faster. Agreement between the analytical and numerical approaches could be increased by two more significant digits by decreasing the integration time step by an order of magnitude, with corresponding increase in run time for the numerical approach.

**Table 2**  
Impulsive Thruster Modeling

APPROACH	TIME STEP (sec)	CPU TIME (sec)	$\omega_y$ (rad)	$\omega_z$ (rad)
Analytical	N/A	0.0106	0.53275603	-0.41630908
Numerical	$10^{-8}$	22.096	0.53275601	-0.41630910
Numerical	$10^{-7}$	2.278	0.53275061	-0.41631581
Numerical	$10^{-6}$	0.246	0.53234719	-0.41681537

However, using this approach, an unexpected complication was discovered. If several factors are concatenated to represent a more complex transfer function, the final output can be found to depend on the order of the transfer function factors. This difficulty arises because, although the input of a block might be essentially constant during an integration time step, the output might not be if the frequency response of the block is relatively high. This output becomes the input to the next transfer function block. The requirement that the input to this next block be essentially constant can break down unless the sampling rate is high. This requires a smaller integration time step and longer computer execution time. This dilemma does not arise in the former numerical integration approach because the very fine time step required avoided difficulties associated with incompatibilities of bandwidth and frequency content as the signal propagated from block to block.



The solution adopted was the conversion of a complex transfer function to a parallel representation instead of a serial representation. The obvious advantage to an equivalent parallel representation is that each block receives the same input at the same time and each block produces its output at the end of the same single time step. These outputs do not become inputs to other autopilot transfer function blocks, but are instead summed to produce the overall output for the overall transfer function. Generally, this technique not only avoids simulation errors arising from time step size incompatibilities with bandwidth, internal lag, and frequency content of the signal propagating through the sequence of transfer function blocks but also has additional benefits. In addition to producing an algorithm that is considerably faster than previous numerical integration approaches, parallel decomposition generally leads to a combination of elementary expressions, whose Laplace inverse and analytical integration are well known.

The general treatment to implement this technique is outlined as follows:

- (1) Writing down the Laplace operator expression *including non-zero initial conditions* from the differential equation description or from the block diagram (which usually will not show the initial conditions)
- (2) Factoring
- (3) Making a formal partial fraction expansion
- (4) Finding the expansion coefficients
- (5) Writing down the expanded Laplace transform.

The latter can then be inverted from standard tables of inverse Laplace transforms to obtain the analytic solutions in parallel decomposition or calculated using the residue theorem of complex variables.

It is worthwhile to emphasize the first of the steps enumerated above. Autopilots, seekers, control actuators, and other components of guided projectiles are conventionally described in block diagrams in terms of the Laplace operator  $s$ . Typically, these block diagrams represent the underlying differential equations only if the initial conditions vanish. This is a convenient shorthand notation but can also be a source of confusion. Recall

that for initial conditions  $y_0 = y(t=0)$ ,  $\dot{y}_0 = \frac{dy}{dt}(t=0)$ , etc., the Laplace transform of the  $n^{\text{th}}$  derivative of a time domain function is given by

$$L[y^{(n)}] = s^n L[y(t)] - s^{n-1} y_0 - s^{n-2} \dot{y}_0 - \dots - y_0^{(n-1)} \quad (1)$$

where  $y^{(n)}$  represents the  $n^{\text{th}}$  derivative of  $y(t)$  with respect to  $t$ , and  $L$  and  $s$  represent the Laplace operator and variable respectively.

There are two methods for implementing this procedure. The first involves the taking of limits and derivatives and requires the factoring of the transfer functions into first-order systems. The second requires the solution of sets of simultaneous equations and does not involve limits or derivatives. Factoring the transfer function into first-order terms is optional in the second method. These techniques are well-known from partial fraction expansions of algebraic expressions.

A concrete illustration follows. Consider an autopilot component represented by the following block diagram in Figure 2. The differential equation corresponding to this block diagram is

$$\begin{aligned} \ddot{y}(t) + \dot{y}(t) - 2y(t) &= \text{Driving term} \\ &= K(t-t_0) + K_0 = Kt + L \end{aligned} \quad (2)$$

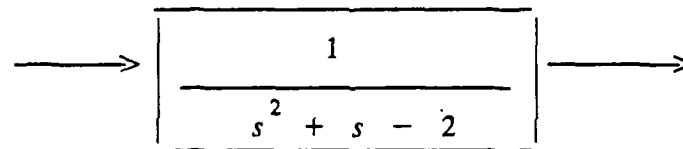


Figure 2. Example of block diagram

$K$ ,  $K_0$  and  $L$  are constants and  $t$  is time. This has the Laplace transform  $K/s^2 + L/s$ . Let  $\mathbf{L}\{y(t)\} = Y(s)$ . The Laplace transform of the differential equation is

$$(s^2 + s - 2)Y(s) - y_0(1 + s) - \dot{y}_0 = \frac{K + sL}{s^2} \quad (3)$$

where the additional terms are due to the initial conditions defined by  $y_0 = y|_{t=t_0}$  and  $\dot{y}_0 = \frac{dy}{dt}|_{t=t_0}$ . This may be written (factoring the denominator)

$$Y(s) = \frac{K + sL + s^2(y_0 + \dot{y}_0) + s^3 y_0}{s^2(s+2)(s-1)} \quad (4)$$

This may be formally expanded as follows:

$$Y(s) = \frac{A}{s^2} + \frac{B}{s} + \frac{C}{s+2} + \frac{D}{s-1} \quad (5)$$

If there are no multiple roots, the partial fraction expansion coefficients can be evaluated one at a time by taking the corresponding factors of the denominator of (5) one at a time and multiplying the right sides of (4) and (5) by that factor and then equating right sides. The resulting expression is taken to the limit as the factor goes to zero. This causes all expansion coefficients but one to drop out. Note that this technique fails when trying to find A or B because of the multiple root. The expression  $\lim_{s \rightarrow 0} sY(s)$  does not exist. If instead one tries  $\lim_{s \rightarrow 0} s^2 Y(s)$ ,  $A = -K/2$  is obtained. To obtain B, take  $\lim_{s \rightarrow 0} \frac{d}{ds} [s^2 Y(s)]$  and  $B = -(K+2L)/4$  results. In this way, using one factor at a time, all the expansion coefficients may be obtained.

The result of a partial fraction decomposition is

$$Y(s) = -\frac{K}{2s^2} - \frac{(K+2L)}{4s} + \frac{(4y_0 - 4\dot{y}_0 - K + 2L)}{12(s+2)} + \frac{(K+L+2y_0+\dot{y}_0)}{3(s-1)} \quad (6)$$

This transition to parallel decomposition or expansion is shown in Figure 3 for this simple case. It is a simple matter to invert this expression into the time domain.

$$y(t) = -\frac{Kt}{2} - \frac{(K+2L)}{4} + \left[ \frac{4y_0 - 4\dot{y}_0 - K + 2L}{12} \right] e^{-2t} + \left[ \frac{K+L+2y_0+\dot{y}_0}{3} \right] e^t \quad (7)$$

An alternate approach is algebraic. The first step in the algebraic approach is always the same as above. Factoring the denominator of the transfer function into monomials and expanding in terms of these monomials is done as before. Write the right side of the formal partial fraction decomposition with a least common denominator and equate to the Laplace expression for the autopilot transfer function.

$$Y(s) = \frac{K + sL + s^2(y_0 + \dot{y}_0) + s^3 y_0}{s^2(s^2 + s - 2)} \quad (8)$$

$$\begin{aligned}
&= \frac{A}{s^2} + \frac{B}{s} + \frac{C}{(s+2)} + \frac{D}{s-1} \\
&= \frac{A(s^2+s-2) + Bs(s^2+s-2) + Cs^2(s-1) + Ds^2(s+2)}{s^2(s^2+s-2)}
\end{aligned}$$

The denominators and numerators are equal. Making use of the linear independence of powers of  $s$ , a set of equations is obtained for the expansion coefficient which must be solved simultaneously. This yields the same result as before.

Alternatively, the investigator may wish to retain some higher order terms rather than reduce all the denominators to monomials, perhaps to retain a physical interpretation of terms.

In summary, by giving up the generality of numerical integration and using closed form solutions to particular differential equations in stepwise fashion, significant savings in computer run time can be obtained. Care must be taken when concatenating several such solutions in sequence. If the product of several sequential transfer functions can be recast into an equivalent network of transfer functions in parallel, difficulties arising from propagation of the signal through a sequential network can be eliminated even when relatively large integration time steps are used. This can be done by making a partial fraction decomposition of the Laplace operator representation. This technique produces a decomposition that leads to terms that can be readily integrated.

(a) Step 1

$$\longrightarrow \left[ \frac{K + sL + s^2(y_0 + \dot{y}_0) + s^3 y_0}{s^2(s^2 + s - 2)} \right] \longrightarrow$$

(b) Step 2

$$\longrightarrow \left[ \frac{K + sL + s^2(y_0 + \dot{y}_0) + s^3 y_0}{s^2(s + 2)(s - 1)} \right] \longrightarrow$$

(c) Steps 3 - 4

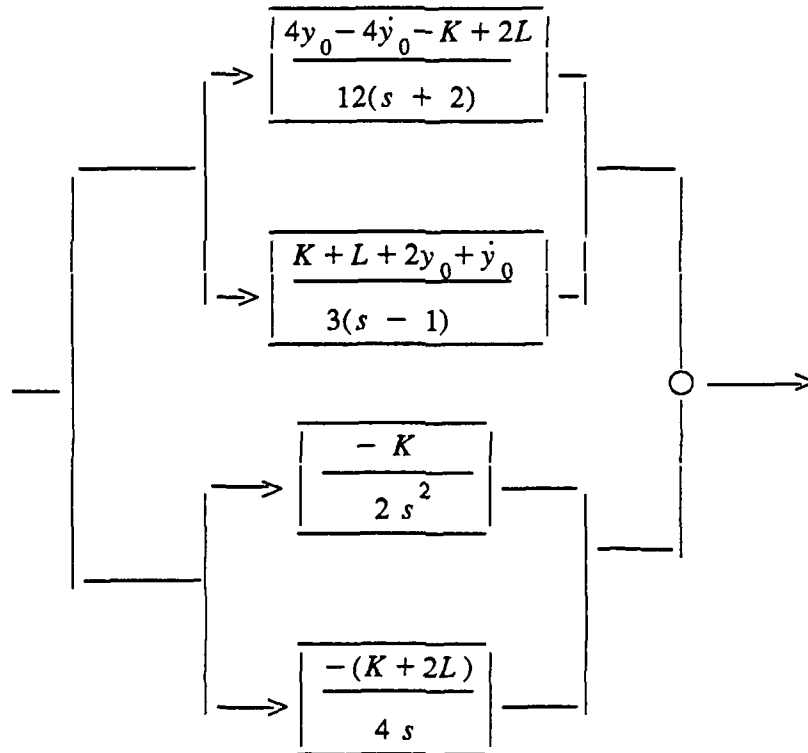


Figure 3. Example of transition to parallel representation

## Bibliography

E. M. Friedman and M. J. Amoruso, "An Analytical Modular Treatment of Autopilots for Guided Projectile Simulation," ARLCD-TR-85025, July 1985.

M. J. Amoruso, "An Analytical Technique for Modeling Gyroscopes in Guided Projectile Simulations," ARAED-TR-86010, June 1986.

E. M. Friedman, M. J. Amoruso, and R. Campbell, "An Analytical Model of an Impulsive Thruster," AD-E401 601, November 1986.

M. J. Amoruso, R. Campbell and E. M. Friedman, "Analytic Transfer Functions in Parallel Configurations," AD-E401 906, March 1989.

# GRÖBNER BASIS OPTIMIZATION

Moss Sweedler  
Department of Mathematics  
White Hall  
Cornell University  
Ithaca, NY 14853  
(607) 255-4373  
E-mail: JC5J@CORNELLA.CIT.CORNELL.EDU

Lee Taylor  
1838 Cedar Drive  
Severn, MD 21144  
(301) 551-5331  
E-mail: MQ9J@CORNELLC.CIT.CORNELL.EDU

**ABSTRACT:** A number of possible optimizations for Gröbner Basis construction are presented. Currently we are developing a system which permits easy experimentation with these and other optimizations. Perhaps as important as the hoped for optimizations, a conceptual framework is discussed, within which these and other possible optimizations are easily presented. The framework should be elaborated by others, as needed to support their own optimization experiments.

**INTRODUCTION:** Here are several related potential optimizations for finding Gröbner bases within  $k[\mathbf{X}] = k[X_1, \dots, X_n]$ , where  $k$  is a field. The main idea is to precipitate internal reduction by finding elements whose lead monomials divide other (lead) monomials. This obviates some S-pairs and may cut down storage requirements. We also use a generalization of: discarding the S-pair of  $f$  and  $g$  when the lead monomials of  $f$  and  $g$  are relatively prime. Our optimizations are supported by a novel approach to Gröbner basis construction. Typically, Gröbner basis construction is formulated in terms of two sets  $G$  and  $\mathbf{P}$ .  $G$  is the forming Gröbner basis and  $\mathbf{P}$  is a set of particular S-pairs from  $G \times G$  which remain to be reduced. If all of the S-pairs in  $\mathbf{P}$  reduce to zero over  $G$ , then  $G$  is a Gröbner basis. One traditional optimization question is avoiding unnecessary S-pair reductions. I.e. how to keep  $\mathbf{P}$  small. Unnecessary reductions waste computation. An unnecessarily large set  $\mathbf{P}$  is a waste of memory.

The underlying idea of our approach is formulated in terms of three sets,  $G$ ,  $J$  and  $\mathbf{P}$ .  $\mathbf{P}$  is a set of particular S-pairs from  $G \times G$ .  $(G, J, \mathbf{P})$  has the following property:

if all of the S-pairs in  $\mathbf{P}$  reduce to zero over  $G \cup J$  and if all of the  
 \* S-pairs from  $G \times J$  and  $J \times J$  reduce to zero over  $G \cup J$  then  $G \cup J$  is  
 a Gröbner basis

By dividing the forming Gröbner basis into two sets  $G$  and  $J$ , it is only necessary to **explicitly** keep track of S-pairs from  $G \times G$ . The S-pairs from  $G \times J$  and  $J \times J$  are known implicitly because  $G$  and  $J$  are known. This saves memory. Part of our approach is the notion of **allowable moves**. An allowable move might remove an element from  $\mathbf{P}$  at the expense of adding an element to  $J$ . Reduction of an element of  $\mathbf{P}$  over  $G \cup J$  is such an example. An allowable move might move an element from  $J$  to  $G$  and add elements to  $\mathbf{P}$ . An allowable move might move a number of elements from  $G$  to  $J$  and discard elements of  $\mathbf{P}$ . Allowable moves preserve (\*). The objective is to use allowable moves to reach a stage where  $\mathbf{P}$  and  $J$  are empty. When  $\mathbf{P}$  becomes empty there are no longer explicit S-pairs from  $G \times G$  which need to be reduced, and when  $J$  becomes empty there are no longer implicit S-pairs from  $G \times J$  and  $J \times J$  to reduce and  $G$  is a Gröbner basis by (\*).

$(G, J, \mathbf{P})$  satisfying (\*) is the underlying idea of our approach to optimization, but we have added two refinements:

$J$  is subdivided into two sets  $H$  and  $R$

two properties are added to (\*) which are preserved by the allowable moves

Our prospective optimizations are untested and other refinements of the underlying idea of  $(G, J, \mathbf{P})$  satisfying (\*) may prove to be more effective. We encourage researchers in the area of Gröbner basis optimization to experiment with their own refinements to  $(G, J, \mathbf{P})$  satisfying (\*).

At the present time we are developing a system to test these and related potential optimizations. The system MACAULAY by Bayer and Stillman is a Gröbner Basis based computer algebra system but is not designed for experimenting with variations on the fundamental algorithms. A selection of papers addressing Gröbner basis optimization can be found in the references.

**THE SETTING:** Suppose we have an implicit term ordering which allows us to form/find  $\text{LM}(f)$  — the lead monomial of  $f$ . Let:  $\text{MLCM}(f, g)$  —



the Monomial LCM of  $f$  and  $g$  — denote  $\text{LCM}(\text{LM}(f), \text{LM}(g))$ . Define the **NT** order on  $k[\mathbf{X}]$  in terms of the original implicit order, as follows: for  $f, g \in k[\mathbf{X}]$ ,  $f < g$  in the **NT** order if

$$\text{total degree } f < \text{total degree } g$$

or

$$\begin{aligned} &\text{total degree } f = \text{total degree } g \\ &\text{and } \text{LM}(f) > \text{LM}(g) \text{ in the original implicit order} \end{aligned}$$

Here is the promised generalization of: discarding the S-pair of  $f$  and  $g$  when the lead monomials of  $f$  and  $g$  are relatively prime. For  $p = (f, g) \in k[\mathbf{X}] \times k[\mathbf{X}]$ ,  $S(p) = S(f, g) =$  S-polynomial for the pair  $(f, g)$ , formed as follows:

Let  $m$  be the monomial  $\text{MLCM}(f, g)$ . Write  $f = q_1 m + r$  and  $g = q_2 m + s$ , where  $m$  does not divide any of the monomials of  $r$  or  $s$ . For  $S(f, g)$  use:

$$q_2 f - q_1 g = q_2 r - q_1 s$$

For  $\mathbf{P} \subset k[\mathbf{X}] \times k[\mathbf{X}]$ ,  $S(\mathbf{P}) = \{S(p) | p \in \mathbf{P}\}$ .

We assume there is an appropriate notion of reduction of a given element over a set. Typically this is repeated reduction of the lead term of the given element over the set or repeated reduction of all terms of the given element over the set. We are experimenting with both. Whichever notion of reduction is used, if a given element is fully reduced over a set, then the lead monomial of the given element must not be divisible by the lead monomial of any element of the set.

**THE FRAMEWORK:** The approach we are about to describe involves several stages of Gröbner basis construction. Frequently Gröbner basis construction is described in terms of one set — the forming Gröbner basis — and another set — the S-pairs which remain to be checked. The easiest way to describe our ideas is to split the forming Gröbner basis into three

parts and explicitly keep track of the remaining S-pairs for the first of the three sets.

A Gröbner frame<sup>1</sup> is the following:

Three sets  $G, H, R \subset k[\mathbf{X}]$  and a set  $\mathbf{P} \subset G \times G$ , where

1.  $G \cup H \cup R$  is a Gröbner basis if every element in

$$S(\mathbf{P}) \cup S((G \cup H \cup R) \times (G \cup H \cup R) \setminus G \times G)$$

reduces to zero over  $G \cup H \cup R$ .

2. For every  $(f, g) \in \mathbf{P}$ ,  $\text{LM}(f)$  and  $\text{LM}(g)$  are NOT relatively prime.

3. Every element of  $G \cup H \cup R$  is fully reduced with respect to  $G \cup H$ .

**EXAMPLE: GETTING STARTED.** Given a subset of  $k[\mathbf{X}]$ , call the set  $R$ . Let  $\mathbf{P}$ ,  $G$  and  $H$  be empty sets.

Given a Gröbner frame, the object is to use only the ALLOWABLE MOVES, described below, to decrease  $\mathbf{P}$ ,  $H$  and  $R$  to the empty set. When this is achieved,  $G$  is a Gröbner basis. Starting with three sets  $G, H, R \subset k[\mathbf{X}]$  and a set  $\mathbf{P} \subset G \times G$ , forming a Gröbner frame, they still form a Gröbner frame after an allowable move.

#### ALLOWABLE MOVES:

**MOVE P:** If  $\mathbf{P}$  is not empty, choose  $p \in \mathbf{P}$ , set  $\mathbf{P} = \mathbf{P} \setminus \{p\}$ , form  $S(p)$ , fully reduce  $S(p)$  over  $G \cup H \cup R$ . If the final reductum  $r$  is not zero, set  $R = R \cup \{r\}$ .

**MOVE H:** If  $H$  is not empty, choose  $h \in H$ , set  $H = H \setminus \{h\}$ . Let  $M$  be the ideal in the monoid of monomials generated by  $\text{MLCM}(g, h)$  as  $g$  runs over  $G$ . Let  $G' \subset G$  be chosen such that  $\{\text{MLCM}(g, h) | g \in G'\}$  generates  $M$ . When the  $\text{MLCM}(g, h)$ 's are being computed, make note of those

---

<sup>1</sup> One frame in the movie of the forming Gröbner basis.

$g$  such that  $\text{LM}(g)$  and  $\text{LM}(h)$  are relatively prime. Let  $G''$  be the elements of  $G'$  where  $\text{LM}(g)$  and  $\text{LM}(h)$  are not relatively prime. Set  $\mathbf{P} = \mathbf{P} \cup (G'' \times \{h\})$  and set  $G = G \cup \{h\}$ .

**MOVE R:** If  $R$  is not empty, choose  $s \in R$ , set  $R = R \setminus \{s\}$ . Let  $R'$  start as the empty set. For each  $m \in R$ , reduce  $m$  with respect to  $\{s\} \cup G \cup H$ , starting with  $\{s\}$ . If the final reductum  $r$  is not zero, set  $R' = R' \cup \{r\}$ . When done considering all  $m \in R$ , set  $R = R'$ . Let  $H'$  start as the empty set. For each  $h \in H$ , reduce  $h$  with respect to  $\{s\} \cup G \cup (H \setminus \{h\})$ , starting with  $\{s\}$ . If the final reductum  $r$  has the same lead monomial as  $h$ , set  $H' = H' \cup \{r\}$ . Otherwise, if  $r$  is non-zero, set  $R = R \cup \{r\}$ . When done considering all  $h \in H$ , set  $H = H' \cup \{s\}$ . Let  $G'$  start as the empty set. For each  $g \in G$ , reduce  $g$  with respect to  $\{s\} \cup (G \setminus \{g\}) \cup H$ , starting with  $\{s\}$ . If the final reductum  $r$  has the same lead monomial as  $g$ , then set  $G' = G' \cup \{r\}$ . Replace each  $(f, g) \in \mathbf{P}$  by  $(f, r)$ . Replace each  $(g, f) \in \mathbf{P}$  by  $(r, f)$ . Otherwise if  $r$  is non-zero but has lead monomial less than that of  $g$ , set  $R = R \cup \{r\}$ . Delete each  $(f, g) \in \mathbf{P}$  from  $\mathbf{P}$ . Delete each  $(g, f) \in \mathbf{P}$  from  $\mathbf{P}$ . When done considering all  $g \in G$ , set  $G = G'$ .

**MOVE's**  $R$ ,  $H$  and  $\mathbf{P}$  involve choosing elements from  $R$ ,  $H$  and  $\mathbf{P}$ . Among others, we are experimenting with the following priorities:

1. When  $R$  is not empty do **MOVE R**.
2. When  $R$  is empty but  $H$  is not empty, do **MOVE H**.
3. When  $R$  and  $H$  are empty, do **MOVE P**.

Our goal is to precipitate internal reduction by finding elements whose lead monomial divides many other lead monomials. These priorities are expected to precipitate internal reduction at relatively little computational expense. We elaborate on this theme when discussing **MOVE R**.

For **MOVE R**. Here one chooses an element  $r \in R$  and reduces elements of  $G \cup H \cup R$  over  $\{r\}$ . One could tentatively pick each element  $r \in R$  and count how many elements of  $G \cup H \cup R$  would reduce over that  $\{r\}$ . Then use the  $r$  which causes the most reduc-

tion as the actual choice. This would be computationally expensive. There might be something along this line using clever data structures and updating information cleverly which achieves this. At present we are investigating: choose the element of  $R$  whose lead monomial is as small as possible in the NT order. The hope is this will pick elements  $r \in R$  which tend to cause reduction.

For **MOVE H**. We have no particular best guess at this point.

For **MOVE P**. Ideally, one would choose  $(f, g) \in \mathbf{P}$  where the lead monomial of the final reductum of  $S(f, g)$  causes as much internal reduction as possible. Or, following the simplification for **MOVE R**, choose  $(f, g) \in \mathbf{P}$  where the lead monomial of the final reductum of  $S(f, g)$  is as small as possible in the NT order. Even this simplification seems much too computationally expensive. A further simplification is: choose  $(f, g) \in \mathbf{P}$  where  $\text{MLCM}(f, g)$  is as small as possible in the NT order.  $\text{MLCM}(f, g)$  gets computed (and can be saved) for each pair  $(f, g)$  when building up  $\mathbf{P}$ . Potential optimization aside,  $S(f, g)$  need only be computed when the pair  $(f, g)$  is selected in **MOVE P**. Pairs may be discarded from  $\mathbf{P}$  without ever computing (or reducing)  $S(f, g)$ . Thus the question: for purposes of optimization, how much work should be done with  $S(f, g)$  for pairs  $(f, g) \in \mathbf{P}$  which have not been selected in **MOVE P**?

When the minimal guidelines, based on small lead monomials in the NT order, together with our other prospective optimizations, are followed for small hand computations, we have been pleased with how few S-pairs ever are computed.

**REFERENCES:** The following is the current bibliography. The authors would welcome any additional references for future bibliographies.

Auzinger, W. and Stetter, H. (1988) An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations, International Series of Numerical Mathematics. 86, Birkhauser, Basel, 11-30.

Bajaj, C., Garrity, T. and Watten, J. (1988) On the applications of

- the multi-equational resultant, Technical Report CSD-TR-826, Purdue University.
- Bayer, D. (1982) The division algorithm and the Hilbert scheme, Ph.D. Thesis, Harvard University, Order number 82-22588, University Microfilms International, 300 N. Zeeb Rd. Ann Arbor, MI 48106.
- Bayer, D. and Stillman, M. (1986) 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 (B. Char ed.) 157-162.
- Bayer, D. and Stillman, M. (1987) Refining orders by the reverse lexicographic order, Duke Math. Journal, 55, 321-328.
- Buchberger, B. (1965) Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal, Ph.D. Thesis, University of Innsbruck.
- Buchberger, B. (1970) Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems, Aequationes Mathematicae 4 374-383.
- Buchberger, B. (1979) A criterion for detecting unnecessary reductions in the construction of Gröbner bases, Proc. EUROSAM'79, Marseille, Springer Verlag Lecture Notes in Computer Science, Vol. 72, pp. 3-21.
- Mayr, E. and Meyer, A. (1982) Complexity of the word problem for commutative semigroups and polynomial ideals. Adv. in Math. 46, 305-329.
- Schwartz, N. (1988) Stability of Grobner bases, J. Pure and Applied Algebra 53, 171-186.
- Sturmfels, B. (1989) Dynamic versions of the Buchberger algorithm. preprint.
- Weispfennig, V. (1989) Constructing universal Grobner bases. Technical Report, MIP 8901, Universität Passau.
- Yap, C. (1988) A new lower bound construction for the word problem for commutative Thue systems, Manuscript.

# The Relationship Between Linear and Nonlinear Variational Models of Coherent Phase Transitions\*

*Robert V. Kohn*

Courant Institute  
251 Mercer Street  
New York, NY 10012

## ABSTRACT

Variational models of phase transitions seek to explain the observed fine scale structures of phase mixtures as being due to bulk energy minimization. One theory of this type, geometrically linear in character, has been developed in the metallurgical literature by Khachaturyan and others. A second, apparently different, geometrically nonlinear theory has been developed in the mathematical literature by Ball, James, and others. We show that Khachaturyan's theory is roughly the linearization of Ball and James' approach. We also discuss how Khachaturyan's method permits the explicit relaxation of certain double-well energies in the linear setting. The corresponding calculation for triple-well energies remains incomplete.

## 1. Introduction

Coherent mixtures of crystalline solids have long been studied using elasticity. The metallurgical literature has primarily been based on linear theory, see e.g. [25,26,31-33,41,45,48,51-56,60,61]. Recent mathematical work, on the other hand, has taken a geometrically nonlinear viewpoint, see e.g. [4-6,8,11-15,18-22,27-30,34]. There is naturally a connection between these two approaches, and indeed the link is much stronger than has heretofore been recognized. The purpose of this article is to explore that connection in some detail, focussing particularly on the nonlinear variational model of Ball and James [4,27] and on work done in the linearized context by Roitburd, Khachaturyan, and Shatalov

\*Supported by ARO contract DAAL03-89-K-0039, DARPA contract F49620-87-C-0065, ONR grant N00014-88-K-0279 and NSF grant DMS-8701895

[31,33,51].

Coherent phase mixtures arise, for example, from martensitic phase transitions and in the early stages of decomposition processes. They have rather characteristic fine scale structures, often involving laminar arrangements of phases or distributions of like-shaped inclusions. A central goal of the variational theory is to explain the origin of these microstructures. A comprehensive survey is beyond the scope of this article, due to the vastness of the literature and also the limitations of the author's expertise. Nevertheless, we attempt a brief introduction.

It is precisely the condition of coherence that permits an analysis based on elasticity. Briefly, this condition assures that the atoms' actual positions are related to their locations in a reference lattice by a continuous elastic deformation. More detailed discussions of this point will be found in [42-44], where the central notion is the "network constraint," and in [11,12], where the discussion is based on the "Born rule."

One well-known approach in the linear context has its origins in the work of Eshelby [16,17], which gives the elastic field due to an elliptical inclusion of one phase in an otherwise uniform second phase. This leads to an approximate formula for the elastic energy of a multi-phase mixture, either through a mean field theory or by taking a dilute distribution of ellipsoids as the phase geometry. One can try to predict the inclusion shape by minimizing this approximate energy, see e.g. [45,56]. Here we will not deal with such approximate theories, but rather with calculations that are (mathematically, if not physically) exact.

A different approach was developed independently by Khachaturyan [31] and Roitburd [51] for the two-component case, and subsequently generalized to more than two components by Khachaturyan and Shatalov [33]. Their theory is geometrically exact, in the sense that it makes no hypotheses about the phase geometry, and it computes the elastic fields exactly. One pays a price for this generality, however: their method requires that the phases all have the same elastic moduli. For mixtures of two such phases, this work gives a formula for the extremal elastic energy as a function of the stress-free strains, the volume fraction of each phase, and the elastic moduli [32]. It shows moreover that this extremal energy is always achieved by a layered microstructure. Subsequent work has noted that other microgeometries may also be extremal, depending on the symmetry of the individual phases. For three or more phases the analysis is less complete: the treatment of

[31-33] does not yield a formula for the extremal energy of a general three-phase mixture.

The preceding work is all linear in character. Recently, a number of authors have explored models based on nonlinear elasticity. One approach leads to an elastic energy density with infinitely many local minima [14,15]; the associated "relaxed" energy is unfortunately rather degenerate [6,18-20,22]. We prefer the viewpoint of Ball and James [4,5,27], in which the elastic energy density has one "well" for each phase or phase variant. The relaxation of such an energy has yet to be computed; rather, attention has been focussed on determining where it achieves its minimum value. This is sufficient for studying transformations that do not involve internal stress, a class which includes many martensitic phase transitions.

The linear theory of Khachaturyan *et al.* and the nonlinear one of Ball and James are superficially quite different: the former deals with phase microgeometry directly, while the latter lets it enter through the structure of energy minimizing sequences. Both approaches involve the minimization of an elastic energy, however, and each provides a rationale for the more phenomenological "crystallographic theory of martensitic" [7,59]. Thus it is natural to look for a relationship between them.

The connection is in fact quite close. Roughly speaking, Khachaturyan's calculation is equivalent to the relaxation of a linearized version of the energy studied by Ball and James. Our goal is to explain this relationship, and hopefully to bridge the language barrier which currently separates the two theories.

We begin, in Section 2, with the linearization of the Ball-James theory. The linear analogue of their energy turns out to be a minimum of paraboloids, differing as to their shape, height, and the locations of their minima. Each paraboloid is the graph of the linearly elastic energy for a separate phase or phase variant.

In Section 3 we explain what one means by the "relaxed" or "macroscopic" energy  $QW$  associated to a given energy density  $W$ . The functions that emerge from Section 2 have the special form  $W = \min \{W^1, \dots, W^N\}$ . For such functions we introduce a new concept, the "relaxation at fixed volume fraction"  $Q_\theta W$ . It gives the macroscopic energy of the system when both the average strain and the volume fractions of the phases are fixed. The standard relaxation  $QW$  is just  $\inf_\theta Q_\theta W$ , in which  $\theta$  varies over all possible volume



fractions (see Proposition 3.1); thus knowledge of  $Q_\theta W$  for all  $\theta$  effectively determines  $QW$ .

For a system consisting of two linear phases with the same elastic moduli, the calculation of Roitburd and Khachaturyan amounts to the determination of  $Q_\theta W$ . This is the central link between the linear and nonlinear viewpoints, and it is discussed in Section 4. A more complete discussion concerning the relaxation of double-well energy functions will be found in [36].

The corresponding analysis for three or more linear phases is presented in Section 5. Our approach is essentially the same as that of Khachaturyan and Shatalov [33], and we get no further than they do. In selected cases, when the stress-free strains are related appropriately, one can determine the minimum value of  $Q_\theta W$ ; this is the linear analogue of work by Ball and James [4]. In general, however, the calculation of  $Q_\theta W(\xi)$  remains open. The new notion of H-measures, recently introduced independently by Tartar [58] and Gerard [24], may be useful in this context: as we shall explain, calculating  $Q_\theta W$  is equivalent to minimizing a certain functional over the H-measures associated to certain characteristic functions.

In its use of Fourier analysis, Khachaturyan's calculation of  $Q_\theta W$  bears a strong resemblance to recent work in homogenization [2,38,50]; thus we are in essence using homogenization to compute the relaxations of certain energy integrands. This link between relaxation and homogenization has in fact been noted before [39,40]. The main difference between those discussions and the present one is that here the phases have different stress-free strains and the same elastic moduli.

Our attention is concentrated entirely on the minimization of bulk energy; one expects such an analysis to be qualitatively correct if the effects of surface energy are sufficiently small. The latter are presumably important for determining the length scale and periodicity of the microstructure; they are also thought to be the reason for the appearance of plate-like inclusions when the theory predicts fine-scale layering. Formal treatments of these effects will be found in [4,32]; there has been little rigorous analysis, but see [21,35] for some preliminary steps in that direction. In addition, surface energy is presumably responsible for selecting between distinct microstructures with the same bulk energy (see Remark 4.4).

Taken together, Proposition 3.1 and Theorem 4.1 determine the relaxation of a "two-well" energy describing a system of two linearly elastic phases with the same elastic moduli. For the special cases of an isotropic elastic law in two space dimensions, that relaxation was previously computed by Lurie and Cherkaev [47]. Their analysis is quite different from ours, being based on the method of "polyconvexification" rather than Fourier analysis.

The hypothesis of equal elastic moduli is apparently a good approximation for many two-phase systems: microstructures consistent with Khachaturyan's theory are observed in a wide variety of systems [32]. However, this approximation is certainly not always valid. We have recently extended Khachaturyan's calculation to the case of two phases with different elastic moduli, provided that the elasticity tensors are in a certain sense well-ordered [37]. This extension is based on the Hashin-Shtrikman variational principle; it is similar to [2,38,50] except that the phases have different stress-free strains.

## 2. Linearization of the Ball-James Theory

Ball and James have developed a model for martensitic phase transitions based on finite elasticity [4]. Their idea is to minimize a non-elliptic energy function which has a separate "well" for each phase or phase variant. We focus on the case of a cubic-tetragonal phase transition such as that of InTl, following [4].

The elastic energy has the form

$$E[u] = \int_{\Omega} W_T(\nabla u) dx, \quad (2.1)$$

where  $\Omega \subset \mathbb{R}^3$  is a reference domain,  $u: \Omega \rightarrow \mathbb{R}^3$  is an elastic deformation, and  $T$  represents temperature. The temperature could vary from point to point, i.e.  $T = T(x)$ ; but it is considered given, not to be varied in the minimization of  $E$ . For simplicity, we shall assume for the duration of this discussion that  $T$  is constant. The energy density satisfies the condition of frame indifference:

$$W_T(RF) = W_T(F) \quad \text{for } R \in SO(3). \quad (2.2)$$

( $SO(3)$  is the group of orientation preserving rotations of  $\mathbb{R}^3$ .) As a consequence, the local minima of  $W_T$  must occur on orbits of  $SO(3)$ .

For a cubic-tetragonal phase transition,  $W_T$  has four "wells," one corresponding to the austenite (cubic) phase and the others to the three symmetry-related variants of martensite (the tetragonal phase). There is an exchange of stability at the transformation temperature  $T_c$ : for  $T > T_c$  the absolute minimum is in the austenite well, whereas for  $T < T_c$  it is in the martensite wells. We take austenite at  $T = T_c$  as the reference configuration, with spatial axes aligned with the axes of symmetry. Then the minima of  $W_T$  at  $T = T_c$  are the orbits  $\{R\}$ ,  $\{RA_1\}$ ,  $\{RA_2\}$ , and  $\{RA_3\}$  where  $R$  ranges over  $SO(3)$  and

$$A_1 = \begin{pmatrix} 1 + \eta & 0 & 0 \\ 0 & 1 - \delta & 0 \\ 0 & 0 & 1 - \delta \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 - \delta & 0 & 0 \\ 0 & 1 + \eta & 0 \\ 0 & 0 & 1 - \delta \end{pmatrix} \quad A_3 = \begin{pmatrix} 1 - \delta & 0 & 0 \\ 0 & 1 - \delta & 0 \\ 0 & 0 & 1 + \eta \end{pmatrix} \quad (2.3)$$

are the three symmetry-related transformation strains. According to [4],  $\eta \approx .026$  and  $\delta \approx .013$  for InTl.

The main quantities one can measure are the transformation strains and the linear elastic moduli of each phase. These determine the locations of the relative minima of  $W_T$  and its behavior near those minima. One might also impose the condition that  $W(F) \rightarrow \infty$  as  $\det F \rightarrow 0$ , but basically the form of  $W_T$  is open when  $F$  is far from a natural state. One approach is to use a simple polynomial function for  $W_T$ , see e.g. [8,13]. However, we prefer to view each phase as having "its own" energy function, with  $W_T$  being the minimum of the lot:

$$W_T(F) = \min\{W_T^j(F), j = 0, 1, 2, 3\}. \quad (2.4)$$

Here  $W_T^0$  corresponds to austenite, and  $W_T^i, i = 1, 2, 3$ , to martensite. At  $T = T_c$ ,  $W_T^0$  is minimized at the orbit of the identity and  $W_T^i$  at that of  $A_i$ . Since the martensite phases are symmetry-related, their energies satisfy

$$W_T^2(F) = W_T^1(FR_{12}), \quad W_T^3(F) = W_T^1(FR_{13}) \quad (2.5)$$

for some rotations  $R_{12}$  and  $R_{13}$  in the symmetry group of the cube but not in that of the tetragon. It follows from (2.5) that  $R_{12}^T A_2 R_{12} = A_1 = R_{13}^T A_3 R_{13}$ .

By frame indifference, each of these energies should depend only on  $(F^T F)^{1/2}$ . It is a natural approximation to assume that they are quadratic functions of  $(F^T F)^{1/2}$ . If we suppose for simplicity that the stress-free strains and elastic moduli are independent of  $T$ , this

leads to

$$W_T^0(F) = \langle \alpha_0[(F^T F)^{1/2} - I], (F^T F)^{1/2} - I \rangle + \Phi_a(T)$$

for the austenite, and

$$W_T^i(F) = \langle \alpha_i[(F^T F)^{1/2} - A_i], (F^T F)^{1/2} - A_i \rangle + \Phi_m(T)$$

for the martensite. Here  $\Phi_a(T)$  and  $\Phi_m(T)$  are the energies of the stress-free states as a function of temperature;  $\alpha_j (j = 0, \dots, 3)$  is a symmetric linear map acting on symmetric tensors; and  $\langle A, B \rangle = \text{Tr}(AB)$  is the standard inner product of symmetric tensors. We shall see presently that  $\alpha_j$  is precisely the Hooke's law of the  $j$ th phase. The symmetry relations (2.5) imply that  $\alpha_2$  and  $\alpha_3$  arise from  $\alpha_1$  through the action of  $R_{12}$  and  $R_{13}$  respectively, acting in the usual way on symmetric tensors; for example,

$$\langle \alpha_1 R_{12}^T \xi R_{12}, R_{12}^T \xi R_{12} \rangle = \langle \alpha_2 \xi, \xi \rangle$$

for every symmetric tensor  $\xi$ .

Now let us linearize  $W_T$ . This is done by taking

$$F = I + \epsilon f, \quad A_i = I + \epsilon a_i,$$

$$\Phi_m(T) = \epsilon^2 \phi_m(T), \quad \Phi_a(T) = \epsilon^2 \phi_a(T),$$

and expanding to principal order in  $\epsilon$ . Since  $(F^T F)^{1/2} = I + \frac{\epsilon}{2}(f + f^T) + O(\epsilon^2)$ , one easily obtains

$$W_T^0(F) = \epsilon^2 \left\{ \left\langle \alpha_0 \left( \frac{f + f^T}{2} \right), \frac{f + f^T}{2} \right\rangle + \phi_a(T) \right\} + O(\epsilon^3)$$

$$W_T^i(F) = \epsilon^2 \left\{ \left\langle \alpha_i \left( \frac{f + f^T}{2} - a_i \right), \frac{f + f^T}{2} - a_i \right\rangle + \phi_m(T) \right\} + O(\epsilon^3).$$

Writing  $\xi = \frac{1}{2}(f + f^T)$  for the linear strain tensor, we see that the linearization of (2.4) is

$$\bar{W}_T(\xi) = \min \left\{ \bar{W}_T^j(\xi), j = 0, 1, 2, 3 \right\} \quad (2.6)$$

with

$$\bar{W}_T^0(\xi) = \langle \alpha_0 \xi, \xi \rangle + \phi_a(T)$$

$$\bar{W}_T^i(\xi) = \langle \alpha_i(\xi - a_i), \xi - a_i \rangle + \phi_m(T), i = 1, 2, 3.$$

Thus the graph of the linearized energy  $\bar{W}_T$  is the minimum of a family of paraboloids, each having its vertex at a different linear strain.

The preceding argument is easily extended to allow the transformation strains and Hooke's law to depend on temperature, and a similar analysis can clearly be done for other types of phase transitions. Notice that the linearization process requires  $A_i - I$  as well as  $F - I$  to be small; thus it is only reasonable to use a linearized theory when the lattice parameters of the phases are close to one another.

The linearization just performed is of course only formal. It is not at all clear that the original elastic energy (2.1) behaves like its linearized analogue (2.6). However, the known results are indicative of a very strong connection. For example, it is conjectured in the nonlinear context that if a Young-measure limit of gradients is supported on two wells, then those wells must be rank-one related [30]. We shall prove a very similar result for the linearized setting in Section 4 (see Proposition 4.4).

### 3. Relaxation of multiple-well energy functions

This section explains the idea of relaxation in the context of linear elasticity; the central notion is the **quasiconvexification**  $QW$  associated to an energy density  $W$ . For energies of the form  $W = \min \{W^1, \dots, W^N\}$  we introduce the related notion of the **quasiconvexification at fixed volume fractions**,  $Q_\theta W$ .

To explain why relaxation is of interest, consider a system of two linearly elastic phases with Hooke's laws  $\alpha_i$  and stress-free strains  $a_i$ ,  $i = 1, 2$ :

$$W_T(\xi) = \min\{W_T^1(\xi), W_T^2(\xi)\},$$

$$W_T^i(\xi) = \langle \alpha_i(\xi - a_i), \xi - a_i \rangle + \phi_i(T). \quad (3.1)$$

Here  $\xi$  is the linear elastic strain,  $T$  is temperature, and  $\phi_i(T)$  is the minimum energy of the  $i$ th phase at temperature  $T$ . We assume that the phases exchange stability at  $T = T_c$ , say  $\phi_1 < \phi_2$  for  $T < T_c$  and  $\phi_1 > \phi_2$  for  $T > T_c$ .

Suppose that such a system is held in a variable temperature field  $T = T(x)$ , with no body loads or surface tractions. The elastic energy is then

$$E[u] = \int_{\Omega} W_T(x) (e(u)) dx, \quad (3.2)$$

with  $e(u) = \frac{1}{2}(\nabla u + \nabla u^t)$ . At first glance the minimization of  $E$  might seem trivial: phase 1 is preferred for  $T < T_c$  and phase 2 for  $T > T_c$ , so it is tempting to look for a solution of the form

$$e(u) = \begin{cases} a_1 & T < T_c \\ a_2 & T > T_c \end{cases}$$

This does not work, however: for such a deformation to exist,  $a_1 - a_2$  must here the form  $n \otimes m + m \otimes n$ , and the surface  $\{T = T_c\}$  must consist of hyperplanes normal to either  $n$  or  $m$ . A second idea would be to consider the optimality conditions for (3.2); but this, too, is ill-conceived, since it is based on the assumption that a solution exists. In fact, it is not clear that the minimum of (3.2) is achieved; rather, it is possible for a minimizing sequence to develop oscillatory spatial gradients. Physically, this arises because a fine-scale mixture of both phases may lead to a lower energy than either pure phase. (We view the set where  $W_T^1(e(u)) < W_T^2(e(u))$  as being occupied by phase 1, and its complement by phase 2.) Indeed,  $E[u]$  is most interesting when its minimum is not achieved: that is the case when energy minimization requires a mixture of the two phases. The minimizing sequences for  $E[u]$  determine the preferred microstructures for phase mixtures.

The technique of relaxation is, in essence, a method for constructing minimizing sequences of nonconvex variational problems such as (3.2). The relaxed problem has a similar form, with  $W_T$  replaced by its "quasiconvexification"  $QW_T$ :

$$\int_{\Omega} QW_T(x) (e(u)) dx. \quad (3.3)$$

Though the relaxed problem need not be convex, its minimum is achieved; indeed, its minimizers are precisely the weak limits of minimizing sequences of the original problem. The relaxed integrand is defined, for each  $T$ , by

$$QW_T(\xi) = \inf_{v|_{\partial U} = \xi \cdot x} \frac{1}{|U|} \int_U W_T(e(v)) dx. \quad (3.4)$$

We need not specify the domain  $U$ , since the value of the infimum is the same for all domains with reasonably regular boundaries [3,9,10].

The introduction of the relaxed energy is physically quite natural. We think of  $W_T$  as the "microscopic" energy function, and view  $QW_T$  as an associated "macroscopic" energy: it gives, in essence, the minimum average energy when the average strain is  $\xi$ . Given knowledge of the minimizing sequences for (3.4) and a minimizer of (3.3), one easily constructs a minimizing sequence for the original energy  $E$ : this is done, roughly speaking, by superimposing the oscillations prescribed by (3.4) upon the slowly-varying strain of the solution to (3.3). In particular, the preferred phase microstructures associated to a given strain  $\xi$  at temperature  $T$  are determined by minimizing sequences for (3.4). We refer to Section 2 of [40] for an expository discussion of the basic facts about relaxation, and to [1,3,9,10] for more comprehensive treatments including proofs. (These references discuss functions of  $\nabla u$  rather than functions of  $e(u) = \frac{1}{2}(\nabla u + \nabla u')$ . But they require no coercivity hypotheses, so the results apply *a priori* in the context of linear elasticity.)

The definition of  $QW_T$ , (3.4), applies generally, whatever the form of the energy  $W_T$ . From Section 2, however, we see that for modelling phase transitions it is natural to consider energies of the special form

$$W(\xi) = \min\{W^1(\xi), \dots, W^N(\xi)\}. \quad (3.5)$$

(We suppress the parameter  $T$  for simplicity of notation; the particular form of  $W^i$  will not matter for what follows). For such  $W$ , we now define the quasiconvexification with fixed volume fractions,  $Q_\theta W$ .

Let  $\mathcal{V}$  denote the set of all possible volume fractions:

$$\mathcal{V} = \{\theta = (\theta_1, \dots, \theta_N): \theta_i \geq 0, \sum \theta_i = 1\}.$$

Fixing a region  $U$  of  $\mathbb{R}^3$ , we say a partition  $U = U_1 \cup \dots \cup U_N$  has volume fraction  $\theta$  if  $|U_j| = |U| \cdot \theta_j$ ,  $1 \leq j \leq N$ . It is convenient to represent such a partition by its "marker functions"

$$\chi_j(x) = \begin{cases} 1 & x \in U_j \\ 0 & \text{otherwise} \end{cases};$$

note that  $\chi_i \chi_j = \delta_{ij}$ ,  $\sum \chi_j = 1$ , and

$$\frac{1}{|U|} \int_U \chi_j = \theta_j, \quad 1 \leq j \leq N.$$

For  $\theta \in \mathcal{V}$ , we set

$$Q_\theta W(\xi) = \inf_{\{\chi_j\}} \inf_{v|_{\partial U} = \xi \cdot x} \frac{1}{|U|} \int_U \sum \chi_j W^j(e(v)) dx, \quad (3.6)$$

where  $\{\chi_j\}_{j=1}^N$  ranges over marker functions associated to partitions of  $U$  with volume fraction  $\theta$ . As was the case for (3.4), we need not specify the domain  $U$ : the value of the infimum is independent of  $U$ . (The proof is parallel to that for  $QW$ , see for example Proposition 2.3 of [3].)

The following proposition asserts that if  $Q_\theta W$  is known for every  $\theta \in \mathcal{V}$ , then the determination of  $QW$  requires only a finite-dimensional optimization.

**Proposition 3.1:**

$$QW(\xi) = \inf_{\theta \in \mathcal{V}} Q_\theta W(\xi).$$

**Proof:** Clearly

$$\inf_{\theta \in \mathcal{V}} Q_\theta W(\xi) = \inf_{\{\chi_j\}} \inf_{v|_{\partial U} = \xi \cdot x} \frac{1}{|U|} \int_U \sum \chi_j W^j(e(v)), \quad (3.7)$$

where  $\{\chi_j\}$  now range over the marker functions associated to all partitions, regardless of volume fraction. The right side of (3.7) is not altered if we interchange the order of the two minimizations. But for fixed  $v$ ,

$$\inf_{\{\chi_j\}} \frac{1}{|U|} \int_U \sum \chi_j W^j(e(v)) = \frac{1}{|U|} \int_U W(e(v))$$

with  $W$  given by (3.5): the optimal partition has

$$\chi_j = 1 \text{ where } W^j(e(v)) = \min_{1 \leq i \leq N} \{W^i(e(v))\}.$$

Thus

$$\inf_{\theta \in \mathcal{V}} Q_\theta W(\xi) = \inf_{v|_{\partial U} = \xi \cdot x} \frac{1}{|U|} \int_U W(e(v)) = QW(\xi).$$

■

We require one more lemma concerning the general notion of quasiconvexification. The integrands  $QW$  and  $Q_\theta W$  were defined above using the Dirichlet boundary condition  $v|_{\partial U} = \xi \cdot x$  on an arbitrary domain  $U$  (see (3.4) and (3.6)). However there is an



equivalent characterization involving the averaging of periodic functions. This will be convenient in Section 4, where we will use Fourier analysis to calculate  $Q_0 W$  for certain two-well energies. We choose  $C = [0, 2\pi]$  the unit cell in  $\mathbb{R}^3$ ;  $\bar{f}$  denotes the average value of a  $C$ -periodic function  $f$ .

**Lemma 3.2:** The quasiconvexification has the alternate characterization

$$QW(\xi) = \inf_{\phi \text{ per}} \bar{f} W(\xi + e(\phi)) dx, \quad (3.8)$$

in which  $\phi$  ranges over all  $C$ -periodic maps from  $\mathbb{R}^3$  to  $\mathbb{R}^3$ . Similarly,

$$Q_0 W(\xi) = \inf_{\chi_j = \theta_j} \inf_{\phi \text{ per}} \bar{f} \sum \chi_j W^j(\xi + e(\phi)) dx, \quad (3.9)$$

where  $\{\chi_j\}$  range over periodic marker functions associated with partitions of  $C$ , and  $\phi$  ranges over  $C$ -periodic maps from  $\mathbb{R}^3$  to  $\mathbb{R}^3$ .

**Proof (sketch):** We assume that  $W$  is continuous with  $L^p$  growth. It is easy to see that (3.8)  $\leq$  (3.4): we may take  $U = C$  in (3.4) and write  $v = \xi + \phi$  with  $\phi|_{\partial C} = 0$ , then extend  $\phi$  periodically to get an admissible test field for (3.8). Conversely, if  $\phi$  is any periodic test field, then  $\phi_N(x) = \frac{1}{N}\phi(Nx)$  is again periodic for any integer  $N$ , and it gives the same value as  $\phi$  when substituted into (3.8). If  $N$  is large then we can modify  $\phi_N$  on a thin transition layer to make it vanish at  $\partial C$ , while leaving its energy (the right side of (3.8)) virtually unchanged; therefore (3.4)  $\leq$  (3.8). The argument that (3.6) = (3.9) is essentially the same. ■

A complete proof of Lemma 3.2 will be found in [36]. The fact that periodic test fields are sufficient to test for quasiconvexity was previously noted in [3].

#### 4. Calculation of $Q_\theta W$ for two linear phases with the same elastic moduli.

In this section we present the calculation of  $Q_\theta W$ , when  $W$  is the energy associated to a system of two linear phases with distinct stress-free strains  $a_1, a_2$  but the same tensor of elastic moduli  $\alpha$ :

$$W(\xi) = \min\{W^1(\xi), W^2(\xi)\}$$

$$W^2(\xi) = \langle \alpha(\xi - a_i), \xi - a_i \rangle \quad i = 1, 2. \quad (4.1)$$

Our method is basically the same as used by Khachaturyan in [31,32]; however we emphasize the role of certain projection operators rather than that of the Green's function associated to  $\alpha$ .

We begin with some notation. Let  $S$  be the 6-dimensional space of symmetric tensors. For any  $k \in \mathbb{R}^3$ , let

$$V(k) = \{k \otimes v + v \otimes k : v \in \mathbb{R}^3\}, \quad (4.2)$$

which is a 3-dimensional subspace of  $S$ . For any subspace  $V$  of  $S$  we write  $\pi_V \xi$  for the orthogonal projection of  $\xi$  onto  $V$ .

**Theorem 4.1:** Let  $W$  be given by (4.1), and let  $\theta = (\theta_1, \theta_2) \in \mathcal{V}$ . Then

$$Q_\theta W(\xi) = \theta_1 W^1(\xi) + \theta_2 W^2(\xi) - \theta_1 \theta_2 g \quad (4.3)$$

with

$$g = \max_{|n|=1} |\pi_{\alpha^{1/2}V(n)} \alpha^{1/2}(a_1 - a_2)|^2. \quad (4.4)$$

Whenever  $n^*$  is extremal for (4.3), a laminar microstructure with  $n^*$  as the layer normal gives an optimal phase arrangement.

**Proof:** In view of Lemma 3.2, we must minimize

$$\int [\chi_1 W^1(\xi + e(\phi)) + \chi_2 W^2(\xi + e(\phi))] dx \quad (4.5)$$

over periodic marker functions  $\chi_1, \chi_2 = 1 - \chi_1$  and over periodic deformation fields  $\phi$ . Elementary manipulation using (4.1) transforms the integrand of (4.5) into

$$\chi_1 W^1(\xi) + \chi_2 W^2(\xi) + \langle \alpha e(\phi), e(\phi) \rangle$$

$$+ 2 \langle \xi, \alpha e(\phi) \rangle - 2 \langle \chi_1 a_1 + \chi_2 a_2, \alpha e(\phi) \rangle .$$

Since  $\int \chi_i = \theta_i$ , the integrals of the first two terms are determined:

$$\int \chi_1 W^1(\xi) + \chi_2 W^2(\xi) = \theta_1 W^1(\xi) + \theta_2 W^2(\xi).$$

Since  $\phi$  is periodic and  $\chi_1 + \chi_2 = 1$  we can integrate by parts in the last two terms to get

$$2 \int \langle \xi, \alpha e(\phi) \rangle - \langle \chi_1 a_1 + \chi_2 a_2, \alpha e(\phi) \rangle = -2 \int \chi_1 \langle a_1 - a_2, \alpha e(\phi) \rangle .$$

Thus to prove (4.3) we must establish that

$$\inf_{\int \chi_i = \theta_i} \inf_{\phi} \int \langle \alpha e(\phi), e(\phi) \rangle - 2 \langle \chi_1(a_1 - a_2), \alpha e(\phi) \rangle = -\theta_1 \theta_2 g, \quad (4.6)$$

with  $g$  defined by (4.4).

Fixing  $\chi$ , we shall compute the minimum over  $\phi$  in (4.6) using Fourier analysis. Since our functions are periodic with period  $2\pi$  in each variable, their Fourier transforms are supported on the lattice of integers  $Z^3$ , e.g.

$$\chi_i(x) = \sum_{k \in Z^3} \hat{\chi}_i(k) e^{ik \cdot x}.$$

The integral in (4.6) can be rewritten as

$$\int |\alpha^{1/2} e(\phi)|^2 - 2 \chi_1 \langle \alpha^{1/2}(a_1 - a_2), \alpha^{1/2} e(\phi) \rangle ,$$

where  $\alpha^{1/2}$  is the square root of  $\alpha$ , itself a positive definite symmetric map on the space  $S$  of symmetric tensors. By Plancherel's formula, this is equal to

$$\sum_{k \in Z^3} |\alpha^{1/2} e(\hat{\phi})|^2 - 2 \operatorname{Re} \langle \alpha^{1/2}(a_1 - a_2) \bar{\chi}_1, \alpha^{1/2} e(\hat{\phi}) \rangle , \quad (4.7)$$

in which  $\hat{\chi}_1 = \hat{\chi}_1(k)$ , etc., and  $\langle \cdot, \cdot \rangle$  is the symmetric inner product on complex matrices. Choosing  $\phi$  to minimize (4.6) is the same as choosing  $\hat{\phi}$  to minimize (4.7), which may be done separately at each  $k$ . Frequency 0 is special: it contributes nothing to (4.7), since  $e(\hat{\phi})(0) = 0$  for any periodic  $\phi$ . When  $k \neq 0$ , the optimal value of  $\eta = \alpha^{1/2} e(\hat{\phi})(k)$  is obtained by minimizing

$$|\eta|^2 - 2 \operatorname{Re} \langle \eta \overline{\hat{\chi}_1(k)}, \alpha^{1/2}(a_1 - a_2) \rangle \quad (4.8)$$

over the space of all possible values of  $\alpha^{1/2}e(\hat{\phi})(k)$ , which is the complexification of  $\alpha^{1/2}V(k)$ . The necessary linear algebra is presented as Lemma 4.2 below; the optimal  $\eta$  turns out to be

$$\eta = \hat{\chi}_1(k) \pi_{\alpha^{1/2}V(k)} \alpha^{1/2}(a_1 - a_2),$$

and substitution into (4.8) gives the value

$$-|\hat{\chi}_1(k)|^2 |\pi_{\alpha^{1/2}V(k)} \alpha^{1/2}(a_1 - a_2)|^2.$$

Thus for given  $\chi_1$ ,

$$\inf_{\phi} (4.5) = - \sum_{k \neq 0} |\hat{\chi}_1(k)|^2 |\pi_{\alpha^{1/2}V(k)} \alpha^{1/2}(a_1 - a_2)|^2.$$

Next we must minimize this expression over  $\chi_1$ . Since the subspace  $V(k)$  depends only on  $k/|k|$ , it is immediate from (4.4) that

$$|\pi_{\alpha^{1/2}V(k)} \alpha^{1/2}(a_1 - a_2)|^2 \leq g,$$

with equality if and only if  $n = k/|k|$  is extremal for (4.4). Since  $\int \chi_1 = \theta_1$ , another application of Plancherel's formula gives

$$\sum_{k \neq 0} |\hat{\chi}_1(k)|^2 = \int (\chi_1 - \theta_1)^2 = \theta_1 \theta_2.$$

It follows that

$$- \sum_{k \neq 0} |\hat{\chi}_1(k)|^2 |\pi_{\alpha^{1/2}V(k)} \alpha^{1/2}(a_1 - a_2)|^2 \geq -\theta_1 \theta_2 g, \quad (4.9)$$

with equality if  $k/|k|$  is extremal for (4.4) whenever  $\hat{\chi}(k) \neq 0$ .

To complete the proof of (4.6), we must show that equality can be approached in (4.9). To this end, note that a layered geometry has its Fourier transform concentrated on the line parallel to the layer normal; in other words, if  $\chi_1(x) = f(x \cdot k)$  with  $f$  periodic, then  $\hat{\chi}_1$  is supported on the line through  $k$ . For  $k \in \mathbb{Z}^3$  this  $\chi_1$  is  $C$ -periodic provided that  $f$  has period  $2\pi$ , and  $\int \chi_1 = \theta_1$  provided that  $\int f = \theta_1$ . As  $k/|k|$  approaches an extremal for (4.4), this layered phase arrangement establishes the optimality of the lower bound (4.9), completing the proof. ■

In minimizing (4.8), we made use of the following lemma from linear algebra:

**Lemma 4.2.** Let  $V$  be a subspace of a finite dimensional real vector space  $S$ , and let  $V_C \subset S_C$  be their complexifications. For any  $\xi \in S$  and any complex number  $c$ ,

$$\min_{\eta \in V_C} |\eta|^2 - 2 \operatorname{Re} \langle c\eta, \xi \rangle = -|c|^2 |\pi_V \xi|^2, \quad (4.10)$$

The extremal  $\eta$  being  $\eta^* = \bar{c} \pi_V \xi$ .

**Proof:** The function  $f(\eta) = |\eta|^2 - 2 \operatorname{Re} \langle c\eta, \xi \rangle$  is convex, and

$$\frac{\partial f}{\partial \eta_j} = \bar{\eta}_j - c \xi_j, \quad \frac{\partial f}{\partial \bar{\eta}_j} = \eta_j - \bar{c} \xi_j.$$

The proposed extremal  $\eta^*$  belongs to  $V_C$ , and the directional derivative of  $f$  at  $\eta^*$  vanishes in directions  $\dot{\eta} \in V_C$ :

$$\begin{aligned} D_{\dot{\eta}} f(\eta^*) &= \sum \frac{\partial f}{\partial \eta_j} \dot{\eta}_j + \frac{\partial f}{\partial \bar{\eta}_j} \dot{\bar{\eta}}_j \\ &= \langle c(\pi_V \xi - \xi), \dot{\eta} \rangle + \langle \bar{c}(\pi_V \xi - \xi), \dot{\bar{\eta}} \rangle \\ &= \langle \pi_V \xi - \xi, \operatorname{Re}(c\dot{\eta}) \rangle = 0 \end{aligned}$$

since  $\operatorname{Re}(c\dot{\eta}) \in V$ . It follows that the minimum of  $f$  on  $V_C$  is achieved at  $\eta^*$ , and substitution yields (4.10). ■

Theorem 4.1 gives a formula for  $Q_\theta W$  in terms of a single real constant  $g$ , which must be determined through an optimization over  $S^2$  (see (4.4)). If  $a_1 - a_2$  is dyadic, i.e. if

$$a_1 - a_2 = p \otimes q + q \otimes p \quad (4.11)$$

for some  $p, q \in \mathbb{R}^3$ , then clearly  $g = |\alpha^{1/2}(a_1 - a_2)|^2$ , and the extremals for (4.3) are  $n = \pm p/|p|$  and  $n = \pm q/|q|$ . This is the case when the two phases can coexist in their stress-free states, separated by hyperplanes orthogonal to  $p$  or  $q$ . If (4.11) does not hold then  $g < |\alpha^{1/2}(a_1 - a_2)|^2$ , and the optimization (4.4) is more subtle. It has been

examined, and the extremal choices of  $n$  have been identified, under various symmetry hypotheses on  $\alpha$  [26,41,48,55,60,61]; the most complete such study is [41].

We have observed that there is always an extremal geometry which is layered. The extremal geometry is generally not unique, however. The proof of Theorem 4.1 shows that a geometry is optimal exactly if  $\hat{\chi}_1$  is concentrated on the extremals of (4.4). Similarly, a sequence of microstructures is asymptotically optimal if the associated marker functions  $\chi_1^j, \chi_2^j = 1 - \chi_1^j$  "have their Fourier transforms concentrated at the extremals of (4.4), asymptotically as  $j \rightarrow \infty$ ." This is best made precise by using the notion of an H-measure, recently introduced and explored in a much more general setting by Tartar [58] and Gerard [24]. In our periodic setting, the H-measure associated to a periodic marker function  $\chi_1$  is the measure  $\mu$  on  $S^2$  defined by

$$\mu = \sum_{k \neq 0} |\hat{\chi}_1(k)|^2 \delta_{k/|k|}, \quad (4.12)$$

where  $k$  ranges over  $\mathbb{Z}^3$  and  $\delta_n$  is the Dirac measure concentrated at  $n$ . With this terminology, we can assert:

**Proposition 4.3.** Consider a sequence of microstructures corresponding to marker functions  $\chi_1^j, \chi_2^j = 1 - \chi_1^j, j = 1, 2, \dots$ , and let  $\mu_j$  be the H-measure associated to  $\chi_1^j$  as in (4.12). Then the sequence  $\{\chi_1^j\}$  is asymptotically optimal for (4.5) if and only if every weak limit of the sequence  $\{\mu_j\}$  is supported on the set of extremals for (4.4).

The proof is an easy extension of that given for Theorem 4.1, so it is omitted.

**Remark 4.4.** In the homogenization literature, many composites with extremal characteristics have been found in the class of "sequentially laminated" microstructures, see e.g. [2,23,38,46,49,50,57]. The relevance of such microstructures to coherent phase mixtures was noted long ago by Roitburd, who called them "polydomain structures of second or higher order" [51-54]. Using this construction, one can construct a large variety of optimal phase microgeometries if (4.4) has at least two linearly independent extremals. See [36] for details.

■

The relaxation of (4.1) is entirely determined by Proposition 3.1 and Theorem 4.1. We refer to [36] for a detailed discussion of the properties of QW. As an application of

Theorem 4.1, however, let us show here that  $QW$  still has double-well structure if the stress-free strains  $a_1$  and  $a_2$  are incompatible.

**Proposition 4.1.** Let  $W$  be given by (4.1), and assume that  $a_1 - a_2$  does not have the form  $p \otimes q + q \otimes p$ . Then  $QW(a_1) = QW(a_2) = 0$  and

$$QW(\xi) > 0 \text{ for } \xi \neq a_1, a_2. \quad (4.13)$$

**Proof:** As a first step, we claim that the formula for  $Q_\theta W$  can be expressed as

$$Q_\theta W(\xi) = |\alpha^{1/2}[\xi - \bar{a}(\theta)]|^2 + \theta_1 \theta_2 h \quad (4.14)$$

with

$$\begin{aligned} \bar{a}(\theta) &= \theta_1 a_1 + \theta_2 a_2 \\ h &= |\alpha^{1/2}(a_1 - a_2)|^2 - g. \end{aligned}$$

Indeed, the equivalence of (4.3) and (4.14) is a matter of elementary manipulation, making use of the identity

$$|\alpha^{1/2}(\xi - \bar{a})|^2 = \theta_1 |\alpha^{1/2}(\xi - a_1)|^2 + \theta_2 |\alpha^{1/2}(\xi - a_2)|^2 - \theta_1 \theta_2 |\alpha^{1/2}(a_1 - a_2)|^2.$$

Notice that  $h > 0$  as a consequence of our hypothesis on  $a_1 - a_2$ .

Next, we apply Proposition 3.1 to obtain that

$$QW(\xi) = \min_{\theta} \{ |\alpha^{1/2}[\xi - \bar{a}(\theta)]|^2 + \theta_1 \theta_2 h \}. \quad (4.15)$$

The minimum in (4.15) is achieved, since  $\theta = (\theta_1, \theta_2)$  varies over the compact set  $\{\theta_1 \geq 0, \theta_2 \geq 0, \theta_1 + \theta_2 = 1\}$ .

Now the assertions of the theorem follow easily. It is obvious that  $QW \geq 0$ . If, for some  $\xi$ ,  $QW(\xi) = 0$ , then the optimal  $\theta$  for (4.15) would give

$$|\alpha^{1/2}[\xi - \bar{a}(\theta)]|^2 + \theta_1 \theta_2 h = 0.$$

Both terms are non negative, so they must both vanish. We conclude that  $\xi = \bar{a}(\theta)$  and  $\theta_1 \theta_2 = 0$ . Therefore either  $\theta = (1, 0)$ ,  $\xi = a_1$ , or else  $\theta = (0, 1)$ ,  $\xi = a_2$ . Thus  $QW$  is strictly positive except at  $a_1$  and  $a_2$ . ■

5. Toward the calculation of  $Q_\theta W$  for a system of many phases, all with the same elastic moduli.

It is natural to ask whether the method of Section 4 might be applied to an energy function describing more than two phases. The answer at present is no. One can certainly begin the same way, but it is not clear how to minimize the resulting expression in Fourier space over the class of all microstructures. The calculation given here is equivalent to that of Khachaturyan and Shatalov in [32,33]. We present it to clarify the relation between the linear and nonlinear variational theories, and to focus attention on this as a significant open problem.

We consider a system of  $N$  phases, with stress-free strains  $\{a_i\}_{i=1}^N$  and the same elastic law  $\alpha$ :

$$W(\xi) = \min\{W^1(\xi), \dots, W^N(\xi)\}$$

$$W^i(\xi) = \langle \alpha(\xi - a_i), \xi - a_i \rangle \quad 1 \leq i \leq N. \quad (5.1)$$

Our starting point is once again Lemma 3.2: for any vector  $\theta = (\theta_1, \dots, \theta_N)$  of volume fractions,

$$Q_\theta W(\xi) = \inf \int \sum_{j=1}^N \chi_j W^j(\xi + e(\phi)) \, dx. \quad (5.2)$$

Here  $\phi$  ranges over periodic deformations, and  $\{\chi_j\}$  are the periodic marker functions associated to any subdivision of the unit cell with the specified volume fractions. Expanding the integrand of (5.2) gives

$$\sum \chi_j W^j(\xi) + \sum 2\chi_j \langle \xi - a_j, \alpha e(\phi) \rangle + \langle \alpha e(\phi), e(\phi) \rangle,$$

since  $\sum \chi_j = 1$ . The average of this expression is

$$\sum \theta_j W^j(\xi) + \int [\langle \alpha e(\phi), e(\phi) \rangle - 2 \sum \chi_j \langle a_j, \alpha e(\phi) \rangle],$$

since the part of the middle term involving  $\xi$  integrates by parts to zero. Thus the essential problem in calculating  $Q_\theta W$  is the evaluation of

$$\inf_{\chi_j = \theta_j} \int \inf_{\phi} [\langle \alpha e(\phi), e(\phi) \rangle - 2 \sum \chi_j \langle a_j, \alpha e(\phi) \rangle] dx. \quad (5.3)$$

As before, we may evaluate the minimum over  $\phi$  by Fourier analysis: fixing  $\{\chi_j\}$ , we seek



$$\inf_{\phi} \int |\alpha^{1/2} e(\phi)|^2 - 2 \langle \alpha^{1/2} \sum \chi_j a_j, \alpha^{1/2} e(\phi) \rangle dx. \quad (5.4)$$

Taking the Fourier transform then minimizing at each frequency as in (4.6)-(4.7) shows that (5.4) equals

$$- \sum_{k \neq 0} |\pi_{\alpha^{1/2} V(k)} (\alpha^{1/2} \sum \hat{\chi}_j(k) a_j)|^2 \quad (5.5)$$

This can be written in terms of H-measures as follows: for  $1 \leq i, j \leq N$ , let  $\mu = (\mu_{ij})$  be the Hermitian matrix valued measure on  $S^2$  with components

$$\mu_{ij} = \sum_{k \neq 0} \hat{\chi}_i(k) \overline{\hat{\chi}_j(k)} \delta_{k/|k|}.$$

Then (5.5) can be rearranged to give

$$- \sum_{i,j=1}^N \int_{S^2} \langle \pi_{\alpha^{1/2} V(k)} \alpha^{1/2} a_i, \pi_{\alpha^{1/2} V(k)} \alpha^{1/2} a_j \rangle d\mu_{ij}. \quad (5.6)$$

Thus calculating  $Q_{\theta} W$  is equivalent to **minimizing (5.5) over all marker functions  $\{\chi_j\}$  with the specified volume fractions; alternatively, it is equivalent to minimizing a certain linear functional over the class of all H-measures associated to such marker functions.** For the case of two phases ( $N=2$ ) this is precisely what we did in Section 4; for three or more phases we are presently unable to give a formula (or a finite-dimensional optimization) for (5.5)-(5.6).

There is one special case when the calculation can be completed: that is when  $\alpha$  is isotropic and each  $a_j$  is a multiple of the identity, say  $a_j = \lambda_j I$ . It is well-known that under these circumstances the elastic energy is independent of phase geometry. To recover this result, we observe that (5.5) becomes

$$- \sum_{k \neq 0} \sum_{i,j} \lambda_i \lambda_j \hat{\chi}_i(k) \overline{\hat{\chi}_j(k)} |\pi_{\alpha^{1/2} V(k)} (\alpha^{1/2} I)|^2. \quad (5.7)$$

Since  $\alpha$  is isotropic  $\pi_{\alpha^{1/2} V(k)} (\alpha^{1/2} I)$  cannot depend on  $k$ , and

$$\sum_{k \neq 0} \sum_{i,j} \lambda_i \lambda_j \hat{\chi}_i(k) \overline{\hat{\chi}_j(k)} = \int |\sum \lambda_i (\chi_i - \theta_i)|^2,$$

which obviously depends only on  $\theta = (\theta_1, \dots, \theta_N)$ . Thus (5.5) depends only on  $\theta$ , not on the phase geometry.

Even when  $Q_\theta W$  cannot be computed explicitly, it still makes sense to ask where  $Q_\theta W = 0$ . The point is that  $Q_\theta W(\xi) \geq 0$  for all  $\xi$  (since  $W(\xi) \geq 0$ ); so one can establish that  $Q_\theta W(\xi_\theta) = 0$  for some  $\xi_\theta$  by displaying a microstructure which achieves this. In other words, to show that  $Q_\theta W(\xi_\theta) = 0$  it suffices to prescribe test fields  $\phi$  and  $\{\chi_j\}$  for use in (5.2) which have  $\sum \chi_j W^j(\xi_\theta + e(\phi)) \approx 0$ . The class of sequentially laminated microstructures provides a powerful tool for such constructions; it has been used in [32,33] (in the linearized setting) and in [4,5] (in the geometrically nonlinear context) to show that  $Q_\theta W(\xi_\theta) = 0$  for certain  $\xi_\theta$ , when  $W$  models a cubic-tetragonal phase transition such as that discussed in Section 2.

An intriguing open question is whether the extremum of (5.7) can always be found within the class of sequentially layered microstructures. We hope that the answer is affirmative. Such a result seems, however, beyond the power of the existing mathematical methods.

## REFERENCES

- [1] E. Acerbi and N. Fusco, "Semicontinuity problems in the calculus of variations," Arch. Rat. Mech. Anal. 86, 1984, pp. 125-145.
- [2] M. Avellaneda, "Optimal bounds and microgeometries for elastic two-phase composites," SIAM Journal Appl. Math. 47, 1987, pp. 1216-1228.
- [3] J.M. Ball and F. Murat, " $W^{1,p}$  - quasiconvexity and variational problems for multiple integrals," Journal Funct. Anal. 58, 1984, pp. 225-253.
- [4] J. M. Ball and R.D. James, "Fine phase mixtures as minimizers of energy," Arch. Rat. Mech. Anal. 100, 1987, pp. 13-52.
- [5] J. M. Ball and R.D. James, "Proposed experimental tests of a theory of fine microstructure and the two well problem," to appear.
- [6] M. Chipot and D. Kinderlehrer, "Equilibrium configurations of crystals," Arch. Rat. Mech. Anal. 103, 1988, pp. 237-277.

- [7] J. W. Christian, The Theory of Transformations in Metals and Alloys , Pergamon Press, 1975.
- [8] C. Collins and M. Luskin, "The computation of the austenitic-martensitic phase transition," in proc. conf. on phase transitions, Nice, 1989, M. Rascle ed., to appear.
- [9] B. Dacorogna, "Quasiconvexity and relaxation of nonconvex variational problems," Journal Funct. Anal. 46, 1982, pp. 102-118.
- [10] B. Dacorogna, Direct Methods in the Calculus of Variations , Springer-Verlag, 1989.
- [11] J. L. Ericksen, "Changes in symmetry in elastic crystals," IUTAM Symp. Finite Elasticity , D. Carlson and R. Shield, eds, M. Nijhoff, 1981, pp. 167-177.
- [12] J. L. Ericksen, "The Cauchy and Born hypotheses for crystals," in Phase Transformations and Material Instabilities in Solids , M. Gurtin, ed., Academic Press, 1984, pp. 61-78.
- [13] J. L. Ericksen, "Constitutive theory for some constrained elastic crystals," Int. Journal Solids Structures 22, 1986, pp. 951-964.
- [14] J. L. Ericksen, "Stable equilibrium configurations of elastic crystals," Arch. Rat. Mech. Anal. 94, 1986, pp. 1-14.
- [15] J. L. Ericksen, "Twinning of crystals I," in Metastability and Incompletely Posed Problems , S. Antman et al. eds, Springer-Verlag, 1987, pp. 77-94.
- [16] J. D. Eshelby, "The determination of the elastic field of an ellipsoidal inclusion, and related problems," Proc. Roy. Soc. Lon. 241 A, 1957, pp. 376-396.
- [17] J. D. Eshelby, "The elastic field outside an ellipsoidal inclusion," Proc. Roy. Soc. Lon. 252A, 1959, pp. 561-569.
- [18] I. Fonseca, "Variational methods for elastic crystals," Arch. Rat. Mech. Anal. 97, 1987, pp. 189-220.
- [19] I. Fonseca, "Stability of elastic crystals," in Non-Classical Continuum Mechanics , R. Knops and A. Lacey eds, Cambridge University Press, 1987, pp. 187-196.
- [20] I. Fonseca, "The lower quasiconvex envelope of the stored energy function for an elastic crystal," Journal Math. Pures et Appl. 67, 1988, pp. 175-195.

- [21] I. Fonseca, "Interfacial energy and the Maxwell rule," Arch. Rat. Mech. Anal. , to appear.
- [22] I. Fonseca and L. Tartar, "The displacement problem for elastic crystals," Proc. Roy. Soc. Edinburgh , Ser.A, to appear.
- [23] G.A. Francfort and F. Murat, "Homogenization and optimal bounds in linear elasticity," Arch. Rat. Mech. Anal. 94, 1986, pp. 307-334.
- [24] P. Gerard, "Compacite par compensation et regularite 2- microlocale," preprint.
- [25] M. Grinfel'd, "Continuum methods in the theory of phase transitions in solids," Phys. Earth and Planetary Interiors 50, 1988, pp. 99-109.
- [26] M. Hong, D.E. Wedge, and J.W. Morris, "The state and habit of the  $Fe_{16}N_2$  precipitate in b.c.c. iron: elastic theory," Acta Metallurgica 32, 1984, pp. 279-288.
- [27] R. D. James, "The arrangement of coherent phases in a loaded body," in Phase Transformations and Material Instabilities in Solids , M. Gurtin, ed, Academic Press, 1984, pp. 79-98.
- [28] R. D. James, "Displacive phase transformations in solids," Journal Mech. Phys. Solids 34, 1986, pp. 359-394.
- [29] R. D. James, "The stability and metastability of quartz," in Metastability and Incompletely Posed Problems , S. Antman et al. eds, Springer-Verlag, 1987, pp. 147-176.
- [30] R. D. James and D. Kinderlehrer, "Theory of diffusionless phase transitions," in proc. conf. on phase transitions, Nice, 1989, M. Rascle ed., to appear.
- [31] A. G. Khachaturyan, "Some questions concerning the theory of phase transformations in solids," Soviet Physics-Solid State 8, 1967, pp. 2163-2168.
- [32] A. G. Khachaturyan, Theory of Structural Transformations in Solids , John Wiley and Sons, 1983.
- [33] A. G. Khachaturyan and G. A. Shatalov, "Theory of macroscopic periodicity for a phase transition in the solid state," Soviet Physics IETP 29, 1969, pp. 557-561.
- [34] D. Kinderlehrer, "Twinning of crystals II," in Metastability and Incompletely Posed Problems , S. Antman et al eds, Springer-Verlag, 1987, pp. 135-146.

- [35] D. Kinderlehrer and G. Vergara-Caffarelli, "The relaxation of functionals with surface energies," IMA preprint No. 491, 1989.
- [36] R. V. Kohn, "Relaxation of double-well energy functions," in preparation.
- [37] R. V. Kohn, "Relaxation of the elastic energy for a system of two coherent phases with well-ordered elastic moduli," in preparation.
- [38] R. V. Kohn and R. Lipton, "Optimal bounds for the effective energy of a mixture of isotropic, incompressible, elastic materials," Arch. Rat. Mech. Anal. 102, 1988, pp. 331-350.
- [39] R. V. Kohn and G. Strang, "Optimal design and relaxation of variational problems, I-III," Comm. Pure Appl. Math. 34, 1987, pp. 113-137, 139-182, and 353-377.
- [40] R. V. Kohn and M. Vogelius, "Relaxation of a variational method for impedance computed tomography," Comm. Pure Appl. Math. 40, 1987, pp. 745-777.
- [41] E. Kostlan and J. W. Morris, "The preferred habit of a coherent thin-plate inclusion in an anisotropic elastic solid," Acta Metallurgica 35, 1987, pp. 2167-2175.
- [42] F. Larche and J. W. Cahn, "A linear theory of thermochemical equilibrium of solids under stress," Acta Metallurgica 21, 1973, pp. 1051-1063.
- [43] F. Larche and J. W. Cahn, "A nonlinear theory of thermochemical equilibrium of solids under stress," Acta Metallurgica 26, 1978, pp. 53-60.
- [44] F. C. Larche and J. W. Cahn, "Thermomechanical equilibrium of multiphase solids under stress," Acta Metallurgica 26, 1978, pp. 1579-1589.
- [45] J. K. Lee, D. M. Barnett, and H. I. Aaronson, "The elastic strain energy of coherent ellipsoidal precipitates in anisotropic crystalline solids," Metallurgical Trans. 8A, 1977, pp. 963-970.
- [46] K. A. Lurie and A. V. Cherkaev, "Exact estimates of the conductivity of a binary mixture of isotropic components," Proc. Roy. Soc. Edinburgh 104A, 1986, pp. 21-38.
- [47] K. A. Lurie and A. V. Cherkaev, "On a certain variational problem of phase equilibrium," in Material Instabilities in Continuum Mechanics, J. M. Ball, ed., Oxford University Press, 1988, pp. 257-268.

- [48] W. E. Mayo and T. Tsakalakos, "The influence of elastic strain energy on the formation of coherent hexagonal phases," Metallurgical Trans. 11A, 1980, pp. 1637-1644.
- [49] G. Milton, "Modelling the properties of composites by laminates," in Homogenization and Effective Moduli of Materials and Media, J. Ericksen et al. eds, Springer-Verlag, 1986, pp. 150-175.
- [50] G. W. Milton and R. V. Kohn, "Variational bounds on the effective moduli of anisotropic composites," Journal Mech. Phys. Solids 36, 1988, pp. 597-629.
- [51] A. L. Roitburd, Kristallografiya 12, 1967, pp. 567ff. (In Russian)
- [52] A. L. Roitburd, "The domain structure of crystals formed in the solid phase," Sov. Phys. - Solid State 10, 1969, pp. 2870-2876.
- [53] A. L. Roitburd, "Domain structure caused by internal stresses in heterophase solids," Phys. Stat. Sol. (a) 16, 1973, pp. 329-339.
- [54] A. L. Roitburd, "Martensitic transformation as a typical phase transformation in solids," in Solid State Physics 33, Academic Press, 1978, pp. 317-390.
- [55] A. L. Roitburd and N. S. Kosenko, "Orientational dependence of the elastic energy of a plane interlayer in a system of coherent phases," Phys. Stat. Sol. (a) 35, 1976, pp. 735-746.
- [56] R. Schneck, S. I. Rokhlin, and M. P. Dariel, "Criterion for predicting the morphology of crystalline cubic precipitates in a cubic matrix," Metallurgical Trans., 16A, 1985, pp. 197-202.
- [57] L. Tartar, "Estimations fines de coefficients homogeneises," in Ennio de Giorgi's Colloquium, P. Kree, ed., Pitman, 1985, pp. 168-187.
- [58] L. Tartar, "H-measures: a new approach for studying homogenization, oscillations, and concentration effects in partial differential equations," Proc. Roy. Soc. Edinburgh Ser. A, to appear.
- [59] C. M. Wayman, Introduction to the Crystallography of Martensitic Transformations, MacMillan, 1964.
- [60] S. H. Wen, E. Kostlan, M. Hong, A. G. Khachaturyan, and J. W. Morris, Jr., "The preferred habit of a tetragonal inclusion in a cubic matrix," Acta Metallurgica 29,

1981, pp. 1247-1254.

[61] J. A. Wert, "The strain energy of a disc-shaped GP zone," *Acta Metallurgica* 24, 1976, pp. 65-71.

# Relation Between Microscopic and Macroscopic Properties in Crystals Undergoing Phase Transformation\*

R. D. James  
Department of Aerospace Engineering and Mechanics  
University of Minnesota  
Minneapolis, MN 55455

Army Conference on Applied Mathematics and Computing  
1989

**Abstract.** When cooled below a certain critical temperature  $\theta_c$ , many crystals undergo a structural phase transformation marked by the appearance of *microstructure*. In the simplest case, this microstructure consists of fine parallel bands, ranging in thickness from a few atomic spacings in NiMn (termed *microwinning* by Baele, van Tendeloo and Amelinckx [2]), to a few microns in InTi. In some of these materials the application of limited displacements to the boundary of the crystal causes a rearrangement of the microstructure. In this talk I describe recent attempts to understand why such microstructures form and how imposed deformations are accommodated by rearrangements of the microstructure. The ideas suggest a new approach to micromagnetics described in Section 5.

## 1. Introduction

The great usefulness of the classical field theories of elasticity, hydrodynamics, thermodynamics and electromagnetism arises from their ability to accurately predict, from a knowledge only of boundary and initial conditions and a few material parameters, the complex fields of deformation, stress, temperature and electromagnetic fields in a deforming material. On the whole, these theories have failed to make similar predictions about materials containing domains or defects. Alternatively, the historical practice has been to make rather restrictive assumptions about the geometry and arrangement of defects and then to calculate something about them using linear theory. This has led to a hodgepodge of special theories of defects having the inherent limitation that they are unable to deal with any situation not envisaged by the severe geometric restrictions assumed at the outset. In particular, they are generally unable to explore conditions that might give rise to new and unusual microstructures important to the development of advanced materials. They are also extremely limited in coping with dynamic situations.

The point of view adopted in the research described here is that the domain structure itself should be predicted from some equations without *a priori* geometric restrictions. This point of view is not new and was expressed nicely by L.M. Brown [8, 9] in his books on the domain structure of magnetic materials; he termed this approach *micromagnetics*. In the magnetic case, the competing theory involving geometric (and other) restrictions is called *domain theory*. Similarly, in the area of martensitic transformations the *crystallographic theory of martensite* has served the subject well, but primarily as a way to understand only a rather special microstructure among many that are actually observed. A new theory, in some ways analogous to micromagnetics, developed by J.M. Ball and the author ([3, 4], see also Chipot and Kinderlehrer [7]), is designed to avoid the geometric restrictions adopted by the crystallographic theory and to offer possibilities for prediction

---

\* Sponsored by the U.S. Army Research Office. Section 5 of this paper was presented at the "Elasticity Retreat", South Pomfret, Vermont, August 23, 1989.



of complex microstructures. In this talk I describe recent predictions of this theory and a plan for an experimental test of the theory.

These recent predictions all concern unloaded crystals. In a companion paper in this volume, R. Kohn [13] discusses recent results on the microstructure of loaded crystals and relates these results to the metallurgical literature.

I return to micromagnetics in Section 5. Despite the attractiveness of Brown's philosophy, his calculations met with limited success. The reason for this has been explored recently by D. Kinderlehrer and the author [10] and seems to arise from the fact that the free energy he adopts does not have a minimum in the conventional sense.

## 2. Austenite/Martensite Interface

The crystallographic theory of martensite, due to Wechsler, Lieberman and Read [19] and Bowles and Mackenzie [6], is a theory for the description of a special microstructure known as the austenite/martensite interface. This is a microstructure pictured in Figure 1 consisting of fine bands of martensite (twins) on one side of the interface and homogeneous austenite on the other side of the interface. It predicts the normal to the interface (the so-called habit plane), the proportion of volume occupied by one twin relative to the other, the relative orientation of the twinned martensite lattice relative to the austenite lattice, and the macroscopic deformation of the twinned martensite relative to the austenite. By "macroscopic deformation" is meant the homogeneous deformation which takes the austenite lattice to the twinned martensite lattice, neglecting the small zig-zags produced by the twins themselves. Figure 1 shows a picture of the atoms in a typical austenite/martensite interface. This picture was generated using lattice parameters appropriate to the alloy  $\text{Ni}_{62}\text{Al}_{38}$ , and only Ni atoms are shown. Recent high resolution electron micrographs of Schryvers [16] show many of the features of Figure 1 including accurate atomic periodicity of the twins.

The crystallographic theory of martensite is based on the following assumptions. A certain twin system is presumed given; this means that orthogonal vectors  $\mathbf{a}$  and  $\mathbf{n}$  are given where  $\mathbf{I} + \mathbf{a} \otimes \mathbf{n}$  is the shear that maps the transformed lattice into its twin. It is recognized that the martensite (to the left of the interface in Figure 1) is finely twinned so that its macroscopic deformation is in fact

$$\mathbf{P}_\lambda := \mathbf{I} + \lambda \mathbf{a} \otimes \mathbf{n} ,$$

where  $\lambda \in (0, 1)$  represents the volume fraction of one variant of martensite relative to its twin. A positive-definite symmetric matrix  $\mathbf{U}$  is given which represents the pure stretch, or *Bain strain*, associated with the transformation from austenite to untwinned martensite. The average deformation of the twinned martensite body is assumed to have the form  $\mathbf{E} = \mathbf{I} + \mathbf{b} \otimes \mathbf{m}$ ,  $\mathbf{m}$  being the normal to the austenite/martensite interface. The basic equation of the crystallographic theory is then written [see, e.g., 19]

$$\mathbf{E} = \mathbf{R} \mathbf{U} \mathbf{P}_\lambda , \tag{2.1}$$

or, using the definitions given above,

$$\mathbf{I} + \mathbf{b} \otimes \mathbf{m} = \mathbf{R} \mathbf{U} (\mathbf{I} + \lambda \mathbf{a} \otimes \mathbf{n}) . \tag{2.2}$$

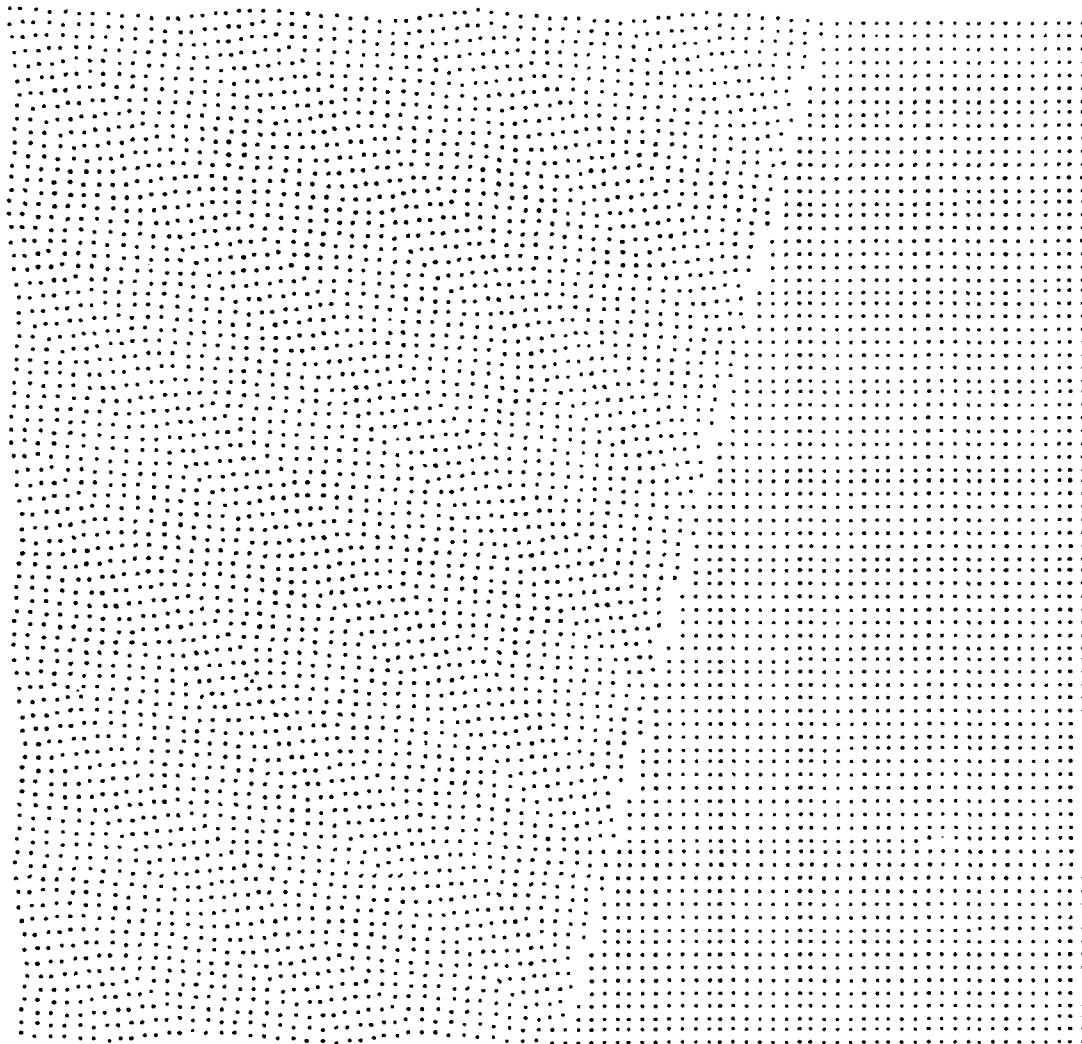


Figure 1. Austenite-martensite interface drawn with lattice parameters appropriate to  $\text{Ni}_{62}\text{Al}_{38}$ .

Here  $\mathbf{R} \in \text{SO}(3)$  is an arbitrary rotation matrix. In applications of equation (2.2), the matrix  $\mathbf{U}$  and the vectors  $\mathbf{a}$  and  $\mathbf{n}$  are given and (2.2) is solved for  $\mathbf{b} \in \mathbb{R}^3$ ,  $\mathbf{m} \in \mathbb{R}^3$ ,  $\mathbf{R} \in \text{SO}(3)$ ,  $\lambda \in (0, 1)$ . An excellent treatment of the crystallographic theory is given in the original paper of Wechsler, Lieberman and Read [19].

An alternative view of this microstructure is given by Ball and James [3]. They consider a free energy density  $\phi(\mathbf{D}\mathbf{y}, \theta)$  where  $\mathbf{y}:\Omega \rightarrow \mathbb{R}^3$  describes the deformation, either ordinary elastic deformation or the deformation associated with transformation, and  $\theta$  represents the temperature. Principles of frame-indifference and lattice symmetry are used to restrict the form of  $\phi$ . With symmetry assumptions appropriate to a cubic-to-tetragonal phase transformation, there emerge three scalar functions of temperature,  $\alpha(\theta)$ ,  $\eta_1(\theta)$  and  $\eta_2(\theta)$ , and a fixed orthonormal basis  $\{\mathbf{e}_i\}$  such that

$$\begin{aligned} \text{for } \theta \geq \theta_c \quad \phi(\alpha(\theta)\mathbf{1}, \theta) &\leq \phi(\mathbf{F}, \theta) \quad \forall \mathbf{F} \quad , \\ \text{for } \theta \leq \theta_c \quad \phi(\mathbf{U}_i(\theta), \theta) &\leq \phi(\mathbf{F}, \theta) \quad \forall \mathbf{F} \quad , \end{aligned} \tag{2.3}$$

where  $\mathbf{U}_i(\theta) := \eta_1(\theta)\mathbf{1} + (\eta_2(\theta) - \eta_1(\theta))\mathbf{e}_i \otimes \mathbf{e}_i$  (no sum).

Thus  $\phi$  has “potential wells” and, after accounting for the condition of frame-indifference ( $\phi(\mathbf{R}\mathbf{F}, \theta) = \phi(\mathbf{F}, \theta) \quad \forall \mathbf{F}$  and  $\forall \mathbf{R} \in \text{SO}(3)$ ) we are led to adopt the terminology,

$\alpha(\theta)\text{SO}(3)$ .....Austenite well  
 $\text{SO}(3)\mathbf{U}_1(\theta) \cup \text{SO}(3)\mathbf{U}_2(\theta) \cup \text{SO}(3)\mathbf{U}_3(\theta)$ .....Martensite wells

Microstructures in *stable equilibrium* at the temperature  $\theta$  are minimizers  $\mathbf{y}(\mathbf{x})$  of the total energy

$$E_\theta[\mathbf{y}] = \int_{\Omega} \phi(\mathbf{D}\mathbf{y}(\mathbf{x}), \theta) d\mathbf{x} \quad . \tag{2.4}$$

This formulation does not involve the geometric restrictions mentioned in the Introduction.

The energy  $E_\theta[\cdot]$  fails the condition of weak lower semicontinuity in  $W^{1,\infty}(\Omega, \mathbb{R}^3)$ . The effect of this is that there are minimizing sequences  $\mathbf{y}^{(k)} \xrightarrow{*} \mathbf{y}$  in  $W^{1,\infty}$  with

$$E_\theta[\mathbf{y}] > \inf_k E_\theta[\mathbf{y}^{(k)}] \quad .$$

These sequences involve finer and finer oscillations which model the phenomenon of microtwinning as pictured in Figure 1.

How does the austenite/martensite interface emerge as a minimizing sequence? There is a huge variety of minimizing sequences to the problem, just as there is a huge variety of microstructures observed in transformed crystals. To discuss the austenite/martensite interface we gather from Figure 1 that essentially three deformation gradients participate in this microstructure. As a way of quantifying this restriction, the concept of a *Young measure* is useful. The basic theorem on Young measures states that to every sequence  $\mathbf{y}^{(k)} \xrightarrow{*} \mathbf{y}$  (in  $W^{1,\infty}$ ) we can assign a family of probability measures  $\nu_{\mathbf{x}}, \mathbf{x} \in \Omega$ , such that for every continuous function  $f: M^{3 \times 3} \rightarrow \mathbb{R}$ ,

$$f(\mathbf{D}\mathbf{y}^{(k)}) \xrightarrow{*} \int_{M^{3 \times 3}} f(\mathbf{G}) d\nu_{\mathbf{x}}(\mathbf{G}) \quad .$$

A version of this theorem which is ideally suited to these problems of microstructure is given by Ball [5]. An easy consequence of this theorem is that (with mild growth assumptions on  $\phi$ ) for any minimizing sequence of  $E_\theta[\cdot]$ , the support of the Young measure lies on the potential wells of  $\phi(\cdot, \theta)$ . This support may be thought of as the set of deformation gradients that “participate” in the microstructure.

Hence, the austenite-martensite interface should be modeled by a minimizing sequence  $y^{(k)}$  whose Young measure  $\nu_x$  is supported on three matrices  $F^+$ ,  $F^-$ ,  $I$  where  $F^+ - F^- = \hat{a} \otimes n$  for some  $a \in \mathbb{R}^3$ ,  $n \in \mathbb{R}^3$ , and  $F^+$  and  $F^-$  belong to the martensite wells (it is easily calculated that the martensite wells have such rank-1 connections). In the spirit of "no geometric restrictions," we prefer to say *nothing* about how the sets on which  $Dy^k$  takes on (approximately) the values  $I$ ,  $F^+$  and  $F^-$  are arranged. Under the condition only that  $\nu_x$  is supported on  $\{I, F^+, F^-\}$ , James and Kinderlehrer [11] prove that there is  $\lambda \in (0, 1)$  and vectors  $b$  and  $m$  such that

$$\lambda F^+ + (1-\lambda)F^- = I + b \otimes m$$

But (2.5) is precisely the equation of the crystallographic theory of martensite when we recognize that

$$\begin{aligned} \lambda F^+ + (1-\lambda)F^- &= F^- + \lambda \hat{a} \otimes n \\ &= RU_i(I + \lambda a \otimes n) \end{aligned}$$

where  $i=1, 2$  or  $3$  and  $a = (RU_i)^{-1} \hat{a}$ . It is found that this vector  $a$  is exactly the one used as input to the crystallographic theory. Further information on the geometry of this microstructure (still obtained from the same assumption) is given by James and Kinderlehrer [11], and this information is in complete agreement with Figure 1 when the calculation is specialized to the measured lattice parameters  $\eta_1(\theta_c)$  and  $\eta_2(\theta_c)$  appropriate to  $Ni_{62}Al_{38}$ .

Note that in the analysis above, it was assumed that two of the matrices ( $F^+$  and  $F^-$ ) differ by a rank-1 matrix. It is still not known whether there exists a microstructure for three matrices with no rank-1 connections, but an example of James and Kohn [12] shows that there exist microstructures with Young measure supported on four matrices having no rank-1 connections.

### 3. The Two-Well Problem

From the point of view of "no geometric restrictions," the crystallographic theory of martensite provides only a weak test of theory. With this in mind Ball and James [4] consider the problem of what are all possible macroscopic deformations that can be realized by microstructures involving just two variants. Physically, these deformations should be associated with flat regions on stress-strain curves, common in materials that undergo reversible martensitic transformations and thought to be associated with rearrangements of the variants. Shape-memory materials are interesting materials which easily rearrange their variants in response to imposed distortions, and it is hoped that such calculations may reveal the reason for this. However, presently only two variants have been considered, while most of the good shape-memory materials have six or twelve variants.

The simplest version of the problem is as follows. The two variants are defined by the set of matrices

$$\mathfrak{M} = SO(3)U_1 \cup SO(3)U_2 \quad (3.1)$$

where  $U_1 = U_1^T > 0$  and  $U_2 = U_2^T > 0$  are distinct  $3 \times 3$  matrices. We assume that  $\det U_2 = \det U_1$  and that there is an  $R \in SO(3)$  such that

$$RU_2 - U_1 = a \otimes n \quad (3.2)$$

These assumptions are satisfied in the typical case of two variants of the less symmetric phase (in which case there is an  $\hat{R}$  belonging to the point group of the more symmetric phase such that  $\hat{R}U_2\hat{R}^T = U_1$ ).

To be definite, we assume that

$$\begin{aligned} U_1 &= \eta_1 \mathbf{1} + (\eta_2 - \eta_1) \mathbf{e}_1 \otimes \mathbf{e}_1, \\ U_2 &= \eta_1 \mathbf{1} + (\eta_2 - \eta_1) \mathbf{e}_2 \otimes \mathbf{e}_2, \end{aligned} \quad (3.3)$$

as in the cubic-to-tetragonal base, although the analysis of [4] operates under the more general assumptions listed above. Consider the problem

$$\inf_{y \in W^{1,\infty}(\Omega)} \int_{\Omega} \phi(Dy(x)) dx, \quad (3.4)$$

where  $\phi$  has strict potential-well minima on  $\mathcal{M}$ . Under mild restriction on  $\phi$ , any minimizing sequence  $y^{(k)} \xrightarrow{*} y$  in  $W^{1,\infty}(\Omega; \mathbb{R}^3)$  has the property that its Young measure  $\nu_x$  is supported on  $\mathcal{M}$ . Some of these sequences will also satisfy the linear boundary conditions\*

$$y^{(k)}(x) = Fx, \quad x \in \partial\Omega. \quad (3.5)$$

The two-well problem is the problem of finding all matrices  $F$  such that  $\text{supp } \nu_x \subset \mathcal{M}$  a.e.. That is, what are all possible macroscopic deformations  $F$  that can be achieved by mixtures of the two variants  $SO(3)U_1$  and  $SO(3)U_2$ ?

The answer to this question is as follows. Let

$$\begin{aligned} \hat{\mathbf{e}}_1 &= (\eta_1^{-1} \mathbf{e}_2 + \eta_2^{-1} \mathbf{e}_1) / (\eta_1^{-2} + \eta_2^{-2})^{1/2}, \\ \hat{\mathbf{e}}_3 &= (\eta_1 \mathbf{e}_2 - \eta_2 \mathbf{e}_1) / (\eta_1^2 + \eta_2^2)^{1/2}, \\ \hat{\mathbf{e}}_2 &= \hat{\mathbf{e}}_3 \times \hat{\mathbf{e}}_1 \\ \delta &= \frac{1}{2} (\eta_2^2 - \eta_1^2) (\eta_1^2 + \eta_2^2)^{-1/2} (\eta_1^{-2} + \eta_2^{-2})^{1/2}. \end{aligned} \quad (3.6)$$

Then,  $F$  can be achieved by a mixture of two variants if and only if

---

\* The analysis of [4] is not restricted to linear boundary conditions.

$$F^T F = U_1 (1 + \delta \hat{e}_1 \otimes \hat{e}_3) C (1 + \delta \hat{e}_3 \otimes \hat{e}_1) U_1, \tag{3.7}$$

where  $C=C^T$  satisfies

$$\begin{aligned} \det C &= 1, \\ C \hat{e}_2 &= \hat{e}_2, \end{aligned} \tag{3.8}$$

and  $C_{11} = \hat{e}_1 \cdot C \hat{e}_1$  and  $C_{33} = \hat{e}_3 \cdot C \hat{e}_3$  lie in the hatched region shown in Figure 2.

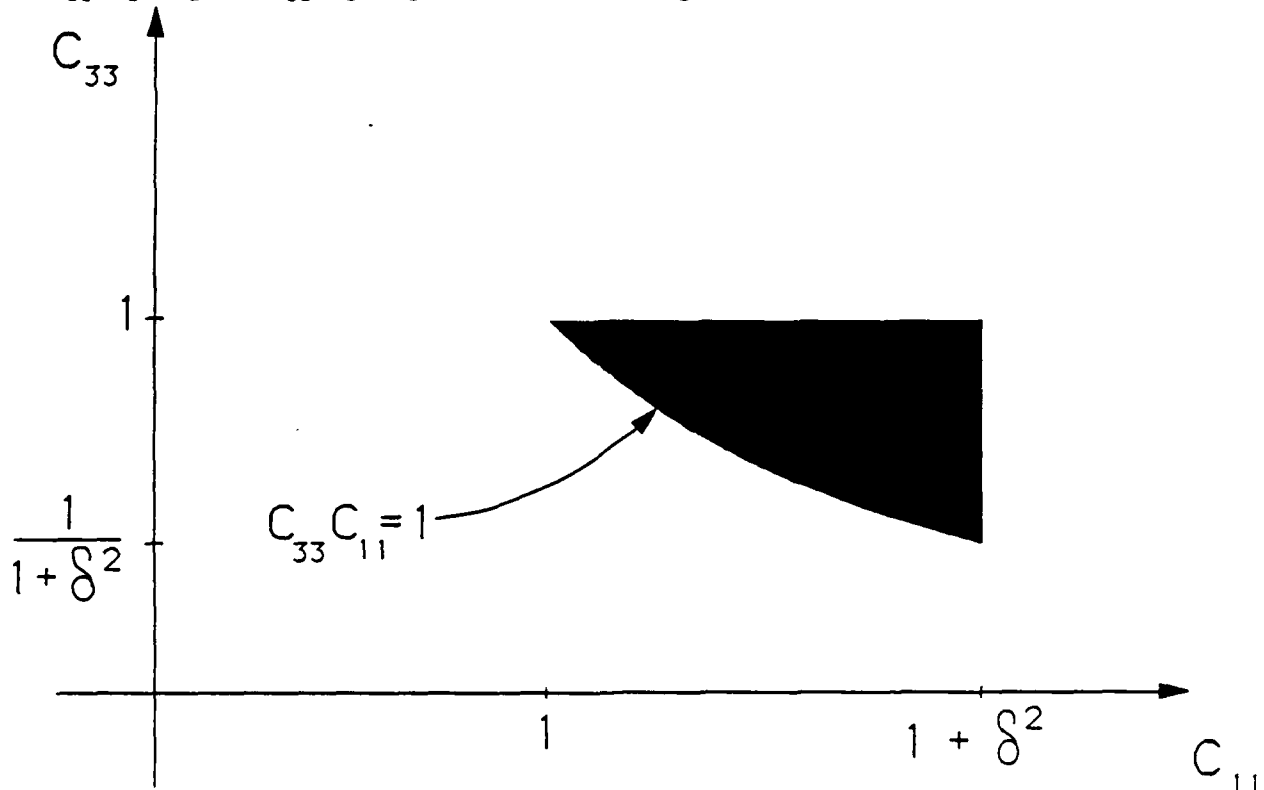


Figure 2. Deformations achievable by mixing two variants.

The proof of this result consists of two parts. The first part makes use of the weakly continuous functions, those functions  $f$  having the property that

$$F(Dy^{(k)}) \overset{*}{\rightharpoonup} f(Dy) \tag{3.9}$$

whenever  $y^{(k)} \overset{*}{\rightharpoonup} y$  in  $W^{1,\infty}(\Omega, \mathbb{R}^3)$ . It is known that the weakly continuous functions for sequences  $y^{(k)} \rightarrow \mathbb{R}^3$  are

$$G, \text{ cof}G, \det G. \tag{3.10}$$

Here,  $\text{cof}G$  denotes the matrix of cofactors of  $G$ . The weakly continuous functions give restrictions on a Young measure with support on  $\mathbb{M}$  of the form

$$\begin{aligned}
 Dy &= \int_{\mathfrak{M}} G dv_x(G) , \\
 \text{cof} Dy &= \int_{\mathfrak{M}} \text{cof} G dv_x(G) , \\
 \det Dy &= \int_{\mathfrak{M}} \det G dv_x(G) ,
 \end{aligned} \tag{3.11}$$

which in turn lead to restrictions on the Young measure like

$$\text{cof} \left[ \int_{\mathfrak{M}} G dv_x(G) \right] = \int_{\mathfrak{M}} \text{cof} G dv_x(G) . \tag{3.12}$$

The first part of the proof exploits (3.12) and the analogous restriction for det to yield the result summarized by Figure 2.

The second half of the proof consists of showing that each point in the domain of Figure 2 is achievable by some microstructure. This follows from an explicit calculation of a family of sequences. The construction proceeds by selecting suitable matrices  $A, B, C$  from  $\mathfrak{M}$  with the properties

$$\begin{aligned}
 \text{rank}(A-B) &= 1 , \\
 \text{rank}[\lambda A + (1-\lambda)B - C] &= 1 ,
 \end{aligned} \tag{3.13}$$

for some  $\lambda \in (0, 1)$ . The conditions (3.13) are sufficient that there be a sequence  $y^{(k)}$ , which essentially describes the "layers within layers" microstructure shown in Figure 3a and satisfies the boundary conditions (3.5).

There is, however, great nonuniqueness in this calculation, and, for example, the microstructures shown in Figure 3b also suffice\*. In Figure 3 we have also shown typical macroscopic deformations associated with these microstructures. Of course, the angles between the layer groups and the relative volume fractions change with  $F$ .

#### 4. Micromagnetics

Ferromagnetic materials often exhibit fine microstructures consisting of magnetic domains. Furthermore, it is of interest in such materials to have methods of relating microscopic to macroscopic properties, both in the case of atomic/microstructural and microstructural/macroscopic properties. Since Maxwell's equations are linear, there is no difficulty averaging solutions unless it is necessary to average nonlinear functions of the solutions, such as the electromagnetic energy. The div-curl lemma and the method of compensated compactness (*e.g.*, Tartar [18]) show which nonlinear functions can be meaningfully averaged. Here, our interest is not in nonlinear functions of the fields, but rather in the nonlinearity introduced by the

---

\* Remark due to D. Kinderlehrer.

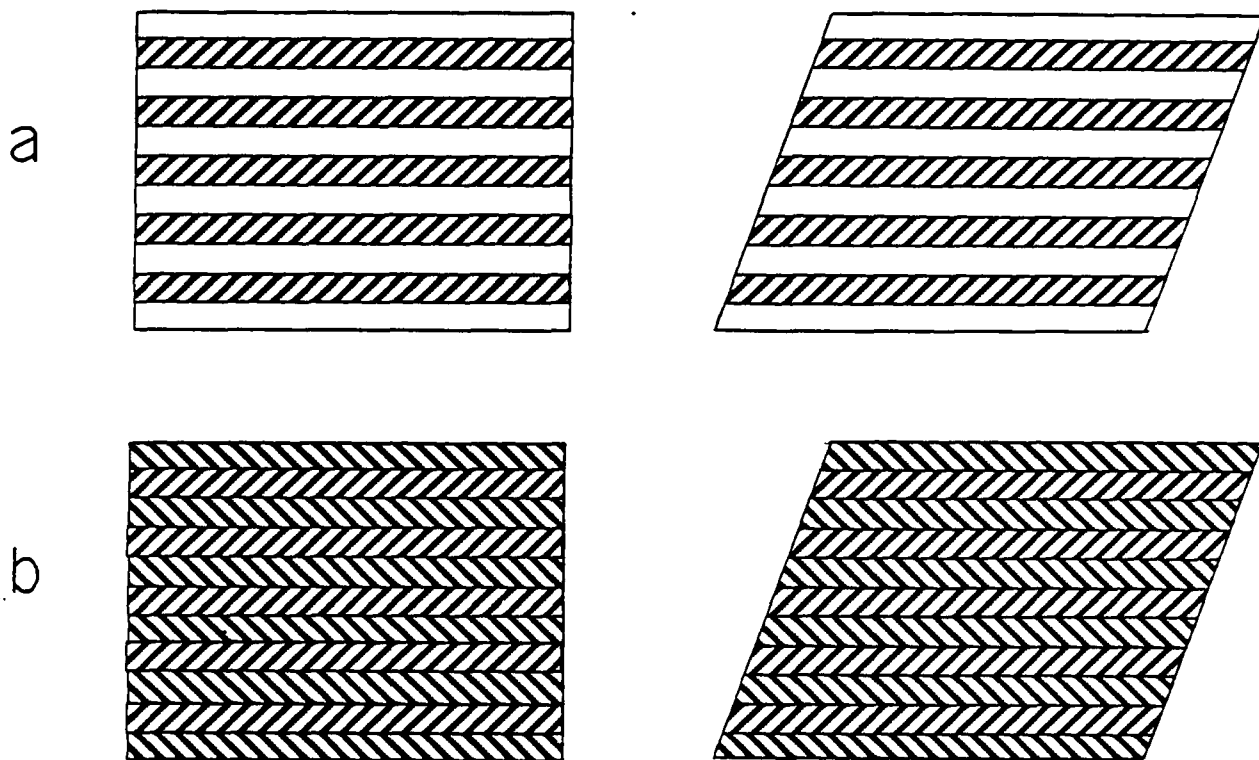


Figure 3. Microstructures which suffice to achieve all possible linear boundary conditions that can be achieved by mixing two variants compatibly.

constitutive properties of ferromagnetic materials. This nonlinearity is responsible for domain structure.

The results of this section are from recent work of James and Kinderlehrer [10]. These results provide an example of the phenomenon of *frustration* (the only rigorous example we know of), this being the phenomenon whereby a material has a defective ground state. Mathematically, the defects arise from the failure of weak convergence to preserve the constraint.

Usually, the total free energy of a ferromagnetic material [14] is given as a sum of exchange energy, anisotropy energy and magnetostatic energy:

$$E[\mathbf{M}] = \int_{\Omega} c|\mathbf{DM}|^2 \, dx + \int_{\Omega} \phi(\mathbf{M}) \, dx + \frac{1}{2} \int_{\mathbb{R}^n} |\mathbf{Du}|^2 \, dx, \tag{4.1}$$

where  $\mathbf{M}:\Omega \rightarrow \alpha(\theta)S^2$ , and the magnetostatic potential  $u$  and  $\mathbf{M}$  are related by

$$\operatorname{div} (-\mathbf{Du} + \mathbf{M}\chi_{\Omega}) = 0 \text{ on } \mathbb{R}^n. \tag{4.2}$$



We hold the temperature constant and therefore put  $\alpha(\theta)=1$  (so  $|M|=1$ ), without loss of generality. It is usually assumed that the anisotropy energy  $\phi$  is even and quadratic in the direction cosines of  $M$  and that  $\phi$  exhibits crystallographic symmetry; we discuss the two cases:

1. Uniaxial  $\phi(M_0)=\phi(-M_0)<\phi(M) \quad \forall M \neq M_0$ ,
2. Cubic  $\phi(\pm M_i)<\phi(M) \quad \forall M \neq M_j$ ,  
 $i, j=1, \dots, n$ ,  
 $(\{M_i\})$  is an orthonormal basis).

The exchange energy can be thought of as giving rise to a surface energy on domain boundaries. The associated calculation is very similar to the calculation of the surface energy on an interface between fluid phases using the van der Waals theory. Anzellotti, Baldo and Visintin [1] give a modern treatment of this calculation; the correct asymptotic scaling can either be obtained from treatments of the van der Waals theory (e.g., Sternberg [17]) or from the famous 1935 paper of Landau and Lifshitz [14]. Typical domain patterns in large bodies reveal a huge surface area of domain walls so we shall temporarily put  $c=0$ .

Hence, we consider the problem

$$\inf_{\substack{M \in L^\infty \\ |M|=1}} \left[ \int_{\Omega} \phi(M(x)) dx + \frac{1}{2} \int_{\mathbb{R}^n} |Du|^2 dx \right], \tag{4.3}$$

subject to

$$\operatorname{div}(-Du + M\chi_{\Omega}) = 0, \quad u \in H^1(\mathbb{R}^n, \mathbb{R}). \tag{4.4}$$

Does (4.3) have an attained absolute minimum? In the uniaxial case, the answer is no if  $\Omega$  is a smooth bounded domain. Intuitively, this can be seen from the following argument. To make both the magnetostatic and anisotropy energies small, we would like to put

$$\begin{aligned} Du(x) &= 0 \quad \text{a.e. } x \in \mathbb{R}^n, \\ M(x) &= \pm M_0 \quad \text{a.e. } x \in \Omega. \end{aligned} \tag{4.5}$$

Equation (4.5)<sub>1</sub> shows that  $M$  should be a divergence-free field in the weak sense. We can construct divergence-free fields  $M=\pm M_0$  on  $\Omega$  by constructing columnar domains with boundaries everywhere parallel to  $M_0$ , but these will not be divergence-free on  $\mathbb{R}^n$  since the boundary condition

$$[[M]] \cdot n = 0 \quad \text{on } \partial\Omega \tag{4.6}$$

is not satisfied at points where the tops of the columns meet  $\partial\Omega$ . However, if we consider finer and finer columns of equal volume (that fill  $\Omega$ ), then the average value of  $M$  will be nearly zero on  $\Omega$  and therefore will approximately satisfy (4.6). Minimizing sequence  $M^{(k)}$  can be constructed using exactly this idea, and for such sequences

$$M^{(k)} \rightharpoonup 0 \text{ in } L^\infty(\Omega, \mathbb{R}^n),$$

$$u^{(k)} \rightarrow 0 \text{ in } L^2(\mathbb{R}^n, \mathbb{R}).$$

The weak limit of  $M^{(k)}$  is zero which does not satisfy the constraint  $|M|=1$ . This minimizing sequence serves to show that the value of the infimum in (4.3) is zero, so any minimizer would have to satisfy (4.4) and (4.5); as might be anticipated from (4.6), the equations (4.4) and (4.5) have no solution  $u \in H^1, M \in L^\infty$ .

The typical microstructure of uniaxial ferromagnets (of mm size or greater) consists of fine columnar domains parallel to the easy axis (*i.e.*, parallel to  $M_0$ ). A huge variety of the cross-sectional shapes is observed.

The cubic case is quite different, as might be anticipated from the textbook picture (Figure 4a) of domains in an iron crystal whose boundaries are (100) planes. Clearly, this picture embodies a minimizer since  $M$  minimizes both the anisotropy energy and is divergence-free on  $\mathbb{R}^3$ . At first, this suggests that the minimum is attained only on special domains  $\Omega$  but Figure 4b suggests otherwise. Figure 4b shows a portion of  $\partial\Omega$  and a divergence-free unit vector field  $M$  on  $\Omega$ . This field gets finer and finer at  $\partial\Omega$  as indicated in the figure. Note that  $M$  averages to zero as  $\partial\Omega$  is approached from inside  $\Omega$ , and it can be shown that for such a vector field,

$$u(x) = 0 \text{ a.e. } x \in \mathbb{R}^2.$$

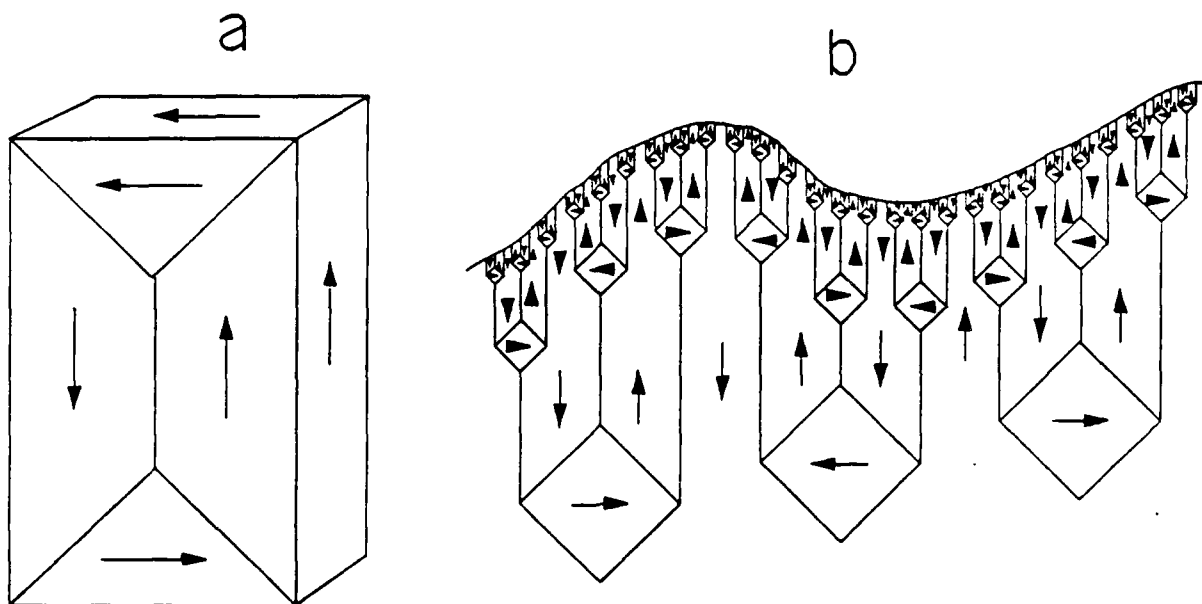


Figure 4. Minimizing domain structures for cubic ferromagnets with exchange energy omitted.

Hence, this vector field  $M \in L^\infty(\Omega, \mathbb{R}^2)$  represents an attained absolute minimum. A more refined argument [10] based on Vitali's Covering Theorem and the domain structure of Figure 4a suffices to prove attainment for any bounded open set  $\Omega$  with  $\text{meas}(\partial\Omega)=0$ .

Domain splitting near the boundary of cubic magnets is common, making the interpretation of observations of domain patterns on  $\partial\Omega$  extremely difficult. However, the phenomenon of often explained from a different perspective by an analysis of Lifshitz [15] which has origins in the magnetostrictive contribution to energy. A discussion of this point can be found in reference [10].

### References

1. G. Anzellotti, S. Baldo and A. Visitin, Asymptotic behavior of the Landau-Lifshitz model of ferromagnetism, preprint.
2. I. Baele, G. van Tendeloo and S. Amelinckx, Microtwinning in Ni-Mn resulting from the  $\beta \rightarrow \theta$  martensitic transformation, *Acta Metall.* **35** (1987), p. 401-412.
3. J. M. Ball and R. D. James, Fine phase mixtures as minimizers of energy, *Arch. Rational Mech. Anal.* **100** (1987), p. 13-52.
4. J. M. Ball and R. D. James, Proposed experimental tests of a theory of fine microstructure and the two-well problem, preprint.
5. J. M. Ball, A version of the fundamental theorem for Young measures, *Conf. on "Partial Differential Equations and Continuum Models of Phase Transitions,"* (ed. D. Serre), Springer-Verlag, to appear.
6. J. S. Bowles and J. K. MacKenzie, The crystallography of martensitic transformations, I and II. *Acta Metall.* **2** (1954), p. 129-137, 138-147.
7. M. Chipot and D. Kinderlehrer, Equilibrium configurations of crystals, *Arch. Rational Mech. Anal.* **103** (1988), p. 237-277.
8. L. M. Brown, *Micromagnetics*, John Wiley and Sons (Interscience), 1963.
9. L. M. Brown, Magnetoelastic Interactions, Springer Tracts in Natural Philosophy 9, Springer-Verlag, 1966.
10. R. D. James and D. Kinderlehrer, Frustration in ferromagnetic materials, preprint.
11. R. D. James and D. Kinderlehrer, Theory of diffusionless phase transformations, *Proc. Conf. on "Partial Differential Equations and Continuum Models of Phase Transitions"* (ed. D. Serre), Springer-Verlag, to appear.
12. R. D. James and R. V. Kohn, Note on Young measures supported on four matrices, preprint.
13. R. V. Kohn, this volume.
14. L. Landau and E. Lifshitz, *Physik. Z. Sowjetunion* **8** (1935), p. 153-169.
15. E. Lifshitz, On the magnetic structure of iron, *J. Physics* **8** (1944), p. 337-346.

16. D. Schryvers, R.U.C.A., Universiteit Antwerpen, personal communication.
17. P. Sternberg, The effect of a singular perturbation on nonconvex variational problems, Ph.D. thesis, New York University (1986).
18. L. Tartar, The compensated compactness method applied to systems of conservation laws, in *Systems of nonlinear partial differential equations* (ed. J. M. Ball), Reidel, 1983.
19. M. S. Wechsler, D. S. Lieberman and T. A. Read, On the theory of formation of martensite. *Trans. AIME J. Metals* 197 (1953), p. 1503–1515.

# CONCURRENT SPECIFICATIONS AND THEIR GUREVICH-HARRINGTON GAMES AND REPRESENTATION OF PROGRAMS AS STRATEGIES

Alexander Yakhnis  
Mathematical Sciences Institute  
Caldwell Hall  
Cornell University  
Ithaca, NY 14853

**ABSTRACT.** We suggest a novel way to view concurrent (possibly perpetually executing) programs. Non-deterministic choice is allowed. We regard program execution as a play of a game of two players, which we call a computational game. One player (Prog, which stands for "programmer") submits sets of instructions for another player (Comp, which stands for "computer") to execute. A program is represented by a strategy of Prog, a program specification is represented by a winning condition. Our approach stems from the work of Rabin, Gurevich and Harrington on S2S, and Buchi on game determinacy. We relate to a programming language a computational game and give two examples of the simplest programs viewed as strategies in such a game. Programming language constructs (including concurrent connection), correspond to operations over strategies producing new strategies. These operations permit to easily relate to each program a strategy that is denoted by it. The operations are defined informally here and more accurately in the sequel to this paper. Here we list properties of such operations over strategies. The properties of the operations allow to do program verification proofs if the program specification is represented as a winning condition for a computational game. We illustrate the program verification by using Park's example. The concurrent program specification requirements of mutual exclusion and absence of lockouts are represented by Gurevich-Harrington winning conditions. These requirements can be verified for any given program using the above properties. The idea of using techniques from the decidability results belongs to Prof. Anil Nerode who also was the first to my knowledge to clearly state that the programs could be understood as strategies in certain games. He also suggested many other valuable ideas.

The sequel to this work, "Extraction of Concurrent Programs from Gurevich-Harrington Strategies," is written by Vladimir Yakhnis and is included in this collection.

## INTRODUCTION.

Game theory is a traditional branch of Mathematics, Logic and, more recently, Computer Science. Von Neumann and Morgenstern in 1930's and 1940's developed finite mathematical games for several players. Games for two players have been put in their present form as "infinite games with perfect information" in the work of Gale and Stewart (1953). Such games have been extensively used in descriptive set theory [M], model theory and mathematical logic. The connection between decidability of certain theories, determinacy of games and automata was explored by Rabin [R], Gurevich and Harrington [GH], and Buchi [B].

The words "strategy" and "program behavior" are often used in the context of computer programming. But this terms also have precise game-theoretic meaning. We systematically interpret such computer programming terms as computer program, execution sequence and program specification, game-theoretically. Then program development and program verification become

precise game theoretical notions of finding a winning strategy or proving that a strategy is winning.

It appears to us that game theoretical meaning of a program as a strategy is more natural than the meanings supplied by other approaches. Interestingly, this does not mean that matters become easy immediately. This is because the relevant game theoretical problems are different than the ones typically considered in the theory of Gale–Stewart games. For example, instead of finding out whether the game is determinate (i.e. whether one of the players wins) we not only want to know who exactly wins and what is the winning strategy, but we also like to know whether it is possible to find a winning strategy in the designated class of strategies.

There are some examples of such theorems in game theory. They are [BL] 1967 and [B] 1983. But the winning strategies, that are described there, are so complex that we were unable to use them in order to produce concrete programs.

In 1982 Gurevich and Harrington published a proof of game determinacy theorem for a class of games which we shall call GH games. This theorem served as a tool in their celebrated short proof of Rabin's theorem. Their proof contains ingenious descriptions of winning strategies, but the strategies are not explicitly given. Using their methods, we developed a sufficient conditions for a given player to win which also gives a wide class of explicit winning strategies. Our purpose is to use these strategies in constructing concurrent programs

We briefly compare our game–theoretical meaning of programs to that of *temporal logic* (Manna & Wolper [MW], Lamport [L1], Gabbay et al [GA], Manna & Pnueli [MP2]), *automata* (Manna & Pnueli [MP1], Alpern & Schneider [AS]), *Gries–Owicki* (Owicki & Gries [OG], Lamport [L2]), and *denotational semantics* (de Bakker & Zucker [BZ]).

We regard any program specification given informally or formally as a winning condition in a game that we associate with a given programming language.

Unlike Manna & Wolper [MW], we do not extract a program from a model of its (temporal) specification. Instead we rely on a theorem yielding

1. Sufficient conditions for a given player to win;
2. A large class of winning strategies in case the conditions hold. A program is then designed from a winning strategy.

Following Buchi [B], we employ automata with output to represent strategies, while Manna & Pnueli [MP1] and Alpern & Sneider [AS] use accepting automata to define execution sequences satisfying the program specification. Automata are not absolutely essential for our present approach and could be replaced by non–deterministic partial strategies.

In the temporal logic, Gries–Owicki and Manna–Pnueli–Alpern–Sneider approaches the meaning of a program is the set of its execution sequences. These approaches provide formalisms for specifying properties of execution sequences and for program verification on that basis.

In contrast, we think that our notions of program as denoting a strategy and an execution sequence of a program as representing a play consistent with the strategy more naturally reflect programming practice and with full mathematical precision. Our notion of play contains more information than that of execution sequence, for example information about computer delaying the execution of submitted instructions. This information permits us to define interleaving naturally, and to specify a wider variety of properties for the program execution process. The essential ingredient of our approach is the use of theorems for finding winning strategies in classes of games that arise from program specifications.

Denotational semantics views a program as a function over a mathematical structure designed for a given programming language. Our approach shares this feature. A program is represented by a function (i.e. a strategy) over a game tree, which is a mathematical structure encapsulating rules of the

game corresponding to a programming language. In the denotational semantics we have seen thus far the mathematical structures which give the meaning to concurrency are so complex that even reasonably simple programs and proofs have cumbersome denotations. We think that our approach gives a simple and very intuitive denotation.

## 1. PROGRAMS AS STRATEGIES

Gale-Stuart games are played by two *players*. We consider in the present paper the following version of their games which we call *computational games*. The plays of a game can be either infinitely long or finite. We shall call the players *Prog* and *Comp*. They alternate in making moves. Prog plays first. A move of a player is to choose and append a letter  $a \in \Sigma$  of a given alphabet  $\Sigma$  to a sequence obtained from previous moves. The resulting sequence of moves is a *play* of a game:

$$a_0, a_1, a_2, \dots, a_n, \dots$$

A finite initial segment of a play is called a *position* of the game. The set  $T$  of all positions is called the *game tree*. Positions where Prog (Comp) makes a move are called  $\text{Pos}(P)$  ( $\text{Pos}(C)$ ). They are the positions with even (odd) length. The game must specify rules restricting possible moves of players and a winning condition. We introduce the rules gradually in subsequent examples. A *winning condition* or a *winning set* for a given player is a collection of plays satisfying all the restrictions on moves. We say that a player *wins a play* if it lies in the winning set of the player. A play is finite if Prog made a special move end. This is the last move of a play. If a play does not include the move end, it is infinite.

The usual intuition describes a program as some orderly way of submission of instructions for a computer to execute. In our framework each of the submissions constitute a move of Prog and each execution of an instruction is a move of Comp. Informally, a Prog's move is a set of instructions (as opposed to a singleton in deterministic systems). An empty set of instructions is called skip. A Comp's move indicates which instruction (if any) has just been executed. The move corresponding to the absence of executions is called wait. In contrast to a Prog's move, only execution of at most one instruction is allowed. Multiple executions are simulated by all possible orderings of the respective consecutive moves.

However apart from the instructions, each program contains a number of directives governing the order in which the instructions are executed. In the case of a sequential program these directives are the actual order of the instructions, go to's, the structured statements like *if...then...else*, *while...do* and so on. In the case of a concurrent program they are *cobegin...coend*'s, *par*'s, *fork*'s, implicit directives contained in *semaphores* and so on. In our framework these directives govern the behavior of Prog in the course of a play and therefore constitute a strategy for Prog.

A *strategy* for a given player is a function from all positions of the player, into the set of moves allowed for the player. We say that a player *uses* a strategy  $f$  from some position  $p$  on, if at any later position  $q$  of a play, where the player has to make his move, the player chooses a move from the set  $f(q)$ . We also say that a play, where a player uses a strategy  $f$  from some position  $p$  on, is *consistent* with the strategy at  $p$ . We consider the "*state-strategies*" developed by Buchi and the "*strategies with restricted memory*" developed by Gurevich and Harrington.

We shall show how to find for each program a finite state-strategy for Prog which represents precisely the directives for the order of computations prescribed by the program. We think of this finite state-strategy as the *meaning* of the program.

We relate to every program building construct an operation over state-strategies. This is illustrated by examples.

Suppose that we are given a PASCAL-like programming language  $L$ . Every program which we consider consists of assignments, constructors **if ... then ... else's**, **while ... do's**, and a concurrent constructor. Concurrent constructs have the form  $\text{cobegin } P_1; \dots; P_n \text{ coend}$ , where  $P_1, \dots, P_n$  are programming blocks, i.e. either subroutines, blocks **begin ... end** or single assignments. In distinction to the other approaches, we do not make any assumption about the appearance or nesting of the concurrent constructs in the programs. They may appear anywhere and may be nested in any possible way.

Let  $A$  be a finite set of instructions from  $L$  including  $x:=1$ .

EXAMPLE 1. Suppose that the program **begin  $x:=1$  end** is being executed. We define the following computational game. The Prog's moves constitute the set  $\mathcal{P}(A) \cup \{\text{end}\}$ , called Prog alphabet  $\Sigma^P$ . We shall often use the notation **skip** for the empty move of Prog. The Comp moves constitute the set  $A \cup \{\text{wait}\}$ , called Comp alphabet  $\Sigma^C$ . The (disjoint) sum  $\Sigma^P \cup \Sigma^C$  is called the *game alphabet*  $\Sigma$ . The following sequence is an example of a play.

**$\{x:=1\}$ , wait, skip, wait, skip,  $x:=1$ , end**

The following two rules are restricting the moves of players.

(Rule 1) Each Comp move other than wait has to be a member of some previous Prog move, s.t. Comp has not yet used in his previous moves. We call the set of all such permissible Comp moves at a position  $p$  by  $\text{Avail}(p)$ .

(Rule 2) Any move of Prog at a position  $p$  must be disjoint with  $\text{Avail}(p)$ . I.e. Prog may not include an instruction in his move if he has submitted this instruction previously and Comp has not executed it yet.

For the play above the function  $\text{Avail}(p)$  assumes the following values at positions of the play (beginning with the root).

$\emptyset, \{x:=1\}, \{x:=1\}, \{x:=1\}, \{x:=1\}, \{x:=1\}, \emptyset, \emptyset$

Prog and Comp moves may be defined by means of a special Moore automaton called *Exec*, which models the states of an operational system and a computer memory. The automaton alphabet is  $\Sigma$ . A state of *Exec* has two components. The first component, called a *machine state* is an assignment of values to all variables occurring in  $A$ . We may assume for example that the values are rational numbers. The set of all machine states we designate  $S^E$ . The second component is a subset of  $A$ . It is intended to represent  $\text{Avail}(p)$ . So the set of *Exec* states is  $S^E \times \mathcal{P}(A)$ . The set of initial states is  $S^E \times \{\emptyset\}$ . The transition table  $M$  is described as follows. If  $(s, U) \in S^E \times \mathcal{P}(A)$  and  $\sigma \in \Sigma^P$ , then  $M((s, U), \sigma) = (s, U \cup \sigma)$ , where **end** is identified with  $\emptyset$ . If  $\delta \in \Sigma^C$ , then  $M((s, U), \delta) = (s', U - \{\delta\})$ , where wait is identified with  $\emptyset$  and  $s'$  is a machine state that results from the execution of the instruction  $\delta$  in the machine state  $s$ .

There are two output functions defined on states of *Exec*.  $f^P(s, U) = \mathcal{P}(A - U) \cup \{\text{end}\}$  and  $f^C(s, U) = U \cup \{\text{wait}\}$  which represent vacuous strategies for Prog and Comp respectively. The value of the vacuous strategy for a player ( $\text{Vac}^P$  for Prog and  $\text{Vac}^C$  for Comp) at a position  $p$  is the set of all possible moves the player can make according to the game rules. If  $(s, U)$  is the last state of the *Exec*'s run on  $p$ , then  $\text{Vac}^P(p) = f^P(s, U)$  and  $\text{Vac}^C(p) = f^C(s, U)$ , depending on whether  $p \in \text{Pos}(P)$  or  $p \in \text{Pos}(C)$ . In fact  $f^C$  is somewhat more complicated than we just described. We shall give more precise explanations later (Section 2).



If  $s$  is a machine state,  $y$  is a variable and  $v$  is some value then  $s[y/v]$  is the state which assigns to all distinct from  $y$  variables the same value as does  $s$  and which assigns  $v$  to  $y$ .

Let  $(s, \emptyset)$  be an initial Exec's state. Then the  $(s, \emptyset)$ -run of Exec over the above play is  $(s, \emptyset), (s, \{x:=1\}), (s, \{x:=1\}), (s, \{x:=1\}), (s, \{x:=1\}), (s, \{x:=1\}), (s[x/1], \emptyset), (s[x/1], \emptyset)$ .

Call an Exec's run over a position (or a play) an Exec run over the position (or the play) as a word in the game alphabet.

Now we shall explain the notion of a non-deterministic state-strategy for a player  $\Omega$ . Let  $\mathcal{Q}$  be a Moore automaton with input over the game alphabet  $\Sigma$ , with the set of states  $S$ , set of initial states  $S_{in}$ , a deterministic transition table  $M$  and the output function  $F: S \rightarrow \mathcal{P}(\Sigma^\Omega)$ . If  $s_0 \in S_{in}$  then  $\mathcal{Q}$  and  $s_0$  induce the following strategy  $f$ . Let  $p \in \text{Pos}(\Omega)$ . Run  $\mathcal{Q}$  on  $p$  from  $s_0$ . If  $s$  is the last state of the run, take  $f(p) = F(s)$ . It is easy to see that  $\text{Vac}^P$  and  $\text{Vac}^C$  above are non-deterministic state-strategies.

For a deterministic state-strategies the above is simplified since it is sufficient for  $\mathcal{Q}$  in this case to take as an input only the moves of the opponent.

Informally, the program `begin x:=1 end` represents the following Prog's behavior in the computational game:

1. Submit the set of of instructions  $\{x:=1\}$ ;
2. Wait until the instruction is completed, by making skip move;
3. Finish the play by submitting end move.

We shall describe a deterministic state-strategy corresponding to `begin x:=1 end` by giving its state diagram.

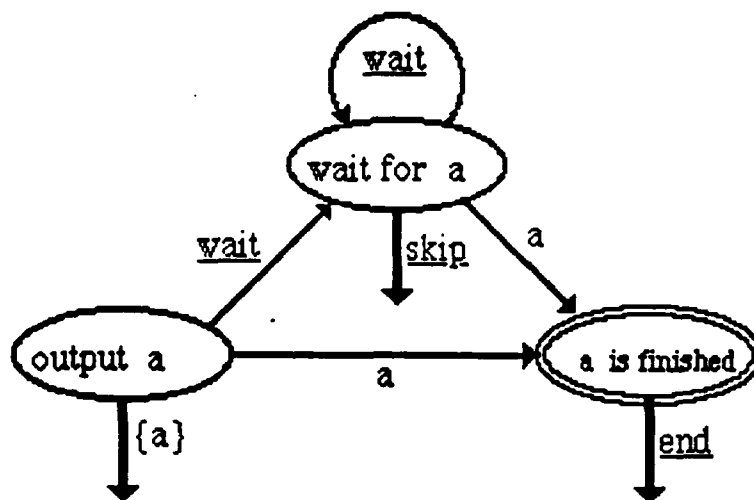


FIG 1.

On Fig. 1  $a$  represents `begin x:=1 end`, ovals represent states, thin arrows represent the input and thick arrows represent the output.

Let  $W$  be a collection of plays and  $T$  be the game tree.  $\Gamma = \langle T, \Omega, W \rangle$  designates a game where  $\Omega$  wins a play  $\mu$  iff  $\mu \in W$  and, conversely,  $1-\Omega$  wins a play  $\mu$  iff  $\mu \notin W$ .  $W$  is called the *winning condition* for  $\Omega$ . If the complement of  $W$  is designated  $W^c = \text{Play}(T) - W$  then we can also write  $\Gamma = \langle T, 1-\Omega, W^c \rangle$ .

Now we shall reduce a program specification for the example to the notion of a winning condition. Let  $(s, \emptyset)$  be an initial Exec's state. The program specification "the program terminates

with  $x=1$ " is translated as "the set of all finite (i.e. terminated by end) plays  $\mu$  s.t. the last machine state on  $\mu$  resulting from the  $(s, \emptyset)$ -run of Exec on  $\mu$  is  $s[x/1]$ ." Let us call this set of plays  $W^a$ , where  $a$  stands for begin  $x:=1$  end.

We adopt the following convention, unless we say otherwise. In all computational games Comp loses every infinite play, where he refuses to execute an instruction, submitted by Prog. More precisely, Comp loses every infinite play, for which there is an instruction which is contained in the set Avail( $q$ ) for all  $q$  after some position  $p$ . Let  $W^r$  be the set of all such plays. Here  $r$  stands for Comp refusal to complete some submitted instruction.

Therefore we represent the program specification above by the game  $\Gamma = \langle T, \text{Prog}, W^a \cup W^r \rangle$ .

Let  $f$  be an  $\Omega$ -strategy. We say that  $f$  is *perpetual* at  $p$  if for any nonterminal position  $q \in \text{Pos}(\Omega)$  consistent from  $p$  with  $f$  we have  $f(q) \neq \emptyset$ . We say that  $f$  is *conditionally winning* at  $p$  a game  $\Gamma$ , if every play, containing  $p$  and consistent after  $p$  with  $f$  is in  $W$ . Finally, we say that  $f$  is *winning  $\Gamma$  at  $p$* , if  $f$  is perpetual at  $p$  and  $f$  is conditionally winning  $\Gamma$  at  $p$ . If  $p$  is the root  $e$ , then we omit references to positions in the above definitions. Lachlan in [LAC] 1970 also used the notion of a perpetual strategy, though he named it differently.

It is easy to see that the strategy begin  $x:=1$  end wins  $\Gamma$ .

EXAMPLE 2. We begin with the program cobegin  $x:=1, x:=2$  coend. We assume that  $x:=1$  and  $x:=2$  are in  $A$ , and so Prog's and Comp's alphabets are the same as in the example 1.

We shall use abbreviations  $a$  for  $x:=1$ ,  $b$  for  $x:=2$ .

The program represents the following Programmer behavior in the computational game.

1. Submit the the set of of instructions  $\{x:=1, x:=2\}$ .
2. Wait until each submitted instruction is completed, by making skip move.
3. Finish the play by submitting end move.

Below is its state diagram.

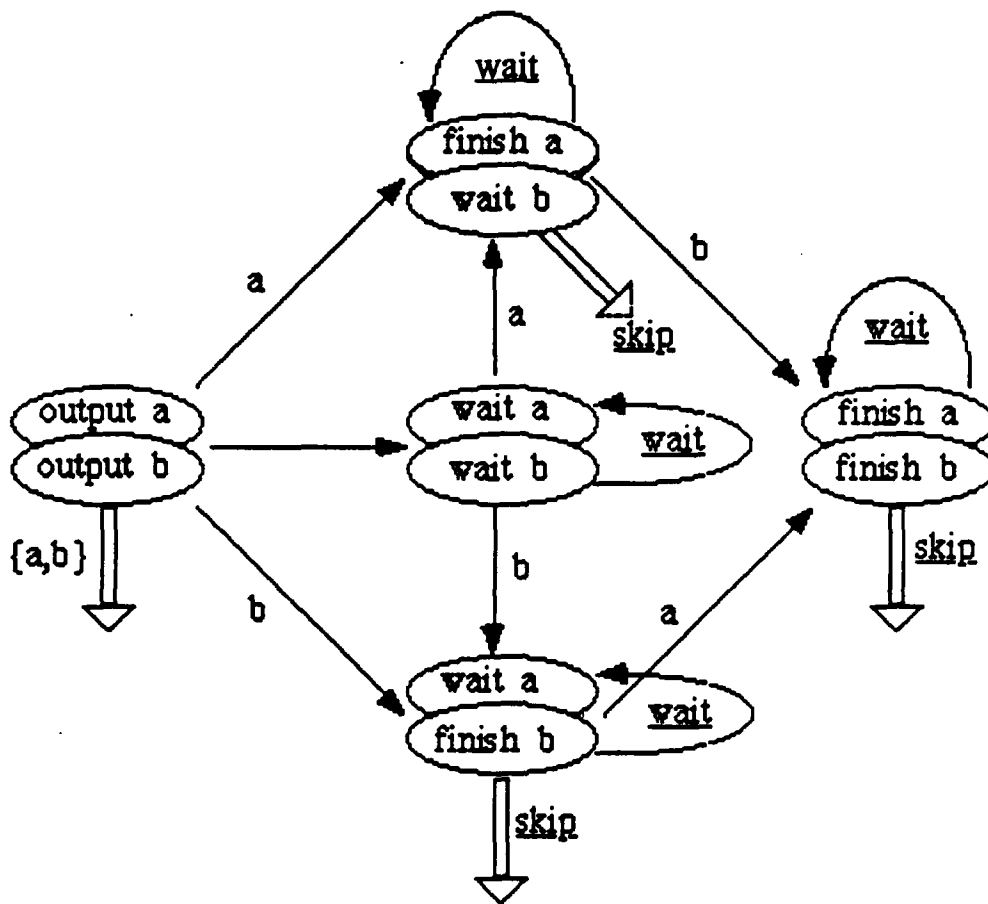


FIG. 2

On the diagram above, big arrows correspond to the output (i.e. moves of Prog) and small arrows correspond to the input (i.e. moves of Comp).

Now, let  $W^b$  be defined by replacing  $s[x/1]$  with  $s[x/2]$  in the definition of  $W^a$  in example 1. Then the program specification "the program terminates with  $x=1$  or  $x=2$ " corresponds to the game  $\langle T, \text{Prog}, W^a \cup W^b \cup W^r \rangle$ . It is easy to see that the above strategy wins this game.

The consideration of the last example is based on the notion of a mutual atomicity. Suppose that  $A_1, \dots, A_n$  are respectively the collections of assignments from the blocks  $P_1, \dots, P_n$  from the concurrent construct above. We assume here, as in other models of concurrency, that  $A_1, \dots, A_n$  must be mutually atomic. In the simplest case, the assignments  $a_1$  and  $a_2$  are called *mutually atomic* if the result of a concurrent execution of  $a_1$  and  $a_2$  is always the same as that of the sequential execution of either  $a_1; a_2$  or  $a_2; a_1$ .

Assume that  $x=0$  and that `cobegin  $x:=x+1$ ;  $x:=x+1$  coend` is being executed. Suppose that there are two processors each of which independently computes  $x:=x+1$ . Assume that any assignment consists of two separate steps, i.e. (1) computing of the right-hand part and (2) putting the result into the location which corresponds to the left-hand part. Then the following scenario is feasible. At moment  $t_1$  both processors compute  $x+1$  (which is equal to 1), at moment  $t_2$  the first processor puts the result in  $x$  and at moment  $t_3$  the second processor puts the result in  $x$ . It is easy to see that after all these operations are finished, we have  $x=1$ . Since `begin  $x:=x+1$ ;  $x:=x+1$  end` gives us

$x=2$ , we must conclude that  $x:=x+1$  and  $x:=x+1$  are not mutually atomic.

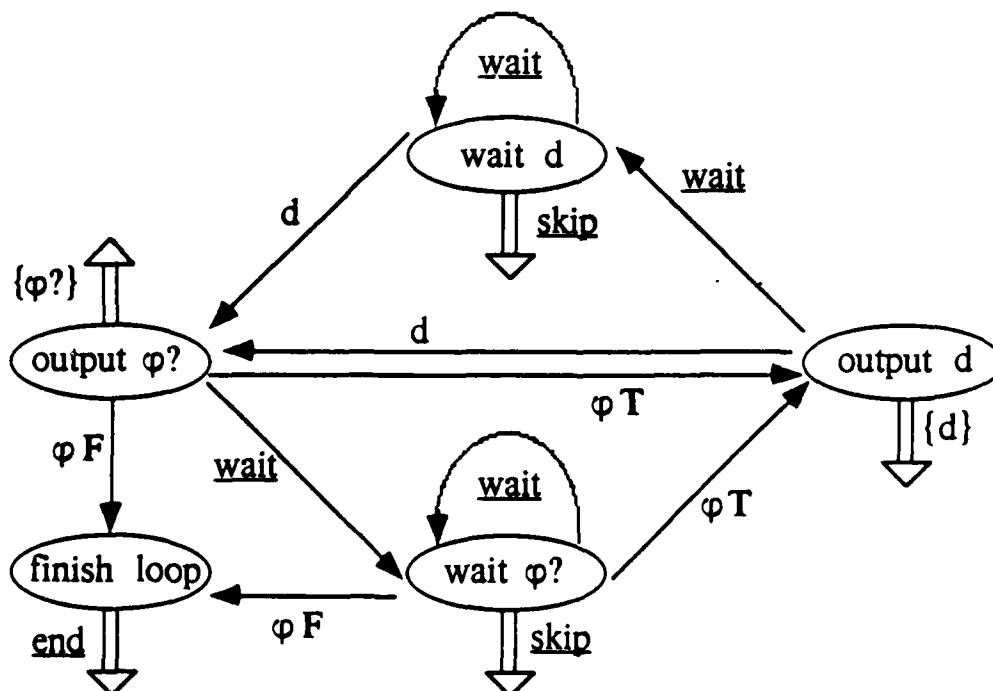
Our game-theoretical approach allows us to deal with programs without the assumption of mutual atomicity as above. However, the model in this case is more complicated and we shall not consider it here.

## 2. VERIFICATION OF THE PARK'S EXAMPLE.

This section is intended to give an informal example of a direct verification proof based on game-theoretic notions. This proof ultimately rests on the material of the next section, providing basis for any proofs of this sort. But there are other possibilities to do game-theoretic verification.

Since the Park's example involves a predicate  $x=0$ , we extend a computational game alphabet to account for predicates. We often use the term instruction for either a predicate or true instruction. Let  $\Phi$  be a collection of predicates admissible by the programming language. The Prog's moves constitute the set  $\Sigma^P = \mathcal{P}(A \cup \Phi) \cup \{\text{end}\}$ . The notation skip is still used for the empty move of Prog. The Comp moves constitute the set  $\Sigma^C = A \cup (\Phi \times \{t, f\}) \cup \{\text{wait}\}$ , based on the set  $\{t, f\}$  of truth values. The game alphabet is  $\Sigma = \Sigma^P \cup \Sigma^C$ . The Rules 1 and 2 of Section 1 are easily applicable in the present context, if the references there to Comp's moves, which include predicates, are understood as referring only to the predicate component of the move. The Exec's alphabet and transition table are extended to account for the larger game alphabet as follows. If  $\delta = \langle \varphi, b \rangle$  is a Comp's move, then  $M((s, U), \delta) = (s, U - \{\varphi\})$ , where  $\varphi$  holds in a machine state  $s$  iff  $b=t$ . I.e. a Comp's move, containing a predicate, also contains its truth value at an Exec's machine state  $s$  at which the move has been made.

The program of the form **while**  $\varphi$  **do**  $d$ , where  $d$  is an instruction, is used in the example. The following state-diagram describes the respective state-strategy.



Note that with each state-strategy  $u$  there is an naturally associated alphabet  $\Sigma^u$ , which is a union of its input and output alphabets. This permits us to say that a move occurs in  $u$  meaning that it occurs in the above alphabet of  $u$ .

We consider the program  $g = \text{begin } x:=0; y:=0; \text{par}(\text{begin } x:=1 \text{ end}, \text{while } x=0 \text{ do } y:=y+1) \text{ end}$ . It is required to show that  $g$  terminates in an Exec's state  $s$  satisfying  $(x=1 \text{ and } \exists n \in \omega \ y=n)$ . Call the latter predicate  $\psi(x,y)$ . We shall restate this as the winning condition of the computational game. The computational game alphabet can be restricted only to be based on  $A = \{x:=0, x:=1, y:=y+1\}$  and  $\Phi = \{x=0\}$ . Let  $s(p)$  denote the Executive state arising after the last Comp's move of a position  $p$ . Let  $p^-$  denote a position obtained from  $p$  by removing its last move. Let  $W = \{\mu: \mu \text{ is a terminal play and } s(\mu^-) \models \psi(x,y)\}$ . Let  $W^f$  be the Computer refusal set for the game. This is the set of all infinite plays satisfying the condition, that for every play from some position of it there is an item always occurring in a set Avail and never occurring in all subsequent Computer moves. Computer always loses the set  $W^f$ . So we are to show that  $g$  wins the set  $W^f \cup W$ .

**PROPOSITION 2.1.** The strategy  $g$  wins the set  $W^f \cup W$ .

**Proof.** Note that  $g$  is a perpetual strategy, because the strategy submits only instructions defined on all Exec's machine states. It remains to show that  $g$  is conditionally winning. It is sufficient to show that if  $\mu$  is a play, where Prog uses  $g$ , and is not in  $W^f$  then  $\mu$  is in  $W$ . The use of  $g$  involves first the use of the strategy  $h = \text{begin } x:=0; y:=0 \text{ end}$ , it follows that there is a position  $p$  of  $\mu$  where  $h$  is used last and  $s(p) \models (x=0 \text{ and } y=0)$ . We may write  $\mu = p \cdot \eta$ , where  $\eta$  is a play that begins at  $p$ .

(P1) The initial Exec's state  $s_0(\eta)$  for a play  $\eta$  is  $s(p)$ . Hence it satisfies  $s_0(\eta) \models (x=0 \text{ and } y=0)$ .

We shall show that a play  $\eta$  has a position  $q$  satisfying

(P2) all positions  $r$  of  $\eta$  following (and including)  $q$  satisfy  $s(r) \models (x=1)$  and

(P3) all positions  $r$  of  $\eta$  (strictly) preceding  $q$  satisfy  $s(r) \models (x=0)$ .

These will be used to show that  $\eta$  is a finite play. It then immediately follows from (P2) that the terminating Exec's state satisfies the first conjunct of  $\psi(x,y)$ . Then we shall show that this Exec's state satisfy the remaining conjunct also.

Since immediately after the use of  $h$  has been completed,  $g$  submits the instruction  $x:=1$  and  $\mu \notin W^f$ , it follows that Comp makes a move  $\delta = (x:=1)$  in  $\mu$  after  $p$ , i.e.  $\delta$  occurs in  $\eta$ . It immediately follows that (P2) and (P3) hold at a position  $q$  of  $\eta$  arising after this move, because there are no other moves involved in  $g$ , that can affect the variable  $x$ .

Observe that Prog uses the strategy  $\text{par}(\ , \ )$  along  $\eta$ . Denote  $u = \text{begin } x:=1 \text{ end}$  and  $v = \text{while } x=0 \text{ do } y:=y+1$ . It can be noticed that all Prog's moves in  $\eta$  occur either in  $u$  or in  $v$ . If Comp's moves from  $u$  are replaced by wait and Prog's moves from  $u$  are deleted from  $\eta$ , the resulting play is such that Prog uses  $u$  in it. It follows from (P1) that this play is terminating. Hence  $\eta$  is terminating. So is  $\mu$ .

It remains to show that  $s(\eta^-) \models (y \in \omega)$ . We'll show by induction that this holds for all prefixes of  $\eta^-$ . The induction base holds due to (P1). Every Comp's move in  $\eta$  either does not effect the variable  $y$  or is an instruction  $y:=y+1$ , which preserves the desired property. This completes the

induction and the proof. ■

### 3. PROPERTIES OF OPERATIONS OVER STATE-STRATEGIES.

The operations over state-strategies corresponding to program connecting constructs: sequential connection of programs, concurrent connection of programs, conditional connection of programs and conditional repetition of a program are defined in the sequel to this paper. Here we state the properties of the operations over state-strategies. The operations over strategies are named similar to respective program connecting constructs.

We need less restrictive notion of a game than a computational game, to state the properties of concurrent connection of strategies. We call such a game a *free computational game*. It differs from computational game in omitting all references to the automaton Exec in the game rules. Note that Rule 1 and 2 are still valid for a free computational game.

We first define a split of any position of a *free game* in respect to (Prog's) strategies  $g$  and  $h$  and an arbitrary position  $q$  of the game. Let  $\Sigma^g$  and  $\Sigma^h$  be the alphabets associated with the strategies as explained in Section 2. We assume these alphabets to be disjoint. A split of a letter  $\delta$  from the Computer alphabet  $\Sigma^C$  are two unique letters:  $Sp^g(\delta) = (\text{if } \delta \in \Sigma^g \text{ then } \delta \text{ else wait})$  and  $Sp^h(\delta) = (\text{if } \delta \in \Sigma^h \text{ then } \delta \text{ else wait})$ . A split of a letter  $\sigma$  from the Programmer alphabet  $\Sigma^P$  are two unique letters:  $Sp^g(\sigma) = \sigma \cap \Sigma^g$  and  $Sp^h(\sigma) = \sigma \cap \Sigma^h$ . If  $r$  is any word let  $Sp^g(r)$  be a word obtained from  $r$  by replacing any its letter by  $Sp^g$  of it. If  $r = q \cdot r'$  call  $Sp^g(r) = q \cdot Sp^g(r')$  and  $Sp^h(r) = q \cdot Sp^h(r')$ .

For any play  $\mu$  and its position  $r$  call  $\mu_r$  a play arising by deleting the prefix  $r$  from  $\mu$ . For any natural number  $n$  let  $\mu_n$  be a play obtained by deleting a prefix of length  $n$  from  $\mu$ .

#### PROPOSITION 3.1

A. If a play  $\mu$  of a free computational game is consistent with concurrent connection of two strategies  $g$  and  $h$  then there are two plays  $\mu'$  and  $\mu''$  consistent with  $g$  and  $h$  respectively and in the respective free games and such that

1. a play  $\mu$  is terminal if and only if  $\mu'$  and  $\mu''$  are terminal.
2.  $\mu'$  and  $\mu''$  are both not terminal  $\Rightarrow$  for every  $n \in \omega$   $\mu'(n) = Sp^g(\mu(n))$ ,  
 $\mu''(n) = Sp^h(\mu(n))$ .
3.  $\mu'$  is terminal,  $\mu''$  is not  $\Rightarrow$  there is a prefix  $r$  of  $\mu$  s.t.  $(\mu')^- = Sp^g(r)$ ,  
 $\mu''_{\text{length}(r)} = \mu_r$  for every  $n < \text{length}(r)$   $\mu''(n) = Sp^h(\mu(n))$ .
4.  $\mu''$  is terminal,  $\mu'$  is not  $\Rightarrow$  there is a prefix  $r$  of  $\mu$  s.t.  $(\mu'')^- = Sp^g(r)$ ,  
 $\mu'_{\text{length}(r)} = \mu_r$  for every  $n < \text{length}(r)$   $\mu'(n) = Sp^g(\mu(n))$ .

Below  $\mu'$  and  $\mu''$  are both terminal. Then one of the following three possibilities holds.

5.  $(\mu')^- = Sp^g(\mu^-)$ ,  $(\mu'')^- = Sp^h(\mu^-)$ .
6.  $(\mu')^- = Sp^g(r)$ ,  $(\mu'')^-_{\text{length}(r)} = (\mu^-)_r$  for some proper prefix  $r$  of  $\mu^-$  and for every  $n < \text{length}(r)$   $\mu''(n) = Sp^h(\mu(n))$ .

7.  $(\mu'')^- = \text{Sp}^h(r^-)$ ,  $(\mu')^-_{\text{length}(r)} = (\mu^-)_r$  for some proper prefix  $r$  of  $\mu^-$  and for every  $n < \text{length}(r)$   $\mu'(n) = \text{Sp}^g(\mu(n))$ .

B. For every state-strategies  $g, h, u$   $(g||h)||u$  is equivalent to  $g|(h||u)$  in a free computational game ■

**Sequential connection of two strategies.** The sequential use of two strategies  $h$  and  $g$ , written  $h; g$ , is defined when for any final state of  $h$  if any state that can shift into a final state it can shift only into final states. The strategy  $h; g$  is a strategy whose use consists in the use of the first strategy until this strategy reaches (if ever) its final state in a course of a play, at this position of a play the final state is forgotten and Programmer uses the second strategy from its initial state.

**PROPOSITION 3.2**

A. If a play  $\mu$  of a free computational game is consistent with a sequential connection of strategies  $h; g$  at a position  $p$  then either

1.  $\mu$  is infinite and is consistent with  $h$  from  $p$  or
2.  $\mu = \eta^- \cdot \xi$  for some finite play  $\eta$  consistent with  $h$  from  $p$  and some play  $\xi$  consistent with  $g$  from the root. Also the initial Executive state for a play  $\xi$  has to coincide with an element of  $s(\eta^-)$ .

B. For any state-strategies  $g, h, u$   $(g \cdot h) \cdot u$  is equivalent to  $g \cdot (h \cdot u)$ . ■

**The conditional connection of two strategies.** The strategy  $\varphi?$  evaluates the predicate  $\varphi$  and memorizes the obtained truth values by terminating in two distinct states.

It is convenient to consider two related strategies. The first strategy is denoted  $\varphi^t?$  It differs from the strategy  $\varphi?$  by being undefined whenever a Comp's move is not  $\langle \varphi, t \rangle$  or wait. The second strategy is denoted  $\varphi^f?$  It differs from the strategy  $\varphi?$  by being undefined whenever a Comp's move is not  $\langle \varphi, f \rangle$  or wait.

Conditional use of strategies  $g, h$  depending on the predicate  $\varphi$  beginning from a given position  $p$  consists of the following. Use the strategy  $\varphi?$  from  $p$ . If the final state, reached by  $\varphi?$  in the course of its use, corresponds to true outcome, use  $g$  from this position on, if the final state corresponds to the outcome false use  $h$ . We denote this strategy by  $\text{if } \varphi \text{ then } g \text{ else } h$ .

The following property characterizes the conditional use of state-strategies.

**PROPOSITION 3.3** A play  $\mu$  of a free computational game is consistent with the state-strategy  $\text{if } \varphi \text{ then } g \text{ else } h$  from a position  $q$  iff

$\mu = \eta^- \cdot x \cdot \xi$ , where  $\eta$  is terminal and is consistent with  $\varphi?$  from  $q$ ,  $x$  is a Comp's move and  $x = \langle \varphi, t \rangle$  or  $\langle \varphi, f \rangle$  and either

- a.  $\xi$  is consistent with  $g$  at root and  $x = \langle \varphi, t \rangle$  or
- b.  $\xi$  is consistent with  $h$  at root and  $x = \langle \varphi, f \rangle$ . ■

**Use of a strategy while certain condition holds.** The strategy which uses a given strategy  $g$  while a predicate  $\varphi$  holds is denoted by **while**  $\varphi$  **do**  $g$ . Its use in a play consists in a use of  $\varphi?$  and termination, if the final state of  $\varphi?$  is reached, corresponding Comp's evaluation of the predicate to false. Otherwise,  $g$  is used sequentially after  $\varphi?$ . If  $g$  reaches one of its final states,  $\varphi?$  is used again and the preceding part of the description applies.

The characteristic property of a repetitive use of a strategy follows.

**PROPOSITION 3.4** A play  $\mu$  of a free computational game is consistent with the strategy **while**

$\varphi$  do  $g$  from  $q$  iff

1. there is  $n \geq 1$  and plays  $\xi_1, \xi_2, \dots, \xi_n$  such that  $\mu = \xi_1^- \cdot \xi_2^- \cdot \dots \cdot \xi_{n-1}^- \cdot \xi_n^-$  and for  $i < n$  each  $\xi_i$  is consistent with  $\varphi^t \cdot g$  and is terminal, and  $\xi_n$  is either consistent with  $\varphi^f \cdot g$  or with  $\varphi^t \cdot g$  and in the latter case  $\xi_n$  is infinite or
2. there is infinite sequence of finite plays  $\xi_1, \xi_2, \dots, \xi_n, \dots$  such that each  $\xi_i = \delta^- \cdot \kappa^-$  for some finite play  $\delta$  consistent with  $\varphi^t \cdot g$  from the root and some finite play  $\kappa$  consistent with  $g$  from the root (both  $\delta$  and  $\kappa$  depend on  $\xi_i$ ) and  $\mu = \xi_1^- \cdot \xi_2^- \cdot \dots \cdot \xi_n^- \cdot \dots$  ■

#### 4. GUREVICH-HARRINGTON GAMES AND THE MUTUAL EXCLUSION PROBLEM

Gurevich and Harrington (GH) considered winning conditions in the form of a Boolean combination of the sets  $[C_1], \dots, [C_n]$ , where  $C_i$  is a subset of the game tree and  $[C_i]$  is the set of plays with infinitely many intersections with  $C_i$ . They proved that one of the players has a winning strategy with restricted memory. In many cases their strategies with restricted memory are non-deterministic finite state-strategies and as such they could be simulated by (concurrent) programs.

However, since GH determinacy result neither include a criterion for a winning player nor explicit description of the winning strategy, it is not by itself sufficient for our purpose to find a concurrent program corresponding to a specification. By analysing their proof we have found a sufficient condition for a given player to win and an explicit description of a winning strategy. This would be given in the second part of the talk. We have found that this condition encompasses the specification of the Mutual Exclusion Problem if stated in game-theoretic terms. We shall now convert the specification for a mutual exclusion problem into a Gurevich-Harrington winning condition.

Suppose that we are given  $n$  parallel processes of the form

```
repeat
  criti;
  remi;
until false
```

and assume that each critical section  $\text{crit}$  requires a use of a resource  $t$  while the remainder  $\text{rem}$  does not have such a requirement. Further assume that we have only  $k \leq n$  units of the resource  $t$ .

The classical mutual exclusion problem is to modify the processes so as to insure the following.

1. No more than  $k$  processes can be in their critical sections at the same time. ("The absence of clashes" requirement.)
2. No process can wait for indefinitely long in order to be allowed to enter its critical section. ("The absence of lockouts or deadlocks" requirement.)
3. There is  $m \in \omega$  such that if a process is waiting for the permission to enter its critical section then during this period of time no more than  $m$  other processes are allowed to enter their respective critical zones. ("The bound" requirement.)
4. If some of the processes would stay in the respective remainder sections for indefinitely long, this should not effect the execution of the other processes. (The requirement on tolerance to failure.)

In order for the solution to be possible, the following precondition is required.

5.  $k$  processes must not occupy the respective critical zones indefinitely long. (An assumption



of absence of failure while at least  $k$  processes are in their critical sections.)

In order to represent the above requirements in the form of a winning condition we first assume that in addition to assignments the set  $A$  (see section 2) also includes elements  $\{crit_i, rem_i : i=0, \dots, n-1\}$ . Let the alphabet  $\Sigma$  be as above. For the convenience, we shall not distinguish between subsets  $\sigma \subseteq A$  and their characteristic functions. So, for example, for  $\sigma \subseteq A$  we shall write  $\sigma(crit_i)=1$  if  $crit_i \in \sigma$  and  $\sigma(crit_i)=0$  if  $crit_i \notin \sigma$ .

For the simplicity we shall formalize the first two most important requirements omitting the rest. Then there is the following correspondence between the subsets of the game tree and these requirements.

1.  $\underline{MutExcl} = \{p : \text{for each prefix } q \text{ of } p \text{ s.t. } \sum_{i \in n} Avail(p)(crit_i) \leq k\}$  corresponds to the positions never violating the first requirement;
2.  $\underline{NoLock}_k = \{p : \text{if } p \text{ contains a Comp's move } \delta \text{ s.t. } \delta = rem_i \text{ then } p \text{ contains a Prog's move } \sigma \text{ after } \delta \text{ s.t. } \sigma(crit_i) = 1 \text{ and if } p \text{ contains a Comp's move } \delta \text{ s.t. } \delta = crit_i \text{ then } p \text{ contains a Prog's move } \sigma \text{ after } \delta \text{ s.t. } \sigma(rem_i) = 1\}$  corresponds to the second requirement;

The winning set corresponding to the requirements 1, 2 and disregarding all others is the following Gurevich-Harrington set.  $W = ([\underline{MutExcl}] \cap (\bigcap_{i \in n} [\underline{NoLock}_i]))$ .

#### REFERENCES.

- [HG] Yuri Gurevich and Leo Harrington, Trees, Automata and Games, Proc. of the 14<sup>th</sup> Annual ACM Symposium on Theory of Computing, 60-65, 1982.
- [AS] B. Alpern, F. B. Schneider, Proving Boolean Combinations of Deterministic Properties, Symposium on Logic in Computer Science, June, 131-137, 1987.
- [B] J Richard Buchi, State-Strategies for Games in  $F_{\sigma\delta} \cap G_{\delta\sigma}$ , The Journal of Symbolic Logic, vol. 48, no 4, Dec. 1983.
- [BL] J Richard Buchi and Lawrence H. Landweber, Solving Sequential Conditions by Finite State Strategies. Purdue University, CSD TR 14, 1967.
- [BA] M. Ben-Ari, Principles of Concurrent Programming, Prentice-Hall International, Inc., 1982.
- [BZ] J. W. de Bakker, J. I. Zucker, Processes and the Denotational Semantics of Concurrency, Information and Control, vol. 54, 70-120, 1982.
- [E] Herbert B. Enderton, A Mathematical Introduction to Logic. Academic Press, 1972.
- [GA] Dov. Gabbay, Amir Pnueli, Saharon Shelah, Jonathan Stavi, 7<sup>th</sup> Annual ACM Symposium on Principals of Programming Languages, 163-173, 1980.
- [G] David Gries, The Science of Programming. Springer-Verlag, 1983.
- [GL] V. M. Glushkov, Automata theory and formal microprogram transformations, Kibernetika, Vol. 1, No. 5, pp. 1-9, 1965.
- [LAC] A. H. Lachlan, On Some Games which Are Relevant to the Theory of Recursively Enumerable Sets., Annals of Mathematics, vol 91, no 2, 291-310, March 1970.
- [L1] Leslie Lamport, Specifying Concurrent Program Modules, ACM Transactions on Programming Languages and Systems, vol. 5, no. 2, April, 190-222, 1983.

- [L2] Leslie Lamport, The "Hoare Logic" of CSP, and All That, *ACM Transactions on Programming Languages and Systems*, vol. 6, no. 2, April, 281-296, 1984.
- [M] Yiannis N. Moschovakis, *Descriptive Set Theory*. North-Holland, 1980 (see chapter The playful universe).
- [M1] Yiannis N. Moschovakis, A game-theoretic modeling of concurrency. February 23, 1989.
- [MP1] Zohar Manna, Amir Pnueli, Specification and Verification of Concurrent Programs by  $\forall$ - automata, 14<sup>th</sup> Annual ACM Symposium on Principles of Programming Languages, 1-12, 1987.
- [MP2] Zohar Manna, Amir Pnueli, Verification of Concurrent Programs: The Temporal Framework, *The Correctness Problem in Computer Science*, Academic Press, New York, 1981.
- [MW] Zohar Manna, Pierre Wolper, Synthesis of Communicating Processes from Temporal Logic Specifications, *ACM Transactions on Programming Languages and Systems*, vol. 6, no. 1, January, 68-93, 1984.
- [OG] Susan Owicki, David Gries, An Axiomatic Proof Technique for Parallel Programs 1, *Acta Informatica* 6, 319-340, 1976.
- [P] R. Parikh, Propositional Game Logic, 24th Annual Symposium on Foundation of Computer Science, 195-200, 1983.
- [R] M. O. Rabin, Decidability of Second Order Theories and Automata on Infinite Trees, *Transactions of the American Mathematical Society*, vol. 141, 1-35, 1969.
- [V] M. Y. Vardi, Verification of Concurrent Programs: The Automata-Theoretic Framework, *Symposium on Logic in Computer Science*, June, 167-176, 1987.

# EXTRACTION OF CONCURRENT PROGRAMS FROM GUREVICH-HARRINGTON GAMES

Vladimir Yakhnis  
Mathematical Sciences Institute  
Caldwell Hall  
Cornell University  
Ithaca, NY 14853

**ABSTRACT.** To a programming language we assign a class of two player games called computational games. In each game the first player, called Prog, gives sets of instructions for another player, called Comp, who executes them. Programs are strategies of Prog, and program specifications are winning conditions. We construct an algebra of strategies which constitute all meanings of the programs (including concurrent programs) in the language. As winning conditions we consider the ones used by Gurevich and Harrington (GH) in their celebrated short proof of Rabin's Theorem. We give a new Theorem providing a sufficient condition for a given player to win a GH game and a wide class of explicit winning strategies. To create the class of winning strategies, we are using a new notion of Priority Automata, generalizing GH's notion of Latest Appearance Record (LAR). This sufficient condition is applicable to Mutual Exclusion Problem formulated as a GH game. So, using the Theorem, we can find a class of winning strategies for Prog in a corresponding computational game. Using the above algebra of strategies it is then possible to find such strategy from the class of winning strategies which corresponds to a concurrent program. The idea of using techniques from the decidability results belongs to Prof. Anil Nerode who also was the first to my knowledge to clearly state that the programs could be understood as strategies in certain games. He also suggested many other valuable ideas.

This work is a sequel to "Concurrent specifications and their Gurevich-Harrington games and representation of programs as strategies," written by Alexander Yakhnis, which is included in this collection. We assume familiarity with this paper which we call Part 1.

## 1. NON-DETERMINISTIC STATE-STRATEGIES

We shall show how to associate with every program from the language, introduced in Part 1 of the sequel, a *strategy* for the first player in the game environment defined above. Though it is possible to consider only non-deterministic strategies (in the usual sense), it appeared to be convenient to deal with state-strategies similar to those used by Buchi. We modified the Buchi's notion of a state-strategy by making it non-deterministic.

### A run of a state-strategy

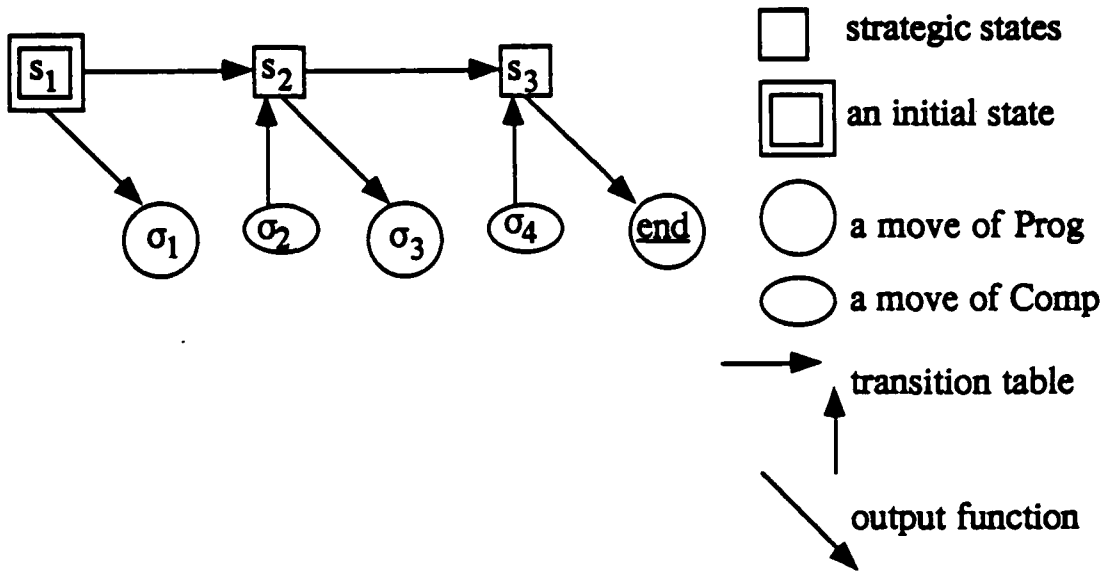


FIG. 1

A *non-deterministic state-strategy* for Prog in our game environment is the following finite state Moore automaton  $F = \langle S, M, S_{in}, P, f \rangle$ , where

1.  $S$  is a set of strategic states;
2.  $M: S \times \Sigma \rightarrow \mathcal{P}(S)$  is the transition table<sup>(1)</sup>, where  $M(s, \sigma) \cap P \neq \emptyset \Leftrightarrow M(s, \sigma) \subseteq P$ ;
3.  $S_{in} \neq \emptyset$  is the set of initial states;
4.  $P$  is the (possibly empty) set of final states;
5. A function  $f: S \rightarrow \Sigma^0$  is the strategic function, where for every state  $s \in S$  ( $s \in P \Leftrightarrow f(s) = \text{end}$ ).

Intuitively, a state-strategy  $F$  for Prog works as follows. The automaton  $F$  moves along the play changing states accepting only moves of Comp, thereby changing its states according to the transition table and producing a run. With every state  $s$  we connect the output  $f(s)$ . Whenever a position  $p$  of Prog with the resultant state  $s$  of the run is reached, Prog uses  $f(s)$  as its current move. This is illustrated on Fig.1. Since, however, a state-strategy may be non-deterministic and therefore could have more than one run on  $p$ , we make an agreement that if we are using  $F$  in a play then shall use a unique run of  $F$  on this play. This is clarified by the definitions of consistency and perpetuality.

Let  $\sigma_0 \dots \sigma_n \in \Sigma^*$  and  $s \in S_{in}$ . We say that  $\sigma_0 \dots \sigma_n$  is *s-consistent* with  $F$  if there an *s-run*  $s_0 s_1 \dots s_n$  on  $\sigma_0 \dots \sigma_{n-1}$  s.t.  $s_0 = s$ ,  $f(s_0) = \sigma_0$  and for all  $i \in [1, n-1]$  if  $\sigma_0 \dots \sigma_i \in \text{Pos}(\text{Prog})$  then  $f(s_{i+1}) = \sigma_{i+1}$ . In this case we call  $s_0 s_1 \dots s_n$  a *strategic run* of  $F$  on  $\sigma_0 \dots \sigma_{n-1}$ . It is easy to see that the empty string  $\lambda$  is *s-consistent* with  $F$  and that any initial state constitute a strategic run on  $\lambda$ . We say that  $\sigma_0 \dots \sigma_n$  is *consistent* with  $F$  if  $\sigma_0 \dots \sigma_n$  is *s-consistent* with  $F$  for some  $s \in S_{in}$ . Similarly we define *s-consistency* and consistency on infinite strings. Note that if  $\alpha$  is a finite string and there is a strategic run of  $F$  on  $\alpha$  then  $\alpha$  is consistent with  $F$ , but that the

(1) We sometimes shall write  $M(s, \sigma) \rightarrow s'$  instead of  $s' \in M(s, \sigma)$ .

converse may not be true. For infinite strings the consistency and the existence of a strategic run are equivalent notions.

We say that the state-strategy  $F$  is *perpetual* if for every position  $p$  and every strategic run  $r$  of  $F$  on  $p$  (if any) the following is true. If  $p \in \text{Pos}(\text{Comp})$  then for any legal move  $\sigma$  of  $\text{Comp}$  from the position  $p$ , the run  $r$  can be continued on  $p \cdot \sigma$ . If  $p \in \text{Pos}(\text{Prog})$  and  $s$  is the resultant state of  $r$  then  $f(s)$  is defined,  $f(s)$  is a legal move of  $\text{Prog}$  from the position  $p$  and if  $f(s) \neq \text{end}$  then the run  $r$  can be continued on  $p \cdot f(s)$ .

Suppose we have a game  $\Gamma = \langle \Sigma, T, \text{Prog}, W \rangle$  which is a part of the game environment. We say that the state-strategy  $F$  is *conditionally winning over*  $\Gamma$  if for every consistent with  $F$  play  $\mu$  we have  $\mu \in W$ . We say that the state-strategy  $F$  is *winning over*  $\Gamma$  if  $F$  is perpetual and is conditionally winning over  $\Gamma$ .

Let  $F = \langle S, M, S_{\text{in}}, P, f \rangle$ ,  $F' = \langle S', M', S'_{\text{in}}, P', f' \rangle$  be state-strategies and suppose that all the states from  $S$  and  $S'$  are reachable. An injection  $\varphi: S \rightarrow S'$  is called a *homomorphism* from  $F$  into  $F'$  if there is a relabelling  $g$  s.t.  $g \circ f = f' \circ \varphi$  (in the sense that  $g \circ f$  is defined iff  $f' \circ \varphi$  is defined),  $\varphi(S_{\text{in}}) \subseteq S'_{\text{in}}$ ,  $\varphi(P) \subseteq P'$  and for all  $s, t \in S$  and  $\sigma \in \Sigma$  ( $M(s, \sigma) \rightarrow t \Rightarrow M'(\varphi(s), g(\sigma)) \rightarrow \varphi(t)$ )  $\wedge$  ( $\sigma$  is a move of  $\text{Comp}$  and  $M'(\varphi(s), g(\sigma)) \neq \emptyset \Rightarrow M(s, \sigma) \neq \emptyset$ )  $\wedge$  ( $M'(\varphi(s), g(f(s))) \neq \emptyset \Rightarrow M(s, f(s)) \neq \emptyset$ ). If there is a homomorphism from  $F$  into  $F'$  we say that  $F$  is a *refinement* of  $F'$ . Since the set of strategic states is finite, it is easy to see that  $\varphi: F \rightarrow F'$  is an isomorphism if and only if  $\varphi: S \rightarrow S'$  is a bijection.

PROPOSITION 1. Let  $F$  be a refinement of  $F'$ . Then if  $F'$  is perpetual (conditionally winning) then so is  $F$ . ■

## 2. AN ALGEBRA OF STATE-STRATEGIES

Now we can build a calculus of strategies. In the definitions below  $F_0 = \langle S_0, M_0, S_{0\text{in}}, P_0, f_0 \rangle$  and  $F_1 = \langle S_1, M_1, S_{1\text{in}}, P_1, f_1 \rangle$  are arbitrary non-deterministic state strategies for  $\text{Prog}$  and  $F = \langle S, M, S_{\text{in}}, P, f \rangle$  are the ones being defined, unless specified otherwise. We assume that  $S_0$  is disjoint with  $S_1$  (which could be achieved by renaming of the states) and that  $s_0$  and  $s_1$  are respectively elements of  $S_0$  and  $S_1$ , unless specified otherwise. We also assume that  $f_0(S_0)$  is disjoint with  $f_1(S_1)$ , which could be achieved by relabelling of the instructions.

### Atomic Strategies.

Let  $a \in \mathcal{A}$ . The automaton  $[a]$  consists of the following components.

1.  $S = \{(\text{output } a), (\text{wait for } a), (a \text{ is finished})\}$ ;
2.  $M((\text{output } a), \underline{\text{wait}}) = (\text{wait for } a)$ ,  
 $M((\text{output } a), a) = (a \text{ is finished})$ ,  
 $M((\text{wait for } a), \underline{\text{wait}}) = (\text{wait for } a)$ ,  
 $M((\text{wait for } a), a) = (a \text{ is finished})$ ,

3.  $S_{in} = \{(\text{output } a)\}$ ;
4.  $P = \{\text{finish}\}$ ;
5.  $f(\text{output } a) = \{a\}$ ,  $f(\text{wait for } a) = \text{skip}$ ,  $f(a \text{ is finished}) = \text{end}$ .

**Composition**  $F_0 \cdot F_1$ .

1.  $S = (S_0 - P_0) \cup S_1$ ;
2. if  $P_0 \cap M_0(s_0, \sigma) = \emptyset$  then  $M(s_0, \sigma) = M_0(s_0, \sigma)$ ,  
if  $P_0 \cap M_0(s_0, \sigma) \neq \emptyset$  then  $M(s_0, \sigma) = S_{lin}$ ,  
 $M(s_1, \sigma) = M_1(s_1, \sigma)$ ;
3.  $S_{in} = S_{0in}$ ;
4.  $P = P_1$ ;
5.  $f(s_0) = f_0(s_0)$  and  $f(s_1) = f_1(s_1)$ .

The following propositions shows that the result of composition behaves as a sequential application of strategies.

**PROPOSITION 2.** Suppose  $F_0 \cdot F_1 = F$  and  $v \in V$ . Then a position  $p$  associated with a machine state  $v$  is consistent with  $F$  iff  $p$  is non-terminal and is consistent with  $F_0$  or there are positions  $q_0$  and  $q_1$  associated respectively with machine states  $v_0$  and  $v_1$  s.t.  $v = v_0$ ,  $v_1$  is the resultant state of the  $v_0$ -run of Exec on  $q_0$ ,  $q_0 \cdot \text{end}$  is consistent with  $F_0$ ,  $q_1$  is consistent with  $F_1$  and  $q = q_0 \cdot q_1$ .

**PROPOSITION 3.** The composition is associative up to an isomorphism.

**Concurrent connection**  $F_0 \parallel F_1$ .

1.  $S = S_0 \times S_1$ ;
2. Suppose that  $\sigma$  is a non-empty move of Comp. Then if  $M_0(s_0, \sigma) \rightarrow t$  then  
 $M(\langle s_0, s_1 \rangle, \sigma) \rightarrow \langle t, s_1 \rangle$  and if  $M_1(s_1, \sigma) \rightarrow t$  then  
 $M(\langle s_0, s_1 \rangle, \sigma) \rightarrow \langle s_0, t \rangle$ .  
If  $s_0$  and  $s_1$  are not final, then  
 $M(\langle s_0, s_1 \rangle, \text{wait}) = M_0(s_0, \text{wait}) \times M_1(s_1, \text{wait})$   
else if  $s_0$  is not final, then

$$M(\langle s_0, s_1 \rangle, \text{wait}) = M_0(s_0, \text{wait}) \times \{s_1\}$$

else if  $s_1$  is not final, then

$$M(\langle s_0, s_1 \rangle, \text{wait}) = \{s_0\} \times M_1(s_1, \text{wait});$$

$$3. S_{in} = S_{0in} \times S_{1in};$$

$$4. P = P_0 \times P_1;$$

5. If at least one of  $s_0$  and  $s_1$  is not final then  $f(\langle s_0, s_1 \rangle) = f_0(s_0) \cup f_1(s_1)$  where at most one of the functions is undefined. In this case we treat the undefined function (if any) as if it gives skip. Also in this case we treat end as skip. If  $s_0$  and  $s_1$  are both final then  $f(\langle s_0, s_1 \rangle) = \text{end}$ .

The following definitions and propositions shows the intuitive behavior of the concurrent connection.

Suppose  $\beta$ ,  $\beta_0$  and  $\beta_1$  are strings. We say that  $p$  is an *interleaving* of  $q_0$  and  $q_1$  if there are order preserving maps  $h_i: q_i \rightarrow p$  for  $i \in \{0, 1\}$  s.t.  $h_i(\text{skip}) = \text{skip}$ ,  $h_i(\text{wait}) = \text{wait}$ ,  $p = h_0(q_0) \cup h_1(q_1)$  and if  $\sigma_i$  is an occurrence of a non-empty character in  $q_i$  for  $i \in \{0, 1\}$  then  $h_0(\sigma_0) \neq h_1(\sigma_1)$ .

**PROPOSITION 4.** Suppose  $p = \sigma_0 \cdot \alpha_0 \cdot \text{wait} \dots \sigma_{n-2} \cdot \alpha_{n-2} \cdot \text{wait} \cdot \sigma_{n-1} \cdot \alpha_{n-1}$  where for all  $i \in n$   $\sigma_i \in \Sigma^0$  and  $\alpha_i$  is a string in  $\Sigma^1 - \{\text{wait}\}$  is a non-terminal position consistent with  $F_0 \parallel F_1$ . Then there are non-executable positions  $q_0 = \sigma'_0 \cdot \alpha'_0 \cdot \text{wait} \dots \sigma'_{n-2} \cdot \alpha'_{n-2} \cdot \text{wait} \cdot \sigma'_{n-1} \cdot \alpha'_{n-1}$  and  $q_1 = \sigma''_0 \cdot \alpha''_0 \cdot \text{wait} \dots \sigma''_{n-2} \cdot \alpha''_{n-2} \cdot \text{wait} \cdot \sigma_{n-1} \cdot \alpha_{n-1}$  s.t.  $n = \max(k, m)$ , for all  $i \in n$   $\sigma'_i = \sigma''_i \cap \sigma_i$  and  $\alpha_i$  is an interleaving of  $\alpha'_i$  and  $\alpha''_i$ , where if  $i > k-1$  (or  $i > m-1$ ) then  $\sigma'_i$  (or  $\sigma''_i$ ) is identified with  $\emptyset$  and  $\alpha'_0$  (or  $\alpha''_0$ ) is identified with  $\lambda$ ,  $q_0$  is consistent with  $F_0$  and  $q_1$  is consistent with  $F_1$ . Moreover, if  $p \cdot \text{end}$  is consistent with  $F_0 \parallel F_1$  then  $q_0 \cdot \text{end}$  is consistent with  $F_0$  and  $q_1 \cdot \text{end}$  is consistent with  $F_1$ . ■

**PROPOSITION 5.** The concurrent connection is commutative and associative up to an isomorphism of state-strategies. ■

**Conditional** if exp then  $F_0$  else  $F_1$

Let  $e$  be a Boolean expression.

$$1. S = S_0 \cup S_1 \cup \{(\text{output } e), (\text{wait for } e)\};$$

2. If  $s \in \{(\text{output } e), (\text{wait for } e)\}$  then

$$M(s, \text{wait}) = (\text{wait for } e),$$

$$M(s, (e, \text{true})) = S_{0in},$$

$$M(s, (e, \text{false})) = S_{1in},$$

- $$M(s_0, \sigma) = M_0(s_0, \sigma),$$
- $$M(s_1, \sigma) = M_1(s_1, \sigma);$$
3.  $S_{in} = \{(output\ a)\};$
  4.  $P = P_0 \cup P_1;$
  5.  $f(output\ e) = \{e\}, f(wait\ for\ e) = skip, f(s_0) = f_0(s_0)$  and  
 $f(s_1) = f_1(s_1).$

The following proposition shows that the conditional behaves as intuitively expected.

**PROPOSITION 6.** Suppose if  $e$  then  $F_0$  else  $F_1 = F, v \in V$  and  $e^V$  (i.e. evaluation of  $e$  in  $v$ ) is defined. Then a position  $p$  associated with a machine state  $v$  is consistent with  $F$  iff  $p$  is not terminal and it is consistent with  $[e]$  or the following is true. If  $e^V = true$  then there are positions  $p'$  and  $q_0$  associated with machine state  $v$  s.t.  $p' \cdot \underline{end}$  is consistent with  $[e]$  and  $q_0$  is consistent with  $F_0$ . If  $e^V = false$  then there are positions  $p'$  and  $q_1$  associated with machine state  $v$  s.t.  $p' \cdot \underline{end}$  is consistent with  $[e]$  and  $q_1$  is consistent with  $F_1$ .

### *Loop while $e$ do $F_0$*

1.  $S = S_0 \cup \{(output\ e), (wait\ for\ e), finish\};$
2. If  $s \in \{(output\ e), (wait\ for\ e)\} \cup P_0$  then  
 $M(s, \underline{wait}) = (wait\ for\ e),$   
 $M(s, (e, true)) = S_{0in},$   
 $M(s, (e, false)) = finish$   
 If  $s_0 \in S_0 - P_0$  then  
 $M(s_0, \sigma) = M_0(s_0, \sigma);$
3.  $S_{in} = \{(output\ a)\};$
4.  $P = \{finish\};$
5. For  $s \in \{(output\ e)\} \cup P_0$   $f(s) = e, f(wait\ for\ e) = skip,$  for  $s_0 \in S_0 - P_0$   $f(s_0) = f_0(s_0)$  and  
 $f(finish) = \underline{end}.$

The following proposition shows that the loop behaves as intuitively expected.

**PROPOSITION 7.** Suppose while  $e$  do  $F_0 = F$  and  $v_0$  is a machine state. Then a position  $p$  with a  $v_0$ -run of  $r$  of Exec is consistent with  $F$  iff there are positions  $p_0, q_0, \dots, p_{n-1}, q_{n-1}$ , where  $p_{n-1}$  is not empty and  $q_{n-1}$  may be empty, and the states  $v_0, \dots, v_{n-1}$  s.t.  $p_i, q_i$  are associated with  $v_i$  which is the resultant state of  $r$  on  $p_0, q_0, \dots, p_i$  and  $p_0, q_0, \dots, p_i, q_i$ , also  $p_i$  is consistent with  $[e]$  and  $q_i$  is consistent with  $F_0$  for  $i \in n$  and  $p = p_0 \cdot q_0 \cdot \dots \cdot p_{n-1} \cdot q_{n-1}$ .



Moreover,  $p$  is terminal iff  $p_{n-1}$  is terminal and  $q_{n-1}$  is empty.

### 3. HOW TO LOOK FOR WINNING STRATEGIES

#### Priorities and the Strategies Based on Them.

Let  $\Sigma$  be a finite alphabet and  $T \subseteq \Sigma^*$ . The sets  $C_1, \dots, C_n \subseteq T$  are called basic GH sets. We define a function which extracts from a position the history of meeting the basic GH sets. Let  $\Sigma = \{0,1\}^n$ . Define the coding function  $\text{Code}: T \rightarrow \Sigma$  by  $\text{Code}(p)_i = 1$  if  $p \in C_i$  and  $\text{Code}(p)_i = 0$  otherwise. This map is converted into  $\text{Code}': T \rightarrow \Sigma^*$  by putting  $\text{Code}'(e) = \text{Code}(e)$  and  $\text{Code}'(p \cdot a) = \text{Code}'(p) \cdot \text{Code}(p \cdot a)$ .

We'll explain one of the ways we combine several strategies into one. Let  $\mathcal{U} = \langle S, S_0, M \rangle$  be a finite non-deterministic automaton over the alphabet  $\Sigma$ . A run of  $\mathcal{U}$  on  $\Sigma^*$  is a function  $r: \Sigma^* \rightarrow S$  s.t.  $r(e) \in S_0$  and for all  $\alpha \in \Sigma^*$  and  $\sigma \in \Sigma$ ,  $r(\alpha \cdot \sigma) \in M(r(\alpha), \sigma)$ . We combine a given finite collection  $\{f^s : s \in S\}$  of strategies into a strategy  $f$  using  $\Sigma^*$ -runs of  $\mathcal{U}$  as follows. Suppose  $r$  is a run of  $\mathcal{U}$ . For any  $p \in \text{Pos}(\Omega)$  let  $f(p) = f^{r(\text{Code}'(p))}(p)$ . We shall call  $f$  a strategy induced by  $\mathcal{U}$ . Such strategies are also called *induced by automata (relative to  $C_1, \dots, C_n$ )*. Thus for any  $\mathcal{U}$  and  $\{f^s : s \in S\}$  as above we have a class of strategies parametrized by runs of  $\mathcal{U}$ .

There is a special class of automata which we are going to use in constructing the strategies. The purpose of such automata is to construct a strategy which allows us to meet every  $C_i$  infinitely often. While making moves, we are building a run  $r$  of  $\mathcal{U}$  over the  $\text{Code}'$ -image of the play. At any position  $p$  of the play our immediate goal is to reach  $C_{h(r(\text{Code}'(p)))}$ . First, we wish to guarantee that the goal is not changed until it is fulfilled. Second, we wish to guarantee that if for the run  $r$  the goal is reached infinitely often, then the play meets *every*  $C_i$ ,  $i \in \{1, \dots, n\}$ , infinitely often. Moreover, we would like to insure a "fair"<sup>(2)</sup> treatment of any  $C_i$  in the sense that for some fixed  $m \in \omega$ , we should not be able to reach all the other sets more than  $m$  times in a row without reaching  $C_i$  in between. The following definition satisfies these requirements.

Let  $\mathcal{U}$  be a finite non-deterministic Moore automata with the set of states  $S$ , the output function  $h: S \rightarrow \{1, \dots, n\}$  and the transition table  $M$ , accepting all words in the alphabet  $\{0,1\}^n$ . If  $r$  is a run of  $\mathcal{U}$ ,  $\alpha$  is a word in  $\{0,1\}^n$ ,  $b \in \{0,1\}^n$  and  $b_{h(r(\alpha))} = 1$ , we say that  $r$  *reaches the goal* on  $\alpha \cdot b$ .

We call  $\mathcal{U}$  a *priority automaton* if its output function, called the *priority function* of  $\mathcal{U}$ , satisfies the following two properties.

1. For any state  $s$  and any letter  $b \in \{0,1\}^n$ , if  $b_{h(s)} = 0$  then for any  $s' \in M(s, b)$   $h(s) = h(s')$ ;
2. There is some  $m \geq n$  (which we call a bound) s.t. for any run  $r$  of  $\mathcal{U}$  and words  $\alpha \leq \beta$ , if  $r$  reaches its goal at least  $m$  times on  $\{\beta' : \alpha \leq \beta' \leq \beta\}$  then  $h$  takes *all* the values from  $\{1, \dots, n\}$  on  $r(\{\beta' : \alpha \leq \beta' \leq \beta\})$ .

<sup>(2)</sup>This is similar to "fair" treatment of concurrent processes in Computer Science.

**Theorem 1.** Let  $Q \subseteq T$ ,  $\mathcal{A}$  be a priority automata with an output function  $h$ ,  $F_i = \text{Decr}(C^i - Q, \Omega)$  and  $\varphi_i(q) \Leftrightarrow q \in \text{Dom}^1(C^i - Q, \Omega)$ . for  $i \in \{1, \dots, n\}$  and  $G_1, \dots, G_n$  be such  $\Omega$ -strategies that

1.  $e \notin Q$ ;
2. For  $i \in n$  and any position  $q \in Q \cup \text{Dom}^1(C^i - Q, \Omega)$ ,  $G_i$  wins  $\Gamma'$  against  $\text{Avoid}(C^i - Q, 1 - \Omega)$ .

Then for any run  $r$  of  $\mathcal{A}$ , the following strategy  $f$  wins  $\Gamma$  from every position  $p \notin Q$ .

Let  $p \notin q$  and  $i = h(r(\text{Code}'(p)))$ . If  $p \in \text{Dom}^1(C^i - Q, \Omega)$ , define  $f(p) = F_i(p)$  and otherwise define  $f(p) = G_i(p)$ . See the Appendix for the notions of  $\text{Dom}^1$ ,  $\text{Decr}$  and  $\text{Avoid}$ . ■

### Gurevich-Harrington's LAR and examples of non-deterministic priority automata

The notion of priority automata is a generalization of GH's Latest Appearance Record (LAR). Our version of LAR is a somewhat modified form of LAR from [GH].

The alphabet of  $\text{LAR}(n)$  and all other priority automata in this section is  $\Sigma = \{0, 1\}^n$ . Let  $\text{Order}(n) = \{s \in \{1, \dots, n\}^* : \text{for all } i \in \{1, \dots, n\}, i \text{ occurs in } s \text{ at most once}\}$ . Note that  $\text{Order}(n)$  includes the empty word  $e$ .  $\text{Order}(n)$  is going to be the set of states of  $\text{LAR}(n)$ . Let us define the transition table  $M$ .

Let  $\sigma \in \Sigma$  and  $s \in \text{Order}(n)$ . Let  $X = \{i : \sigma_i = 1\}$ ,  $s' \in \{1, \dots, n\}^*$  be an increasing sequence of elements of  $X$  and  $s''$  be the result of crossing out of  $s$  of all elements of  $X$ . We define  $M(s, \sigma) = s'' \cdot s'$ . The output function  $h$  is defined by  $h(e) = 0$  and  $h(s) = \text{First}(s)$  if  $s \neq e$ .

**PROPOSITION 3.1.**  $\text{LAR}(n)$  is bounded by  $n$ . ■

$\text{LAR}(n)$  is a deterministic priority automaton.

The following priority automata are non-deterministic and both have been discovered by game theoretic analysis of well known concurrent programs.  $\text{MOD}(n)$  stems from Eisenberg and McGuire's algorithm (see [PS]) and  $\text{SUB}(n)$  stems from Morris' algorithm (see [MOR]).

Our next priority automaton is designated  $\text{MOD}(n)$  from "modulo  $n$ ". Its set of states is  $\{1, \dots, n\}$ . The transition table is defined as follows. For  $s \in \{1, \dots, n\}$  and  $\sigma \in \{0, 1\}^n$ ,  $M(s, \sigma) = \{i\}$  if  $\sigma_i = 0$ ,  $M(s, \sigma) = \{1, \dots, n\}$  if for all  $j \in \{1, \dots, n\}$   $\sigma_j = 1$ , and  $M(s, \sigma) = \{j\}$ , where  $j - i \pmod n = \min\{k - i \pmod n : k \in \{1, \dots, n\}\}$ , otherwise. The output function is the identity function on  $\{1, \dots, n\}$ .

**PROPOSITION 3.2.**  $\text{MOD}(n)$  is bounded by  $n$ . ■

The last example of a priority automaton is designated  $\text{SUB}(n)$  from "subset of  $n$ ". Its set of states is  $S = \{\langle X, i \rangle : i \in X \subseteq \{1, \dots, n\}\}$ . The transition table is defined as follows. For  $s = \langle X, i \rangle \in S$  and  $\sigma \in \{0, 1\}^n$ , let  $X' = \{j \in X : \sigma_j = 0\}$ . Then  $M(s, \sigma) = \{\langle X', i \rangle\}$  if  $\sigma_i = 0$ ,  $M(s, \sigma) = \{\langle X', j \rangle : j \in X'\}$  if  $X' \neq \emptyset$  and  $\sigma_j = 1$ , and  $M(s, \sigma) = \{\langle \{1, \dots, n\}, j \rangle : j \in \{1, \dots, n\}\}$  otherwise. The output function is  $h(\langle X, i \rangle) = i$ .

**PROPOSITION 3.3.**  $\text{SUB}(n)$  is bounded by  $2^n$ . ■

#### 4. THE MUTUAL EXCLUSION PROBLEM

In Part I of the sequel (see the abstract), the winning condition is represented by  $W = ([\text{MutExcl}] \cap (\bigcap_{i \in \mathbb{N}} [\text{NoLock}_i]))$ .

Since this set is exactly of the form used in Theorem 1, it is easy to see that the theorem provides a class of winning strategies for this problem.

#### Appendix. RANKS, DOMAINS AND STRATEGIES BASED ON THEM

Let  $A$  be a finite alphabet and  $T \subseteq A^*$  be a tree without leaves. We call  $T$  a *game tree* and its elements *positions*. If a position  $p$  is a *prefix* of a position  $q$  we shall write  $p \leq q$ .

We shall consider games with the following rules. 0 and 1 are respectively first and second players. If  $p$  is a position with even (odd) length then 0 (1) chooses a letter  $a \in A$  s.t.  $p \cdot a \in T$ . We designate the set of all even (odd) positions as  $\text{Pos}(0)$  ( $\text{Pos}(1)$ ). A *play* is an infinite sequence produced by the above rules. The collection of all plays is designated  $\text{Play}(T)$ .

Let  $X \subseteq T$  and  $\Omega$  be a player. A position  $q$  is a *child* of  $p$  if  $q = p \cdot a$  for some  $a \in A$ . Denote by  $p_T$  the *set of children* of  $p$ . A *non-deterministic strategy* for  $\Omega$  is a function  $f: \text{Pos}(\Omega) \rightarrow \mathcal{P}(T)$  s.t. for any  $p \in T$   $f(p)$  is a subset of the set of children of  $p$ . From now on a 'strategy' means a 'non-deterministic strategy', unless we say otherwise. The two most simple strategies are called *vacuous* they are defined by  $\text{Vac}(\Omega)(p) = p_T$  and  $\text{Vac}(1-\Omega)(q) = q_T$ . We say that a strategy  $f$  is *defined* on  $X \subseteq T$  if for any  $p$  in  $X$   $f(p) \neq \emptyset$  and for any  $p$  outside  $X$   $f(p) = \emptyset$ . For example  $\text{Vac}(\Omega)$  is defined on  $\text{Pos}(\Omega)$  and  $\text{Vac}(1-\Omega)$  is defined on  $\text{Pos}(1-\Omega)$ . We say that a strategy  $f$  is *defined at least* on  $X \subseteq T$  if for any  $p$  in  $X$   $f(p) \neq \emptyset$  without any supposition for the behavior of  $f$  outside  $X$ . We say that an  $\Omega$ -strategy  $f'$  is a *refinement* of  $f$  if for all  $p \in \text{Pos}(\Omega)$   $f'(p) \subseteq f(p)$ . Informally, we shall write  $f' \subseteq f$ . The set of *consistent* with  $f$  positions is defined as follows. The empty word  $e$  is consistent with  $f$ . For any consistent with  $f$  position  $p$  if  $p \in \text{Pos}(1-\Omega)$  then any child of  $p$  is consistent with  $f$  and if  $p \in \text{Pos}(\Omega)$  then any  $q \in f(p)$  is consistent with  $f$ . Consistency after position  $p_0$  is defined by replacing  $e$  above by  $p_0$ . Consistency of plays is defined similarly. Sometimes we shall informally use words 'reach', 'meet' and so on instead of speaking in terms of consistency.

All the strategies considered here are based on the following concept of " $\Omega$ -rank inside an  $\Omega$ -strategy and against a  $(1-\Omega)$ -strategy". Let  $f$  be an  $\Omega$ -strategy and  $g$  be a  $(1-\Omega)$ -strategy and  $X \subseteq T$ . We shall inductively define a partial function  $\text{Rank}(X, \Omega, f/g): T \rightarrow \omega$  as follows.

1. For all  $p \in X$ ,  $\text{Rank}(X, \Omega, f/g)(p) = 0$ ;
2. If  $p \in \text{Pos}(\Omega)$ ,  $\text{Rank}(X, \Omega, f/g)$  is defined on at least one child of  $p$  from  $f(p)$ , and  $n$  is the minimal value of  $\text{Rank}(X, \Omega, f/g)$  on  $f(p)$ , then  $\text{Rank}(X, \Omega, f/g)(p) = n + 1$ ;
3. If  $p \in \text{Pos}(1-\Omega)$ ,  $g(p) \neq \emptyset$ ,  $\text{Rank}(X, \Omega, f/g)$  is defined on all children of  $p$  which are in  $g(p)$ , and  $n$  is the maximal value of  $\text{Rank}(X, \Omega, f/g)$  on  $g(p)$ , then  $\text{Rank}(X, \Omega, f/g)(p) = n + 1$ .

**PROPOSITION 1.1.** For any  $X \subseteq T$ ,  $p \in T$  and  $n \geq 0$   $\text{Rank}(X, \Omega, f/g)(p) = n$  iff there is a strategy for  $\Omega$  which is a refinement of  $f$  and which allows to reach  $X$  starting from  $p$  with at most  $n$  moves while  $1-\Omega$  uses  $g$  after  $p$ .

GH's original notion of "rank" can be represented in our notation in the form  $\text{Rank}(X, \Omega, \text{Vac}(\Omega)/\text{Vac}(1-\Omega))$ . The idea of playing against a fixed strategy of the opponent, they explored in their "Sewing Lemma". In contrast, we do not use the "Sewing Lemma", but rely on modified GH's "domains" based on "rank for  $\Omega$  inside an  $\Omega$ -strategy and against a  $(1-\Omega)$ -strategy".

It may appear at first that introduction of  $f$  and  $g$  into ranks and the subsequent notions is redundant since  $f$  and  $g$  may form a subtree of  $T$  and, in this subtree, our rank is just the GH's rank. However closer observation shows that since  $f$  and  $g$  are not necessarily everywhere defined, i. e. we allow  $f(p)=\emptyset$  or  $g(p)=\emptyset$ ,  $f$  and  $g$  may not necessarily form a subtree. Rather, they may form many disjoint subtrees and finite parts of subtrees. Since we would like to define uniform strategies in areas where  $f$  and  $g$  are defined, the notion of rank in full generality appears to be useful.

We designate  $\text{Dom}(X, \Omega, f/g) = \{p \in T : \text{Rank}(X, \Omega, f/g)(p) \geq 0\}$  and  $\text{Dom}^1(X, \Omega, f/g) = \{p \in T : \text{Rank}(X - \{p\}, \Omega/g)(p) \geq 1\}$ .

PROPOSITION 1.2.  $\text{Dom}^1(X, \Omega, f/g) \subseteq \text{Dom}(X, \Omega, f/g)$ ,  
 $\text{Dom}^1(X, \Omega, f/g) \cup X = \text{Dom}(X, \Omega, f/g)$ ,  $\text{Dom}(\text{Dom}(X, \Omega, f/g), \Omega, f/g) = \text{Dom}(X, \Omega, f/g)$  and  
 $\text{Dom}(\text{Dom}^1(X, \Omega, f/g), \Omega, f/g) = \text{Dom}^1(X, \Omega, f/g)$ .

The following strategies are used as building blocks for all the strategies considered here.  $\text{Decr}(X, \Omega, f/g)(p) = \{q : q \in f(p) \text{ and } \text{Rank}(X - \{p\}, \Omega, f/g)(q) \leq \text{Rank}(X - \{p\}, \Omega, f/g)(p)\}$  is an  $\Omega$ -strategy. Note that if  $q$  descends from  $p$  then  $\text{Rank}(X - \{p\}, \Omega, f/g)(q) = \text{Rank}(X, \Omega, f/g)(q)$ .  $\text{Avoid}(X, 1-\Omega, g/f)(p) = \{q : q \in g(p) \text{ and } q \notin \text{Dom}(X, \Omega, f/g)\}$  is a  $(1-\Omega)$ -strategy.

PROPOSITION 1.3. If  $p \in \text{Dom}^1(X, \Omega, f/g)$  then  $\text{Decr}(X, \Omega, f/g)$  allows to reach  $X$  after  $p$  if  $1-\Omega$  uses  $g$  after  $p$ . Moreover while  $\Omega$  is using  $\text{Decr}(X, \Omega, f/g)$  and  $1-\Omega$  is using  $g$ , the play stays inside  $\text{Dom}^1(X, \Omega, f/g)$  at least until it reaches  $X$  first time after  $p$ . If  $p \notin \text{Dom}^1(X, \Omega, f/g)$  then  $\text{Avoid}(X, 1-\Omega, g/f)$  allows to *never* reach  $\text{Dom}(X, \Omega, f/g)$  (and hence also  $X$ ) after  $p$  if  $\Omega$  uses  $f$  after  $p$ .

## REFERENCES.

- [HG] Yuri Gurevich and Leo Harrington, Trees, Automata and Games, Proc. of the 14<sup>th</sup> Annual ACM Symposium on Theory of Computing, 60-65, 1982.
- [AS] B. Alpern, F. B. Schneider, Proving Boolean Combinations of Deterministic Properties, Symposium on Logic in Computer Science, June, 131-137, 1987.
- [B] J Richard Buchi, State-Strategies for Games in  $F_{\sigma\delta} \cap G_{\delta\sigma}$ , The Journal of Symbolic Logic, vol. 48, no 4, Dec. 1983.
- [BL] J Richard Buchi and Lawrence H. Landweber, Solving Sequential Conditions by Finite State Strategies. Purdue University, CSD TR 14, 1967.
- [BA] M. Ben-Ari, Principles of Concurrent Programming, Prentice-Hall International, Inc.,

1982.

- [BZ] J. W. de Bakker, J. I. Zucker, Processes and the Denotational Semantics of Concurrency, *Information and Control*, vol. 54, 70–120, 1982.
- [E] Herbert B. Enderton, *A Mathematical Introduction to Logic*. Academic Press, 1972.
- [GA] Dov. Gabbay, Amir Pnueli, Saharon Shelah, Jonathan Stavi, 7<sup>th</sup> Annual ACM Symposium on Principals of Programming Languages, 163–173, 1980.
- [G] David Gries, *The Science of Programming*. Springer-Verlag, 1983.
- [GL] V. M. Glushkov, Automata theory and formal microprogram transformations, *Kibernetika*, Vol. 1, No. 5, pp. 1–9, 1965.
- [LAC] A. H. Lachlan, On Some Games which Are Relevant to the Theory of Recursively Enumerable Sets., *Annals of Mathematics*, vol 91, no 2, 291–310, March 1970.
- [L1] Leslie Lamport, Specifying Concurrent Program Modules, *ACM Transactions on Programming Languages and Systems*, vol. 5, no. 2, April, 190–222, 1983.
- [L2] Leslie Lamport, The "Hoare Logic" of CSP, and All That, *ACM Transactions on Programming Languages and Systems*, vol. 6, no. 2, April, 281–296, 1984.
- [M] Yiannis N. Moschovakis, *Descriptive Set Theory*. North-Holland, 1980 (see chapter The playful universe).
- [M1] Yiannis N. Moschovakis, A game-theoretic modeling of concurrency. February 23, 1989.
- [MP1] Zohar Manna, Amir Pnueli, Specification and Verification of Concurrent Programs by  $\forall$ - automata, 14<sup>th</sup> Annual ACM Symposium on Principals of Programming Languages, 1–12, 1987.
- [MP2] Zohar Manna, Amir Pnueli, *Verification of Concurrent Programs: The Temporal Framework*, *The Correctness Problem in Computer Science*, Academic Press, New York, 1981.
- [MW] Zohar Manna, Pierre Wolper, Synthesis of Communicating Processes from Temporal Logic Specifications, *ACM Transactions on Programming Languages and Systems*, vol. 6, no. 1, January, 68–93, 1984.
- [OG] Susan Owicki, David Gries, An Axiomatic Proof Technique for Parallel Programs 1, *Acta Informatica* 6, 319–340, 1976.
- [P] R. Parikh, Propositional Game Logic, 24th Annual Symposium on Foundation of Computer Science, 195–200, 1983.
- [R] M. O. Rabin, Decidability of Second Order Theories and Automata on Infinite Trees, *Transactions of the American Mathematical Society*, vol. 141, 1–35, 1969.
- [V] M. Y. Vardi, *Verification of Concurrent Programs: The Automata-Theoretic Framework*, *Symposium on Logic in Computer Science*, June, 167–176, 1987.

# UNSTABLE INTERFACES AND ANOMALOUS WAVES IN COMPRESSIBLE FLUIDS

JOHN W. GROVE†, RALPH MENIKOFF‡, QIANG ZHANG§

**Abstract.** The gravitational acceleration of a heavier fluid into a lighter fluid causes unstable modes to grow in the interface between the two fluids and leads to their eventual chaotic mixing. This phenomena is known as the Rayleigh-Taylor instability. We discuss the development and validation of a model for the long term dynamics of the mixing boundary layer. The model uses a simplified description of the dynamics of the bubbles of lighter fluid rising in the heavier fluid. Validation of the model is achieved by comparison with experiments and full scale two dimensional simulations of the mixing process.

We also discuss the production of anomalous waves during the interaction of shock waves with fluid interfaces. The focus here is on the case when the shock passes from a medium of high to low acoustic impedance. Curvature of either of the interacting waves causes the diffraction patterns produced during the collision to bifurcate from locally self-similar pseudo-stationary configurations to unsteady anomalous reflections. This process is analogous to the transition from a regular to a Mach reflection where the reflected wave is a rarefaction instead of a shock. These bifurcations are incorporated into a front tracking code that gives an accurate description of the wave interactions. Numerical results for two illustrative cases are described; a planar shock passing over a bubble, and an expanding shock impacting a planar contact.

## 1. Introduction.

This report treats two aspects of computational fluid dynamics, the unstable behavior of gravity driven mixing, and the diffraction of shock waves through fluid interfaces. The latter process is itself associated with a related instability, known as the Richtmyer-Meshkov instability, that occurs in shock accelerated interfaces.

The Rayleigh-Taylor instability is a fingering instability between two fluids with different densities. If the interface between the two fluids is planar and perpendicular to the direction of the applied external forces, then such a system is in a state of unstable equilibrium when the light fluid supports the heavier. Any small perturbation of the fluid interface will upset this unstable equilibrium leading to the formation of rising bubbles of the light fluid and falling spikes of the heavier. As the mixing process develops, spikes can pinch off to form droplets.

---

†Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794. Supported in part by the U. S. Army Research Office, grant no. DAAL03-89-K-0017.

‡Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545. Supported by the U. S. Department of Energy.

§Courant Institute of Mathematical Sciences, New York University, New York, NY 10012. Supported in part by NSF Grant DMS-8619856

Typeset by  $\mathcal{A}\mathcal{M}\mathcal{S}$ -TEX

The mixing of two fluids under the influence of gravity was first investigated by Rayleigh [26] and later by Taylor [31]. Since then a variety of computational and analytic methods have been used to study this classical problem. These include; nonlinear integral equations [4], [ 7] boundary integral techniques [33] conformal mapping [21], dynamical modeling [12], [ 30], vortex-in-cell methods [32], [ 36] higher order Godunov methods [34], and front tracking [8], [ 11], [ 13]. Most of this work has been carried out for incompressible fluids or in the limit of a single fluid in a vacuum. For a review of the Rayleigh-Taylor instability and its applications to science and engineering, see reference [29]. We will present here results on the behavior of a single unstable mode (one bubble), as well as the interaction of multiple bubbles.

The front tracking method was used for the direct simulation of the mixing process. We conducted a series of computational experiments for periodic arrays of single bubbles (the single mode case) as well as for multiple bubble interactions. Tracking the fluid interface offered several advantages, it eliminated numerical diffusion at the interface, and it allowed an accurate measurement to be made of the interface velocity.

Our analysis consisted of modeling the motion of the tip of a spike or bubble in a single mode system by an ordinary differential equation, and applying these results to the interaction of multiple bubbles. We found that in a chaotic flow the interaction between the different bubbles causes the magnitude of the terminal velocity of a large bubble to be greater than that predicted by the single bubble theory. This led us to formulate a superposition model in which larger bubbles "capture" the velocity of nearby smaller bubbles. We found agreement between the velocities predicted by this simplified model and those obtained by direct numerical simulations, although the agreement is better for large Atwood numbers and low compressibility than in the opposite case.

The second part of this report treats the diffraction patterns produced by the collision of a shock wave with a fluid interface. This process produces a variety of complicated wave diffractions [1], [ 2], [ 18]. In the simplest case these consist of pseudo-stationary self-similar waves that can be described by solutions to Riemann problems for the supersonic steady-state Euler equations. In more complicated cases and in particular when one or both of the colliding waves is curved, these regular diffraction patterns can bifurcate into complex composites of individual wave interactions between the scattered waves.

The goal here is to understand the particular bifurcation behavior of the collision of a shock in a dense fluid with an interface between the dense fluid and a much lighter one. Two basic cases are considered. The collision of a shock in water with a bubble of air, and the diffraction of a cylindrically expanding underwater shock wave with the water's surface. It will be seen that initially these interactions produce regular shock diffractions with reflected Prandtl-Meyer waves. Subsequently these regular waves bifurcated to form anomalous waves that are analogous to non-

centered Mach reflections whose reflected waves are rarefactions. We will describe a method to include this analysis into a front tracking numerical method that allows enhanced resolution computations of these interactions.

## 2. The Equations of Motion.

In the absence of heat conduction and viscosity, fluid flow is governed by the Euler equations that describe the laws of conservation of mass, momentum and energy respectively.

$$(2.1a) \quad \partial_t \rho + \nabla \cdot (\rho \mathbf{q}) = 0,$$

$$(2.1b) \quad \partial_t (\rho \mathbf{q}) + \nabla \cdot (\rho \mathbf{q} \otimes \mathbf{q}) + \nabla P = \rho \mathbf{g},$$

$$(2.1c) \quad \partial_t (\rho \mathcal{E}) + \nabla \cdot \rho \mathbf{q} (\mathcal{E} + VP) = \rho \mathbf{q} \cdot \mathbf{g}.$$

Here,  $\rho$  is the mass density,  $\mathbf{q}$  is the particle velocity,  $\mathbf{g}$  is the gravitational acceleration,  $\mathcal{E} = \frac{1}{2}|\mathbf{q}|^2 + E$  is the total specific energy,  $E$  is the specific internal energy, and  $P$  is the pressure. The equilibrium thermodynamic pressure  $P(V, E)$ , where  $V = 1/\rho$  is the specific volume, is referred to as the equation of state and describes the fluid properties. The numerical examples below used either the polytropic equation of state,

$$(2.2) \quad P(V, E) = (\gamma - 1)\rho E,$$

or the stiffened polytropic equation of state, [16], [25]

$$(2.3) \quad P(V, E) = \Gamma_0 \rho (E - E_\infty) - (\Gamma_0 + 1)P_\infty,$$

where  $\gamma$ ,  $\Gamma_0$ ,  $E_\infty$ , and  $P_\infty$  are positive constants. In particular, all of the Rayleigh-Taylor simulations used a polytropic equation of state with  $\gamma = 1.4$ .

System (2.1) is hyperbolic with characteristic modes corresponding to the propagation of sound waves and fluid particles through the medium. The sound waves propagate in all directions from their source with a sound speed  $c$  with respect to the fluid, where  $c^2 = \partial P / \partial \rho$  at constant entropy. Another important measure of sound propagation is the Lagrangian sound speed or acoustic impedance given by  $\rho c$ .

## 3. Motion of single mode bubbles and spikes.

For a given equation of state, the two fluid mixing problem is characterized by two dimensionless quantities. The first of these is a relative measure of the difference in densities between the two fluids, the Atwood number  $A = \frac{\rho_h - \rho_l}{\rho_h + \rho_l}$ . The second measures the compressibility of the heavy fluid. If  $\lambda$  is the wavelength of the perturbation and  $c_h$  is the speed of sound in the heavy fluid, we define the dimensionless compressibility to be  $C^2 = \frac{\lambda g}{c_h^2}$ . Our goal is to study the overall behavior



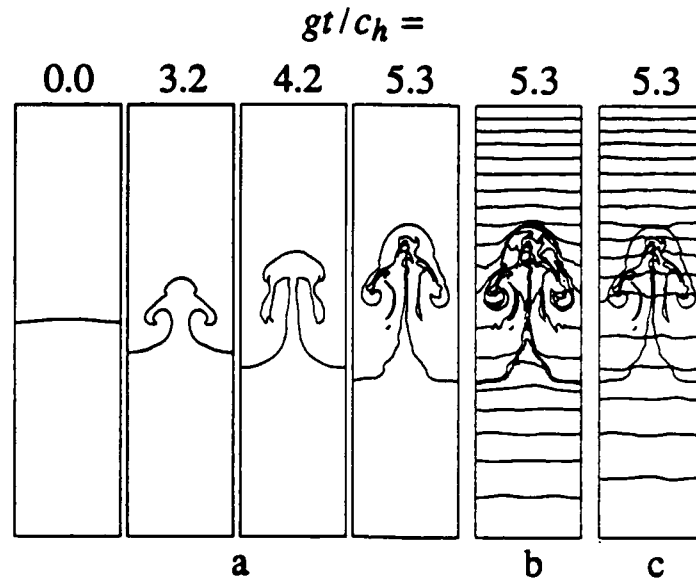


Fig. 1 Plots of the interface position, and density and pressure contours for  $A = 1/5$ ,  $C^2 = 0.5$ , and  $\gamma = 1.4$  in a  $1 \times 6$  computation domain with a  $40 \times 240$  grid. Only the upper two thirds of the computational region is shown in the plot because nothing of interest occurs in the remainder of the computation. The interface position for successive time steps is shown in (a) while (b) and (c) show contours of density and pressure respectively. Gravity is directed downward.

of the unstable mixing between the two fluids for a range of Atwood numbers and compressibilities.

For a polytropic equation of state, the equilibrium solution of the Euler equations is an exponentially stratified distribution of density and pressure along the direction of gravity. We used the solution to a linearized perturbation of this equilibrium solution [3], [8] to provide the Cauchy data for a full Euler simulation. Here we consider the single mode system, which is a periodic array of bubbles and spikes. The top and bottom of the computational domain are reflecting boundaries.

Figs. 1 and 2 show computational results for two different simulations with  $C^2 = 0.5$  and  $A = \frac{1}{5}$  and  $\frac{9}{11}$  respectively. If the Atwood number is small (Fig. 1), two interpenetrating fingers of similar shape are formed with secondary instabilities appearing along the side of the spike. As  $A \rightarrow 0$ , the pattern of the two fluids becomes symmetric with a phase difference  $\pi$ . For larger Atwood numbers (Fig. 2), the spike is thinner with less roll up shed off the edge of its tip. If the compressibility is high, the velocity of the bubble or spike will eventually become supersonic relative to the heavy material but will remain subsonic in the light material. We refer to

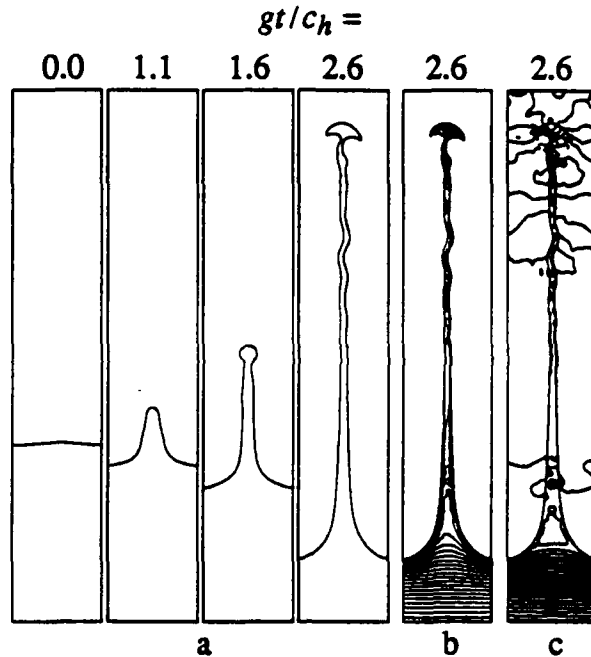


Fig. 2 Plots of the interface position, density and pressure contours for  $A = 0.01$ ,  $C^2 = 0.5$ , and  $\gamma = 1.4$  in a  $1 \times 10$  computation domain with a  $20 \times 200$  grid. Only the upper four fifths of the computational region is shown in the plot because nothing of interest occurs in the remainder of the computation. The interface position for successive time steps is shown in (a) while (b) and (c) show contours of density and pressure respectively. Gravity is directed downward.

[8] for the details of these studies.

A bubble or spike that arises from a small amplitude disturbance goes through three regimes; an initial stage governed by the linearized equations, a period of free fall, and a final terminal velocity phase. During the linear stage, the velocity grows exponentially with time. We denote this growth rate by  $\sigma$ . In the free fall regime the velocity varies linearly with time and the acceleration reaches a maximum absolute value called the renormalized gravity  $g_R$ . Finally the velocity approaches a limiting value (terminal velocity  $v_\infty$ ) with a decay rate  $b$ . These three regimes are illustrated in Fig. 3 which shows plots of the spike velocity and acceleration verses time.

By using curve fitting through the three growth regimes of the spike or bubble, it is possible to describe its motion the ordinary differential equation

$$(3.1) \quad \frac{dv}{dt} = \frac{\sigma v(1 - \frac{v}{v_\infty})}{\frac{\sigma}{b} \frac{v}{v_\infty} + (1 - \frac{v}{v_\infty}) + [\frac{\sigma v_\infty}{g_R} - (1 + \sqrt{\frac{\sigma}{b}})^2] \frac{v}{v_\infty} (1 - \frac{v}{v_\infty})},$$

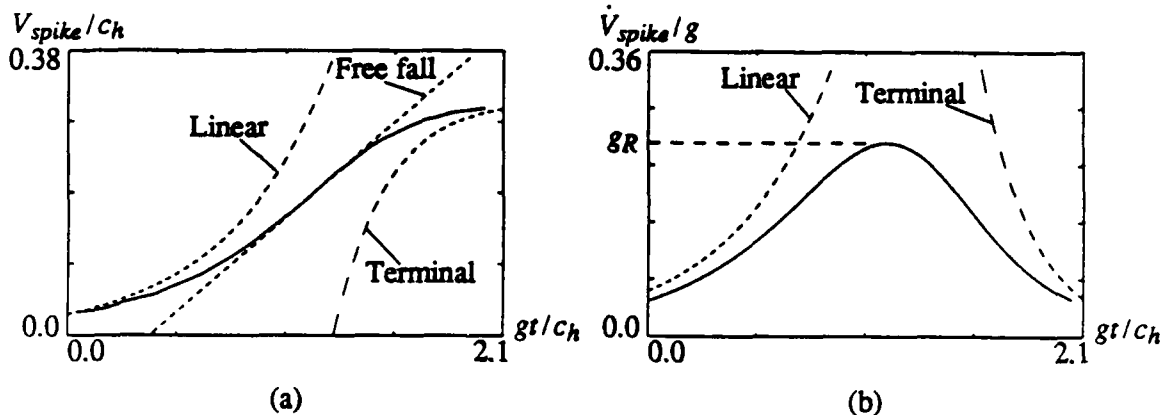


Fig. 3 The comparison of the spike velocity and the spike acceleration obtained numerically with the asymptotic behavior in each regime.  $A = 1/3$ ,  $C^2 = 0.5$  and  $\gamma = 1.4$ . The solid lines are the numerical results obtained by using a  $80 \times 640$  grid in a  $1 \times 8$  computational domain.

which has solution

$$(3.2) \quad t - t_0 = \frac{1}{\sigma} \ln\left(\frac{v_t}{v_0}\right) + \left[\frac{1}{gR} - \left(\frac{1}{\sqrt{\sigma}} + \frac{1}{\sqrt{b}}\right)^2 \frac{1}{v_\infty}\right](v_t - v_0) - \frac{1}{b} \ln\left(\frac{v_\infty - v_t}{v_\infty - v_0}\right).$$

The first term in (3.2) is the contribution from the linear regime, the second is that of the free fall regime, and the third comes from the asymptotic terminal velocity. Extensive validation of this model has been performed for a range of Atwood numbers and compressibilities. The dependency of  $\sigma$ ,  $gR$ ,  $b$  and  $v_\infty$  on  $A$  and  $C$  is described in [35]. Fig. 4 shows a comparison between a numerical simulation of the full two dimensional Euler equations and the curve given by (3.2).

From a dimensional argument, the terminal velocity of the bubble should be proportional to  $\sqrt{\lambda g}$ , where the constant of proportionality  $c_1$  only depends on the dimensionless parameters  $A$ ,  $C$  and  $\gamma$ . Fig. 5 shows a plot of  $c_1$  for a range of Atwood numbers and compressibilities. We see that  $c_1$  depends strongly on  $C$  and for small fixed values of  $C$  is approximately  $\sqrt{A}$ . We did not explore the dependence of  $c_1$  on  $\gamma$  in this study.

#### 4. Interaction between bubbles.

Multiple bubble interactions are initialized with an ensemble of bubbles of different wavelengths. When started at small amplitudes, shorter wavelength bubbles have higher growth rates than the larger bubbles. However the short wavelength bubbles saturate out at smaller terminal velocities than the larger ones. Thus while the small bubbles initially run faster, the larger ones catch up and overtake them emerging on the outer envelope of the interface between the fluids. It was discovered that bubble interaction causes the terminal velocity of the large bubbles to

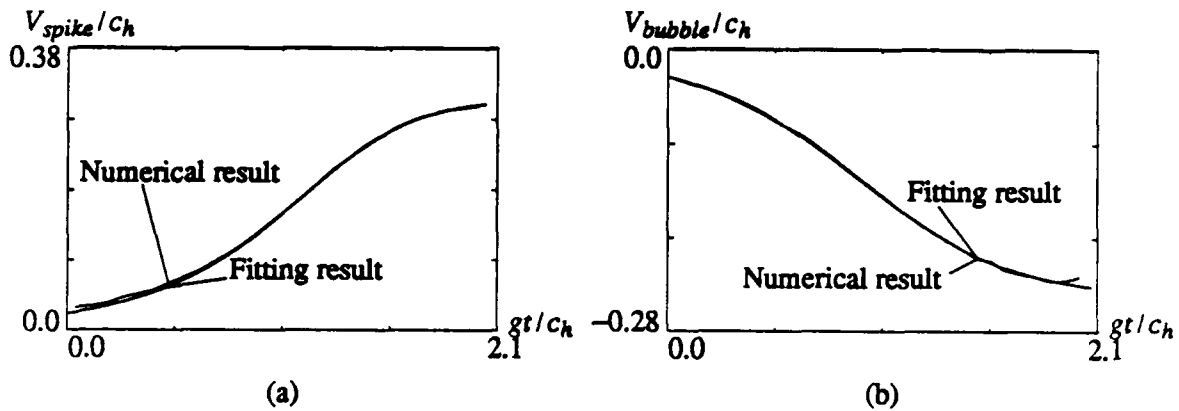


Fig. 4 Plots of spike velocity and bubble velocity versus time superimposed over the best three parameter fit to the solution of the ODE model. The parameter values are  $A = 1/3$ ,  $C^2 = 0.5$ , and  $\gamma = 1.4$ . The numerical results are obtained by using a  $80 \times 640$  grid in a  $1 \times 8$  computational domain.

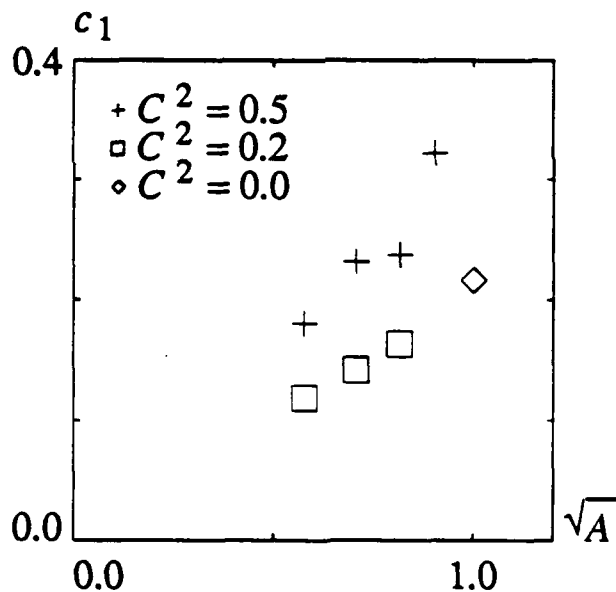


Fig. 5 The dependence of  $c_1$  on  $A$  and  $C$ . Note that  $c_1$  has a strong dependence on  $C$ . For a given value of  $C^2$  the dependence on  $A$  is approximately  $\sqrt{A}$  in systems of low compressibility. The value of  $c_1$  for an incompressible fluid ( $C^2 = 0$ ) is taken from reference [21].

exceed the prediction based on the single bubble theory for a bubble of comparable wavelength. As a large bubble overtakes a smaller one, it absorbs the velocity of

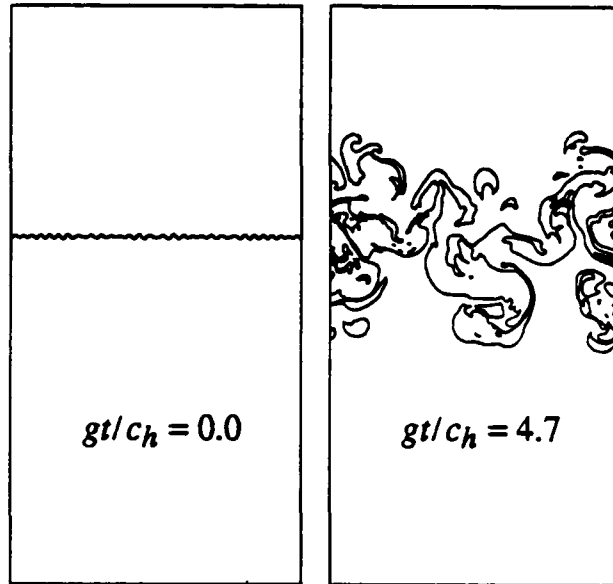


Fig. 6 Plots of interfaces in a random disturbance simulation of the Rayleigh-Taylor instability. The density ratio is  $A = 1/3$  and the compressibility is  $C^2 = 0.1$ . The acceleration of the bubble envelope is in good agreement with the experiment of Read for 1.5 generations of bubble merger. The acceleration decreases after this time due to the multiphase connectivity, which is different in the exactly two dimensional computation from the approximately two dimensional experiments. Gravity is directed downward.

that bubble which in turn is washed away downstream. We call this process bubble merger since it reduces the number of bubbles in the outer envelope. This number is reduced by a factor of  $2^n$  after  $n$  generations of bubble merger, a phenomenon that was observed in the experiments of Read [27] as well as in our numerical simulations [11]. The interface configuration of a random multiple bubble system is shown in Fig. 6. We see that the small structures (bubbles) merge into large structures.

We propose a simple superposition model for the bubble velocity in the chaotic regime. The basic idea is to treat the envelope of the bubbles as a single bubble of long wavelength. The velocity of individual bubbles as well as the bubble envelope are first computed based on the single bubble theory, the hypothesis is that to leading order the total velocity of each bubble is the sum of its single bubble theory velocity and the velocity of the envelope. More advanced bubbles are in phase with this envelope so the superposition is constructive and their velocity is increased. On the other hand a less advanced bubble is out of phase with the envelope causing its net velocity to be decreased. During the initial small amplitude regime, the envelope's longer wavelength causes its velocity to be dominated by the individual

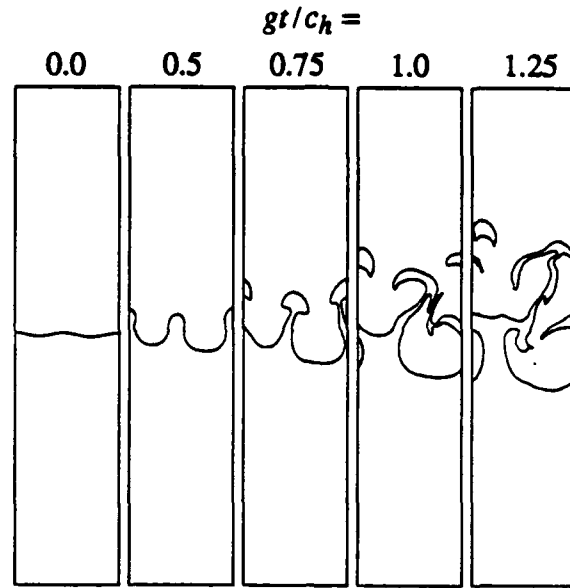


Fig. 7 Successive times in a two bubble merger process. The compressibility and Atwood number for this case are  $C^2 = 0.1$  and  $A = 2/3$  respectively. It can be seen that the large bubble overtakes the smaller one at  $gt/c_h = 1.0$ . The velocity of the large bubble is accelerated during the merger while the velocity of the small bubble is reversed, see Fig. 8.

bubbles, but at later times the envelope velocity is the main contribution to the bubble velocity.

We compared the results of this superposition model with the experimental results of Read [27] and our numerical simulations of the full Euler equations. The relative error between the superposition theory and the experimental or computational data was less than 20% for systems with  $A > \frac{2}{3}$  and  $C^2 \leq .1$ , and about 30% for systems with small Atwood numbers or large compressibility. In the latter case, the density stratification of the fluids cause the superposition principle to break down in finite time [11].

Fig. 7 shows the interface between two fluids at successive times in a two bubble merger process and Fig. 8 shows a comparison of the velocities of these bubbles and the predictions obtained from the superposition model. The behavior of the small bubble velocity clearly indicates the contribution from the envelope. At first the single mode bubble velocity dominates since the envelope has a small growth rate. The bubble stops accelerating when the small bubble and the envelope have equal but opposite velocities. After that, the envelope velocity dominates and the small bubble de-accelerates and is washed away downstream.

It would be expected from the existence of a terminal velocity in the single bubble

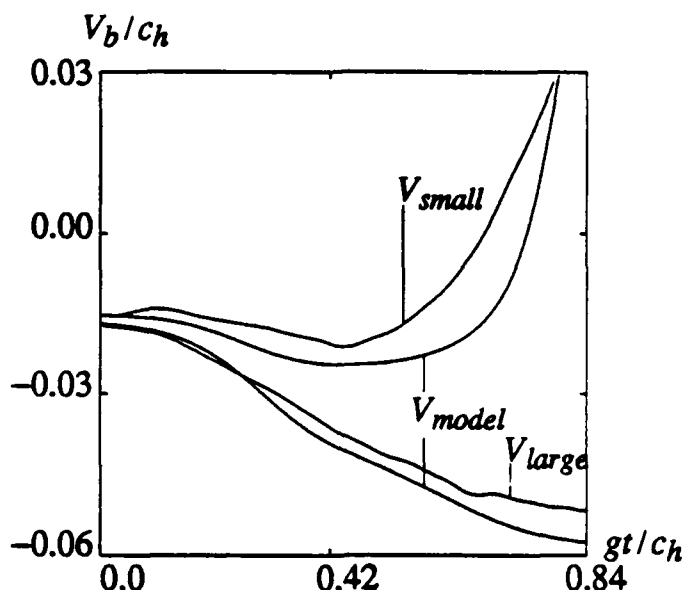


Fig. 8 A plot of bubble velocities vs. time for the two bubble merger simulation. The small bubble is accelerated at the beginning and is then decelerated after about  $gt/c_h = 0.42$ . The small bubble is washed downstream after its velocity is reversed, while the large bubble is under constant acceleration. The smooth curves represent the bubble motion as predicted by the superposition model.

theory that the asymptotic position of a bubble would be proportional to time. However for a chaotic flow interactions cause the radius of a large bubble to increase, consequently raising its terminal velocity. This led to the prediction that position of the tip of a large bubble is proportional to  $t^2$ ,  $z = \alpha A g t^2$ . In his experiments, Read [27] reported a range of values for  $\alpha$ , a typical value being  $\alpha = .06$ . Youngs [34] and Zufuria [36] reported values of  $\alpha$  ranging from  $0.04 \sim 0.05$  and  $0.05 \sim 0.06$  respectively, based on their numerical simulations. Our simulations indicated that  $\alpha$  is not a constant. Rather it varies during the interaction from an early value of  $0.055 \sim 0.065$  to  $0.038 \sim 0.044$  at late stages of the interaction [11]. The reduction of  $\alpha$  from about .06 to about .04 is due to the multi-connectivity of the interface in the deep chaotic regime. In Young's numerical simulations [34], the interface between two fluids was not tracked so that effective multi-connectivity occurred in the early stages of his simulations. This may explain the small values of  $\alpha$  which be observed. The discrepancy between the value of  $\alpha$  at late times in our numerical simulations and the value observed in Read's experiments results from the difference between an exact two dimensional numerical simulation and an approximately two dimensional experiment. In Read's experiments the ratio of width to thickness was six to one, and the isolated segments of fluids in the  $x - z$  plane for the computations

might be connected in the third dimension ( $y$  direction). Such discrepancies may be resolved in three dimensional calculations which will provide a more realistic approximation of the experimental conditions.

When compressibility effects are significant, the stratification of the density in the unperturbed fluid causes the effective Atwood number at the tip of a bubble to decrease as the bubble moves into the heavy fluid. The bubble velocity is non-monotone and may even reverse directions. Since this factor was not taken into account in the single mode theory, our superposition theory breaks down when the effective Atwood number has been substantially reduced. To get a better understanding of the phenomenon of velocity turnover in a single mode system and the failure of the superposition hypothesis in a multi-mode system, we use the initial density distribution of light and heavy fluid to approximate the effective dynamic Atwood number  $A_e$ . For a flat interface, the density distribution is

$$(4.1) \quad \rho_i(z) = \rho_i(0) \exp\left(\frac{\gamma g z}{c_i^2}\right), \quad i = l, h.$$

When a bubble reaches the position  $z$ , we approximate the effective Atwood number as

$$(4.2) \quad A_e(z) = \frac{\rho_h(z) - \rho_l(z)}{\rho_h(z) + \rho_l(z)} = \frac{(1 + A) \exp\left(\gamma C^2 \frac{2A}{1+A} \frac{z}{\lambda}\right) - (1 - A)}{(1 + A) \exp\left(\gamma C^2 \frac{2A}{1+A} \frac{z}{\lambda}\right) + (1 - A)}.$$

For a single mode system, the turnover phenomenon should occur before the effective Atwood number  $A_e$  vanishes. For a multi-mode system, the superposition model is applicable as long as  $A_e \approx A = A_e(z = 0)$ . In Fig. 9, we plot the approximate effective Atwood number versus  $\frac{z}{\lambda}$ . Since  $A_e$  decreases more rapidly in a system with a small Atwood number or large compressibility, the superposition model fails at a small value of  $\frac{z}{\lambda}$  in these systems.

One should not confuse the turnover of the bubble velocity in a single mode system with the turnover of the velocity of a small bubble in the multi-mode system. The former is due the stratified density distribution and latter is due to the interactions between bubbles, i.e., the contribution of the envelope velocity to the total velocity of the small bubble.

## 5. Elementary Wave Nodes and the Supersonic Steady State Riemann Problem.

We now turn our attention to an investigation of wave interactions between shock waves and fluid interfaces. An elementary wave node is a point of interaction between two waves that is both stationary and self-similar [14]. Gravity will be neglected here since the interactions considered occur on short time-scales. It can be shown [10], [19¶405-409] that there are four basic elementary nodes. These are the crossing of two shocks moving in opposite directions (cross node), the overtaking



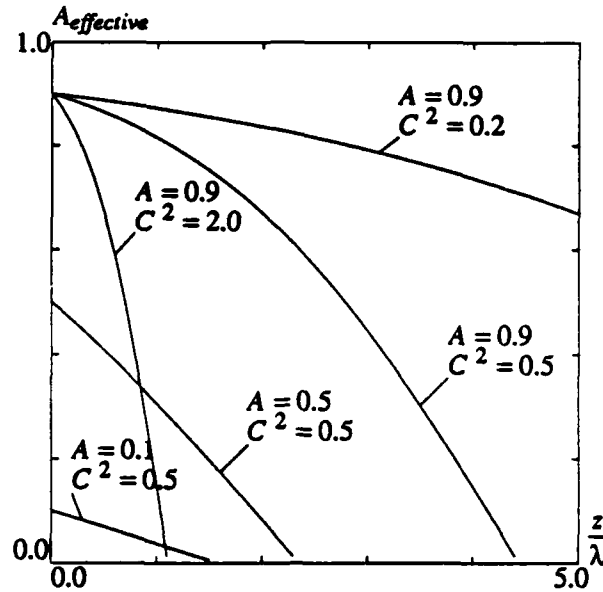


Fig. 9 The plot of approximate effective Atwood number as the bubble reaches position  $z$ .  $A_e$  decreases more rapidly in the system with small initial Atwood number or large compressibility than in the system with large initial Atwood number and small compressibility. The decreasing of the effective Atwood number is the source the turnover phenomenon in single mode system and the failure of the superposition model in multi-mode systems.

of one shock by another moving in the same direction (overtake node), the collision of a shock with a fluid interface (diffraction node), and the splitting of a shock wave due to interaction with other waves or boundaries to produce a Mach reflection (Mach node). All of these waves are characterized by the solution of a Riemann problem for a steady state flow, where the data is provided by the states behind the interacting waves. We will primarily be concerned with the diffraction node, but bifurcations in this node will lead to the production of all of the other elementary nodes.

For a stationary planar flow, system (2.1) reduces to a  $4 \times 4$  system that is hyperbolic in the restricted variables provided the Mach number  $M = |\mathbf{q}|/c$  is greater than one, i.e., the flow is supersonic. The streamlines or particle trajectories define the time-like direction. The hyperbolic modes in this case are associated with two families of sound waves, and a linearly degenerate double characteristic family. If  $\theta$  and  $q$  are the polar coordinates of the particle velocity  $\mathbf{q}$ , then the sonic waves have characteristic directions with polar angles  $\theta \pm A$ , where  $A$  is the Mach angle,  $\sin A = M^{-1}$ . Waves of these families are either stationary shock waves or steady state centered rarefaction waves called Prandtl-Meyer waves. Waves of

the degenerate family are a combination of a contact discontinuity and a vortex sheet across which the pressure and flow direction  $\theta$  are continuous while the other variables may experience jumps.

Following the general analysis of systems of hyperbolic conservation laws [20], we see that the wave curve for a sonic wave family consists of two branches corresponding to either a shock or a simple wave. The shock branch is commonly called a shock polar [6¶294–317] and forms a closed and bounded loop where the two sonic families meet at the point where the stationary shock is normal to the incoming flow. If we let the state ahead of the wave be denoted by the subscript 0, a straightforward derivation of the Rankine-Hugoniot equations for the system (2.1) shows that the thermodynamics of the states on either side of the shock are related by the Hugoniot equation

$$(5.1) \quad E = E_0 + \frac{P + P_0}{2}(V_0 - V).$$

A similar derivation applied to the steady state Euler equations shows that the flow velocities on either side of a stationary oblique shock satisfy

$$(5.2) \quad \frac{1}{2}q^2 + H = \frac{1}{2}q_0^2 + H_0,$$

where  $H = E + PV$  is the specific enthalpy. The jump in the flow direction is given by

$$(5.3) \quad \tan(\theta - \theta_0) = \pm \left[ \frac{P - P_0}{\rho_0 q_0^2 - (P - P_0)} \right] \cot \beta.$$

Here  $\beta$  is the angle between the incoming streamline and the shock wave, and is given by  $\sin \beta = \sigma/q_0$ , where  $\sigma = V_0 m$  is the wave speed of the shock wave with respect to the fluid ahead and  $m$  is the mass flux across the shock.  $m^2 = -\Delta P/\Delta V$ . The difference between the flow direction on either side of the shock is called the turning angle of the wave.

The same analysis when applied to the simple wave curves shows that the entropy is constant inside a Prandtl-Meyer wave. The flow speed and flow direction are related by (5.2) where  $H = H(P, S_0)$  and

$$(5.4) \quad \theta = \theta_0 \mp \int_{P_0}^P dP \frac{\cos A}{\rho c q} \Big|_S.$$

In analogy to the shock polar defined by (5.1)–(5.3) we will call this locus of states the rarefaction polar.

It is easily checked that the two branches of (5.4) are respectively associated with the  $\theta \pm A$  characteristic directions in the sense of Lax [20]. Similarly it can be shown

[15] that for most equations of state, the two branches of (5.3) are also associated with the  $\theta \pm A$  characteristics in the sense of Lax provided the state downstream from the shock is supersonic. Since  $\theta$  and  $P$  are constant across waves of the degenerate middle family, the Riemann problem for a stationary two-dimensional flow can be solved by finding the intersection of the projections of the wave curves in the  $\theta - P$  phase plane.

There are two major differences between the solution to the Riemann problem for a stationary flow and that of a one-dimensional unsteady flow. The Mach number behind the shock wave is given by

$$(5.5) \quad M = \frac{m}{\rho c} \left( 1 + \frac{\rho^2}{\rho_0^2} \cot^2 \beta \right)^{1/2} .$$

For most equations of state [22]  $m < \rho c$  and is a monotone function of the pressure along the shock Hugoniot. Thus if  $\beta$  is sufficiently close to  $\frac{\pi}{2}$  the flow behind the shock will be subsonic and the steady Euler equations cease to be hyperbolic. The second reason is that for an normal angle of incidence, the turning angle through the shock is zero. This means that the two branches of the shock polar meet at this point forming a closed and bounded loop. These two issues together imply a loss of existence and uniqueness for the solution to the two dimensional stationary Riemann problem. This means that that a bifurcation must occur from a stationary solution to a time dependent solution of the full two dimensional Euler equations.

The actual shape and properties of the shock and rarefaction polars depends on the equation of state. We will make no use of a specific choice of equation of state in our analysis, but we will need to assume that the equation of state satisfies appropriate conditions to guarantee that the shock polar has a unique point at which the state behind the shock becomes sonic, and a unique local extremum in the turning angle. These conditions are satisfied by most ordinary equations of state, and in particular by the polytropic and stiffened polytropic equations of state used in the numerical examples.

## 6. Anomalous Reflection.

As was mentioned above, the simplest case of shock diffraction is that in which the flow near a point of diffraction is scale invariant and pseudo-stationary. This will be the case provided the flow is sufficiently supersonic when measured in a frame that moves with the point [15]. Then the data behind the incoming waves define Riemann data for the downstream scattering of the interacting waves. A representative shock polar diagram for a regular shock diffraction producing a reflected Prandtl-Meyer wave is shown in Fig. 10.

Diffractions of these types have been studied experimentally by several investigators [1], [2], [17], [18], as well as numerically [5], [15]. Longer time simulations of the resulting surface instabilities in the fluid interface (called the Richtmyer-Meshkov instability [23], [28]) are found in [15], [24], [34]. One of the interferograms, Fig. 14 of [18] shows an irregular wave pattern that corresponds to what we

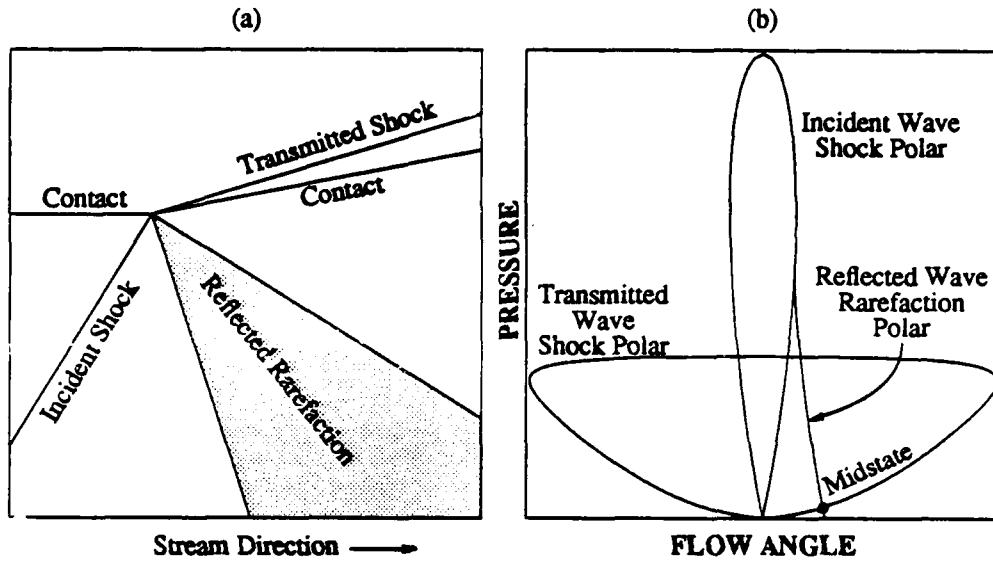


Fig. 10 A sketch of the wave pattern and polar diagrams for a regular shock-contact diffraction that produces a reflected rarefaction wave.

call an anomalous reflection. In this wave the angle between the incident shock and the material interface is such that the state behind the shock has become subsonic.

We consider the perturbation of a regular shock diffraction that produces a reflected Prandtl-Meyer wave. Suppose that initially the state behind the incident shock is close to but slightly below the sonic point on the incident shock polar. We allow the incident angle to increase while keeping the other variables constant so that the state behind the incident shock passes above the sonic point. Such a situation might occur as a shock diffracts through a bubble as illustrated in Fig. 11. When this happens, the solution can no longer be self-similar since a Prandtl-Meyer wave can only occur in supersonic flow. Instead the reflected wave begins to overtake and interact with the incident shock, Fig. 11c. This interaction dampens and curves the incident shock near its base on the fluid interface allowing the flow immediately behind the node to return to a supersonic condition. The single point of interaction bifurcates into a degenerate overtake node where the leading edge of the reflected rarefaction overtakes the incident shock, and a sonic diffraction node at the fluid interface. This interaction is a two-dimensional version of the one-dimensional overtaking of a shock by a rarefaction. The composite configuration is in many ways analogous to a regular Mach reflection. In this case the reflected wave is a Prandtl-Meyer wave and instead of a single point of Mach reflection the interaction is spread over the region where the rarefaction interacts with the incident shock. The "Mach" stem can be regarded as the entire region from the point where the incident shock is overtaken by the rarefaction to its base on the fluid interface.

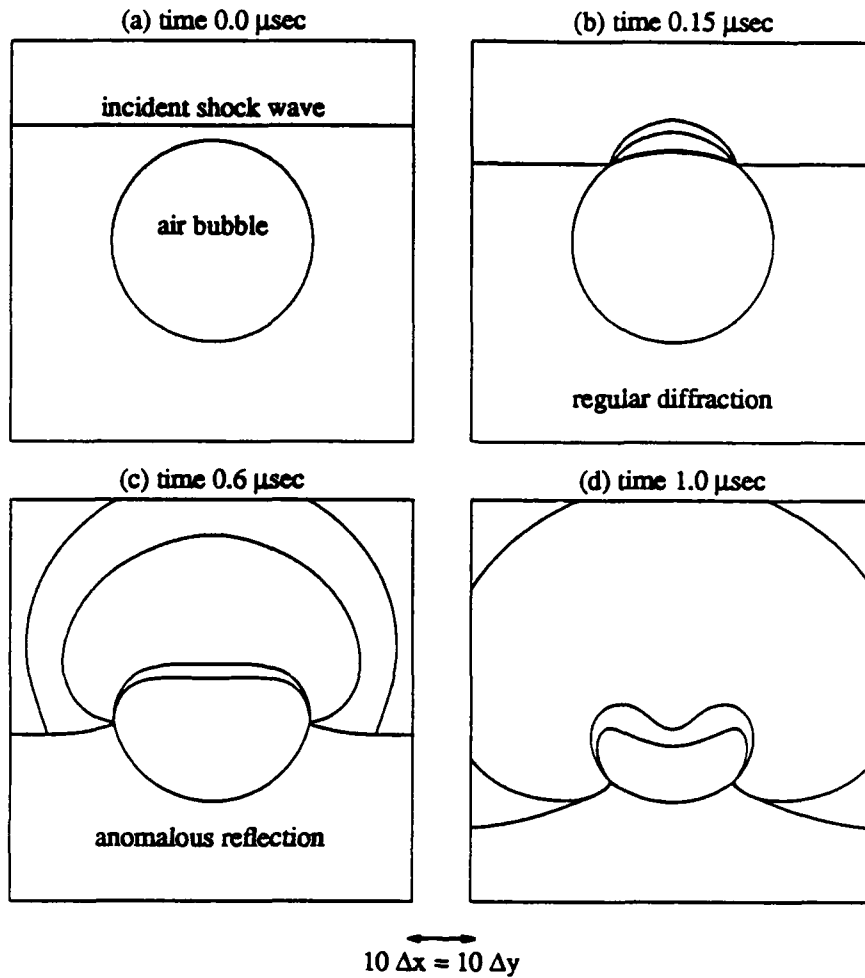


Fig. 11 The collision of a shock wave in water with an air bubble. The fluids ahead of the shock are at normal conditions of 1 atm. pressure, with the density of water 1 g/cc and air 0.0012 g/cc. The pressure behind the incident shock is 10 Kbar with a shocked water density of 1.195 g/cc. The grid is  $60 \times 60$ .

If we allow the incident angle to increase further we will eventually see a second bifurcation in the solution, Fig. 11d. As the material interface continues to diverge from the incident shock, the Mach number near the trailing edge of the reflected rarefaction continues to decrease. The characteristics behind the incident shock are almost parallel to the shock interface near the base of the anomalous reflection. The flow there becomes nearly one-dimensional and the rarefaction wave eventually overtakes the incident shock. If there is a great difference in the acoustic impedance between the two materials as in the numerical cases studied here, this second bifurcation will occur as the strength of the incident shock at the fluid interface reduces

to zero. The now non-centered rarefaction breaks loose from the fluid interface and begins to propagate away. This second configuration is also analogous to a Mach reflection. Here the Mach node corresponds to the interaction region between the rarefaction and incident shock, while the Mach stem is the degenerate wave portion from the trailing edge of the rarefaction to the fluid interface.

### 7. 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.

The tracking of a regular shock diffraction was described in [15]. The first step in the propagation is the computation of the velocity of the diffraction node with respect to the computational (lab) reference frame. Suppose at time  $t$  the node is located at point  $p_{00}$ . The node position at time  $t + dt$  is found by computing the intersection between the two propagated segments of the incident waves. If this new node position is  $p_0$ , then the node velocity is given by  $(p_0 - p_{00})/dt$ . This velocity defines the Galilean transformation into a frame where the node is at rest. When the state behind the incident shock is supersonic in this frame, it together with the state on the opposite side of the fluid 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 will occur if the calculated node velocity is such that the state behind the incident shock is subsonic in the frame of the node. If the reflected wave is a Prandtl-Meyer wave this will result in an anomalous reflection. The front tracking implementation of this bifurcation is a straightforward application of the analysis described in the previous section.

First the leading edge of the reflected rarefaction is allowed to break loose from the diffraction node. The intersection  $p_1$  between the propagated rarefaction leading edge and the incident shock is computed and a new overtake node is installed at  $p_1$  by disconnecting the rarefaction leading edge from the diffraction node and connecting it to  $p_1$ .

If this reflected rarefaction edge is untracked, then  $p_1$  is found by calculating the characteristic through the old node position corresponding to the state behind the incident shock and computing the intersection of its propagated position with the propagated incident shock. This characteristic makes the Mach angle  $A$  with the streamline through the node. Since the bifurcation occurs between times  $t$  and  $t + dt$ ,  $M \geq 1$  at time  $t$  and  $A$  is real. This wave moves with sound speed in its normal direction. In this case no new overtake node is tracked.

We are now ready to compute the states and position of the point of shock diffraction after the bifurcation. As was mentioned previously, the rarefaction expands onto the incident shock causing it to weaken. This in turn slows down the node causing the incident shock to curve into the fluid interface. The diffraction node will slow

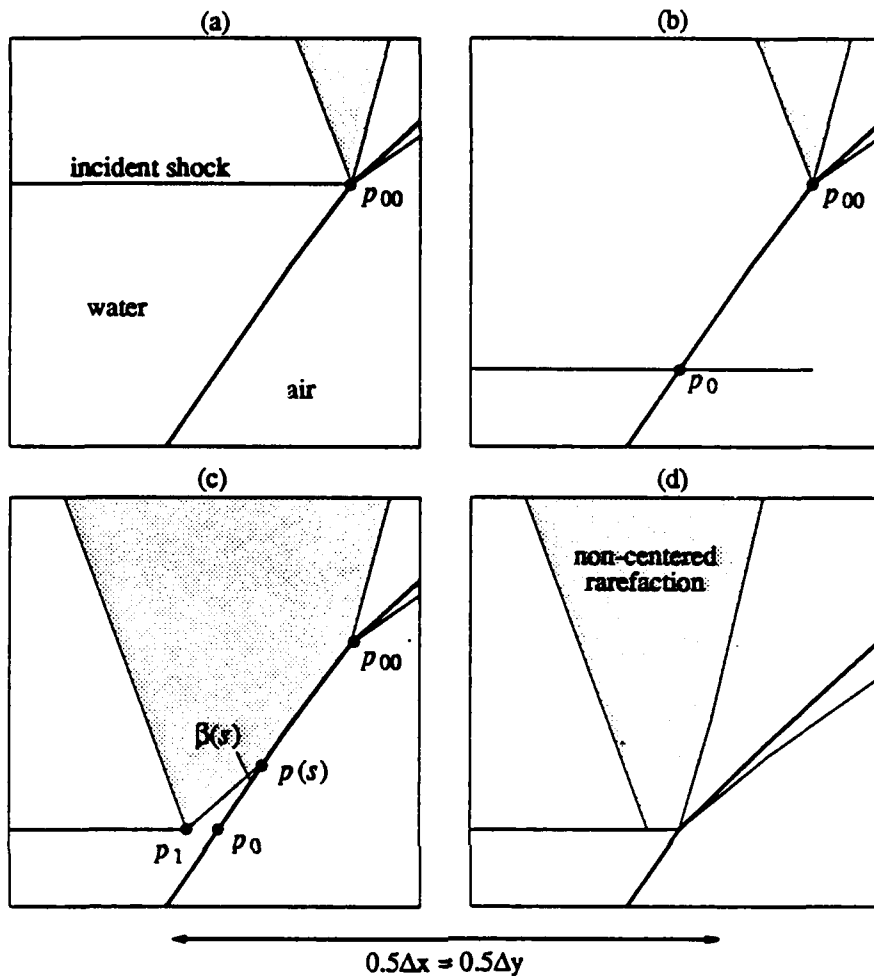


Fig. 12 A diffraction node initially at  $p_{00}$  bifurcates into an anomalous reflection. The predicted new node position at  $p_0$  yields a Mach number of 0.984 behind the incident shock. The leading edge of the reflected Prandtl-Meyer wave breaks away from the diffraction node to form an overtake node at  $p_1$ . The propagated position of the diffraction node is adjusted to return the flow to sonic behind the node.

down to the point where the state immediately behind the node becomes sonic. After this the configuration near the node can be computed using the regular case analysis.

The adjusted propagated node position is computed as follows, see Fig. 12. For each number  $s$  sufficiently small, 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  $p_1$ . Let  $v(s)$  be the node velocity found by moving the diffraction node to position  $p(s)$ , and let  $q(s)$  be the velocity of the flow ahead of the incident shock in the frame that moves with velocity  $v(s)$  with respect to the lab frame. The mass flux across this shock is given by

$$(7.1) \quad m(s) = \rho_0 |q(s)| \sin \beta(s) .$$

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  $s^*$  is the root of the equation  $M(s^*) = 1$ . 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 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, *i.e.*, 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. This is a purely numerical effect since physically the upstream state is always supersonic. However in nearly sonic cases such numerical undershoot can occur. If such a situation is detected the trailing edge of the reflected rarefaction wave is disengaged from the anomalous reflection node and installed at a new overtake node on the incident shock. The residual shock strength for the portion of the incident shock behind the rarefaction wave is small and the diffraction node at the material interface reduces to the degenerate case of a sonic signal diffracting through a material interface.

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. As before the tracked characteristic is disengaged from the diffraction node and a new overtake node is installed at the point of intersection. The residual shock strength at the node is non-zero so the diffraction at the material interface produces an additional expansion wave behind the original one. This new expansion wave is not tracked.

It is possible to make a few remarks about the amount of tracking required for these problems. Since the front tracking method is coupled to a finite difference method for the solution away from the tracked interface (the interior solver), there is always an option between tracking a wave or allowing it to be captured. Of course



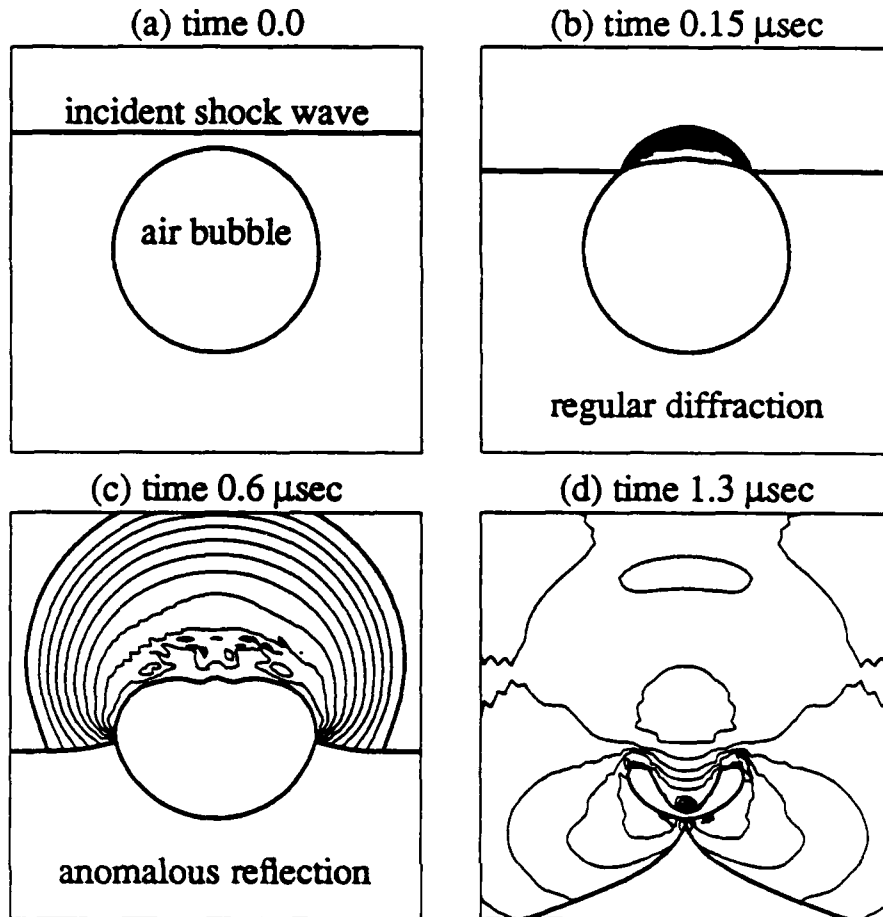
capturing can result in a considerable loss in resolution in the waves as compared to tracking [9], but it will also simplify the resolution of the interactions. The secondary bifurcations described above are only tracked when the trailing edge of the reflected Prandtl-Meyer wave is tracked. The current algorithm is structured so that at a minimum the two interacting incoming waves are tracked. At this extreme none of the outgoing waves are tracked and no explicit bifurcations in the tracked interface occur. More commonly, the material interface separates different fluids and so must be tracked on both sides of the interaction. Also, instabilities in the finite difference approximation can affect the accuracy of the solution near the node, especially for stiff materials such as water. Tracking the additional waves seems to considerably reduce these problems. Tracking also allows the use of a much coarser grid, which is important when the diffraction occurs in a small but important zone of a larger simulation. It allows the entire region of diffraction to extend only over only a fraction of a grid block. These remarks show 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.

## 8. Numerical Examples.

Fig. 13 shows a series of frames documenting the collision of a 10 Kbar shock wave with a bubble of air in water. Note in this case the trailing edge of the reflected Prandtl-Meyer wave is not tracked. The states ahead of the incident shock are at one atmosphere pressure and standard temperature. Under these conditions, water is about a thousand times as dense as air. During the initial stage of the interaction regular diffraction patterns are produced.

In less than half of a microsecond an anomalous reflection has formed, and by one microsecond 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. Long time simulations are expected to show the initial bubble split, and the resulting bubbles going into oscillation as they are overcompressed and then expand. This process is important in the transfer of energy as a shock passes through a bubbly fluid. The first diffraction considerably dampens the shock, and much of this energy will eventually be returned to the shock wave in the form of compression waves generated by the expanding bubbles.

Fig. 14 shows the diffraction of an expanding underwater shock wave through the water's surface. Initially a ten Kbar cylindrically expanding shock wave with a radius of one meter is placed two meters below the water's surface. The interior of the shock wave contains a bubble of hot dense gas. The states exterior to the shock are ambient at one atmosphere pressure and normal temperature. A gravitational acceleration of one  $g$  has been added in this case, but due to the rapid time scale on which the diffractions occur the effect of gravity is negligible. Here the entire reflected Prandtl-Meyer wave is captured rather than tracked. The pressure contour plots show that by six milliseconds an anomalous reflection has developed as



$\overleftrightarrow{10 \Delta x = 10 \Delta y}$   
 Fig. 13  $\text{Log}(1 + \text{pressure})$  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 tracked interface is shown in a dark line. The grid is  $60 \times 60$ .

indicated in the blowup of Fig. 14b shown in Fig. 15. 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. When the bubble reaches the surface it expands into the atmosphere leading 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 example of the elementary wave called the cross node where two oblique shocks collide.

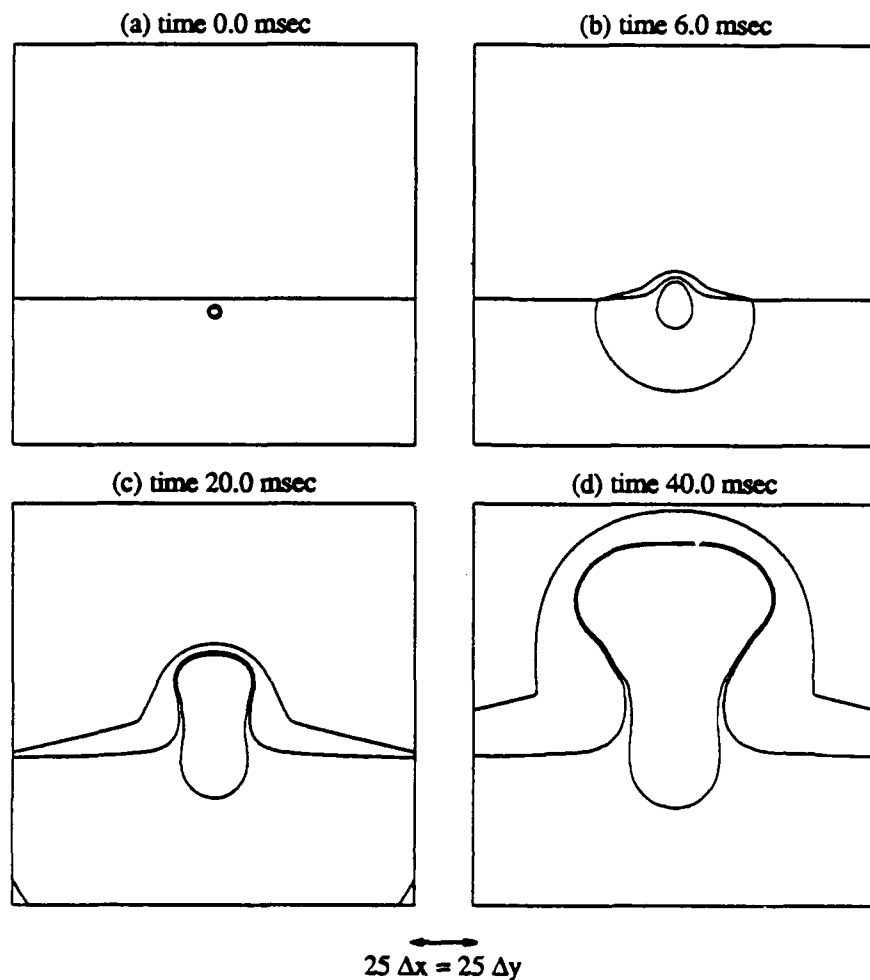
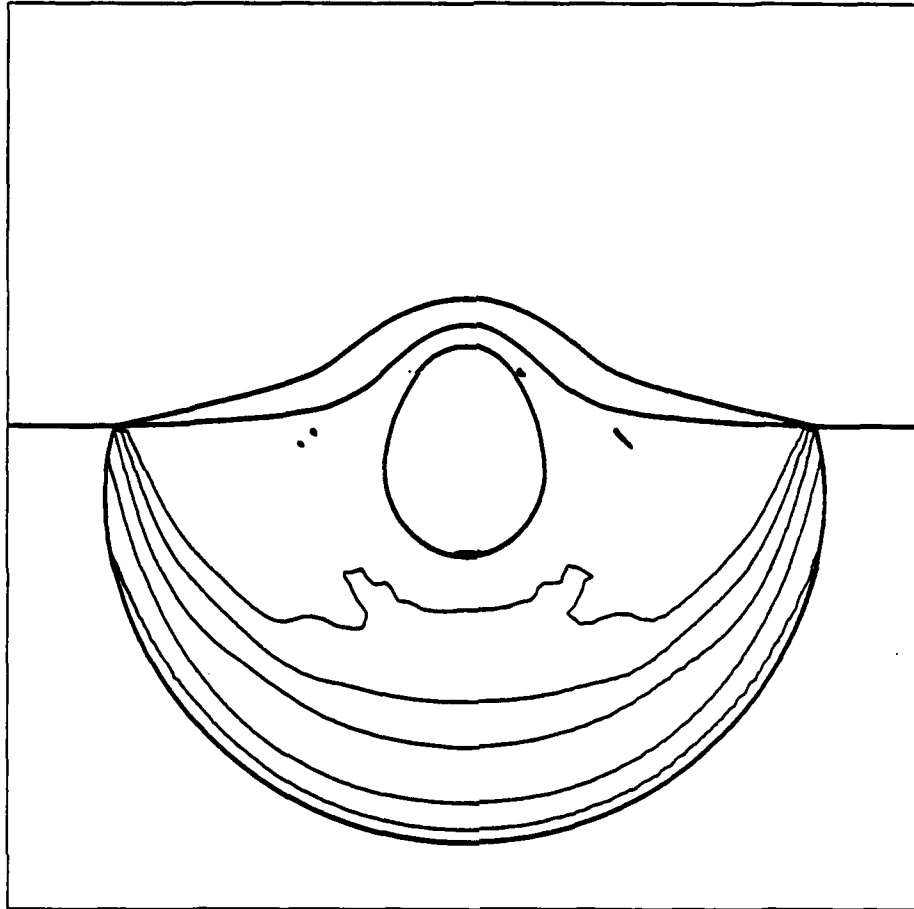


Fig. 14 An underwater expanding shock wave diffracting through the water's surface. An expanding shock wave with an internal pressure of 10 Kbars and initial radius of 1 meter is installed at a depth of 2 meters below the water's surface. The external conditions are ambient at one atmosphere pressure and normal densities for the air and water. The boundary conditions are constant Dirichlet at the initial ambient values. The grid is  $150 \times 150$ .

The water in the simulations described above is modeled by the stiffened polytropic equation of state with  $\Gamma_0 = 6$ ,  $E_\infty = 0$ , and  $P_\infty = 3000$  atm. The air is treated as a polytropic gas with  $\gamma = 1.4$ .

### 9. Summary.

We have seen that the process of bubble growth and interaction in gravity driven mixing can be modeled using a simplified description of the bubble dynamics at least in the small compressibility regime. In this regime, the model agrees with



$$25 \Delta x = 25 \Delta y$$

Fig. 15 A blowup of Fig. 14.1b showing pressure contours scaled from 0.001 - 10 Kbars. The tracked interface is shown superimposed in a dark line over the pressure contours.

experiments and computer simulations well into the chaotic regime. It should be possible to include results of this simplified model in statistical models that can study the interaction of large numbers of bodies.

We also studied the diffraction of a shock through a material interface from a medium of high to low acoustic impedance. The bifurcations that occur during the diffraction were analyzed in terms of polar diagrams for steady supersonic flow. This analysis was incorporated into a front tracking code to allow enhanced resolution computations of the interactions. The particular simulations studied were the diffraction of a planar shock in water through an air bubble, and the diffraction of an expanding shock in water through the water's surface. In both cases the anomalous reflection bifurcation plays an important role in correctly computing the

flow.

We would like to thank James Glimm, Xiao Lin Li, and David H. Sharp who have been actively involved in the analysis of the Rayleigh–Taylor problem and the Institute of Mathematics and its Applications for providing us with the use of a CRAY-2 for portions of our study of the single Rayleigh–Taylor mode problem.

#### REFERENCES

- [1] A. M. ABD-EL-FATTAH AND L. F. HENDERSON, *Shock Waves at a Fast-Slow Gas Interface*, J. Fluid Mech., 86 (1978a), pp. 15–32.
- [2] ———, *Shock Waves at a Slow-Fast Gas Interface*, J. Fluid Mech., 89 (1978b), pp. 79–95.
- [3] I. B. BERNSTEIN AND D. L. BOOK, *Effect of Compressibility on the Rayleigh-Taylor Instability*, Phys. Fluids, 26 (1983), pp. 453–458.
- [4] G. BIRKHOFF AND D. CARTER, *Rising Plane Bubbles*, J. Math. Mech., 6 (1957), p. 769.
- [5] P. COLELLA, L. F. HENDERSON AND E. G. PUCKETT, *A Numerical Study of Shock Wave Refraction at a Gas Interface*, LLNL Preprint UCRL-100260, 1989.
- [6] R. COURANT AND K. O. FRIEDRICHS, *Supersonic Flow and Shock Waves*, Springer Verlag, New York, 1948.
- [7] P. R. GARABEDIAN, *On Steady-State Bubbles Generated by Taylor Instability*, Proc. R. Soc. London A, 241 (1957), pp. 423–431.
- [8] C. L. GARDNER, J. GLIMM, O. MCBRYAN, R. MENIKOFF, D. SHARP AND Q. ZHANG, *The Dynamics of Bubble Growth for Rayleigh-Taylor Unstable Interfaces*, Phys. of Fluids, 31 (1988), pp. 447–465.
- [9] J. GLIMM, J. GROVE AND X. L. LI, *Three Remarks on the Front Tracking Method*, Proceedings of the conference in Taormina Sicily, 1987.
- [10] J. GLIMM, C. KLINGENBERG, O. MCBRYAN, B. PLOHR, D. SHARP AND S. YANIV, *Front Tracking and Two Dimensional Riemann Problems*, Adv. Appl. Math., 6 (1985), pp. 259–290.
- [11] J. GLIMM, X. L. LI, R. MENIKOFF, D. H. SHARP AND Q. ZHANG, *A Numerical Study of Bubble Interactions in Rayleigh-Taylor Instability for Compressible Fluids*, To appear.
- [12] J. GLIMM AND X. L. LI, *On the Validation of the Sharp-Wheeler Bubble Merger Model from Experimental and Computational Data*, Phys. Fluids, 31 (1988), pp. 2077–2085.
- [13] J. GLIMM, O. MCBRYAN, D. SHARP AND R. MENIKOFF, *Front Tracking Applied to Rayleigh Taylor Instability*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 230–251.
- [14] J. GLIMM AND D. SHARP, *An S-matrix Theory for Classical Nonlinear Physics*, Foundations of Physics, 16 (1986), pp. 125–141.
- [15] J. W. GROVE, *The Interaction of Shock Waves with Fluid Interfaces*, Adv. Appl. Math., 10 (1989), pp. 201–227.
- [16] F. H. HARLOW 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.
- [17] L. F. HENDERSON, *On the Refraction of Longitudinal Waves in Compressible Media*, LLNL Report UCRL-53853, 1988.
- [18] R. G. JAHN, *The Refraction of Shock Waves at a Gaseous Interface*, J. Fluid Mech., 1 (1956), pp. 457–489.
- [19] L. LANDAU AND E. LIFSHITZ, *Fluid Mechanics*, Addison-Wesley, Reading, Mass., 1959.
- [20] P. LAX, *Hyperbolic Systems of Conservation Laws II*, Comm. Pure Appl. Math., 10 (1957), pp. 537–556.

- [21] R. MENIHOFF AND C. ZEMACH, *Rayleigh-Taylor Instability and Use of Conformal Maps for Ideal Fluid Flow*, J. Comp. Phys., 51 (1983), pp. 28-64.
- [22] R. MENIHOFF AND B. PLOHR, *Riemann Problem for Fluid Flow of Real Materials*, Revs. Mod. Phys., 61 (1989), pp. 75-130.
- [23] E. E. MESHKOV, *Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaz.*, 5 (1969), p. 151.
- [24] K. O. MIKAELIAN, *Simulation of the Richtmyer-Meshkov Instability and Turbulent Mixing in Shock-Tube Experiments*, LLNL Preprint UCID-21328, Jan. 1988.
- [25] B. PLOHR, *Shockless Acceleration of Thin Plates Modeled by a Tracked Random Choice Method*, AIAA J., 26 (1988), pp. 470-478.
- [26] L. RAYLEIGH, *Investigation of the Character of the Equilibrium of an Incompressible Heavy Fluid of Variable Density*, in *Scientific Papers*, II, Cambridge Univ. Press, Cambridge, England, 1900, p. 200.
- [27] K. I. READ, *Experimental Investigation of Turbulent Mixing by Rayleigh-Taylor Instability*, Physica D, 12 (1984), p. 45.
- [28] R. D. RICHTMYER, *Taylor Instability in Shock Acceleration of Compressible Fluids*, Comm. Pure and Appl. Math, 13 (1960), pp. 297-319.
- [29] D. H. SHARP, *An Overview of Rayleigh-Taylor Instability*, Physica D, 12 (1984), pp. 3-18.
- [30] D. H. SHARP AND J. A. WHEELER, *Late Stage of Rayleigh-Taylor Instability*, Institute of Defense Analyses. Unpublished Technical Report, 1961.
- [31] G. I. TAYLOR, *The Instability of Liquid Surfaces When Accelerated in a direction Perpendicular to Their Planes I*, Proc. R Soc. London A, 201 (1950), pp. 192-196.
- [32] G. TRYGGVASON, *Numerical Simulations of the Rayleigh-Taylor Instability*, J. Comp. Phys., 75 (1988), pp. 253-282.
- [33] C. P. VERDON, R. L. MCCRORY, R. L. MORSE, G. R. BAKER, D. I. MEIRON AND S. A. ORSZAG, *Nonlinear Effects of Multifrequency Hydrodynamic Instabilities on Ablatively Accelerated Thin Shells*, Phys. Fluids, 25 (1982), pp. 1653-1674.
- [34] D. L. YOUNGS, *Numerical Simulation of Turbulent Mixing by Rayleigh-Taylor Instability*, Physica D, 12 (1984), pp. 32-34.
- [35] Q. ZHANG, *A Model for the Motion of Single Bubble and Spike in Rayleigh-Taylor Instability*, To appear.
- [36] J. A. ZUFIRIA, *Vortex-in-Cell Simulation of Bubble Competition in Rayleigh-Taylor Instability*, Phys. Fluids, 31 (1988), pp. 440-446.

# CHARACTERISTICS AND STABILITY IMPLICATIONS OF A STREAMWISE VORTEX IN BOUNDED SHEAR FLOW

Joseph D. Myers  
Department of Mathematics  
United States Military Academy  
West Point, New York 10996

Frederick H. Abernathy  
Division of Applied Sciences  
Harvard University  
Cambridge, Massachusetts 02138

September 19, 1989

## Abstract

The characteristics of a single streamwise vortex embedded in Poiseuille flow have been analyzed both numerically and analytically. On short time scales, velocity profiles are found to evolve similarly to those previously derived in unbounded Couette flow. As wall effects begin to be felt at later times, a perturbation solution is derived whose profiles and decay rates are found to agree with calculations. Both solutions are used to answer questions about the strength and motion of the vortex, the presence and strength of any counter-rotating vortices induced at the viscous wall, and the possible (or impossible) role of such induced vortices in transition. Time scale arguments are used to make inferences about the stability of various initial distributions of streamwise vorticity, and in particular to derive a neutral curve (critical Reynolds number vs vertical position in flow) with a minimum critical Reynolds number close to 1000.

## Contents

1	INTRODUCTION	2
2	PROBLEM FORMULATION	5
2.1	GEOMETRY	5
2.2	ASSUMPTIONS	5
2.3	EQUATIONS	5
2.4	BOUNDARY CONDITIONS	7
2.5	INITIAL CONDITIONS	7
3	COMPUTATIONAL SOLUTION	9
4	BENCHMARKS	9
5	STREAMWISE VELOCITY PROFILES	12
6	COUNTER-ROTATING VORTEX	16

<b>7</b>	<b>FINITE DISK OF VORTICITY</b>	<b>21</b>
<b>8</b>	<b>DUPLICATING YANG'S SECONDARY STRUCTURE</b>	<b>24</b>
<b>9</b>	<b>STABILITY IMPLICATIONS</b>	<b>26</b>
<b>10</b>	<b>STABILITY OF CORE VS UPFLOW REGION</b>	<b>29</b>
<b>11</b>	<b>CONCLUSIONS</b>	<b>34</b>
<b>A</b>	<b>PERTURBATION SOLUTION</b>	<b>35</b>
	A.1 CROSSFLOW SOLUTION . . . . .	35
	A.2 STREAMWISE SOLUTION . . . . .	40
<b>B</b>	<b>EFFECT OF FREE SURFACE</b>	<b>45</b>
<b>C</b>	<b>COMPARISON WITH SEMI-INFINITE DOMAIN</b>	<b>52</b>

## 1 INTRODUCTION

Laminar-turbulent transition is still an incompletely understood phenomenon. Observations disprove the conjecture that transition is governed by the Reynolds number,  $R$ . Linear stability theory correctly explains phenomena within its regime (eg, instability of plane parabolic flow to infinitesimal perturbations above a critical Reynolds number  $R_c = 5772$  based on surface velocity and flow depth), but it turns out that Tollmein-Schlichting waves are neither necessary nor sufficient for transition; not necessary since subcritical transition is observed at  $R$  down to about 1000, and not sufficient in the sense that, when present, their most unstable mode grows on a diffusive time scale, and in the absence of a secondary three-dimensional perturbation, these modes do not become significant until very large times (eg, Nishioka et al, 1975). 2-D finite amplitude perturbations predict  $R_c \approx 2800$  in plane parabolic flow, but fail to account for observed transition down to  $R_c \approx 1000$ , and produce profiles that always saturate in amplitude before decaying back to laminar (Herbert, 1977). Additionally, some other flows (eg, plane Couette) exhibit no 2-D finite amplitude growing modes. 2-D finite amplitude waves have been found to have a strong secondary instability in the presence of a small 3-D disturbance at  $R > 1200$  in rough agreement with observations (Orszag and Patera, 1983), but there is no rational mechanism to generate these finite amplitude waves in a subcritical flow. Statistical methods have the problem of closure from a theoretical standpoint; empirically, they perform well in known regimes, but extrapolate poorly to new ones, and tell us little of the physics of transition. At a very practical level, an understanding of the physical processes involved is desirable for use when, for example, designing for reduced drag or designing for increased mixing.

Much attention has been devoted recently to investigating the role of streamwise vorticity in shear flow transition. The experiments of Klebanoff et al (1962) showed that 2-D waves developed a spanwise waviness that generated 3-D streamwise vortices, which in turn broke down as part of the transition cycle. Taylor vortices, Görtler vortices, the trailing vortices from a boundary roughness element, sweeps from the outer flow that pull up spanwise vortex tubes which are subsequently stretched by the shear to form hairpin vortices - all these are observed to be possible precursors to transition. The common element in all is the presence of the streamwise vortex in shear flow (eg, Figure 1).

As the simplest abstraction, Pearson and Abernathy (1984) investigated a single infinitely long vortex aligned in the flow direction in an unbounded uniform shear flow (Figure 2). They found that



the full 3-D incompressible Navier-Stokes equations reduced to one ODE in a similarity variable. This solution depends on the vortex Reynolds number ( $R_v \equiv \frac{\Gamma_0}{2\pi\nu}$  - a measure of the vortex strength; specifically, the ratio of the circulation rate to diffusion rate) and not on the flow Reynolds number,  $R$ . The flow Reynolds number sets the local shear, which in turn sets the time scale of the instability. Figure 3 shows the perturbed streamwise velocity profiles as caused by vortices of different strengths. The effect of the vortex is to rotate low velocity fluid from bottom to top and high velocity fluid from top to bottom. The resulting inflectional profiles are reminiscent of those treated in inviscid stability theory. A linear stability analysis of these profiles in a viscous fluid shows them to be unstable for  $R_v > 2 - 3$  (and for any  $R$ ). Yang and Abernathy (1987) generated what they believed to be a single vortex in plane parabolic flow and produced similar velocity profiles, but observed additional structure near the bottom viscous wall. Suri and Abernathy (1988) investigated the effect of a diffuse array of vortices above a viscous wall, but found profiles that took much longer to develop and were less inflectional.

These results raise additional questions. Do Pearson's inflectional profiles generalize to more realistic (eg, bounded, non-Oseen vortex) flows? Or are these profiles just peculiar artifacts of starting with a delta function of vorticity, with more realistic initial conditions yielding much less perturbation of the streamwise flow field? Even if more realistic initial conditions yield similar profiles, does the viscous wall reduce circulation enough to make the vortices (and thus their reduced inflectional profiles) just innocuous observers during the transition sequence? How do Suri's velocity profiles tie in with those of Pearson? Does a single streamwise vortex kick up enough added structure at the viscous wall so as to not remain a single vortex, and thereby cast doubt on Yang's conclusions about the role of the single vortex in transition? If  $R$  is not the controlling factor in transition, what is its role? This paper attempts to answer these questions.



Figure 1: Boundary layer transition on a flat plate (Werlé 1980). Note the initially counter-rotating vortex pair at left, as evidenced by the contracting flow near the plate and the upflow region at the center.

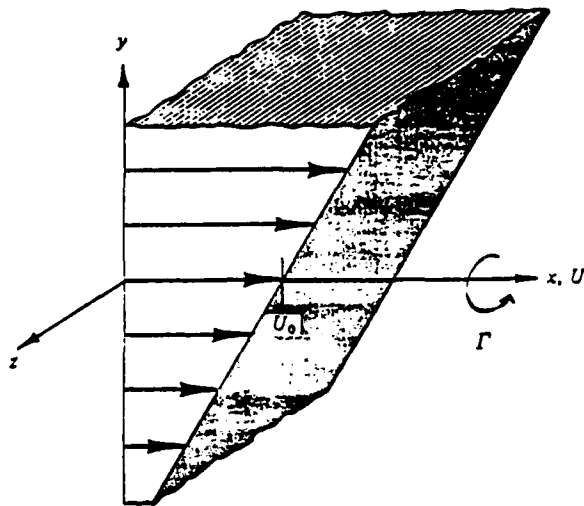


Figure 2: Pearson and Abernathy's (1984) undisturbed flow configuration. Flow is an unbounded uniform shear. A line vortex of strength  $R_v$  is inserted, coincident with the  $x$  axis, at time zero.

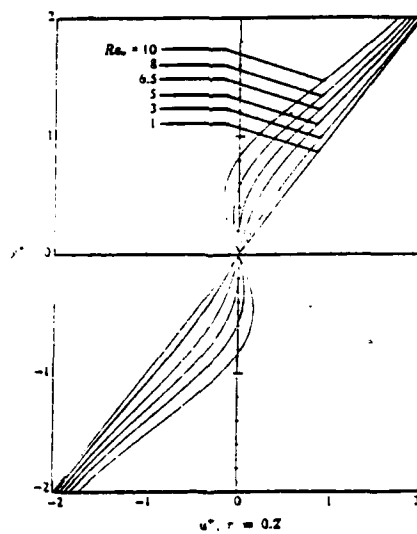


Figure 3: Dimensionless vertical profiles of the streamwise velocity after being perturbed by stream-wise vortices of various strengths  $R_v$ . Profiles are calculated after a constant time interval. From Pearson and Abernathy (1984).

## 2 PROBLEM FORMULATION

### 2.1 GEOMETRY

We begin by adding another layer of realism to Pearson: we consider a bounded region with a viscous wall, a parabolic mean profile, and initial distributions of vorticity other than potential. Specifically, we consider Poiseuille flow of infinite extent in the streamwise ( $x$ ) and spanwise ( $z$ ) directions over a viscous plane. A free surface is located at height  $h$  with streamwise velocity  $u_0$ . The wall (and  $x$  axis) are at an angle  $\theta$  from horizontal, and it is the body force due to the component of gravity in the  $x$  direction that generates the mean velocity profile. Embedded in the mean flow is a periodic array of vortex pairs aligned in the streamwise direction, each of initial circulation  $\Gamma_0$ . The ratio of the average vortex spacing to flow depth is the aspect ratio,  $a$ . (See Figure 4). The case of a single streamwise vortex in an infinitely wide box is obtained in the limit  $a \rightarrow \infty$ . Since we will often be interested in this limit, we plot the independent variable  $\frac{1}{a}$  rather than  $a$  in order to put this limit at the origin. Vertical streamlines dividing a vortex from its images on the left and right are located at  $z = \pm \frac{a}{2}$ . These streamlines can equivalently be thought of as slippery walls - slippery in the sense that boundary conditions (Section 2.4 below) require no shear stress on them. For later reference, we denote the left wall as  $\partial_{left}$ , the free surface as  $\partial_{top}$ , the right wall as  $\partial_{right}$ , and the bottom wall as  $\partial_{bottom}$ .

### 2.2 ASSUMPTIONS

1. Like Pearson, Yang, and Suri, we assume no  $x$  dependence in the flow at  $t = 0$ . This is based on the observation that the flow variables change slowly in the  $x$  direction compared to changes in the  $y$  and  $z$  directions (eg, Figure 1).
2. We assume that the free surface is constrained by a normal force to a horizontal plane. This assumption allows us to eliminate one variable (free surface position) from the equations, but also means that additional circulation is lost through the constrained free surface (which is now essentially another slippery boundary caused by an image vortex). This assumption seems warranted based on experimental observations; in Appendix B, we demonstrate its consistency.
3. At  $t = 0$  we are capable of using any initial distribution of vorticity, but for simplicity of analysis we choose to use either a delta function of vorticity (like Pearson and Yang), a fully diffused (in a sense to be explained later) vortex (like Suri), or a cylindrical tophat function of vorticity as a new intermediate case.

### 2.3 EQUATIONS

We begin with the 3-D incompressible Navier-Stokes equations. We nondimensionalize with respect to the characteristic variables listed in Table 1. We choose these characteristic quantities on the purely pragmatic grounds that they result in the greatest simplification in the appearance of subsequent formulas. All variables throughout this paper are dimensionless, except where noted.

The lack of  $x$  dependence at  $t = 0$  implies that there will be no  $x$  dependence at any later time. This eliminates the streamwise velocity  $u$  from the crossflow equations, resulting in a partial decoupling first noticed by Mitchner (1952) and by Stuart (1965). This allows a stream function-vorticity formulation in the crossflow directions. The equations become:

Crossflow:

$$\omega_t + v\omega_y + w\omega_z = \frac{1}{R_\nu}(\omega_{yy} + \omega_{zz}) \quad (1)$$

	Crossflow ( $R_\nu \neq 0$ )	Streamwise ( $R_\nu \neq 0$ )	Streamwise ( $R_\nu = 0$ )
Time	$\frac{h^2}{\nu R_\nu}$	$\frac{h^2}{\nu R_\nu}$	$\frac{h^2}{\nu R}$
Length	$h$	$h$	$h$
Velocity	$\frac{R_\nu}{R} u_0$	$u_0$	$u_0$
Acceleration	$\frac{\nu^2 R_\nu^2}{h^3}$	$\frac{\nu^2 R R_\nu}{h^3}$	$\frac{\nu^2 R^2}{h^3}$
Stream Function	$\nu R_\nu$	-	-
Potential	$\frac{R}{R_\nu} u_0 h$	-	-
Vorticity	$\frac{\nu R_\nu}{h^2}$	$\frac{\nu R}{h^2}$	$\frac{\nu R}{h^2}$
Circulation	$\frac{\Gamma_0}{2\pi}$	$u_0 h$	$u_0 h$
Pressure/Stress	$\frac{\mu \nu R_\nu^2}{h^2}$	$\frac{\mu \nu R_\nu^2}{h^2}$	$\frac{\mu \nu R^2}{h^2}$

Nondimensional parameters are:

$$R_\nu = \frac{\Gamma_0}{2\pi\nu} \quad (\text{ratio of spanwise convective effects to diffusive effects})$$

$$R = \frac{u_0 h}{\nu} \quad (\text{ratio of streamwise convective effects to diffusive effects})$$

Table 1: Characteristic variables and nondimensional parameters

$$\psi_{yy} + \psi_{zz} = -\omega \quad (2)$$

$$w = \psi_y$$

$$v = -\psi_z$$

Streamwise:

$$u_t + v u_y + w u_z = \frac{2}{R_\nu} + \frac{1}{R_\nu} (u_{yy} + u_{zz}) \quad (3)$$

The body force term  $\frac{2}{R_\nu}$  in the streamwise direction is the component of gravity in the  $x$  direction that establishes and works to maintain the original parabolic profile.

## 2.4 BOUNDARY CONDITIONS

On  $\partial_{left}$ ,  $\partial_{top}$ , and  $\partial_{right}$  (slippery walls):  $\bar{\omega} = 0$ . On  $\partial_{bottom}$  (viscous wall):  $u = v = w = 0$ . These yield the boundary conditions:

$$\begin{array}{ll} \text{Equation 1: } \omega = 0 & \text{On } \partial_{left}, \partial_{top}, \partial_{right} \\ w = 0 & \text{On } \partial_{bottom} \end{array}$$

$$\text{Equation 2: } \psi = 0 \quad \text{On } \partial_{left}, \partial_{top}, \partial_{right}, \partial_{bottom}$$

$$\begin{array}{ll} \text{Equation 3: } u_y = 0 & \text{On } \partial_{top} \\ u_z = 0 & \text{On } \partial_{left}, \partial_{right} \\ u = 0 & \text{On } \partial_{bottom} \end{array}$$

## 2.5 INITIAL CONDITIONS

We initialize in both the streamwise and crossflow directions:

$$\text{Equation 1: } \omega = \omega_0(z, y) \quad (\text{for any specified distribution } \omega_0)$$

$$\text{Equation 3: } u = 2y - y^2$$

**Geometry:**

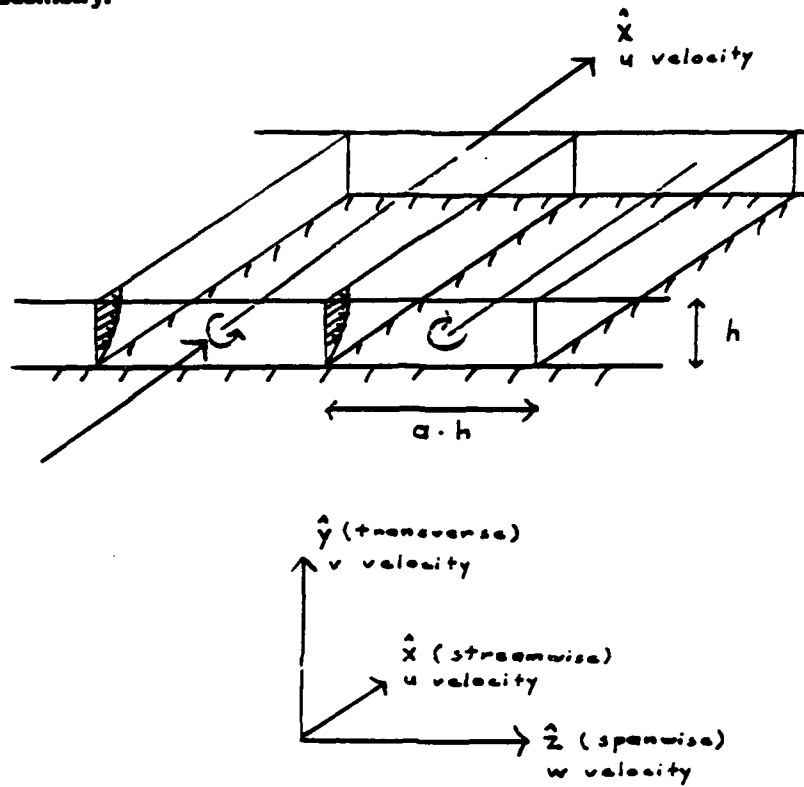


Figure 4: Geometry adopted for this investigation. We consider an initially parabolic flow of infinite extent in  $x$  and  $z$  above a plane viscous wall. At time zero, an array of counter-rotating vortex pairs is begun, aligned in the  $x$  direction. Similar to previous flow table experiments, except the free surface is constrained to a flat plane.

### 3 COMPUTATIONAL SOLUTION

At this point, there is no obvious way to attack the problem analytically. Unlike Pearson's exploitation of the Oseen vortex, we have no expression for the evolving bounded vortex, and because of the boundaries there does not seem to be much likelihood of finding a similarity solution for the streamwise velocity. Therefore we turn to a computational solution of the above equations. The methodology chosen is:

- Initialize
- Step forward  $\omega$  on interior (Equation 1)
- Step forward  $u$  (Equation 3)
- Solve for  $\psi$  (Equation 2)
- Calculate  $\omega$  on viscous boundary (Thom's Method)

We loop through as many time steps as desired.

### 4 BENCHMARKS

An exact solution by Taylor (1923) and Suri (1988) for the spanwise flowfield of a certain vortex will serve as a useful point of reference in places to come, so we present it here. The problem is similar to the crossflow part of our problem listed above (Equations 1 and 2), except for the bottom boundary condition and the initial condition:

$$\omega_t + v\omega_y + w\omega_z = \frac{1}{R_\nu}(\omega_{yy} + \omega_{zz}) \quad (4)$$

$$\psi_{yy} + \psi_{zz} = -\omega \quad (5)$$

$$w = \psi_y$$

$$v = -\psi_z$$

#### BOUNDARY CONDITIONS:

Boundary conditions differ in that the bottom is another slippery plane, like the top:

$$\text{Equation 4: } \omega = 0 \quad \text{On } \partial_{left}, \partial_{top}, \partial_{right}, \partial_{bottom}$$

$$\text{Equation 5: } \psi = 0 \quad \text{On } \partial_{left}, \partial_{top}, \partial_{right}, \partial_{bottom}$$

#### INITIAL CONDITIONS:

$$\omega_0 = \frac{\pi^3}{2a} \cos \frac{\pi z}{a} \cos \pi y \quad (6)$$

Once the above problem has been solved, the pressure P is obtained to within an additive constant from:

$$P_z = -w_t - v\omega_y - w\omega_z + \frac{w_{yy} + \omega_{zz}}{R_\nu}$$

$$P_y = -v_t - v\omega_y - w\omega_z + \frac{v_{yy} + \omega_{zz}}{R_\nu}$$

The solution to the above problem is:

$$\omega = \frac{\pi^3}{2a} \cos \frac{\pi z}{a} \cos \pi y \exp - \frac{(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \quad (7)$$

$$\psi = \frac{\pi}{2a(1 + \frac{1}{a^2})} \cos \frac{\pi z}{a} \cos \pi y \exp - \frac{(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \quad (8)$$

$$w = -\frac{\pi^2}{2a(1 + \frac{1}{a^2})} \cos \frac{\pi z}{a} \sin \pi y \exp - \frac{(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \quad (9)$$

$$v = \frac{\pi^2}{2a(1 + \frac{1}{a^2})} \sin \frac{\pi z}{a} \cos \pi y \exp - \frac{(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \quad (10)$$

$$P = -\frac{\pi^4}{16a(1 + \frac{1}{a^2})} (\cos \frac{2\pi z}{a} + \frac{1}{a^2} \cos 2\pi y) \exp - \frac{2(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \quad (11)$$

This is an exact solution for the given initial condition, and our computations show that it is asymptotically correct for others (above a non-viscous bottom surface); for example, Figure 5 shows the residual between Suri's solution and our computational solution for an initially potential vortex. Suri's solution becomes an increasingly better approximation as time increases.

By making the computational bottom wall slippery, we can compare the results of the above numerical procedure with Suri's exact solution. Figure 6 shows the evolution of  $\omega$  at  $z = 0$  for  $a = 1$  and  $R_\nu = 5$  as given by our numerical solution; there is no visible difference (except discretization) between this and a plot of Suri's solution. The same holds true for plots of other flow variables.

In Appendix A, we derive a solution for a vortex above a viscous wall which is analogous to Equations 7 through 11 above. For example, the zero'th order expression for the vorticity is:

$$\omega = \frac{\pi^2}{a} \frac{\sqrt{k - \frac{\pi^2}{a^2}}}{1 - \cos \sqrt{k - \frac{\pi^2}{a^2}}} \cos \frac{\pi z}{a} \sin \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \exp - \frac{kt}{R_\nu}$$

where  $k$  is determined from

$$\frac{\tan \sqrt{k - \frac{\pi^2}{a^2}}}{\sqrt{k - \frac{\pi^2}{a^2}}} = \frac{\tanh \frac{\pi}{a}}{\frac{\pi}{a}}$$

This solution is similar to Equation 7 in that both share a cosine in  $z$  and both have a trigonometric function in  $y$  that goes to zero at  $\partial_{top}$ . Suri's function (a cosine) goes to zero on  $\partial_{bottom}$ , yielding a slippery bottom surface, whereas our function (a sine with displaced argument) has a shorter period, introducing just enough negative vorticity near  $\partial_{bottom}$  so as to enforce the viscous boundary condition. This solution above a viscous boundary serves as another benchmark; the residual between it and our computational solution is presented in Appendix A. Again, agreement is very good.

Another test against an analytic result can be made by making the computational bottom wall viscous and starting a potential vortex at the center of a square ( $a = 1$ ) box. Since  $\Gamma = \int \vec{v} \cdot d\vec{l}$  and the crossflow component of the velocity  $\vec{v}$  is symmetric on the four walls, we know that turning on the viscous wall at  $t = 0$  will instantaneously decrease the circulation to exactly 75% of its initialized value. Numerically, we measure the circulation via Green's Theorem:  $\Gamma = \iint \omega dA$ , and find the predicted drop to occur to a high degree of accuracy (eg, Figure 33). The circulation is immediately decreased by 25% due to the vortex sheet of opposite-signed vorticity generated when the viscous wall is activated.



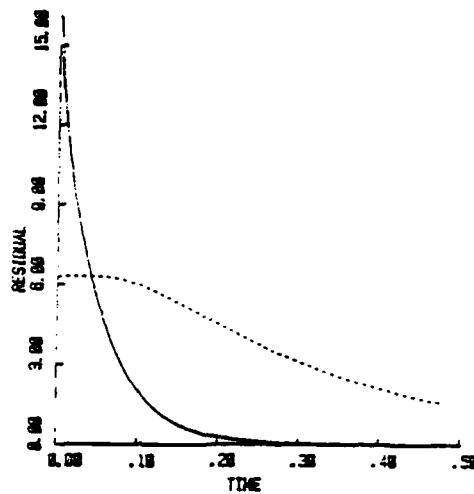


Figure 5: Residual between Suri's solution (Equation 7) beginning from a fully diffused vortex and our computational solution for flow above a slippery wall beginning from an initially potential vortex. After sufficient time the actual initial condition is irrelevant, and Suri's solution is a good approximation. Dotted line is the circulation of an initially potential vortex for comparison.

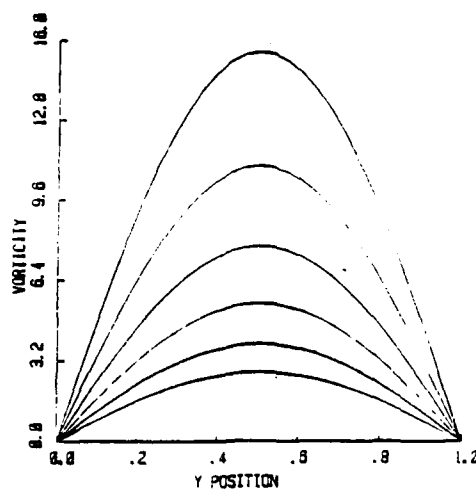


Figure 6: Vorticity of a fully diffused vortex above a slippery wall at different times as calculated by our computational procedure, superimposed on Suri's exact solution (Equation 7). Agreement is good enough to make the two sets of curves coincident.

## 5 STREAMWISE VELOCITY PROFILES

At short times, our numerical solution produces streamwise velocity profiles that, locally, are qualitatively and quantitatively the same as Pearson and Yang (Figures 7 to 10). At a vortex strength of  $R_\nu \approx 6.5$ , we calculate a vertical slope of zero ( $\frac{du}{dy} = 0$ ), in agreement with Pearson. Velocity profiles become more inflectional at higher  $R_\nu$ . The streamwise velocity profiles are slightly asymmetric (the magnitude of the perturbation on the upflow/low speed side is greater than that of the perturbation on the downflow/high speed side) due to the parabolic profile, just as observed by Yang. These areas of agreement occur in the range  $t < \frac{0.4y^2}{\nu}$  (here,  $t$  and  $y$  are dimensional). This is a diffusive time, corresponding to the diffusion of the vortex core to the boundary.

At longer times, the velocity gradient at the vortex center rotates back to its original orientation and value, and the original undisturbed flow is restored (Figure 11). This is because circulation is lost at the boundaries (both viscous and slippery). Since Pearson and Yang worked in unbounded domains, their circulation remained constant (even though they eventually had a finite amount of vorticity of vanishingly small density spread over an infinite area), so their undisturbed flows were never restored. In our present bounded domain, we see the gradient at the center of the vortex begin to be restored soon after the core hits the wall (eg, Figure 20).

Suri's solution (Equations 7 to 11) yields a decay rate of  $\exp -\frac{(1+\frac{1}{2})\pi^2}{R_\nu}$  for a vortex above a slippery wall. How much faster does the vortex decay when above a viscous wall? In Appendix A, we show that the asymptotic decay rate is  $\exp -\frac{kt}{R_\nu}$ , where  $k$  is determined from

$$\frac{\tan \sqrt{k - \frac{\pi^2}{a^2}}}{\sqrt{k - \frac{\pi^2}{a^2}}} = \frac{\tanh \frac{\pi}{a}}{\frac{\pi}{a}}$$

This yields the same decay rate in an infinitely skinny box ( $a = 0$ ) where the viscous wall is negligible, but yields a decay rate almost twice that of Suri's when in a semi-infinite domain ( $a \rightarrow \infty$ ). This means that the flow variables decay like  $\exp -5t/sec$  ( $t$  is dimensional here) in typical transition experiments.

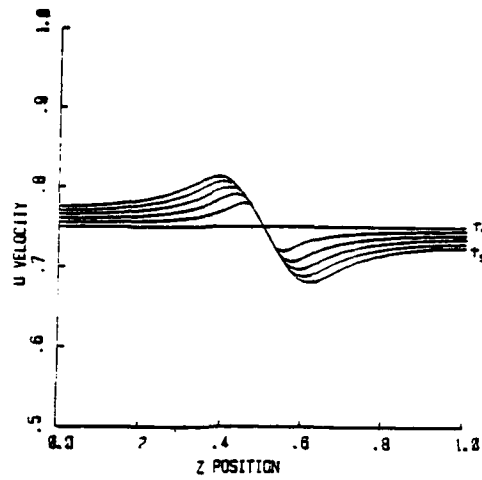


Figure 7: Streamwise velocity profiles at successive times. Total elapsed time is small. Flow is initially parabolic above a viscous wall, vortex is initially potential. Note the greater magnitude of the low-speed perturbation, due to the parabolic mean flow.

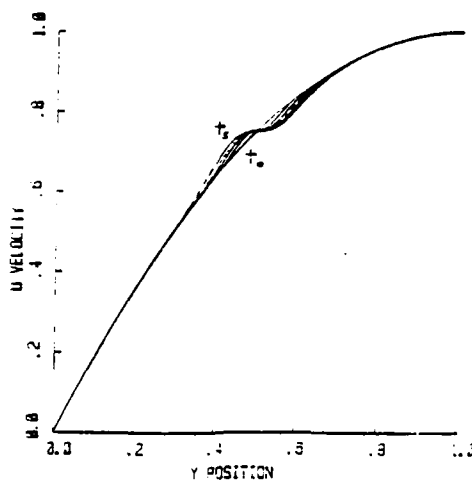


Figure 8: Vertical streamwise velocity profiles at successive times for the flow in Figure 7. Vertical profile is initially parabolic; greater perturbations occur after longer times. Note how  $\frac{\partial u}{\partial y} \approx 0$  at this vortex strength of  $R_v = 6.5$ , in agreement with Pearson and Abernathy (1984).

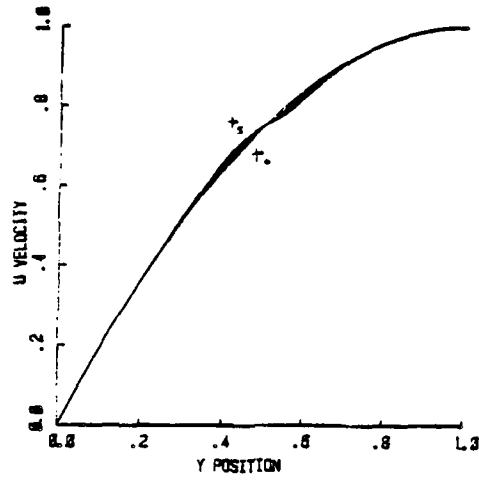


Figure 9: Vertical streamwise velocity profiles at successive times for a vortex of strength  $R_v = 4$ .

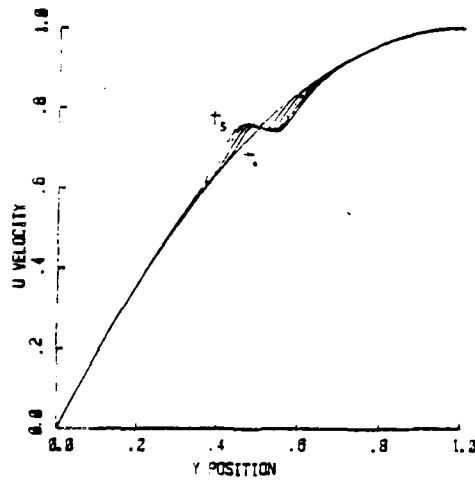


Figure 10: Vertical streamwise velocity profiles at successive times for a vortex of strength  $R_v = 10$ .

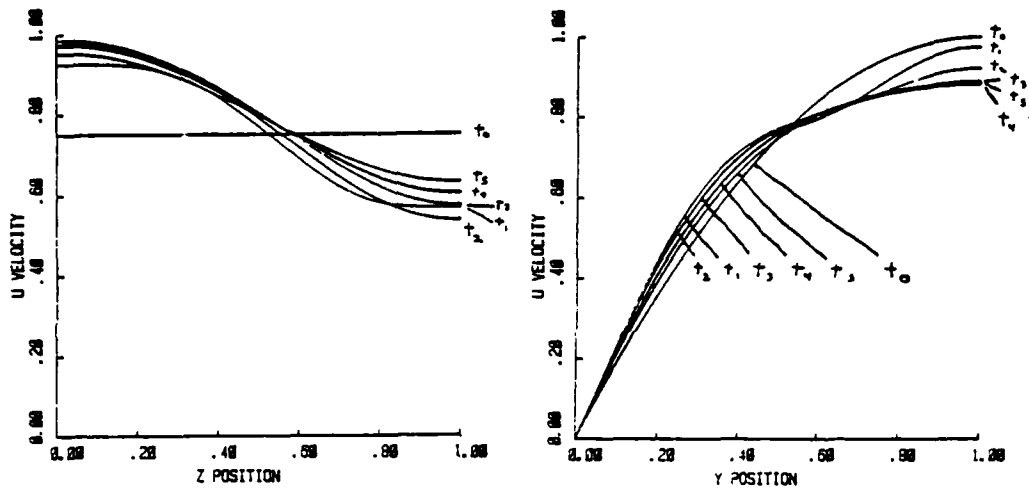


Figure 11: Streamwise velocity profiles at later times for an initially potential vortex,  $R_v=5$ , in initially parabolic flow. Note the decay back toward the initial profile.

## 6 COUNTER-ROTATING VORTEX

One of the primary motivations for undertaking this investigation was to look into the ramifications of Yang's secondary structure: is it a counter-rotating vortex induced by the main vortex above the viscous wall? Does it affect the stability of the flow, and thereby reopen questions about the existence and stability of single streamwise vortices in real flows?

At  $a = 1$ , no counter-rotating vortex is observed below some threshold  $R_\nu$  (resolution-dependent:  $\approx 10$  on an  $81 \times 81$  grid), whereas at higher  $R_\nu$  a small counter-rotating vortex is induced in the lower right corner (for a counterclockwise main vortex). At larger aspect ratios (eg,  $a = 4$ ), the secondary vortex is much more prominent; it rises from the wall to the same height as the main vortex. Another counter-rotating vortex (much weaker than the first) is also induced in the lower left corner (Figure 12). If the main vortex is started close to the wall, then this left vortex is initially stronger than the right, but the right vortex becomes the stronger as the main vortex lifts away from the wall (Figure 13). At higher  $R_\nu$ , the left vortex appears later, and the right vortex develops from a point on the bottom wall closer to the main vortex (Figure 14).

There are two alternative explanations for this induced counter-rotating vortex on the right side:

1. Negative vorticity is generated at the viscous wall on the bottom and is convected to the right. The counter-rotating vortex begins at the wall on the right when this region of negative vorticity has become sufficiently concentrated with respect to the crossflow velocity of the main vortex.
2. The pressure is high on the right and left sides where the velocity is low, and low between the vortex and the bottom wall where the velocity is higher. This constitutes an unfavorable pressure gradient on the right side, resulting in separation.

In Appendix A, we indicate that this counter-rotating vortex will always be induced for main vortices above a threshold strength that is dependent on the aspect ratio. At  $a = \infty$ , this threshold is small, but non-zero. We find these induced vortices to vanish after a finite time. Also, since we are comparing results with experiments conducted at  $a = \infty$ , it is a potential concern that the side walls sometimes seem to play a part in the development of these induced vortices (eg, Figures 12 and 13). In Appendix C, we show that there is no practical difference between the strengths of the vortex induced in a box of aspect ratio  $a \approx 4$  and the strength of the induced vortex in a semi-infinite domain ( $a = \infty$ ).

At high  $R_\nu$ , the peak circulation of the induced vortex is one order of magnitude smaller than that of the main vortex (Figures 15 and 16). This small circulation has no observable effect on the streamwise velocity profile (Figures 17 and 18). At more moderate  $R_\nu$ , the counter-rotating vortex is at least two orders of magnitude weaker than the main vortex in circulation; within these ranges, the induced vortex is insignificant and has no effect on the streamwise velocity profile.

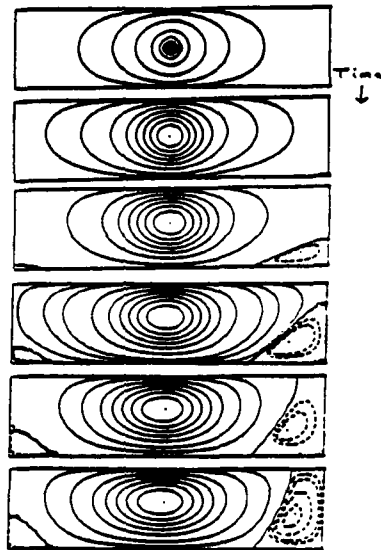


Figure 12: Development of the induced vortex over time for an initially potential vortex of strength  $R_v = 5$ . Box is of aspect ratio  $a = 4$ .

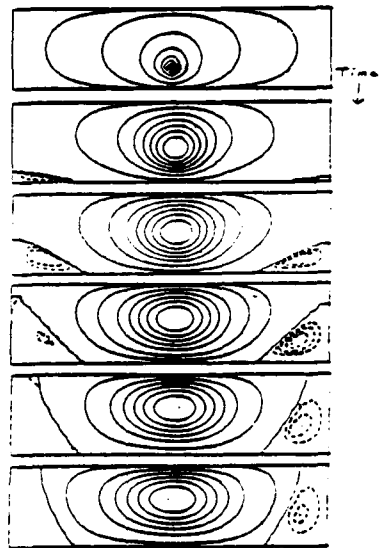


Figure 13: Development of the induced vortex over time for an initially potential vortex of strength  $R_v = 5$  started near the wall. The left induced vortex develops first, then the right induced vortex overtakes it in strength as the main vortex moves away from the wall.  $a = 4$ .

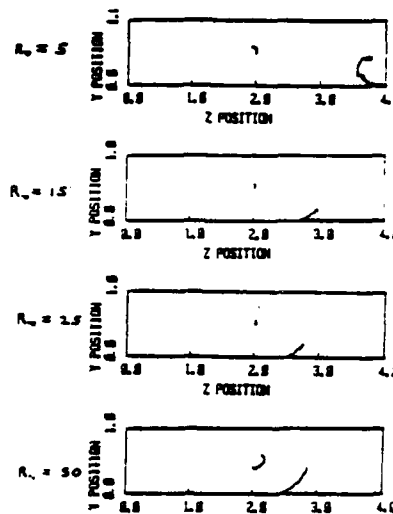


Figure 14: Motion of the centers of the main and induced vortices for main vortices of different strengths,  $R_v$ . In this figure, all vortices begin at the lowest vertical position shown and rise over time.  $a = 4$ .

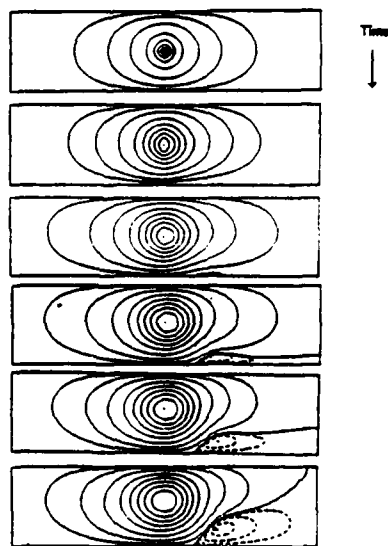


Figure 15: Development of the induced counter-rotating vortex over time. The main vortex is initially potential,  $R_v=25$ , rotating counterclockwise.  $a = 4$



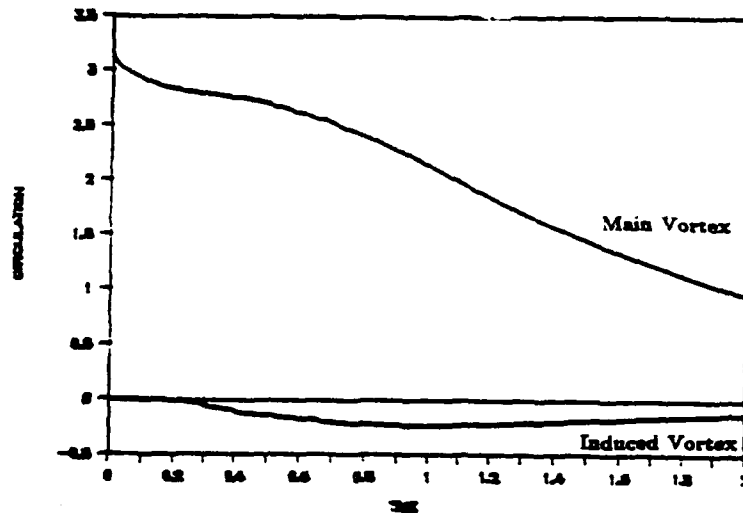


Figure 16: Strength of the main ( $R_v = 25$ ) and induced vortices of Figure 15. Even for a main vortex of this large strength, the induced vortex is at least one order of magnitude weaker.

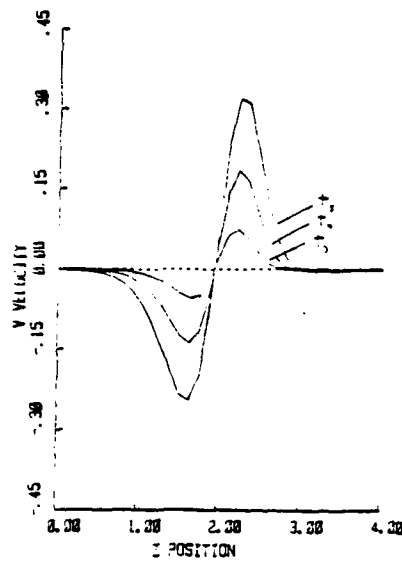


Figure 17: Vertical velocity  $v$  through the center of the induced vortex shown in Figure 15. Note the insignificance of this velocity in the induced vortex (the negative values near  $z \approx 3.2$ ) and the dominant vertical velocities underneath the main vortex ( $z \approx 1.5 - 2.2$ ).

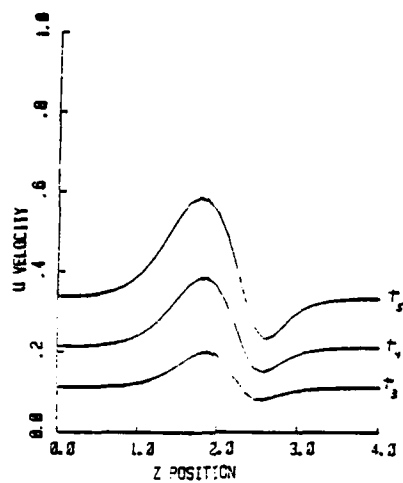


Figure 18: Streamwise velocity profiles through the center of the induced vortex shown in Figure 15. Velocities are greater at successive times since we follow the induced vortex as it rises. The large perturbations near  $z \approx 2$  are due to the main vortex. Note the absence of any effect due to the induced vortex (near  $z \approx 3.2$ ).

## 7 FINITE DISK OF VORTICITY

We have previously used both a delta function and a fully diffused distribution of vorticity as our initial condition. However, we suspect that streamwise vortices in real flows begin with a finite size on the order of the originating disturbance. For small initial disturbances, we expect that finite size to be closer to a delta function than to being fully diffused, but what significant differences are there in subjecting the initially undisturbed flow to a vortex of small but finite structure as opposed to a delta function?

For simplicity, we assume a vortex that begins with constant vorticity inside a cylinder of radius  $r_0$  and with zero vorticity outside. This seems a reasonably more accurate initial view of the vortex, yet simple enough for clear analysis. For this finite-sized vortex, we see that the streamwise velocity profiles are established more slowly over a larger core (Figure 19), rather than instantaneously at the center and then at immediately following moments at the edge of the growing core, as for the Oseen vortex.

The streamwise velocity gradients in the core have not finished growing in this figure. If we plot the slope at the vortex center, we see that the delta function of vorticity sets the slope immediately (first time step) to a constant value. Vortices with successively larger radii (but the same  $R_\nu$ ) take successively longer times to set the slope at the center, but they all eventually set the slope at the vortex center to the same constant value (Figure 20). After the core senses the walls, the slope decays back to the undisturbed value in the same fashion for all cases. Note that it is possible for the slope to begin decaying before its maximum value has been reached, eg, if a finite-sized vortex is placed with its core edge sufficiently close to a wall.

Do larger finite-sized vortices generate stronger induced vortices than smaller vortices (or even delta functions) of the same strength? We have always observed the larger vortices to be weaker, although they may be asymptotically equal, barring interference from the walls.

How does the time to set the gradient inside the core depend on the vortex strength  $R_\nu$ ? Figure 21 shows that the time to set the gradient is independent of  $R_\nu$ , except when the vortex is so strong that the gradient is rotated by more than  $90^\circ$ , ie, when  $R_\nu > 6.5$ .

Summarizing, the diffusion time across the initial vortex radius sets another time scale (in addition to the shear) on the growth of the instability, with delta functions setting the fastest time scales and larger radii structures setting slower time scales. This time required to set the gradient is independent of  $R_\nu$  for  $R_\nu < 6.5$ . For walls that are sufficiently far removed, vortices of all different radii (but at the same  $R_\nu$ ) eventually (on their respective time scales) achieve the same core slopes and velocity profiles, decay at the same time and rate, and therefore affect the flow equivalently. For vortices sufficiently close to a wall, decay may stabilize a large-radius vortex while allowing a smaller one (of the same  $R_\nu$  and whose edge is the same distance from the wall) to go unstable (Figure 22).

Therefore, sufficiently far from walls, the initial distribution of vorticity does not matter in the long run. Near walls, the initially-potential vortex is potentially the most destabilizing since its profiles develop the quickest and stand the best chance of going unstable before wall damping comes into play, and so represents a worst case.

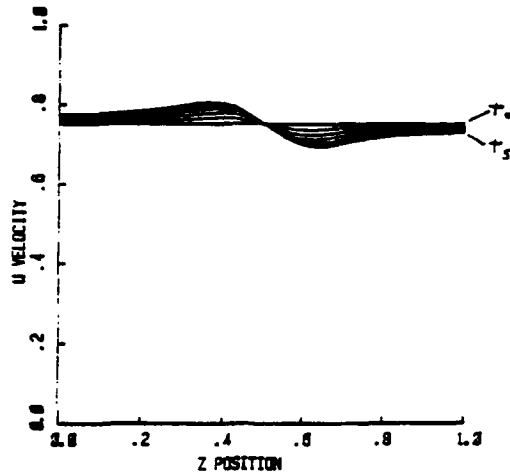


Figure 19: Streamwise velocity profiles for a vortex which begins with finite radius = .1,  $R_v = 6.5$ . Note how slowly the profile develops at the center of the core.

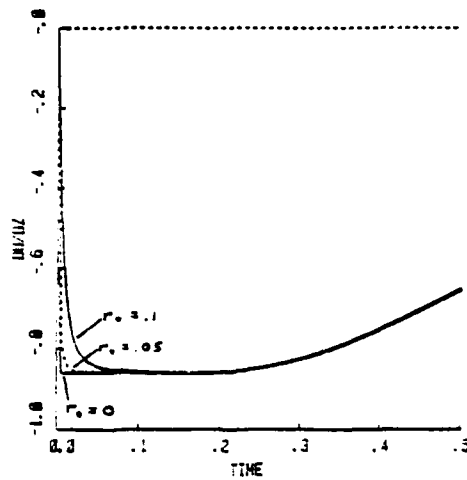


Figure 20: Perturbation of the streamwise velocity at the vortex center for vortices of different initial radii ( $r_0$ ), but the same strength ( $R_v = 6.5$ ). Larger radii vortices develop more slowly, but all achieve the same eventual profile. Note how the original velocity gradient begins to be restored after the vortex core diffuses to the wall at  $t \approx .2$ .

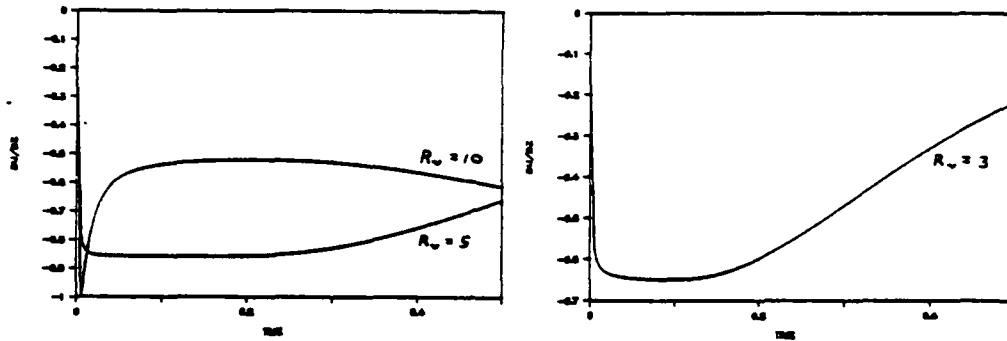


Figure 21: Perturbation of the streamwise velocity at the vortex center for vortices of different  $R_v$ , but the same initial finite radius ( $= .05$ ). Note that the time to develop is independent of  $R_v$ , until the gradient is rotated by more than  $180^\circ$ .

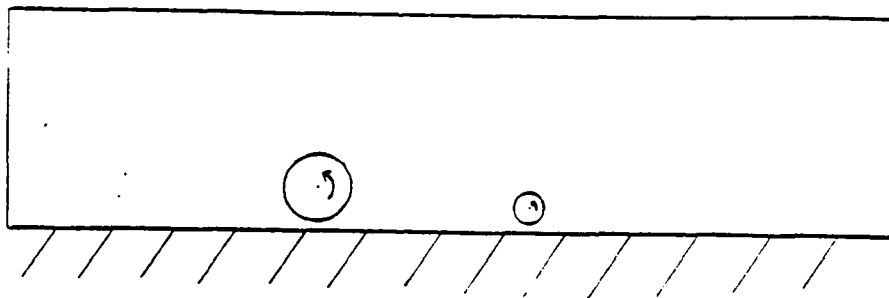


Figure 22: Two vortices of equal strength and distance from the viscous wall. Because the larger vortex develops on a slower time scale, it may be stabilized by the wall while the smaller one goes unstable.

## 8 DUPLICATING YANG'S SECONDARY STRUCTURE

Since the appearance of Yang's secondary structure was one of the main motivations of this work, it is reasonable to attempt to duplicate it in order to find some clues as to its nature. From our preceding results, we know it is not an induced vortex above a plane viscous wall. We try to duplicate Yang's structure by artificially adding another vortex of opposite sign to the flow. After trials with many initial sizes, strengths, and locations, we make two observations: the difference in the observed core sizes of the two vortices necessitates at least one of them starting at finite size in order to achieve this size differential, and also that the second vortex can only exist for a very short time when very near the wall - either the adjacent velocity field must permit it to grow away from the wall (which it generally does not for an artificially introduced vortex), or the second vortex must be started away from the wall.

The configuration which finally comes closest to duplicating Yang's structure is a large finite disk of vorticity (corresponding to his single vortex), and a weak (20% as strong as the main) counter-rotating delta function of vorticity at the edge of the main vortex core (Figure 23). This yields streamwise velocity profiles similar to Yang's data, with effects of the secondary structure undetected far from the wall and more pronounced closer (but not at) the viscous wall (Figure 24).

To be this strong and to appear in this location, it seems a reasonable guess that this secondary structure is induced in the  $90^\circ$  angle between the bottom wall and the vortex generator, rather than above just a plane viscous boundary (Figure 25). This secondary structure might be reduced by softening the sharp corner between the generator and the table. In any case, this second vortex seems to be only about 20% as strong as the main vortex, and therefore should have negligible effect on stability for vortex strengths within our typical range of interest.

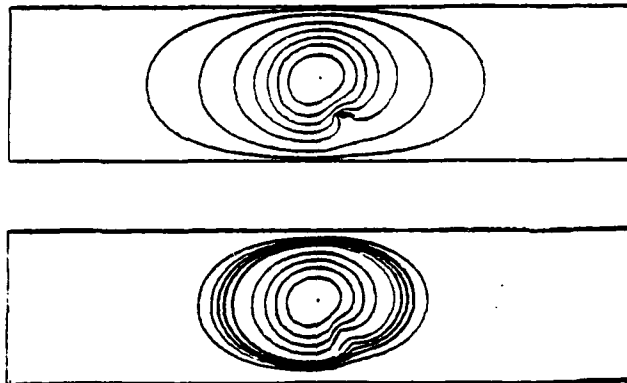


Figure 23: Streamlines of an unequal counter-rotating vortex pair at  $t = 0$  and at a later time. The main vortex is more diffuse and five times stronger than the induced vortex.

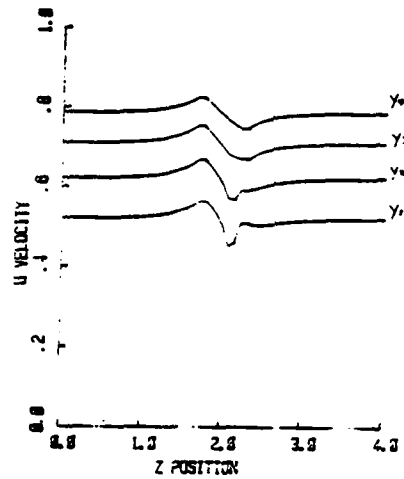


Figure 24: Streamwise velocity profiles at various vertical positions, but at the same fixed time, for the vortices of Figure 23. These profiles are qualitatively the same as Yang's.

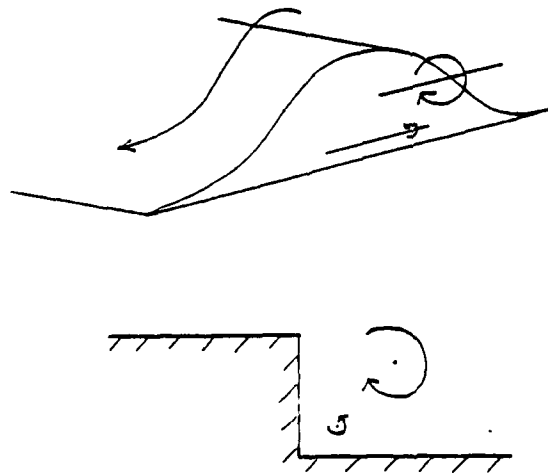


Figure 25: A possible geometry for generating the vortices and profiles of Figures 23 and 24. This counter-rotating vortex induced in the viscous corner of Yang's vortex generator may be the source of his observed secondary structure.

## 9 STABILITY IMPLICATIONS

We can combine previous empirical and analytic data with some hypotheses to yield some simple-minded implications about the stability of our flow.

Yang (as reported by Pearson (1985)) observes transition of streamwise vortices to typically occur at a dimensional time of  $t_{unstable} \approx \frac{40}{\kappa}$ . Pearson's eigenvalue calculations show that the most unstable mode:

$$\begin{array}{ccc} R_\nu = 8 & R_\nu = 5 & R_\nu = 3 \\ \lambda = .115 & \lambda = .096 & \lambda = .073 \end{array}$$

$$\text{grows 10X by } t = \frac{30}{\kappa} \quad \frac{34}{\kappa} \quad \frac{42}{\kappa}$$

$$\text{grows 100X by } t = \frac{50}{\kappa} \quad \frac{58}{\kappa} \quad \frac{73}{\kappa}$$

This confirms that  $t_{unstable} \approx \frac{40}{\kappa}$  may be close. Another piece of supporting evidence comes from Suri's observations of the growth of an instability in a low speed streak. The instability grows over a dimensional distance  $x = 15 - 10 = 5\text{cm}$  (his p.139). His observations are taken at  $y = .095\text{cm}$ , at which  $u = 70 \frac{\text{cm}}{\text{sec}}$  (his p.138). At  $u = 70 \frac{\text{cm}}{\text{sec}}$ :  $\frac{\partial u}{\partial y} = \frac{(94-10) \frac{\text{cm}}{\text{sec}}}{14\text{cm}} = \frac{600}{\text{sec}}$ . Therefore the nondimensional time for transition is:  $t^+ = \frac{x}{u} \frac{\partial u}{\partial y} = \frac{5\text{cm}}{70 \frac{\text{cm}}{\text{sec}}} \cdot \frac{600}{\text{sec}} = 42.9$ . Again, this supports the notion that  $t_{unstable} \approx \frac{40}{\kappa}$ . Another observation that we will use is derived from our previous computations: for a vortex located at  $y = .5$ , we observe wall damping affect the vortex center at a dimensional time of  $t_{damp} \approx \frac{.04h^2}{\nu}$ .

Assume that an unbounded flow with an embedded streamwise vortex undergoes transition (due to linear instability) for  $R_\nu$  greater than some critical value ( $\sim 3$  according to Pearson) at time  $t_{unstable}$ . Also assume that wall damping prevents this and stabilizes the flow if and only if the vortex center feels the wall before time  $t_{damp}$ . Since we have previously shown that flow variables decay like  $\sim \exp(-5t/\text{sec})$  (for typical values of the flow parameters) after  $t_{damp}$ , this seems a reasonable assumption. These assumptions yield three cases:

1.  $t_{unstable} > t_{damp} \Rightarrow$  stable
2.  $t_{unstable} = t_{damp} \Rightarrow$  marginally stable
3.  $t_{unstable} < t_{damp} \Rightarrow$  unstable

For the marginally stable case of  $t_{unstable} = t_{damp}$ :

$$\frac{40}{\kappa} \equiv \frac{40h^2}{R_\nu} = \frac{.04h^2}{\nu} \Rightarrow R_c = 1000$$

Of course, changes in the empirical data or in the  $y$  location will change this value of  $R_c$ . We can look at an arbitrary  $y$  location in a parallel shear flow:

$$t_{unstable} = \frac{f}{u_y} \frac{h}{u_0} \tag{12}$$

(where  $f$  was 40 above).

$$t_{damp} = \frac{4g(y - y_0)^2 h^2}{\nu} \tag{13}$$



where there is a single damping wall at  $y_0$  (and where  $g$  was .04 above).

The neutral curve is again located at  $t_{unstable} - t_{damp} = 0$ :

$$(y - y_0)^2 u_y = \frac{f}{4gR} \quad (14)$$

For a given velocity profile, we can take either  $R_c$  or  $y$  and solve for the other via this expression.

To maximize or minimize the instability, we would place the vortex at the location where  $(t_{unstable} - t_{damp})_y = 0$ , ie, where the time to damp the disturbance most exceeds the time required for the disturbance to succumb to linear instability, or vice-versa. Therefore:

$$(y - y_0)u_y^2 + \frac{f}{8gR}u_{yy} = 0 \quad (15)$$

Again, knowledge of either  $R$  or the most/least stable  $y$  position in a given parallel shear flow allows us to solve for the other. These equations were derived for a single damping wall at  $y = y_0$ . A similar analysis could be done for more complicated geometries.

In the case of our usual Poiseuille flow above a single damping surface at  $y = 0$  and an unconstrained free surface at  $y \approx 1$ , a neutral curve of  $R$  vs vertical position is found from Equation 14 to be:

$$y^3 - y^2 + \frac{f}{8gR} = 0$$

Using the data  $\frac{L}{g} = \frac{40}{.04} = 1000$ :

$$y^3 - y^2 + \frac{125}{R} = 0 \quad (16)$$

The discriminant vanishes when

$$\frac{125}{R} \left( \frac{125}{4R} - \frac{1}{27} \right) = 0$$

$$\Rightarrow R_c = 843.75$$

Similarly, the locations at which a disturbance is most/least stable are found from Equation 15:

$$y^3 - 2y^2 + y - \frac{f}{16gR} = 0$$

or with empirical values for  $\frac{L}{g}$ :

$$y^3 - 2y^2 + y - \frac{125}{2R} = 0 \quad (17)$$

This discriminant vanishes at  $R = 421.875$

These curves are sketched in Figure 26. The curve  $T \equiv t_{unstable} - t_{damp} = 0$  is the neutral curve; a disturbance on this curve will be neutrally stable. Unstable disturbances fall to the right of this curve, stable disturbances fall to the left.  $R_c$ , the minimum possible flow Reynolds number that can sustain instability, is 843.75 for sufficiently strong disturbances located at a dimensional height of  $y = \frac{2}{3}h$ . The upper part of the curve  $T_y = 0$  denotes, for a given  $R$ , the most destabilizing position for a disturbance. Likewise, the lower part denotes the (locally) most stable position. The "most unstable" curve passes through  $R_c$ , as it must. The most unstable position is always in the upper third of the flow (when  $R > R_c$ ).  $R_c \rightarrow \infty$  at  $y = 0$  (cannot go unstable right against the viscous bottom wall) and at  $y = 1$  (the local shear is zero, so it takes infinitely long to go unstable). It is interesting that when near the wall, the most stable position is not right at the wall, but slightly above it.

Note from Equation 14 that the empirical number  $\frac{L}{y}$  is just a multiplicative constant that sets the scale of the abscissa. Therefore if its value is actually somewhat different from what we have inferred, the only change to Figure 26 would be to re-scale the numbers along the  $R$  axis - all else would remain unchanged.

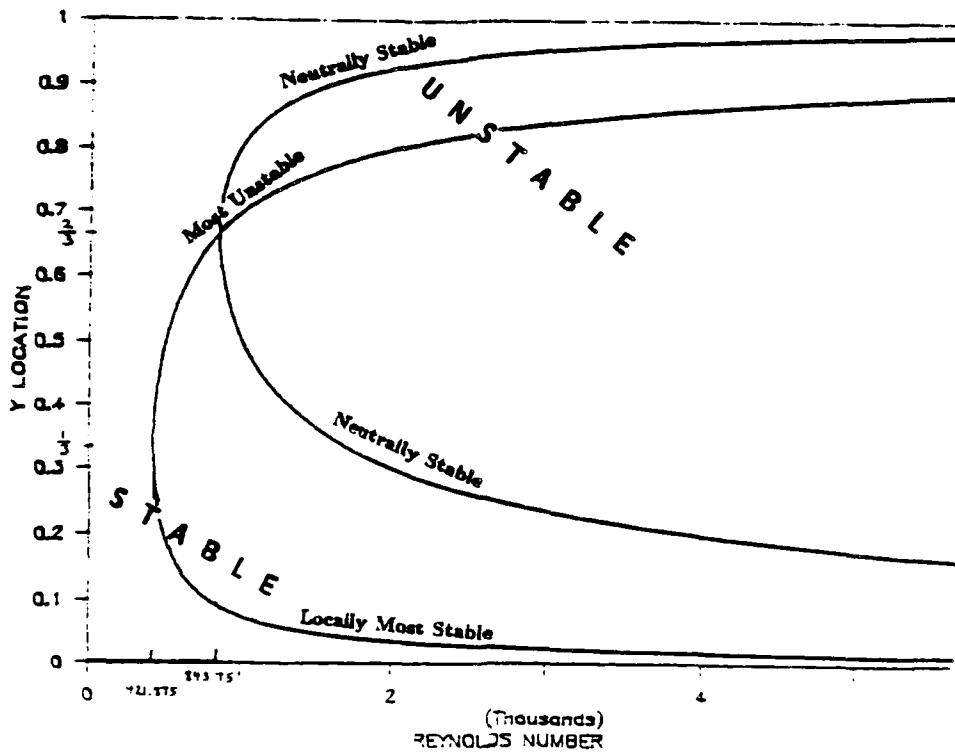


Figure 26: Neutral curve of the neutrally stable Reynolds number (for the most destabilizing mode of a disturbance) at a given distance from the viscous wall.  $R_c$  occurs at 843.75. The curve on which a disturbance is most stable/most unstable passes through this point, and bifurcates at  $R = 421.375$ .

## 10 STABILITY OF CORE VS UPFLOW REGION

Many researchers have been preoccupied with identifying and monitoring the development of instability in the upflow region between a counter-rotating vortex pair (such as between the two legs of a hairpin vortex). Yet we have already seen the highly inflectional profiles that can develop within the core of a single vortex. Applying the same time scale assumptions and arguments used in deriving the previous neutral curve (Section 9), we can investigate which vortex geometry permits the middle upflow region to go unstable first, and which geometry permits the two vortex cores to be the first to go unstable.

We consider a typical vortex pair, each leg of strength  $R_v = 5$ , as shown in Figure 32. Over time, the two legs lift each other and push each other away (Figure 27). Figure 28 shows how the inflectional velocity profiles develop at the vortex center for different initial distributions of vorticity; as we have seen before, more diffuse initial distributions of vorticity (at the same total circulation) perturb the streamwise velocity profiles more slowly so that the maximum perturbation occurs at a later time. For boundaries that are sufficiently far removed, this maximum value is the same for all initial distributions of vorticity. Since boundaries are close at hand here (ie, the virtual boundary caused by the companion vortex), viscous damping may set in before the maximum streamwise perturbation is achieved, and thus more diffuse vorticity distributions create smaller maximum perturbations in the streamwise velocity.

Inflectional profiles also develop in the upflow region at the  $z$  location exactly between the vortex cores. There may be zero, one, or two inflection points at this  $z$  location, and they may vary in vertical location over time (Figure 29). For analysis, we assume that any instability at this  $z$  location will occur at that  $y$  where  $\frac{\partial^2 u}{\partial y^2}$  is a minimum (if  $\frac{\partial^2 u}{\partial y^2} \neq 0$  for all  $y$ ), or at that  $y$  where  $\frac{\partial^2 u}{\partial y^2} = 0$  (if that occurs at a unique  $y$ ), or at the minimum  $y$  at which  $\frac{\partial^2 u}{\partial y^2} = 0$  (if there are two such  $y$  locations). This last criterion is selected because the inflection at minimum  $y$  is located at or near the vertical location of the vortex centers where the shear has been perturbed the most. Figure 30 shows how this most unstable  $y$  location (defined via the above criteria) moves as a function of time for different initial distributions of vorticity. The dips in  $y$  that appear from just after  $t = 0$  to near  $t \approx .2$  are where the flow goes from  $\frac{\partial^2 u}{\partial y^2} > 0$  at all  $y$  to  $\frac{\partial^2 u}{\partial y^2} < 0$  near the middle (yielding two inflection points) - we then follow the development and eventual disappearance of the lower (in  $y$ ) point where  $\frac{\partial^2 u}{\partial y^2} = 0$ . Note that very diffuse vortex pairs (radius = .5) never develop an inflection point in the upflow region. When observing how  $\frac{\partial u}{\partial y}$  develops over time at the  $y$  locations shown in Figure 30 we find, as expected, that more concentrated vortices produce the greatest perturbations in slope from equilibrium.

The time required for the core to go unstable is:

$$t_{core} = \frac{40}{\sqrt{\frac{\partial u^2}{\partial y_{core}^2} + \frac{\partial u^2}{\partial z_{core}^2}}}$$

(as in the previous neutral curve derivation, Equation 12). The time required for the upflow region to go unstable is:

$$t_{upflow} = \frac{40}{\frac{\partial u}{\partial y_{upflow}}}$$

(since  $\frac{\partial u}{\partial z_{upflow}} = 0$ ). Figure 31 shows how  $\Delta t \equiv t_{core} - t_{upflow}$  develops over time.  $\Delta t > 0$  means that the upflow region will go unstable first, and  $\Delta t < 0$  means that the vortex cores will go unstable first. Fortunately, the curves are fairly horizontal (ie, time independent), and we can identify the

dividing case  $\Delta t = 0$  as being roughly at vortex radius  $\approx .25$  (Figure 32). Initial distributions of vorticity more diffuse than this will cause the upflow region to go unstable first. Distributions more concentrated than this will cause the vortex cores to go unstable first. This is consistent with experimental results; Suri's diffuse array of vortices experiences its initial rms increases in the low speed streak (his p.139), whereas Yang's single vortex begins as a more compact structure with no apparent twin, and goes unstable first at the core.

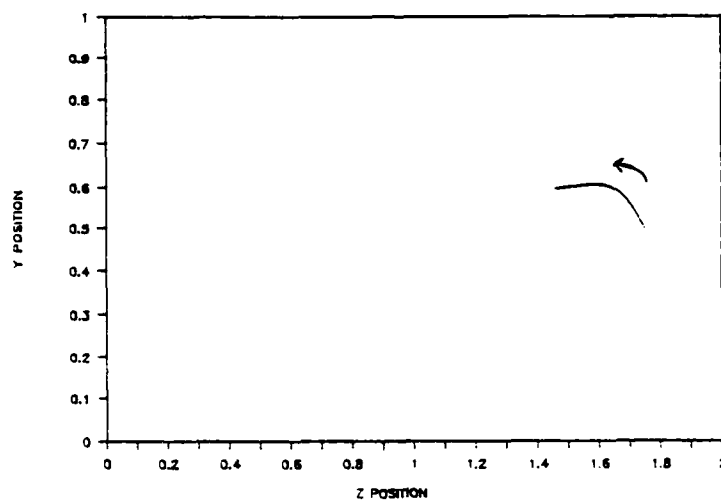


Figure 27: Motion of the center of one of a pair of equal counter-rotating vortices over time. The vortex shown is rotating counterclockwise, resulting in an upflow region in the neighborhood of  $z = 2$ . The motion of the other vortex is obtained by reflecting about the line  $z = 2$ .

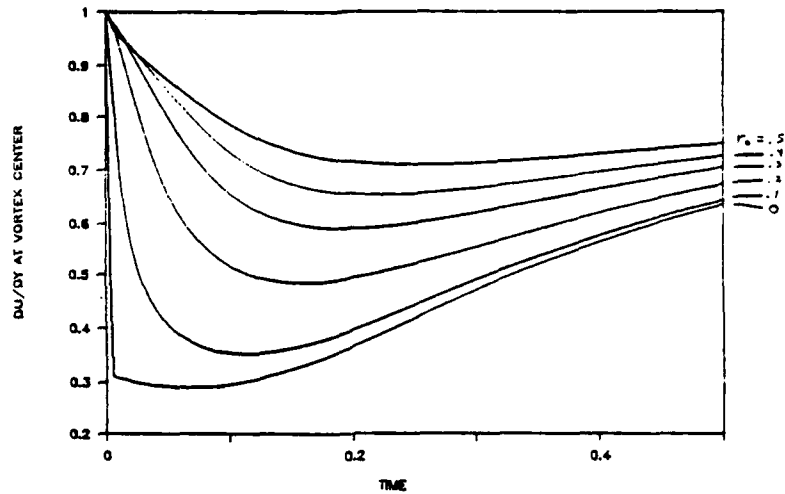


Figure 28: Perturbation of the streamwise velocity field at the center of the vortex core due to vortices of different radii. Note that larger radii vortices have less effect before the original profile begins to be restored.

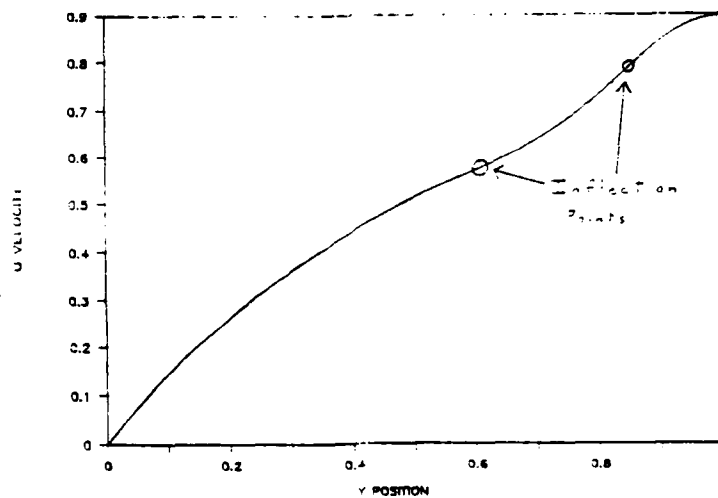


Figure 29: Vertical streamwise velocity profile in the upflow region between two counter-rotating vortices soon after  $t = 0$ . Note the presence of two inflection points, with the lower one closer to the two vortex centers where the flow has been most perturbed.

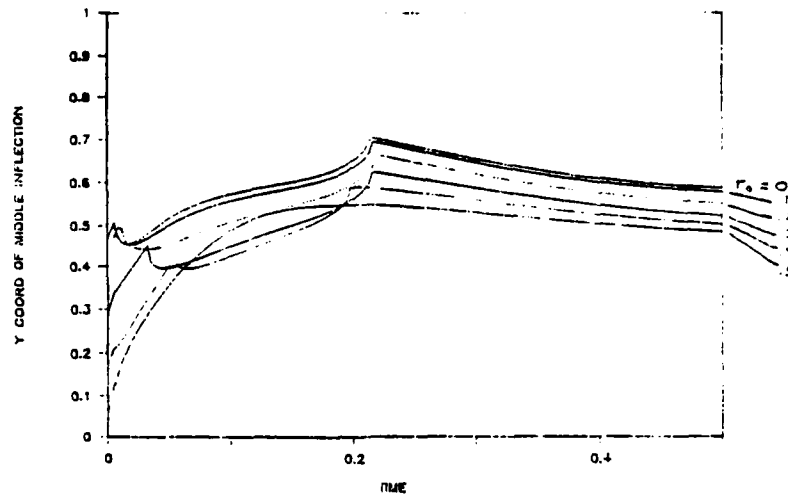


Figure 30: Vertical position of the lower inflection point in the upflow region between two counter-rotating vortices. The dips at  $t < .2$  are where the lower and upper inflection points appear and separate from each other (and we follow the lower). At all other times there is no inflection point; we then plot the location where the second derivative achieves its minimum value.

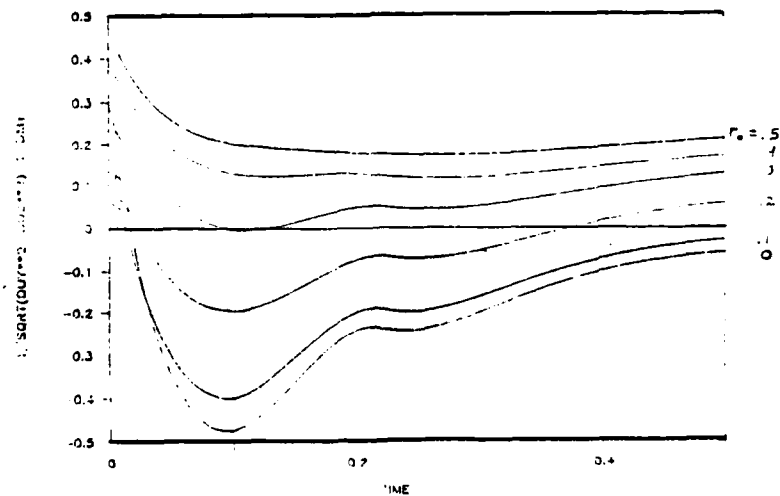


Figure 31: Time required for the core to go unstable minus the time required for the upflow region to go unstable. These times are derived from the assumptions of Section 9. Note that these times are approximately equal for vortices whose initial diameter is equal to their separation.

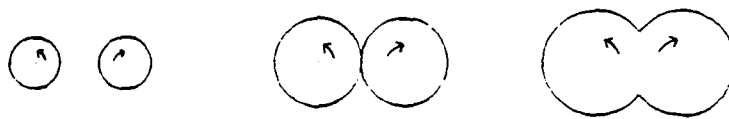


Figure 32: Three counter-rotating vortex pairs of constant separation and strength. Figure 31 shows that the small concentrated pair on the left will go unstable in the cores first, the large diffuse pair at right will go unstable in the middle upflow region first, and the center pair is the intermediate case where cores and upflow region go unstable together.

## 11 CONCLUSIONS

Before the vortex core hits the wall ( $t < \frac{0.4y^2}{\nu}$ ), we find that the crossflow and streamwise velocity profiles are locally the same as those derived by Pearson. After hitting the wall, we find that flow variables decay back to their undisturbed values, typically  $\sim \exp(-5t/sec)$ . It is wall influence that is responsible for restoring the original profile. We find a perturbation solution, analogous to a previous analytic solution, that describes the flow well after this time.

Sufficiently far from walls, any initial distribution of vorticity has the same long-term effect on stability. Close to a wall, smaller (more concentrated) vortices may be more destabilizing than larger (more diffuse) ones of the same circulation.

Pearson's velocity profiles are quite remarkable in that they show how just a tiny bit of streamwise vorticity (eg, a circulation of  $\sim .4 \frac{cm^2}{sec}$  for liquid water) can destabilize a shear flow. The fact that we achieve these same velocity profiles above a viscous wall and the fact that we find them to be relatively insensitive to initial vortex structure allows us to conclude that the single streamwise vortex is a promising model for describing laminar-turbulent transition in real flows and for investigating (in a very idealized fashion) one of the most commonly recurrent structures in the self-sustenance cycle of turbulence.

A streamwise vortex above a plane viscous wall whose strength is greater than some small (but non-zero) threshold induces a counter-rotating vortex. This counter-rotating vortex exists for a finite time, and is insignificant to stability within our typical parameter range of interest. So once a single streamwise vortex is created, the secondary structure is irrelevant to the stability of the flow (within the parameter range for which the main vortices are normally observed), and can be safely omitted from analysis.

Yang's structure seems to be a second counter-rotating vortex of smaller size and  $\sim 20\%$  strength of the main vortex, possibly induced in the  $90^\circ$  corner between the bottom viscous wall and the side of the vortex generator. There is no effect on the streamwise velocity profiles or stability, so he really does seem to have generated, investigated, and drawn conclusions about an essentially single streamwise vortex.

The role of the Reynolds number in transition seems to be to set the local shear, and thus the time scale of the instability, as a disturbance grows toward instability in a race against viscous damping. At the critical Reynolds number, the time to go unstable is the same as the time required for wall damping to be felt. Above  $R_c$ , the Reynolds number sets the local shear in the vicinity of a disturbance high enough so that the disturbance goes unstable on a time scale shorter than the viscous time scale required for wall damping, so the local instability "beats" wall damping. Coupled with some empirical data, this interpretation yields a critical Reynolds number for Poiseuille flow of  $R_c \approx 844$ .

In situations with a pair of equal counter-rotating vortices, the initial instability may develop either in the common upflow region (diffuse vortices) or in the core centers (more concentrated vortices), depending on the initial vortex structure.

The first author gratefully acknowledges the support provided by the Department of the Army and the West Point Education Center through the Tuition Assistance Program.



## A PERTURBATION SOLUTION

Suri found a solution for the crossflow variables for a vortex above a slippery wall (Equations 7-11); his solution is exact for a vortex that begins fully diffused in a specified way (Equation 6), and is asymptotically accurate for other initial distributions of vorticity. It is desirable to have an analogous expression for a vortex above a viscous wall, both for further simple analytic modeling, and to see what such a solution can tell us about our current counter-rotating vortex concerns.

### A.1 CROSSFLOW SOLUTION

As shown in Figure 33, for a non-viscous bottom wall, the vortex circulation is constant until the core hits the wall. The circulation then decays exponentially. For a viscous bottom, the circulation instantaneously drops due to the vortex sheet generated at  $t = 0$ , then decreases as negative vorticity continues to be generated at the wall. When the core reaches the wall, the circulation again appears to assume an exponential decay (albeit at a different rate). A semilog plot of circulation vs time confirms this exponential decay.

With this additional clue, analytic progress is possible. Under our two original assumptions (concerning  $x$  independence and the free surface shape), the analytic crossflow problem is as in Equations 1 and 2:

$$\omega_t + u\omega_y + w\omega_z = \frac{1}{R_\nu}(\omega_{yy} + \omega_{zz})$$

$$\begin{aligned} \text{Boundary Conditions: } \omega &= 0 && \text{On } \partial_{left}, \partial_{top}, \partial_{right} \\ w &= 0 && \text{On } \partial_{bottom} \end{aligned}$$

$$\psi_{yy} + \psi_{zz} = -\omega$$

$$\text{Boundary Conditions: } \psi = 0 \quad \text{On } \partial_{left}, \partial_{top}, \partial_{right}, \partial_{bottom}$$

$$w = \psi_y$$

$$v = -\psi_z$$

Since the circulation and other flow variables decay approximately exponentially after the core reaches the wall:

$$\omega(z, y, t) = \omega'(z, y) \exp -\frac{kt}{R_\nu} \quad (18)$$

and similarly for the other flow variables. Substitution yields:

$$(\nabla^2 + k)\omega' = \epsilon R_\nu \exp -\frac{kt}{R_\nu} (v'\omega'_y + w'\omega'_z) \quad (19)$$

where

$$\epsilon \equiv \exp -\frac{k(t-t_0)}{R_\nu}$$

where  $t_0$  is the time at which the exponential decay sets in (ie. soon after wall effects begin to be felt), and where we use the 2-D Laplacian:

$$\nabla^2 \equiv \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Adopting a regular perturbation expansion:

$$\omega' = \omega_0 + \epsilon\omega_1 + \epsilon^2\omega_2 + \dots \quad (20)$$

and similarly for the other flow variables. Note that assuming an expansion of this form makes our previous time derivative (Equation 19) only approximately correct since our expansion parameter  $\epsilon$  is also a function of time, but computations tell us to expect that after a sufficient time, the time dependence will be carried mainly in the first few modes of this (essentially) weakly nonlinear expansion; therefore we anticipate only a small error (to be checked shortly) and proceed. This yields a sequence of coupled (through the boundary conditions) elliptic problems:

$$\begin{cases} (\nabla^2 + k)\omega_0 = 0 \\ \nabla^2\psi_0 = -\omega_0 \end{cases} \quad (21)$$

$$\begin{cases} (\nabla^2 + k)\omega_1 = R_\nu \exp -\frac{kia}{R_\nu}(v_0\omega_{0,y} + w_0\omega_{0,z}) \\ \nabla^2\psi_1 = -\omega_1 \end{cases} \quad (22)$$

$$\begin{cases} (\nabla^2 + k)\omega_2 = R_\nu \exp -\frac{kia}{R_\nu}(v_0\omega_{1,y} + w_0\omega_{1,z} + v_1\omega_{0,y} + w_1\omega_{0,z}) \\ \nabla^2\psi_2 = -\omega_2 \end{cases} \quad (23)$$

⋮

Boundary conditions for each  $\omega_i$  and  $\psi_i$  component are all individually the same as given above for  $\omega$  and  $\psi$ . Note that these boundary-value problems do not allow the specification of an initial condition. Suri's solution (Equations 7 to 11) shares this feature. Soon after the core has reached the boundaries, information about initial conditions has been diffused away.

These equations are linear and separable. The solution to the zero'th order equations (21) is:

$$\omega_0 = A \cos \frac{\pi z}{a} \sin \sqrt{k - \frac{\pi^2}{a^2}}(y - \frac{1}{2}) \quad (24)$$

$$\psi_0 = A \cos \frac{\pi z}{a} \left( \frac{1}{k} \sin \sqrt{k - \frac{\pi^2}{a^2}}(y - \frac{1}{2}) - \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{k \sinh \frac{\pi}{a}} \sinh \frac{\pi}{a}(y - \frac{1}{2}) \right) \quad (25)$$

$$\begin{aligned} w_0 = & A \cos \frac{\pi z}{a} \left( \frac{\sqrt{k - \frac{\pi^2}{a^2}}}{k} \cos \sqrt{k - \frac{\pi^2}{a^2}}(y - \frac{1}{2}) \right. \\ & \left. - \frac{\pi \sin \sqrt{k - \frac{\pi^2}{a^2}}(y - \frac{1}{2})}{ak \sinh \frac{\pi}{a}} \cosh \frac{\pi}{a}(y - \frac{1}{2}) \right) \end{aligned} \quad (26)$$

$$v_0 = A \frac{\pi}{a} \sin \frac{\pi z}{a} \left( \frac{1}{k} \sin \sqrt{k - \frac{\pi^2}{a^2}}(y - \frac{1}{2}) - \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{k \sinh \frac{\pi}{a}} \sinh \frac{\pi}{a}(y - \frac{1}{2}) \right) \quad (27)$$

$$P_0 = A \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{k \sinh \frac{\pi}{a}} \sin \frac{\pi z}{a} \cosh \frac{\pi}{a}(y - \frac{1}{2}) \quad (28)$$

where  $k$  is determined from the eigenvalue relationship:

$$\frac{\tan \sqrt{k - \frac{\pi^2}{a^2}}}{\sqrt{k - \frac{\pi^2}{a^2}}} = \frac{\tanh \frac{\pi}{a}}{\frac{\pi}{a}} \quad (29)$$

For an initially fully diffused (in the sense that it starts with the same form as  $\omega_0$  above) vortex above a viscous wall:

$$A = \frac{\pi^2}{a} \frac{\sqrt{k - \frac{\pi^2}{a^2}}}{(1 - \cos \sqrt{k - \frac{\pi^2}{a^2}})}$$

The solution to the first order equations (22) is:

$$\begin{aligned} \omega_1 = & A^2 R_\nu \exp -\frac{kt_0}{R_\nu} \sin \frac{2\pi z}{a} (B \sin \sqrt{k - \frac{4\pi^2}{a^2}} (y - \frac{1}{2}) \\ & - \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{4k \sinh \frac{\pi}{a}} \sin \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \cosh \frac{\pi}{a} (y - \frac{1}{2})) \end{aligned} \quad (30)$$

$$\begin{aligned} \psi_1 = & A^2 R_\nu \exp -\frac{kt_0}{R_\nu} \sin \frac{2\pi z}{a} (\frac{B}{k} \sin \sqrt{k - \frac{4\pi^2}{a^2}} (y - \frac{1}{2}) \\ & - \frac{k + \frac{2\pi^2}{a^2}}{4k^2(k + \frac{3\pi^2}{a^2})} \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \sin \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \cosh \frac{\pi}{a} (y - \frac{1}{2}) \\ & - \frac{\pi \sqrt{k - \frac{\pi^2}{a^2}}}{2ak^2(k + \frac{3\pi^2}{a^2})} \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \cos \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \sinh \frac{\pi}{a} (y - \frac{1}{2}) \\ & + (\frac{(k + \frac{4\pi^2}{a^2}) \sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{8k^2(k + \frac{3\pi^2}{a^2}) \sinh^2 \frac{\pi}{a}} - \frac{B \sin \sqrt{k - \frac{4\pi^2}{a^2}}}{k \sinh \frac{2\pi}{a}}) \sinh \frac{2\pi}{a} (y - \frac{1}{2})) \end{aligned} \quad (31)$$

$$\begin{aligned} w_1 = & A^2 R_\nu \exp -\frac{kt_0}{R_\nu} \sin \frac{2\pi z}{a} (\frac{B \sqrt{k - \frac{4\pi^2}{a^2}}}{k} \cos \sqrt{k - \frac{4\pi^2}{a^2}} (y - \frac{1}{2}) \\ & + \frac{\pi(k - \frac{4\pi^2}{a^2})}{4ak^2(k + \frac{3\pi^2}{a^2})} \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \sin \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \sinh \frac{\pi}{a} (y - \frac{1}{2}) \\ & - \frac{(k + \frac{4\pi^2}{a^2}) \sqrt{k - \frac{\pi^2}{a^2}} \sin \sqrt{k - \frac{\pi^2}{a^2}}}{4k^2(k + \frac{3\pi^2}{a^2}) \sinh \frac{\pi}{a}} \cos \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \cosh \frac{\pi}{a} (y - \frac{1}{2}) \\ & + \frac{2\pi}{a} (\frac{(k + \frac{4\pi^2}{a^2}) \sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{8k^2(k + \frac{3\pi^2}{a^2}) \sinh^2 \frac{\pi}{a}} - \frac{B \sin \sqrt{k - \frac{4\pi^2}{a^2}}}{k \sinh \frac{2\pi}{a}}) \cosh \frac{2\pi}{a} (y - \frac{1}{2})) \end{aligned} \quad (32)$$

$$\begin{aligned} v_1 = & -A^2 R_\nu \exp -\frac{kt_0}{R_\nu} \frac{2\pi}{a} \cos \frac{2\pi z}{a} (\frac{B}{k} \sin \sqrt{k - \frac{4\pi^2}{a^2}} (y - \frac{1}{2}) \\ & - \frac{k + \frac{2\pi^2}{a^2}}{4k^2(k + \frac{3\pi^2}{a^2})} \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \sin \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \cosh \frac{\pi}{a} (y - \frac{1}{2}) \\ & - \frac{\pi \sqrt{k - \frac{\pi^2}{a^2}}}{2ak^2(k + \frac{3\pi^2}{a^2})} \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \cos \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2}) \sinh \frac{\pi}{a} (y - \frac{1}{2}) \end{aligned}$$

$R_\nu$	$a$		$k$ (Numerical)	$k$ (Analytic)
5	1	(slippery wall)	19.7550	19.7390
5	1	(viscous wall)	26.2775	26.2798
5	2	"	21.2513	21.2702
3	3	"	20.2251	20.6118
8	3	"	19.8104	20.6118
5	4	"	20.1585	20.4145
-	$\infty$	"	-	20.1907

Table 2: Computed vs analytic decay coefficients ( $k$ )

$$+ \left( \frac{(k + \frac{4\pi^2}{a^2}) \sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{8k^2(k + \frac{3\pi^2}{a^2}) \sinh^2 \frac{\pi}{a}} - \frac{B \sin \sqrt{k - \frac{4\pi^2}{a^2}}}{k \sinh \frac{2\pi}{a}} \right) \sinh \frac{2\pi}{a} (y - \frac{1}{2}) \quad (33)$$

$$B = \frac{\sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{2(k + \frac{3\pi^2}{a^2}) \left( -\frac{a}{\pi} \sqrt{k - \frac{4\pi^2}{a^2}} \cos \sqrt{k - \frac{4\pi^2}{a^2}} + \frac{2 \sin \sqrt{k - \frac{4\pi^2}{a^2}}}{\tanh \frac{2\pi}{a}} \right)} \quad (34)$$

Note the turning point at  $k = \frac{4\pi^2}{a^2}$ , ie, at  $a \approx 1.301$ .

Given the geometrically increasing effort required and the exponentially decreasing importance of the higher order modes, the first order solution seems a good stopping point. Note that up through the first order solution,  $t_0$  (the time at which  $\epsilon$  becomes small) cancels out of the solution.

Note the similarity of Equation 24 to Suri's solution (Equation 7) for a diffused vortex above a slippery wall. Both solutions for  $\omega$  share a cosine term in the  $z$  direction which goes to zero at the slippery side walls, and a trigonometric term in  $y$  that goes to zero at the top surface. In the  $y$  direction, the half-period of Suri's cosine is equal to the flow depth in order to enforce the slippery boundary condition on the bottom wall. Our function (a sine with displaced argument) has a shorter period, introducing just enough negative vorticity in the region near the viscous wall so as to enforce the no-slip boundary condition.

We verify the accuracy of this approximate solution by calculating the residual between the zeroth and first order solutions (Equations 24 and 30 substituted into Equations 18 and 20) and the numerical solution (Figure 34). We see that the residual of this approximate solution quickly becomes small (compared to the vortex circulation). Now that we have demonstrated its accuracy, it is appropriate to determine the implications of this analytic solution.

Flow variables decay like  $\sim \exp - \frac{kt}{R_\nu}$  where  $k$  is determined from Equation 29. Table 2 shows values of the nondimensional decay rate coefficient ( $k$ ) as found from numerical simulation and as calculated analytically. Agreement is good. Discrepancies are attributable to the fact that we are calculating the decay rate at a finite time, and are not extrapolating to the asymptotic limit. This is why, for example, the vortex at  $R_\nu = 8$  inside a box of aspect ratio  $a = 3$  is off by more than a vortex at  $R_\nu = 3$  at the same aspect ratio: the stronger vortex is reaching the asymptotic decay rate more slowly, but both share the same limit. For a single vortex in a box of infinite aspect ratio, we find the decay rate to be  $\approx \frac{20.1907}{R_\nu}$  ( $\approx \frac{20.1907\pi}{h^2}$  dimensionally). (Actually at this aspect ratio, the nondimensional decay rate =  $\frac{c^2}{R_\nu}$ , where  $c$  is defined in Equation 35 below). In typical transition experiments, such as Yang's and Suri's flow table experiments, this means that flow variables decay like  $\sim \exp(-5t/sec)$  ( $t$  is dimensional here). We can put bounds on  $k$ .  $k \geq \frac{\pi^2}{a^2}$ , else Equation 29

has the unique solution  $k = 0$ , ie, the trivial solution. Now  $k \geq \frac{\pi^2}{a^2}$  implies that we must use the second or higher branch of the  $\tan$  function. (Our relation actually admits a countably infinite number of eigenvalues  $k$ , one for each positive branch of the  $\tan$  function). But solutions on higher branches will decay faster (each branch at least  $4\pi^2$  faster than the next lower one), so the lowest possible branch will dominate in observations. Therefore, we should use the second branch of the  $\tan$  function. As Equation 29 is written, it equates the slope of the secant line through the origin and a point on the  $\tanh$  curve to the corresponding secant line for the  $\tan$  function. The argument  $\sqrt{k - \frac{\pi^2}{a^2}}$  must equal the value of the independent variable at the point of intersection, and so must take on a value in the range  $[\pi, c]$ , where  $c$  is determined from

$$\tan c = c \quad (35)$$

on the second branch (meaning that  $c \approx 4.49$ ). Therefore bounds on  $k$  are  $\frac{\pi^2}{a^2} + \pi^2 \leq k \leq \frac{\pi^2}{a^2} + c^2$ . The actual solution  $k$ , along with its upper and lower bounds, is plotted in Figure 35. Note that the lower bound is just Suri's decay rate (Equations 7 to 11) for a slippery bottom boundary.  $k$  achieves this lower bound as  $a \rightarrow 0$  (ie, in an infinitely skinny box) which makes sense: we would expect the bottom viscous wall to have negligible effect as it is made infinitely small. From the figure, we also observe that  $k$  is almost constant for  $a > 4$ ; when the side walls are sufficiently far away, decay is dominated by the viscous wall.

Decay rates comprise an interesting additional benchmark. Figure 36 shows how the decay rate of the circulation quickly approaches the theoretical value of Equation 7.1.12. Suri's decay rate for a fully diffused vortex above a slippery wall (Equations 7 to 11) is roughly only half this value. A streamwise analysis (see section A.2 following) similar to that performed above for the crossflow shows that streamwise flow variables should have an asymptotic decay rate of  $k_s = \frac{\pi^2}{R_\nu} (\frac{1}{4} + \frac{1}{a^2})$ . Figure 36 also shows the agreement between this rate and the calculated decay; convergence is slower because decay is ultimately driven through spanwise processes which act via coupling in the convective terms. Again, agreement in both of these cases lends another degree of confidence to both the computational and analytic solutions.

Analysis of Equations 24 to 29 shows that the vortex moves (asymptotically) to  $x = 0$  and to a  $y$  location determined from:

$$\frac{\cos \sqrt{k - \frac{\pi^2}{a^2}} (y - \frac{1}{2})}{\cosh \frac{\pi}{a} (y - \frac{1}{2})} = \frac{\cos \sqrt{k - \frac{\pi^2}{a^2}}}{\cosh \frac{\pi}{a}} \quad (36)$$

This  $y$  position is plotted in Figure 37. When  $a = 0$  (infinitely skinny box), then  $y = .5$  as for a slippery bottom, which again seems consistent. When  $a \rightarrow \infty$  (infinitely long box), then  $y = 2 - \frac{2\pi}{c}$  ( $\approx .6$ ). Therefore if a vortex is started nearly centered in  $y$  (which Yang does), then it will raise only slightly, explaining why he detects no movement.

To investigate the induced vortex, we check for separation at the viscous wall:  $\lim_{y \rightarrow 0} (\psi_0 + \epsilon \psi_1) = 0$ , or after reducing several indeterminate forms:  $\lim_{y \rightarrow 0} (\omega_0 + \epsilon \omega_1) = 0$ . In other words, a point of separation is also a point with zero vorticity. We deduce that separation point(s) is(are) located at:

$$z = \frac{a}{\pi} \sin^{-1} \frac{\exp \frac{kt}{R_\nu} \sin \sqrt{k - \frac{\pi^2}{a^2}}}{2.4 R_\nu (B \sin \sqrt{k - \frac{4\pi^2}{a^2}} + \frac{\sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{4k \tanh \frac{\pi}{a}})}$$

This induced vortex is present at  $t = 0$ ; this solution does not model its beginning well, but does better at later times.

The induced vortex disappears into the lower right corner at

$$t = \frac{R_\nu}{k} \ln \frac{2AR_\nu(-B \sin \sqrt{k - \frac{\pi^2}{a^2}} + \frac{\sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{4k \tanh \frac{\pi}{a}})}{\sin \sqrt{k - \frac{\pi^2}{a^2}}}$$

The induced vortex appears only if the main vortex is above a threshold strength:

$$R_\nu > \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{2A(-B \sin \sqrt{k - \frac{\pi^2}{a^2}} + \frac{\sin^2 \sqrt{k - \frac{\pi^2}{a^2}}}{4k \tanh \frac{\pi}{a}})} \quad (37)$$

This threshold strength is graphed in Figure 38. Note that  $a = 0$  (infinitely skinny box) yields a threshold of  $R_\nu \rightarrow \infty$ , which again seems consistent. It is interesting that there exists a small but nonzero threshold  $R_\nu$  for  $a \rightarrow \infty$ . However, the threshold is low enough ( $\approx 1.2$ ) that we will generate a counter-rotating vortex for any interesting main vortex.

The preceding expressions simplify for the case  $a \rightarrow \infty$ : Separation point(s):

$$z = \frac{2k \exp \frac{kt}{R_\nu}}{AR_\nu \sin \sqrt{k}}$$

Conditions for separation:

$$R_\nu > \frac{2}{\pi} (1 - \sec \sqrt{k}) \frac{1}{p}$$

$$t < \frac{R_\nu}{k} \ln \frac{\pi R_\nu p}{2(1 - \sec \sqrt{k})}$$

Here,  $p$  is an empirical phase factor that extrapolates the exponential decay back to  $t = 0$ . In this case the separating streamline between the two vortices is a vertical line through the above  $z$  position. The induced vortex moves to the same  $y$  position as the main ( $\approx .6$ ) before receding exponentially to the right.

## A.2 STREAMWISE SOLUTION

Still under the assumption that the crossflow variables decay mostly like  $\exp -\frac{kt}{R_\nu}$  we have:

$$v_t + vu_y \exp -\frac{kt}{R_\nu} + wu_z \exp -\frac{kt}{R_\nu} = \frac{2}{R_\nu} + \frac{1}{R_\nu} (u_{yy} + u_{zz})$$

Assuming that the streamwise velocity has the same type of exponential decay after a certain time (but at a possibly different rate):

$$u = 2y - y^2 + u' \exp -k_s t$$

Substituting and retaining zero'th order terms yields:

$$(\nabla^2 + R_\nu k_s) u' = 0$$

Boundary Conditions:  $u'(z, 0) = 0$   
 $u'_y(z, 1) = 0$   
 $u'_z(\pm \frac{a}{2}, y) = 0$

The solution is:

$$u = 2y - y^2 + A \sin \frac{\pi x}{a} \sin \frac{\pi y}{2} \exp -k_s t \quad (38)$$

where

$$k_s = \frac{\pi^2}{R_v} \left( \frac{1}{4} + \frac{1}{a^2} \right) \quad (39)$$

Agreement with computations is shown in Figure 36.

In terms of dimensional variables, the transient part of Equation 38 decays like  $\exp -\left(\frac{1}{4} + \frac{1}{a^2}\right) \frac{\pi^2 \nu t}{h^2}$ . Note that this zero'th order solution is independent of  $R$ ; ie, the asymptotic streamwise decay rate is fully driven by the crossflow geometry. Also note that the streamwise decay rate is essentially independent (related only through  $a$ ) of the crossflow decay rate.

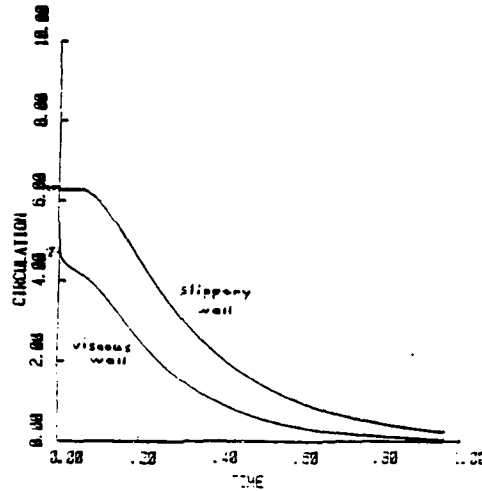


Figure 33: Circulation of an initially potential vortex above both a slippery and a viscous bottom wall. Above a viscous wall, note how the circulation instantaneously drops from  $2\pi$  to  $1.5\pi$  (since this box is of aspect ratio  $a = 1$ ) due to the vortex sheet generated at the wall. Decay quickly becomes exponential.

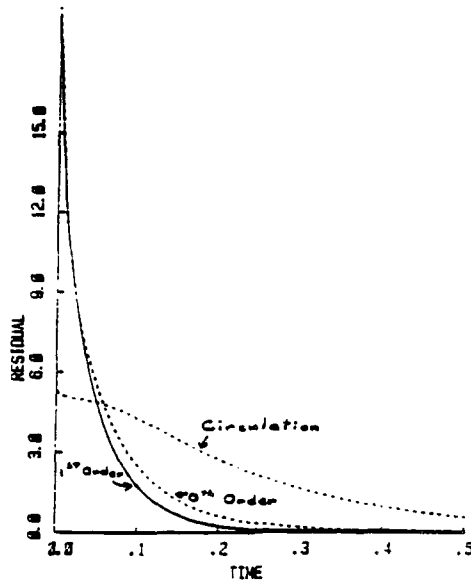


Figure 34: Residual between our computational solution for an initially potential vortex above a viscous wall and the zero'th and first order perturbation solutions. The circulation of the initially potential vortex is included for reference.

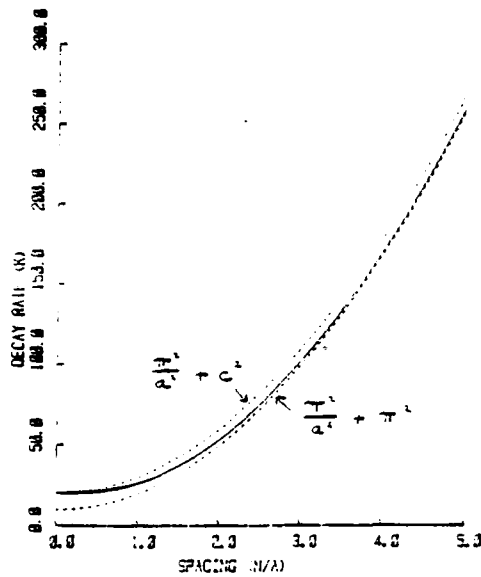


Figure 35: Asymptotic decay rate for spanwise variables above a viscous wall as a function of the aspect ratio of the vortex array (Equation 29), along with upper and lower bounds. The lower bound is just Suri's decay rate for spanwise variables above a slippery wall.



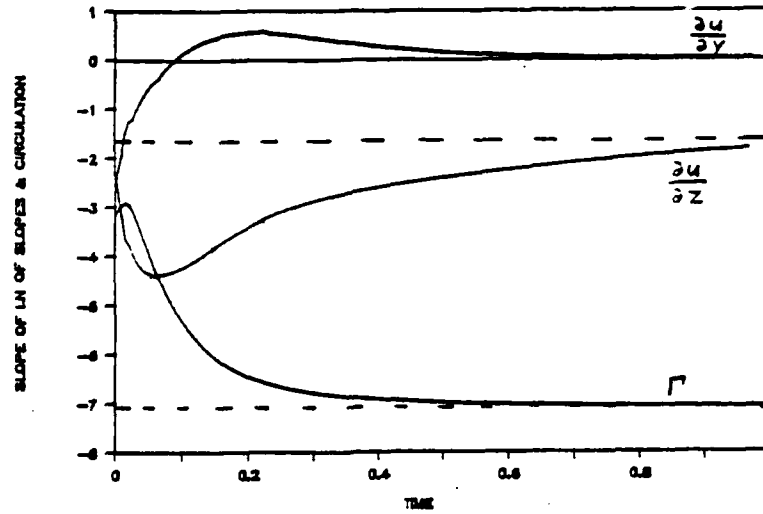


Figure 36: Convergence of streamwise ( $\frac{\partial u}{\partial z}$ ) and spanwise ( $\Gamma$ ) decay rates to the asymptotic values (dashed lines) of Equation 29 and Equation 39.  $R_v = 3$ ,  $a = 2$ .

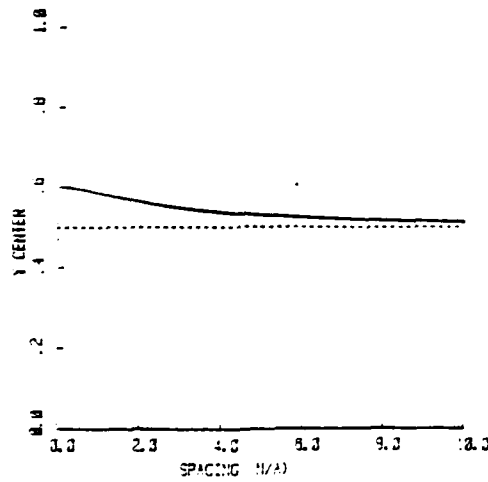


Figure 37: Asymptotic vertical position of the vortex center as a function of aspect ratio,  $a$ , as given by Equation 36.

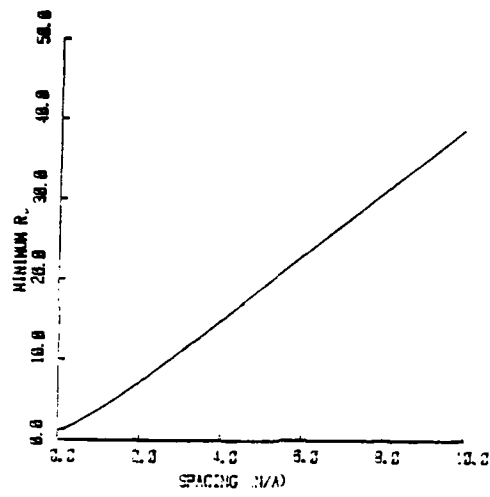


Figure 38: Minimum main vortex strength required to induce a counter-rotating vortex at the viscous wall as a function of aspect ratio, from Equation 37.

## B EFFECT OF FREE SURFACE

To make the numerical problem more tractable, we previously assumed that the free surface was constrained by a normal force to a flat slippery surface. This allowed us to remove one variable (free surface position) from the problem, but at the price of violating the  $\tau_{ij}n_j = 0$  (no normal stress) boundary condition. This assumption seems to be well-justified by experiment; however, at this point, we will test the consistency of this assumption with the calculated solution. We do this during the flow evolution by periodically calculating what the shape of the free surface would be if we were to let it deform to satisfy the above no-normal-stress boundary condition. The shapes we obtain are not fed back into the flow evolution calculation; rather, if the shapes we find are always negligibly different from a flat surface, then we will have justified the flat-surface assumption throughout the computation.

First we calculate the pressure throughout the flow from:

$$\nabla^2 P = 2(w_x v_y - w_y v_x) \quad (40)$$

$$\begin{aligned} \text{Boundary Conditions: } P_x &= 0 && \text{on } \partial_{left} \text{ and } \partial_{right} \\ P_y &= 0 && \text{on } \partial_{top} \text{ (and on } \partial_{bottom} \text{ if slippery)} \\ P_y &= \frac{v_{xy}}{R_\nu} && \text{on } \partial_{bottom} \text{ (if viscous)} \end{aligned}$$

This determines  $P$  to within an additive constant (which may be time dependent). Then we calculate the normal stress on all boundaries from:

$$\tau_{ij} = -P\delta_{ij} + \frac{1}{R_\nu}(u_{i,j} + u_{j,i}) \quad (41)$$

At this point, we choose the additive constant for  $P$  such that

$$\int_{\partial_{top}} \tau_{ij} dn_j = 0 \quad (42)$$

This condition is necessary to satisfy continuity; rationale will be given shortly. Note that our previous slippery boundary condition on the flat plane  $\partial_{top}$  (ie,  $\omega = 0$  coupled with  $\hat{n} = j$ ) ensures that there is no shear stress on  $\partial_{top}$ . We now relieve the normal stress on  $\partial_{top}$  through two mechanisms.

A first mechanism for relieving normal stress is to remove an appropriate mass of fluid from positions with downward normal stress and to add an appropriate mass of fluid to positions with upward normal stress such that the weight of the fluid removed/added balances the excess/deficient normal stress. Therefore:

$$\tau = \frac{R}{R_\nu} \Delta h g \cos \theta \quad (43)$$

where  $\Delta h$  is the nondimensional position of the free surface relative to the constrained surface,  $g$  is the nondimensional acceleration due to gravity, and  $\theta$  is the angle of  $\partial_{bottom}$  (ie, the  $x$  axis) from horizontal. This angle is the source of fluid motion in the  $x$  direction; the component of gravity in the  $x$  direction is the body force that works to maintain the parabolic profile. We previously found this body force to be  $\frac{2}{R_\nu}$  (Equation 3), so we find that:

$$g \equiv \frac{2}{R_\nu \sin \theta}$$

We can now solve for our nondimensional change in free surface height:

$$\Delta h = \frac{\tau R_\nu^2 \tan \theta}{2R} \quad (44)$$

This is a first approximation to the curvature of the free surface. We can pause here and make a crude order estimate as to its magnitude. Substituting Equations 28 and 27 into Equation 41, Equation 41 into Equation 44, and evaluating at  $\partial_{top}$  in the neighborhood of  $a \sim \mathcal{O}(1)$  yields  $\Delta h_{max} \sim .8 \frac{R_v}{R}$ . So within our typical range of interest we have  $\Delta h \sim \mathcal{O}(.001)$ . This crude estimate indicates that curvature effects at the free surface are negligible.

Equation 44 is a first approximation to the position of the free surface, yet there is another effect that should be considered. By adding and subtracting fluid at the upper surface, we have necessarily stretched the streamlines in some locations and compressed them in others, thus changing the velocities and ultimately both the pressure and the stress. This will further change the free surface height. We can account for this effect by appropriately perturbing our first approximation. The calculation begins by finding the point on  $\partial_{top}$  where:

$$\tau = -P + \frac{2v_y}{R_v} = 0$$

(from Equation 41); in other words, we first find the point on the free surface where there is no deformation. The existence of at least one such point is guaranteed by applying the mean value theorem to our condition used to find the additive constant for the pressure (Equation 42). The normal stress at a nearby point on  $\partial_{top}$  is:

$$\tau_n = -P_0 - \widetilde{P}'_y \Delta h - \widetilde{P}'_z \Delta z + \frac{2v'_y}{R_v} + \frac{2R\Delta h}{R_v^2 \tan \theta} \quad (45)$$

where  $P_0$  is the pressure at our starting point, tildes denote derivatives to be evaluated by centered differences between the previous and current points, and primes denote quantities evaluated after the free surface has been perturbed to relieve the normal stress due to streamline stretching. Note that by construction, we would have  $\tau_n = 0$  at all points on  $\partial_{top}$  if we were to omit the primes (ie, neglect the stretching of streamlines). For small variation in the free surface height (our original conjecture, to be verified through consistency), we can accurately assume the distance between streamlines to be modified by a factor  $\frac{1+\Delta h}{1}$ , therefore the spanwise velocity  $w$  will be modified by  $\frac{1}{1+\Delta h}$  and the transverse velocity  $v$  will be similarly modified through continuity. Spatial derivatives are unaffected. Equation 45 can now be expressed in terms of quantities that are more easily computed:

$$\begin{aligned} \tau_n = & -P_0 + \Delta h \left( \frac{2R}{R_v^2 \tan \theta} \right) \\ & + \frac{1}{1+\Delta h} \left( \frac{2v_y}{R_v} - \Delta z (\widetilde{w w}_z + \widetilde{P}_z) \right) + \frac{\Delta h}{1+\Delta h} \left( -\frac{\widetilde{v_{yy}}}{R_v} \right) \\ & + \frac{1}{(1+\Delta h)^2} (\Delta z \widetilde{w w}_z) \end{aligned} \quad (46)$$

In deriving this equation, we have neglected vertical velocities introduced by free surface curvature. We find our final free surface height by setting  $\tau_n = 0$  in the above equation and iteratively solving for  $\Delta h$ . Then using the newly calculated point, we repeat for the following point, and so on for all points on  $\partial_{top}$ .

We calculate stresses preliminary to calculating the free surface position, and it is interesting to note in passing the total stress and torque acting on the box of fluid. Figure 39 shows the total stress acting in the  $z$  (spanwise) direction. Initially, the vortex applies a large jerking force to the right as the vortex-induced flow pushes on the viscous wall, then the force changes direction as the vortex applies increased stress against the left wall, and the net effect is then a small and decaying

force to the left. The stress in the  $y$  direction is a tiny downward force, but this is practically zero. The net torque on the box grows quickly in the counterclockwise direction, then decays (Figure 40). In order to hold the box steady, equal and opposite forces and torques would have to be applied to the box, ie, a large leftward force as the box jerks to the right and then a small decaying force to the left, along with a decaying clockwise torque. Of course, when considering an array of vortices, the torques and horizontal forces are all balanced by those of the images; the very small downward force is the only resultant force.

Figure 41 shows the height of the free surface in a box with an aspect ratio of one, and Figure 42 shows the free surface in a box of aspect ratio equal to eight. In all computations at  $R_\nu$  within our range of interest ( $< 15$ ), the contribution due to streamline stretching is totally insignificant and the gravity effect dominates. We retain both in the computation, but the gravity effect expression for the free surface height is accurate to a high degree of accuracy. Therefore  $\Delta h$  is proportional to  $\tau$  (Equation 44). To preserve continuity we need  $\int_{-\frac{a}{2}}^{\frac{a}{2}} \Delta h dz = 0$ , and so this is the reason (as previously promised at Equation 42) that we used the condition  $\int_{-\frac{a}{2}}^{\frac{a}{2}} \tau_n(z, 1) dz = 0$  to choose the additive constant for the pressure. The error in not including the streamline stretching effect in the additive constant (via iteration) is negligible.

The free surface in Figure 41 is similar to a sine curve slightly displaced from center. This is consistent with the analytic solution found previously; substituting Equation 28 (at  $\partial_{top}$ ) into Equation 41, Equation 41 into Equation 44, and using Equation 27 to evaluate  $v_y$ , we have:

$$\Delta h = \frac{AR_\nu^2 \tan \theta}{2R} \left( \frac{2\pi}{akR_\nu} \sqrt{k - \frac{\pi^2}{a^2}} - \left(1 + \frac{2\pi}{akR_\nu}\right) \frac{\sin \sqrt{k - \frac{\pi^2}{a^2}}}{\sinh \frac{\pi}{a}} \right) \sin \frac{\pi z}{a} \exp - \frac{kt}{R_\nu} \quad (47)$$

We have used just the zero'th order terms here; they should suffice since the vortex diffuses to all four boundaries relatively quickly in a square box. So the perturbation solution predicts exactly this centered sinusoid for the free surface shape. This is slightly different from the shape of the free surface derived from Suri's expression (Equations 10 and 11) for the pressure distribution above a slippery bottom:

$$\Delta h = \frac{R_\nu^2 \pi^3 \tan \theta}{4a^2 R (1 + \frac{1}{a^2})^2} \left( \frac{\pi}{8} \left( \cos \frac{2\pi z}{a} - \frac{1}{a^2} \right) \exp - \frac{2(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \right. \\ \left. - \frac{2}{R_\nu} \sin \frac{\pi z}{a} \exp - \frac{(1 + \frac{1}{a^2})\pi^2 t}{R_\nu} \right) \quad (48)$$

Note that his pressure distribution is symmetric rather than asymmetric. It also decays twice as fast as the convective stress term. Both expressions for the free surface height (Equations 47 and 48) share the same shape as  $a \rightarrow 0$  and  $t$  becomes large (for both expressions:  $\Delta h \sim \tan \theta \sin \frac{\pi z}{a} \exp - \frac{\pi^2 t}{a^2 R_\nu}$ ).

The free surface in Figure 42 is derived from a box of large aspect ratio. The dip centered above the vortex at  $z = 4$  is due mainly to the low pressure caused by the relatively high velocities at the core edge, and is aided by the downward normal stress from the convective term. The raised lip to the right of this low pressure dip is totally due to the upward normal stress from the convective term. We have essentially Stokes flow in the two extreme ends of the box, therefore the free surface height is governed by the pressure. In this case, pressure is high on the left, raising the surface, and low on the right, lowering it. One crude way of understanding this pressure behavior is to consider a two-compartment box (Figure 43), with the vortex at the center acting to pump fluid from right to left at the top, and from left to right at the bottom. The viscous wall impedes the flow at the bottom by exerting a force to the left. At the top there is no such restriction, and the vortex continues to pump

fluid to the left unimpeded. This pressurizes the left compartment and depressurizes the right. As the vortex has not yet reached the side walls, we would have to take more terms of the perturbation solution in order to derive an acceptable analytic expression to describe the figure, though one might convince oneself that the region above the core is qualitatively described by Equation 47.

Figure 44 shows the free surface above a very strong vortex ( $R_v = 25$ ). Here, the shape of the free surface is totally dominated by the very high velocities, and thus very low pressures, inside the core. The shape is almost a symmetric cosine, centered over the core. This is consistent with Equation 48 (which should also be the form of Equation 47 when higher order terms are included); the high  $R_v$  in the denominator of the asymmetric stress term makes it negligible, leaving the symmetric part of the pressure. In this case, the stress distribution above the viscous wall is actually well-described by the solution above a slippery wall, because the very great strength of the vortex renders viscous wall effects relatively unimportant. As circulation decreases, the slippery wall solution becomes increasingly less applicable.

Our ultimate purpose was to see whether or not the deformation of the free surface was a large effect. We can see now that it is not. At  $R_v$  within our range of interest, the maximum deformation is less than .3%. This is why free surface deformation in laminar flow has not been detected in the course of Yang's or Suri's experiments - they estimate their observational threshold to be  $\approx 5\%$ . Therefore they may be able to detect deformations due to vortices of strength  $R_v \sim 50$  (providing the flow remained laminar), but not much weaker. If these deformations later come within our observational grasp, the above work predicts that increases in height will be detected above the downflow region (high speed streak), and decreases in height above the upflow region (low speed streak). This small value of the deformation is the result of the factor  $\frac{R_v^2}{R}$  in Equation 47. In our range of interest for observed laminar flows ( $R_v \sim 5$ ,  $R \sim 2500$ ), the deformation is then  $\Delta h \sim \mathcal{O}(\frac{1}{100})$ .

We conclude that the deviation of the free surface from a slippery flat plane is indeed negligible for values of  $R_v$  and  $R$  within our range of interest, and that our previous assumption of such a slippery flat plane for  $\partial_{top}$  is well justified.

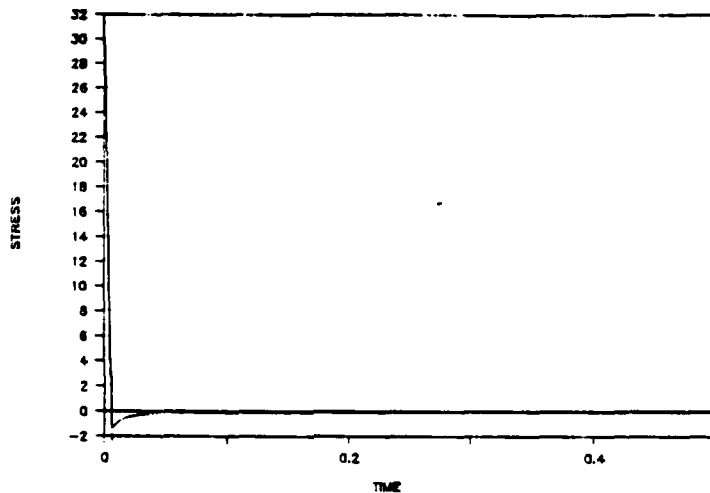


Figure 39: Horizontal stress on a vortex cell,  $a = 4$ , due to an initially potential vortex above a plane viscous wall.

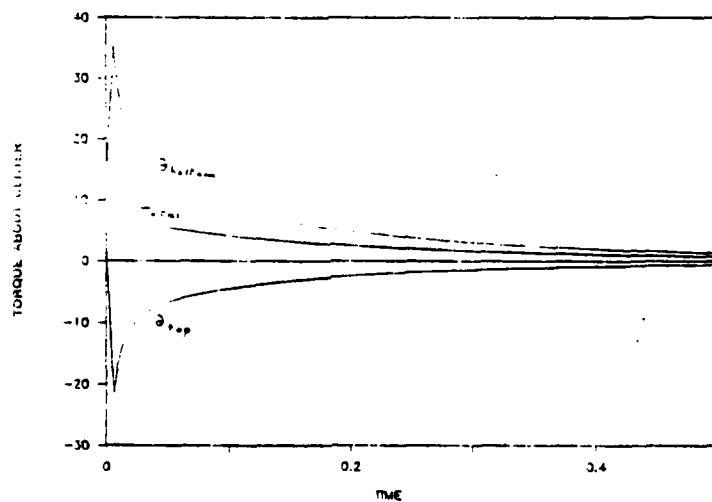


Figure 40: Torque about center of vortex cell due to normal forces on  $\partial_{top}$ , torque due to normal and tangential forces on  $\partial_{bottom}$ , and the total torque due to an initially potential vortex above a plane viscous wall. Positive torque is counterclockwise.

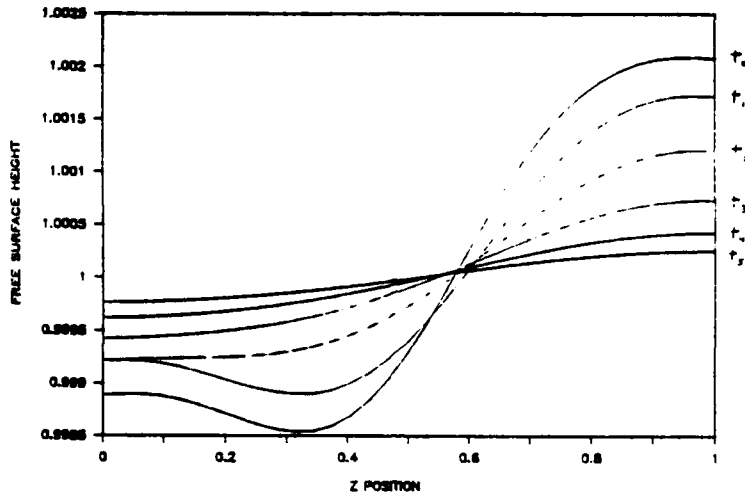


Figure 41: Free surface height at successive times due to an initially potential vortex above a plane viscous wall.  $\alpha = 1$ ,  $R_\nu = 5$ .

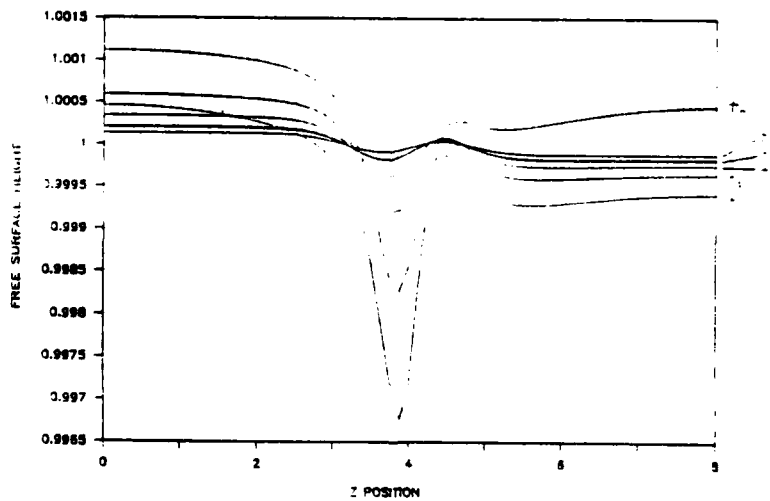


Figure 42: Free surface height over time due to an initially potential vortex above a plane viscous wall.  $\alpha = 3$ ,  $R_\nu = 5$ .



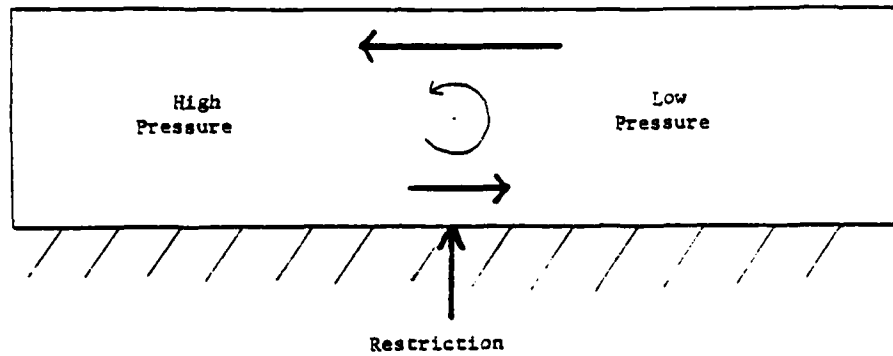


Figure 43: Development of high and low pressure regions in the Stokes flow regions near a vortex above a viscous wall.

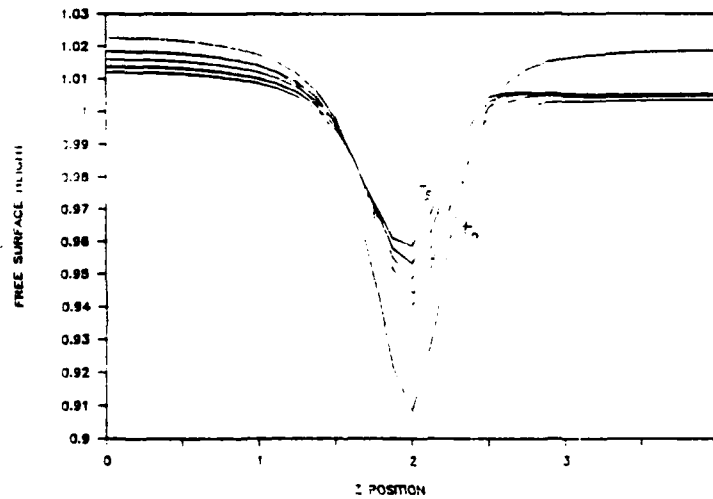


Figure 44: Free surface height at successive times due to an initially potential vortex above a plane viscous wall.  $n = 4$ ,  $R_v = 25$ .

## C COMPARISON WITH SEMI-INFINITE DOMAIN

Previous computations have all involved flow inside a box of some finite aspect ratio. When studying the induced vortex, it has often seemed that the side walls played some part in its development, particularly at small  $R_\nu$  (eg, Figures 12 and 13). It is now desirable to remove the side walls and consider the flow in a semi-infinite domain (bounded in  $y$ , unbounded in both  $x$  and  $z$ ) which is more representative of Yang's water table flow. Specifically, we want to see if our previous conclusions about the induced vortex (existence, strength, effect on the streamwise flowfield) still hold.

To do this, we map the unbounded coordinate  $z \in (-\infty, \infty)$  onto  $\xi \in [-\frac{1}{2}, \frac{1}{2}]$  via the mapping  $\xi = \frac{1}{\pi} \tan^{-1} z$ . Equations 1 to 3 then transform as:

Crossflow:

$$\omega_x + v\omega_y + w\omega_z \frac{\cos^2 \pi\xi}{\pi} = \frac{1}{R_\nu} (\omega_{yy} + \omega_{\xi\xi} \frac{\cos^4 \pi\xi}{\pi^2} - \omega_\xi \frac{\cos^2 \pi\xi \sin 2\pi\xi}{\pi}) \quad (49)$$

$$\psi_{yy} + \psi_{\xi\xi} \frac{\cos^2 \pi\xi}{\pi} - \psi_\xi \frac{\cos^2 \pi\xi \sin 2\pi\xi}{\pi} = -\omega \quad (50)$$

$$w = \psi_y \quad (51)$$

$$v = -\psi_\xi \frac{\cos^2 \pi\xi}{\pi} \quad (52)$$

Streamwise:

$$u_x + v u_y + w u_z \frac{\cos^2 \pi\xi}{\pi} = \frac{2}{R_\nu} + \frac{1}{R_\nu} (u_{yy} + u_{\xi\xi} \frac{\cos^4 \pi\xi}{\pi^2} - u_\xi \frac{\cos^2 \pi\xi \sin 2\pi\xi}{\pi}) \quad (53)$$

Boundary conditions and initial conditions are unchanged.

At an initial vortex strength of  $R_\nu = 5$ , we see that a major induced vortex appears on the right and a minor one appears on the left (Figure 45). (Note that machine resolution stops short of  $\pm\infty$ ). Now the major induced vortex initially forms very far to the right (as close to  $z = \infty$  as machine resolution allows), the center moves to within a distance of two flow-depths of the main vortex, and then rises (Figure 46). However, the circulation of this induced vortex is still negligible - three orders of magnitude smaller than that of the main vortex (Figure 47). Despite its presence, vorticity contours remain almost symmetric in  $z$  (Figure 48). It has no noticeable impact on the streamwise velocity field (Figure 49).

At stronger  $R_\nu$  (eg,  $R_\nu = 15$ ) the major induced vortex is stronger, but remains two orders of magnitude weaker than the main vortex (Figure 50). Vorticity contours are more asymmetric due to increased convection of negative vorticity generated at the viscous wall (Figure 51). The effect of the induced vortex on the streamwise velocity field is still negligible (Figure 52). At this higher  $R_\nu$ , it is interesting to note the highly inflectional profiles produced in the vortex core. This rotation of the shear angle by a little more than  $180^\circ$  is as predicted by Pearson for this  $R_\nu$ .

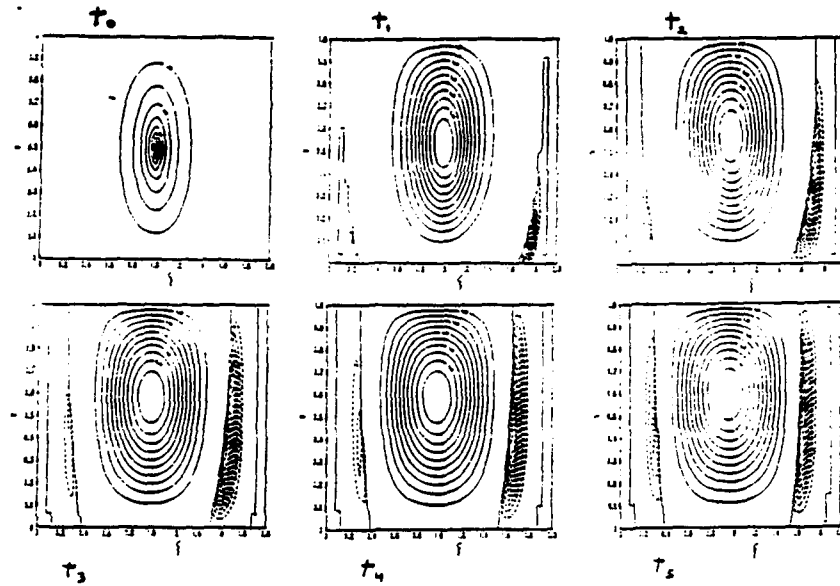


Figure 45: Development of the main and induced vortex streamlines over time,  $R_\nu=5$ . The vortex cell is unbounded in  $z$  and is mapped onto a finite domain. Smaller contour intervals are used for the induced vortex.

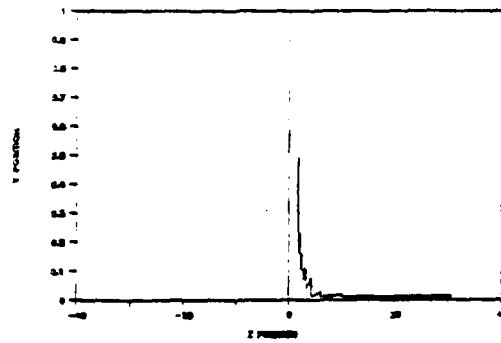


Figure 46: Motion of the predominant induced vortex over time. In the early stages, tangential velocities are very weak and the vortex center is not well defined. In this plot, the unbounded coordinate  $z$  has been truncated.

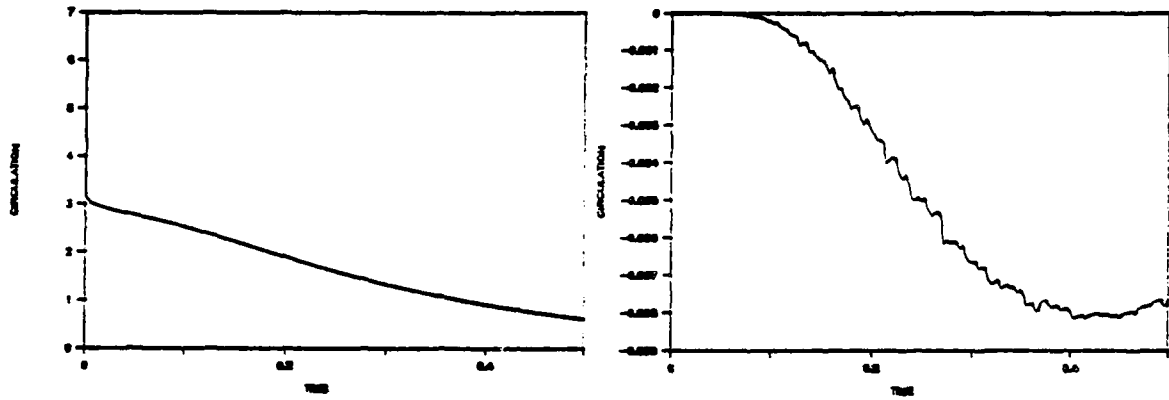


Figure 47: Circulations of the main vortex (left graph) and right induced vortex (right graph) of Figure 45. The circulation of the induced vortex has a small 'noisy' component because of uncertainty in the location of the separating streamline between it and the main vortex.

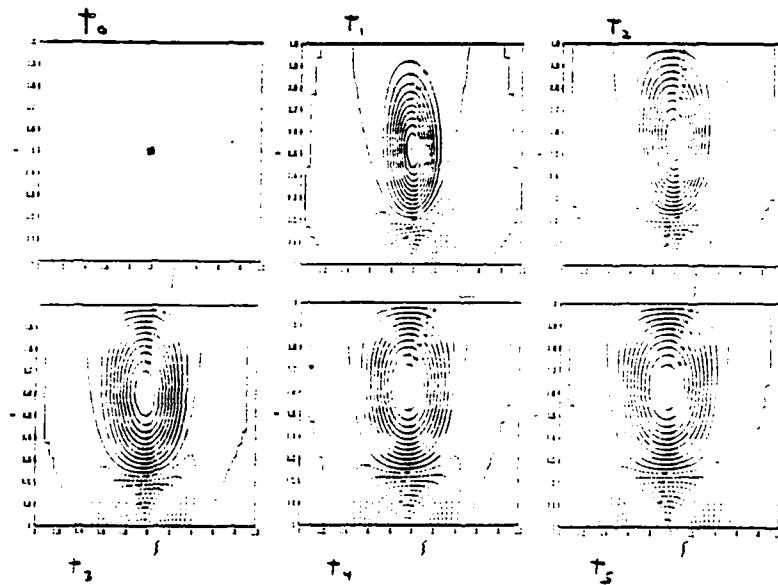


Figure 48: Vorticity contours for the vortices of Figure 45. The initial condition is a delta function of vorticity.

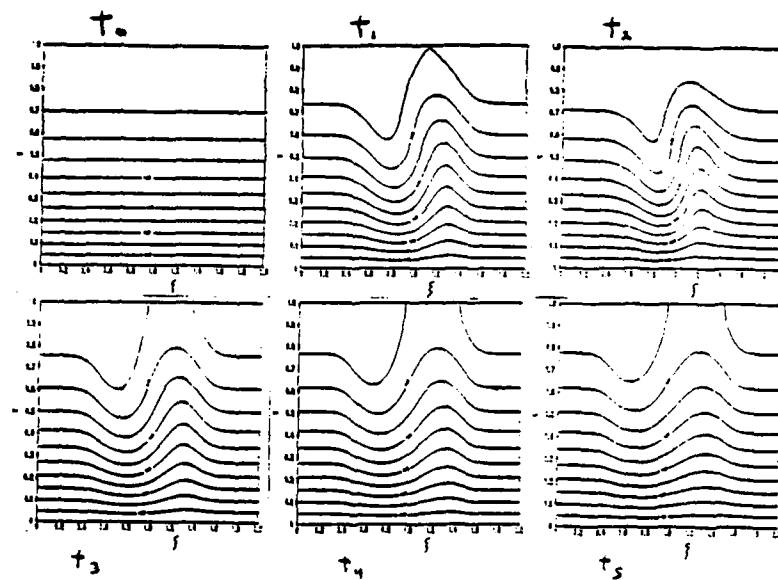


Figure 49: Streamwise velocity contours for the vortices of Figure 45. Note that the induced vortex causes no noticeable perturbation ( $z \approx 1.8$ ).

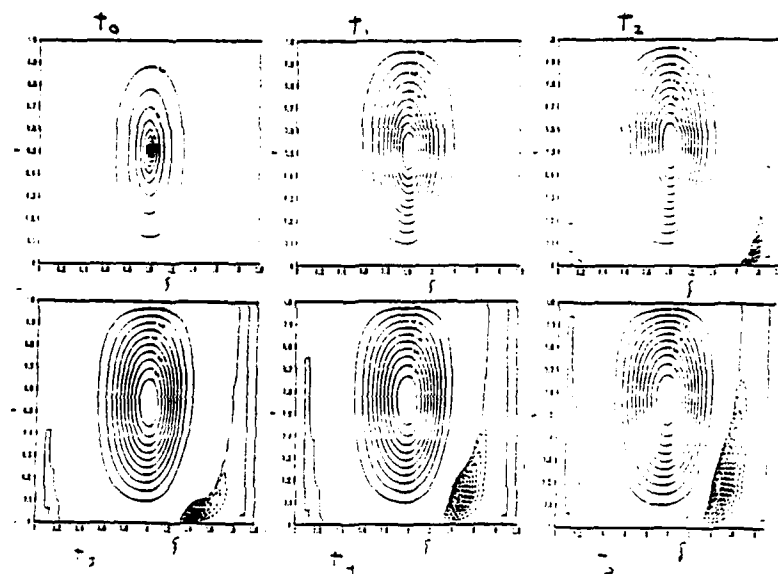


Figure 50: Development of the main and induced vortex streamlines over time at  $R_v = 15$ .

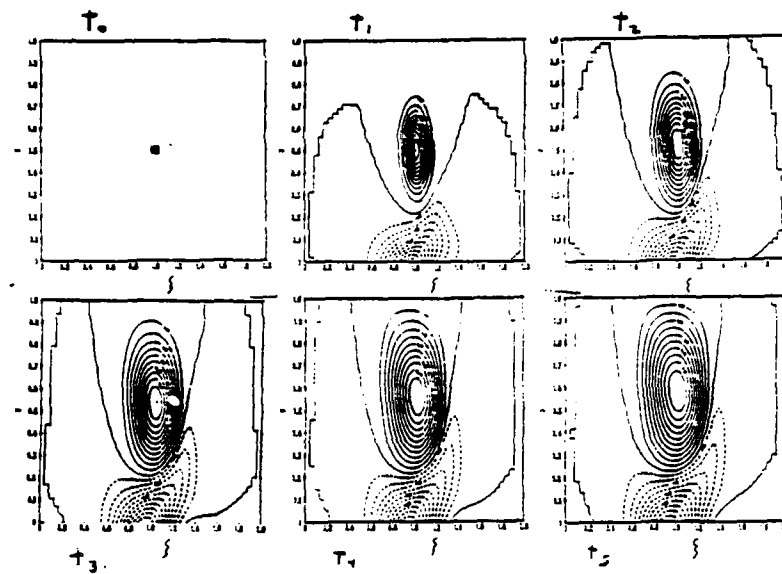


Figure 51: Vorticity contours for the vortices of Figure 50. Initial condition is a delta function.

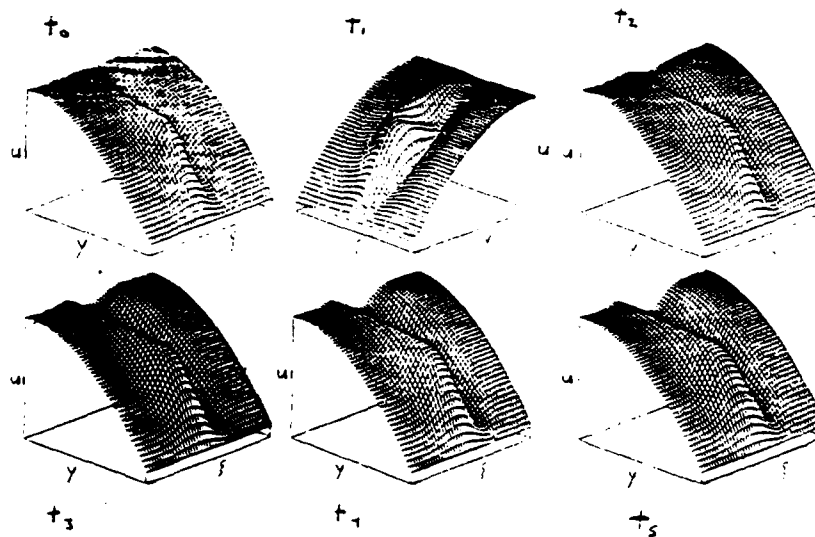


Figure 52: Streamwise velocity for the vortices of Figure 50. Note the lack of noticeable effect of the induced vortices, and the highly inflectional profiles in the core of the main vortex.

## REFERENCES

1. Herbert, T. (1977). Finite amplitude stability of plane parallel flows, Proceedings of the AGARD Symposium on Laminar-Turbulent Transition, Paper 3.
2. Klebanoff, P., Tidstrom, K., and Sargent, L. (1962). The three-dimensional nature of boundary layer instability. *J. Fluid Mech.*, 12, 1-34.
3. Mitchner, M. (1952). The propagation of turbulence into a laminar boundary layer. PhD Thesis, Division of Applied Sciences, Harvard University.
4. Nishioka, M., Iida, S., and Ichikawa, Y. (1975). An experimental investigation of the stability of plane Poiseuille flow. *J. Fluid Mech.*, 72, 731-751.
5. Orszag, S., and Patera, A. (1983). Secondary instability of wall-bounded shear flows. *J. Fluid Mech.*, 128, 347-385.
6. Pearson, C. (1985). A class of instabilities associated with streamwise vorticity. PhD Thesis, Division of Applied Sciences, Harvard University.
7. Pearson, C., and Abernathy, F. (1984). Evolution of the flow field associated with a streamwise diffusing vortex. *J. Fluid Mech.*, 146, 271-283.
8. Stuart, J. (1965). N.P.L. Report 1147. AGARD Report 514.
9. Suri, A. (1988). Streamwise vortices in shear flow transition. PhD Thesis, Division of Applied Sciences, Harvard University.
10. Taylor, G. I. (1923). On the decay of vortices in a viscous fluid. *Philosophical Magazine*, XLVI, 671-674.
11. Werlé, H. (1980). *Recherche Aérospatiale*, 35-49.
12. Yang, Z. (1987). A single streamwise vortical structure and its instability in shear flows. PhD Thesis, Division of Applied Sciences, Harvard University.

# Generalized Rewriting in Type Theory\*

David A. Basin  
Department of Computer Science,  
Cornell University, Ithaca, NY 14853  
basin@cs.cornell.edu

## Abstract

While type theories such as Nuprl are expressive logics for theorem proving, they present difficulties for designers of term rewriting systems. The two most serious difficulties are: 1) Nuprl does not provide a global equality. Instead users rewrite over arbitrary user-defined relations. 2) Each rewrite step must be proved valid. In general, these proofs cannot be recursively generated.

We have overcome these difficulties and designed a package that works well in practice. Our solution is an extensible system for directing and validating relational inferences. The heart of our package is a set of operators that use a user-supplied lemma database to create new rewrites from old ones. These routines place no restrictions on relations; a rewrite's success depends on the strength of the database. Overall, the package allows rewrites to be pieced together in numerous ways, providing the user with a tool to construct sophisticated rewrite strategies.

## 1 Introduction

Our research addresses rewriting in Nuprl, a sequent calculus formulation of a constructive type theory similar to Martin-Löf's[13]. While most systems assume that their rewrite functions are reliable, in Nuprl, every rewrite must be proven correct. This is difficult as Nuprl's expressive power yields undecidable typing problems: There is no effective procedure that determines if a term belongs to a given type, or determines the type of a given term. As a result it is undecidable whether even simple rewrites are valid. For example, each type comes with its own equality and a proof of  $t =_T t'$ <sup>1</sup> necessitates a proof that both  $t$  and  $t'$  are members of  $T$ . Furthermore, the validity of standard congruence (substitutability) reasoning must be proven. If  $t =_T t'$  then to substitute  $t'$

\*This work was supported, in part, by an IBM Fellowship.

<sup>1</sup>We shall follow the following syntactic conventions:  $T, A, B, \dots$  shall represent types;  $Q, R,$  and  $S$  relations;  $t, r, s, \dots$  terms and  $x, y, z, \dots$  variables.



for  $t$  in  $B[t]$  one must show that for every  $x$  and  $y$  equal in  $T$ ,  $B[x]$  and  $B[y]$  are equal types. This too is undecidable. We explore these problems and their implications for rewriting in Section 2.

Despite these difficulties, we have implemented a rewrite package that works well in practice. Our approach provides operators that construct *relational conversions*. Given a sequent  $p$  and a term  $t$ , a conversion yields a triple: a relation  $R$ , a term  $t'$ , and an ML program called a *tactic*, which should prove  $t R t'$  under the assumptions in  $p$ . As with tactics, a conversion may fail; but if it succeeds, then  $t R t'$  may be used as an assumption.

Basic conversions are generally constructed from lemmas of the appropriate form (which is roughly that of a universally quantified chain of implications that ends in a relation  $t R t'$ ). We provide an operator that takes a lemma and a relation and constructs a conversion which, if it succeeds, rewrites an instance of the left-hand side of the relation in the lemma into a corresponding instance of the right hand side. Conversions themselves are data-values, and higher order combinators provide a means to compose them and form new conversions. For example, *THENC* is a combinator that sequentially composes two conversions and *SubConv* applies a conversion to the immediate subterms of a term. A user-supplied lemma database is used in the construction of tactics that prove these rewrites valid. *SubConv*, for example, uses this database to produce a tactic that generates congruence proofs. No restrictions are placed on relations; they need not even be equivalence relations. However, whether a conversion succeeds depends on the strength of the database. The overall package is highly modular; conversions can be put together in numerous ways, allowing the user to construct sophisticated rewrite strategies. Section 3 details our implementation and illustrates the power of our approach with examples.

Our package is similar to Paulson's[14]<sup>2</sup>. Both are collections of ML programs that can be applied in a modular higher-order style and many of the combinators are functionally identical. However, our package differs in two important respects. First, Paulson's package allows rewriting only over two relations: term equality and formula equivalence. These are special cases of our general relational approach. Second, he provides separate notions of conversions (*conv* and *fconv*) for each relation, each with its own fixed strategy for proving rewrites valid. Our implementation allows the user to extend relational inference simply by expanding the lemma database. Of course, such extensibility is necessary in light of undecidable typing, but we have found the resulting generality useful in practice.

Our package is currently used by the author and other researchers at Cornell working in hardware verification. We have successfully constructed a variety of rewrite procedures including a predicate calculus simplifier and associative/commutative term normalizers. We have also explored rewriting in con-

---

<sup>2</sup>It also bears similarities to ideas of Howe and Steeg[9] who implemented rewriting in Nuprl over the (type parametrized) equality relation.

structive set theory. In this domain, set equality is a user defined predicate. Congruence reasoning requires proofs that each set theoretic operator respects this predicate, and these proofs must be composed to justify each rewrite. Without a package such as ours, the burden of constructing such justifications by hand would be unmanageable.

Our approach to rewriting has applications outside of Nuprl and the related theories of Martin-Löf. Logics with decidable typing but weak equality, such as the Calculus of Constructions[4], also require that users define their own equalities and prove their rewrites valid. We believe that as such logics become more popular, extensible rewrite systems such as ours will become integrated into their use.

## 2 Nuprl.

In this section, we highlight those aspects of Nuprl relevant to rewriting. The interested reader may consult [3] for a more complete account.

The basic objects of reasoning in Nuprl are types and members of types. The rules of Nuprl deal with sequents, objects of the form

$$x_1:H, x_2:H_2, \dots, x_n:H_n \gg G.$$

To judge a sequent true essentially means that when given members  $x_i$  of  $H_i$ , one can construct an inhabitant of the goal  $G$ . Nuprl's rules are applied in a top-down fashion. That is, they allow us to refine a goal, such that a proof of the goal may be constructed from proofs of the subgoals. Nuprl provides two kinds of inference rules: primitive rules and ML programs called tactics. Nuprl tactics are similar to those in LCF[6]: Given a sequent as input, they apply primitive inference rules and other tactics to the sequent. Tactics serve as derived inference rules; their correctness is justified by the way the type structure of ML is used.

Nuprl's type theory is expressive; its intent is to facilitate the formalization of constructive mathematics. Types are stratified into a cumulative hierarchy of universes. A term is a well-formed type if and only if it inhabits some universe  $U_i$ . Type constructors include dependent function space, dependent product, disjoint union, equality, set type, and quotient type. Each type comes with its own equality relation. The equality type is a three place relation  $t_1 =_T t_2$  that is inhabited exactly when  $t_1$  and  $t_2$  are equal members of  $T$ . Propositions are represented via the propositions-as-types correspondence: A proposition is true if and only if the type associated with it is inhabited.

A special case of type checking in Nuprl is deciding if a program meets its specification. As a result, a number of typing problems are undecidable. These include: type membership ( $t \in T$ ); type inference; and type well-formedness ( $T \in U_i$ ). The first two problems imply that there is no uniform way of proving the identity rewrite valid. Given a term  $t$ , we cannot infer a type  $T$  it inhabits;

even if we are given a  $T$ , we cannot prove  $t \approx_T t$  as that requires a proof of  $t \in T$ . Type well-formedness is a semantic property as there are syntactically legitimate yet meaningless terms. As we shall see in the next section, proofs of certain well-formedness goals may be thought of as proofs of relational congruence.

### 3 Generalized Rewriting

Rewriting is the process of finding a term  $t'$  that is somehow simpler than a given term  $t$ . Moreover, the terms must stand in some relation. Our rewrite functions are called conversions. Given a sequent  $p$  and a term  $t$ , a conversion returns a *rewrite triple*: a relation  $R$ , a term  $t'$ , and a tactic  $tac$  that proves the assertion  $t R t'$  under the assumptions in  $p$ . We call  $R$  the *rewrite relation*. We take an abstract view of relations. A relation is specified by a constructor that maps sequents and pairs of terms to terms, and a destructor that breaks down terms into pairs. Aside from the ability to construct and destruct relations, no other properties are assumed. Conversions are applied with Nuprl's cut rule: If  $c$  is a conversion, and  $c(p)(t)$  returns the triple  $\langle R, t', tac \rangle$ , then  $t R t'$  may be proved with  $tac$  and used as an assumption.

In the remainder of this section we describe how conversions are constructed and composed. Our approach provides operators that construct conversions using primitive inference rules and lemmas and provides combinators that build conversions from simpler ones. This modular higher-order approach to rewriting originated with Paulson who provides an account in [14]. Rather than duplicate Paulson's account, we shall instead focus on what is novel about our approach: how we coordinate rewrites over arbitrary relations and our use of lemmas to direct inference and validation in a theory with undecidable typing problems.

#### 3.1 Basic Conversions

Basic conversions are of two types: primitive and lemma-based. Primitive conversions, such as the identity conversion  $IdConv$  and (values of)  $IdConvWithR$ , prove their rewrites valid with primitive inference rules.  $IdConvWithR$  takes a relation  $R$  as an argument and returns the conversion  $\lambda p. \lambda t. \langle R, t, tac \rangle$ . A Nuprl *library*, an ordered collection of definitions, theorems, and other objects, is used to construct a database of lemmas expressing relational properties. For the identity rewrite,  $IdConvWithR$  searches this database for a lemma of the form

$$\gg t R t.$$

If no such lemma is found, the relation is assumed irreflexive and the conversion fails. Otherwise the conversion succeeds and  $tac$  proves  $t R t$  with the found lemma.  $IdConv$  is similar to  $IdConvWithR$ . But instead of taking a relational argument, type inference routines described in [7] infer a  $T$  for the relation  $\approx_T$ .

The conversion fails when  $T$  cannot be inferred, but this has not been a problem in our experience.

Other primitive conversions include *FailConv*, the always failing conversion, *ReduceConv*, which replaces redices by their contracta, and *Simplify* which performs arithmetic simplification.

Lemma-driven conversions construct rewrite triples from lemmas in a user's library. These lemmas take the form

$$\forall x_1 : T_1. \forall x_2 : T_2. \dots \forall x_n : T_n. A_1 \Rightarrow A_2 \Rightarrow \dots \Rightarrow A_m \Rightarrow s R s'.$$

The function *LemmaToConv* takes two arguments, a lemma  $l$ , in which  $m$ , the number of assumptions, is zero, and a relation  $R$ , and returns a conversion  $\lambda p. \lambda t. \langle R, t', tac \rangle$ . *LemmaToConv* matches  $t$  against the left operand of the relation  $s R s'$  in  $l$ . If the match fails, the conversion fails. Otherwise, the variables  $x_1, \dots, x_n$  are bound by the match and used to instantiate  $s'$  which is returned as  $t'$ . When executed the tactic  $tac$  applies  $l$  to prove  $t R t'$ .

Conditional rewrites are constructed similarly. *ImpLemmaToConv* takes as inputs a lemma  $l$ , which may have assumptions, a relation  $R$ , and a tactic. Before constructing its rewrite triple, *ImpLemmaToConv* applies the input tactic to the instantiated assumptions  $A_1, \dots, A_m$  and fails if this application fails.  $Tac$  appropriately combines both the lemma  $l$  and the input tactic.

Both *LemmaToConv* and *ImpLemmaToConv* are driven by first order matching. We have also implemented powerful versions that are driven by second order matching. As second order matches are not necessarily unique, these functions take an additional argument, a match discriminator function that chooses an appropriate match from a set of matches.

### 3.2 Composing Conversions

Given functions that construct basic conversions, the next step is to provide operators that combine them. The two basic combinators are *ORELSEC* and *THENC*. The former provides selective composition and the latter a method of sequentially composing conversions.

*ORELSEC* is based on ML failure. In ML, the expression  $e_1 ? e_2$  computes the value of  $e_1$  and if that fails it computes the value of  $e_2$ . Thus, we define  $(c_1 \text{ ORELSEC } c_2)(p)(t)$  as

$$c_1(p)(t) ? c_2(p)(t).$$

*THENC* uses a generalized kind of transitivity reasoning to sequentially compose conversions. Unlike selective composition, this requires use of the lemma

<sup>3</sup>There are functions *RevLemmaToConv* and *RevImpLemmaToConv* which match against the right operand.

database. Given conversions  $c_1$  and  $c_2$ ,  $(c_1 \text{ THENC } c_2)(p)(t)$  computes the triples

$$\begin{aligned} \langle R_1, t_1, tac_1 \rangle &= c_1(p)(t) & \text{and} \\ \langle R_2, t_2, tac_2 \rangle &= c_2(p)(t_1). \end{aligned}$$

If either conversion fails, then *THENC* fails. Otherwise, *THENC* uses  $R_1$  and  $R_2$  as keys and searches the database for a sequencing lemma of the form

$$t R_1 s, s R_2 r \gg t R_3 r.$$

If such a lemma is found, the triple  $\langle R_3, t_3, tac \rangle$  is returned, where  $tac_1$ ,  $tac_2$ , and the lemma are appropriately combined into  $tac$ .

There is no need for the relations to be identical. For example,

$$t < s, s \leq r \gg t < r$$

is a valid sequencing lemma. Our implementation insists that there is at most one sequencing lemma for any pair of relations; more general approaches allowing multiple sequencing lemmas are possible. Such approaches would be implemented analogously to the congruence reasoning routines described in Section 3.3.2.

*ORELSEC* and *THENC*, along with *FailConv* and *IdConv*, provide the basis for multi-way choice and repetition. The operator *FirstC* returns the first successful conversion from its argument list  $[c_1; \dots; c_n]$ . It is equivalent to

$$c_1 \text{ ORELSEC } \dots \text{ ORELSEC } c_n$$

and is defined recursively in terms of *ORELSEC* and *FailConv*. *RepeatC*( $c$ ) repeatedly applies the conversion  $c$  until failure. It is defined recursively as

$$(c \text{ THENC } (\text{RepeatC}(c))) \text{ ORELSEC } \text{IdConv}.$$

With *FirstC* and *RepeatC* it is easy to construct a general chaining procedure where relational inferences are combined to the extent justifications are provided in the database.

### 3.3 Congruence Conversions

#### 3.3.1 Congruence Proofs

Subterm rewriting requires the construction of tactics that generate congruence proofs. These proofs can be subtle and in practice more difficult to construct than reflexivity and transitivity proofs. We shall first examine these proofs in the simplest possible setting, equality congruence, and then consider general relational congruence.

Given  $t =_T t'$ , to conclude  $r[t/x] =_{T'} r[t'/x]$ , we must prove that equal members  $x$  in  $T$  yield equal members  $r$  in  $T'$  ( $x$  may be free in  $r$ ), i.e.,

$$\frac{\gg t =_T t' \quad x:T \gg r \in T'}{\gg r[t/x] =_{T'} r[t'/x]}$$

The extra premise, that  $r$  is functional in  $x$ , takes the form of a membership goal and, as discussed in Section 2, is in general undecidable.

Congruence proofs are by induction on the structure of Nuprl terms. A term is a tree. It is specified by an  $n$ -ary term constructor and its  $n$  subterms. If the subterms  $t_i$  of a term  $T$  are rewritten via equality to  $t'_i$ , congruence is proven by decomposing  $T$  into its immediate subterms and proving that  $T$  is functional in the types of these terms. The  $t_i$  are then either proved equal to  $t'_i$  by reflexivity (*equality*), proved equal by assumption (*hypothesis*), or recursively decomposed. We illustrate this with the following example. Given  $r =_A s$ , we prove

$$\gg f(g(r)) =_C f(g(s))$$

under the assumptions  $f: B \rightarrow C$ , and  $g: A \rightarrow B$  as follows.

1	...	$\gg f(g(r)) =_C f(g(s))$	<i>by intro using <math>B \rightarrow C</math></i>
1.1	...	$\gg f =_{B \rightarrow C} f$	<i>by equality</i>
1.2	...	$\gg g(r) =_B g(s)$	<i>by intro using <math>A \rightarrow B</math></i>
1.2.1	...	$\gg g =_{A \rightarrow B} g$	<i>by equality</i>
1.2.2	...	$\gg r =_A s$	<i>by hypothesis</i>

In the above example, observe that each step in the congruence proof can be viewed as justifying a rewrite with respect to some (different) equality relation. As we "peel" through the various term constructors, our rewrite relation changes.

This approach of proving functionality by recursive decomposition generalizes to congruence proofs for arbitrary relational rewrites. Suppose  $T$  is a term containing  $n \geq 1$  subterms  $t_i$  that are rewritten with respect to  $n$  (possibly distinct) relations. Viewing  $T$  as a term tree, for each term constructor  $\theta$  in the path from each  $t_i$  to the root of  $T$  we must prove that  $\theta$  is functional in the relations used to rewrite its immediate subterms. Formally,  $\theta$  is said to be  $R_0$ -functional in the relations  $R_1, \dots, R_n$  whenever

$$\theta(t_1, \dots, t_n) R_0 \theta(t'_1, \dots, t'_n) \tag{1}$$

is provable under the assumptions

$$\begin{array}{l} t_1 \quad R_1 \quad t'_1, \\ \vdots \\ t_n \quad R_n \quad t'_n \\ C_1 \\ \vdots \\ C_m. \end{array} \tag{2}$$

The  $C_1, \dots, C_m$  are additional auxiliary conditions which may be required of the  $t_i$ . We call the lemma that proves that Equation 1 follows from the assumptions in 2, a *congruence lemma*. A congruence proof consists of recursively decomposing a term by applying congruence lemmas until the relations among the resulting subterms follow by reflexivity or from a given conversion.

As a simple example, suppose we are given the term  $\forall x_1:T_1 \dots \forall x_n:T_n. t_1$  and the conversion  $c$  where  $c(p)(t_1) = \langle \Leftrightarrow, t'_1, tac_1 \rangle$ . Then  $n$  applications of the congruence lemma

$$\begin{aligned} >> \forall T:U_1. \forall P, Q:T \rightarrow U_1. (\forall x:T. (P(x) \Leftrightarrow Q(x))) \\ &\Rightarrow (\forall x:T. P(x) \Leftrightarrow \forall x:T. Q(x)), \end{aligned}$$

which proves that the relation  $\forall$  is  $\Leftrightarrow$ -functional in the relation  $\Leftrightarrow$ , reduces proving

$$\forall x_1:T_1 \dots \forall x_n:T_n. t_1 \Leftrightarrow \forall x_1:T_1 \dots \forall x_n:T_n. t'_1$$

to proving  $t_1 \Leftrightarrow t'_1$ , which is proved by the  $tac_1$ .

### 3.3.2 Inferring Rewrite Relations

In our implementation, the function *SubConv* provides the basis for subterm rewriting. *SubConv(c)* is a conversion that applies the conversion  $c$  to the immediate subterms of a term. The tactic it produces justifies the subterm rewrite. Repeated application of *SubConv* allows rewriting of arbitrary subterms.

To construct a rewrite triple, *SubConv* use the lemma database as a source for congruence lemmas. However, there may be more than one choice of  $R_0$  for a given  $\theta$  and  $R_1, \dots, R_n$ . For example, suppose that  $\theta$  is the multiplication operator  $*$ , and  $R_1$  and  $R_2$  are  $=$  and  $<$ . Then with an additional inequality condition, the rewrite relation  $R_0$  may be either  $<$  or  $\neq$ . That is, both

$$a = c, b < d, a > 0 >> a * b < c * d \quad (3)$$

and

$$a = c, b < d, a > 0 >> a * b \neq c * d \quad (4)$$

are valid congruence lemmas. Moreover, in the proper proof context, each of these lemmas could find application.

How one determines which congruence lemma is applicable is an important question. If *SubConv* chooses an improper rewrite relation, then a rewrite may either eventually fail, as it will be unable to create a tactic that constructs a congruence proof, or the relation chosen may be too weak, rendering the rewrite useless. Hence, it is important to have an effective strategy for selecting rewrite relations. In this section we outline two methods for controlling relational inference: a powerful but computationally expensive method, and a simplified heuristic method that works well in practice. Our implementation is based on the latter.

The most general method of subterm rewriting is to construct all possible rewrites as permitted by the congruence lemmas contained in the database. Such a method necessitates a generalized type of conversion that returns not a single rewrite triple, but rather a set of triples. In such a setting, if a term  $t$ 's outer most operator is the  $n$ -ary constructor  $\theta$ , then  $SubConv(c)(p)(t)$  would:

1. Associate with each of the  $n$  immediate subterms  $t_i$  of  $t$  either the set of rewrite triples returned by  $c(p)(t_i)$ , or if this fails, the singleton triple returned by  $IdConv(p)(t_i)$ . If all  $n$  subterm rewrites fail, then  $SubConv$  fails.
2. Form all possible  $n$ -tuples where the  $i$ th element of the tuple comes from the triple set associated with the term  $t_i$ . For each such tuple, search the lemma database for a congruence lemma which states that for some  $R_0$ ,  $\theta$  is  $R_0$ -functional in the tuple's  $n$  rewrite relations. If such a lemma exists, use it to construct a new rewrite  $\langle R_0, \theta(t'_1, \dots, t'_n), tac \rangle$  where  $tac$  applies to  $t$  and the  $tac_i$ .
3. Return the set of new triples or fail if the set is empty.

The effect of the above construction is that  $SubConv$  generates, bottom-up, all possible congruence proofs. The resulting set of triples can be used to selectively add new facts to the hypothesis list of the sequent and for subsequent inference.

While this approach makes the fullest use of the lemma database, its time complexity is exponential in the depth of the rewritten subterms. Our solution to this combinatorial explosion rests on the observation that in most cases, when there is a choice among rewrite relations, it suffices to pick the strongest relation. For example, it is preferable to know that two types are equal instead related by bi-implication (if and only if). Similarly, bi-implication is a stronger relation than implication, and less-than is stronger than less-than-or-equal. So, for example, one generally prefers to use the congruence lemma given by Equation 3 over Equation 4.

Our approach, which is linear in the depth of the rewritten subterms, always returns the strongest possible rewrite relation. We use a user provided table to determine relative relational strength.  $SubConv(c)(p)(t)$  produces a single rewrite triple as follows:

1. The function  $c(p)$  is applied to each subterm  $t_i$ . For each subterm this yields the triple  $\langle R_i, t'_i, tac_i \rangle$ , or, failing that, the triple  $IdConv(p)(t_i)$ . If all  $n$  conversions fail, then  $SubConv$  fails.
2. The operator  $\theta$  and the relations  $R_i$  are used to index into the library for congruence lemmas that specify relations  $R_0$  such that  $\theta$  is  $R_0$ -functional in the  $R_i$ . When more than one such lemma is found, the one with the strongest  $R_0$  is chosen. If no such lemma is found,  $SubConv$  fails.



3. A tactic *tac* is constructed that applies the lemma found in the previous step. This reduces the proof of  $t R_0 t'$  (where  $t' = \theta(t'_1, \dots, t'_n)$ ) to the subgoals  $t_i R_i t'_i$ . *Tac* then proves each such subgoal by *tac<sub>i</sub>*.
4. The rewrite triple  $\langle R_0, t', tac \rangle$  is returned.

### 3.3.3 Subterm Traversal

*SubConv* provides the basis for traversing terms recursively. It is now easy to write an operator *Depth(c)* that recursively rewrites all subterms of a term in depth-first order. Its recursive definition is

$$((\text{SubConv } (\text{Depth}(c))) \text{ THENC } \text{RepeatC}(c)) \text{ ORELSEC } (\text{RepeatC}(c)).$$

Similarly, top-down rewriting is accomplished by *Top(c)* whose recursive definition is

$$\text{ProgressC}(\text{RepeatC}(c)) \text{ ORELSEC } (\text{SubConv } (\text{Top}(c))).^4$$

### 3.3.4 An Example

Let *c* be a conversion that rewrites *t* to *t'* under the less-than relation  $<$ . Suppose *m* is non-negative, the sequent *p* contains appropriate well-formedness hypotheses, and the database contains the congruence lemmas given by

$$a = c, b \leq d \gg a < b \Rightarrow c < d \quad (5)$$

and Equation 3. Then

$$\text{Top}(c)(p)(n < m * t)$$

returns the triple

$$\langle \Rightarrow, n < m * t', tac \rangle. \quad (6)$$

Tracing *Top*'s recursive execution, we find that after two calls to *SubConv c(p)(t)* returns the triple  $\langle \langle, t', tac_1 \rangle$ . The second *SubConv* uses Equation 3 and the previous triple (as well as the triple where *m* is rewritten by the identity conversion) and returns the triple  $\langle \langle, m * t', tac_2 \rangle$ . Finally, the first call to *SubConv* uses Equation 5 and returns the final triple, Equation 6.

The end result of the above rewrite is that  $n < m * t \Rightarrow n < m * t'$  may be added to the hypothesis list and this new hypothesis can be used for other rewrites or forward and backchain deductions. This example demonstrates how *SubConv* uses the lemma database to reason about inequalities. It also proves that our approach is strictly stronger than Paulson's.

<sup>4</sup> *ProgressC(c)* is a combinator that fails when *c* behaves like *IdConv*.

## 4 Conclusion

Our system is a practical solution to an important theorem proving problem. It dramatically raises the level of user/system interaction; moreover, it provides a foundation for building high-level proof procedures such as decision procedures based on equational reasoning and term normalization. Examples are given in [1,2].

One area for future research is rewrite efficiency. Our approach essentially justifies rewrites twice: once when using the lemma database to construct rewrites, and again when the tactics are executed. An alternative is to build proofs directly; however, this approach is wasteful when conversions fail and *ORELSEC* uses failure as selective composition. Another possibility is to reflect Nuprl's meta-language into the object language and prove conversions correct. Such an approach would obviate the need to construct tactics as the rewrite is formally proved valid. Howe[8] has had some success building a partial-reflection library and verifying basic rewrite strategies. Unfortunately his library is limited in scope and has its own efficiency problems which stem from inefficiencies in Nuprl's evaluator. There is a research effort at Cornell to design a reflected object language that will better support this style of rewriting.

## Acknowledgements

The author is grateful to Stuart Allen, Robert Constable, Doug Howe, and Michael Slifker for helpful discussions.

## References

- [1] David A. Basin. Equality of terms containing associative-commutative functions and commutative binding operators is isomorphism complete. Technical Report 89-1020, Cornell University, 1989.
- [2] David A. Basin and Peter Delvecchio. Verification of combinational logic in Nuprl. In *Hardware Specification, Verification and Synthesis: Mathematical Aspects*, Ithaca, New York, 1989.
- [3] R.L. Constable et al. *Implementing Mathematics with the Nuprl Proof Development System*. Prentice Hall, 1986.
- [4] Thierry Coquand and Gérard Huet. The Calculus of Constructions. *Information and Computation*, pages 95-120, 1988.
- [5] R.J. Cunningham and A.J.J Dick. Rewrite systems on a lattice of types. *Acta Informatica* 22, pages 149-169, 1985.

- [6] Michael J. Gordon, Robin Milner, and Christopher P. Wadsworth. *Edinburgh LCF: A Mechanized Logic of Computation*, volume 78 of *Lecture Notes in Computer Science*. Springer-Verlag, 1979.
- [7] Douglas J. Howe. *Automating Reasoning in an Implementation of Constructive Type Theory*. PhD thesis, Cornell University, 1988.
- [8] Douglas J. Howe. Computational metatheory in Nuprl. In *9th International Conference On Automated Deduction*, pages 238–257, Argonne, Illinois, 1988.
- [9] Douglas J. Howe and Evan Steeg. Rewrite.ml. Nuprl Internal Document, 1985.
- [10] Gérard Huet. Confluent reductions: Abstract properties and applications to term rewriting systems. *Journal of the Association for Computing Machinery*, pages 797–821, October 1980.
- [11] Gérard Huet. Deduction and computation. In *Fundamentals of Artificial Intelligence*, pages 39–74. Springer-Verlag, 1986.
- [12] Paul Jacquet. Program synthesis by completion with dependent subtypes. In *9th International Conference On Automated Deduction*, pages 550–562, Argonne, Illinois, 1988.
- [13] Per Martin-Löf. Constructive mathematics and computer programming. In *Sixth International Congress for Logic, Methodology, and Philosophy of Science*, pages 153–175, Amsterdam, 1982. North Holland.
- [14] Lawrence C. Paulson. A higher-order implementation of rewriting. *Science of Computer Programming*, 3:119–149, 1983.
- [15] G. Smolka, W. Nutt, J. Meseguer, and J.A. Goguen. Order-sorted equational computation. In *Colloquium on the Resolution of Equations in Algebraic Structures*, Lakeway, Texas, May 4-6 1987.

# SOLVING THE EULER EQUATIONS USING ADAPTIVE MESH MOTION AND REFINEMENT<sup>1</sup>

*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*

Department of Computer Science  
Rensselaer Polytechnic Institute  
Troy, NY 12180-3590

and

U.S. Army Armament, Munition, and Chemical Command  
Armament Research and Development Center  
*Close Combat Armaments Center*  
Benét Laboratory  
Watervliet, NY 12189-4050

**ABSTRACT.** We use an adaptive mesh moving and refinement finite volume method to solve the transient Euler equations of compressible flow in one and two space dimensions. Numerical solutions are generated by a MacCormack scheme with Davis's artificial viscosity model. Richardson's extrapolation is used to calculate estimates of the local discretization error which can be used to control mesh motion and refinement. Questions regarding the optimal combination of adaptive strategies and the characterization of the initial mesh are investigated. Results indicate that local mesh refinement with and without mesh moving provide dramatic improvements in accuracy over uniform mesh solutions; that mesh motion provides good results on relatively fine initial meshes; that each problem has an optimal initial mesh and that it is more efficient to begin with a coarser than optimal mesh and refine rather than starting with too fine a mesh; and that a combination of both the adaptive strategies produced the most accurate solutions.

---

<sup>1</sup> This research was partially supported by the SDIO/IST under management of the U. S. Army Research Office under Contract Number DAAL 03-86-K-0112.

**1. INTRODUCTION.** Our goal is to develop reliable, robust, and efficient software for solving hyperbolic partial differential equations. With this in mind, Arney and Flaherty [4] developed an adaptive procedure combining mesh motion and mesh refinement for solving one- and two-dimensional vector systems of time-dependent partial differential equations. The solution, mesh motion, and local refinement procedures were explicit and independent of each other; thus, modules can easily be replaced.

Arney and Flaherty's [4] method solves vector systems of hyperbolic conservation laws having the form

$$\mathbf{u}_t + \mathbf{f}_x(x, y, t, \mathbf{u}) + \mathbf{g}_y(x, y, t, \mathbf{u}) = 0, \quad (1a)$$

with initial conditions

$$\mathbf{u}(x, y, 0) = \mathbf{u}_0(x, y), \quad (1b)$$

and with appropriate well-posed boundary conditions on a one- or two-dimensional domain  $\Omega$ . Their adaptive approach consists of moving a coarse "base" mesh of quadrilateral cells to follow fronts and reduce dispersive errors. Recursive refinement of mesh cells is performed when necessary to satisfy a prescribed local error tolerance. Solutions are generated using MacCormack's [10] finite volume scheme coupled with Davis's [8] artificial viscosity model to make the scheme total variation diminishing (TVD). Local motion and refinement indicators on each cell of the mesh are used to control mesh motion and refinement, respectively. They used an estimate of the local discretization error obtained by Richardson's extrapolation [2,11] as the mesh refinement indicator. For the examples presented in this paper, we used a normalized solution gradient as the mesh movement indicator, although other choices are possible as long as the indicator is large where additional resolution is required and small where less resolution is desired. An automatic time step adjustment feature, based on maximizing the Courant stability condition, is also provided in our algorithm.

The generation of a proper initial mesh is important for the efficiency of any adaptive algorithm. Initially we create a uniform mesh on  $\Omega$  having a specified number of nodes without considering the possibility of any high-error regions. A global mesh refinement is performed on the first time step to estimate the discretization error of the initial data. The nodes of the mesh are then placed to equidistribute this error estimate. As time evolves, these nodes are dynamically moved to reduce dispersive errors.

Arney and Flaherty [4,5] perform mesh motion based on an intuitive approach by identifying computational cells having large motion indicators and clustering them into isolated regions that are presumed to contain similar solution characteristics. The center of motion indicators of each clustered region is moved so as to follow the dynamics of the solution. Remaining portions of the mesh are moved according to an algebraic function so as to produce a smooth grid having minimal distortion. Most mesh points cannot move independently but must be coupled to their immediate neighbors. The amount of movement is determined by a function which ensures that the center  $r_m(t)$  of error clusters moves according to the differential equation

$$\ddot{r}_m + \lambda \dot{r}_m = 0, \quad (2)$$

used by Coyle et al. [7]. Clustered regions created at one time step can subsequently be destroyed when a dynamic phenomena subsides. Similarly, two or more clusters can be united when structures of the solution intersect.

Results obtained by using Arney and Flaherty's [1,3,4,5] adaptive algorithm in one and two dimensions indicated that, in some instances, proper mesh motion was capable of dramatically reducing errors for a modest increase in the cost of computation. In general, however, mesh motion alone cannot produce solutions that satisfy arbitrarily prescribed accuracy requirements. They, therefore, combined mesh motion with a local temporal and spatial cellular mesh refinement strategy [4,6]. The space-time cells of a mesh that violated the prescribed error tolerance were gathered into clusters and were recursively bisected in space and time. The problem was solved locally on the successively smaller domains created by the clustering and refinement. Initial and boundary data for any refined mesh were determined by interpolation from their "parent" coarser mesh. Error tolerances involved control of the local error per unit time step and were, thus, halved at each refinement to account for the binary temporal refinement.

A dynamic tree structure, where fine grids are regarded as offspring of coarser ones, is used to manage the data associated with the motion and refinement strategies. Solutions were generated by a preorder traversal of the tree; thus, solutions on all fine meshes preceded those on coarser ones.

Our results on solving shock problems for the one- and two-dimensional Euler equations are presented in Section 2. We explore the relationship of the base mesh to the level of refinement. We found, for example, that it is more effective to begin with a coarse mesh and perform more refinement than to create a finer mesh which needs less refinement. Effective mesh motion, on the other hand, required a finer base mesh rather than a coarser one. The combination of mesh motion and refinement produced the best results. Local refinement with and without mesh moving provide substantial improvements in accuracy per unit cost relative to computations on uniform stationary mesh solutions.

**2. NUMERICAL EXPERIMENTS.** Computer codes for one- and two-dimensional problems based on Arney and Flaherty's [4] algorithm have been implemented in FORTRAN on an IBM 3090-200S computer and tested on several problems [1,4]. In this paper, we consider examples involving solutions of the Euler equations for a one-dimensional shock tube and a two-dimensional piston problem. The Euler equations for a perfect compressible fluid are studied in the conservative form

$$u_t + f_x(u) + g_y(u) = 0, \quad (3a)$$

where

$$u = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix}, \quad f(u) = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(e + p) \end{bmatrix}, \quad g(u) = \begin{bmatrix} \rho v \\ \rho vu \\ \rho v^2 + p \\ v(e + p) \end{bmatrix}. \quad (3b,c,d)$$

Here,  $\rho$  is the fluid density;  $u$  and  $v$  are the Cartesian components of the velocity vector;  $e$  is the total internal energy per unit volume; and the subscripts  $x$ ,  $y$ , and  $t$  denote partial differentiation with respect to the spatial coordinates and time, respectively. The pressure  $p$  is evaluated according to the ideal gas equation of state as

$$p = (\gamma - 1)[e - \rho(u^2 + v^2)/2], \quad (4)$$

where  $\gamma$  is the specific heat ratio of the fluid. Computational experiments were conducted with  $\gamma = 1.4$ . Solution accuracy is appraised in the  $L_1$  norm

$$\|e(\cdot, \cdot, t)\|_1 = \max_{1 \leq j \leq 4} \int_{\Omega} |e_j(x, y, t)| dx dy, \quad (5)$$

where  $e_j(x, y, t)$  is a piecewise constant approximation of  $u_j(x, y, t) - U_j$  obtained by using values at cell centers.

*Example 1.* Consider Sod's [12] one-dimensional shock tube problem which consists of solving (3,4) with  $v = 0$  and  $\partial(\cdot)/\partial y = 0$  subject to the initial conditions

$$\begin{bmatrix} \rho(x, 0) \\ p(x, 0) \\ u(x, 0) \end{bmatrix} = \begin{cases} [1.0, 1.0, 0.0]^T, & \text{if } -0.2 \leq x \leq 0.5 \\ [0.125, 0.1, 0.0]^T, & \text{if } 0.5 < x \leq 1.5 \end{cases} \quad (6)$$

A diaphragm at  $x = 0.5$  separates two regions of a tube that contain gases at different densities and pressures. The two regions are in a constant state and both fluids are initially at rest. At time  $t = 0$  the diaphragm is ruptured and three waves are generated: a shock moving with velocity 1.7522, a contact discontinuity moving with velocity 0.9275, and an expansion wave centered between  $0.5 - 1.1832t \leq x \leq 0.5 - 0.0703t$ . The exact solution [13] of this problem is

$$\begin{bmatrix} u(x, t) \\ \rho(x, t) \\ p(x, t) \end{bmatrix} = \begin{cases} [0.0, 1.0, 1.0]^T, & \text{if } \eta \leq -1.1832 \\ [0.9860 + \eta/1.2, (1 - u/5.9161)^5, \rho^{1.4}]^T, & \text{if } -1.1832 \leq \eta \leq -0.0703 \\ [0.9275, 0.4263, 0.3031]^T, & \text{if } -0.0703 \leq \eta < 0.9275 \\ [0.9275, 0.2656, 0.3031]^T, & \text{if } 0.9275 < \eta < 1.7522 \\ [0.0, 0.125, 0.1]^T, & \text{if } 1.7522 < \eta \end{cases} \quad (7)$$

where  $\eta = (x - 0.5)/t$ .

The "base" mesh is the coarsest mesh used to solve a problem. It reflects the scale on which dominant temporal and spatial changes in the solution occur. Selecting too coarse a base mesh will result in excessive refinement. Selecting too fine a base mesh will be inefficient. At present, selection of the base mesh is at the discretion of the user and in this first experiment we hope to provide guidance for this choice as well as for future automated base mesh selection procedures. Six cases having base meshes of  $N = 2^k$ ,  $k = 3, 4, \dots, 8$ , cells were solved on  $0 < t \leq 0.35$ . The maximum number of refinement levels, the initial time step, and the local discretization error were set at  $8-k$ ,  $3 \times 2^{9-k} \times 10^{-4}$ , and  $2^{5-k} \times 10^{-5}$ ,  $k = 3, 4, \dots, 8$ , respectively, so that the finest allowable discretization and local error tolerance were constant for all six cases.

$N$	Error Tolerance ( $\times 10^5$ )	Max. No. Refinement Levels	Normalized CPU Time (Effort)	No. Space-Time Cells	Effort per Unit Accuracy ( $\times 10^3$ )
8	4.0	5	1.295	28162	3.71
16	2.0	4	1.066	23026	2.17
32	1.0	3	1.000	21006	2.24
64	0.5	2	1.104	21396	2.67
128	0.25	1	1.533	25996	3.89
256	0.125	0	4.104	63744	6.70

Table 1. Normalized CPU time, number of space-time cells, and effort per unit accuracy at  $t = 0.35$  with different initial base meshes for Example 1. The parameters are adjusted so that the finest discretization and the corresponding local error tolerance are constant for all cases.

Results for the normalized CPU time, the number of space-time cells, and the effort per unit accuracy are reported in Table 1 for each of the six cases. Effort per unit accuracy is the product of the normalized CPU time and the  $L_1$  error at terminal time (0.35 in this case). In Figure 1, we show how the effort per unit accuracy varies with the logarithm of the number of cells in the base mesh. It is preferable to select a coarser base mesh than a finer one since, with our procedures, refinement of a coarse mesh will decrease the effort/accuracy ratio. The number of space-time cells vary in approximately the same ratio as the CPU time suggesting that the overhead associated with data management is minimal.

Error Tolerance ( $\times 10^5$ )	Normalized CPU Time	No. of Space-Time Cells	$\ e\ _1 (\times 10^3)$
128.0	1.000	910	25.7
32.0	4.473	7532	12.7
8.0	9.370	19322	6.20
2.0	15.610	34562	3.03

Table 2. Normalized CPU time, number of space-time cells, and global  $L_1$  error at  $t = 0.35$  as a function of the local error tolerance for Example 1 using local mesh refinement.

We continued our experiments by solving this problem on  $-0.2 \leq x \leq 1.5$  for  $0 < t \leq 0.35$  using local mesh refinement on 16-element base meshes, an initial time step



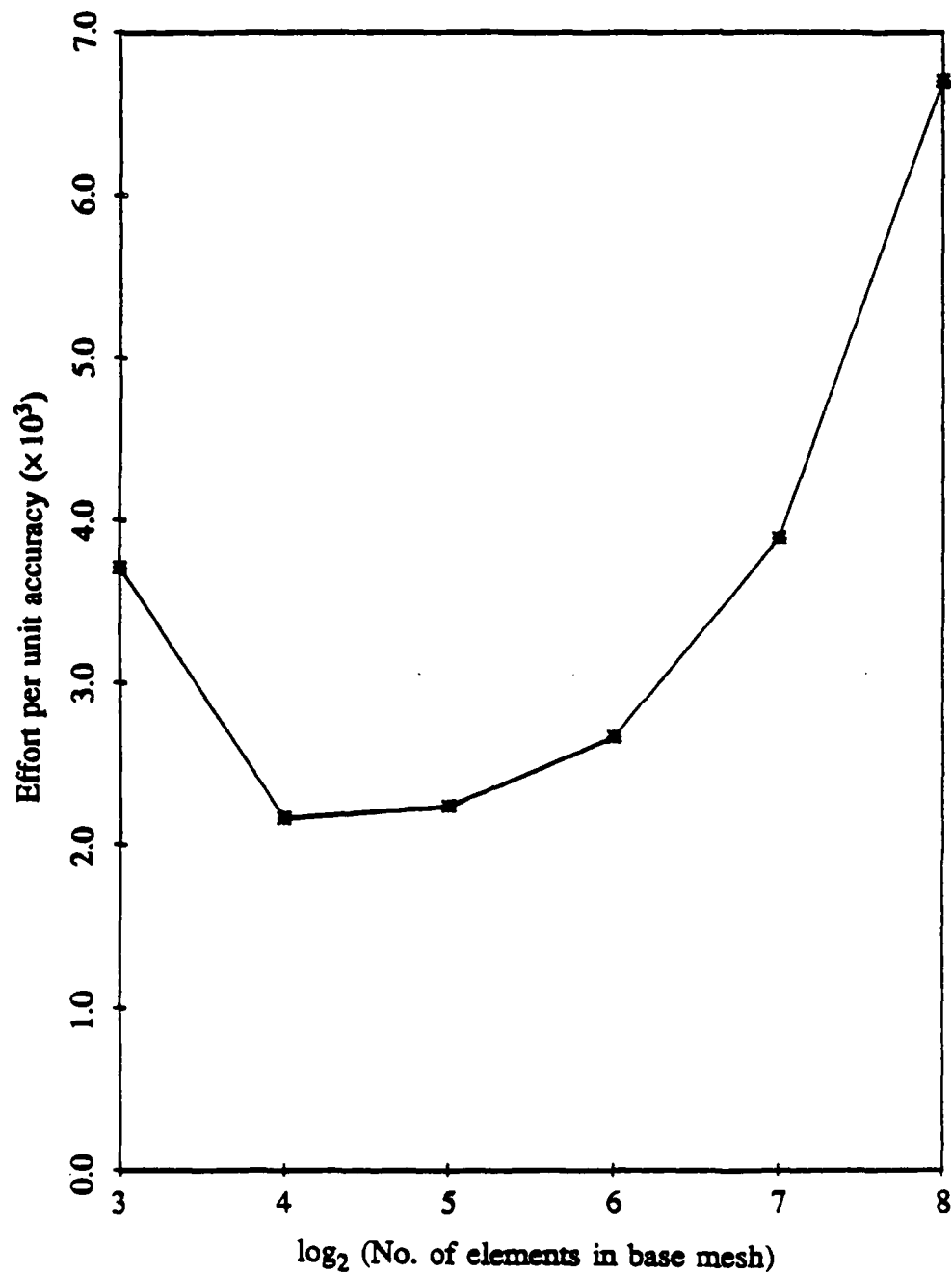


Figure 1. Effort per unit accuracy vs. number of elements in the base mesh for Example 1.

of 0.0035, and with varying error tolerances. Refinement was restricted to a maximum of four levels to avoid excessive refinement near shocks. The normalized CPU time, the number of space-time cells used to solve the problem, and the errors in  $L_1$  at  $t = 0.35$  are presented in Table 2 as functions of the local discretization error tolerance. For small tolerances, the CPU times and the number of space-time cells increase at approximately

the same rate as the  $L_1$  error decreases, again indicating a minimal overhead associated with refinement. The decrease in the local pointwise error tolerance is quadratic when compared with the actual global  $L_1$  error, which is what one would expect for problems having smooth solutions. The result apparently carries over to this shock problem.

Strategy	Normalized CPU Time	No. of Space-Time Cells	$\ e\ _1 (\times 10^3)$
Uniform Mesh	1.000	576	30.7
Coarse Mesh Motion	2.026	1152	16.9
Refinement	19.009	34562	3.03
Motion & Refinement	26.532	44602	1.88
Fine Mesh Motion	8.584	12690	4.37

Table 3. Normalized CPU time, number of space-time cells, and global  $L_1$  error at  $t = 0.35$  for adaptive and standard solutions of Example 1.

The third experiment involves comparing adaptive solutions obtained using mesh motion, local mesh refinement, and mesh motion plus local refinement with one obtained on a uniform mesh. In each case, a 16-element base mesh and an initial time step of 0.0035 was selected. An error tolerance of 0.00002 was used for those solutions that involved refinement. A fifth solution involving motion of a finer 50-element mesh was also generated. Data similar to that presented in Table 2 is displayed in Table 3 comparing the results of different adaptive strategies with those on a stationary uniform mesh. In Figures 2 to 6 we display the calculated density as a function of  $x$  at  $t = 0, 0.09, 0.18, 0.27,$  and  $0.35$ , the meshes used, and the time steps selected for each of the solutions shown in Table 3. The uniform mesh solution shown in Figure 2 exhibits excessive diffusion at the shock, at the contact surface, and in the expansion region. However, the time step increases rapidly in accordance with the Courant condition. A larger initial time step could clearly have been used; however, we wanted to use the same initial time step for all the cases. In Figure 3 we show that the moving mesh procedure follows the dominant features of the solution. Results are clearly superior to those in Figure 2, but the mesh is too coarse to obtain good resolution everywhere. The results in Figure 6 demonstrate that far better resolution is obtained when a finer mesh consisting of 50 elements is used; however, this mesh did not move correctly in the expansion region because the mesh movement indicator is too small there. The initial mesh generator distributes a specified number of nodes  $N$  based on the initial data. In this case, the initial data has a jump discontinuity at  $x = 0.5$ , so nodes were clustered around that point and then gradually spread across the domain. There are too many nodes in the expansion region in relation to the small magnitude of the movement indicator to produce adequate motion there. A static rezone of the mesh could alleviate this problem. The time steps of both solutions with mesh moving (Figures 3 and 6) are erratic for small times while the mesh is

adjusting itself to the three breaking waves. Time steps increase at the same rate as those for the uniform mesh solution of Figure 2 when a coarser mesh is used. Incorrect motion of the fine mesh in the expansion region (Figure 6) prevented a similar increase of the time step. The results depicted in Figure 4 show that refinement was correctly performed at all critical points of the calculation. In each case, shocks are captured sharply with the correct speed. As expected with Davis's [8] artificial viscosity model, diffusive effects are more pronounced near the contact surface than at the shock. Results obtained using both mesh motion and refinement are depicted in Figure 5. The results have improved somewhat but at the cost of a significantly higher computational effort relative to the solution of Figure 4. This suggests that mesh motion, with or without refinement, is not competitive with refinement alone. Additional experimentation is needed to determine a better combination of mesh moving and refinement.

*Example 2.* Consider the solution of the Euler equations (3,4) in a region exterior to an infinite cylindrical piston that is expanding radially creating a radially expanding shock wave. We ignore the cylindrical symmetry and solve this problem in one quadrant of the two-dimensional rectangular domain  $-0.05 \leq x, y \leq 0.05$  with the two-dimensional algorithm of Arney and Flaherty [4]. Self-similar solutions of this test problem are obtained by solving a pair of ordinary differential equations (by numerical integration) for the radial velocity and acoustic speed [9].

We solved this problem for  $0 < t \leq 0.0096$  with the piston initially positioned at a radius of 0.016023 and having a velocity of 1.6185. Numerical solutions were calculated on a  $26 \times 26$  spatial mesh (i) without adaptation, (ii) with one level of local refinement, and (iii) with mesh motion and one level of refinement. Contours of the density at  $t = 0.0096$  are presented for the exact and three numerical solutions in Figure 7. The spatial meshes produced by the two adaptive strategies at  $t = 0.0096$  are shown in Figure 8.

Clearly one level of refinement is not sufficient to adequately resolve the structure of this solution. We were forced to limit our computations to this level because of memory restrictions on our computing system. Nevertheless, local refinement with and without mesh moving provide improvements in accuracy over uniform stationary mesh solutions. Detailed quantitative comparisons have yet to be performed; however, qualitatively, the expanding shock is sharper in both adaptive solutions. The combination of mesh motion and refinement provides additional improvement.

**3. CONCLUSIONS.** We have applied an adaptive mesh motion and refinement method for time-dependent partial differential equations to the one- and two-dimensional Euler equations. Our method can be used with several numerical methods and local error indicators to produce solutions that satisfy prescribed local tolerances. Mesh motion is global and is performed at every time step. Mesh refinement is cellular and can be used on irregular or moving meshes of quadrilateral cells.

Our results indicate that mesh refinement can be used to achieve prescribed levels of accuracy. Refinement is easy, recursive, and works well. It appears to be computationally efficient for a given accuracy level. Proper mesh movement improved the computed results. Refinement has a definite advantage over mesh motion in that it is inferred in an a posteriori manner from a preliminary solution whereas our mesh motion is applied in an

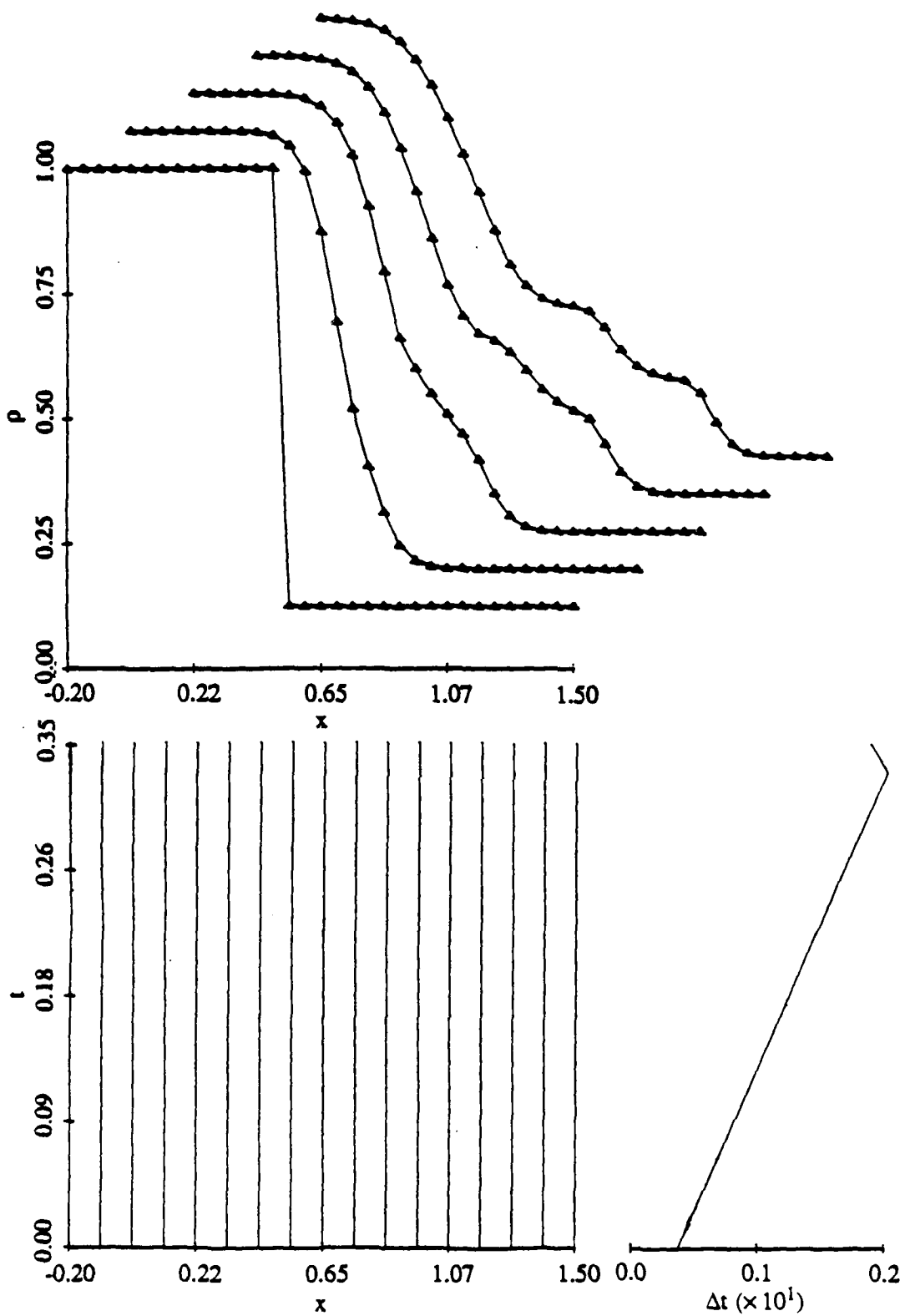


Figure 2. Solutions, mesh trajectories, and time step profile for computations performed with a stationary uniform mesh of 16 cells for Example 1.

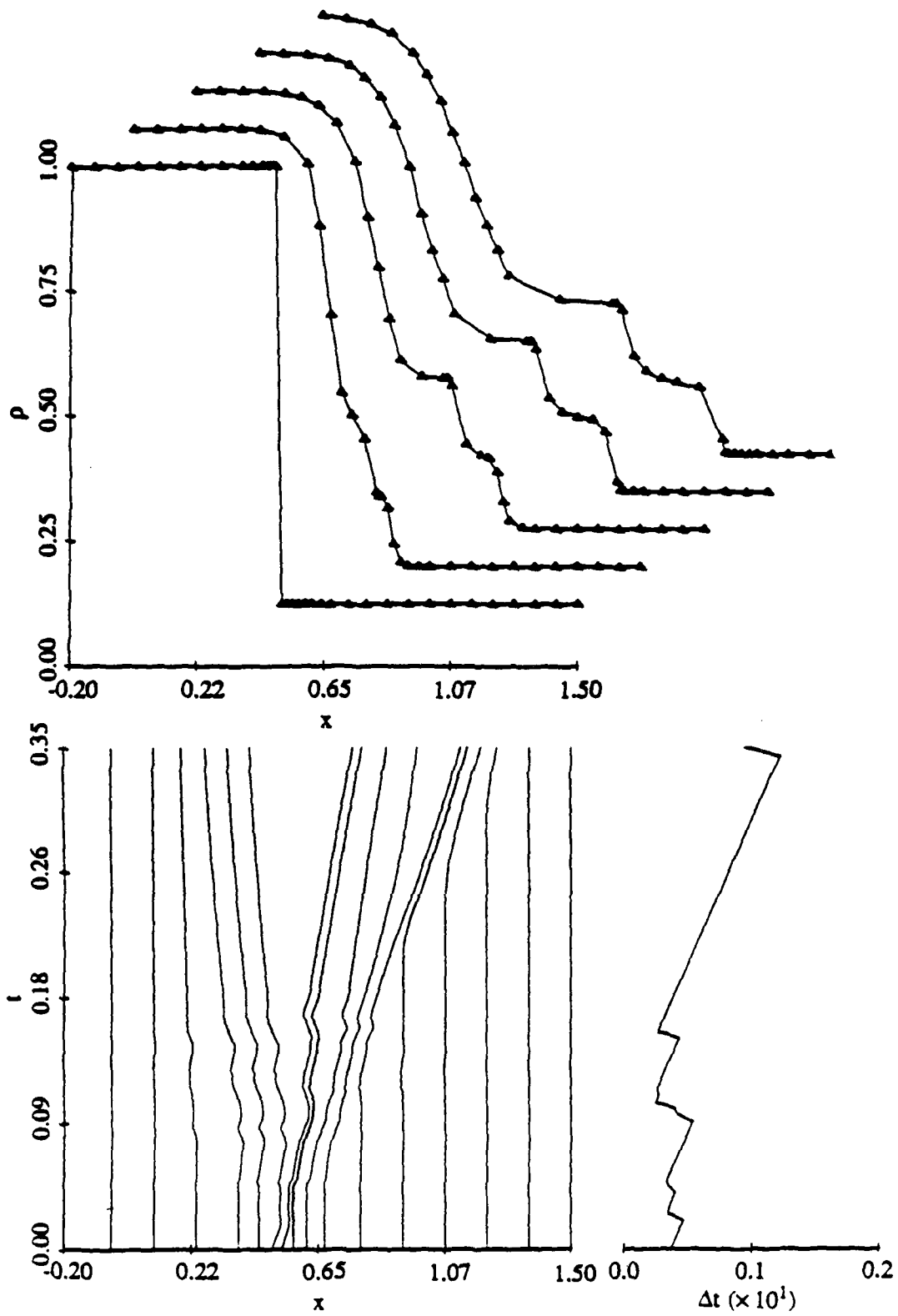


Figure 3. Solutions, mesh trajectories, and time step profile for computations performed with adaptive mesh motion on a mesh of 16 cells for Example 1.

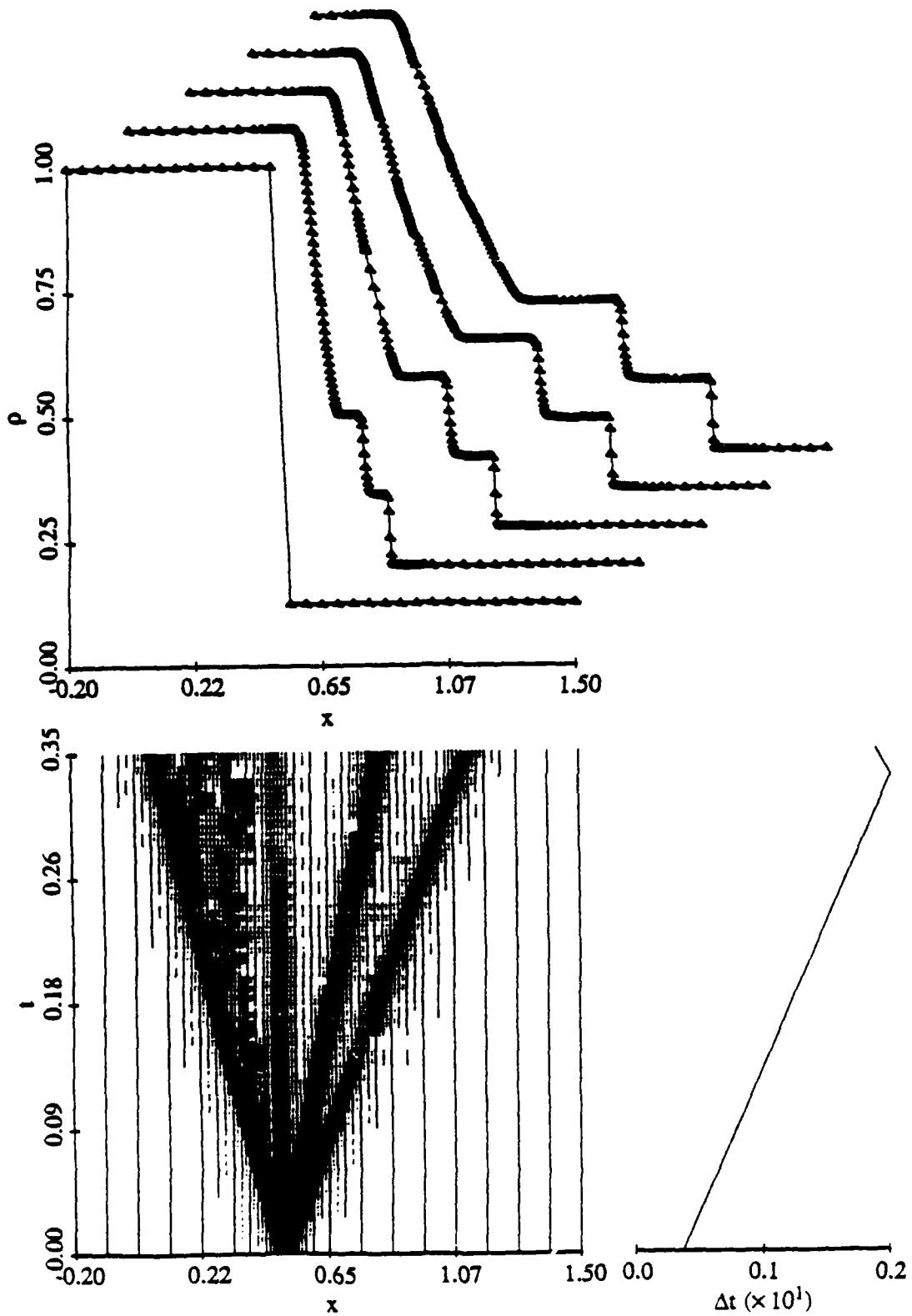


Figure 4. Solutions, mesh trajectories, and time step profile for computations performed with adaptive local mesh refinement for Example 1.

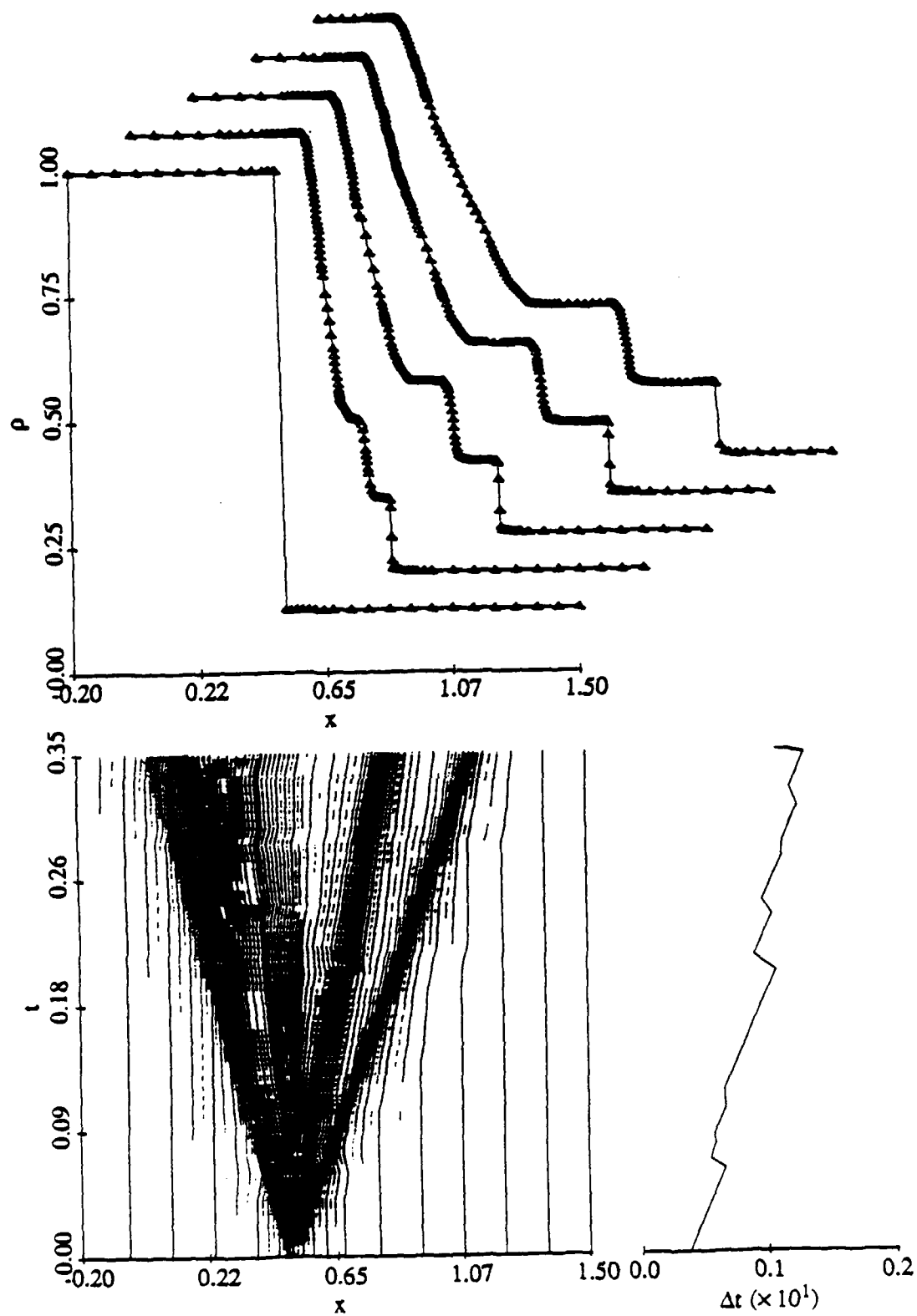


Figure 5. Solutions, mesh trajectories, and time step profile for computations performed with both adaptive mesh motion and local mesh refinement for Example 1.

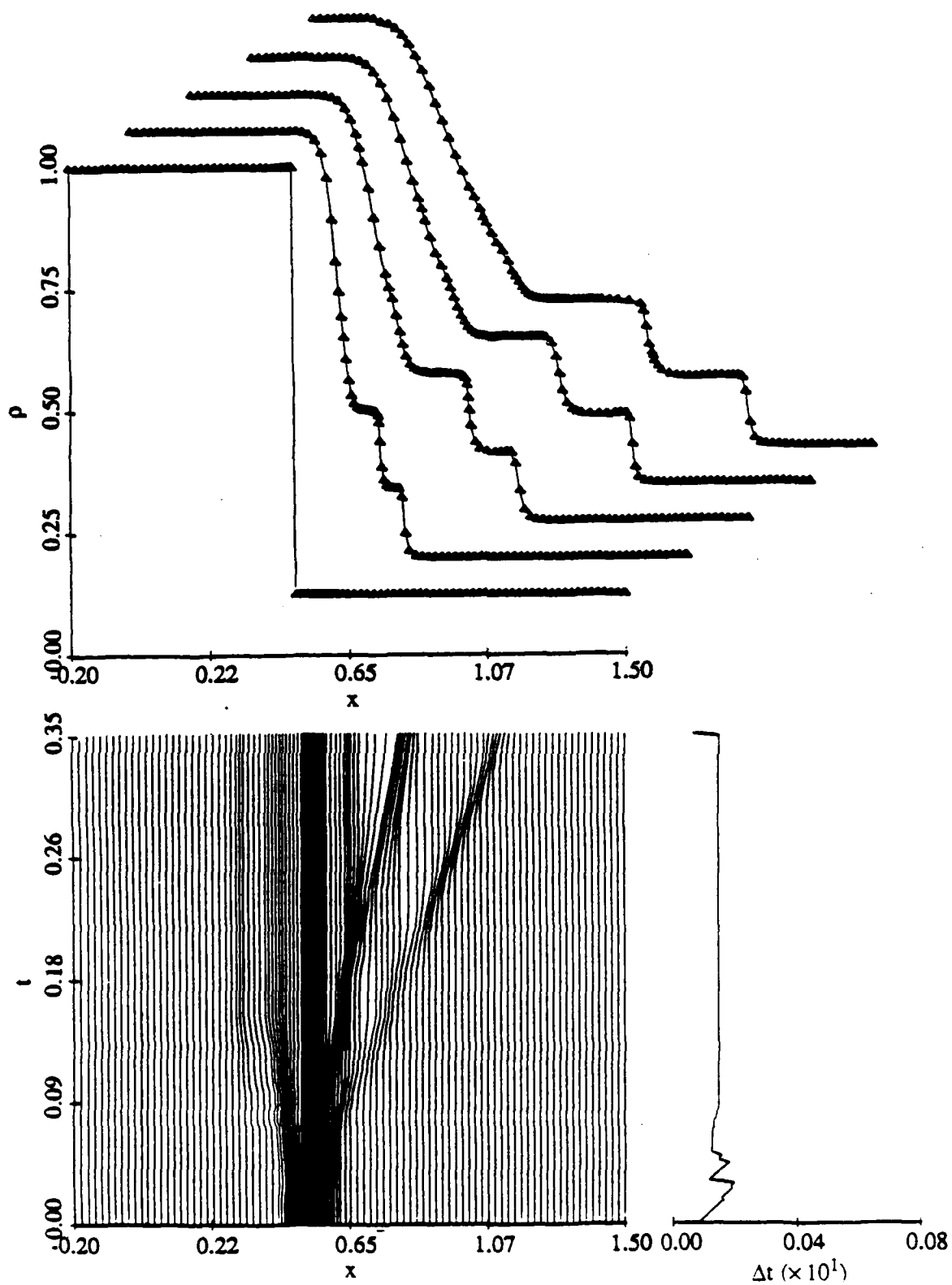


Figure 6. Solutions, mesh trajectories, and time step profile for computations performed with adaptive mesh motion on a mesh of 50 cells for Example 1.



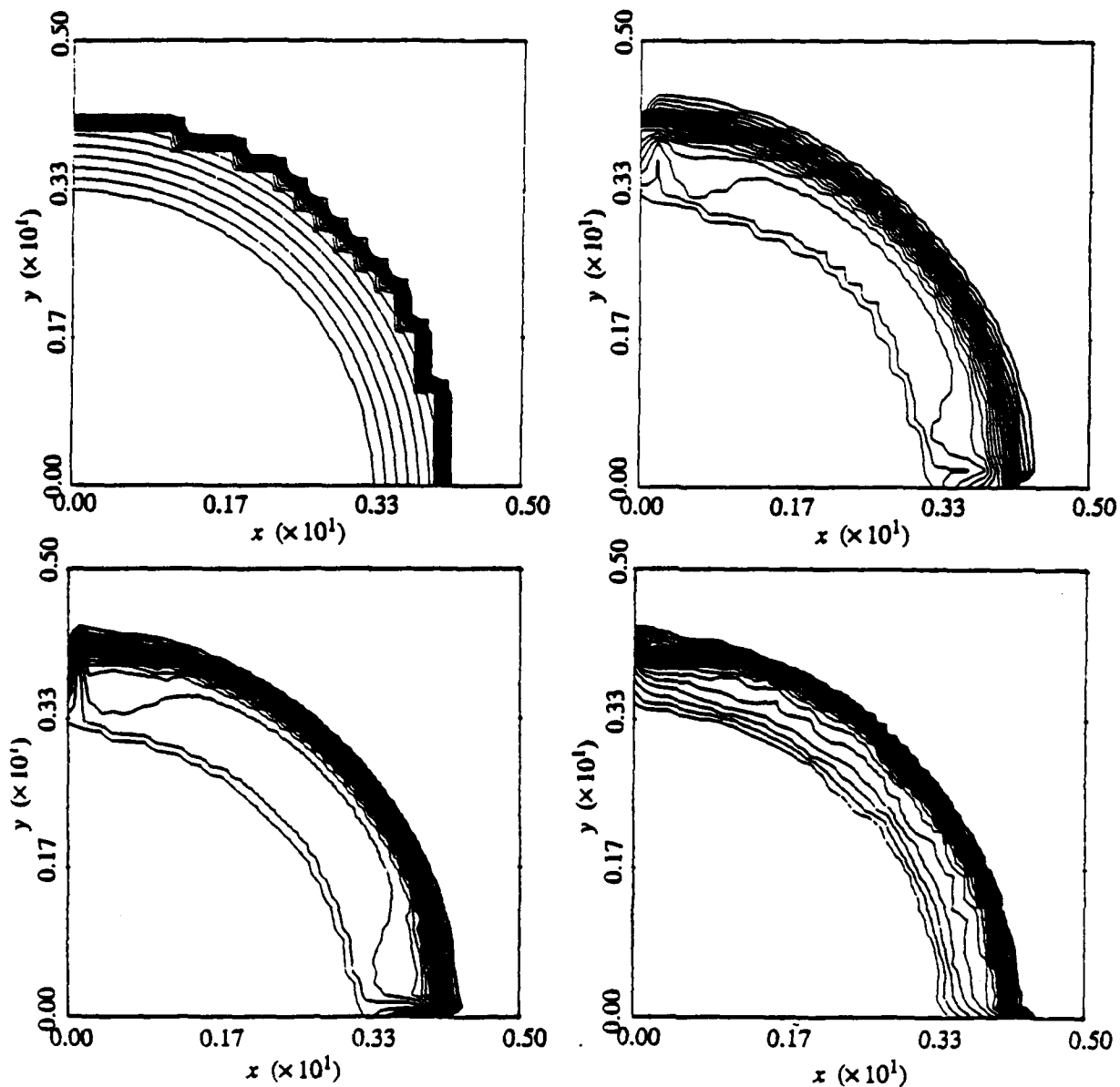


Figure 7. Density contours for Example 2 at  $t = 0.0096$  obtained from the exact solution (upper left) and by computed solutions on a uniform stationary mesh (upper right), a uniform stationary base mesh with one level of refinement (lower left), and a moving base mesh with one level of refinement (lower right).

a priori fashion by extrapolating the mesh behavior of the previous two base time steps. As a result, mesh refinement may be inefficient but it never leads to anomolous behavior. On the other hand, incorrect mesh motion can easily mess a local nonuniformity in the solution that evolves suddenly. Such incorrect motion restricts the size of the time steps and diminishes the overall efficiency of the adaptive method. These difficulties can largely be overcome by combining mesh motion with mesh refinement and static mesh redistribution. Further experimentation and analysis are needed in order to determine

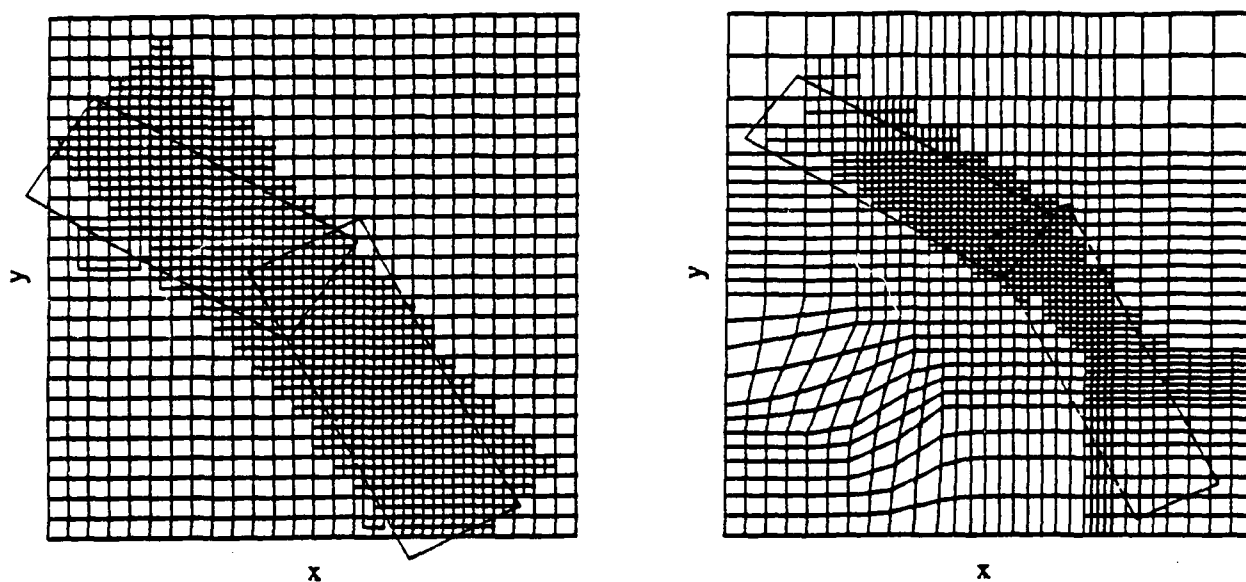


Figure 8. Spatial meshes at  $t = 0.0096$  for Example 2 using one level of local mesh refinement on a uniform stationary base mesh (left) and a moving base mesh (right).

optimal combinations of these strategies.

We used the first example to demonstrate that each problem has an optimal initial base mesh size and that it is always computationally efficient to adaptively refine beginning with a less than optimal mesh rather than starting with too fine a mesh. This example also showed that for mesh motion to be effective, a fine base mesh is absolutely necessary. A combination of both the adaptive strategies of mesh motion and refinement produced the best results but at the cost of a significantly higher computational effort. The second example demonstrates that our adaptive mesh procedures extend to two-dimensional problems.

We are currently developing higher-order explicit finite volume methods to replace the second-order MacCormack scheme. The present Richardson's extrapolation-based error indicator is expensive and we are seeking ways of replacing it by using p-refinement techniques. Such methods have been shown to have an excellent cost performance ratio when used in conjunction with finite element methods. We are also working on a modification of our algorithm which allows a variety of geometries. Our adaptive techniques must be able to take advantage of the latest advances in vector and parallel computing hardware. The tree is a highly parallel structure and we are developing solution procedures that exploit this in a variety of shared and distributed memory parallel computing environments; however, it is difficult to parallelize mesh motion because of its global nature. Cells assigned to a particular processor may migrate to the domain of other neighboring processors and cause non-trivial bookkeeping problems. Mesh motion is also difficult to

perform in higher dimensions. We are, therefore, actively considering hp-type techniques in parallel environments.

## REFERENCES.

1. Arney, D.C., Biswas, R., and Flaherty, J.E., "Adaptive Mesh Experiments for Hyperbolic Partial Differential Equations," *Trans. Sixth Army Conf. Appl. Math. and Comput.*, (1989), pp. 1051-1074.
2. 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. Appl. Math. and Comput.*, (1988), pp. 437-458.
3. Arney, D.C., Carofano, G., and Misner, E., "An Adaptive Mesh Method for Solving Blast Problems Using the Eule. Equations," *Trans. Sixth Army Conf. Appl. Math. and Comput.*, (1989), pp. 1115-1132.
4. Arney, D.C. and Flaherty, J.E., "An Adaptive Mesh Moving and Local Refinement Method for Time-Dependent Partial Differential Equations," to appear in *Trans. Math. Software*.
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 Local Mesh Refinement Method for Time Dependent Partial Differential Equations," *Appl. Num. Math.*, 5 (1989), pp. 257-274.
7. Coyle, J.M., Flaherty, J.E., and Ludwig, R., "On the Stability of Mesh Equidistribution Strategies for Time-Dependent Partial Differential Equations," *J. Comput. Phys.*, 62 (1986), pp. 26-39.
8. Davis, S., "A Simplified TVD Finite Difference Scheme via Artificial Viscosity," *SIAM J. Sci. Stat. Comput.*, 8 (1987), pp. 1-18.
9. 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.
10. MacCormack, R.W., "The Effect of Viscosity in Hypervelocity Impact Cratering," *AIAA Paper 69-354*, (1969).
11. Richardson, L.F., "The Deferred Approach to the Limit, I. Single Lattice," *Trans. Roy. Soc. London*, 226 (1927), pp. 299-349.
12. Sod, G., "A Survey of Several Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws," *J. Comp. Phys.*, 27 (1978), pp. 1-31.
13. Whitham, G.B., *Linear and Nonlinear Waves*, Wiley-Interscience, New York, 1974.

# Line Iterative Methods for Cyclically Reduced Non-Self-Adjoint Elliptic Problems\*

Howard C. Elman

Department of Computer Science and  
Institute for Advanced Computer Studies  
University of Maryland  
College Park, MD 20742

Gene H. Golub

Department of Computer Science  
Stanford University  
Stanford, CA 94305

## Abstract

We present the results of a study of line iterative methods for solving linear systems arising from finite difference discretizations of non-self-adjoint elliptic partial differential equations on two-dimensional domains. The methods consist of performing one step of cyclic reduction, followed by solution of the resulting reduced system by line relaxation. We consider both one-line and two-line relaxation methods, and we present analytic and experimental results showing that these classes of methods are highly effective for solving the convection-diffusion equation. The paper summarizes results from [2] and [3], where further details can be found.

## 1. Introduction.

We consider iterative methods for solving linear systems of the type that arise from two-cyclic discretizations of two-dimensional elliptic partial differential equations. Such systems can be ordered using "black ordering" so that they have the form

$$(1.1) \quad \begin{pmatrix} D & C \\ E & F \end{pmatrix} \begin{pmatrix} u^{(r)} \\ u^{(b)} \end{pmatrix} = \begin{pmatrix} v^{(r)} \\ v^{(b)} \end{pmatrix}$$

where  $D$  and  $F$  are diagonal matrices. If block elimination is used to decouple the "red" points  $u^{(r)}$  from the "black" points  $u^{(b)}$ , the result is a *reduced system*

$$(1.2) \quad [F - ED^{-1}C]u^{(b)} = v^{(b)} - ED^{-1}v^{(r)}.$$

Let

$$(1.3) \quad S = F - ED^{-1}C, \quad s = v^{(b)} - ED^{-1}v^{(r)}.$$

In this paper, we describe a study of relaxation methods for solving (1.2) when (1.1) comes from a finite-difference discretization of the constant coefficient convection-diffusion equation

$$(1.4) \quad Au = -\Delta u + \sigma u_x + \tau u_y = f$$

---

\*Supported by the U. S. Army Research Office.

with Dirichlet boundary conditions. We consider several orderings of the rows and columns of  $S$ , based on either a *one-line* ordering or a *two-line* ordering of the reduced grid. Line methods of these types have been considered for solving the original problem (1.1) in [1], [8], and for the reduced system in [4], [7], [8]. For the ordering strategies considered, the reduced matrices have block Property A so that Young's analysis of iterative methods [11] is applicable. We use this analysis to determine the convergence properties of block Jacobi, Gauss-Seidel and successive overrelaxation (SOR) methods for solving the discrete convection-diffusion equation, in terms of discrete cell Reynolds numbers  $\sigma h/2$  and  $\tau h/2$ . In addition, we present the results of numerical experiments showing convergence behavior not revealed by the analysis. Together, the analytic and numerical results show that the two types of orderings lead to very effective methods for solving (1.4).

An outline of the paper is as follows. In §2, we describe two discretization schemes for (1.4), and we examine the truncation error associated with taking the reduced system as an approximation of (1.4). In §3, we present the one-line and two-line orderings for the unknowns of (1.2), including variants based on block red-black groupings of unknowns, and we outline the convergence analysis for line relaxation methods. In §4, we describe some numerical experiments that confirm and supplement the convergence results.

## 2. The convection-diffusion equation and the reduced system.

Consider the two-dimensional convection-diffusion equation (1.4), posed on the unit square  $\Omega \in (0, 1) \times (0, 1)$  with Dirichlet boundary conditions  $u = g$  on  $\partial\Omega$ . Discretization by a five-point finite difference operator leads to a linear system

$$Au = v$$

where  $u$  now denotes a vector in a finite dimensional space. We discretize on a uniform  $n \times n$  grid using standard second order differences for the Laplacian [10], [11], and either centered or upwind differences for the first derivatives. With  $u$  ordered lexicographically in the natural ordering as  $(u_{1,1}, u_{2,1}, \dots, u_{n,n})^T$ , the coefficient matrix has the form

$$(2.1) \quad A = \text{tri}[A_{j,j-1}, A_{jj}, A_{j,j+1}].$$

Here,  $\text{tri}[X_{j,j-1}, X_{jj}, X_{j,j+1}]$  is the (block) tridiagonal matrix whose  $j$ 'th row contains  $X_{j,j-1}$ ,  $X_{jj}$  and  $X_{j,j+1}$  on its subdiagonal, diagonal and superdiagonal, respectively. We omit the subscripts when there is no ambiguity. The entries of (2.1) are

$$A_{j,j-1} = bI, \quad A_{jj} = \text{tri}[c, a, d], \quad A_{j,j+1} = eI.$$

where  $I$  is the identity matrix,  $a, b, c, d$  and  $e$  depend on the discretization, and all blocks are of order  $n$ . Let  $h = 1/(n + 1)$ . After scaling by  $h^2$ , the matrix entries are given by

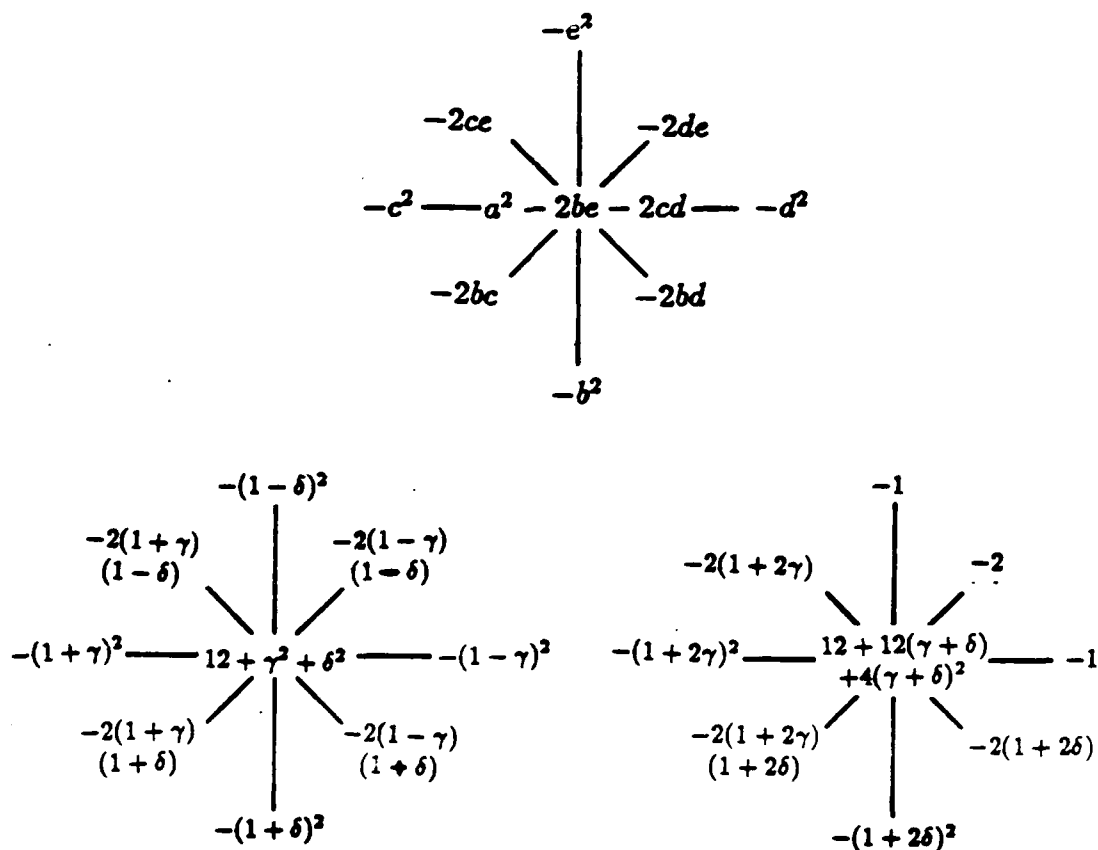
$$\begin{aligned} a &= 4, & b &= -(1 + \delta), & c &= -(1 + \gamma), \\ d &= -(1 - \gamma), & e &= -(1 - \delta), \end{aligned}$$

for the centered difference scheme, where  $\gamma = \sigma h/2$  and  $\delta = \tau h/2$ ; and

$$a = 4 + 2(\gamma + \delta), \quad b = -(1 + 2\delta), \quad c = -(1 + 2\gamma), \\ d = -1, \quad e = -1,$$

for the upwind scheme. At the  $(i, j)$  grid point, the right hand side satisfies  $v_{ij} = h^2 f_{ij}$  where  $f_{ij} \equiv f(ih, jh)$ .

In [2], we showed that the reduced matrix  $S$  is a *skewed* nine-point operator. At all grid points except those bordering  $\partial\Omega$ , the computational molecule has the form (after scaling by  $a$ ) given in Fig. 2.1.



**Fig. 2.1:** Computational molecules for the reduced system. Top: general case. Bottom left: centered differences. Bottom right: upwind differences.

Suppose centered differences are used to discretize the first derivative terms. At the  $(i, j)$  grid point, the discrete operator satisfies  $\frac{1}{h^2}[Au]_{ij} = [Au]_{ij} + O(h^2)$ , i.e. the truncation error of the discretization is of order  $h^2$ . The following result shows that the reduced system (1.2) can also be viewed as a discretization of (1.4) with truncation error of order  $h^2$ . The proof is based on Taylor series, see [3]. A similar analysis shows that the reduced system for the upwind scheme approximates (1.4) with truncation error  $O(h)$ .

**THEOREM 1.** For the centered difference discretization of (1.4), let  $\tilde{S}$  and  $\tilde{s}$  denote the reduced matrix and right hand side obtained by multiplying the reduced system by  $a$  ( $= 4$ ).

Then for  $2 \leq i, j \leq n - 1$ ,  $\tilde{S}$  satisfies

$$\frac{1}{8h^2}[\tilde{S}u]_{ij} = -\left[\left(1 + \frac{\sigma h^2}{8}\right)u_{xx} + \left(1 + \frac{\tau h^2}{8}\right)u_{yy}\right] + \sigma u_x + \tau u_y + O(h^2),$$

and  $\tilde{s}$  satisfies

$$\frac{1}{8h^2}\tilde{s}_{ij} = f_{ij} + O(h^2).$$

### 3. Convergence of line relaxation relaxation methods.

The performance of iterative methods for solving (1.2) depends on the ordering of the underlying grid. In this section, we define the one-line orderings and two-line orderings, and outline the convergence analysis of the resulting iterative methods.

For the one-line orderings, grid points are grouped by diagonal lines oriented at a  $45^\circ$  angle with the horizontal and vertical axes. For the purpose of discussion, we fix the orientation to be along the NW—SE direction. In the *natural* one-line ordering, the  $n - 1$  diagonal lines are numbered starting from one corner (e.g. the SW) from 1 to  $n - 1$ , and individual points are numbered from bottom to top along the lines. An example for  $n = 7$  is shown in the left side of Fig. 3.1, where the line indices are shown outside  $\partial\Omega$ . In the *red-black* variant, the lines with odd indices from the natural ordering are ordered first, followed by those with even indices. The individual grid points are renumbered to be consistent with this reordering. An example for  $n = 7$  is shown in the right side of Fig. 3.1.

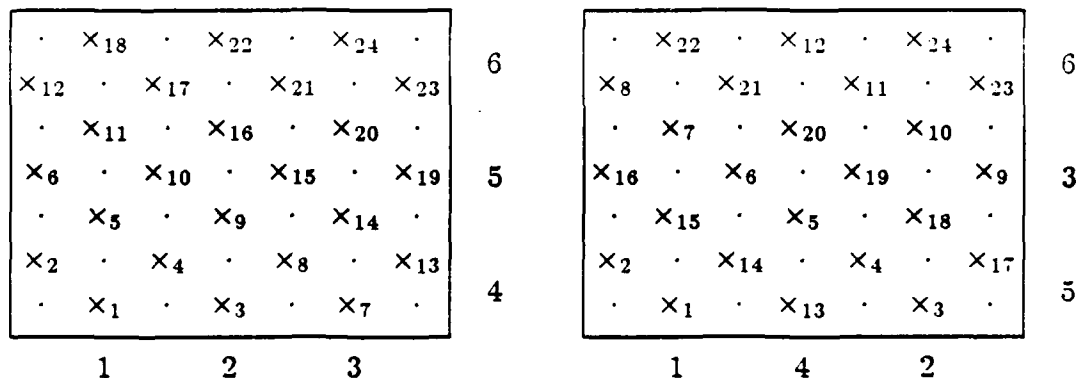


Fig. 3.1: The reduced grid derived from a  $7 \times 7$  grid, with natural one-line (left) and red-black one-line (right) orderings.

In the two-line orderings, points in the reduced grid are grouped by pairs of horizontal or vertical lines. Examples with horizontal lines, for  $n = 6$ , are shown in Fig. 3.2. The left side of the figure shows a *natural* two-line ordering, and the right side shows a *red-black*

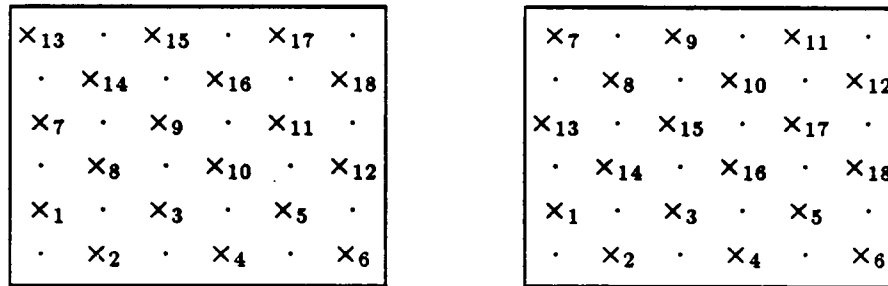


Fig. 3.2: The reduced grid derived from a  $6 \times 6$  grid, with natural two-line (left) and red-black two-line (right) orderings.

two-line ordering. We restrict our attention to two-line orderings with horizontal lines; generalization to vertical lines is straightforward.

For all these orderings, we use a splitting of the reduced matrix,

$$S = D - C,$$

where  $D$  is a block diagonal matrix whose individual blocks come from the underlying lines. Thus,  $D$  contains tridiagonal blocks for the one-line orderings and pentadiagonal blocks for the two-line orderings. Consider the block Jacobi iterative method

$$u_{k+1}^{(b)} = B u_k^{(b)} + D^{-1} s,$$

where  $B = D^{-1}C$  is the block Jacobi iteration matrix. The standard measure of the effectiveness of this method is the spectral radius  $\rho(B)$ ; the iteration is *convergent* provided  $\rho(B) < 1$ , and convergence tends to be more rapid if  $\rho(B)$  is closer to 0 [10]. The following results determine bounds on  $\rho(B)$  for both the centered difference and upwind difference schemes, and each of the four orderings defined above, see [2], [3] for proofs.

**THEOREM 2.** *For the centered difference scheme, if  $|\gamma| < 1$  and  $|\delta| < 1$ , then the spectral radii of the one-line block Jacobi iteration matrices for the reduced system are bounded by*

$$\frac{(\sqrt{1 - \gamma^2} + \sqrt{1 - \delta^2})^2}{8 - (\sqrt{1 - \gamma^2} + \sqrt{1 - \delta^2})^2 + 2\sqrt{(1 - \gamma^2)(1 - \delta^2)}(1 - \cos(\pi h))}.$$

*For the upwind difference scheme, the spectral radii are bounded by*

$$\frac{(\sqrt{1 + 2\gamma} + \sqrt{1 + 2\delta})^2}{2(2 + \gamma + \delta)^2 - (\sqrt{1 + 2\gamma} + \sqrt{1 + 2\delta})^2 + 2\sqrt{(1 + 2\gamma)(1 + 2\delta)}(1 - \cos(\pi h))}.$$

**THEOREM 3.** *For the centered difference scheme, if  $|\gamma| < 1$  and  $|\delta| < 1$ , then the spectral radii of the two-line block Jacobi iteration matrices for the reduced system are bounded by*

$$\frac{(1 - \delta^2) \cos 2\pi h + 2\sqrt{(1 - \gamma^2)(1 - \delta^2)} \cos \pi h}{3 - (\sqrt{1 - \gamma^2} + \sqrt{1 - \delta^2})^2 - (1 - \gamma^2) + 2\sqrt{(1 - \gamma^2)(1 - \delta^2)}(1 - \cos \pi h) + 2(1 - \gamma^2)(1 - \cos^2 \pi h)} + o(h^2).$$



For the upwind difference scheme, the spectral radii are bounded by

$$\frac{(1 + 2\delta) \cos 2\pi h + 2\sqrt{(1 + 2\gamma)(1 + 2\delta)} \cos \pi h}{2(2 + \gamma + \delta)^2 - (\sqrt{1 + 2\gamma} + \sqrt{1 + 2\delta})^2 - (1 + 2\gamma) + 2\sqrt{(1 + 2\gamma)(1 + 2\delta)}(1 - \cos \pi h) + 2(1 + 2\gamma)(1 - \cos^2 \pi h)} + o(h^2).$$

Note that these results apply for both the natural and red-black variants of the line orderings. The bounds of Theorem 3 are smaller than those of Theorem 2. For the centered difference schemes, the restrictions on  $\gamma$  and  $\delta$  coincide with the conditions guaranteeing that the discrete solution is nonoscillatory. Some bounds applicable when both  $|\gamma| > 1$  and  $|\delta| > 1$  can be found in [2], [3].

For all orderings considered, the reduced matrices  $S$  have block Property A, so that Young's analysis of relaxation methods applies. In particular, let  $C = L + U$  where  $L$  and  $U$  are strictly lower triangular and upper triangular, respectively, and let  $\mathcal{L}_\omega = (D - \omega L)^{-1}[(1 - \omega)D + \omega U]$  denote the block SOR iteration matrix. Then  $\rho(\mathcal{L}_1) = \rho(B)^2$ , and for all of the cases handled by Theorems 2 and 3, the choice

$$(3.1) \quad \omega^* = \frac{2}{1 + \sqrt{1 - \rho(B)^2}}$$

minimizes  $\rho(\mathcal{L}_\omega)$  with respect to  $\omega$ , with  $\rho(\mathcal{L}_{\omega^*}) = \omega^* - 1$ .

#### 4. Numerical experiments.

In this section, we present the results of numerical experiments that confirm and supplement the convergence analysis. We compare the bounds on spectral radii of iteration matrices with computed spectral radii, and we examine the performance of the Gauss-Seidel and SOR methods for solving the reduced system arising from the centered difference discretization of the convection-diffusion equation. All computations were performed on a VAX-3600 in double precision Fortran. The reduced matrices were computed using PCGPAK [9]. All spectral radii were computed using the QZ algorithm in EISPACK [5], [6].

Table 4.1 shows the computed spectral radii of the one-line Gauss-Seidel iteration matrices for  $h = 1/32$  and various choices of the parameters  $\gamma$  and  $\delta$ . Table 4.2 shows analogous data for the two-line Gauss-Seidel iteration matrices. For the one-line ordering, the results under the heading  $\delta = 0$  in Table 4.1 are identical to those occurring when  $\gamma = 0$  and  $\delta$  has the values in the first column of the table. It is evident from the tables that the analytic bounds are very close to the computed values (except for the case of large  $\gamma = \delta$ , where the analytic bounds come from [2], [3]), and that the computed spectral radii are considerably smaller than one.

Figs. 4.1 - 4.3 summarize the performance of the block iterative methods for solving various examples of the discrete convection-diffusion equation (1.4) with Dirichlet boundary conditions. In all cases, centered differences were used to discretize the first derivative

	$\delta = 0$		$\gamma = \delta$	
$\gamma$	Computed	Bound	Computed	Bound
.2	.89	.92	.85	.85
.4	.69	.72	.52	.52
.6	.45	.46	.22	.22
.8	.21	.22	.05	.05
1.0	.05 <sup>1</sup>	0	0	0
1.2	.04	—	.03	.05
1.4	.06	—	.10	.23
1.6	.08	—	.19	.61
1.8	.11	—	.27	1.25
2.0	.15	—	.35	2.25

Table 4.1: Spectral radii and bounds for the one-line Gauss-Seidel iteration matrices, centered differences,  $h = 1/32$ .

	$\gamma = \xi, \delta = 0$		$\gamma = 0, \delta = \xi$		$\gamma = \delta = \xi$	
$\xi$	Computed	Bound	Computed	Bound	Computed	Bound
.2	.86	.90	.85	.90	.77	.81
.4	.63	.66	.62	.65	.42	.44
.6	.38	.40	.34	.36	.16	.16
.8	.18	.19	.12	.12	.03	.03
1.0	.06 <sup>2</sup>	.02	0	0	0	0
1.2	.04	—	.04	—	.02	.03
1.4	.06	—	.09	—	.05	.13
1.6	.07	—	.13	—	.09	.34
1.8	.07	—	.18	—	.12	.71
2.0	.07	—	.22	—	.16	1.27

Table 4.2: Spectral radii and bounds for the two-line Gauss-Seidel iteration matrices, centered differences,  $h = 1/32$ .

terms, and the mesh size was  $h = 1/32$ , so that the order of the linear system was  $N = 961$ . The curves in the figures represent the average iteration counts for three test problems, determined by three initial guesses with random values in the interval  $[-1, 1]$ . In all cases, the right hand side  $s$  was identically zero. The convergence criterion was  $\|r_i\|_2 / \|r_0\|_2 \leq 10^{-6}$ , where  $r_i = s - Su_i^{(b)} = -Su_i^{(b)}$  is the residual at the  $i$ 'th iteration.

The left side of each of these figures contains results for the one-line orderings, and the right side contains results for the two-line orderings. Fig. 4.1 corresponds to the case  $\delta = 0$  (i.e. only the  $u_x$  first order term was present in (1.4)), Fig. 4.2 to  $\gamma = 0$  (only  $u_y$ ).

<sup>1,2</sup> We believe that these eigenvalue computations are affected by ill-conditioning, and that this is why the computed spectral radii exceed the asymptotic bounds.

and Fig. 4.3 to  $\gamma = \delta (u_x \text{ and } u_y)$ . The results are for the block Gauss-Seidel method with the natural, and red-black orderings. In addition, results for the block SOR method with the natural ordering are shown for some choices of  $\gamma$  and  $\delta$ . For SOR, we used the optimal value of  $\omega$  determined by (3.1), where  $\rho(B)^2$  is taken from the computed values of Tables 4.1 - 4.2.

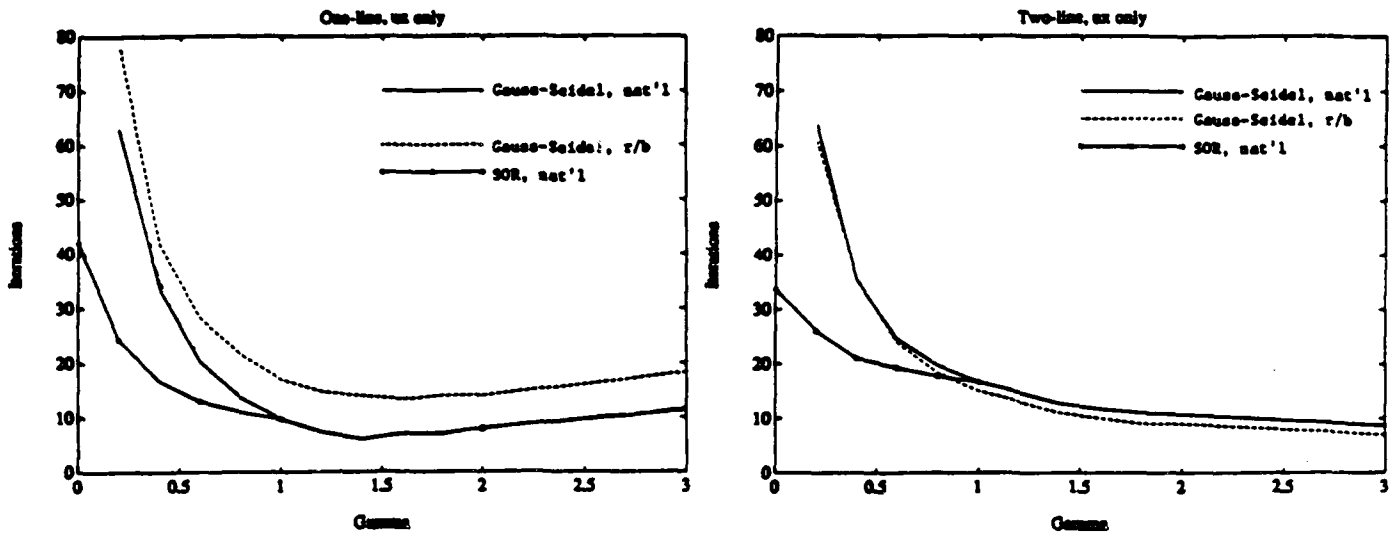


Fig. 4.1: Average iteration counts,  $h = 1/32$ ,  $\delta = 0$ .

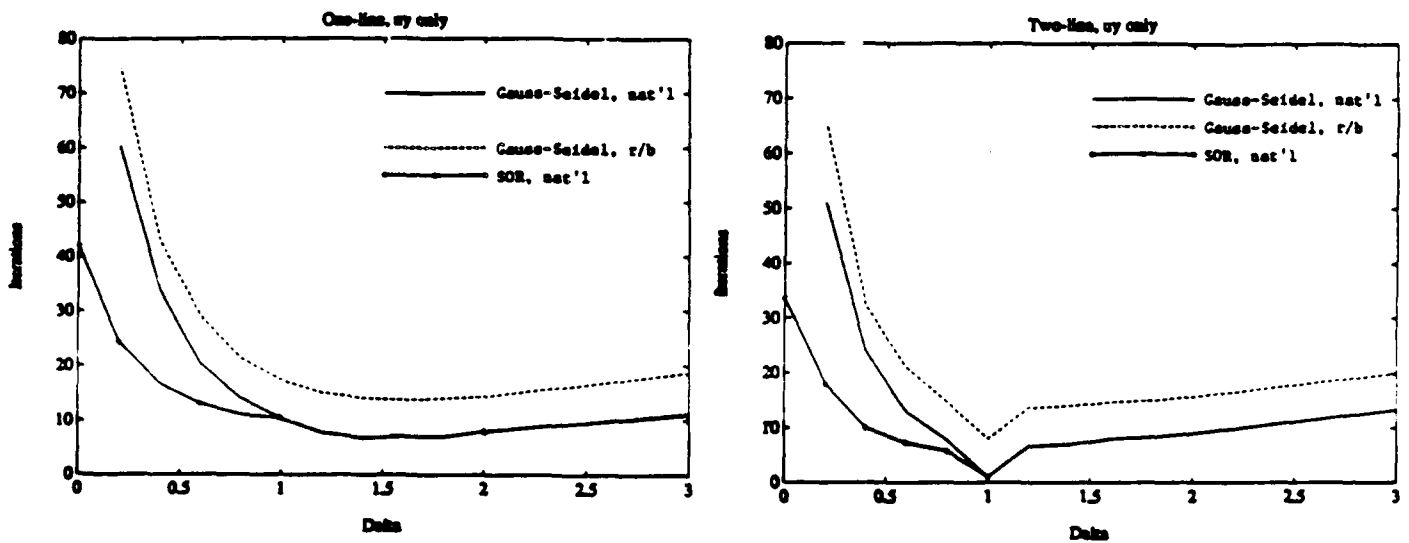


Fig. 4.2: Average iteration counts,  $h = 1/32$ ,  $\gamma = 0$ .

We make the following observations on these results. In most cases, the Gauss-Seidel method requires thirty or fewer iterations to reach the stopping criterion. The best results are obtained when  $\gamma$  or  $\delta$  are near one, and performance typically improves as  $|\gamma|$  or

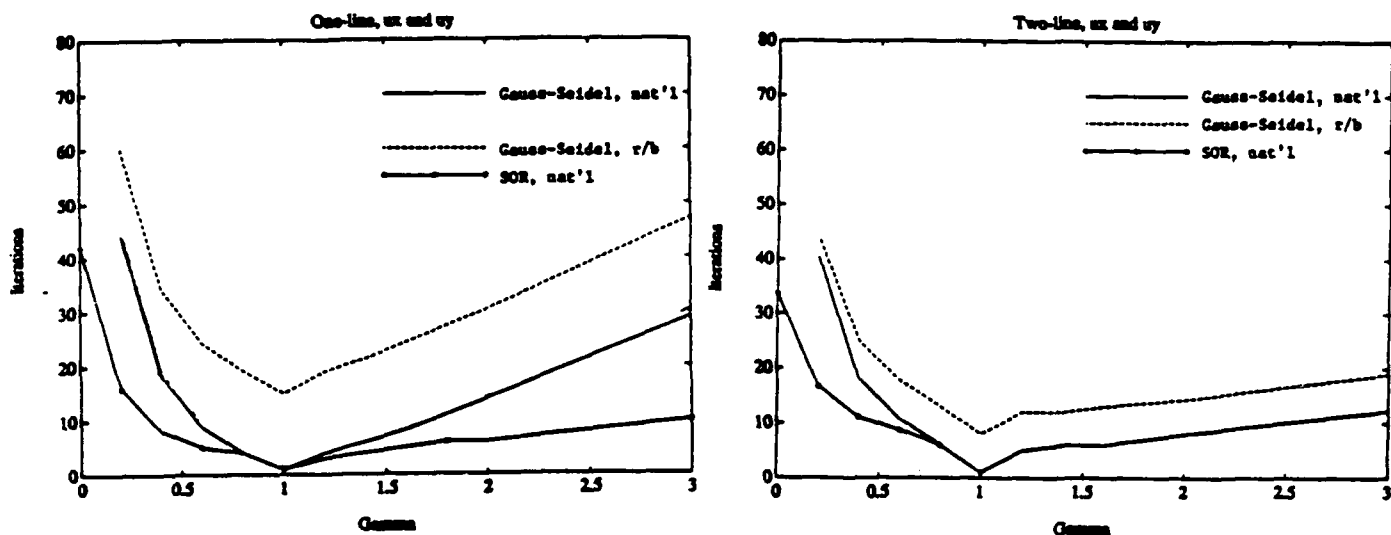


Fig. 4.3: Average iteration counts,  $h = 1/32$ ,  $\gamma = \delta$ .

$|\delta| \rightarrow 1$ . For all values of  $\gamma$  and  $\delta$  tested, the self-adjoint case ( $\gamma = \delta = 0$ ) required the largest number of Gauss-Seidel iterations. In these cases, for which the results are not shown on the figures, the stopping criterion was typically not reached after 150 iterations. The best results for large  $\gamma$  or  $\delta$  are for the two-line orderings with  $\delta = 0$  (Table 4.2 and Fig. 4.1). This is because as  $|\gamma|$  grows,  $S$  essentially consists of its block diagonal  $D$  plus a small perturbation. For large  $\delta$  and  $\gamma = 0$ , a vertical two-line splitting would give better results than the horizontal splitting used. SOR was much more effective than Gauss-Seidel when the latter was slow. We examined SOR only in cases where the spectrum of the block Jacobi iteration matrix is real, i.e. where either  $|\gamma| < 1$  and  $|\delta| < 1$  or (for the one-line ordering [2])  $|\gamma| > 1$  and  $|\delta| > 1$ . Thus, (3.1) applies. In variable coefficient problems of a similar character, it would be realistic to use an adaptive method to estimate the optimal value of  $\omega$  (see e.g. [11]).

#### References.

- [1] R. C. Y. Chin and T. A. Manteuffel, An analysis of block successive overrelaxation for a class of matrices with complex spectra, *SIAM J. Numer. Anal.* 25:564-585, 1988.
- [2] Iterative Methods for Cyclically Reduced Non-Self-Adjoint Linear Systems, Report UMIACS-TR-88-87, University of Maryland, Nov. 1988. To appear in *Math. Comp.*
- [3] Iterative Methods for Cyclically Reduced Non-Self-Adjoint Linear Systems II, Report UMIACS-TR-89-45, University of Maryland, Jun. 1989. Submitted to *Math. Comp.*
- [5] R. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler, *Matrix Eigensystem Routines: EISPACK Guide Extension*, Springer-Verlag, New York, 1972.
- [6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
- [4] L. A. Hageman and R. S. Varga, Block iterative methods for cyclically reduced matrix equations, *Numer. Math.* 6:106-119, 1964.
- [7] S. V. Parter, On estimating the "rates of convergence" of iterative methods for elliptic

- difference equations, *Trans. Amer. Math. Soc.* 114:320-354, 1965.
- [8] S. V. Parter and M. Steuerwalt, Block iterative methods for elliptic and parabolic difference equations, *SIAM J. Numer. Anal.* 19:1173-1195, 1982.
  - [9] PCGPAK User's Guide, Version 1.04, Scientific Computing Associates, New Haven, CT, 1987.
  - [10] R. S. Varga, *Matrix Iterative Analysis*, Prentice-Hall, New Jersey, 1962.
  - [11] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

## A COMPUTER SIMULATION OF THE FREELY-ASSOCIATING NEOCORTEX

M. Johnson, R. Scanlon, and M. Cipollo  
U.S. Army Armament Research, Development, and Engineering Center  
Close Combat Armaments Center  
Benet Laboratories  
Watervliet, NY 12189-4050

**ABSTRACT.** It is hypothesized that when signal energy is blocked by the thalamic reticular nucleus, the neocortex, being dynamically unstable, continues to turn on codons. Successive codons have the characteristic that they share some neurons. A computer simulation with video output gives us a grasp of the implications of this hypothesis.

**STATEMENT OF THE PROBLEM.** The investigation of the application of neural nets to machine intelligence involves the examination of patterns of activity. These patterns of activity are most difficult to comprehend. Quantification does not lead to insight. This paper describes a technique that allows subjective decisions about the effectiveness of a given configuration.

**BACKGROUND.** This paper had its genesis in a series of programs called GROWER [1] that were implemented on a coarse-grained parallel computer made up of 40 transputers. These programs investigated the mechanics by which a neocortical net could be interconnected solely by the effects of the incoming signal energy. The more active neurons were under pressure to extend their axons, and the less active were receptive to the acquisition of synapses. Following a period of growth, there was selective stabilization of synapses depending on sensory experience. The patterns of activity were compared with one another by computing a scalar, the cosine between the patterns when each was characterized as a vector. While these scalars allow one to see whether a given strategy is successful, there is a feeling that one would like to get a more direct intuition of the resemblance between patterns.

When GROWER is running and we give our attention to one input, we are aware that a unique pattern of activity results, yet to qualify the relationship by eye is beyond us. Furthermore, if we attend to a given pattern of activity, we can not say how much or in what way it is related to other patterns we have seen. One solution is to use recognizable patterns for the cortical activity ... patterns that we can recognize after intervening activity ... patterns for which we can form a subjective opinion about their correspondence with one another.

Human faces make up such a set. We have a specific mental faculty that allows us to recognize faces. Prosopagnosia is the loss of this ability. Patients can exhibit this highly specific form of visual agnosia following injury to the underside of the occipital lobe extending forward to the inner side of the temporal lobes of either or both cerebral hemispheres [2]. Such an unfortunate said, "I clearly see the details of your face, your mouth, your nose, but it is like a blur .... I am no longer able to see a face as a whole" [3]. The peculiar specificity of this ability led Prof. Kohonen to choose faces

for his patterns [4], as is done in this investigation. Not for one second do we think that such little faces float about on the human cortex. We use them so that anyone can form his own opinion, at a glance, about how one pattern resembles another. "It has to be made completely clear in the beginning that this is not an attempt to model the visual system of animals; optical images are only used because their quality can easily be esteemed visually" [4, p. 9].

Thinking is the subjective aspect of a freely-associating neocortex. There is always activity in the living cerebrum, such as pulses flowing along the axons. If the thalamus is relaying incoming sensory signals to the cortex, then the activity is this energy coursing through the cortex. If the thalamus blocks the sensor input, then the cortical activity is originally traceable to the last signal input, but because of the unstable nature of the cortex, the following activity becomes less and less predictable, but still logical after the act.

If we look at fluctuating transmembrane voltages, then all the neurons are active all the time. If we look at discrete events, such as axonal pulses, their asynchronous nature is mind-boggling. However, if we count the pulses during a time period, we have a number, and we can say that some neurons are more active than others during that period. Those neurons that are more active make up a codon. Of course this is a relative condition; still it does exist, and we can always set a level that distinguishes more from less. A codon is the material aspect of what the mind is subjectively aware of as a thought or a mental image. Unfortunately, if the cortex were exposed and the activity made visible, we could not perceive the relationship of the activity to the environment. It would appear as a coruscating, twinkling of a myriad of lights, transposed beyond any human insight. This is exactly the problem we have with GROWER but on a much, much larger scale.

Kohonen quotes Aristotle:

"Mental items (ideas, perceptions, sensations, or feelings) are connected in memory under the following conditions:

- (1) If they occur simultaneously, ('spatial contact').
- (2) If they occur in close succession ('temporal contact').
- (3) If they are similar.
- (4) If they are contrary." [4, p. 3]

Our aim is to show how simply association, and therefore thinking can be implemented. We use only (1) and (3). The extension to (2) and (4) is immediate. The life's experience of our simulated cortex consists of a series of faces. All the pixels of one face are presented simultaneously. This is how we make use of (1). The faces have areas, large and small, in which they share a pixel configuration with one or more other faces. In this connection it should be noted that the background is just as much a part of the "face" as the eyes and nose. If two faces have a large white or black area in the same part of the background, then this group of pixels constitutes a "similarity" in the sense of (3).

It is held that our imaginings are composed solely of past experiences, but in possible novel combinations. We can imagine an animal with a lion's head, a goat's body, and a snake's tail. We have experienced all these separately, but not in this combination. This is called a chimera. On the other hand, it is impossible to visualize any object as it might appear to an organism sensitive to ultraviolet radiation. The congenitally blind can not visualize; they can, however, conceive a spatial extension of the ability to touch. So we hope that we will find our machine cortex displaying a chimera.

When all the neurons in a column are off, and some or all the neighboring columns contain a firing neuron, then the most likely neuron to turn on is the one that shares a history with most of these firing neurons. This means that if most of the locally firing neurons share a "face," then this face will most likely be continued in the activated column. However, if this majority is shared by two or more "faces," then it is not at all clear what will ensue. It is our hypothesis that this is exactly what happens in the neocortex when we say we are thinking. The program is a successful simulation in the sense that chimeras do occur.

This completes the logic of the simulation. At any time, we could simulate the thalamus passing through new sensory input or chopping it off.

At this point, it is traditional to mathematically demonstrate the relative likelihood of various scenarios. This is exactly where we suggest that investigations of intelligent machines have gone astray in the past. A mathematical demonstration is substituted for laboratory experience with such machines, and as a result we have statements such as "We also must study the brain at a theoretical level that investigates the computations that are necessary to perform its functions" [5].

The brain does not compute. The incoming signal energy flows through the brain, but there is no computation--and no need of computation.

APPROACH TO THE PROBLEM. We implemented a computer program called THINKER that models a selection of the neocortex as a 64- by 64-array of columns. These are neural columns, not matrix columns. Each column contains eight principal neurons that are mutually inhibitory so that they form a competitive group with one being dominant during a jiffy. Brain jiffies come eight to ten a second.

THINKER consists of a main module and 14 subroutines written in Fortran, and organized around named common statements. It is best explained by analyzing the common statements:

```
LOGICAL*1 CORTEX, ON, SMALON, SAVE, EOF
COMMON/SLAB/ CORTEX(8,7,7,8,64,64)
COMMON/INPUT/ IPIX,KPIX(64,64), EOF
COMMON/WRKNG/ CONINC, CONDEC, CONCHK, CONDX(8,64,64),
1 ON(8,64,64)
COMMON/FRAME/ SMALON(64,64), KSGCNT(8,8), BGCNDX(8,8),
1 KACTIV(2,3)
COMMON/OUTPUT/ SAVE(64,64)
```



Enough programming maturity is assumed in the reader so that he knows that LOGICAL and CHARACTER bytes can be manipulated as eight-bit integers.

Sensory input is provided by a TV camera via a frame grabber with 256 by 256 pixels of 256 gray levels. To match the computational space available (16 Mbytes), this is reduced to 64 by 64 pixels of gray levels. At this resolution, faces can be recognized. NEXT furnishes these frames, one at a time, by placing the pixels in KPIX and incrementing IPIX. EOF is returned as false as long as input frames are available.

The primary visual cortex (area 17) is simulated as 64 by 64 columns (barrels) containing 32,768 neurons. A column contains eight 'computer' neurons, each of which represents a group of biological neurons. A group has a gray level as its output. Each neuron has an efferent inhibitory synapse on the other seven neurons in its column and efferent excitatory synapses on the 384 neurons closest to it, except the neurons in its own column. The array, CORTEX(I,J,K,L,M,N), contains (implicitly and explicitly) all the necessary working information. M and N locate the efferent column, and L the efferent neuron within the column. J and K identify the afferent column, and I the afferent neuron. The byte pointed to contains the potentiation of the corresponding synapse, and thus the facilitation may have 256 values. A zero value implicitly represents a missing (discarded or never formed) synapse. The inhibitory synapses are left implicit, as the output is based on the winner-take-all scheme.

During the experiential phase (the life experience of the simulated organism), faces are presented to the cortex by PRESNT. A face causes one neuron in each column to be excited according to the gray level of the corresponding pixel. The 4096 neurons that are excited have any mutual synapses potentiated. This is called Hebbian learning after a hypothesis of A.O. Hebb. We see that there is a limit of 255 faces that could be experienced before there is a possibility that some synapse has reached its limit of potentiation. This is not unreal; we should recall the trouble we have with twins. In real practice, many thousands of faces could be distinguished, but the reality of time available limited us to at most a couple of dozen. The only effect on the organism of this life experience is the potentiated synapses. There are no representations, no computations that the brain 'must perform,' and no need of them. We hold this potentiation to be equivalent to all that occurs during experience in the mammalian brain.

Now for free-running association. The signal energy is cut off. The activity of the cortex drops. The inactive neurons replenish their molecules. As the activity drops almost to nothing, those fully-replenished neurons in the vicinity of the last active neurons are triggered by the afferents they have from this dying activity. They, in turn, excite neurons in their vicinity. For easy viewing, we arrange things so that this activity spreads from a point (wavelike). This is an artifice, solely for academic reasons. If the pattern came on in dispersed points, as it would when the corpus callosum is involved, it would be impossible to "see" in the same way as random patterns can not be "seen."

Following the experiential phase, there is a period of random association. This follows the hypothesis that in the mammalian brain the thalamus

periodically (8 to 10 times a second) interrupts the flow of sensory input to the cortex, and during this period the cortex associates freely. If motor output does not ensue, this period is extended and the process called thinking arises.

In DEPLET a concept of "molecular depletion" is introduced. We speculate that neurons are always more or less active. A period of greater activity "uses up" certain molecules faster than they can be replenished. After such a period, a neuron becomes refractory and enters a period of lesser activity. As long as signal energy is coming in, a neuron can be exercised to the point of exhaustion. But when the signal energy is interdicted by the thalamus, the internal condition of a cortical neuron becomes more important, and it is not likely to fire if the store of needed molecules is mostly depleted.

The condition of each neuron is kept in CONDX(I,J,K). Initially it is set to 1.0. During each cycle, DEPLET is called and CONDX(I,J,K) adjusted. The status of the neuron is kept in ON(I,J,K). This logical variable is set to true if the neuron is on, and false otherwise. If the neuron is on, then CONDX is decremented by CONDEC. If CONDX falls to zero, then the neuron is turned off. If the neuron is off at entrance to DEPLET, CONDX is incremented by CONINC. The end of the refractory period is signaled by the rise of CONDX above a check point, CONCHK. Satisfactory values are

CONDEC=0.05  
CONINC=0.1  
CONCHK=0.4

If a neuron is turned off, a special check is made to see if this neuron is the center of an association. If it is, its indices are removed from KACTIV. The effects of this action are noted under 'major branch' below. Next, a call is made to ANLSYS. The cortex is divided into 8 by 8 subregions, each containing 8 by 8 columns, or 512 neurons. ANLSYS analyzes the condition of this subdivided cortex. If at least one neuron in a column is on, the corresponding SMALON(J,K) is set true, otherwise false. Because of the mutual inhibition involved in a column and the winner-take-all approach, the specifications of the simulation are that no more than one neuron in a column can be on at any given time. A tally of the number of neurons in a subregion that are on is placed in KBGCNT(JA,KA). A summation of the condition of all the neurons in the subregion (on or off) is placed in BGCNDX(JA,KA). If BGCNDX should be divided by KBGCNT, it would give the average condition of the neurons in that subregion.

At this point, a major branch in the association cycle occurs--the status of KACTIV(1,1) is checked. This array contains the indices of neurons that are the center of active associations. Initially the array is set to zero. It can become nonzero in FRAMER and be reset to zero in DEPLET. If KACTIV(1,1) is zero, the program branches to FRAMER. If it is nonzero, the branch is to THINK. We start with the branch to FRAMER.

It is hypothesized that the cortex maintains a level of activity that fluctuates within limits (unless in a pathological state). The governance of this condition is unknown, but biologically reasonable. We simulated this governance by monitoring the activity of our cortex. It is divided into 64 regions.

When the activity in each region is found to be at or below 20 active neurons, a new wave of activity is started by choosing the region that showed the most activity. If there is a tie, a random choice is made.

FRAMER checks KBGCNT. If at least one subregion has more than 20 neurons on in its 64 columns, a return is made. Otherwise, the subregion that has the most active neurons is selected. If no subregion has a single neuron on, then one column of the entire cortex is selected at random. A random neuron of this column is turned on, and SMALON of the column made true. Otherwise, a nonactive column in the identified subregion is selected at random. In either case the indices of the selected column are placed in KACTIV. At this point, if FRAMER has selected a nonactive column, a call is made to ASSOC.

ASSOC analyzes the selected column. The excitation afferent on each neuron of the column with CONDX greater than, or equal to, CONDEC is summed. This is taken to be the product of CORTEX and CONDX of the afferent neuron. The neuron that has the greatest excitation is turned on (ON is made true), and SMALON of the column is made true. If no neuron meets these tests, a return is made to FRAMER and then to CYCLE. Eventually a neuron is selected in ASSOC or randomly chosen in FRAMER.

Now we see that the next time we come to the major branch, KACTIV(1,1) is nonzero, and the path to THINK is taken. In THINK, the column recorded in KACTIV is taken as a center and a radius is set to one. A call is made to ASSOC for each column that is within this radius. This process is continued with the radius incremented by one until ASSOC has turned on 64 neurons. This arbitrary number was chosen because of viewer requirements in the final output. Alternatively, a return is made if the entire cortex has been swept with less than 64 neurons turned on.

This completes the logic of the association cycle. A TV frame is saved by recording a pixel for each column with the gray scale value of the active neuron (if there is one). Thirty frames are required for each second of the final videotape. The program is run long enough to give enough frames to make an acceptable viewing period.

We should note that the action of various minor routines that provide housekeeping functions have been skipped over as extraneous to the logical flow of the simulation.

RESULTS. THINKER was run many times with various selections of faces. The output was viewed on a video monitor. During the association phase, portions of various faces could be seen. Blending of the features on one face with those of another could be seen. The consensus of viewers is that the action of the network can be followed subjectively. The hoped-for chimeras appeared. The linking of one activity to another could sometimes be seen. A videotape of a typical run was made and presented at the conference.

CONCLUSIONS. This technique is an excellent way to judge the performance of a complicated neural net. The ability of the human eye and brain to recognize a portion of a face, even if presented fleetingly, goes far beyond any practical numerical description of a pattern.

The faces presented constitute the life history of the simulated organism. The extension of time-dependent sequences and moving objects is straightforward, but computationally expensive.

#### REFERENCES

1. Mark Johnson and Raymond Scanlon, "Parallel Processing of Fully Plastic Neural Networks," Proceedings of ISMM International Symposium, Computer Applications in Design, Simulation, and Analysis, Reno, Nevada, 22-24 February 1989.
2. N. Geschwind, "Specializations of the Human Brain," Scientific American, Vol. 241, 1979, pp. 180-199.
3. H. Hecain, "Clinical Symptomatology in Right and Left Hemisphere Lesions," in: Interhemisphere Relations and Cerebral Dominance, (V.B. Mountcastle, ed.), John Hopkins Press, Baltimore, Maryland, 1962.
4. Teuvo Kohonen, Self-Organization and Associative Memory, Springer-Verlag, New York, 1984.
5. Ellen C. Hildreth and John M. Hollerbach, "Artificial Intelligence: Computational Approach to Vision and Motor Control," Handbook of Physiology, Section 1: The Nervous System, Volume V: Higher Functions of the Brain, American Physiological Society, Bethesda, Maryland, 1987, p. 605.

## On topological complexity of solving polynomial equations of special type.

A. Libgober \*

Department of Mathematics  
University of Illinois at Chicago  
P.O.B. 4348, Chicago, Ill, 60680

**Abstract.** *The notion of Smale's topological complexity is reviewed. Topological and algebro-geometrical problems arising from finding topological complexity for solving polynomial equations with several vanishing coefficients formulated. Partial results toward their solutions are stated with an outline of proofs.*

In [S] S. Smale introduced the notion of topological complexity of an algorithm which provides an information on the *structure* of possible algorithms for solving a given problem rather than on their implementation time. Roughly speaking one assumes that the computation tree consists of nodes and connecting edges and that the nodes are either input nodes (having no incoming edges), or computation nodes (having one incoming and one outgoing edge), or branching nodes (having one incoming and two outgoing edges) or leaves (halts with no outgoing edges). The topological complexity of an algorithm is the number of branching nodes in its computation tree (or the number of leaves minus one).

In the same work S. Smale shows how the low bound for the topological complexity can be reduced to purely topological problems. For an algorithm for finding with accuracy  $\epsilon$  the roots of a polynomial from a family of polynomials  $F$  one can state that the topological complexity is greater or equal than the Schwartz genus of the covering map which relates to an ordered collection of roots of a polynomial from  $F$  without multiple roots the collection of its coefficients. Here by the Schwartz genus of a map  $f : X \rightarrow Y$  one means the minimal number  $k$  such that  $Y$  affords a cover with  $k$  open sets  $U_1, \dots, U_k$ , ( $Y = \bigcup_{i=1}^k U_i$ , such that  $f$  has a section over each  $U_i$ , i.e. for each  $i$  there exist a continuous map  $g_i : U_i \rightarrow X$  such that  $f \circ g_i = id$ ).

The Schwartz genus can be estimated from below as the maximal length of a non zero cup product of elements in  $Ker(H^i(Y, Z_2) \rightarrow H^i(X, Z_2))$ . One can use here twisted coefficients instead  $Z_2$  (cf. [Sch]). Using this method S. Smale ([S]) obtained  $(\log_2 n)^{2/3}$  as the lower bound for the topological complexity for finding with accuracy  $\epsilon$  the roots of the polynomial equation with one unknown. On the other hand in the case when  $Y$  is a quotient of  $X$  by a free action of a discrete group  $G$  one can use the homological genus of any  $G$ -module  $A$  as a lower bound for the Schwartz genus of the quotient map. The  $A$ -homological genus of a principal  $G$ -bundle  $f : X \rightarrow Y$  with the fibre a discrete group  $G$  with corresponding classifying map  $c : Y \rightarrow K(G, 1)$  ( $K(G, 1)$  is the Eilenberg MacLane space of the group  $G$ ) is the minimal integer  $i$  such that the canonical map  $H^j(K(G, 1), A) \rightarrow H^j(Y, c^*(A))$  is trivial for  $j \geq i$  ([Sch]). Using this V. Vasiljev [V] obtained as a lower bound for the Smale's problem  $n - \min_p(D_{p(n)})$  where  $D_{p(n)}$  is the sum of the digits in  $p$ -adic expansion of  $n$  and the minimum is taken over all primes  $p$ . He used as  $A$  the group of integers  $Z$  with the action of the symmetric group corresponding to the sign representation.

\* Supported by NSF and U.S. Army grants.

It seems it would be interesting to estimate the topological complexity of the solving some special classes of polynomial equations, for example polynomial equations with several vanishing coefficients, or answer similar questions for systems of polynomial equations (the latter was addressed in [L]). The application of the Smale's theory requires rather detailed information on the topology of the complements to discriminants in the space of special types of polynomials which seems is not available at the moment. This is the problem which we begin to address here. Specifically the following should be answered.

*Problem 1.* What is the fundamental group of the space of polynomials with several vanishing coefficients? Do the cohomology of this space depend only on this fundamental group? i.e. is the space of polynomials with vanishing coefficients is the Eilenberg MacLane space.

*Problem 2.* What are the cohomology with various (twisted) coefficients of the space of polynomials with several vanishing coefficients? What is their relationship with the cohomology of symmetric group?

If one considers the space of all monic polynomials then the answer to problem 1 goes back to E. Artin ([A]) and Fadell and Neuwirth [FN]: the fundamental group of the space of monic polynomials without multiple roots is the braid group  $B_n$  on  $n$  strings and this space is the Eilenberg MacLane space of  $B_n$ . The cohomology of the symmetric group surjects on the cohomology of the braid group in the case of cohomology with  $Z_2$  coefficients ([S]) or coefficients in sign representation of symmetric group ([V]).

Here we shall only indicate a solution for trinomials. First note that in the case of polynomials with several vanishing coefficients of the form

$$x^n + a_{i_1}x^{i_1} + a_{i_2}x^{i_2} \dots + a_{i_k} \quad (1)$$

the discriminant hypersurface is rather different than the discriminant of the space of all monic polynomials of degree  $n$ : in may become reducible and have different than in generic case degree (when the degree is  $2n - 2$ ).

*Examples of discriminants:*

1) For

$$x^5 - ax^2 - bx - c$$

the discriminant is

$$-27a^4b^2 + 2250a^2bc^2 - 1600ab^3c + 3125c^4 + 256b^5 + 108ca^5$$

2) For

$$x^6 + ax^3 + bx + c$$

the discriminant is

$$27000b^3ac^2 - 1350b^3ca^3 + 108a^5b^3 + 3125b^6 + 34992a^2c^4 - 87483 + 729c^2a^6 - 46656c^5$$

3) For

$$x^6 - ax^3 - bx^2 + c$$

the discriminant is

$$c^9 - 1024b^6 - 13824b^3c^2 + 108a^4b^3 - 46656c^4 + 729a^6c + 34992a^2c^3 - -8748a^4c^2 - 8640a^2b^3c$$

More generally one has the following:

**Theorem A.** *The discriminant of the family of polynomials of the form (1) has at most two irreducible components. The number of irreducible components is two if and only if  $i_{k-1} \neq 1$  and in this case one of components is the linear subspace  $a_{i_k} = 0$ . The degree of the discriminant is  $n + i_1 - i_{k-1}$ .*

(The first part of this theorem is obtained in [FS]). In the case of trinomials

$$x^n + ax^k + b \quad (2)$$

one can give complete answer to the problem 1 above.

**Theorem B.** *The fundamental group of the space of polynomials of form (2) with no multiple roots is the group of an algebraic link of the type explicitly determined by  $n$  and  $k$ . In particular if  $k = 1$  then the fundamental group of the space of polynomials of form (2) without multiple roots is the group of the torus knot of type  $(n, n - 1)$  i.e. admits a presentation with two generators  $g_1, g_2$  and one relator  $g_1^n = g_2^{n-1}$ . This space is the Eilenberg MacLane space for any  $n$  and  $k$ .*

**Remark:** For  $k = 1$  by virtue of having so simple presentation for the fundamental group one can easily describe the homomorphism of it into the braid group induced by embedding space of polynomials of form (2) into the space of all polynomials of degree  $n$ . If  $s_1, \dots, s_{n-1}$  are the standard generators of  $B_n$  then this homomorphism is given by  $g_1 \rightarrow s_1 \dots s_{n-1}, g_2 \rightarrow s_1 \dots s_{n-1} s_1$ . In particular this map is surjective. This in turn implies that the Galois group of generic trinomial equation in characteristic zero is the full symmetric group. (cf. [Sm] with much milder restrictions on characteristic of the ground field). This argument can be carried out in the case  $k > 1$  as well.

*Sketch of the proof* First notice that the equation of the reduced discriminant of the polynomial (2) is

$$b((-1)^{n-k-1} k^k (n-k)^n a^n - n^2 b^{(n-k)}) = 0 \quad (3)$$

if  $k > 1$  (cf. [S]). This follows from the fact that a polynomial has multiple root if and only if it and its derivative have common root. One can eliminate  $x$  from  $x^n + ax^k + b = 0, nx^{n-1} + kax^{k-1} = 0$  by replacing last equation by  $x^{n-k} = -ka/n$  (this is possible assuming  $x \neq 0$  which is the case provided  $b \neq 0$ .  $b = 0$  clearly belongs to support of discriminant if and only if  $k > 1$  which accounts for the first factor in (3)), substituting this in the first equation and replacing it by expression for  $x < k$  in terms of  $a$  and  $b$  after which elimination of  $x$  gives the second factor in (3). Now the complex curve  $D$  defined by (3) is invariant under  $C^*$  action on  $C^2$  which implies that the complement to  $D$  in  $C^2$  is equivalent to complement in 3-sphere to the link of the only singularity of the curve  $D$  namely the singularity at the origin. The Milnor fibration of the link of singularity of  $D$  exhibits the complement to the link of the singularity of  $D$  as a fibration over the circle with the real punctured surface as a fibre which implies that the complement to the curve  $D$  is the Eilenberg-MacLane space. In the case  $k = 1$  the equation of the discriminant is given by the vanishing of the second factor in (3). This equation after change of variables looks like  $u^n = v^{n-1}$ . The link of singularity of this curve is the torus knot of type  $(n, n - 1)$  and the description of the fundamental group of

the torus knot cited above is the well known one. The details of the proof of both theorems above and the cohomology calculations involved in the problem 2 will appear elsewhere.

*Acknowledgment.* The author thanks prof. S.Smale for introducing him to this subject and to Department of Mathematics of Columbia University for its hospitality which made it possible.

#### References.

- [A] E. Artin, *Theory of Braids*, *Ann. of Math.* 48, pp.101-126, (1947)
- [FN] E.Fadell and L. Neuwirth, *Configuration spaces*, *Math.Scand*, 10 pp 111-118. (1962)
- [FS] M.Freed and J.Smith, *Irreducible discriminant components of the coefficient spaces*, *Acta Arithm.* XLIV (1984) p.59-71.
- [L] H.Levine, *A Lower bound for the Topological complexity of Poly(D,n)*, *Journal of Complexity* 5, 34-44, (1989).
- [S] S.Smale, *On the topology of algorithms I*, *Journal of complexity*, (3) (1987).
- [Sch] A.S.Schwartz, *Genus of fibre bundles*, *Proc. of Moscow Math. Soc.* 10, 217-272 (1961).
- [Sm] J.Smith, *General trinomials having symmetric Galois group*, *Proc. Amer. Math. Soc.* 63(1977) p.208-217. (1977).



# SYMBOLIC UNCOUPLING AND EFFICIENT SOLUTION OF TREE-STRUCTURED LINEAR EQUATION SYSTEMS

Roger A. Wehage  
System Simulation and Technology Division  
AMSTA-RY  
United States Army Tank-Automotive Command  
Warren, Michigan 48397-5000

## ABSTRACT

A general method is presented for symbolically uncoupling a special class of augmented linear equations with tree topology defined by sparse nonsingular incidence or connectivity matrices. These equations, expressed in terms of excess and shared generalized state variables, are characteristic of  $p$ -element open loop systems. This paper presents an algorithm based on optimal block matrix permutation and factorization to precisely follow system topology and recursively generate a symbolic set of fully uncoupled equations yielding all variables in order  $p$  ( $O(p)$ ) operations. However, the operations coefficient can be relatively large for many problems, and recursion may inhibit full exploitation of vector and parallel processors. Thus an equivalent, compact and highly coupled set of generalized equations is obtained by eliminating the excess variables. The generalized matrix of this set of equations is symbolically manipulated into its natural factors using the previous recursive algorithm to get a new  $O(p)$  to  $O(p^2)$  solution. This algorithm has a much smaller operations coefficient and can more effectively exploit vector and parallel processing. Iterative refinement is also added to avoid many of the recursive decomposition steps required at each function evaluation. This allows even greater exploitation of vector and parallel processors. The algorithms are also modified to allow any number of the generalized states to be specified and to account for any degree of singularity or redundancy in the system equations.

## 1. INTRODUCTION

The increase in digital computer capacity and the development of advanced numerical methods has stimulated the desire to model and analyze large scale systems. When the equations must be solved thousands of times, direct numerical methods are unsuitable because of the excessive computer processing required to manipulate the resulting matrices. The extensive computational overhead and limited computer speed has prompted new searches for more efficient algorithms.

In general, formulations which incorporate the maximum number of variables yield the largest, least coupled augmented equation systems. Open-loop or tree-structured equations of this type can be solved recursively in  $O(p)$  operations (in many cases the minimum possible) with careful algorithm implementation [1-3]. However, the constant in front of  $O(p)$  can be relatively large, making recursion less effective than direct decomposition as the degree of system parallelism increases. A combined algorithm exploiting the sparsity of highly uncoupled augmented equations, compactness of generalized equations, iterative refinement, and vector and parallel processing can offer substantial computational advantages for many applications.

This paper presents a brief overview of a method for symbolically representing system topology by two sparse connectivity matrices. It is shown how these matrices loosely couple the augmented system equations and how they can be used to direct the recursive elimination and back substitution process. The connectivity matrix inverses can be used to transform these augmented equations into a maximally coupled generalized set of equations. A recursive

algorithm can then be employed to generate symbolic nonsingular natural factors of the resulting generalized coefficient matrix. The algorithms are also modified so any number of the generalized states can be routinely specified, and singularities and redundant equations can be handled. An iterative refinement algorithm is combined with the natural factors to exploit vector and parallel processors, yielding even more efficient solutions for many applications.

## 2. THE RECURSIVE ALGORITHM

Consider the following loosely coupled augmented equations

$$\begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13} \\ 0 & H_{42} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_2 \\ b_1 \\ b_4 \end{bmatrix} \quad (1)$$

or the equivalent highly coupled generalized equations

$$H_{42} R_{22} A_{21} R_{11} H_{13} x_3 = b_4 + H_{42} R_{22} [b_2 - A_{21} R_{11} b_1] \quad (2)$$

where

$$R_{11} = C_{11}^{-1} \quad (3)$$

$$R_{22} = C_{22}^{-1} \quad (4)$$

$$x_1 = R_{11} [b_1 + H_{13} x_3] \quad (5)$$

and

$$x_2 = -R_{22} [b_2 - A_{21} x_1] \quad (6)$$

(Throughout this paper the symbols 0 and I will be used to indicate respective zero and identity matrices whose dimensions are implied by the accompanying matrices and vectors.) Equation 2 may be obtained directly from Eq. 1 by eliminating the excess vectors  $x_1$  and  $x_2$  using Eqs. 5 and 6. Both Eqs. 1 and 2 represent the same system in terms of generalized state vector  $x_3$ . The remaining vectors  $x_1$ ,  $x_2$ ,  $b_1$ ,  $b_2$  and  $b_4$  evolve according to the basic system definitions. Subscripts 1 and 2 associate vectors and matrices with dual spaces where the dimensions may be different. Vectors  $x_3$  and  $b_4$  are dual subspaces of the respective spaces 1 and 2. The dimensions of spaces 1 and 2 are generally equal, as are their respective subspaces 3 and 4. Matrix  $A_{21}$  may be symmetric and semidefinite, positive semidefinite or even positive definite. If this is true and  $C_{22} = C_{11}^T$  and  $H_{42} = H_{13}^T$ , then the overall generalized coefficient matrix in Eq. 2 will be symmetric and will have one of the above properties.

The big challenge is to represent the governing equations for coupled systems in the augmented form of Eq. 1 or factored generalized form of Eq. 2. Intuitively one strives to formulate equations in terms of the minimum possible number of variables. This approach results in equations similar to Eq. 2, but unfortunately in the form  $Ax=b$  where the internal structure of Eqs. 1 or 2 have been lost. Thus it is important to change one's viewpoint of the problem and first represent individual components of coupled systems as separate entities in terms of the maximum number of state variables. These excess variables are obviously not required for successful formulation of the problem as indicated by the generalized equations, however, they are essential for identification and formulation of the augmented equations.

With these thoughts in mind, it may be possible to convert many different formulations into augmented form to take advantage of the algorithms presented in this paper.

Let the topology of a  $p$ -element system be defined by two sparse  $p$  by  $p$  matrices  $C_1$  and  $C_2$  containing only  $\pm 1$ 's and 0's (see Fig. 1 and Eqs. B8 and B9 of Appendix B). Matrix  $C_1$  defines a tree representing forward communication or coupling from parent element to child and matrix  $C_2$  defines a tree representing backward communication or coupling from children to parent elements. By special orientation of the communicating element interfaces (all oriented positively outward from the tree roots) and element naming convention (all named consecutively outward from the roots), row  $*$  of  $C_1$  ( $*$  = a, b, ... p) specifies that child  $*$  receives communication from parent  $*-1$ . Since child  $*$  always appears after parent  $*-1$  in the naming sequence, this convention causes  $C_1$  to be lower triangular with a unity determinant. Any row of  $C_1$  corresponding to a child not influenced by a parent will have a single 1 at its diagonal. In reality, this element forms the root of a new tree in  $C_1$ . The remaining rows will have exactly one 1 at the diagonal and one -1 to the left of the diagonal.

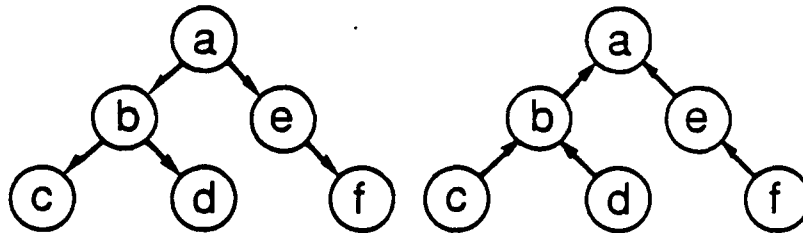


Figure 1. Forward and backward communicating trees of a six-element system

Similar to matrix  $C_1$ , row  $*$  of  $C_2$  ( $*$  = a, b, ... p) specifies which children  $*+1$  communicate back to parent  $*$  (there may be none, one or many). Since  $*+1$  is always greater than  $*$ ,  $C_2$  will always be upper triangular with 1's on the diagonal. If parent  $*$  receives no communication from any child, the corresponding row  $*$  of  $C_2$  will have only a single 1 at its diagonal. This is a terminal or leaf element in  $C_2$ . The introduction of  $C_2 = C_1^T$  allows unidirectional communication which may be useful in many applications.

Since the system is composed of  $p$  elements, there is a natural partitioning of vectors  $x_1$ ,  $x_2$ ,  $x_3$ ,  $b_1$ ,  $b_2$  and  $b_4$  into  $p$  corresponding subvectors. Likewise block diagonal matrices  $A_{21}$ ,  $H_{13}$  and  $H_{42}$  are partitioned into  $p$  submatrices consistent with the subvector dimensions. Block lower triangular matrix  $C_{11}$  has the same block sparsity pattern as  $C_1$  and its submatrix dimensions are compatible with the components of  $x_1$ . In a similar manner, block upper triangular matrix  $C_{22}$  has the same block sparsity pattern as  $C_2$  and its submatrix dimensions are compatible with the components of  $x_2$ . If  $C_2 = C_1^T$ , the generalized coefficient matrix in Eq. 2 will be block-symmetric and this matrix is represented by an undirected graph or simply graph. Otherwise it is represented by a directed graph or digraph [4]. In either case, all of the algorithms developed in this paper apply. A given subvector of a vector, or submatrix of a block diagonal matrix is referenced by appending an additional subscript to the corresponding

symbol. For example  $x_{1*}$  refers to the  $*$ th subvector of  $x_1$  and  $A_{21*}$  refers to the  $*$ th block diagonal submatrix of  $A_{21}$ .

With this convention, Eq. 1 can be described in more detail as follows. A typical submatrix equation in the first set of equations in Eq. 1 is

$$A_{21*-1} x_{1*-1} - x_{2*-1} + \sum_{*on} E_{22*} x_{2*} = b_{2*-1} \quad (7)$$

where submatrices  $E_{22*}$  are transformations or projections (not necessarily orthogonal or invertible) which take  $x_{2*}$  into  $x_{2*-1}$  coordinates. The dimensions of vectors  $x_{2*}$  and  $x_{2*-1}$  may be different. These matrices depend on the individual coordinates selected to represent the system being modeled. The  $-E_{22*}$  block submatrices fall into the off-diagonal block locations of  $C_{22}$  in accordance with the location of off-diagonal -1's in  $C_2$ . The summation in Eq. 7 results because parent  $*-1$  may receive communication from none, one or many children, hence the notation " $*$  on  $*-1$ ". The corresponding block row  $*-1$  of  $-C_{22}$  in Eq. 1 has just the right number of  $E_{22*}$ 's to pick up and transform the subvectors of children influencing  $*-1$ .

In a similar manner a typical submatrix equation in the second set of equations in Eq. 1 is

$$x_{1*} - E_{11*} x_{1*-1} - H_{13*} x_{3*} = b_{1*} \quad (8)$$

where submatrix  $E_{11*}$  is a transformation or projection (again not necessarily orthogonal or invertible) which takes  $x_{1*-1}$  into  $x_{1*}$  coordinates. The  $-E_{11*}$ 's fall into the off-diagonal block locations of  $C_{11}$  in accordance with the location of off-diagonal -1's in  $C_1$ .

Finally, a typical submatrix equation in the last set of equations in Eq. 1 is simply

$$H_{42*} x_{2*} = b_{4*} \quad (9)$$

For Eq. 1 to have a solution, Eqs. 2 to 6 indicate that it is necessary and sufficient for the generalized matrix  $H_{42} R_{22} A_{21} R_{11} H_{13}$  to be nonsingular so  $x_3$  can be evaluated. This implies that it is necessary for matrix  $H_{42}$  to at least have full row rank, for matrix  $H_{13}$  to have full column rank and for the subspace dimensions 3 and 4 to be the same, but this is not sufficient. Situations in which the coefficient matrix in Eq. 1 or 2 is singular will be addressed later. Thus Eqs. 2 to 6 and Eq. 1 are equivalent and this paper is concerned with the symbolic generation of the natural factors of the generalized matrix in Eq. 2 using a recursive algorithm for solving Eq. 1.

Using the topological information stored in matrices  $C_{11}$  and  $C_{22}$ , the following natural symbolic factors

$$L_{33}^{-1} D_{34} U_{44}^{-1} = [H_{42} R_{22} A_{21} R_{11} H_{13}]^{-1} \quad (10)$$

and

$$U_{44} D_{34}^{-1} L_{33} = H_{42} R_{22} A_{21} R_{11} H_{13} \quad (11)$$

will be obtained in a later section. Matrices  $L_{33}$  and  $U_{44}$  and their inverses are block lower and upper triangular with 1's on the diagonal and have respective block sparsity patterns identical to  $R_{11}$  and  $R_{22}$ , and  $D_{34}$  is block diagonal. This presumes that matrix  $D_{34}$  exists, which may not be true in many cases, such as when subspaces 3 and 4 have different dimensions. Such singularity problems will be addressed later. However, in the following development, assume that there are no singularities and that all matrices can be evaluated.

The basic recursive algorithm for solving Eq. 1 was developed in [1] and is repeated here with modifications. The labels in the following algorithm, e.g. ROL1 stand for "Recursive Open Loop step 1", etc. Recall that special ordering and orientation of the joints and elements made  $C_{11}$  block lower triangular and  $C_{22}$  block upper triangular with all 1's on the diagonal. Thus, solving equations of the form  $C_{11} x_1 = b_1$  is best done by evaluating the first subvector of  $x_1$ , and successively evaluating adjacent subvectors in forward order, in this case from a to p. Likewise  $C_{22} x_2 = b_2$  is solved from last subvector to first or in reverse order from p to a. The first situation amounts to optimal traversal of the  $C_1$  tree from root toward leaves and the second from leaves toward root in the  $C_2$  tree. The sequence in step ROL4 of the following algorithm is executed in forward order and steps ROL2, ROL3 and ROL5 in reverse order. Because of the special ordering of element names, decrementing or incrementing \* in the algorithm means "move to the adjacent symbol and the corresponding adjacent row and column in  $C_{11}$  or  $C_{22}$ ." Since adjacent elements may not have adjacent symbols, as noted earlier, reference to \*-1 means "select the adjacent or parent element closer to the root of the  $C_1$  tree and the corresponding parent row or column of  $C_{11}$ , not necessarily the previous element, row or column in the sequence." The symbol ← in steps ROL2, ROL3 and ROL5 means "assign the quantity in the right expression to the left expression." This is equivalent to summing projected quantities from the one or more communicating children onto their parent and it does not disturb the natural sequence of the recursive algorithm. If the system should have more than one tree, then each can be processed independent of the others by repeating the following algorithm with a different set of matrices. As an aid to understanding this and the following algorithms, a comprehensive example is developed in Appendix B.

### RECURSIVE OPEN LOOP ALGORITHM

ROL1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$  and  $C_{22}$

ROL2.0 Initialize  $A_{21}^0 = A_{21}$

ROL2.1 For \* = p to a repeat

$$(B_{41*} = H_{42*} A_{21*}^0)$$

$$B_{23*} = A_{21*}^0 H_{13*}$$

$$D_{34*} = [B_{41*} H_{13*}]^{-1}$$

$$F_{31*} = D_{34*} B_{41*}$$

$$F_{24*} = B_{23*} D_{34*}$$

$$A_{21*}^0 \leftarrow A_{21*}^0 + E_{22*} [A_{21*}^0 - B_{23*} F_{31*}] E_{11*}$$

Skip the last equation when \* corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an I since the corresponding submatrix  $E_{11*}$  or  $E_{22*}$  is zero

ROL3.0 Initialize  $b_2^0 = b_2$

ROL3.1 For \* = p to b repeat

$$(x_{3*}^0 = D_{34*} [b_{4*} + H_{42*} b_{2*}^0] - F_{31*} b_{1*})$$

$$b_{2*}^0 \leftarrow b_{2*}^0 + E_{22*} [b_{2*}^0 - A_{21*}^0 [b_{1*} + H_{13*} x_{3*}^0]]$$

Skip the second equation when \* corresponds to any column of  $C_{22}$  with only an I

ROL3.2  $x_{3a} = D_{34a} [b_{4a} + H_{42a} b_{2a}^0] - F_{31a} b_{1a}$

ROL4  $x_{1a} = b_{1a} + H_{13a} x_{3a}$

ROL4.1 For \* = b to p repeat

$$(x_{3*} = x_{3*}^0 - F_{31*} [E_{11*} x_{1*}])$$

$$x_{1*} = [E_{11*} x_{1*} + b_{1*} + H_{13*} x_{3*}]$$

The vector  $E_{11*} x_{1*} = 0$  when \* corresponds to any row of  $C_{11}$  with only an I

ROL5.0 Initialize  $x_2 = -b_2 + A_{21} x_1$

ROL5.1 For \* = p to b repeat

$$(x_{2*} \leftarrow x_{2*} + E_{22*} x_{2*})$$

Skip this equation when \* corresponds to any column of  $C_{22}$  with only an I

### 3. NATURAL FACTORS OF THE GENERALIZED MATRIX

Optimal block permutation and U-L factorization applied to the coefficient matrices of tree-structured systems result in an absolute minimum block fill pattern in the U and L matrices [4]. This is achieved by selecting forward elimination and back substitution sequences which precisely follow tree topology and never jump across (from branch to branch) unprocessed elements. Elimination and back substitution each require p sets of the recursive operations in the above algorithm. Instructions for completing recursive forward sweep step \* in ROL4.1

come from cell row \* of  $C_{11}$  and can be represented by an elementary matrix  $C_{11*}$  which corresponds to matrix  $C_{11}$  with all off-diagonal cell entries zeroed except in cell row \* (leaving exactly one (or none in case of a root or non communicating parent) off-diagonal block matrix

-  $E_{11*}$  in cell row \*) [5]. Matrix  $C_{11*}$  is a factor of  $C_{11}$  and not a submatrix, and thus has the same dimension as  $C_{11}$ . It follows that the  $p$  sequential instructions for the complete recursive process can be factored into the product

$$C_{11} = \prod_{*=1}^p C_{11*} \quad (12)$$

Further inspection of each elementary matrix  $C_{11*}$  reveals that its inverse can be obtained simply by reversing the single off-diagonal block matrix sign. Thus from Eq. 3 it follows that

$$R_{11} = \prod_{*=1}^p C_{11*}^{-1} \quad (13)$$

In a similar manner

$$C_{22} = \prod_{*=1}^p C_{22*} \quad (14)$$

and

$$R_{22} = \prod_{*=1}^p C_{22*}^{-1} \quad (15)$$

As above, elementary matrix  $C_{22*}$  corresponds to  $C_{22}$  with all off-diagonal entries zeroed except in cell column \* (which again leaves exactly one (or none in case of a root or noncommunicating child) off-diagonal block matrix -  $E_{22*}$  in cell column \*). Again,  $C_{22*}$  is not a submatrix but a factor of  $C_{22}$ .

Matrix  $C_{11*}$  can be envisioned as selecting and coupling subvectors  $x_{1*}$  and  $x_{1*,1}$  into the sum  $x_{1*} - E_{11*} x_{1*,1}$  in the composite arrays of subvectors. Matrix  $C_{11}$  couples the entire set of subvector sums where the components are either equal to or added to other subvectors with \* subscripts (see Eqs. 8 and 1). For example, suppose  $x_{1*} - E_{11*} x_{1*,1} = b_{1*} + \dots$  then  $C_{11} x_1 = b_1 + \dots$  represents the coupled system of subvectors. Reverse order of products in Eq. 13 still yields a nonsingular, lower triangular matrix with 1's on the diagonal, but the degree of sparsity in  $R_{11}$ , relative to  $C_{11}$ , is a function of the degree of parallelism, whereas  $C_{11}$  sparsity is a function only of the number of trees and elements in the system. The minimum fill pattern in  $C_{11}$  compared to  $R_{11}$  is what makes recursive algorithms so attractive for solving highly sequential problems on serial processors.

In a similar manner  $C_{22}$  couples a set of equations with subvector sums of the form

$x_{2^{*+1}} = \sum_{*on} E_{22^{*+}} x_{2^{*+}}$  which are equal to or added to other subvectors with  $*-1$  subscripts (see Eqs. 1 and 7).

The following matrices are assumed to operate on subvectors whose topology is defined by  $C_1$ . Let

$$B_{1i} = \text{diag}[B_{1ia}, B_{1ib}, \dots, B_{1ip}] \quad (i = 1, 3) \quad (16)$$

and consider a set of subvectors of the form  $-B_{31^{*+}} E_{11^{*+}} x_{1^{*+1}}$  which are equal to or added to other subvectors with  $*$  subscripts. The product  $B_{31}[C_{11} - I] x_1$  represents the coupled system of subvectors. Likewise the product  $[I + B_{11}[C_{11} - I]] x_1$  represents the coupled system of subvectors of the form  $x_{1^{*+}} = B_{11^{*+}} E_{11^{*+}} x_{1^{*+1}}$  which are equal to or added to other subvectors with  $*$  subscripts. In the latter case with  $i = 1$  the submatrices in Eq. 16 must all be square, but not necessarily nonsingular.

In a similar manner the following matrices

$$B_{2j} = \text{diag}[B_{2ja}, B_{2jb}, \dots, B_{2jp}] \quad (j = 1, 2, 4) \quad (17)$$

operate on subvectors whose topology is defined by  $C_2$ . The product  $[C_{22} - I] B_{2j} b_j, j = 1, 4$

couples subvectors  $-\sum_{*on} E_{22^{*+}} B_{2j^{*+}} b_{j^{*+}}, j = 1, 4$  and  $[I + [C_{22} - I] B_{22}] b_2$  couples subvectors

$b_{2^{*+1}} = \sum_{*on} E_{22^{*+}} B_{22^{*+}} b_{2^{*+}}$ , both which are equal to or added to other subvectors with  $*-1$  subscripts.

As above, with  $j = 2$ , the submatrices in Eq. 17 must all be square, but not necessarily nonsingular.

In summary, the matrix identities are:

$$\text{For } x_{1^{*+}} = E_{11^{*+}} x_{1^{*+1}} \rightarrow (*), \text{ use } C_{11} x_1 \quad (18)$$

$$\text{For } x_{2^{*+1}} = \sum_{*on} E_{22^{*+}} x_{2^{*+}} \rightarrow (*-1), \text{ use } C_{22} x_2 \quad (19)$$

$$\text{For } -B_{31^{*+}} E_{11^{*+}} x_{1^{*+1}} \rightarrow (*), \text{ use } B_{31}[C_{11} - I] x_1 \quad (20)$$



$$\text{For } x_{1*} - B_{11*} E_{11*} x_{1*} \rightarrow (*), \text{ use } [I + B_{11} [C_{11} - I]] x_1 \quad (21)$$

$$\text{For } - \sum_{*on}^{*-1} E_{22*} B_{2*} b_{j*}, j = 1, 4 \rightarrow (* - 1), \text{ use } [C_{22} - I] B_{2*} b_{j*}, j = 1, 4 \quad (22)$$

$$\text{For } b_{2*} - \sum_{*on}^{*-1} E_{22*} B_{22*} b_{2*} \rightarrow (* - 1), \text{ use } [I + [C_{22} - I] B_{22}] b_2 \quad (23)$$

These special shifting matrix structures insure that all subvectors are placed into the correct locations in the composite vector arrays. Subvector products such as  $B_{1*} x_{1*} \rightarrow (*)$  or  $B_{j*} x_{j*} \rightarrow (* - 1)$  are not shifted and the corresponding matrix product  $B_j x_j$  applies in both cases. Matrix  $B_{31} [C_{11} - I]$  is block lower triangular with zero matrices on the block diagonal, is generally rectangular and always singular. Matrix  $I + B_{11} [C_{11} - I]$  is always nonsingular and lower triangular with 1's on the diagonal. In a similar manner matrix  $[C_{22} - I] B_{2*}, j = 1, 4$  is block upper triangular with zero matrices on the block diagonal, is generally rectangular and always singular. Matrix  $I + [C_{22} - I] B_{22}$  is always nonsingular and upper triangular with 1's on the diagonal. Again, these simplifications are due to the special preordering and orientation of the elements.

With these tools, steps ROL3.1 and ROL4.1 of the previously developed recursive algorithm will now be used to obtain the natural factors. First write the equation in step ROL3.1 as

$$b_{2*}^p - \sum_{*on}^{*-1} E_{22*} P_{22*} b_{2*}^o = b_{2*}^o - \sum_{*on}^{*-1} E_{22*} [F_{24*} b_{4*} + A_{21*}^p b_{1*}] \quad (24)$$

where

$$P_{22*} = I - B_{23*} D_{34*} H_{42*} = I - F_{24*} H_{42*} \quad (25)$$

is a projection matrix and

$$A_{21*}^p = A_{21*}^o - F_{24*} H_{42*} A_{21*}^o = P_{22*} A_{21*}^o \quad (26)$$

is a projected coefficient matrix [1]. A second set of equations similar to Eqs. 25 and 26, which will be useful later are

$$P_{11*} = I - H_{13*} D_{34*} B_{41*} = I - H_{13*} F_{31*} \quad (27)$$

and

$$A_{21*}^p = A_{21*}^o - A_{21*}^o H_{13*} F_{31*} = A_{21*}^o P_{11*} \quad (28)$$

Now use the above identities to obtain

$$[I + [C_{22} - I] P_{22}] b_2^0 = b_2 + [C_{22} - I] [F_{24} b_4 + A_{21}^0 b_1] \quad (29)$$

or

$$b_2^0 = [I + [C_{22} - I] P_{22}]^{-1} [b_2 + [C_{22} - I] [F_{24} b_4 + A_{21}^0 b_1]] \quad (30)$$

where the submatrices of

$$B_{23} = \text{diag} [B_{23a}, B_{23b}, \dots, B_{23p}] \quad (31)$$

$$F_{24} = \text{diag} [F_{24a}, F_{24b}, \dots, F_{24p}] \quad (32)$$

$$F_{31} = \text{diag} [F_{31a}, F_{31b}, \dots, F_{31p}] \quad (33)$$

$$D_{34} = \text{diag} [D_{34a}, D_{34b}, \dots, D_{34p}] \quad (34)$$

$$P_{11} = \text{diag} [P_{11a}, P_{11b}, \dots, P_{11p}] \quad (35)$$

$$P_{22} = \text{diag} [P_{22a}, P_{22b}, \dots, P_{22p}] \quad (36)$$

and

$$A_{21}^0 = \text{diag} [A_{21a}^0, A_{21b}^0, \dots, A_{21p}^0] \quad (37)$$

used in the above equations are obtained from the recursive equations in step ROL2.1 of the above algorithm and Eqs. 25 to 28.

From step ROL4.1 with

$$x_{3*} = D_{34*} [b_{4*} + H_{42*} b_{2*}^0] - F_{31*} [b_{1*} + E_{11*} x_{1*}] \quad (38)$$

it follows that

$$x_3 = D_{34} [b_4 + H_{42} b_2^0] - F_{31} [b_1 - [C_{11} - I] x_1] \quad (39)$$

This equation cannot be evaluated directly because  $x_1$  also depends on  $x_3$ . However, one can substitute Eq. 5 to eliminate  $x_1$  and rearrange to obtain

$$[I + F_{31} [R_{11} - I] H_{13}] x_3 = D_{34} [b_4 + H_{42} b_2^0] - F_{31} R_{11} b_1 \quad (40)$$

In light of the earlier discussions, the matrix in front of  $x_3$  in Eq. 40 is nonsingular and lower triangular with 1's on the diagonal and Eqs. 2, 30 and 40 can now be used to find a symbolic representation of the natural factors of the generalized coefficient matrix in Eq. 2. First invert the left matrix in Eq. 40, substitute Eq. 30 and set the arbitrary quantities  $b_1$  and  $b_2$  to zero giving

$$x_3 = [I + F_{31} [R_{11} - I] H_{13}]^{-1} D_{34} [I + H_{42} [I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}] b_4 \quad (41)$$

Do likewise for Eq. 2 giving

$$x_3 = [H_{42} R_{22} A_{21} R_{11} H_{13}]^{-1} b_4 \quad (42)$$

Since  $b_4$  is also arbitrary it follows by equating coefficients in Eqs. 41 and 42 that

$$[H_{42} R_{22} A_{21} R_{11} H_{13}]^{-1} = [I + F_{31} [R_{11} - I] H_{13}]^{-1} D_{34} [I + H_{42} [I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}] \quad (43)$$

Equation 43 is the desired factored representation for Eq. 10 since the matrix to the left of  $D_{34}$  is lower triangular with 1's in the diagonal and the matrix to the right is upper triangular and also has 1's on the diagonal. Thus it follows that

$$L_{33} = I + F_{31} [R_{11} - I] H_{13} \quad (44)$$

and

$$U_{44} = [I + H_{42} [I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}]^{-1} \quad (45)$$

Equation 45 is inconvenient to evaluate, but the results from [2] suggest the following alternative expressions

$$L_{33} = [I + F_{31} [C_{11} - I] [I + P_{11} [C_{11} - I]]^{-1} H_{13}]^{-1} \quad (46)$$

and

$$U_{44} = I + H_{42} [R_{22} - I] F_{24} \quad (47)$$

Equations 46 and 47 are verified in Appendix A. The matrices in Eqs. 44 and 47 will be used to provide the optimal solution algorithm for Eq. 2.

While Eqs. 44 and 47 represent the most effective implementation of the natural factors, an interesting different view of these factors can be obtained by first noting that

$$\begin{aligned} F_{31} H_{13} &= D_{34} B_{41} H_{13} \\ &= D_{34} D_{34}^{-1} \\ &= I \end{aligned} \quad (48)$$

and

$$\begin{aligned} H_{42} F_{24} &= H_{42} B_{23} D_{34} \\ &= D_{34}^{-1} D_{34} \\ &= I \end{aligned} \quad (49)$$

Then these equations can be used to express the factors as

$$\begin{aligned} L_{33} &= F_{31} R_{11} H_{13} \\ &= D_{34} [H_{42} A_{21}^0 R_{11} H_{13}] \\ &= [H_{42} A_{21}^0 H_{13}]^{-1} [H_{42} A_{21}^0 R_{11} H_{13}] \end{aligned} \quad (50)$$

and

$$\begin{aligned}
U_{44} &= H_{42} R_{22} F_{24} \\
&= [H_{42} R_{22} A_{21}^{\circ} H_{13}] D_{34} \\
&= [H_{42} R_{22} A_{21}^{\circ} H_{13}] [H_{42} A_{21}^{\circ} H_{13}]^{-1}
\end{aligned} \tag{51}$$

where

$$D_{34} = [H_{42} A_{21}^{\circ} H_{13}]^{-1} \tag{52}$$

Finally, substitute these equations into Eq. 11 to give

$$\begin{aligned}
U_{44} D_{34}^{-1} L_{33} &= [H_{42} R_{22} A_{21}^{\circ} H_{13}] [H_{42} A_{21}^{\circ} H_{13}]^{-1} [H_{42} A_{21}^{\circ} R_{11} H_{13}] \\
&= [H_{42} R_{22}] [A_{21}^{\circ} H_{13} [H_{42} A_{21}^{\circ} H_{13}]^{-1} H_{42} A_{21}^{\circ}] [R_{11} H_{13}] \\
&= [H_{42} R_{22}] A_{21} [R_{11} H_{13}]
\end{aligned} \tag{53}$$

#### 4. THE GENERALIZED SOLUTION ALGORITHM

The first step in determining  $x_3$  in Eq. 2, given all the necessary matrices and vectors is to find  $y_4$  from

$$U_{44} y_4 = b_4 + H_{42} R_{22} [b_2 - A_{21} R_{11} b_1] \tag{54}$$

Since  $U_{44}$  is upper triangular, evaluate the subvectors of  $y_4$  from bottom to top. Next evaluate the subvectors of  $x_3$  from top to bottom by solving

$$L_{33} x_3 = D_{34} y_4 \tag{55}$$

Matrix  $L_{33}$  always has the same block fill pattern as  $R_{11}$  and it was noted earlier that the degree of fill is strictly a function of system topology. Thus the overhead in Eqs. 54 and 55 can vary from  $O(p^2)$  for serial trees where  $R_{11}$  and  $L_{33}$  have maximum fill below the diagonal, to  $O(p)$  for completely parallel systems in which case  $R_{11}$  and  $L_{33}$  have minimum fill below the diagonal.

The revised computational algorithm is now presented where the symbol POL stands for "Parallel Open Loop".

#### PARALLEL OPEN LOOP ALGORITHM

POL1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$  and  $C_{22}$

POL2.0 Initialize  $A_{21}^{\circ} = A_{21}$

POL2.1 For  $\star = p$  to  $a$  repeat

$$( B_{41\star} = H_{42\star} A_{21\star}^{\circ} )$$

$$B_{23\star} = A_{21\star}^{\circ} H_{13\star}$$

$$D_{34\star} = [B_{41\star} H_{13\star}]^{-1}$$

$$F_{31\star} = D_{34\star} B_{41\star}$$

$$F_{24\star} = B_{23\star} D_{34\star}$$

$$A_{21\rightarrow}^{\circ} \leftarrow A_{21\rightarrow}^{\circ} + E_{22\rightarrow} [A_{21\rightarrow}^{\circ} - B_{22\rightarrow} F_{31\rightarrow}] E_{11\rightarrow}$$

Skip the last equation when \* corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an I since the corresponding submatrix  $E_{11\rightarrow}$  or  $E_{22\rightarrow}$  is zero

POL3 Evaluate

$$L_{33} = I + F_{31} [R_{11} - I] H_{13}$$

$$U_{44} = I + H_{42} [R_{22} - I] F_{24}$$

POL4 Solve

$$U_{44} y_4 = b_4 + H_{42} R_{22} [b_2 - A_{21} R_{11} b_1] \text{ for } y_4$$

$$L_{33} x_3 = D_{34} y_4 \text{ for } x_3$$

POL5 If desired, evaluate

$$x_1 = R_{11} [b_1 + H_{13} x_3]$$

$$x_2 = -R_{22} [b_2 - A_{21} x_1]$$

## 5. SYMBOLIC FACTORS OF PARTITIONED AND SINGULAR MATRICES

Inspection of the matrix in Eq. 1 reveals that the submatrix

$$\begin{bmatrix} A_{21} & -C_{22} \\ C_{11} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & R_{11} \\ -R_{22} & R_{22} A_{21} R_{11} \end{bmatrix} \quad (56)$$

is nonsingular, independent of the rank of  $A_{21}$  (see more discussion in Appendix B). Thus  $x_1$  and  $x_2$  can be determined, regardless of the rank deficiency of the coefficient matrix in Eq. 1 or 2. Rank deficiency in these matrices is reflected in the rank deficiency of the individual matrices

$$D_{34\rightarrow}^{-1} = H_{42\rightarrow} A_{21\rightarrow}^{\circ} H_{13\rightarrow} \quad (57)$$

in the recursive steps ROL2.1 or POL2.1. If a given  $D_{34\rightarrow}^{-1}$  is singular, the corresponding Eq. 38 must be returned to the form

$$D_{34\rightarrow}^{-1} x_{3\rightarrow} = H_{42\rightarrow} A_{21\rightarrow}^{\circ} H_{13\rightarrow} x_{3\rightarrow}$$

$$= b_{4\rightarrow} + H_{42\rightarrow} b_{2\rightarrow}^{\circ} - B_{31\rightarrow} [b_{1\rightarrow} + E_{11\rightarrow} x_{1\rightarrow-1}] \quad (58)$$

Equation 58 indicates that components of  $x_{3\rightarrow}$  equal in number to the column rank deficiency of  $D_{34\rightarrow}^{-1}$  cannot be computed and must be supplied and/or a number of equations equal the row rank deficiency of  $D_{34\rightarrow}^{-1}$  are dependent and must be checked for consistency and/or eliminated. In the first situation, the undetermined components of  $x_{3\rightarrow}$  are assumed to be specified or the problem will be ill posed. In addition, one may wish to specify or drive one or more of the components

of  $x_{3*}$  even though  $D_{34*}^{-1}$  may have full column rank. In either situation, these known quantities can be moved to the right hand side of the equations and eliminated from the recursive solution process. With this knowledge  $x_{3*}$ ,  $b_{4*}$ ,  $H_{13*}$  and  $H_{42*}$  can be partitioned according to free and specified variables, and independent and redundant equations as

$$x_{3*} = \begin{bmatrix} x_{3*}^i \\ x_{3*}^s \end{bmatrix} \quad (59)$$

$$b_{4*} = \begin{bmatrix} b_{4*}^i \\ b_{4*}^d \end{bmatrix} \quad (60)$$

$$H_{13*} = [H_{13*}^i \quad H_{13*}^s] \quad (61)$$

and

$$H_{42*} = \begin{bmatrix} H_{42*}^i \\ H_{42*}^d \end{bmatrix} \quad (62)$$

In a similar manner, Eq. 1 can be partitioned as

$$\begin{bmatrix} A_{21} & -C_{22} & 0 & 0 \\ C_{11} & 0 & -H_{13}^i & -H_{13}^s \\ 0 & H_{42}^i & 0 & 0 \\ 0 & H_{42}^d & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3^i \\ x_3^s \end{bmatrix} = \begin{bmatrix} b_2 \\ b_1 \\ b_4^i \\ b_4^d + b_4^s \end{bmatrix} \quad (63)$$

The dimensions of  $x_3^i$  and  $b_4^i$  will always be the same, matrix  $H_{13}^i$  will have full column rank and matrix  $H_{42}^i$  will have full row rank. The slack variables in vector  $b_4^s$  are introduced to insure equality in the last set of dependent redundant equations.

The first three sets of equations may be rewritten as

$$\begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13}^i \\ 0 & H_{42}^i & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3^i \end{bmatrix} = \begin{bmatrix} b_2 \\ b_1 + H_{13}^s x_3^s \\ b_4^i \end{bmatrix} \quad (64)$$

and the last set of equations as

$$b_4^s = -b_4^d + H_{42}^d x_2 \quad (65)$$

By construction, the coefficient matrix in Eq. 64 is now nonsingular and the general equation equivalent to Eq. 2 with nonsingular coefficient matrix is

$$H_{42}^i R_{22} A_{21} R_{11} H_{13}^i x_3^i = b_4^i + H_{42}^i R_{22} [b_2^i - A_{21} R_{11} [b_1^i + H_{13}^i x_3^i]] \quad (66)$$

The "Recursive Open Loop algorithm for Singular" (ROLS) matrices may now be stated. For simplicity, the following algorithm assumes that all singularities are known in advance and that the partitioning into free and specified variables is known. A more general algorithm would detect rank deficiencies in submatrices  $D_{34}^i$  at each step and take appropriate action as necessary.

### RECURSIVE OPEN LOOP ALGORITHM FOR SINGULAR MATRICES

ROLS1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$ ,  $C_{22}$  and  $x_3^s$

ROLS2.0 Initialize  $A_{21}^o = A_{21}$

ROLS2.1 For  $\star = p$  to  $a$  repeat

$$(B_{41\star}^i = H_{42\star}^i A_{21\star}^o$$

$$B_{23\star}^i = A_{21\star}^o H_{13\star}^i$$

$$D_{34\star}^{ii} = [B_{41\star}^i \ H_{13\star}^i]^{-1}$$

$$F_{31\star}^i = D_{34\star}^{ii} B_{41\star}^i$$

$$F_{24\star}^i = B_{23\star}^i D_{34\star}^{ii}$$

$$A_{21\star-1}^o \leftarrow A_{21\star-1}^o + E_{22\star} [A_{21\star}^o - B_{23\star}^i F_{31\star}^i] E_{11\star}$$

Skip the last equation when  $\star$  corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an  $l$  since the corresponding submatrix  $E_{11\star}$  or  $E_{22\star}$  is zero

ROLS3.0 Initialize  $b_2^o = b_2$

ROLS3.1 For  $\star = p$  to  $b$  repeat

$$(x_{3\star}^i = D_{34\star}^{ii} [b_{4\star}^i + H_{42\star}^i b_{2\star}^o] - F_{31\star}^i [b_{1\star}^i + H_{13\star}^i x_{3\star}^s])$$

$$b_{2\star-1}^o \leftarrow b_{2\star-1}^o + E_{22\star} [b_{2\star}^o - A_{21\star}^o [b_{1\star}^i + H_{13\star}^i x_{3\star}^s + H_{13\star}^i x_{3\star}^i]]$$

Skip the second equation when  $\star$  corresponds to any column of  $C_{22}$  with only an  $l$

ROLS3.2  $x_{3a}^i = D_{34a}^{ii} [b_{4a}^i + H_{42a}^i b_{2a}^o] - F_{31a}^i [b_{1a}^i + H_{13a}^i x_{3a}^s]$

ROLS4  $x_{1a} = b_{1a} + H_{13a}^i x_{3a}^i$

ROLS4.1 For  $\star = b$  to  $p$  repeat

$$(x_{3\star}^i = x_{3\star}^o - F_{31\star}^i [E_{11\star} x_{1\star-1}])$$

$$x_{1\star} = x_{1\star-1} + b_{1\star} + H_{13\star}^i x_{3\star}^i$$

The vector  $E_{11\star} x_{1\star-1} = 0$  when  $\star$  corresponds to any row of  $C_{11}$  with only an  $l$

ROLS5.0 Initialize  $x_2 = -b_2 + A_{21} x_1$

ROLS5.1 For  $\star = p$  to  $b$  repeat

$$(x_{2\star-1} \leftarrow x_{2\star-1} + E_{22\star} x_{2\star})$$

Skip this equation when  $\star$  corresponds to any column of  $C_{22}$  with only an I

$$b_{4\star}^d = -b_{4\star}^d + H_{42\star}^d x_{2\star})$$

Skip this equation when  $H_{42\star}^d$  is null

Following the same procedures as above, the revised "Parallel Open Loop algorithm for Singular" (POLS) matrices becomes:

### PARALLEL OPEN LOOP ALGORITHM FOR SINGULAR MATRICES

POLS1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$ ,  $C_{22}$  and  $x_3^s$

POLS2.0 Initialize  $A_{21}^o = A_{21}$

POLS2.1 For  $\star = p$  to  $a$  repeat

$$(B_{41\star}^i = H_{42\star}^i A_{21\star}^o)$$

$$B_{23\star}^i = A_{21\star}^o H_{13\star}^i$$

$$D_{34\star}^{ii} = [B_{41\star}^i \ H_{13\star}^i]^{-1}$$

$$F_{31\star}^i = D_{34\star}^{ii} B_{41\star}^i$$

$$F_{24\star}^i = B_{23\star}^i D_{34\star}^{ii}$$

$$A_{21\star-1}^o \leftarrow A_{21\star-1}^o + E_{22\star} [A_{21\star}^o - B_{23\star}^i F_{31\star}^i] E_{11\star}$$

Skip the last equation when  $\star$  corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an I since the corresponding submatrix  $E_{11\star}$  or  $E_{22\star}$  is zero

POLS3 Evaluate

$$L_{33}^{ii} = I + F_{31}^i [R_{11} - I] H_{13}^i$$

$$U_{44}^{ii} = I + H_{42}^i [R_{22} - I] F_{24}^i$$

POLS4 Solve

$$U_{44}^{ii} y_4^i = b_4^i + H_{42}^i R_{22} [b_2 - A_{21} R_{11} [b_1 + H_{13}^i x_3^s]] \text{ for } y_4^i$$

Solve

$$L_{33}^{ii} x_3^i = D_{34}^{ii} y_4^i \text{ for } x_3^i$$

POLS5 If desired, evaluate

$$x_1 = R_{11} [b_1 + H_{13} x_3]$$

$$x_2 = -R_{22} [b_2 - A_{21} x_1]$$

$$b_4^s = -b_4^d + H_{42}^d x_2$$



These algorithms show that the recursive process makes it a simple matter to efficiently eliminate any number of embedded singularities and redundant equations, and to find symbolic factors for the largest nonsingular submatrix.

## 6. ITERATIVE REFINEMENT

Consider the linear system of equations  $Ax=b$  and suppose the product of lower triangular matrix  $L$  and upper triangular matrix  $U$  approximates  $A$ . Then solve the equivalent system of equations  $LUx^{(1)}=b$  for  $x^{(1)}$ , the first approximation to  $x$ . The residual vector for this first iteration is

$$r^{(1)} = b - Ax^{(1)} \quad (67)$$

Inverting  $A$  in Eq. 67 yields  $A^{-1}r^{(1)} = x - x^{(1)}$  and implies that a correction  $\Delta x^{(1)}$  to  $x^{(1)}$  may be approximated by solving

$$LU\Delta x^{(1)} = r^{(1)} \quad (68)$$

giving a second approximation for  $x$

$$x^{(2)} = x^{(1)} + \Delta x^{(1)} \quad (69)$$

In general the iterative process repeats to the  $(k)$ th step as

$$r^{(k)} = b - Ax^{(k)} \quad (70)$$

$$LU\Delta x^{(k)} = r^{(k)} \quad (71)$$

and

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \quad (72)$$

One can show by induction at the  $k^{\text{th}}$  iteration that

$$r^{(k)} = [I - A(LU)^{-1}]^k b \quad (73)$$

which implies that the spectral radius of  $I - A(LU)^{-1}$  should be less than 1 [4]. Clearly if  $LU=A$  then  $r^{(1)}=0$ . As the product  $LU$  deviates from  $A$ , the rate of convergence decreases and the number of iterations to an acceptable solution increases. An excessive number of iterations indicates the need to update  $L$  and  $U$ .

Iterative refinement is useful in the above algorithms especially for slowly varying systems because the costly steps in ROL2, POL2 and POL3 can be avoided most of the time. If the iterations converge quickly, this can yield substantial savings in computer time. Furthermore iterative refinement allows the POL algorithm to more effectively exploit vector and parallel processors since less time is spent in the serial operations necessary to evaluate the

matrices in steps POL2 and POL3. The problem is to find equations to inexpensively evaluate the residuals for the above algorithms and to incorporate them. Residual calculations are always based on updated quantities, not approximations, since the residuals must go to zero as the iteration converges to the correct solution. The residuals for the first algorithm are easily obtained with the help of Eq. 1. Let  $x_3^{(k)}$  be the  $k$ th approximation to the solution and compute  $x_1$ ,  $x_2$  and  $r_4^{(k)}$  from

$$C_{11} x_1 = b_1 + H_{13} x_3^{(k)} \quad (74)$$

$$C_{22} x_2 = -[b_2 - A_{21} x_1] \quad (75)$$

and

$$r_4^{(k)} = b_4 - H_{42} x_2 \quad (76)$$

To see that this is the correct residual, these equations may be combined with the help of Eqs. 5 and 6 to obtain

$$r_4^{(k)} = b_4 + H_{42} R_{22} [b_2 - A_{21} R_{11} b_1] - H_{42} R_{22} A_{21} R_{11} H_{13} x_3^{(k)} \quad (77)$$

Which is simply the residual of Eq. 2. Thus the residual in Eq. 76 or 77 is appropriate for both of the algorithms. According to Eq. 71, Eq. 77 implies that

$$H_{42} R_{22} A_{21} R_{11} H_{13} \Delta x_3^{(k)} = r_4^{(k)} \quad (78)$$

which from Eqs. 1 and 2 leads to

$$\begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13} \\ 0 & H_{42} & 0 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \\ \Delta x_3^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ r_4^{(k)} \end{bmatrix} \quad (79)$$

and

$$x_3^{(k+1)} = x_3^{(k)} + \Delta x_3^{(k)} \quad (80)$$

The quantities  $\Delta x_1$  and  $\Delta x_2$  from Eq. 79 cannot be used to update  $x_1$  and  $x_2$  because there are no residuals associated with them and one cannot be sure that they will satisfy Eqs. 74 and 75. Therefore, they must be computed directly from Eqs. 74 and 75 for evaluating the residual in Eq. 76. Thus Eqs. 74 to 76 and 78 to 80 provide the information necessary to add iterative refinement to the above two algorithms. The modified "Recursive Open Loop algorithm with Iterative refinement" (ROLI) follows as:

### RECURSIVE OPEN LOOP ALGORITHM WITH ITERATIVE REFINEMENT

ROLI1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$  and  $C_{22}$

ROLI2.0 Bypass steps ROLI2.1 and ROLI2.2 unless convergence rate is slow

ROLI2.1 Initialize  $A_{21}^0 = A_{21}$

ROLI2.2 For  $\star = p$  to a repeat

$$(B_{41\star} = H_{42\star} A_{21\star}^{\circ})$$

$$B_{23\star} = A_{21\star}^{\circ} H_{13\star}$$

$$D_{34\star} = [B_{41\star} H_{13\star}]^{-1}$$

$$F_{31\star} = D_{34\star} B_{41\star}$$

$$F_{24\star} = B_{23\star} D_{34\star}$$

$$A_{21\star-1}^{\circ} \leftarrow A_{21\star-1}^{\circ} + E_{22\star} [A_{21\star}^{\circ} - B_{23\star} F_{31\star}] E_{11\star}$$

Skip the last equation when  $\star$  corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an I since the corresponding submatrix  $E_{11\star}$  or  $E_{22\star}$  is zero

ROLI3.0 Initialize  $b_2^{\circ} = b_2$

ROLI3.1 For  $\star = p$  to b repeat

$$(x_{3\star}^{\circ} = D_{34\star} [b_{4\star} + H_{42\star} b_{2\star}^{\circ}] - F_{31\star} b_{1\star})$$

$$b_{2\star-1}^{\circ} \leftarrow b_{2\star-1}^{\circ} + E_{22\star} [b_{2\star}^{\circ} - A_{21\star}^{\circ} [b_{1\star} + H_{13\star} x_{3\star}^{\circ}]]$$

Skip the second equation when  $\star$  corresponds to any column of  $C_{22}$  with only an I

ROLI3.2  $x_{3a}^{(0)} = D_{34a} [b_{4a} + H_{42a} b_{2a}^{\circ}] - F_{31a} b_{1a}$

ROLI4  $x_{1a} = b_{1a} + H_{13a} x_{3a}^{(0)}$

ROLI4.1 For  $\star = b$  to p repeat

$$(x_{3\star}^{(0)} = x_{3\star}^{\circ} - F_{31\star} [E_{11\star} x_{1\star-1}])$$

$$x_{1\star} = [E_{11\star} x_{1\star-1}] + b_{1\star} + H_{13\star} x_{3\star}^{(0)}$$

The vector  $E_{11\star} x_{1\star-1} = 0$  when  $\star$  corresponds to any row of  $C_{11}$  with only an I

LOOP For  $k = 0, 1, \dots$  do to LOOP End

ROLI5.0 Initialize  $x_2 = -b_2 + A_{21} x_1$

ROLI5.1 For  $\star = p$  to b repeat

$$(x_{2\star-1} \leftarrow x_{2\star-1} + E_{22\star} x_{2\star})$$

Skip this equation when  $\star$  corresponds to any column of  $C_{22}$  with only an I

ROLI6 Evaluate

$$r_4^{(k)} = b_4 - H_{42} x_2 \text{ and exit if small}$$

ROLI7.0 Initialize  $rb_2^0 = 0$

ROLI7.1 For  $\star = p$  to  $b$  repeat

$$(\Delta x_{3\star}^0 = D_{34\star} [r_{4\star}^{(k)} + H_{42\star} rb_{2\star}^0])$$

$$rb_{2\star-1}^0 \leftarrow rb_{2\star-1}^0 + E_{22\star} [rb_{2\star}^0 - B_{23\star} \Delta x_{3\star}^0]$$

Skip the second equation when  $\star$  corresponds to any column of  $C_{22}$  with only an I

ROLI7.2  $(\Delta x_{3a}^{(k)} = D_{34a} [r_{4a}^{(k)} + H_{42a} rb_{2a}^0])$

$$x_{3a}^{(k+1)} = x_{3a}^{(k)} + \Delta x_{3a}^{(k)}$$

ROLI8  $(\Delta x_{1a} = H_{13a} \Delta x_{3a}^{(k)})$

$$x_{1a} = b_{1a} + H_{13a} x_{3a}^{(k+1)}$$

ROLI8.1 For  $\star = b$  to  $p$  repeat

$$(\Delta x_{3\star}^{(k)} = \Delta x_{3\star}^0 - F_{31\star} [E_{11\star} \Delta x_{1\star-1}])$$

$$\Delta x_{1\star} = [E_{11\star} \Delta x_{1\star-1}] + H_{13\star} \Delta x_{3\star}^{(k)}$$

The vector  $E_{11\star} \Delta x_{1\star-1} = 0$  when  $\star$  corresponds to any row of  $C_{11}$  with only an I

$$x_{3\star}^{(k+1)} = x_{3\star}^{(k)} + \Delta x_{3\star}^{(k)}$$

$$x_{1\star} = [E_{11\star} x_{1\star-1}] + b_{1\star} + H_{13\star} x_{3\star}^{(k+1)}$$

The vector  $E_{11\star} x_{1\star-1} = 0$  when  $\star$  corresponds to any row of  $C_{11}$  with only an I

LOOP End

The modified "Parallel Open Loop algorithm with Iterative refinement" (POLI) follows as:

### PARALLEL OPEN LOOP ALGORITHM WITH ITERATIVE REFINEMENT

POLI1 Evaluate the components of  $A_{21}$ ,  $H_{13}$ ,  $H_{42}$ ,  $b_1$ ,  $b_2$ ,  $b_4$ ,  $C_{11}$  and  $C_{22}$

POLI2.0 Bypass steps POLI2.1, POLI2.2 and POLI3 unless convergence rate is slow

POLI2.1 Initialize  $A_{21}^0 = A_{21}$

POLI2.2 For  $\star = p$  to  $a$  repeat

$$(B_{41\star} = H_{42\star} A_{21\star}^0)$$

$$B_{23\star} = A_{21\star}^0 H_{13\star}$$

$$D_{34\star} = [B_{41\star} H_{13\star}]^{-1}$$

$$F_{31\star} = D_{34\star} B_{41\star}$$

$$F_{24*} = B_{23*} D_{34*}$$

$$A_{21*}^{\circ} \leftarrow A_{21*}^{\circ} + E_{22*} [A_{21*}^{\circ} - B_{23*} F_{31*}] E_{11*}$$

Skip the last equation when \* corresponds to any row of  $C_{11}$  or column of  $C_{22}$  with only an I since the corresponding submatrix  $E_{11*}$  or  $E_{22*}$  is zero

POLI3 Evaluate

$$L_{33} = I + F_{31} [R_{11} - I] H_{13}$$

$$U_{44} = I + H_{42} [R_{22} - I] F_{24}$$

POLI4 Solve

$$U_{44} y_4 = b_4 + H_{42} R_{22} [b_2 - A_{21} R_{11} b_1] \text{ for } y_4$$

$$L_{33} x_3^{(k)} = D_{34} y_4 \text{ for } x_3^{(k)}$$

LOOP For  $k = 0, 1, \dots$  do to LOOP End

POLI5 Evaluate

$$x_1 = R_{11} [b_1 + H_{13} x_3^{(k)}]$$

$$x_2 = -R_{22} [b_2 - A_{21} x_1]$$

POLI6 Evaluate

$$r_4^{(k)} = b_4 - H_{42} x_2 \text{ and exit if small}$$

POLI7 Solve

$$U_{44} \Delta y_4 = r_4^{(k)} \text{ for } \Delta y_4$$

$$L_{33} \Delta x_3^{(k)} = D_{34} \Delta y_4 \text{ for } \Delta x_3^{(k)}$$

POLI8 Evaluate

$$x_3^{(k+1)} = x_3^{(k)} + \Delta x_3^{(k)}$$

LOOP End

## 7. CONCLUSIONS

Solving the coupled equations of large, multiply connected systems involves many numerical computations which must be carried out efficiently when the equations are solved many times. Until recently, most general purpose programs have assembled the necessary coefficient matrices and relied on well developed external programs to numerically manipulate and solve the resulting linear equation systems. The need for fast or possibly real time solutions has prompted development of recursive strategies to symbolically uncouple the equations. These recursive algorithms are ideally suited for long sequential systems but not for parallel structures. Furthermore their highly recursive nature precludes effective use of parallel or vector processors. To address the above problems, this paper first showed how these basic methodologies could be obtained from an optimal symbolic block matrix factorization, and it presented a recursive algorithm. From this algorithm, symbolic natural factors of an equivalent generalized coefficient matrix were obtained. It was suggested that iterative refinement would allow some of the more computationally intensive recursive operations to be bypassed or transferred to other computers for parallel processing. In this development, some of the recursive steps were eliminated by using others to generate a natural factorization of the

generalized matrix. The resulting algorithm has computational overhead which can vary from  $O(p)$  for highly parallel to  $O(p^2)$  for serial systems. Exploiting iterative refinement and taking advantage of vectorization and parallel processing can effectively reduce many  $O(p^2)$  problems to  $O(p)$ .

## REFERENCES

1. Wehage, R.A., "Recursive Multibody Dynamics by Symbolic Block Factorization," to appear in the ASME Journal of Mechanisms, Transmissions and Automation in Design.
2. Wehage, R.A., "Solution of Multibody Dynamics Using Natural Factors and Iterative Refinement - Part I: Open Kinematic Loops - Part II: Closed Kinematic Loops," presented at the 1989 Design Automation Conference, September 17-20, 1989. To appear in the ASME Journal of Mechanisms, Transmissions and Automation in Design.
3. Wehage, R.A. and Shabana, A.A., "Application of Generalized Newton-Euler Equations and Recursive Projection Methods to Dynamics of Deformable Multibody Systems" presented at the 1989 Design Automation Conference in Montreal, Canada on September 17-20, 1989. To appear in the ASME Journal of Mechanisms, Transmissions and Automation in Design.
4. Duff, I.S., et al., Direct Methods for Sparse Matrices, Clarendon Press, New York, NY, 1986.
5. Strang, G., Linear Algebra and its Applications, Academic Press, Inc., New York, NY, 1976.

## APPENDIX A

Prove Eqs. 46 and 47. Since the matrices in Eqs. 44 to 47 are nonsingular, it is sufficient to show that

$$L_{33} L_{33}^{-1} = [I + F_{31}[R_{11} - I] H_{13}] [I + F_{31}[C_{11} - I][I + P_{11}[C_{11} - I]]^{-1} H_{13}] = I \quad (A1)$$

and

$$U_{44}^{-1} U_{44} = [I + H_{42}[I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}] [I + H_{42}[R_{22} - I] F_{24}] = I \quad (A2)$$

The following identities

$$R_{11} C_{11} = C_{11} R_{11} = I \quad (3)$$

$$R_{22} C_{22} = C_{22} R_{22} = I \quad (4)$$

$$H_{13} F_{31} = I - P_{11} \quad (25)$$

and

$$F_{24} H_{42} = I - P_{22} \quad (27)$$

will be useful when proving Eqs. A1 and A2. Since the expansion and simplification of Eqs. A1 and A2 requires some tricks and juggling of terms, several intermediate steps have been shown to assist the reader in following the proofs. Keep in mind, as throughout this paper, that the identity matrices appearing in these equations will have different dimensions according to the matrices they appear with. Thus

$$\begin{aligned} L_{33} L_{33}^{-1} &= [I + F_{31}[R_{11} - I] H_{13}] [I + F_{31}[C_{11} - I][I + P_{11}[C_{11} - I]]^{-1} H_{13}] \\ &= I + F_{31}[R_{11} - I] H_{13} + F_{31}[I + [R_{11} - I][I - P_{11}]] [C_{11} - I][I + P_{11}[C_{11} - I]]^{-1} H_{13} \\ &= I + F_{31}[R_{11} - I] H_{13} + F_{31}[R_{11} - R_{11} P_{11} + P_{11}] [C_{11} - I][I + P_{11}[C_{11} - I]]^{-1} H_{13} \\ &= I + F_{31}[R_{11} - I] H_{13} - F_{31}[R_{11} - I][I + P_{11}[C_{11} - I]][I + P_{11}[C_{11} - I]]^{-1} H_{13} \\ &= I + F_{31}[R_{11} - I] H_{13} - F_{31}[R_{11} - I] H_{13} \\ &= I \end{aligned} \quad (A3)$$

and

$$\begin{aligned} U_{44}^{-1} U_{44} &= [I + H_{42}[I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}] [I + H_{42}[R_{22} - I] F_{24}] \\ &= I + H_{42}[R_{22} - I] F_{24} + H_{42}[[I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I] F_{24}] [I + H_{42}[R_{22} - I] F_{24}] \\ &= I + H_{42}[R_{22} - I] F_{24} + H_{42}[I + [C_{22} - I] P_{22}]^{-1} [C_{22} - I][P_{22} - P_{22} R_{22} + R_{22}] F_{24} \\ &= I + H_{42}[R_{22} - I] F_{24} - H_{42}[I + [C_{22} - I] P_{22}]^{-1} [I + [C_{22} - I] P_{22}][R_{22} - I] F_{24} \\ &= I + H_{42}[R_{22} - I] F_{24} - H_{42}[R_{22} - I] F_{24} \\ &= I \end{aligned} \quad (A4)$$

which completes the proof.

## APPENDIX B

Consider the requirements for the coefficient matrix of Eq. 1 repeated below to be nonsingular.

$$\begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13} \\ 0 & H_{42} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_2 \\ b_1 \\ b_4 \end{bmatrix} \quad (1)$$

Let

$$\begin{bmatrix} X & -Y \\ Z & 0 \end{bmatrix} = \begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13} \\ 0 & H_{42} & 0 \end{bmatrix} \quad (B1)$$

where

$$X = \begin{bmatrix} A_{21} & -C_{22} \\ C_{11} & 0 \end{bmatrix} \quad (B2)$$

$$Y = \begin{bmatrix} 0 \\ H_{13} \end{bmatrix} \quad (B3)$$

$$Z = \begin{bmatrix} 0 & H_{42} \end{bmatrix} \quad (B4)$$

$$\begin{bmatrix} X & -Y \\ Z & 0 \end{bmatrix}^{-1} = \begin{bmatrix} X^{-1} - X^{-1} Y [Z X^{-1} Y]^{-1} Z X^{-1} & X^{-1} Y [Z X^{-1} Y]^{-1} \\ -[Z X^{-1} Y]^{-1} Z X^{-1} & [Z X^{-1} Y]^{-1} \end{bmatrix} \quad (B5)$$

and

$$X^{-1} = \begin{bmatrix} 0 & R_{11} \\ -R_{22} & R_{22} A_{21} R_{11} \end{bmatrix} \quad (B6)$$

Substituting Eqs. B1 to B4 and B6 into Eq. B5 gives

$$\begin{bmatrix} A_{21} & -C_{22} & 0 \\ C_{11} & 0 & -H_{13} \\ 0 & H_{42} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} R_{11} H_{13} A_{43}^{-1} H_{42} R_{22} & R_{11} - R_{11} H_{13} A_{43}^{-1} H_{42} R_{22} A_{21} R_{11} & R_{11} H_{13} A_{43}^{-1} \\ -R_{22} + R_{22} A_{21} R_{11} H_{13} A_{43}^{-1} H_{42} R_{22} & R_{22} A_{21} R_{11} - R_{22} A_{21} R_{11} H_{13} A_{43}^{-1} H_{42} R_{22} A_{21} R_{11} & R_{22} A_{21} R_{11} H_{13} A_{43}^{-1} \\ A_{43}^{-1} H_{42} R_{22} & -A_{43}^{-1} H_{42} R_{22} A_{21} R_{11} & A_{43}^{-1} \end{bmatrix} \quad (B7)$$

where  $A_{43}$  is the generalized coefficient matrix  $H_{42} R_{22} A_{21} R_{11} H_{13}$  in Eq. 2. For this inverse to exist, it is necessary and sufficient for  $A_{43}$  to be invertible.

While the inverse in Eq. B7 is of theoretical interest, it has no immediate value. However, block diagonal inverses of the above matrix are the basis for the algorithms developed in this paper. A rather simple, yet extensive, six-element example shown in Fig. 1 is given to illustrate and help explain the algorithms. Note that this example contains no closed loops and illustrates both serial and parallel tree structure. Also, the system is described by a



graph since  $C_2 = C_1^T$  and thus all connected elements communicate in both directions. And finally, the coefficient matrix is assumed to be nonsingular. The following matrices for this example are included to illustrate the structure of Eq. 1

$$C_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \quad (B8)$$

$$C_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (B9)$$

$$C_{11} = \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 \\ -E_{11b} & I & 0 & 0 & 0 & 0 \\ 0 & -E_{11c} & I & 0 & 0 & 0 \\ 0 & -E_{11d} & 0 & I & 0 & 0 \\ -E_{11e} & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & -E_{11f} & I \end{bmatrix} \quad (B10)$$

$$C_{22} = \begin{bmatrix} I & -E_{22a} & 0 & 0 & -E_{22b} & 0 \\ 0 & I & -E_{22c} & -E_{22d} & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & -E_{22e} \\ 0 & 0 & 0 & 0 & 0 & I \end{bmatrix} \quad (B11)$$

$$R_1 = C_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (B12)$$

$$R_2 = C_2^{-1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (B13)$$

$$R_{11} = C_{11}^{-1} = \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 \\ E_{11b} & I & 0 & 0 & 0 & 0 \\ E_{11cb} & E_{11c} & I & 0 & 0 & 0 \\ E_{11db} & E_{11d} & 0 & I & 0 & 0 \\ E_{11e} & 0 & 0 & 0 & I & 0 \\ E_{11fb} & 0 & 0 & 0 & E_{11f} & I \end{bmatrix} \quad (B14)$$

$$R_{22} = C_{22}^{-1} = \begin{bmatrix} 1 & E_{22b} & E_{22bc} & E_{22bd} & E_{22be} & E_{22bf} \\ 0 & 1 & E_{22c} & E_{22d} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & E_{22f} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (\text{B15})$$

$$E_{11cb} = E_{11c} E_{11b}, \text{ etc.} \quad (\text{B16})$$

$$E_{22bc} = E_{22b} E_{22c}, \text{ etc.} \quad (\text{B17})$$

$$A_{21} = \begin{bmatrix} A_{21a} & 0 & 0 & 0 & 0 & 0 \\ 0 & A_{21b} & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{21c} & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{21d} & 0 & 0 \\ 0 & 0 & 0 & 0 & A_{21e} & 0 \\ 0 & 0 & 0 & 0 & 0 & A_{21f} \end{bmatrix} \quad (\text{B18})$$

$$H_{13} = \begin{bmatrix} H_{13a} & 0 & 0 & 0 & 0 & 0 \\ 0 & H_{13b} & 0 & 0 & 0 & 0 \\ 0 & 0 & H_{13c} & 0 & 0 & 0 \\ 0 & 0 & 0 & H_{13d} & 0 & 0 \\ 0 & 0 & 0 & 0 & H_{13e} & 0 \\ 0 & 0 & 0 & 0 & 0 & H_{13f} \end{bmatrix} \quad (\text{B19})$$

$$H_{42} = \begin{bmatrix} H_{42a} & 0 & 0 & 0 & 0 & 0 \\ 0 & H_{42b} & 0 & 0 & 0 & 0 \\ 0 & 0 & H_{42c} & 0 & 0 & 0 \\ 0 & 0 & 0 & H_{42d} & 0 & 0 \\ 0 & 0 & 0 & 0 & H_{42e} & 0 \\ 0 & 0 & 0 & 0 & 0 & H_{42f} \end{bmatrix} \quad (\text{B20})$$

$$x_1 = [ x_{1a}^T \ x_{1b}^T \ x_{1c}^T \ x_{1d}^T \ x_{1e}^T \ x_{1f}^T ]^T \quad (\text{B21})$$

$$x_2 = [ x_{2a}^T \ x_{2b}^T \ x_{2c}^T \ x_{2d}^T \ x_{2e}^T \ x_{2f}^T ]^T \quad (\text{B22})$$

$$x_3 = [ x_{3a}^T \ x_{3b}^T \ x_{3c}^T \ x_{3d}^T \ x_{3e}^T \ x_{3f}^T ]^T \quad (\text{B23})$$

$$b_1 = [ b_{1a}^T \ b_{1b}^T \ b_{1c}^T \ b_{1d}^T \ b_{1e}^T \ b_{1f}^T ]^T \quad (\text{B24})$$

$$b_2 = [ b_{2a}^T \ b_{2b}^T \ b_{2c}^T \ b_{2d}^T \ b_{2e}^T \ b_{2f}^T ]^T \quad (\text{B25})$$

$$b_4 = [ b_{4a}^T \ b_{4b}^T \ b_{4c}^T \ b_{4d}^T \ b_{4e}^T \ b_{4f}^T ]^T \quad (\text{B26})$$

Combining the above equations gives the composite representation of Eq. 1 as

$$\begin{bmatrix}
 A_{21a} & 0 & 0 & 0 & 0 & 0 & -I & E_{21b} & 0 & 0 & E_{21c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & A_{21b} & 0 & 0 & 0 & 0 & 0 & -I & E_{21b} & E_{21c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & A_{21c} & 0 & 0 & 0 & 0 & 0 & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & A_{21d} & 0 & 0 & 0 & 0 & 0 & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & A_{21e} & 0 & 0 & 0 & 0 & 0 & -I & E_{21f} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & A_{21f} & 0 & 0 & 0 & 0 & 0 & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 -E_{11b} & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12b} & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & -E_{11c} & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12c} & 0 & 0 & 0 & 0 & 0 \\
 0 & -E_{11d} & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12d} & 0 & 0 & 0 & 0 \\
 -E_{11e} & 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12e} & 0 & 0 \\
 0 & 0 & 0 & 0 & -E_{11f} & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -H_{12f} & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & H_{13a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13d} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13f} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13g} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13h} & 0 & 0 & 0 & 0 & 0 & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{1b} \\
 x_{1c} \\
 x_{1d} \\
 x_{1e} \\
 x_{1f} \\
 x_{2a} \\
 x_{2b} \\
 x_{2c} \\
 x_{2d} \\
 x_{2e} \\
 x_{2f} \\
 x_{3a} \\
 x_{3b} \\
 x_{3c} \\
 x_{3d} \\
 x_{3e} \\
 x_{3f}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{1a} \\
 b_{1b} \\
 b_{1c} \\
 b_{1d} \\
 b_{1e} \\
 b_{1f} \\
 b_{2a} \\
 b_{2b} \\
 b_{2c} \\
 b_{2d} \\
 b_{2e} \\
 b_{2f} \\
 b_{3a} \\
 b_{3b} \\
 b_{3c} \\
 b_{3d} \\
 b_{3e} \\
 b_{3f}
 \end{bmatrix}
 \tag{B27}$$

which could be symbolically inverted using Eq. B7 (and efficiently too if one takes maximum advantage of recursion). However an optimal block U-L factorization (the element symbols can be kept in naturally occurring order) of Eq. B27 is being sought, so first permute it into the following form

$$\begin{bmatrix}
 A_{21a} & -I & 0 & 0 & E_{21b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 I & 0 & -H_{12a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & H_{13a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & A_{21b} & -I & 0 & 0 & E_{21c} & 0 & 0 & E_{21d} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 -E_{11b} & 0 & 0 & I & 0 & -H_{12b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & H_{13b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & A_{21c} & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -E_{11c} & 0 & 0 & I & 0 & -H_{12c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21d} & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -E_{11d} & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & -H_{12d} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 -E_{11e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21e} & -I & 0 & 0 & E_{21f} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13d} & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21f} & -I & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -E_{11f} & 0 & 0 & -H_{12f} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{13e} & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a} \\
 x_{1b} \\
 x_{2b} \\
 x_{3b} \\
 x_{1c} \\
 x_{2c} \\
 x_{3c} \\
 x_{1d} \\
 x_{2d} \\
 x_{3d} \\
 x_{1e} \\
 x_{2e} \\
 x_{3e} \\
 x_{1f} \\
 x_{2f} \\
 x_{3f}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{1a} \\
 b_{1b} \\
 b_{1c} \\
 b_{2a} \\
 b_{2b} \\
 b_{2c} \\
 b_{3a} \\
 b_{3b} \\
 b_{3c} \\
 b_{4a} \\
 b_{4b} \\
 b_{4c} \\
 b_{5a} \\
 b_{5b} \\
 b_{5c} \\
 b_{6a} \\
 b_{6b} \\
 b_{6c} \\
 b_{6d} \\
 b_{6e}
 \end{bmatrix}
 \tag{B28}$$

Note the six block matrices

$$\begin{bmatrix}
 A_{21*} & -I & 0 \\
 I & 0 & H_{13*} \\
 0 & H_{42*} & 0
 \end{bmatrix}
 (* = a, b, \dots, f)
 \tag{B29}$$

on the diagonal and that each has the same block structure as Eq. 1 itself. In general, if any pair of block matrices is coupled, that pair will be coupled by either one block matrix above the

diagonal, or one block matrix below the diagonal or both. In this example, each will be coupled by both because the system is represented by a graph. Furthermore, every coupling matrix above the diagonal has the same block structure and every coupling matrix below has the same structure. The reader is encouraged to carefully study the structure of Eq. B28 because it is the key to the successful implementation of all highly efficient recursive solution algorithms. For block U-L factorization, the matrices in Eq. B29 (or at least one equivalent block matrix at each stage of the elimination process) become the pivotal matrices and must be invertible.

Before inverting a typical matrix in Eq. B29 it will be instructive to illustrate the standard elimination process in block U-L factorization. Let the matrix equations

$$M x = b \quad (B30)$$

be partitioned into

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (B31)$$

where block matrix  $M_{22}$  is small, nonsingular and invertible. Thus it follows that

$$x_2 = M_{22}^{-1} [b_2 - M_{21} x_1] \quad (B32)$$

and

$$[M_{11} - M_{12} M_{22}^{-1} M_{21}] x_1 = [b_1 - M_{12} M_{22}^{-1} b_2] \quad (B33)$$

Generating the coefficient matrix and the right-hand side of Eq. B33 represents the first step of block U-L factorization. The second step treats Eq. B33 as a new matrix equation where it, in turn, is partitioned similar to Eq. B31 and the process continues until the remaining coefficient matrix in Eq. B33 is easily invertible. At this point, the preprocessing for elimination is complete and the equations generated by Eq. B32 can be used for back substitution.

In this example, the first matrix in Eq. B29 which must be inverted is for element f. Since this matrix has the same block structure as the matrix in Eq. 1, it follows by using the same approach as in Eq. B7 that

$$\begin{bmatrix} A_{211} & -I & 0 \\ I & 0 & -H_{13f} \\ 0 & H_{42f} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} H_{42f} & I - H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} H_{42f} A_{211} & H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} \\ -I + A_{211} H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} H_{42f} & A_{211} - A_{211} H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} H_{42f} A_{211} & A_{211} H_{13f} [H_{42f} A_{211} H_{13f}]^{-1} \\ [H_{42f} A_{211} H_{13f}]^{-1} H_{42f} & -[H_{42f} A_{211} H_{13f}]^{-1} H_{42f} A_{211} & [H_{42f} A_{211} H_{13f}]^{-1} \end{bmatrix} \quad (B34)$$

The matrix in Eq. B34 can be further simplified to

$$\begin{bmatrix} A_{211} & -I & 0 \\ I & 0 & -H_{13f} \\ 0 & H_{42f} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} H_{13f} D_{34f} H_{42f} & I - H_{13f} F_{31f} & H_{13f} D_{34f} \\ -[I - F_{24f} H_{42f}] & A_{211} - B_{23f} F_{31f} & F_{24f} \\ D_{34f} H_{42f} & -F_{31f} & D_{34f} \end{bmatrix} \quad (B35)$$



This enormous product generated exactly one nonzero submatrix  $E_{22f} [A_{21f} - B_{22f} F_{31f}] E_{11f}$ . When added to the remaining submatrix of Eq. B28 to form the equivalent of  $M_{11} - M_{12} M_{22}^{-1} M_{21}$  in Eq. B33, it falls directly on top of the submatrix  $A_{21e}$  of the parent element e. This is a consequence of optimal ordering or pivotal strategy and results in a minimal block matrix fill pattern (in this case zero block fills). Thus the entire operation can be described by the single equation from step ROL2.1

$$A_{21e}^e = A_{21e} + E_{22f} [A_{21f} - B_{22f} F_{31f}] E_{11f} \quad (B42)$$

The differences between Eq. B42 and the last equation in step ROL2.1 are due to the initialization of all  $A_{21*}^e$  to  $A_{21*}$  in step ROL2.0 and the fact that a given parent  $A_{21*}^e$  may eventually accumulate quantities from more than one child, such as elements a and b in this example. Equation B42 can be read as "this step of the elimination process is equivalent to projecting  $A_{21f}$  from child f onto parent e across the interface between the two elements." The quantity  $A_{21f} - B_{22f} F_{31f}$  makes it across the interface and then undergoes a transformation  $E_{22f} [A_{21f} - B_{22f} F_{31f}] E_{11f}$  to match the coordinates of element e. The remaining quantity  $H_{42f} B_{22f} F_{31f} H_{13f} = H_{42f} A_{21f} H_{13f} = D_{34f}^1$  gets projected onto the element interface subspace 3 as the coefficient of  $x_{3f}$  (see Eq. 58). The superscript e is used on the left matrix in Eq. B42 to denote it as an effective quantity because the original was modified by the projection process.

Now the equivalent of computing the quantity  $-M_{12} M_{22}^{-1} b_2$  on the right hand side of Eq. B28 follows as

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & E_{22f} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} H_{13f} D_{34f} H_{42f} & I - H_{13f} F_{31f} & H_{13f} D_{34f} \\ -[I - F_{24f} H_{42f}] & A_{21f} - B_{22f} F_{31f} & F_{24f} \\ D_{34f} H_{42f} & -F_{31f} & D_{34f} \end{bmatrix} \begin{bmatrix} b_{2f} \\ b_{1f} \\ b_{4f} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ E_{22f} [(I - F_{24f} H_{42f}) [b_{2f} - A_{21f} b_{1f}] - F_{24f} b_{4f}] \\ 0 \\ 0 \end{bmatrix} \quad (B43)$$

As above, this enormous product also generates exactly one nonzero subvector,  $E_{22f} [(I - F_{24f} H_{42f}) [b_{2f} - A_{21f} b_{1f}] - F_{24f} b_{4f}]$ . When added to the remaining subvector of Eq. B28 to form the equivalent of  $b_1 - M_{12} M_{22}^{-1} b_2$  in Eq. B33, the single term falls directly on top of the subvector  $b_{2e}$  of the parent e. This again is a consequence of the optimal ordering strategy and the entire operation is described by the following two equations from step ROL3.1

$$x_{3f}^{\circ} = D_{24f} [b_{2f} + H_{42f} b_{2f}] - F_{31f} b_{1f} \quad (B44)$$

and

$$b_{2e}^{\circ} = b_{2e} + E_{22e} [b_{2f} - A_{21f} [b_{1f} + H_{13f} x_{3f}^{\circ}]] \quad (B45)$$

The reader can verify by substitution that these two equations are equivalent to the right hand side of Eq. B33. The intermediate step was introduced to reduce the computations necessary for step ROL4.1 of the algorithm. Again, the differences between Eq. B45 and the second equation in step ROL3.1 are due to the initialization  $b_2^{\circ} = b_2$  in step ROL3.0 and the need to accumulate projected quantities from children into parent  $b_{2e}^{\circ}$ . As above, Eqs. B44 and B45 can be read as "this step of the elimination process is equivalent to projecting the right hand side quantities from child f onto parent e across the communicating interface between the two elements." The quantity  $[I - F_{24f} H_{42f}] [b_{2f} - A_{21f} b_{1f}] - F_{24f} b_{4f}$  makes it across the interface and then undergoes a transformation  $E_{22e} [[I - F_{24f} H_{42f}] [b_{2f} - A_{21f} b_{1f}] - F_{24f} b_{4f}]$  to match the coordinates of element e. And as above, the e superscripts in Eqs. B44 and B45 denote these terms as effective quantities.

Now that these computations have been completed, the reduced matrix corresponding to Eq. B33 which fits the mold of Eq. B30, can again be partitioned according to Eq. B31 as

$$\begin{bmatrix}
 A_{21a} & -I & 0 & 0 & E_{22a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & E_{22a} & 0 \\
 I & 0 & -H_{13a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & H_{42a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & A_{21b} & -I & 0 & 0 & E_{22b} & 0 & 0 & E_{22b} & 0 & 0 & 0 & 0 \\
 -E_{11b} & 0 & 0 & I & 0 & -H_{13b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & H_{42b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & A_{21c} & -I & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -E_{11c} & 0 & 0 & I & 0 & -H_{13c} & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42c} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21d} & -I & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -E_{11d} & 0 & 0 & 0 & 0 & 0 & I & 0 & -H_{13d} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42d} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21e} & -I & 0 \\
 -E_{11e} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I & 0 & -H_{13e} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42e} & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a} \\
 x_{1b} \\
 x_{2b} \\
 x_{3b} \\
 x_{1c} \\
 x_{2c} \\
 x_{3c} \\
 x_{1d} \\
 x_{2d} \\
 x_{3d} \\
 x_{1e} \\
 x_{2e} \\
 x_{3e}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{2a} \\
 b_{1a} \\
 b_{4a} \\
 b_{2b} \\
 b_{1b} \\
 b_{4b} \\
 b_{2c} \\
 b_{1c} \\
 b_{4c} \\
 b_{2d} \\
 b_{1d} \\
 b_{4d} \\
 b_{2e}^{\circ} \\
 b_{1e} \\
 b_{4e}
 \end{bmatrix}
 \quad (B46)$$

Element f and all quantities associated with it have been completely eliminated from the reduced system of equations with the introduction of the two modified terms  $A_{21e}^{\circ}$  and  $b_{2e}^{\circ}$  in Eq. B46. This new equation indicates that child e is connected to parent a so the next elimination step will project the modified child e onto its parent a. These steps are summarized as follows

$$B_{410} = H_{420} A_{210}^{\circ} \quad (B47)$$

$$B_{230} = A_{210}^{\circ} H_{130} \quad (B48)$$

$$D_{340} = [B_{410} H_{130}]^{-1} \quad (B49)$$

$$F_{310} = D_{340} B_{410} \quad (B50)$$

$$F_{240} = B_{230} D_{340} \quad (B51)$$

$$\begin{bmatrix} A_{210}^{\circ} & -I & 0 \\ I & 0 & -H_{130} \\ 0 & H_{420} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} H_{130} D_{340} H_{420} & I - H_{130} F_{310} & H_{130} D_{340} \\ -[I - F_{240} H_{420}] & A_{210}^{\circ} - B_{230} F_{310} & F_{240} \\ D_{340} H_{420} & -F_{310} & D_{340} \end{bmatrix} \quad (B52)$$

$$\begin{bmatrix} 0 & E_{220} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} H_{130} D_{340} H_{420} & I - H_{130} F_{310} & H_{130} D_{340} \\ -[I - F_{240} H_{420}] & A_{210}^{\circ} - B_{230} F_{310} & F_{240} \\ D_{340} H_{420} & -F_{310} & D_{340} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -E_{110} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} =$$

$$\begin{bmatrix} E_{220} [A_{210}^{\circ} - B_{230} F_{310}] E_{110} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (B53)$$

$$A_{210}^{\circ} = A_{210} + E_{220} [A_{210}^{\circ} - B_{230} F_{310}] E_{110} \quad (B54)$$



$$\begin{bmatrix} 0 & E_{22a} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} H_{13a} & D_{34a} & H_{42a} & I - H_{13a} & F_{31a} & H_{13a} & D_{34a} \\ -[I - F_{24a} & H_{42a}] & A_{21a}^{\circ} - B_{23a} & F_{31a} & F_{24a} \\ D_{34a} & H_{42a} & -F_{31a} & D_{34a} \end{bmatrix} \begin{bmatrix} b_{2a}^{\circ} \\ b_{1a} \\ b_{4a} \end{bmatrix} = \begin{bmatrix} E_{22a} [(I - F_{24a} & H_{42a}) [b_{2a}^{\circ} - A_{21a}^{\circ} & b_{1a}] - F_{24a} & b_{4a}] \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{B55}$$

$$x_{3a}^{\circ} = D_{34a} [b_{4a} + H_{42a} b_{2a}^{\circ}] - F_{31a} b_{1a} \tag{B56}$$

and

$$b_{2a}^{\circ} = b_{2a} + E_{22a} [b_{2a}^{\circ} - A_{21a}^{\circ} [b_{1a} + H_{13a} x_{3a}^{\circ}]] \tag{B57}$$

Now the new reduced matrix corresponding to Eq. B33 takes the following form

$$\begin{bmatrix} A_{21a}^{\circ} & -I & 0 & 0 & E_{22a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I & 0 & -H_{13a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & H_{42a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & A_{21b} & -I & 0 & 0 & E_{22b} & 0 & 0 & E_{22d} & 0 \\ -E_{11b} & 0 & 0 & I & 0 & -H_{13b} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & H_{42b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & A_{21c} & -I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -E_{11c} & 0 & 0 & I & 0 & -H_{13c} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42c} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & A_{21d} & -I & 0 \\ 0 & 0 & 0 & -E_{11d} & 0 & 0 & 0 & 0 & 0 & I & 0 & -H_{13d} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42d} & 0 \end{bmatrix} \begin{bmatrix} x_{1a} \\ x_{2a} \\ x_{3a} \\ x_{1b} \\ x_{2b} \\ x_{3b} \\ x_{1c} \\ x_{2c} \\ x_{3c} \\ x_{1d} \\ x_{2d} \\ x_{3d} \end{bmatrix} = \begin{bmatrix} b_{2a}^{\circ} \\ b_{1a} \\ b_{4a} \\ b_{2b} \\ b_{1b} \\ b_{4b} \\ b_{2c} \\ b_{1c} \\ b_{4c} \\ b_{2d} \\ b_{1d} \\ b_{4d} \end{bmatrix} \tag{B58}$$

Elements e and f have been completely eliminated and only the two new terms  $A_{21a}^{\circ}$  and  $b_{2a}^{\circ}$  were generated. For reference, the remaining reduced matrices are shown without the intermediate symbolic steps

$$\begin{bmatrix}
 A_{21a}^{\circ} & -I & 0 & 0 & E_{22b} & 0 & 0 & 0 & 0 \\
 I & 0 & -H_{13a} & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & H_{42a} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & A_{21b}^{\circ} & -I & 0 & 0 & E_{22c} & 0 \\
 -E_{11b} & 0 & 0 & I & 0 & -H_{13b} & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & H_{42b} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & A_{21c} & -I & 0 \\
 0 & 0 & 0 & -E_{11c} & 0 & 0 & I & 0 & -H_{13c} \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_{42c} & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a} \\
 x_{1b} \\
 x_{2b} \\
 x_{3b} \\
 x_{1c} \\
 x_{2c} \\
 x_{3c}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{2a}^{\circ} \\
 b_{1a} \\
 b_{4a} \\
 b_{2b}^{\circ} \\
 b_{1b} \\
 b_{4b} \\
 b_{2c} \\
 b_{1c} \\
 b_{4c}
 \end{bmatrix}
 \quad (B59)$$

$$\begin{bmatrix}
 A_{21a}^{\circ} & -I & 0 & 0 & E_{22b} & 0 \\
 I & 0 & -H_{13a} & 0 & 0 & 0 \\
 0 & H_{42a} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & A_{21b}^{\circ} & -I & 0 \\
 -E_{11b} & 0 & 0 & I & 0 & -H_{13b} \\
 0 & 0 & 0 & 0 & H_{42b} & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a} \\
 x_{1b} \\
 x_{2b} \\
 x_{3b}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{2a}^{\circ} \\
 b_{1a} \\
 b_{4a} \\
 b_{2b}^{\circ} \\
 b_{1b} \\
 b_{4b}
 \end{bmatrix}
 \quad (B60)$$

$$\begin{bmatrix}
 A_{21a}^{\circ} & -I & 0 \\
 I & 0 & -H_{13a} \\
 0 & H_{42a} & 0
 \end{bmatrix}
 \begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a}
 \end{bmatrix}
 =
 \begin{bmatrix}
 b_{2a}^{\circ} \\
 b_{1a} \\
 b_{4a}
 \end{bmatrix}
 \quad (B61)$$

At this point, one might envision, as each additional element is eliminated, that the recursive block elimination process is like folding or collapsing one leaf or child element at a time onto its parent to form new equivalent leaves or to completely eliminate branches. Eventually only a single equivalent leaf (or the equivalent root) remains, namely Eq. B61 in this example. Now the matrix in Eq. B61 can be inverted yielding

$$\begin{bmatrix}
 x_{1a} \\
 x_{2a} \\
 x_{3a}
 \end{bmatrix}
 =
 \begin{bmatrix}
 H_{13a} D_{34a} H_{42a} & I - H_{13a} F_{31a} & H_{13a} D_{34a} \\
 -[I - F_{24a} H_{42a}] & A_{21a}^{\circ} - B_{22a} F_{31a} & F_{24a} \\
 D_{34a} H_{42a} & -F_{31a} & D_{34a}
 \end{bmatrix}
 \begin{bmatrix}
 b_{2a}^{\circ} \\
 b_{1a} \\
 b_{4a}
 \end{bmatrix}
 \quad (B62)$$

If one looks back to Eq. 1 or Eqs. 5 and 6, it will be apparent that the components of  $x_{1a}$  and  $x_{2a}$  can be more efficiently obtained from

$$C_{11} x_1 = b_1 + H_{13} x_3 \quad (1a)$$

and

$$C_{22} x_2 = -[b_2 - A_{21} x_1] \quad (1b)$$

or

$$x_1 = R_{11} [b_1 + H_{13} x_3] \quad (5)$$

and

$$x_2 = -F_{22}[b_2 - A_{21} x_1] \quad (6)$$

Thus only the last equation from Eq. B62 is required from this step giving

$$x_{3a} = D_{34a}[b_{4a} + H_{42a} b_{2a}^0] - F_{31a} b_{1a} \quad (B63)$$

which is just the final step ROL3.2 of the recursive algorithm. Now, substituting this result into the first equation of Eq. 1a gives the first step in ROL4 of the recursive algorithm

$$x_{1a} = b_{1a} + H_{13a} x_{3a} \quad (B64)$$

To solve for the remaining unknowns requires a visit to Eq. B32, that is  $x_2 = M_{22}^{-1}[b_2 - M_{21} x_1]$ , for the back (forward in this example) substitution step of the block U-L factorization algorithm. The first step is complete with Eqs. B63 and B64. The next step uses the partitioned Eq. B60 to form, according to Eq. B32

$$\begin{bmatrix} x_{1b} \\ x_{2b} \\ x_{3b} \end{bmatrix} = \begin{bmatrix} H_{13b} D_{34b} H_{42b} & I - H_{13b} F_{31b} & H_{13b} D_{34b} \\ -[I - F_{24b} H_{42b}] & A_{21b}^0 - B_{23b} F_{31b} & F_{24b} \\ D_{34b} H_{42b} & -F_{31b} & D_{34b} \end{bmatrix} \begin{bmatrix} b_{2b}^0 \\ b_{1b} \\ b_{4b} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ -E_{11b} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{1a} \\ x_{2a} \\ x_{3a} \end{bmatrix} \quad (B65)$$

which yields the necessary equation

$$x_{3b} = D_{34b}[b_{4b} + H_{42b} b_{2b}^0] - F_{31b} b_{1b} - F_{31b} E_{11b} x_{1a} \quad (B66)$$

However, if one refers back, for example, to Eq. B44 with b substituted for f, it follows that the term

$$x_{3b}^0 = D_{34b}[b_{4b} + H_{42b} b_{2b}^0] - F_{31b} b_{1b} \quad (B67)$$

was computed earlier in the recursive elimination steps so Eq. B66 can be modified as

$$x_{3b} = x_{3b}^0 - F_{31b}[E_{11b} x_{1a}] \quad (B68)$$

Now, the second block equation from Eq. 1a yields

$$x_{1b} = b_{1b} + [E_{11b} x_{1a}] + H_{13b} x_{3b} \quad (B69)$$

Equation B66 defines the first step ROL4 and Eqs. B68 and B69 define the equations for recursive steps ROL4.1. One more step will be developed to more clearly illustrate this part of the algorithm. Starting with the reduced equation, Eq. B59, the next substitution step based on Eq. 32 is

$$\begin{bmatrix} x_{1c} \\ x_{2c} \\ x_{3c} \end{bmatrix} = \begin{bmatrix} H_{13c} D_{34c} H_{42c} & I - H_{13c} F_{31c} & H_{13c} D_{34c} \\ -[I - F_{24c} H_{42c}] & A_{21c} - B_{23c} F_{31c} & F_{24c} \\ D_{34c} H_{42c} & -F_{31c} & D_{34c} \end{bmatrix} \begin{bmatrix} b_{2c} \\ b_{1c} \\ b_{4c} \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -E_{11b} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{1a} \\ x_{2a} \\ x_{3a} \\ x_{1b} \\ x_{2b} \\ x_{3b} \end{bmatrix} \quad (B70)$$

which yields

$$x_{3c} = x_{3c}^0 - F_{31c} [E_{11c} x_{1b}] \quad (B71)$$

where

$$x_{3c}^0 = D_{34c} [b_{4c} + H_{42c} b_{2c}] - F_{31c} b_{1c} \quad (B72)$$

was evaluated during the elimination step. Finally, the remaining equation

$$x_{1c} = b_{1c} + [E_{11c} x_{1b}] + H_{13c} x_{3c} \quad (B73)$$

comes from the third block equation of Eq. 1a. This process essentially repeats in the reverse order of the elimination steps until all of the leaves have been visited. Analogous to the earlier discussions and noting that the matrix equations at each step of the substitution expand to include one additional element, one can envision substitution as equivalent to unfolding the previously collapsed tree one leaf at a time in the reverse order of its folding until the tree has been completely returned to its original configuration.

Note that the above steps did not evaluate the unknown quantities  $x_2$ . If they are desired, the remaining step ROL5 based on Eq. 1b or Eq. 6 can be used to compute them.

Finally, the natural symbolic factors of the generalized matrix in step POL3 are evaluated to give a better understanding of the equation structure in the POL algorithm.

$$L_{33} = I + F_{31} [R_{11} - I] H_{13}$$

$$= \begin{bmatrix} I & 0 & 0 & 0 & 0 & 0 \\ F_{31b} E_{11b} H_{13a} & I & 0 & 0 & 0 & 0 \\ F_{31c} E_{11cb} H_{13a} & F_{31c} E_{11c} H_{13b} & I & 0 & 0 & 0 \\ F_{31d} E_{11db} H_{13a} & F_{31d} E_{11d} H_{13b} & 0 & I & 0 & 0 \\ F_{31e} E_{11e} H_{13a} & 0 & 0 & 0 & I & 0 \\ F_{31f} E_{11fe} H_{13a} & 0 & 0 & 0 & F_{31f} E_{11f} H_{13e} & I \end{bmatrix}$$

(B74)

and

$$U_{44} = I + H_{42} [R_{22} - I] F_{24}$$

$$= \begin{bmatrix} I & H_{42a} E_{22b} F_{24b} & H_{42a} E_{22bc} F_{24c} & H_{42a} E_{22bd} F_{24d} & H_{42a} E_{22e} F_{24e} & H_{42a} E_{22ef} F_{24f} \\ 0 & I & H_{42b} E_{22c} F_{24c} & H_{42b} E_{22d} F_{24d} & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & H_{42e} E_{22f} F_{24f} \\ 0 & 0 & 0 & 0 & 0 & I \end{bmatrix}$$

(B75)

# APPLICATION OF GENERALIZED NEWTON-EULER EQUATIONS AND RECURSIVE PROJECTION METHODS TO DYNAMICS OF DEFORMABLE MULTIBODY SYSTEMS

Roger A. Wehage  
System Simulation and Technology Division  
United States Army Tank-Automotive Command  
Warren, Michigan 48397-5000

Ahmed A. Shabana  
Department of Mechanical Engineering  
University of Illinois at Chicago  
P.O. Box 4348  
Chicago, Illinois 60680

## ABSTRACT

A general symbolic-based method is presented for solving equations of motion for open-loop kinematic chains consisting of interconnected rigid and deformable bodies. The method utilizes matrix partitioning, recursive projection based on optimal block U-L factorization and generalized Newton-Euler equations to obtain an order  $n$  solution for the constrained equations of motion. Kinematic relationships between the absolute reference, joint and elastic coordinates are used with the generalized Newton-Euler equations for deformable bodies to obtain a large, loosely coupled system of equations. Taking advantage of the inertia matrix structure associated with elastic coordinates yields a recursive solution algorithm whose dimension is independent of the elastic degrees of freedom. The above solution techniques applied to this system of equations yield a much smaller operations count and can more effectively exploit vectorization and parallel processing. The algorithms presented in this paper are illustrated with the aid of cylindrical joints which are easily extended to revolute, prismatic, rigid and other joint types.

## 1. INTRODUCTION

Various techniques for the dynamic analysis of constrained mechanical systems consisting of interconnected rigid and deformable bodies have been reported in the literature. The resulting algorithms can be roughly divided into two main categories depending on the set of coordinates used to derive the kinematic and dynamic equations. The first category employs relative joint coordinates, eliminates constraint reaction forces and yields the smallest, most strongly coupled system of equations. Absolute coordinates and joint reaction forces are used to formulate the dynamic equations of motion in the second category. This approach yields relatively large, moderately coupled systems of equations. However, the exclusive use of absolute coordinates introduces complexities in implementing control algorithms, because the joint variables are not readily available when solving the equations of motion. Furthermore, many of the algorithms implementing this approach require the use of Newton-Raphson iteration to correct for constraint violations.

Multibody mechanical system algorithms generally employ joint models defining topological networks of coupled equations which must be solved by matrix and numerical methods. Featherstone [1] presented a method for calculating the acceleration of a robot in response to given actuator forces. His method is applicable to open-loop chains containing rigid bodies, and revolute and prismatic joints. In this work, he developed an algorithm based on recursive formulas involving quantities called articulated-body inertias which represent the effective inertia properties of multiple rigid bodies. Wehage [2-4] extended and generalized

Featherstone's algorithm by developing a general method for obtaining an order  $n$  solution for arbitrary constrained equations of motion by applying matrix partitioning and recursive projection techniques. He also showed that the recursive algorithms are essentially the result of optimal block U-L factorization applied to the composite inertia coefficient matrix. The joint kinematics, equations of motion and topology of a mechanical system are represented in factored matrix form resulting in a large system of loosely coupled equations amenable to sparse matrix manipulation. Optimal matrix permutation, partitioning and recursive projection techniques are then applied to symbolically unravel and lay out an order  $n$  solution strategy which follows the natural topological profile of the system. The method can be applied to arbitrary open and closed-loop systems in order to generate the necessary uncoupled equations [4].

In an earlier work, Armstrong [5] developed a recursive inertia projection algorithm for robotic systems composed of spherical joints. Stepanenko and Vukobratovic [6] gave explicit procedures for computer generation and integration of the equations of motion using Newton-Euler equations. Orin [7] proposed a number of improvements on the scheme of Stepanenko. The success of Newton-Euler equations applied to recursive robotic manipulator dynamics is attributed to their simplicity, and the ability to express them in closed form. A typical set of recursive kinematic equations can be obtained by starting at an arbitrary link at the end of the kinematic tree and moving inward toward the base. These kinematic relationships along with the Newton-Euler equations yield, by simple matrix products, a compact set of symbolic equations in terms of the joint variables.

In this paper, a general symbolic-based method is developed for solving the equations of motion for mechanical systems consisting of interconnected rigid and deformable bodies. The method utilizes matrix partitioning, recursive projection [2-4] and generalized Newton-Euler equations [8]. The absolute or reference coordinates of each deformable body in the system are expressed in terms of body joint and elastic coordinates. The resulting equations of motion employing absolute coordinates and based on the above-mentioned generalized Newton-Euler equations, contain the nonlinear inertia coupling between the so-called rigid body or reference motion and the small elastic deformations. A significant portion of these equations can be expressed in terms of time-invariant quantities which depend on the assumed displacement field. The kinematic relationships and generalized Newton-Euler equations yield a large system of loosely coupled equations amenable to sparse matrix manipulation. Direct methods employing optimal numerical block U-L factorization for manipulating sparse matrices [9, 10] have been successfully applied to equations of this type, but the overhead of numerical matrix structure analysis can be excessive. This problem is circumvented here by employing optimal symbolic U-L factorization to develop equations which recursively yield the absolute and relative accelerations, and the joint reaction forces. This method requires the inversion or decomposition of relatively small matrices and the numerical integration of a minimum number of coordinates. In those algorithms which use absolute coordinates exclusively, Newton-Raphson iteration is often employed to correct for constraint violations. This technique generally leads to numerical and convergence problems. The method in this paper avoids the use of Newton-Raphson iteration and can easily be implemented on the digital computer.

## 2. RECURSIVE KINEMATIC EQUATIONS

Figure 1 shows two deformable bodies labeled  $i-1$  and  $i$ , and connected by a cylindrical joint. Reference coordinate systems  $X^{i-1}Y^{i-1}Z^{i-1}$  and  $X^iY^iZ^i$  with origins  $O^{i-1}$  and  $O^i$  are introduced to define absolute displacement relative to a global frame. Let global reference position vectors  $R^{i-1}$  and  $R^i$  locate the respective origins. For convenience in describing the

connecting joint, introduce intermediate body-fixed joint coordinate systems  $X_d^{i-1} Y_d^{i-1} Z_d^{i-1}$  and  $X_p^i Y_p^i Z_p^i$  at the joint definition points as also shown in Fig. 1. These intermediate joint coordinate systems are assumed to experience small displacements (due to body deformation) with respect to the reference and other coordinate systems fixed on the same body. Large relative displacements between coordinate systems on different bodies (due to joint displacements) are allowed and are described using joint variables  $\theta^{i,i-1}$  (rotation) and  $\tau^{i,i-1}$  (translation).

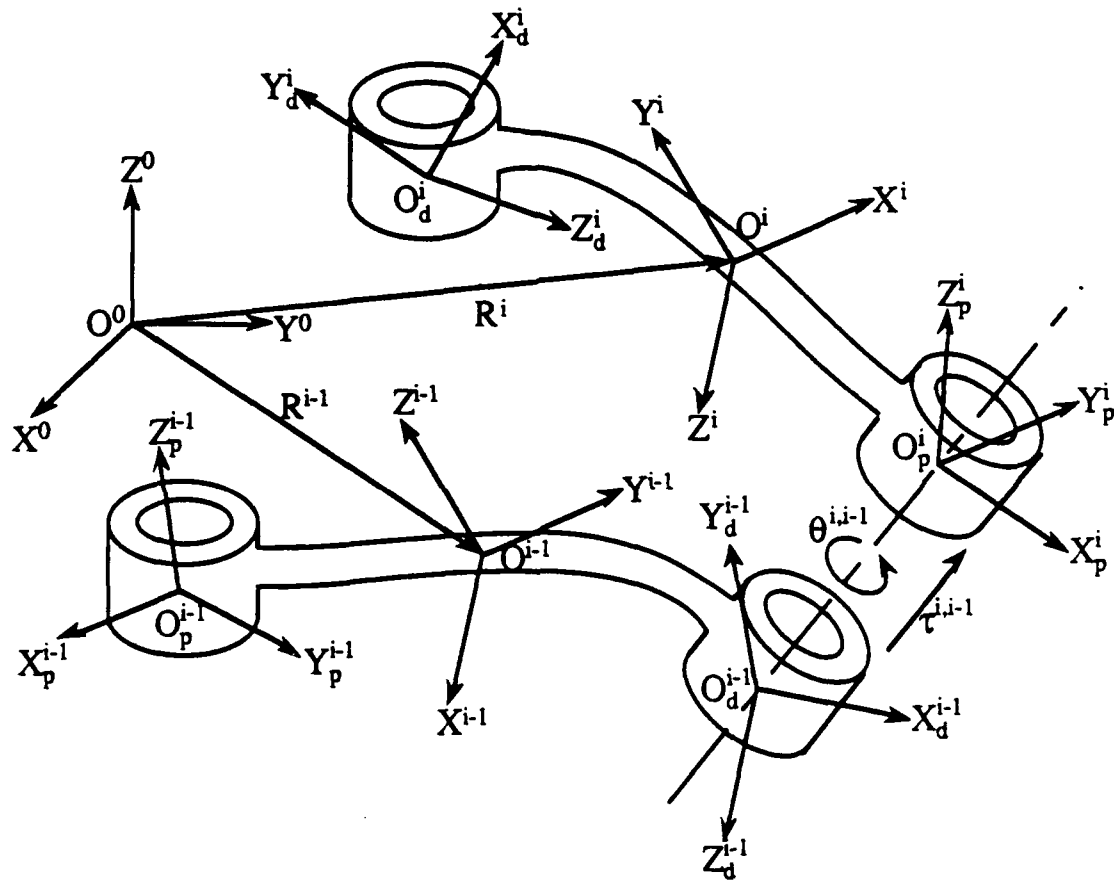


Figure 1. Intermediate coordinate systems

Vectors and matrices can be represented in any coordinate system and throughout this paper, it will be convenient to express them in body reference and global coordinates. Symbols with overbar will denote quantities expressed in global coordinates, otherwise body reference. Let orthonormal matrix  $A^i$  relate global and body reference coordinate systems and vector coordinates as  $\bar{a}^i = A^i a^i$  and  $a^i = A^{iT} \bar{a}^i$ . In this paper, a given vector or matrix associated with body  $i$  will only be expressed in the above two coordinate systems, so additional notation will not be required. The kinematic equations are initially derived in global coordinates and then transformed to body reference coordinates. In general, bold lower and upper case letters denote respective algebraic representation of vectors and matrices. The symbols  $\omega$ ,  $\alpha$  and  $\gamma$  also denote algebraic vector quantities.

**Angular Velocity** The absolute angular velocity of body  $i$  reference coordinate system can be expressed in terms of body  $i-1$  as [8]

$$\bar{\omega}^i = \bar{\omega}^{i-1} + \bar{\omega}_{od}^{i-1} + \bar{v}_{dp}^i \dot{\theta}^{i,i-1} + \bar{\omega}_{po}^i \quad (1)$$

where  $\bar{\omega}^i$  and  $\bar{\omega}^{i-1}$  are the respective angular velocity vectors of body  $i$  and  $i-1$  reference coordinate systems relative to a global inertial frame,  $\bar{\omega}_{od}^{i-1}$  is the intermediate angular velocity of joint coordinate system  $X_d^{i-1} Y_d^{i-1} Z_d^{i-1}$  with respect to body  $i-1$  reference coordinate system,  $\bar{\omega}_{po}^i$  is the intermediate angular velocity of body  $i$  reference coordinate system  $X^i Y^i Z^i$  with respect to the joint coordinate system  $X_p^i Y_p^i Z_p^i$  and  $\bar{v}_{dp}^i = A^i v_{dp}^i$  is a unit vector lying along the joint axis of rotation/translation. The angular velocity vectors  $\bar{\omega}_{od}^{i-1}$  and  $\bar{\omega}_{po}^i$  are the result of small joint coordinate system rotations with respect to the corresponding body reference coordinate systems due to body deformations. It is more efficient to work in body reference coordinate systems because many of the vector and matrix quantities will be constant. In addition, quantities in the body  $i-1$  and  $i$  reference coordinate systems can be related by the matrix  $A^{i,i-1} = A^{iT} A^{i-1}$  or through the basic identity

$$A^{i,i-1} = \left[ I + \bar{v}_{dp}^i \sin \theta^{i,i-1} + 2 \bar{v}_{dp}^{i2} \sin^2 \left( \theta^{i,i-1} / 2 \right) \right] A_0^{i,i-1}$$

where  $A_0^{i,i-1}$  is a constant transformation matrix corresponding to the condition  $\theta^{i,i-1} = 0$  [8].

A skew-symmetric matrix  $\tilde{a}$  equivalent of a vector cross product operator is associated with a [8]. Thus Eq. 1 may be expressed in body  $i-1$  and  $i$  reference coordinates as

$$\omega^i = A^{i,i-1} \left[ \omega^{i-1} + \omega_{od}^{i-1} \right] + v_{dp}^i \dot{\theta}^{i,i-1} + \omega_{po}^i \quad (2)$$

The intermediate angular velocity vectors written in terms of body  $i-1$  and  $i$  elastic coordinates are simply

$$\omega_{od}^{i-1} = S_{\theta_{od}}^{i-1} \dot{q}_f^{i-1} \quad (3)$$

and

$$\omega_{po}^i = S_{\theta_{po}}^i \dot{q}_f^i \quad (4)$$

Constant matrices  $S_{\theta_{od}}^{i-1}$  and  $S_{\theta_{po}}^i$  as defined by Changizi and Shabana [11], depend only on the body shape functions and the relative location of points  $p$  and  $d$  on the bodies. Vectors  $\dot{q}_f^{i-1}$  and  $\dot{q}_f^i$  are the respective body  $i-1$  and  $i$  elastic coordinate derivatives. Throughout this paper, subscripts  $r, j$  and  $f$  will denote respective reference, joint and flexible or elastic coordinates.



**Angular Acceleration** Similar to the above development, one can write the body  $i$  reference angular acceleration vector in terms of body  $i-1$  as

$$\bar{\alpha}^i = \bar{\alpha}^{i-1} + \dot{v}_{dp}^i \ddot{\theta}^{i,i-1} + A^{i-1} S_{\theta od}^{i-1} \ddot{q}_f^{i-1} + A^i S_{\theta po}^i \ddot{q}_f^i + \bar{\gamma}_\theta^i \quad (5)$$

where  $\bar{\alpha}^{i-1}$  and  $\bar{\alpha}^i$  are the respective angular acceleration vectors of body reference coordinate systems  $X^{i-1}Y^{i-1}Z^{i-1}$  and  $X^iY^iZ^i$ . In addition

$$\bar{\gamma}_\theta^i = \ddot{v}_{dp}^i \ddot{\theta}^{i,i-1} + \ddot{\omega}^{i-1} A^{i-1} S_{\theta od}^{i-1} \dot{q}_f^{i-1} + \ddot{\omega}^i A^i S_{\theta po}^i \dot{q}_f^i \quad (6)$$

absorbs components of angular acceleration which are quadratic in first derivatives. The reference accelerations are with respect to the global inertial frame. Equations 5 and 6 may be expressed in body  $i-1$  and  $i$  reference coordinates as

$$\alpha^i = A^{i,i-1} \left[ \alpha^{i-1} + S_{\theta od}^{i-1} \ddot{q}_f^{i-1} \right] + \dot{v}_{dp}^i \ddot{\theta}^{i,i-1} + S_{\theta po}^i \ddot{q}_f^i + \gamma_\theta^i \quad (7)$$

$$\gamma_\theta^i = A^{i,i-1} \ddot{\omega}^{i-1} S_{\theta od}^{i-1} \dot{q}_f^{i-1} + \dot{v}_{dp}^i \ddot{\theta}^{i,i-1} + \ddot{\omega}^i S_{\theta po}^i \dot{q}_f^i \quad (8)$$

or more compactly as

$$\alpha^i = A^{i,i-1} \left[ \alpha^{i-1} + S_{\theta od}^{i-1} \ddot{q}_f^{i-1} \right] + H_{\theta po}^i \ddot{p}^i + \gamma_\theta^i \quad (9)$$

where

$$\ddot{p}^i = \begin{bmatrix} \ddot{\tau}^{i,i-1} & \ddot{\theta}^{i,i-1} & \ddot{q}_f^{i-1} \end{bmatrix}^T = \begin{bmatrix} \ddot{p}_j^i & \ddot{q}_f^i \end{bmatrix}^T \quad (10)$$

and

$$H_{\theta po}^i = \begin{bmatrix} \mathbf{0} & \dot{v}_{dp}^i & S_{\theta po}^i \end{bmatrix} \quad (11)$$

Matrix  $H_{\theta po}^i$  is often called an influence coefficient matrix. Symbols  $\mathbf{I}$  and  $\mathbf{0}$  used in various matrix expressions refer to respective identity and zero matrices whose dimensions are implied by the accompanying matrices and vectors.

**Linear acceleration of body reference origin** The linear acceleration of body  $i$  reference origin  $O^i$  can be written in matrix form as

$$\begin{aligned} \ddot{\mathbf{R}}^i = & \ddot{\mathbf{R}}^{i-1} + \ddot{\alpha}^{i-1} \bar{\mathbf{u}}_{od}^{i-1} + \ddot{\omega}^{i-1} \bar{\omega}^{i-1} \bar{\mathbf{u}}_{od}^{i-1} + 2 \dot{\omega}^{i-1} \dot{\alpha}^{i-1} \bar{\mathbf{u}}_{od}^{i-1} + \ddot{\mathbf{u}}_{od}^{i-1} \\ & + \dot{v}_{dp}^i \bar{\tau}^{i,i-1} + \ddot{\omega}^i \bar{\omega}^i \mathbf{u}_{po}^i + \ddot{\alpha}^i \bar{\mathbf{u}}_{po}^i + 2 \dot{\omega}^i \dot{\alpha}^i \bar{\mathbf{u}}_{po}^i + \ddot{\mathbf{u}}_{po}^i \end{aligned} \quad (12)$$

where  $\bar{\mathbf{u}}_{od}^{i-1}$  is the position vector of joint coordinate system  $X_d^{i-1} Y_d^{i-1} Z_d^{i-1}$  with respect to  $O^{i-1}$  and  $\bar{\mathbf{u}}_{po}^i$  is the position vector of reference coordinate system  $X^i Y^i Z^i$  with respect to intermediate joint coordinate system  $X_p^i Y_p^i Z_p^i$ . Vectors  $\mathbf{u}_{od}^{i-1}$  and  $\mathbf{u}_{po}^i$  can be expressed in terms of respective body  $i-1$  and  $i$  elastic coordinates. In body reference coordinates, Eq. 12 becomes

$$\begin{aligned} \ddot{\mathbf{R}}^i = & \mathbf{A}^{i,i-1} \left[ \ddot{\mathbf{R}}^{i-1} + \ddot{\alpha}^{i-1} \mathbf{u}_{od}^{i-1} + \ddot{\omega}^{i-1} \bar{\omega}^{i-1} \mathbf{u}_{od}^{i-1} + 2 \dot{\omega}^{i-1} \dot{\alpha}^{i-1} \mathbf{u}_{od}^{i-1} + \ddot{\mathbf{u}}_{od}^{i-1} \right] \\ & + v_{dp}^i \bar{\tau}^{i,i-1} + \ddot{\alpha}^i \mathbf{u}_{po}^i + \ddot{\omega}^i \bar{\omega}^i \mathbf{u}_{po}^i + 2 \dot{\omega}^i \dot{\alpha}^i \mathbf{u}_{po}^i + \ddot{\mathbf{u}}_{po}^i \end{aligned} \quad (13)$$

With this knowledge and using Eqs. 1 and 9, Eq. 13 can be written more compactly and in body  $i$  reference coordinates as

$$\ddot{\mathbf{R}}^i + \bar{\mathbf{u}}_{po}^i \ddot{\alpha}^i = \mathbf{A}^{i,i-1} \left[ \ddot{\mathbf{R}}^{i-1} - \bar{\mathbf{u}}_{od}^{i-1} \ddot{\alpha}^{i-1} + \mathbf{S}_{Rod}^{i-1} \ddot{\mathbf{q}}_f^{i-1} \right] + \hat{\mathbf{H}}_{Rpo}^i \ddot{\mathbf{p}}^i + \hat{\gamma}_R^i \quad (14)$$

where

$$\hat{\mathbf{H}}_{Rpo}^i = \begin{bmatrix} v_{dp}^i & 0 & \mathbf{S}_{Rpo}^i \end{bmatrix} \quad (15)$$

is also an influence coefficient matrix and  $\hat{\gamma}_R^i$  absorbs acceleration components which are quadratic in first derivatives.

### 3. KINEMATIC MATRIX EQUATIONS

The first step toward developing recursive kinematic relationships is to express the second derivatives of body  $i$  coordinates explicitly in terms of those of body  $i-1$ . To do this, first combine Eqs. 9 and 14 in matrix form as

$$\begin{bmatrix} \mathbf{I} & \bar{\mathbf{u}}_{po}^i \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \ddot{\mathbf{R}}^i \\ \ddot{\alpha}^i \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{i,i-1} & -\mathbf{A}^{i,i-1} \bar{\mathbf{u}}_{od}^{i-1} & \mathbf{A}^{i,i-1} \mathbf{S}_{Rod}^{i-1} \\ 0 & \mathbf{A}^{i,i-1} & \mathbf{A}^{i,i-1} \mathbf{S}_{\theta od}^{i-1} \end{bmatrix} \begin{bmatrix} \ddot{\mathbf{R}}^{i-1} \\ \ddot{\alpha}^{i-1} \\ \ddot{\mathbf{q}}_f^{i-1} \end{bmatrix} + \begin{bmatrix} \hat{\mathbf{H}}_{Rpo}^i \\ \mathbf{H}_{\theta po}^i \end{bmatrix} \ddot{\mathbf{p}}^i + \begin{bmatrix} \hat{\gamma}_R^i \\ \gamma_\theta^i \end{bmatrix} \quad (16)$$

Then multiply by the coefficient matrix inverse

$$\begin{bmatrix} \mathbf{I} & \tilde{\mathbf{u}}_{po}^i \\ \mathbf{0} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & -\tilde{\mathbf{u}}_{po}^i \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (17)$$

and append the identity

$$\ddot{\mathbf{q}}_f^i = \mathbf{H}_f^i \ddot{\mathbf{p}}^i \quad (18)$$

to yield

$$\begin{bmatrix} \mathbf{R}^i \\ \alpha^i \\ \ddot{\mathbf{q}}_f^i \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{i,i-1} & -[\mathbf{A}^{i,i-1} \tilde{\mathbf{u}}_{od}^{i-1} + \tilde{\mathbf{u}}_{po}^i \mathbf{A}^{i,i-1}] \\ \mathbf{0} & \mathbf{A}^{i,i-1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{H}_{Rod}^{i,i-1} \\ \mathbf{H}_{\theta od}^{i,i-1} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{R}^{i-1} \\ \alpha^{i-1} \\ \ddot{\mathbf{q}}_f^{i-1} \end{bmatrix} + \begin{bmatrix} \mathbf{H}_{Rpo}^i \\ \mathbf{H}_{\theta po}^i \\ \mathbf{H}_f^i \end{bmatrix} \ddot{\mathbf{p}}^i + \begin{bmatrix} \hat{\gamma}_R^i \\ \gamma_\theta^i \\ \mathbf{0} \end{bmatrix} \quad (19)$$

where

$$\begin{bmatrix} \mathbf{I} & -\tilde{\mathbf{u}}_{po}^i \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{i,i-1} & -\mathbf{A}^{i,i-1} \tilde{\mathbf{u}}_{od}^{i-1} & \mathbf{A}^{i,i-1} \mathbf{S}_{Rod}^{i-1} \\ \mathbf{0} & \mathbf{A}^{i,i-1} & \mathbf{A}^{i,i-1} \mathbf{S}_{\theta od}^{i-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{i,i-1} & -[\mathbf{A}^{i,i-1} \tilde{\mathbf{u}}_{od}^{i-1} + \tilde{\mathbf{u}}_{po}^i \mathbf{A}^{i,i-1}] & \mathbf{H}_{Rod}^{i,i-1} \\ \mathbf{0} & \mathbf{A}^{i,i-1} & \mathbf{H}_{\theta od}^{i,i-1} \end{bmatrix} \quad (20)$$

$$\mathbf{H}_{Rod}^{i,i-1} = \mathbf{A}^{i,i-1} \mathbf{S}_{Rod}^{i-1} - \tilde{\mathbf{u}}_{po}^i \mathbf{A}^{i,i-1} \mathbf{S}_{\theta od}^{i-1} \quad (21)$$

$$\mathbf{H}_{\theta od}^{i,i-1} = \mathbf{A}^{i,i-1} \mathbf{S}_{\theta od}^{i-1} \quad (22)$$

$$\mathbf{H}_{Rpo}^i = \hat{\mathbf{H}}_{Rpo}^i - \tilde{\mathbf{u}}_{po}^i \mathbf{H}_{\theta po}^i \quad (23)$$

$$\gamma_R^i = \hat{\gamma}_R^i - \tilde{\mathbf{u}}_{po}^i \gamma_\theta^i \quad (24)$$

and

$$\mathbf{H}_f^i = [\mathbf{0} \quad \mathbf{0} \quad \mathbf{I}] \quad (25)$$

Finally, Eq. 19 may be written more compactly as

$$\mathbf{a}^i = \mathbf{H}_a^{i,i-1} \mathbf{a}^{i-1} + \mathbf{H}_p^i \ddot{\mathbf{p}}^i + \gamma^i \quad (26)$$

where

$$\mathbf{a}^i = \begin{bmatrix} \mathbf{R}^{iT} & \boldsymbol{\alpha}^{iT} & \dot{\mathbf{q}}_f^{iT} \end{bmatrix}^T = \begin{bmatrix} \mathbf{a}_r^{iT} & \dot{\mathbf{q}}_f^{iT} \end{bmatrix}^T \quad (27)$$

$$\mathbf{a}^{i-1} = \begin{bmatrix} \mathbf{R}^{i-1T} & \boldsymbol{\alpha}^{i-1T} & \dot{\mathbf{q}}_f^{i-1T} \end{bmatrix}^T = \begin{bmatrix} \mathbf{a}_r^{i-1T} & \dot{\mathbf{q}}_f^{i-1T} \end{bmatrix}^T \quad (28)$$

$$\mathbf{H}_a^{i,i-1} = \begin{bmatrix} \mathbf{A}^{i,i-1} & -[\mathbf{A}^{i,i-1} \tilde{\mathbf{u}}_{od}^{i-1} + \tilde{\mathbf{u}}_{po}^i \mathbf{A}^{i,i-1}] & \mathbf{H}_{Rod}^{i,i-1} \\ \mathbf{0} & \mathbf{A}^{i,i-1} & \mathbf{H}_{\theta od}^{i,i-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{arr}^{i,i-1} & \mathbf{H}_{arf}^{i,i-1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad (29)$$

$$\mathbf{H}_p^i = \begin{bmatrix} \mathbf{H}_{Rpo}^{iT} & \mathbf{H}_{\theta po}^{iT} & \mathbf{H}_f^{iT} \end{bmatrix}^T = \begin{bmatrix} \mathbf{H}_{prj}^i & \mathbf{H}_{prf}^i \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (30)$$

$$\mathbf{H}_{prj}^i = \begin{bmatrix} \mathbf{v}_{dp}^i & -\tilde{\mathbf{u}}_{po}^i \mathbf{v}_{dp}^i \\ \mathbf{0} & \mathbf{v}_{dp}^i \end{bmatrix} \quad (31)$$

and

$$\mathbf{H}_{prf}^i = \begin{bmatrix} \mathbf{S}_{Rpo}^i & -\tilde{\mathbf{u}}_{po}^i \mathbf{S}_{\theta po}^i \\ \mathbf{S}_{\theta po}^i & \end{bmatrix} \quad (32)$$

**Revolute, Prismatic and Rigid Joints** Formulations for revolute, prismatic and rigid joint kinematic equations, as well as more sophisticated joint types [12] can be obtained as special cases of the cylindrical joint equations. For respective revolute, prismatic or rigid joints,  $\tau^{i,i-1}$ ,  $\theta^{i,i-1}$  or both  $\tau^{i,i-1}$  and  $\theta^{i,i-1}$  are constant.

#### 4. GENERALIZED NEWTON-EULER EQUATIONS OF MOTION

Recently, several formulations have been developed for the dynamic analysis of deformable bodies undergoing large rotations. In this paper, the generalized Newton-Euler equations accounting for all inertia coupling between the reference motion and elastic deformations are used. The generalized Newton-Euler equations presented by [8] in terms of absolute reference and deformation coordinates are given for deformable body  $i$  in its reference frame as

$$\begin{bmatrix} M_{RR}^i & M_{R\theta}^i & M_{Rf}^i \\ M_{\theta R}^i & M_{\theta\theta}^i & M_{\theta f}^i \\ M_{fR}^i & M_{f\theta}^i & M_{ff}^i \end{bmatrix} \begin{bmatrix} \dot{R}^i \\ \alpha^i \\ \dot{q}_f^i \end{bmatrix} = \begin{bmatrix} g_R^i \\ g_\theta^i \\ g_f^i \end{bmatrix} \quad (33)$$

where

$$M_{RR}^i = m^i I \quad (34)$$

$$M_{R\theta}^i = \tilde{m}_c^i \quad (35)$$

$$M_{Rf}^i = \int_{V^i} \rho^i S_f^i dV^i \quad (36)$$

$$M_{\theta\theta}^i = \int_{V^i} \rho^i \tilde{u}^{iT} \tilde{u}^i dV^i \quad (37)$$

$$M_{\theta f}^i = \int_{V^i} \rho^i \tilde{u}^{iT} S_f^i dV^i \quad (38)$$

$$M_{ff}^i = \int_{V^i} \rho^i S_f^{iT} S_f^i dV^i \quad (39)$$

and  $m^i$ ,  $\rho^i$ ,  $V^i$  and  $S_f^i$  are, respectively, body  $i$  total mass, mass density, volume and shape function. Vector  $u^i = u_o^i + u_f^i$  defines the position of any arbitrary point on the deformable body where  $u_o^i$  represents the undeformed position of that point and  $u_f^i = S_f^i q_f^i$  gives the displacement of the point from its undeformed position. The effective mass moment relative to the body reference frame is

$$m_c^i = \int_{V^i} \rho^i u_o^i dV^i + M_{Rf}^i q_f^i \quad (40)$$

The coefficient matrix in Eq. 33 is symmetric and assumed positive definite. The right hand side vectors  $g_R^i$ ,  $g_\theta^i$  and  $g_f^i$  contain externally applied forces and moments, internal elastic and

damping forces and components of inertia forces which are quadratic in first derivatives of the coordinates [8, 13].

Using Eq. 33 and assuming for this discussion that body  $i$  is at the end of a chain of elements so that it contains only one joint common to it and body  $i-1$ , the generalized Newton-Euler equations can be written as

$$\mathbf{M}^i \mathbf{a}^i = \mathbf{g}^i + \mathbf{f}^i \quad (41)$$

where

$$\mathbf{g}^i = \begin{bmatrix} \mathbf{g}_R^{iT} & \mathbf{g}_\theta^{iT} & \mathbf{g}_f^{iT} \end{bmatrix}^T = \begin{bmatrix} \mathbf{g}_r^{iT} & \mathbf{g}_f^{iT} \end{bmatrix}^T \quad (42)$$

and the vector

$$\mathbf{f}^i = \begin{bmatrix} \mathbf{f}_R^{iT} & \mathbf{f}_\theta^{iT} & \mathbf{f}_f^{iT} \end{bmatrix}^T = \begin{bmatrix} \mathbf{f}_r^{iT} & \mathbf{f}_f^{iT} \end{bmatrix}^T \quad (43)$$

contains the internal reaction forces at the joint interface between the two bodies [2, 8, 13].

Using matrix  $\mathbf{H}_p^i$  from Eq. 30, one can write the following equation

$$\mathbf{H}_p^{iT} \mathbf{f}^i = \mathbf{Q}^i \quad (44)$$

where

$$\mathbf{Q}^i = \begin{bmatrix} \mathbf{Q}_j^{iT} & \mathbf{0}^T \end{bmatrix}^T \quad (45)$$

contains the vector of joint generalized forces acting parallel or tangent to the constraint surface. The second equation resulting from Eqs. 44 and 45 yields the dynamic force balance relation

$$\mathbf{H}_{prf}^{iT} \mathbf{f}_r^i + \mathbf{f}_f^i = \mathbf{0} \quad (46)$$

Note that the last part of  $\mathbf{Q}^i$  corresponding to elastic generalized coordinates  $q_f^i$  is zero because all elastic generalized forces were included in  $\mathbf{g}_f^i$ . This arbitrary choice was made to simplify Eq. 46 and will yield the same result as is evident from Eq. 41. Examples of joint generalized forces are actuator forces in prismatic joints, motor torques in revolute joints and friction forces and torques in joints.

## 5. SPARSE MATRIX FORMULATION

In this section, a sparse matrix oriented technique for solving the kinematic and force relationships of the preceding sections is developed. For example, Duff, et al. [10] have shown that optimal block permutation can minimize block matrix fill in U-L factorization which is equivalent to minimizing computational overhead. The purpose of the remaining sections in this paper is to establish patterns which will be applicable to recursive solution of multibody

systems with arbitrary number of bodies. First, Eqs. 26, 41 and 44 can be combined in matrix form to obtain a large, sparse system of equations in terms of the absolute and joint coordinates as

$$\begin{bmatrix} \mathbf{M}^i & -\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{H}_p^i \\ \mathbf{0} & \mathbf{H}_p^{iT} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}^i \\ \mathbf{f}^i \\ \mathbf{p}^i \end{bmatrix} = \begin{bmatrix} \mathbf{g}^i \\ \boldsymbol{\gamma}^i + \mathbf{H}_a^{i,i-1} \mathbf{a}^{i-1} \\ \mathbf{Q}^i \end{bmatrix} \quad (47)$$

If the matrix  $\mathbf{H}_p^i$  has full column rank (which it will have if  $\mathbf{H}_{prj}^i$  has full column rank, see Eq. 30) then Eq. 47 can be solved by block partitioning and using the following basic identities which can be verified by direct matrix multiplication

$$\begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P} \mathbf{A}^{-1} & \mathbf{E} \\ -\bar{\mathbf{E}} & \mathbf{D} \end{bmatrix} \quad (48)$$

$$\mathbf{D} = [\mathbf{B}^T \mathbf{A}^{-1} \mathbf{B}]^{-1} \quad (49)$$

$$\mathbf{E} = \mathbf{A}^{-1} \mathbf{B} \mathbf{D} \quad (50)$$

$$\bar{\mathbf{E}} = \mathbf{D} \mathbf{B}^T \mathbf{A}^{-1} \quad (51)$$

and

$$\mathbf{P} = \mathbf{I} - \mathbf{E} \mathbf{B}^T \quad (52)$$

where  $\mathbf{P}$  is a projection matrix such that  $\mathbf{P}^2 = \mathbf{P}$ .

Setting

$$\mathbf{A} = \begin{bmatrix} \mathbf{M}^i & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \mathbf{H}_p^i \end{bmatrix}$$

in Eq. 47 and noting that

$$\begin{bmatrix} \mathbf{M}^i & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{M}^i \end{bmatrix} \quad (53)$$

it follows that

$$\begin{bmatrix} \mathbf{M}^i & -\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{H}_p^i \\ \mathbf{0} & \mathbf{H}_p^{iT} & \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{G}^i & \mathbf{P}^{iT} & \mathbf{E}^i \\ -\mathbf{P}^i & \mathbf{M}_p^i & \mathbf{F}^i \\ \mathbf{E}^{iT} & -\mathbf{F}^{iT} & \mathbf{D}^i \end{bmatrix} \quad (54)$$

where

$$\mathbf{D}^i = [\mathbf{H}_p^{iT} \mathbf{M}^i \mathbf{H}_p^i]^{-1} \quad (55)$$

$$\mathbf{E}^i = \mathbf{H}_p^i \mathbf{D}^i \quad (56)$$

$$\mathbf{F}^i = \mathbf{M}^i \mathbf{E}^i \quad (57)$$

$$\mathbf{G}^i = \mathbf{E}^i \mathbf{H}_p^{iT} \quad (58)$$

$$\mathbf{P}^i = \mathbf{I} - \mathbf{F}^i \mathbf{H}_p^{iT} \quad (59)$$

$$\mathbf{M}_p^i = \mathbf{P}^i \mathbf{M}^i \quad (60)$$

and

$$\begin{bmatrix} \mathbf{a}^i \\ \mathbf{f}^i \\ \ddots \\ \mathbf{p}^i \end{bmatrix} = \begin{bmatrix} \mathbf{G}^i & \mathbf{P}^{iT} & \mathbf{E}^i \\ -\mathbf{P}^i & \mathbf{M}_p^i & \mathbf{F}^i \\ \mathbf{E}^{iT} & -\mathbf{F}^{iT} & \mathbf{D}^i \end{bmatrix} \begin{bmatrix} \mathbf{g}^i \\ \gamma^i + \mathbf{H}_a^{i,i-1} \mathbf{a}^{i-1} \\ \mathbf{Q}^i \end{bmatrix} \quad (61)$$

Matrix  $\mathbf{P}^i$  is a projection matrix and  $\mathbf{M}_p^i$  is a projected inertia.

Assuming that  $\mathbf{a}^{i-1}$  is known, then Eq. 61 will yield  $\mathbf{a}^i$ ,  $\mathbf{f}^i$  and  $\mathbf{p}^i$ . However, the first two equations from Eq. 61 are not required because it is less expensive to obtain  $\mathbf{a}^i$  and  $\mathbf{f}^i$  directly from Eqs. 26 and 41 once  $\mathbf{p}^i$  is known.



## 6. ALTERNATE MATRIX REPRESENTATION

As pointed out earlier, Eq. 47 yields the largest and least coupled system of equations, where advantage is taken of the identities in Eqs. 48-60. However, some of the intermediate matrix operations involved in evaluating the inverse in Eq. 54 are still quite large. It is possible to obtain a smaller, less sparse equation system by eliminating the second appearance of  $\ddot{q}_f^i$  from  $\ddot{p}^i$  on the left side of Eq. 47 (see Eqs. 10, 18, 25, 30-32, 47 and 61). To this end, write the projection

$$\begin{bmatrix} \ddot{a}^i \\ \ddot{f}^i \\ \ddot{p}^i \end{bmatrix} = \begin{bmatrix} \ddot{a}_r^i \\ \ddot{q}_f^i \\ \ddot{f}_r^i \\ \ddot{f}_f^i \\ \ddot{p}_j^i \\ \ddot{q}_f^i \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \ddot{a}_r^i \\ \ddot{q}_f^i \\ \ddot{f}_r^i \\ \ddot{f}_f^i \\ \ddot{p}_j^i \\ \ddot{q}_f^i \end{bmatrix} \quad (62)$$

Substituting Eq. 62 into Eq. 47, premultiplying by the transpose of the coefficient matrix in Eq. 62, eliminating the resulting null block row and column and permuting gives the following reduced set of equations

$$\begin{bmatrix} \mathbf{M}_{rr}^i & -\mathbf{I} & \mathbf{0} & \mathbf{M}_{rf}^i \\ \mathbf{I} & \mathbf{0} & -\mathbf{H}_{prj}^i & -\mathbf{H}_{prf}^i \\ \mathbf{0} & \mathbf{H}_{prj}^{iT} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_{fr}^i & \mathbf{H}_{prf}^{iT} & \mathbf{0} & \mathbf{M}_{ff}^i \end{bmatrix} \begin{bmatrix} \ddot{a}_r^i \\ \ddot{f}_r^i \\ \ddot{p}_j^i \\ \ddot{q}_f^i \end{bmatrix} = \begin{bmatrix} \mathbf{g}_r^i \\ \mathbf{Y}_r^i + \mathbf{H}_{arr}^{i,i-1} \ddot{a}_r^{i-1} + \mathbf{H}_{arf}^{i,i-1} \ddot{q}_f^{i-1} \\ \mathbf{Q}_j^i \\ \mathbf{g}_f^i \end{bmatrix} \quad (63)$$

where

$$\mathbf{M}_{rr}^i = \begin{bmatrix} \mathbf{M}_{RR}^i & \mathbf{M}_{R\theta}^i \\ \mathbf{M}_{\theta R}^i & \mathbf{M}_{\theta\theta}^i \end{bmatrix} \quad (64)$$

$$\mathbf{M}_{rf}^i = \begin{bmatrix} \mathbf{M}_{Rf}^i \\ \mathbf{M}_{\theta f}^i \end{bmatrix} \quad (65)$$

Observe that the upper left 3 by 3 block of matrices in Eq. 63 has dimension 12 plus the degree of freedom in the joint connecting bodies  $i-1$  and  $i$ , and is nonsingular for well posed problems. This matrix is easily inverted following the steps outlined in Eqs. 48-60.

Furthermore the lower matrix  $M_{ff}^i$  is constant, positive definite and thus all quantities can be evaluated when  $a_r^{i-1}$  is known. In addition, note that the unknown constraint deformation force vector  $f_r^i$  has been eliminated from Eq. 63 but can be evaluated, if desired, from Eq. 46 once  $f_r^i$  has been determined. Use the last equation in Eq. 63 to solve for

$$\ddot{q}_f^i = M_{ff}^{i(-1)} \left[ g_f^i - M_{fr}^i a_r^i - H_{prf}^{iT} f_r^i \right] \quad (66)$$

Note that the symbol  $(-1)$  used above denotes matrix inverse to avoid confusion with the symbol denoting body  $i-1$ . Now substitute into the remaining equations of Eq. 63 and rearrange to get

$$\begin{bmatrix} M_{rr1}^i & -M_{rr0}^{iT} & 0 \\ M_{rr0}^i & M_{rr-1}^i & -H_{prj}^i \\ 0 & H_{prj}^{iT} & 0 \end{bmatrix} \begin{bmatrix} a_r^i \\ f_r^i \\ p_j^i \end{bmatrix} = \begin{bmatrix} g_r^i - M_{ff}^i M_{ff}^{i(-1)} g_f^i \\ \gamma_r^i + H_{arr}^{i,i-1} a_r^{i-1} + H_{arf}^{i,i-1} \ddot{q}_f^{i-1} + H_{prf}^i M_{ff}^{i(-1)} g_f^i \\ Q_j^i \end{bmatrix} \quad (67)$$

where

$$M_{rr1}^i = M_{rr}^i - M_{ff}^i M_{ff}^{i(-1)} M_{fr}^i \quad (68)$$

$$M_{rr0}^i = I + H_{prf}^i M_{ff}^{i(-1)} M_{fr}^i \quad (69)$$

$$M_{rr-1}^i = H_{prf}^i M_{ff}^{i(-1)} H_{prf}^{iT} \quad (70)$$

Let

$$A = \begin{bmatrix} M_{rr1}^i & -M_{rr0}^{iT} \\ M_{rr0}^i & M_{rr-1}^i \end{bmatrix}$$

and

$$B = \begin{bmatrix} 0 \\ H_{prj}^i \end{bmatrix}$$

Then with matrix  $M^i$  positive definite and with full column rank in  $H_{prf}^i$ , it can be shown [14] that matrices  $M_{rr}^i$  and  $M_{rr-1}^i$  are positive definite which guarantees nonsingularity of A. With this, it is straight forward to evaluate a set of matrices similar to Eqs. 55-60 which represent the inverse of the coefficient matrix in Eq. 67 as

$$\begin{bmatrix} M_{rr1}^i & -M_{rr0}^{iT} & 0 \\ M_{rr0}^i & M_{rr-1}^i & -H_{prj}^i \\ 0 & H_{prj}^{iT} & 0 \end{bmatrix}^{-1} = \begin{bmatrix} G_{rr}^i & P_{rr}^{iT} & E_{rj}^i \\ -P_{rr}^i & M_{rrp}^i & F_{rj}^i \\ E_{rj}^{iT} & -F_{rj}^{iT} & D_{jj}^i \end{bmatrix} \quad (71)$$

Similar to Eq. 61, it follows that

$$\begin{bmatrix} a_r^i \\ f_r^i \\ p_j^i \end{bmatrix} = \begin{bmatrix} G_{rr}^i & P_{rr}^{iT} & E_{rj}^i \\ -P_{rr}^i & M_{rrp}^i & F_{rj}^i \\ E_{rj}^{iT} & -F_{rj}^{iT} & D_{jj}^i \end{bmatrix} \begin{bmatrix} g_r^i - M_{rr}^i M_{ff}^{i(-1)} g_f^i \\ \gamma_r^i + H_{arr}^{i,i-1} a_r^{i-1} + H_{arf}^{i,i-1} q_f^{i-1} + H_{prf}^i M_{ff}^{i(-1)} g_f^i \\ Q_j^i \end{bmatrix} \quad (72)$$

where all three equations must be used here because the first two in Eq. 63 depend on  $q_f^i$  which, according to Eq. 66 also depends on  $a_r^i$  and  $f_r^i$ . One might suggest that this problem could have been avoided by first eliminating  $a_r^i$ ,  $f_r^i$  and  $p_j^i$  from Eq. 63. However, this idea was discarded because it requires the repetitive inversion of a much larger matrix the size of  $M_{ff}^i$  in order to evaluate  $q_f^i$  (recall that matrix  $M_{ff}^{i(-1)}$  is constant and must be evaluated only once).

## 7. CONNECTIVITY CONDITIONS AND PROJECTION METHODS

Let body  $i-1$  be located between bodies  $i-2$  and  $i$  in a chain of elements. Then a dynamic equilibrium equation similar to Eq. 41 can be written as

$$M^{i-1} a^{i-1} = g^{i-1} + f^{i-1} - H_a^{i,i-1T} f^i \quad (73)$$

where the transformation  $H_a^{i,i-1T}$  brings the reaction forces at the joint between bodies  $i-1$  and  $i$  to the common reference coordinates and origin of body  $i-1$ . The dynamic equilibrium of bodies  $i$  and  $i-1$  taken together is described by Eqs 41 and 73 combined in matrix form as

$$\begin{bmatrix} M^{i-1} & 0 \\ 0 & M^i \end{bmatrix} \begin{bmatrix} a^{i-1} \\ a^i \end{bmatrix} = \begin{bmatrix} g^{i-1} \\ g^i \end{bmatrix} + \begin{bmatrix} f^{i-1} - H_a^{i,i-1T} f^i \\ f^i \end{bmatrix} \quad (74)$$

Likewise one may extend Eqs. 26 and 44 to obtain

$$\begin{bmatrix} \mathbf{a}^{i-1} \\ \mathbf{a}^i \end{bmatrix} = \begin{bmatrix} \mathbf{H}_a^{i-1,i-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_a^{i,i-1} \end{bmatrix} \begin{bmatrix} \mathbf{a}^{i-2} \\ \mathbf{a}^{i-1} \end{bmatrix} + \begin{bmatrix} \mathbf{H}_p^{i-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_p^i \end{bmatrix} \begin{bmatrix} \mathbf{p}^{i-1} \\ \mathbf{p}^i \end{bmatrix} + \begin{bmatrix} \boldsymbol{\gamma}^{i-1} \\ \boldsymbol{\gamma}^i \end{bmatrix} \quad (75)$$

and

$$\begin{bmatrix} \mathbf{H}_p^{i-1T} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_p^{iT} \end{bmatrix} \begin{bmatrix} \mathbf{f}^{i-1} \\ \mathbf{f}^i \end{bmatrix} = \begin{bmatrix} \mathbf{Q}^{i-1} \\ \mathbf{Q}^i \end{bmatrix} \quad (76)$$

Now Eqs. 74-76 may be combined and permuted into

$$\begin{bmatrix} \mathbf{M}^{i-1} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{H}_a^{i,i-1T} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{H}_p^{i-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_p^{i-1T} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}^i & -\mathbf{I} & \mathbf{0} \\ -\mathbf{H}_a^{i,i-1} & \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & -\mathbf{H}_p^i \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H}_p^{iT} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}^{i-1} \\ \mathbf{f}^{i-1} \\ \mathbf{p}^{i-1} \\ \mathbf{a}^i \\ \mathbf{f}^i \\ \mathbf{p}^i \end{bmatrix} = \begin{bmatrix} \mathbf{g}^{i-1} \\ \boldsymbol{\gamma}^{i-1} + \mathbf{H}_a^{i-1,i-2} \mathbf{a}^{i-2} \\ \mathbf{Q}^{i-1} \\ \mathbf{g}^i \\ \boldsymbol{\gamma}^i \\ \mathbf{Q}^i \end{bmatrix} \quad (77)$$

Following the steps leading to Eqs. 62 and 63, Eq. 77 can also be reduced to

$$\begin{bmatrix} \mathbf{M}_r^{i-1} & -\mathbf{I} & \mathbf{0} & \mathbf{M}_r^{i-1} & \mathbf{0} & \mathbf{H}_{pr}^{i,i-1T} & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{H}_{prj}^{i-1} & -\mathbf{H}_{prf}^{i-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{prj}^{i-1T} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{M}_r^{i-1} & \mathbf{H}_{prf}^{i-1T} & \mathbf{0} & \mathbf{M}_r^{i-1} & \mathbf{0} & \mathbf{H}_{prf}^{i,i-1T} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}_r^i & -\mathbf{I} & \mathbf{0} & \mathbf{M}_r^i \\ -\mathbf{H}_{pr}^{i,i-1} & \mathbf{0} & \mathbf{0} & -\mathbf{H}_{pr}^{i,i-1} & \mathbf{I} & \mathbf{0} & -\mathbf{H}_{prj}^i & -\mathbf{H}_{prf}^i \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H}_{prj}^{iT} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}_r^i & \mathbf{H}_{prf}^{iT} & \mathbf{0} & \mathbf{M}_r^i \end{bmatrix} \begin{bmatrix} \mathbf{a}_r^{i-1} \\ \mathbf{f}_r^{i-1} \\ \mathbf{p}_j^{i-1} \\ \mathbf{q}_r^{i-1} \\ \mathbf{a}_r^i \\ \mathbf{f}_r^i \\ \mathbf{p}_j^i \\ \mathbf{q}_r^i \end{bmatrix} = \begin{bmatrix} \mathbf{g}_r^{i-1} \\ \boldsymbol{\gamma}_r^{i-1} + \mathbf{H}_{pr}^{i-1,i-2} \mathbf{a}_r^{i-2} + \mathbf{H}_{pr}^{i-1,i-2} \mathbf{q}_r^{i-2} \\ \mathbf{Q}_j^{i-1} \\ \mathbf{g}_r^{i-1} \\ \mathbf{g}_r^i \\ \boldsymbol{\gamma}_r^i \\ \mathbf{Q}_j^i \\ \mathbf{g}_r^i \end{bmatrix} \quad (78)$$

The reader is encouraged to carefully study the structure of Eqs. 77 and 78 to identify the minimal coupling between the two major blocks. These equations may be extended to any number of bodies by adding blocks along the major diagonal and corresponding coupling matrices at the row/column intersections corresponding to their joint adjacency. For serial mechanisms, the overall matrix bandwidth will always be the same as in Eqs. 77 and 78. Regardless of the matrix bandwidth (the degree of system serialism or parallelism), the computational overhead per body for open-loop systems will always be the same when the matrix equations have been permuted for optimal U-L factorization. That is, elimination starts in the lower right hand corner and back substitution starts in the upper left hand corner of the composite matrix. To further comprehend the recursive elimination procedure, Eq. 78 is solved for all unknown quantities. This procedure can then be extended to any number of bodies.

Since matrices  $M_{ff}^i$  and  $M_{ff}^{i-1}$  are constant and assumed nonsingular, the accelerations  $\ddot{q}_f^{i-1}$  and  $\ddot{q}_f^i$  are first eliminated non recursively, as in Eq. 66, leaving a system of equations with structure similar to those involving only rigid bodies (refer to Eq. 67). To this end, eliminate  $\ddot{q}_f^{i-1}$  and  $\ddot{q}_f^i$  giving

$$\begin{bmatrix} M_{rr1}^{i-1} & -M_{rr0}^{i-1T} & 0 & 0 & M_{arr0}^{i-1T} & 0 \\ M_{rr0}^{i-1} & M_{rr-1}^{i-1} & -H_{prj}^{i-1} & 0 & M_{arr-1}^{i-1T} & 0 \\ 0 & H_{prj}^{i-1T} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & M_{rr1}^i & -M_{rr0}^{iT} & 0 \\ -M_{arr0}^{i,i-1} & M_{arr-1}^{i,i-1} & 0 & M_{rr0}^i & M_{rr-1}^i & -H_{prj}^i \\ 0 & 0 & 0 & 0 & H_{prj}^{iT} & 0 \end{bmatrix} \begin{bmatrix} a_r^{i-1} \\ f_r^{i-1} \\ p_j^{i-1} \\ a_r^i \\ f_r^i \\ p_j^i \end{bmatrix} = \begin{bmatrix} g_r^{i-1} - M_{ff}^{i-1} M_{ff}^{i-1(-1)} g_f^{i-1} \\ \gamma_r^{i-1} + H_{arr}^{i-1,i-2} a_r^{i-2} + H_{arf}^{i-1,i-2} \ddot{q}_f^{i-2} + H_{prf}^{i-1} M_{ff}^{i-1(-1)} g_f^{i-1} \\ Q_j^{i-1} \\ g_r^i - M_{ff}^i M_{ff}^{i(-1)} g_f^i \\ \gamma_r^i + H_{prf}^i M_{ff}^{i(-1)} g_f^i + H_{arf}^{i,i-1} M_{ff}^{i-1(-1)} g_f^{i-1} \\ Q_j^i \end{bmatrix} \quad (79)$$

where

$$M_{arr0}^{i,i-1} = H_{arr}^{i,i-1} - H_{arf}^{i,i-1} M_{ff}^{i-1(-1)} M_{ff}^{i-1} \quad (80)$$

and

$$M_{arr-1}^{i,i-1} = H_{arf}^{i,i-1} M_{ff}^{i-1(-1)} H_{prf}^{i-1T} \quad (81)$$

The remaining submatrices are obtained from Eqs. 68-70 with  $i$  replaced by  $i-1$ .

Eliminating the unknowns  $\mathbf{a}_r^i$ ,  $\mathbf{f}_r^i$  and  $\mathbf{p}_j^i$  following the procedures outlined earlier yields the reduced system of equations

$$\begin{bmatrix} \mathbf{M}_{rr1}^{ei-1} & -\mathbf{M}_{rr0}^{ei-1T} & \mathbf{0} \\ \mathbf{M}_{rr0}^{ei-1} & \mathbf{M}_{rr-1}^{ei-1} & -\mathbf{H}_{prj}^{i-1} \\ \mathbf{0} & \mathbf{H}_{prj}^{i-1T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}_r^{i-1} \\ \mathbf{f}_r^{i-1} \\ \mathbf{p}_j^{i-1} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_r^{ei-1} - \mathbf{M}_{rr}^{i-1} \mathbf{M}_{ff}^{i-1(-1)} \mathbf{g}_f^{i-1} \\ \gamma_r^{ei-1} + \mathbf{H}_{arr}^{i-1,i-2} \mathbf{a}_r^{i-2} + \mathbf{H}_{arf}^{i-1,i-2} \mathbf{q}_f^{i-2} + \mathbf{H}_{prf}^{i-1} \mathbf{M}_{ff}^{i-1(-1)} \mathbf{g}_f^{i-1} \\ \mathbf{Q}_j^{i-1} \end{bmatrix} \quad (82)$$

where

$$\mathbf{M}_{rr1}^{ei-1} = \mathbf{M}_{rr1}^{i-1} + \mathbf{M}_{arr0}^{i,i-1T} \mathbf{M}_{rrp}^i \mathbf{M}_{arr0}^{i,i-1} \quad (83)$$

$$\mathbf{M}_{rr0}^{ei-1} = \mathbf{M}_{rr0}^{i-1} + \mathbf{M}_{arr-1}^{i,i-1T} \mathbf{M}_{rrp}^i \mathbf{M}_{arr0}^{i,i-1} \quad (84)$$

$$\mathbf{M}_{rr-1}^{ei-1} = \mathbf{M}_{rr-1}^{i-1} - \mathbf{M}_{arr-1}^{i,i-1T} \mathbf{M}_{rrp}^i \mathbf{M}_{arr-1}^{i,i-1} \quad (85)$$

$$\mathbf{g}_r^{ei-1} = \mathbf{g}_r^{i-1} - \mathbf{M}_{arr0}^{i,i-1T} \mathbf{f}_r^{ei} \quad (86)$$

$$\gamma_r^{ei-1} = \gamma_r^{i-1} - \mathbf{M}_{arr-1}^{i,i-1T} \mathbf{f}_r^{ei} \quad (87)$$

and

$$\mathbf{f}_r^{ei} = \begin{bmatrix} -\mathbf{P}_{rr}^i & \mathbf{M}_{rrp}^i & \mathbf{F}_{rj}^i \end{bmatrix} \begin{bmatrix} \mathbf{g}_r^i - \mathbf{M}_{rr}^i \mathbf{M}_{ff}^{i(-1)} \mathbf{g}_f^i \\ \gamma_r^i + \mathbf{H}_{prf}^i \mathbf{M}_{ff}^{i(-1)} \mathbf{g}_f^i + \mathbf{H}_{arf}^{i,i-1} \mathbf{M}_{ff}^{i-1(-1)} \mathbf{g}_f^{i-1} \\ \mathbf{Q}_j^i \end{bmatrix} \quad (88)$$

The superscript "e" used in the above equations means equivalent quantity. Compare the structure of Eqs. 67 and 82. Equations 83-88 clearly show that the elimination process generates equivalent matrix and vector replacement quantities only for the body which holds the eliminated element. That is, properties of the eliminated body are projected across the joint onto its parent. For open kinematic-loop systems and as a consequence of optimal block U-L factorization, each stage of elimination generates a further reduced system of equations whose structure is identical to that of an equivalent system with the corresponding body removed.

Using the procedures developed in this paper, one may generalize Eqs. 66-88 to systems composed of any number of rigid and deformable bodies interconnected by joints. Space limitations do not allow a comprehensive development of recursive solution algorithms for arbitrary systems of interconnected rigid and flexible bodies and this paper does not address the steps required to handle closed kinematic-loop systems [2, 4]. While the coefficient submatrix dimensions in Eqs. 63, 77 and 78 depend on the individual body elastic degrees of freedom, the matrix dimensions in Eqs. 67, 79 and 82 are the same whether bodies are deformable or not. Only the submatrix structures and thus the underlying recursive solution algorithms differ.

## 8. SUMMARY

A method is presented for effective solution of equations of motion for systems of interconnected rigid and deformable bodies. The equations of motion of each body in the system are formulated in terms of the absolute coordinates using generalized Newton-Euler equations. These equations which contain the nonlinear inertia coupling between the rigid body motion and the small elastic deformation are expressed in terms of a set of invariants which depend on the assumed displacement field. Recursive kinematic relationships in which the absolute variables of body  $i$  are expressed in terms of those of body  $i-1$  and joint variables are also developed. The matrix relating absolute and relative coordinates is used to define joint forces which act tangent or parallel to the constraint surfaces. These forces are the generalized forces associated with the joint generalized coordinates. By combining the generalized Newton-Euler equations, the kinematic relationships and the generalized joint force equations, a large system of loosely coupled equations is obtained. Matrix partitioning, optimal block factorization and recursive projection methods can then be employed to obtain an order  $n$  solution for the constrained system equations of motion. The formulation presented in this paper can be applied to arbitrary systems with rigid and flexible elements, and numerous kinematic joint types.

### Acknowledgement

This research was sponsored, in part, by the US Army Research Office, Research Triangle Park, NC.

## REFERENCES

1. Featherstone, W.R., "The Calculation of Robot Dynamics Using Articulated Body Inertias," International Journal of Robotic Research, Vol. 2, No. 1, pp. 13-30, 1983.
2. Wehage, R.A., "Recursive Multibody Dynamics by Symbolic Block Factorization," to appear in the ASME Journal of Mechanisms, Transmissions and Automation in Design.
3. Wehage, R.A., "Symbolic Factors of Linear System Coefficient Matrices for Tree-Structured Systems and Their Efficient Solution," presented at the Seventh Army Conference on Applied Mathematics and Computing, June 6-9, 1989, U.S. Military Academy, West Point, New York, this volume.
4. Wehage, R.A., "Solution of Multibody Dynamics Using Natural Factors and Iterative Refinement - Part I: Open Kinematic Loops - Part II: Closed Kinematic Loops," presented at the 1989 Design Automation Conference, September 17-20, 1989. To appear in the ASME Journal of Mechanisms, Transmissions and Automation in Design.
5. Armstrong, W.W., "Recursive Solution to the Equations of Motion of an n Link Manipulator," Proc. 5th World Congress on Theory of Machines and Mechanisms, Vol. 2, pp. 1343-1346, 1979.
6. Stepanenko, Y. and Vukobratovic, M., "Dynamics of Articulated Open-Chain Active Mechanisms," Math Bioscience, Vol. 28, pp. 137-170, 1976.
7. Orin, D.E., et al., "Kinematic and Kinetic Analysis of Open-Chain Linkages Utilizing Newton-Euler Methods," Math Bioscience, Vol. 43, 1979.
8. Shabana, A.A., Dynamics of Multibody Systems, John Wiley & Sons, New York, NY, 1989.
9. Duff, I.S., et al., Sparse Matrices and their Uses, Academic Press, New York, NY, 1981.
10. Duff, I.S., et al., Direct Methods for Sparse Matrices, Clarendon Press, New York, NY, 1986.
11. Changizi, K.C. and Shabana, A.A., "A Recursive Formulation for the Dynamic Analysis of Open Loop Deformable Multibody Systems," ASME Journal of Applied Mechanics, Vol. 55, pp. 687-693, 1988.
12. Roberson, R.E. and Schwertassek, R., Dynamics of Multibody Systems, Springer-Verlag, New York, NY, 1988.
13. Shabana, A.A. and Chang, C.W., "Connection Forces in Deformable Multibody Dynamics," International Journal of Computers and Structures, to appear, 1989.
14. Cullen, C.G., Matrices and Linear Transformations, Addison Wesley, Reading, MA, 1972.



# AUTOSIM: A COMPUTER LANGUAGE FOR REPRESENTING MULTIBODY SYSTEMS IN SYMBOLIC FORM TO AUTOMATICALLY FORMULATE EFFICIENT SIMULATION CODES

Michael W. Sayers

The University of Michigan  
Transportation Research Institute  
Ann Arbor, Michigan

## Abstract

Object-oriented symbolic computation methods are developed in this paper for describing and analyzing multibody systems, particularly vehicles. Computer data objects are defined for symbolically representing (1) vector/dyadic algebraic expressions, (2) physical components in a multibody system, and (3) program structures needed in a simulation code. With more powerful symbolic manipulation capabilities, all techniques normally employed by human analysts and programmers can be mimicked to obtain efficient numerical simulation codes. These include: selecting "natural" coordinates, dropping negligible terms, and introducing intermediate variables to avoid redundant computations. Also, the description of unusual forces and moments is straightforward when the analysis software can deal with general vector notation. The methods are demonstrated for an example three-dimensional vehicle handling model.

## Introduction

The job of simulating a multibody mechanical system breaks down into two tasks: (1) formulate equations of motion and (2) solve them numerically. The automated numerical solution of differential equations is a well developed area in engineering, and a great deal of software is available for performing this work. It is accomplished by a *simulation code*—a computer program written to numerically simulate a multibody system by integrating nonlinear ordinary differential equations over a small time step hundreds or thousands of times in a "run."

The efficiency of the simulation code is mainly determined by the number of arithmetic operations employed to compute derivatives of state variables at each time step—the equations of motion.

Approaches that are taken to simulate a system can be organized into three categories:

1. Equations of motion of the multibody system are derived by an analyst and translated by a programmer into a *specialized simulation code* that pertains to one particular multibody system.

2. A *generalized simulation code* is used in which the equations have been formulated and programmed once and for all in a generalized fashion.
3. *Symbolic analysis software* is used to aid the analyst and programmer in the formulation of equations and the development of a specialized simulation code.

The manual derivation of the equations of motion for even a modestly complex system (say, five or more degrees of freedom) is a tedious undertaking that involves considerable algebra, a nagging uncertainty of the correctness of the equations, and a considerable programming and debugging effort. To avoid these problems, numerous multibody computer programs exist which build the equations of motion for a particular system automatically, freeing the engineer to concentrate on modelling considerations and parameter values. As indicated by the above categories, some multibody analysis programs operate numerically while other operate symbolically.

### *Generalized Simulation Codes*

The (numerical) generalized simulation codes begin by building a set of equations based on a multibody formulation that has been derived for once and for all, and then they proceed to numerically integrate the equations to simulate the system [1, 2]. These generalized codes are appealing to many engineers because they offer a "complete solution" that handles the entire simulation effort, from model formulation to the numerical integration of equations. Of course, there are some trade-offs made to achieve the generality.

One trade-off is that the generalized codes run slowly relative to specialized simulation codes. A human dynamicist usually tries to obtain equations of motion that are as simple as possible, using a number of techniques that will be detailed later. Further, good programmers can improve computational efficiency when the equations are incorporated into the simulation code. Because the general-purpose simulation code was written for once and for all for all multibody systems, most of the simplification techniques cannot be used. For vehicle simulations, the eventual difference in simulation speed between a special-purpose code and a generalized code can be more than an order of magnitude (preliminary work shows a factor ranging from 10 to over 100). The inefficiency of the general-purpose software precludes its use for highly repetitive design studies and real-time, hardware-in-the-loop operations.

Another trade-off is that the generalized codes are not completely generalized when it comes to introducing force- and moment-producing components. This can be a problem with multibody systems that include elements characterized by semi-empirical models that are not likely to have been fully anticipated by the programmer. E.g., ground vehicles include tires, nonlinear springs, complex shock absorbers, etc. that are modelled differently based on the intended use of the simulation. Assuming that an engineer is able to develop a computer representation of such an element as an *external subroutine*, the subroutine must be incorporated into the multibody simulation. If the simulation program is written by hand, it is a simple matter to incorporate external subroutines. However, for a generalized simulation codes, external subroutines are limited to cases

that were anticipated by the original programmer. Variables needed as inputs to the external subroutine (positions, angles, speeds, etc.) are not always readily available.

### *Symbolic Analysis by Computer*

Symbolic computation offers the potential to combine the high reliability of a general-purpose code with the efficiency and modeling flexibility associated with the development of a new special-purpose code. In this approach, a simulation code is generated by the computer that is similar in structure and efficiency to one written by a human programmer.

There are two approaches that have been taken for performing the symbolic computation needed for analyzing multibody systems:

1. A generic symbolic manipulation language is used by a dynamicist who performs the analysis in the same manner as would be done "by hand," except that the computer aids in performing the algebra.
2. A complete, self-contained multibody analysis program is used to formulate equations automatically, based on a description of how bodies in the multibody system are connected to each other.

Generic symbolic mathematics software (e.g., MACSYMA, REDUCE, Mathematica) have been employed to develop equations of motion for multibody systems [3, 4]. These languages include capabilities far beyond the basic "high-school algebra" needed for analyzing multibody systems, and powerful computers are required for acceptable performance. However, these languages do not include provisions for optimizing numerical analysis computer code.

With a sufficiently detailed multibody formalism, equations of motion can be developed automatically using only rudimentary computer algebra. Self-contained symbolic multibody codes have been written to formulate equations that can be merged into a simulation program (e.g., NEWEUL, SD/FAST) [5, 6, 7, 8]. However, if the symbolic manipulation is too limited, some important simplification methods cannot be applied. Simplification techniques that are not included in the computer algebra can still be applied by including them in the multibody formalism, but there is a loss of modeling flexibility because the formalism must include specific "plans" for dealing with all types of systems being modeled.

This paper describes a new approach to automating the symbolic analysis of multibody systems. A symbolic mathematics language is designed specifically for analyzing multibody systems and generating numerical simulation codes. The language directly represents three aspects of the overall system in symbolic form:

1. vector and dyadic algebra expressions,
2. components of the multibody system (bodies, forces, etc.), and
3. pieces of computer code that goes into the numerical simulation code being generated.

Techniques are presented for representing and manipulating these components as computer data objects. A software package called AUTOSIM has been developed in Lisp at The University of Michigan to apply these techniques and automatically generate simulation codes. AUTOSIM is used to illustrate some of the techniques for an example three-dimensional vehicle handling model.

To provide a background for the symbolic analysis methods, considerations of numerical efficiency are presented. More background is provided by a summary of the multibody dynamics formalism that is automated to derive equations of motion.

## Numerical Efficiency

A simulation code is a computer program that simulates a physical system by numerically integrating differential equations of motion for the system of interest. The integration is performed by using a numerical approximation to integrate the equations over a very small increment of time, which is "stepped" from a start time to a stop time in a simulation run. Numerical efficiency is quantified by the number of arithmetic operations needed to compute derivatives of the state variables of the multibody system at each time step. This efficiency derives from (1) the formulation of the differential equations, and (2) the programming style of the simulation code.

### *Formulation Options*

Choices made by the analyst deriving equations of motion have a direct impact on the complexity of the resulting equations. Some of the techniques that are typically employed to simplify equations are the following:

1. State variables are introduced that are "natural" to the system being analyzed (joint displacements, speeds oriented in body-based directions, Euler angles, etc.), avoiding transformations to a predefined choice (e.g., Cartesian global coordinates).
2. Terms which are known to be zero for the specific system (but which could be non-zero for a more general formulation) are omitted from the equations.
3. Forces and moments that cancel due to symmetry or because they involve no work are eliminated when possible.<sup>1</sup>
4. Equations are written in "factored form," involving products and ratios of sums of terms. For example, the expression  $(A + B + C)^2$  requires two additions and one

---

<sup>1</sup> It should be noted that this technique is not always effective at simplifying equations. By eliminating non-working forces and moments, the number of equations is reduced but the complexity of the equations is increased. The question of whether large sets of simple equations are better or worse than small sets of complicated equations has not been resolved, and is a topic of current research. However, multibody formalisms that include the constraint forces and moments are much more complicated than the one presented in the next section, and have not been yet shown to be effective when implemented symbolically.

integer power; the expanded form  $(A^2 + 2AB + B^2 + 2AC + 2BC + C^2)$  requires five additions, six multiplications, and three integer powers.

5. Terms involving products or powers of quantities known to be "small" are dropped if they are of order 2 or higher. In many mechanical systems, some of the motions are limited such that variables associated with those motions are much smaller than other expressions arising in the equations of motion.
6. Trigonometric functions of small quantities are replaced with truncated Taylor series expansions.

Technique no. 2 (removing zero terms) can only be partially implemented (via the use of sparse matrix operations) for generalized numerical multibody simulation methods. However, virtually all symbolic multibody programs employ it. Techniques 1 - 4 have been used by some programs, and techniques 5 and 6 have not been used in a generalized sense until the implementation described in this paper. (In past work, "small" variables, when used, are built into the multibody formalism. The analyst cannot utilize knowledge that some variables and parameters are small and that others are not.)

### *Programming Options*

A given set of equations can be programmed into a simulation code so as to minimize computation. Techniques routinely employed by human programmers are the following:

7. Complicated expressions that occur in several places are replaced with *intermediate variables*. This technique is particularly important for multibody systems because the equations of motion are inherently redundant. Some of the redundancy is eliminated by using a recursive dynamics analysis method. Even so, inspection of the the equations of motion usually reveals that some subexpressions appear more than once. A human programmer, concerned with numerical efficiency, will try to avoid performing the same computation more than once by saving the results the first time and then using the result when the same computation is called for again.
8. Constant expressions are "precomputed" to avoid performing identical computations over and over with each time step. In previously developed symbolic analysis methodologies, simpler equations are obtained by specifying numerical values, rather than symbols, for parameters. During the manipulation of the symbolic expressions, the numbers are combined and the complexity of the equations is reduced [6, 8]. However, this approach results in a simulation code that is "hard-wired" for one set of parameter values, and which cannot be used for parameter sensitivity studies.

A more general approach is to identify expressions involving constants and introduce *intermediate constants*. In a simulation code, these constants can be precomputed as part of the program initialization.

9. A human programmer will (hopefully) not introduce code that serves no purpose. This obvious technique can be difficult to implement in an automated analysis method. For example, details of the dynamics analysis are often recursive.

Consequently, some expressions are developed so that they can be referenced in a later stage of the recursion. However, if the recursion stops, they may not be needed. As another example, an expression might be developed which is later multiplied by zero. Determining if a particular expression will be needed later can be very difficult at the time the expression is formulated, although it is trivial to do after all equations are formulated.

10. Large matrices are partitioned into smaller matrices, based on the topology of the system, before general numeric matrix solution methods are invoked.

## Equations of Motion for a Multibody System

To provide an idea of the sorts of mathematical operations that must be included in the computer algebra, a dynamics formalism is summarized.<sup>1</sup>

The multibody formalism presently used in AUTOSIM is based on the analysis method of Kane and Levinson [9]. For a holonomic system, or a nonholonomic system in which some speeds are constant, the following four steps are performed:

1. *Position analysis.* For each body in the system, except the inertial reference, develop a direction cosine matrix relating the body to its parent. Also, introduce a *generalized coordinate* for each degree of freedom of the joint connecting the body to its parent. For the entire system there are  $n$  generalized coordinates,  $q_i$  ( $i=1, n$ ).
2. *Velocity analysis.* For each body, introduce a *generalized speed* to account for each degree of freedom. For a nonholonomic system, there are  $p$  generalized speeds,  $u_i$ , ( $i=1, p$ ), where  $p \leq n$ , and  $m$  constant speeds,  $u_i$  ( $m = n - p$ ,  $i=p, n$ ).

For each body B, derive an expression for the derivatives of the generalized coordinates in terms of the generalized speeds. Altogether, there are  $n$  such *kinematical equations*.

Then, for each body B, formulate expressions for the following quantities:

- a.  $n$  partial velocities for the mass center,  $B^*$ , defined as

$$\vec{v}_i^{B^*} = \frac{\partial \vec{v}^{B^*}}{\partial u_i}, \quad (i=1, n) \quad (1)$$

- b.  $n$  partial angular velocities, defined as

$$\vec{\omega}_i^B = \frac{\partial \vec{\omega}^B}{\partial u_i}, \quad (i=1, n) \quad (2)$$

---

<sup>1</sup> This summary does not cover all of the details of how expressions are introduced for a specific system. Rules are applied, based on the topology of the system. A summary of the rules is beyond the scope of this paper.

c. central acceleration remainder, defined as

$$\vec{a}_{rem}^{B*} = \sum_{i=1}^n u_i \frac{d\vec{v}_i^{B*}}{dt} \quad (3)$$

d. angular acceleration remainder, defined as

$$\vec{\alpha}_{rem}^B = \sum_{i=1}^n u_i \frac{d\vec{\omega}_i^B}{dt} \quad (4)$$

3. *Implicit Equations.* The implicit equations of motion are written in matrix form as:

$$\underline{M} \underline{\dot{u}} = \underline{f} \quad (5)$$

where  $\underline{M}$  is the  $p \times p$  mass matrix,  $\underline{\dot{u}}$  is a column array of the  $p$  derivatives of the generalized speeds, and  $\underline{f}$  is a column array called the generalized force array.

a. The elements of the mass matrix are defined as

$$m_{ij} = \sum_{B=1}^{N_{Bodies}} \left( \vec{\omega}_j^B \cdot \vec{I}^B \cdot \vec{\omega}_i^B + m^B \vec{v}_j^{B*} \cdot \vec{v}_i^{B*} \right) \quad (6)$$

where  $m^B$  is the mass of body B and  $\vec{I}^B$  is the inertia dyadic of B.

b. The elements of the generalized force array are defined as

$$f_i = \sum_{B=1}^{N_{Bodies}} \left\{ \begin{aligned} & \left( \sum_{t=1}^{N_{B,T}} \vec{T}_t^B - \vec{\alpha}_{rem}^B \cdot \vec{I}^B - \vec{\omega}^B \times \vec{I}^B \cdot \vec{\omega}^B \right) \cdot \vec{\omega}_i^B \\ & + \left( \sum_{f=1}^{N_{B,F}} \vec{F}_f^B - m^B \vec{a}_{rem}^{B*} \right) \cdot \vec{v}_i^{B*} \end{aligned} \right\} \quad (7)$$

where  $\sum_{t=1}^{N_{B,T}} \vec{T}_t^B$  designates the sum of all torques applied to body B about its

center of mass by force- and moment-producing components and  $\sum_{f=1}^{N_{B,F}} \vec{F}_f^B$

designates the sum of all forces acting on the body. Forces and moments arising from the kinematical constraints need not be included, because they drop out when the dot-product is taken with the partial velocities.

4. *Explicit equations.* The  $p$  implicit equations in eq. 5 are solved to obtain values of the accelerations in  $\underline{\dot{u}}$ .

The above analysis method immediately applies several of the simplification methods described earlier. First, it permits the introduction of "natural" state variables, including generalized speeds that are not derivatives of the generalized coordinates (technique no.

1). If there is reason to think that a certain set of variables is in fact optimal, the analyst is free to use that set. (In contrast, some other multibody formalisms are built upon a pre-defined set of state variables.)

A second potential simplification occurs because non-working forces and moments are never introduced (technique no. 3).

To some extent, the efficient performance of matrix operations (technique no. 10) is also supported. When intermediate variables are introduced appropriately, the symbolic solution of the acceleration equations in step 4 results in an efficiency at least as good as can be obtained from a carefully partitioned formulation. However, it should be noted that a potential drawback of this approach is that the structure of the system is "lost" in the building of a mass matrix which is later decomposed. Recently, a number of recursive "Order-n" formulations have been published that offer greater efficiency for systems with a "chain" topology when the length of the chain exceeds a certain number, generally around  $n=6$  [10, 11, 12]. For models of ground vehicles, the formulation presented here is usually better. However, for systems with chain topologies, a recursive order-n formulation should be considered.

## Representing Symbolic Data

The methods required to manipulate symbolic expressions are derived from the design of the computer data types that are used to represent algebraic expressions and other entities. The AUTOSIM implementation was written in the language Common Lisp [13], called simply "Lisp" in the remainder of the paper.

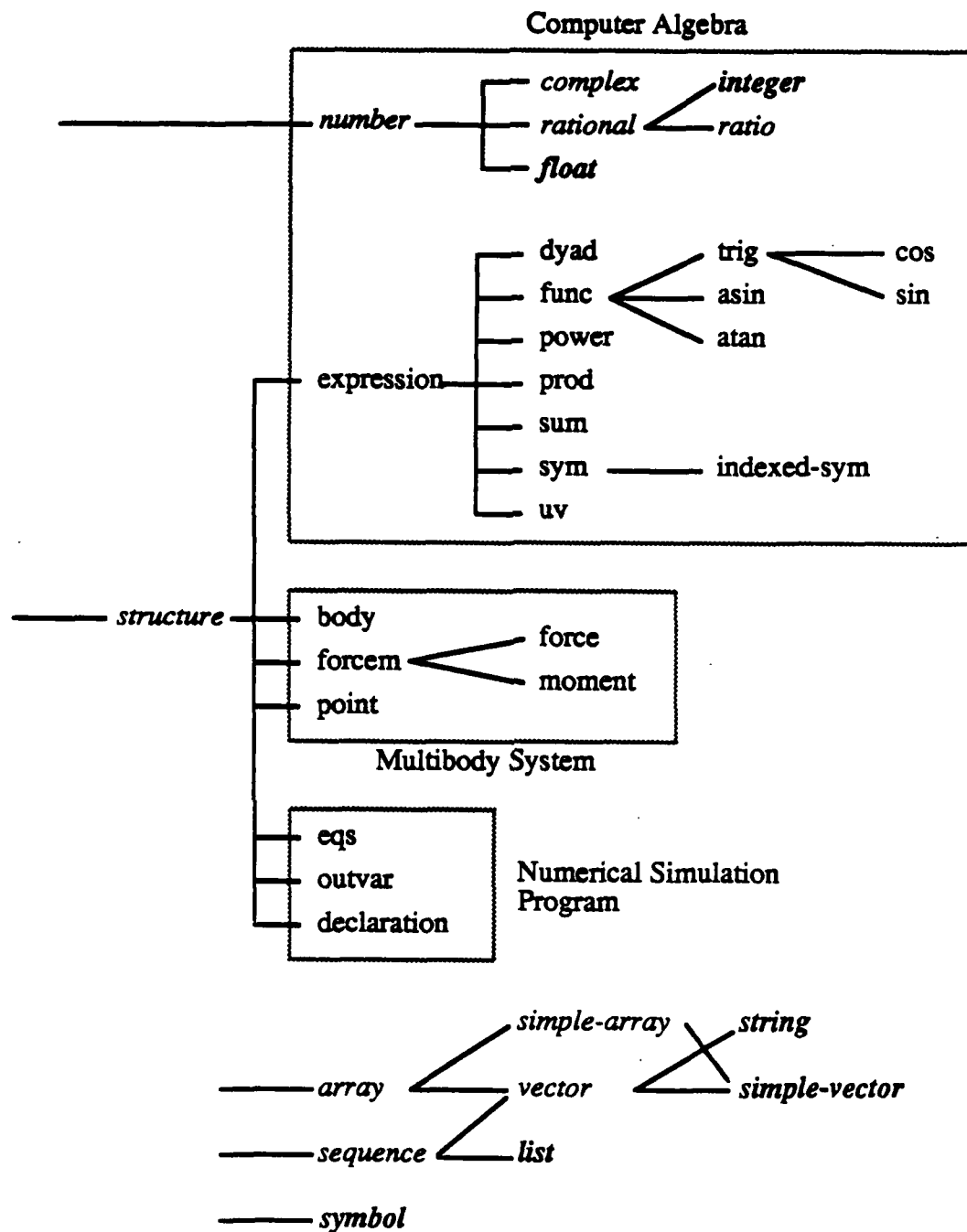
### *Overview of Data Objects*

New data types are implemented in Lisp as *structures*, with *slots* assigned to various entities associated with the data. In AUTOSIM, structures are used as *objects* to support object-oriented programming.<sup>1</sup> Objects facilitate data abstraction by allowing programs to manipulate them without knowledge of the details of their internal representation. Further, generic operators work by obtaining procedures for manipulating objects based on the types of the operands. For example, the generic function `add` works for two arguments by looking at the types of the two arguments, and looking up that pair in a dispatch table of installed specialized functions. The specialized function from the table is then invoked. The specialized function can be very specific in terms of the types of objects it understands, since it need not understand *when* it should be invoked or what to do with *other types* of data. To modify the way a generic function operates on a particular type of object, one or more new specialized functions are "installed" in the system. This style of programming allows new types of objects and new operations to be incorporated into the system without modifying existing software.

---

<sup>1</sup> Extensive object-oriented versions of Lisp are readily available, but are not standardized. To ensure portability, AUTOSIM is written completely in standard Common Lisp. The object-oriented extensions are a part of AUTOSIM.





**Figure 1. Hierarchy of AUTOSIM and Lisp data objects.**

Lisp includes over 40 types of data objects. In addition, new types are included by the use of structures. *Expressions* in AUTOSIM can represent scalars, vector, or dyadics. They are composed of two data types: numbers and expressions. Figure 1 shows a hierarchy of data types used in AUTOSIM, as they relate to data types already in Lisp. Each type of object "inherits" from the type to its immediate left in the figure. For example, an object of type *cos* is also of types *trig*, *func*, and *expression*.

Characteristics of the types `trig`, `func`, and `expression` are "inherited" by objects of type `cos`, and functions that work with objects of type `trig`, `func`, and `expression` also work with objects of type `cos`.

The data objects in the figure are shown in four groups, related to (1) computer algebra, (2) the multibody system, (3) the numerical simulation program, and (4) additional native Lisp objects. All native Lisp forms are shown in italics, and those used extensively in AUTOSIM are shown in bold italics. The multibody analyses and simplification techniques are applied by manipulating these objects.

### *Computer Algebra*

Algebraic expressions are built from number and expression objects, whose characteristics are listed in Table 1.

Of the expressions defined above, four are *elementary* types from which the other *compound* types are built. The elementary types are the `number`, the `sym`, the `indexed-sym`, and the `uv`. When printed as Fortran source code, the `sym` designates a variable and an `indexed-sym` usually designates an array element. Unit-vectors are never written in the final Fortran output, but can be entered and read by the analyst. (They are printed with enclosing square brackets.)

Recall that most of the quantities appearing in the dynamics equations are vectors and dyadics. Virtually all previously developed automated multibody analysis methods define directions ahead of time, so that vectors can be described in terms of arrays of scalar quantities with predefined directions. This approach works fine for the rigid body motions, because expressions can generally be formulated in terms of unit-vectors fixed in the body with which they are associated. However, active forces and moments can assume arbitrary orientations. Introducing arbitrary forces has not been possible with symbolic analysis programs in the past for this reason, limiting the levels of automation that are possible in the modeling. This limit is averted by including unit-vectors as a primitive entity in the computer algebra representation. Vector and dyadic expressions can be introduced using simple mathematics notation, and then manipulated automatically. Also, vector velocities and accelerations can be projected in any direction (via the dot-product operation) to define scalar output variables.

Nested expressions (simplification technique no. 4) are supported in the designs of the compound types. For example, the expressions in the list of factors of a `prod` can be sums, powers, `funcs`, etc. There are no limits to the level of nesting allowed (other than computer memory).

The meta-type `expression` defines a repertoire of qualities associated with all expression types. For example, the units of any expression (if known) are kept in the *units* slot; the name of the expression (if there is one) is kept in the *name* slot; the derivative with respect to time, if known, is kept in the slot *dxdt*.

**Table 1. Summary of AUTOSIM expression types.**

Type	Primary Slots	Definition	Examples
number		number	2, 1/3, -.3333
expression	<i>type, small-order,, sort-code, dxdt, sym-value, const-or- var, units, name</i>	meta-type for all expression objects	
dyad	<i>uv1, uv2</i>	dyad	([A1] . [A2])
func	<i>function, args</i>	function that will be written into numerical program	TIRE(FZ, SLIP)
asin		arc-sine	ASIN(X)
atan		arc-tangent	ATAN2(X, Y)
trig	<i>symbol</i>	sin or cos	
cos		cos	COS(Q(2))
sin		sin	SIN(Q(2))
power	<i>base, exponent</i>	base expression raised to power	U(1)**2
prod	<i>coef, factors</i>	product of numerical coefficient and list of expressions	2.0*M*SIN(Q(1))
sum	<i>terms</i>	sum of expressions	I + M*L**2
sym	<i>symbol, default, hide, exp</i>	symbol for a scalar parameter or variable	M
indexed- var	<i>i</i>	indexed symbol for a scalar parameter or variable	Q(2)
uv	<i>symbol, body, dot-products, cross-products</i>	unit-vector	[A1]

Expressions are classified in several ways besides their object type. The *type* slot tells whether an expression is a scalar, vector, or dyadic. Powers, syms, and indexed-syms always have their *type* slot set to the value scalar. Also, all numbers are by definition scalar. A uv has its slot set to vector, and a dyad is set to dyadic. The prod and sum objects can be any one of the three types, depending on the types of their components.

The *const-or-var* slot tells whether an expression is a constant or a variable. It is mainly used for scalar expressions. to identify expressions that can be precomputed. The value of this slot is set for a sym or an indexed-sym when it is created. When

compound expressions are examined, the *const-or-var* slot is set to *const* if all expressions contained in the compound object are constants; otherwise it is set to *var*.

Some of the other slots are described later, in the context of the algebraic operations used on expressions.

### *Multibody System*

A multibody system is composed of bodies influenced by forces and moments and connected to each other by joints. Points are fixed locations in bodies used to define joint attachments, force attachments, and points of interest needed to define output variables.

**body** — A data structure called a *body* is used to represent each body in the system, together with the kinematics of one joint and a complete coordinate system fixed in the body. Some of the slots in a *body* are shown in Table 2. The right-most column indicates whether the slot is mainly associated with the body, the joint, or the coordinate system. Massless bodies (with the *mass* and *inertia* slots set to zero) can be used to introduce compound joints or intermediate reference frames. Also, bodies with zero degrees of freedom can be used to add (or subtract) mass or inertia to an existing body.

By imposing a one-to-one relationship between bodies and joints, this design for describing a body organizes the multibody system into a *tree topology*. In general, a tree topology consists of abstract entities called *nodes*. One node is the “root node” that starts the tree, and which has no “parent node.” Every other node in the tree is defined as a “child” of a previously defined node. An example tree is shown in Figure 2, for 8 nodes labeled by capital letters. Parent-child relations are shown with lines, with the parent node above the child node(s). The root node is N; nodes A and B have N as their “parent.” Thus, A and B are the “children” of N. B has three children. Nodes G, C, D, and E all have no children, and are called “leaves” of the tree.

For a multibody system, the nodes are rigid bodies, and the connecting lines are joints between the bodies. The *body* object describes the tree topology simply by including slots for the parent and children. For example, if the tree in Figure 2 represents a multibody system, the body labelled B would identify N in its *parent* slot, and the list (C D E) in its *children* slot. The body N—the root node—would contain NIL in its *parent* slot and the list (A B) in its *children* slot.

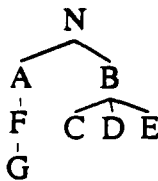


Figure 2. Example Tree.

Methods used previously to represent multibody systems have involved arrays that indicate relationships between bodies. As a minimum, a *body-connection matrix* is needed to indicate which bodies are connected by joints [8, 14]. Other matrices are needed to indicate parent-child relationships and applications of constraint equations. The representation presented here is much

simpler and permits reconstruction of the entire tree starting from any body in the tree, using only *body* objects. It also facilitates analyses that require that the bodies be processed in a certain sequence. For example, lisp code is shown below to apply a function *func* to each body in an order such that the parent is always processed before the child.

**Table 2. Some of the slots in a body.**

<b>Slot Name</b>	<b>Definition</b>	<b>Describes...</b>
<i>symbol</i>	symbol for user to reference body	body, joint
<i>name</i>	descriptive name (string) written in output files of simulation code	body, joint
<i>uvs</i>	array of 3 unit-vectors that define 1-2-3 axis directions of coordinate system	coordinate system
<i>parent</i>	parent body (body)	joint
<i>children</i>	list of bodies that have this body as their parent (list)	joint
<i>cos-matrix</i>	3x3 direction cosine matrix that relates unit-vectors of this body to the unit-vectors of the parent body (array)	coordinate system
<i>mass</i>	expression for mass of body	body
<i>inertia</i>	expression for the inertia dyadic of the body	body
<i>0-point</i>	Origin of coordinate system (also, joint attachment point in this body) (point)	coordinate system, joint
<i>cm-point</i>	center-of-mass location (point)	body
<i>joint-point</i>	joint attachment in parent body (point)	joint
<i>new-rot-vars</i>	rotational generalized coordinates introduced for this body (list)	joint
<i>new-rot-speeds</i>	rotational generalized speeds introduced for this body (list)	joint
<i>new-trans-vars</i>	translational generalized coordinates introduced for this body (list)	joint
<i>new-trans-speeds</i>	translational generalized speeds introduced for this body (list)	joint
<i>abs-w</i>	absolute rotational velocity of this body	coordinate system
<i>abs-vj</i>	absolute velocity of the <i>joint-point</i>	coordinate system
<i>worksheet</i>	another structure used to keep various expressions used for the dynamics analysis method used. For Kane's method, a structure called a kane is used which includes partial velocities, acceleration remainders, etc.	dynamics formalism

```
;;; apply function func to each body from the root down
```

```
(defun apply-func-to-tree-top-down (func body)
  (funcall func body)
  (dolist (b (body-children body))
    (apply-func-to-tree-top-down func b)))
```

The order of processing occurs from parent to child because the function is first applied to the body, and then the `apply-func-to-tree-top-down` function is recursively applied to the children of the body. By reversing two operations in the above function, so that the recursion occurs before the body is processed, the children are always processed first:

```
;;; apply function func to each body from the leaves up
```

```
(defun apply-func-to-tree-bottom-up (func body)
  (dolist (b (body-children body))
    (apply-func-to-tree-bottom-up func b))
  (funcall func body))
```

When bodies are constrained in their motions due to joints, the vector expressions developed for the body motions can be defined recursively, based on the motions of another body and the relative motion between bodies. The above function `apply-func-to-tree-top-down` is representative of the functions employed in AUTOSIM to the dynamics formalism shown earlier.

**Point** — Points are used to define locations of interest in bodies, such as origins of the coordinate systems, centers of mass, attachment points, etc. Each body contains at least three points, as shown before in in Table 2. Additional points can be defined as needed to identify attachment points for forces or as points of interest for output variables. Table 3 shows how a point is defined in the system.

Table 3. Some of the slots in a point.

Slot Name	Definition
<i>symbol</i>	Symbolic name (symbol) for user to identify point
<i>name</i>	descriptive name (string) of point
<i>body</i>	body that contains point
<i>coords</i>	array of 3 coordinates of point in coordinate system of body

**forcem** — Force-producing elements are represented by objects called forces and moment-producing elements are represented by moments. Both types, which inherit from the meta-type `forcem`, are summarized in Table 4.

As each force is introduced by the analyst, it is put into a list of all forces of the multibody system. Similarly, all of the moments are kept in a list. The summations of forces and moments, needed for eq. 7 in the dynamics analysis, are obtained for each body by going through the lists of forces and moments and checking to see if the current body is one of the two bodies contained in the two body slots of the `forcem`. If the body being analyzed is the one contained in the *body1* slot, the `forcem` is applied with

Table 4. Some of the slots in a *forcem*.

Slot Name	Definition
<i>symbol</i>	Symbolic name for user to identify <i>forcem</i>
<i>name</i>	descriptive name of <i>forcem</i> used when writing documentation
<i>dir</i>	expression that gives direction in which <i>forcem</i> acts
<i>exp</i>	expression that gives scalar magnitude of <i>forcem</i>
<i>body1</i>	first body on which <i>forcem</i> acts
<i>body2</i>	second body from which <i>forcem</i> acts
<i>point1</i>	point on line of action of force on body 1 (force only)
<i>point2</i>	point on line of action of force on body 2 (force only)

positive magnitude. When the body is matched with the *body2* slot, the *forcem* is applied with negative magnitude.

The *point1* and *point2* slots in a *force* are used to obtain expressions for the torque applied to a body if the force acts on that body and its line of action does not pass through the center of mass. That is, torque is defined as

$$\vec{T} = \vec{r} \times \vec{f} \quad (8)$$

where  $\vec{r}$  is the position vector going from the center of mass to the point on the body through which the force passes, and  $\vec{f}$  is the force (the product of the expressions in the *dir* and *exp* slots of the *force* object).

#### *Numerical Simulation Program*

In addition to expressions and the multibody system, the numerical simulation program produced as output by AUTOSIM is represented with objects. Three that are the most significant are the types *eqs*, *outvar*, and *declaration*.

**eqs** — a sequence of assignment statements is represented by an object called an *eqs*. Some of the sequences that are generated and manipulated are the kinematical equations, the dynamical equations, the trigonometric functions used in other equations, and the output variables.

**outvar** — information about a variable that will be produced as output by the simulation code is represented by the *outvar* object. It includes a short name, a long name, a generic name, an expression, and units. Before the simulation code is written, the list of *outvars* is processed to ensure that statements are generated to compute all dependent variables defined by the analyst. The labeling information is written by the simulation in such a way that output files can be handled automatically by post-processing software for graphics and analysis.

**declaration** — a list of all variables of a certain type (REAL, INTEGER, etc.) that must be declared in a specific subroutine module of the simulation code is represented in a declaration object.

In its present form, all output source code is written in the Fortran language. However, the representation of the simulation program in `eqs`, `outvar`, and `declaration` objects is not dependent on the language. Generating simulation code in a different language (e.g., C) is mainly a matter of telling these objects to print themselves differently, according to the syntax of the target language.

## Computer Algebra Operations

The mathematical operations needed to derive equations of motion for a multibody system and generate source code for a numerical simulation program can be deduced from the material presented so far. From the point of view of a software implementation, there are four levels of mathematics operations used: (1) operations are implicitly performed when a compound expression object is created (e.g., a `power` object represents an expression raised to a power, a `prod` object represents the multiplication of expressions, etc.), (2) several primitive algebra operations are defined that use information obtained from the expression objects to create a new expression object, (3) higher-level algebra operations are defined in terms of primitive operations, and (4) some operations are performed on computer code that has already been generated. This last category of operations is analogous to a human programmer “looking over” the code he or she has written, to possibly make improvements.

### *Making Expression Objects*

Each definition of a compound expression object implies an operation. The functions that make objects check their arguments and create simpler objects when possible. In fact, significant algebraic simplifications are performed in these operations. Table 5 summarizes simplifications that are performed by creator functions.

The “small”-quantity simplifications all occur in the `make-sum` operation. The term with the minimum order of “smallness” is used as a reference and all other terms are compared to it. Terms whose order is more than the reference by some threshold are dropped. Normally, the threshold for dropping small terms is 2. However, this value can be modified if needed to perform alternate analyses that require higher order terms. For example, AUTOSIM has been used to generate equations needed for a bifurcation stability analysis in which all state variables are “small” and terms are kept up to the fifth order [15].

Pains are taken to ensure that equivalent occurrences of a compound expression always are created the same way. Sums nested within sums and prods within prods are removed. E.g., the sum  $(A + B) + C$  yields  $(A + B + C)$ , rather than  $((A + B) + C)$ . Terms and factors are sorted in the `make-prod` and `make-sum` functions. E.g., the product of B and  $A * C$  is  $A * B * C$  rather than  $B * A * C$ . A sign convention for sums is used that results in a repeatable formulation for a given sum, regardless of how it is obtained.



**Table 5. Simplifications performed by creator functions.**

Creator Function	Simplifications
make-asin make-cos make-sin	<ul style="list-style-type: none"> <li>• if argument is the inverse function, return argument of argument (e.g. <math>\sin(\sin^{-1}x) \rightarrow x</math>)</li> <li>• if argument is a number, evaluate</li> <li>• if argument is small, return truncated Taylor expansion</li> </ul>
make-atan	<ul style="list-style-type: none"> <li>• same simplifications as for make-asin</li> <li>• if there are two arguments, divide both by GCF [e.g., <math>\tan^{-1}(A*X, A*Y) \rightarrow \text{ATAN2}(X, Y)</math>]</li> </ul>
make-power	<ul style="list-style-type: none"> <li>• if base is a power, change exponent</li> <li>• if base is number, evaluate</li> <li>• if base includes small terms, drop if possible</li> </ul>
make-prod	<ul style="list-style-type: none"> <li>• if the coefficient is 0, return 0</li> <li>• if the coefficient is 1 and there is one factor, return the factor</li> <li>•• if any numbers are included as factors, multiply them and include with the coefficient and remove the numbers from the list of factors</li> <li>•• if any factors are prods, multiply coefficients and combine lists of factors (i.e., expand nested prods)</li> <li>•• if any factors can be combined into a power, make the substitution</li> <li>• else, sort factors and create prod object</li> </ul>
make-sum	<ul style="list-style-type: none"> <li>•• compare "small-order" values of terms and remove those which are negligible</li> <li>•• check for trig identities: <math>\sin^2x + \cos^2x \rightarrow 1</math>; <math>1 - \sin^2x \rightarrow \cos^2x</math>; <math>1 - \cos^2x \rightarrow \sin^2x</math></li> <li>•• if any terms are sums, remove them and append terms from nested sums to existing list (i.e., expand nested sums)</li> <li>•• if sym-value of sum would be negative, negate all terms and return negative sum (prod with coefficient of -1)</li> <li>• else, sort terms and create sum object</li> </ul>
<b>Note:</b>	simplifications marked with •• mean that after the simplification is performed, the make- operation is called again recursively using updated arguments.

The expression  $(-A - B - C)$  would never be generated: instead, that result is always represented as  $-(A + B + C)$ .

*Primitive Algebra Operations*

Table 6 summarizes the primitive mathematical operations. These operations involve one or two arguments. In the object-oriented environment, each operator has an

**Table 6. Summary of primitive AUTOSIM mathematics operations.**

Operation	Argument(s)	Description
add	exp1, exp2	add two expressions
const-or-var	exp	is expression constant or variable?
cross	vexp1, vexp2	dot product between two vectors
dot	vexp1, vexp2	dot product between two vectors
dxdt	exp	derivative of expression with respect to time, in the inertial reference frame
gcf	exp1, exp2	find symbolic greatest common factor
mul	exp1, exp2	multiply two expressions
neg	exp	negate expression
partial	exp, symbol	partial derivative of expression with respect to symbol, in the inertial reference frame

associated dispatch table which is used to find a function for dealing with a specific type of expression (for unary functions) or combination of types (for binary operations). For example, to add a `sum` and a `prod`, the combination (`sum prod`) is looked up in the appropriate table, and the function found from the table is applied to the two arguments. Generally, the dispatch functions are small, simple, and specific to one combination of expression types. Hence, they are easy to modify and debug. Also, new types of expression objects and new functions are "installed" in the system without modifying any of the existing software.

Most of the operators in the table work as might be expected. Exceptions and special notes are provided below.

**mul** — When developing expressions through multiplication, products are not expanded, in order to keep factored forms.<sup>1</sup> Further simplifications are attempted—numbers are combined, multiple appearances of an expression are combined into a `power`, multiple powers with the same base expression are combined, etc.

**gcf** — The symbolic "greatest common factor" (GCF) between X and Y is determined. (If X and Y have no factors in common, or one of them is a number, then the GCF is 1.)

**add** — The general method for adding two expressions X and Y is with the formula

$$X + Y = \text{GCF}(X, Y) * (X / (\text{GCF}(X, Y) + Y / \text{GCF}(X, Y))$$

After the GCF is factored out, the results are combined with `make-sum`. For example, when the expressions `A*X` and `B*X**2` are added, the result is `X*(A + B*X)`.

<sup>1</sup> There are applications in which expanded forms are preferred. For example, stability analyses can require coefficients of state variables and their products and powers. The AUTOSIM software does include an option to expand expressions, although this option is not used when the objective is to automatically generate simulation codes.

**dot** — The dot product operation is valid for two vectors, a vector and a dyad, or two dyads. The method used for applying the operation is to recursively expand expressions into multiplications and additions of subexpressions, and dot products of uv/dyad pairs. This approach eventually expands the original dot product to an expression involving operations defined for scalar algebra, together with dot products between unit-vectors. Thus, the only new primitive operation needed is the dot product between two uvs.

Recall that the uv contains a slot called *dot-products*. This contains a table with all pairs of uvs whose dot product is known. Initially, each table contains three entries for the three uvs in the body in which the uv is defined. (The values are 1 for the dot product of the uv with itself and 0 for the other two uvs of the triad.) If the table contains the answer, it is used. Otherwise, the dot product is between two uvs associated with different bodies that have not yet been analyzed, so an analysis is performed.

Each body has a slot with a direction cosine matrix relating the uvs for that body with the uvs of the parent. The uv whose body is furthest “down” the topology tree is transformed into an expression involving the three uvs of its parent body. The dot product is then taken between the new expression and the uv that was “up” the tree.

This method is recursive—the dot operator is defined in terms of itself. It works, because with each recursion, the expressions being considered are simpler, and/or the uvs are closer in the tree. Eventually, the process is guaranteed to stop when both arguments are uvs associated with the same body.

The results of the process are stored in the table of dot-products for one of the original uvs, so that the “tree-climbing” and transformations (via the direction cosine matrices) are not required the next time the dot product is needed.

The method of “tree climbing” ensures that the minimum number of direction transformations is performed for each dot product operation. Thus, trigonometric simplifications are not required for this operation.

Note that the dot-product operator makes use of information from both the uv object from the computer algebra part of the system, and also the body object from the multibody part of the system.

**cross** — The cross product operation is performed using the same recursive approach as described above for the dot product. A uv crossed with a uv is obtained from the table of values in the cross-product slot of either uv if available (with a multiplication by  $-1$  if the table of the second uv is used). Otherwise, the cross-product is formulated using the expansion:

$$\vec{a} \times \vec{b} \rightarrow [(\vec{a} \cdot \vec{b}_1) \vec{b}_1 + (\vec{a} \cdot \vec{b}_2) \vec{b}_2 + (\vec{a} \cdot \vec{b}_3) \vec{b}_3] \times \vec{b} \quad (9)$$

where  $\vec{a}$  is the first uv,  $\vec{b}$  is the second, and  $\vec{b}_1$ ,  $\vec{b}_2$ , and  $\vec{b}_3$  are the unit-vectors for the body containing  $\vec{b}$ . As was the case for the dot product, some of the information needed to perform the operation is obtained from the body object from the *body* slot of the uv object.

**dxdt** — The derivative of an arbitrary expression is determined using elementary rules of calculus to recursively expand the expression into products and sums of simpler expressions and their derivatives. The expansion stops when a sym, indexed-sym, number, or uv is reached. The time derivative of a sym or indexed-sym is zero if the expression is a constant, otherwise it is obtained from the *dxdt* slot.

The time derivative of a uv ( $\vec{u}$ ) is defined as

$$\dot{\vec{u}} \rightarrow \vec{\omega}^{B_1} \times \vec{u} \quad (10)$$

where  $\vec{\omega}^{B_1}$  is the absolute rotational velocity of the body containing  $\vec{u}$ , obtained from the *abs-w* slot of the body found from the *body* slot of the uv.

There are other ways in which the time derivative might be defined. For example, one could project the uv into the coordinate system of the fixed inertial reference and then take derivatives of the scalar components. However, eq. 10 has two strong advantages:

1. it leads to simple expressions, matching the conventional definition of the derivative of a vector fixed in a rotating reference frame.
2. the cross-product operation remains valid after small terms have been dropped and trigonometric functions have been replaced with truncated Taylor series. Thus, simplifications from small angles and small speeds can be made as soon as the small quantities appear in the analysis without causing errors in derivatives taken later.

After the absolute time derivative of an expression is derived, the result is put into the *dxdt* slot for further reference.

**partial** — Partial derivatives are obtained using the same basic process as used for *dxdt*, except that results are not saved and partial derivatives of unit-vectors are zero.

**Table 7. Summary of higher level mathematics operations.**

Operation	Argument(s)	Description
sub	exp1, exp2	negate exp2 and add to exp1
inv	exp	make-power with exponent of -1
square	exp	multiply expression with itself
div	exp1, exp2	invert exp2, then multiply with exp1
dot-plane	vexp1, vexp2	project vexp1 onto plane normal to vexp2
mag	vexp	scalar magnitude of vector, $ \vec{v}  \rightarrow \sqrt{\vec{v} \cdot \vec{v}}$
dir	vexp	direction of vector, i.e., $\vec{v}/ \vec{v} $
angle	vexp1, vexp2, (vexp3)	angle between vexp1 and vexp2, with sign determined by optional vexp3

### Higher Level Operations

Table 7 lists mathematics operations that are derived from the above primitive functions.

Most of the above operators have standard meanings and are implemented according to their definitions. However, two are introduced to aid in describing quantities associated with vehicles and deserve further explanation.

**dot-plane** — This operator describes a short procedure in which a vector is projected onto a plane. Consider vectors  $\vec{a}$  and  $\vec{b}$ . The new vector is defined as  $\vec{a} \cdot (\vec{c}_1 \vec{c}_1 + \vec{c}_2 \vec{c}_2)$  where

$$\vec{c}_1 = \frac{\vec{a} \times \vec{b}}{|\vec{a} \times \vec{b}|} \quad \vec{c}_2 = \frac{\vec{c}_1 \times \vec{b}}{|\vec{c}_1 \times \vec{b}|} \quad (11)$$

**angle** — The angle between two vectors  $\vec{v}_1$  and  $\vec{v}_2$  is determined by defining three unit-vectors and projecting one onto the other two to obtain an expression for the arctangent of the angle. The steps are described below and illustrated in Figure 3:

1. The directions of the two vectors are obtained:

$$\vec{u}_1 = \frac{\vec{v}_1}{|\vec{v}_1|} \quad \vec{u}_2 = \frac{\vec{v}_2}{|\vec{v}_2|} \quad (12)$$

2. A third direction is defined that is orthogonal to  $\vec{v}_1$ :

$$\vec{u}_3 = (\vec{u}_1 \times \vec{u}_2) \times \vec{u}_1 \quad (13)$$

3. The angle,  $\theta$ , is defined as

$$\theta = \tan^{-1} \left( \frac{\vec{u}_3 \cdot \vec{u}_2}{\vec{u}_1 \cdot \vec{u}_2} \right) \text{sign}(\vec{v}_3 \cdot [\vec{u}_1 \times \vec{u}_2]) \quad (14)$$

This method is valid for angles of any size. Results are expressed using the Fortran ATAN2 function, which accepts two arguments and is valid for the range of  $-180^\circ \leq \theta \leq +180^\circ$ . The make-at-an function is used to create the resulting expression, with the possible simplifications noted earlier in Table 5. Note that an optional third vector,  $\vec{v}_3$ , is used to establish the sign of the angle. (The sign function in eq. 14 has a value of  $\pm 1$ , with a sign that matches that of its argument.)

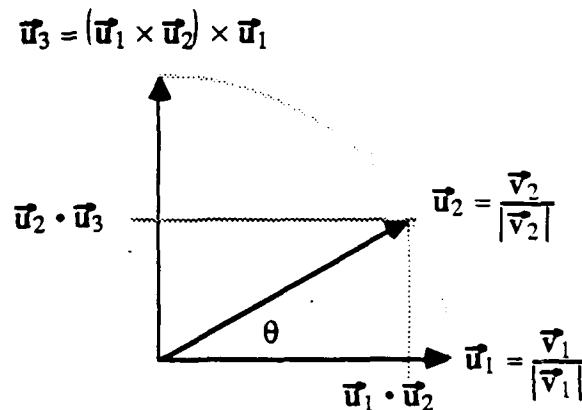


Figure 3. Angle calculation.

### *Multibody Operations*

A few operations for dealing with points and bodies are useful for specifying forces, moments, and dependent variables of interest. These are summarized in Table 8.

**Table 8. Summary of AUTOSIM operations for bodies and points.**

Operation	Argument(s)	Description
rot	body	rotational velocity of body
pos	point	position vector from origin of inertial reference to point
pos	point1, point2	position vector from point2 to point1
vel	point	velocity vector from origin of inertial reference to point
vel	point1, point2	velocity vector going from point2 to point1

The effect of `vel` can be obtained using the `pos` operator together with `dxdt`. However, the result usually involves derivatives of generalized coordinates, whereas the `vel` function provides the result as an expression involving generalized speeds.

Accelerations are obtained by combining the `dxdt` function with `rot` and/or `vel`.

### **Operations on Program Code**

The equation simplifications noted earlier (simplification techniques 8, 9, and 10) are easy to implement after the simulation code has been generated and can be inspected. This means that equations are not written as they are derived, but are kept in computer memory as `eqs` objects.

#### *Introduction of Intermediate Variables and Constants*

The simulation code generated by AUTOSIM includes two sets of intermediate symbols used to replace expressions. One set is for constant expressions and the other is for variables. (Both are called intermediate variables below, since that is how they are implemented in a Fortran program.) A function called `intro-var-if-new` is used to process expressions and introduce new variables as needed. The method is for doing this involves a table that is maintained by the system of all expressions that have been replaced by intermediate variables. The replacements are `indexed-sym` objects, which prints as elements of a Fortran array `PC` (for precomputed constants) or `Z` (for variables). A simplified version of the algorithm in `intro-var-if-new` is as follows:

- If the expression is an `indexed-sym`, a `sym`, or a number, it is returned.
- Else, if the expression is in the table of existing intermediate variables, the corresponding `indexed-sym` is returned.
- Else, if the expression is a constant, define a new `indexed-sym`, put it at the end of the list in the `eqs` object for intermediate constants, put the expression and

symbol into the table of intermediate variables, and return the new indexed-sym.

- Else, if any constant expressions can be factored out, do so. Apply `intro-var-if-new` to the constant part and the variable part, then apply `intro-var-if-new` to the product.
- Else, apply `intro-var-if-new` to all components of the compound expression (arguments in a `func`, factors in a `prod`, etc.), then continue.
  - If the expression is a `prod`, process the scalar factors two at a time. If the `prod` included a factor that is a `uv` or `dyad`, skip over it (intermediate variables are only used to represent scalar expressions). Multiply the first two scalar factors and apply `intro-var-if-new` to the result. Multiply the result with the next scalar factor and apply `intro-var-if-new` to that result. Proceed until all scalar factors have been processed. The definitions of the new indexed-syms are variables, and are placed at the end of an `eqs` object used for the intermediate variables.
  - Else, introduce a new indexed-sym, put its definition at the end of the appropriate `eqs` object, update the table, and return the new indexed-sym.

This algorithm is recursive, and results in a number of intermediate expressions being introduced for a single compound expression. For example, consider the expression  $A*(B*X + C*Y)$ , where A, B, and C are constants and X and Y are variables. Processing this expression with the `intro-var-if-new` function leads to the following `eqs` object for intermediate constants,

```
PC (1) = A*B  
PC (2) = A*C
```

and the following object for intermediate variables:

```
Z (1) = PC (1) *X  
Z (2) = PC (2) *Y  
Z (3) = Z (1) + Z (2)
```

Note that the number of multiplications needed to compute the full expression has been increased from 3 in the original, to 4 with the intermediate variables. However, two of the new multiplications involve constants, leaving only two multiplications that must be performed at each time step during a numerical simulation run.

For the above algorithm to be effective, it is essential that expressions are uniquely identified in the table. For example, if the product  $A*(1 + \text{COS}(Q(1)))$  occurs in one place, we don't want an equivalent expressions such as  $(-\text{COS}(Q(1)) - 1)*A$  to occur in another, because the search of the look-up table will not find the second occurrence. This is why the `make-prod` and `make-sum` functions described earlier ensure that a given product or sum always has the same structure.

The above algorithm always introduces a new intermediate variable whenever an arithmetic operation or function evaluation occurs. For simple multibody systems, this can sometimes degrade computational efficiency, by eliminating possible simplifications that occur by factoring. For example, consider an expression  $A*U(1)$  which is later added to  $A*U(2)$ . If both expressions are replaced by intermediate variables, say  $Z(5)$  and  $Z(15)$ , the sum is  $(Z(5) + Z(15))$ . It requires 2 multiplications, which occur when  $Z(5)$  and  $Z(15)$  are computed. If the intermediate variables were not introduced, the result of the addition would be  $A*(U(1) + U(2))$ —an expression with only one multiplication.

There are some reasons not to introduce a new intermediate variable if that variable will only be used once. First, the equations become almost unreadable by humans. The equations are usually complicated to begin with, and introducing intermediate variables that only appear once compounds the difficulty. Second, some Fortran compilers optimize machine instructions for large expressions, putting temporary intermediate results directly into working registers. For machines with vector processing or other parallel computing capabilities, other techniques are available for the compiler. If an intermediate variable is defined in the source code, the compiler is obliged to save its value by moving it into a RAM location. For these reasons, the method described below for removing unused code is extended to also eliminate any intermediate variables that would only be used once.

### *Removal of Unused Code*

Before the equations are written as output into a Fortran program, they are inspected for intermediate variables that are never used, or used only once. Only equations that contribute to the computation of the accelerations or to the computation of output variables are actually written into the simulation code that is generated by AUTOSIM.

An important part of the design of AUTOSIM is that the three symbolic elements—the *sym*, the *indexed-sym*, and the *uv*—are stored in memory such that there are no copies. For example, the object called “Q(2)” exists in only one place, even though it appears in more than one expression.<sup>1</sup> Recall that one of the slots in the *sym* object is called *hide*. The *hide* slot is used for removing unused code by keeping count of how many times the *sym* actually appears. The *eqs* object only prints equations involving *syms* whose *hide* slots are not set to 0. For example, if an *eqs* contains 100 equations, but only 10 involve *syms* with *hide* counts greater than 0, then only 10 equations are printed. The other 90 equations are still in memory, but are hidden.

To count occurrences, the *hide* slots in all intermediate variables in an *eqs* are set to 0, and then equations used to compute derivatives and output variables are processed with a function called *validate-exp*. The *validate-exp* function operates recursively to “validate” *syms*. If its argument is a *sym* or *indexed-sym*, it increments the count in the *hide* slot, and then applies itself recursively to the expression on the right-hand side

---

<sup>1</sup> Lisp uses *pointers* to reference such objects when they are “contained” in other objects. Thus, when an elementary object is changed, all expressions “containing” that element are updated since their pointers continue to point at the changed object.



of the equation (available from the *exp* slot). If the argument is a compound expression, *validate-exp* applies itself to all of the parts of the expression (arguments in a *func*, factors in a *prod*, etc.)

After the *hide* values have been established for all indexed-syms that appear on the left-hand side of an equation, a second pass is made in which all intermediate variables that are used only once (*hide* = 1) are expanded back into the original expressions.

### Example Vehicle System

A three dimensional vehicle example will now be used to illustrate some of the above methods. The vehicle is shown conceptually as a multibody system in Figure 4. Although the model is relatively simple, it has been shown to predict steering responses that closely match measurements from the test track [16].

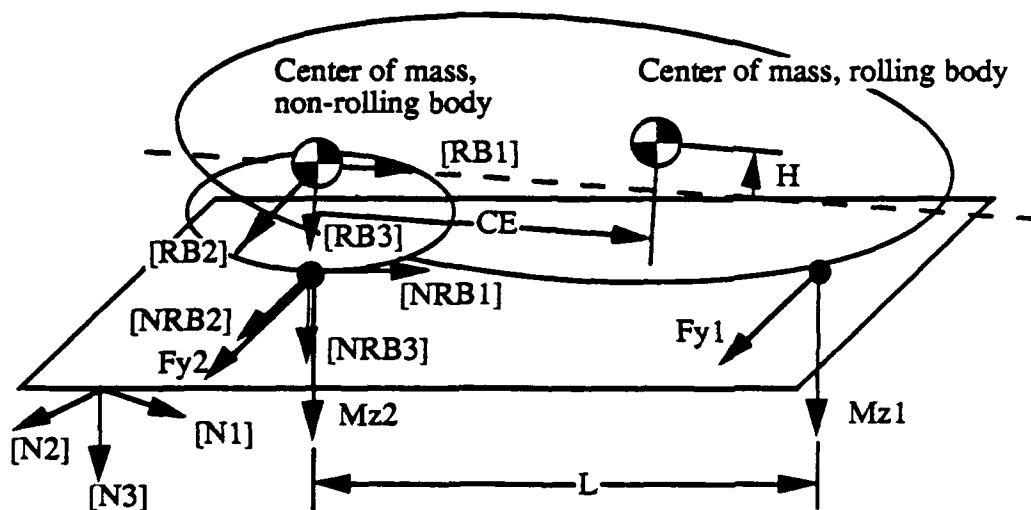


Figure 4. Example vehicle model.

The coordinate system of the inertial reference has its origin in the plane of the road, with axes along directions [N1], [N2], and [N3], where the unit-vector [N3] points down<sup>1</sup>.

The vehicle is modelled as two rigid bodies. One is body NRB that is free to translate and yaw in the plane of the road. The second is body RB, which rolls relative to NRB about a roll axis tilted as shown. This model nominally has four degrees of freedom. However, the forward speed is set constant, limiting the dynamical degrees of freedom to three. The vehicle responds to two applied side forces from the tires ( $F_{y1}$  and  $F_{y2}$ ), two aligning moments ( $M_{z1}$  and  $M_{z2}$ ), gravity, and a roll moment generated by the suspension springs and dampers.

<sup>1</sup> In AUTOSIM unit-vectors are written enclosed with square brackets.

All of the ingredients of the model can be represented using the computer data objects presented earlier. To generate the objects automatically, some Lisp macros and functions are included in AUTOSIM to build the description of the multibody system on the computer. These are summarized in Table 9.

**Table 9. Summary of AUTOSIM commands for building a multibody system description.**

Lisp form	Purpose
add-body	describe one body completely, including its position in the system topology, the kinematics of its joint, and the mass and inertial properties of its rigid body.
add-gravity	apply gravitational force to each body with mass
add-line-force	describe force-producing component (direction of force is known)
add-moment	describe moment-producing component
add-point	identify point of interest on a body
add-strut	describe force-producing component (end points are known)
reset	initialize AUTOSIM and clear previous results
set-speed-constant	specify that a generalized speed is a constant (and thus remove a dynamical degree of freedom)
small	declare syms to have a small-order of 1

The example system is described using these macros in the listing shown in Figure 5. Note that most of the information provided in this paper is not needed to prepare the inputs shown in the listing. The inputs needed to model the example system involve just eight different macros with a fairly simple syntax. The entries are Lisp forms, as described in all Lisp textbooks. The types of Lisp data used in the macros are the symbol, string, array, number, and the *F-string* —a Fortran-style expression entered as string preceded by an exclamation mark, e.g., `!"-kroll*q(4) - croll*u(3)"`. Advanced users can use the programming power of Lisp to define additional variables, use DO loops, etc. However, a knowledge of Lisp is not required to use AUTOSIM.

The specific lines of input shown in the listing of Figure 5 are now described briefly. Every multibody system begins with the inertial reference (N), which in turn contains one point, O, the origin. These objects are established with the form (reset), which also initializes many of the global objects used in AUTOISIM.

The first add-body macro in Figure 5 tells several things about the new body to AUTOSIM: (1) the new body has the inertial reference N as its "parent," (2) the symbolic name for the new body is NRB, (3) a more descriptive name to use in documentation is "non-rolling body," (4) NRB has two translational degrees of freedom relative to the inertial reference, in the directions of axes 1 and 2 ([N1] and [N2]), (5) the center of mass of NRB is a distance HRA above the ground, and (6) NRB has a single rotational degree of freedom about axis 3 ([N3]).

<pre> (reset)  (small thetar ixz)  (add-body n nrb  :name "non-rolling body"  :translate (1 2)  :cm-coords #(0 0 !" -hra")  :parent-rot-axis 3)  (add-body nrb rb  :name "rolling body"  :body-rot-axes (1)  :parent-rot-axis   #(!"cos(thetar)" 0     !"sin(thetar)")  :small-angles (t)  :joint-coords #(0 0 !" -hra")  :cm-coords #(ce 0 !" -h")  :inertia-matrix   #2a((Ixx 0 Ixz)       (0 Iyy 0)       (Ixz 0 Izzr))  (add-point front  "the front axle point" nrb  #(L 0 0))  (small (dot (rot nrb) '[n3])  (dot (vel nrb0) '[nrb2]))  (set-speed-constant  (dot (vel nrb0) '[nrb1]) speed)  (add-gravity) </pre>	<pre> (add-line-force fyl  "Side force, front axle"  !"ca1*(angle([nrb1],               vel(front), [nrb3])               - steer)  - cgl*ccoef1*q(4)"  front [nrb2] o :no-forcem t)  (add-line-force fy2  "Side force, rear axle"  !"ca2*(angle([nrb1], vel(nrb0),               [nrb3]) - krs2*q(4)"  nrb0 [nrb2] o :no-forcem t)  (add-moment mzl  "Aligning moment, front axle"  [n3]  !"cam1*(angle([nrb1],               vel(front), [nrb3])               - steer)"  nrb n :no-forcem t)  (add-moment mz2  "Aligning moment, rear axle"  [n3]  !"cam2*(angle([nrb1], vel(nrb0),               [nrb3]) - krs2*q(4))"  nrb n :no-forcem t)  (add-moment rollm  "roll moment from suspension"  [rb1]  !" -Kroll*q(4) - Croll*u(3)"  rb nrb :no-forcem t) </pre>
--	---

Figure 5. Description of car model in AUTOSIM.

The second add-body macro designates NRB as the parent body and names the new body RB. Further, it indicates that (1) the descriptive name of RB is "rolling body," (2) there is a single rotational degree of freedom, aligned with axis 1 of the coordinate system of RB, [RB1], (3) the rotation axis is oriented in the direction whose coordinates (in the frame of the parent NRB) are (COS(THETAR), 0, SIN(THETAR)) (that is, the axis is inclined down from axis 1 by an angle THETAR towards axis 3), (4) the rotation involves a small angle, (5) the origin of the coordinate system of RB is located at coordinates (0, 0, -HRA) in the coordinate system of NRB, (6) the center of mass is located at coordinates (CE, 0, -H) in the coordinate system of RB, and (7) the inertia

matrix for RB is 
$$\begin{bmatrix} IXX & 0 & IXZ \\ 0 & IYY & 0 \\ IXZ & 0 & IZZR \end{bmatrix}.$$

The AUTOSIM design permits both variables and parameters to be "small." In this example, the parameters THETAR and IXZ were declared to be "small" before the add-

body macros are applied.<sup>1</sup> Thus, when IXZ is multiplied with a “small” variable and the product is added to another parameter, the product is recognized as being numerically negligible and is dropped from the sum.

The body objects created to represent these two bodies are printed in Figure 6 to show the values associated with some of the slots.

Summary of body: NRB	Summary of body: RB
<b>parent:</b> N <b>level:</b> 1 <b>children:</b> (RB) <b>Name:</b> Non-Rolling Body <b>mass:</b> NRBM <b>inertia:</b> IZZNR*([N3].[N3]) <b>unit-vectors:</b> #([NRB1] [NRB2] [N3]) <b>new-trans-vars:</b> (Q(1) Q(2)) <b>new-trans-speeds:</b> (U(1) U(2)) <b>new-rot-vars:</b> (Q(3)) <b>new-rot-speeds:</b> (U(3)) <b>rot-dir-list:</b> ([N3]) <b>trans-dir-list:</b> ([N1] [N2]) <b>joint-pos:</b> (Point NRBJ: Body N: #(0 0 0): attachment point for the non-rolling body) <b>cm-pos:</b> (Point NRBCM: Body NRB: #(0 0 -HRA): center of mass of the non-rolling body) <b>abs-w:</b> U(3)*[N3] <b>abs-vj:</b> (U(1)*[NRB1] + U(2)*[NRB2]) <b>cos matrix:</b> #(COS(Q(3)) SIN(Q(3)) 0) #(-SIN(Q(3)) COS(Q(3)) 0) #(0 0 1.0)	<b>parent:</b> NRB <b>level:</b> 2 <b>children:</b> NIL <b>Name:</b> Rolling Body <b>mass:</b> RBM <b>inertia:</b> (IXX*([RB1].[RB1]) + IXZ*([RB3].[RB1]) + IXZ*([RB1].[RB3]) + IYY*([RB2].[RB2]) + IZZR*([RB3].[RB3])) <b>unit-vectors:</b> #([RB1] [RB2] [RB3]) <b>new-rot-vars:</b> (Q(4)) <b>new-rot-speeds:</b> (U(4)) <b>rot-dir-list:</b> ([RB1]) <b>joint-pos:</b> (Point RBJ: Body NRB: #(0 0 -HRA): attachment point for the rolling body) <b>cm-pos:</b> (Point RBCM: Body RB: #(CE 0 -H): center of mass of the rolling body) <b>abs-w:</b> (U(3)*[N3] + U(4)*[RB1]) <b>abs-vj:</b> (U(1)*[NRB1] + U(2)*[NRB2]) <b>cos matrix:</b> #(1.0 0 THETAR) #(-THETAR*Q(4) 1.0 Q(4)) #(-THETAR -Q(4) 1.0)

Figure 6. Description of body structures for example vehicle.

The macros introduced point objects, generalized coordinates, generalized speeds, and a direction cosine matrix based on the degrees of freedom. Because the parameter THETAR and the roll rotation angle are both “small” angles, the direction cosine matrix of RB does not include any trigonometric functions.

Note that the matrix includes the product -THETAR\*Q(4), which is of order 2. The reason this is included (rather than the number 0) is that a small expression is not set to

<sup>1</sup> The macro operates by finding the s yms with the names THETAR and IXZ (or creating them if they don't already exist), and then setting the *small-order* slot of each to a value of 1.

zero unless it added to another expression of an order that is lower by two or more. For example, if  $\text{THETAR} * Q(4)$  (order=2) is added to the number 1 (order=0), the result is 1.

The definitions of the state variables can also be printed (the `add-body` macro creates a name for every symbol it introduces, and also sets the units). The summaries printed by AUTOSIM are shown in Figure 7.

Generalized Coordinates	
Q(1):	Translation of NRB relative to the attachment point for the non-rolling body along [n1]. (in)
Q(2):	Translation of NRB relative to the attachment point for the non-rolling body along [n2]. (in)
Q(3):	Rotation of the non-rolling body relative to the inertial reference about axis #3. (deg)
Q(4):	Rotation of the rolling body relative to the non-rolling body about axis #1. (deg)
Generalized Speeds (before <code>set-speed-constant</code> macro is used)	
U(1):	Abs. trans. speed of NRB along axis 1. (in/s)
U(2):	Abs. trans. speed of NRB along axis 2. (in/s)
U(3):	Abs. rotation of NRB, axis 3. (deg/s)
U(4):	Rotation of RB relative to NRB, axis 1. (deg/s)

Figure 7. Printed summary of state variables.

Note that generalized speeds for translational velocity are defined that are not derivatives of the generalized coordinates.

The macro `set-speed-constant` removes a dynamical degree of freedom by changing slot values in the `indexed-sym` object that represents a generalized speed, and then renumbering the remaining speeds. The renumbering is performed by changing the *i* slot in all `indexed-sym` objects that represent generalized speeds. In the example, the forward vehicle speed, initially printed as "U(1)" is declared to be a constant called `SPEED`. The macro changes the `const-or-var` slot to `const`, the `dxdt` slot to 0, the `exp` slot to `SPEED`, and the *i* slot to 0.

Printing of expressions is performed recursively, with every type of object having an associated print function. If an object is changed such that it prints differently, all expressions containing that object will also print with the "updated" form. The print-function associated with `indexed-sym` objects prints the expression in the `exp` slot if the *i* value is 0. Thus, all expressions that contain the generalized speed originally named U(1) will now print that object as "SPEED." The generalized speeds have been renumbered and appear as shown in Figure 8.

Because AUTOSIM will freely rename objects, the analyst must be careful when referring to state variables by name. For example, the speed that was originally called U(2) is now U(1). The possibility of erroneously naming the wrong variable can be

**Generalized Speeds (After set-speed-constant macro is used)**

SPEED: Abs. trans. speed of NRB along axis 1. (in/s)  
U(1): Abs. trans. speed of NRB along axis 2. (in/s)  
U(2): Abs. rotation of NRB, axis 3. (deg/s)  
U(3): Rotation of RB relative to NRB, axis 1. (deg/s)

**Figure 8. Printed summary of generalized speeds after speed is set constant.**

eliminated by referring to speeds generically as scalar projections (dot products) of velocities. The `small` and `set-speed-constant` macros near the bottom of the left column in Figure 5 illustrate this. The `small` macro is applied to two expressions: the first is the rotational velocity of NRB dotted with the unit-vector [N3] (i.e., the yaw rate), and the second is the velocity of the origin of NRB dotted with the unit-vector [NRB2] (i.e., the lateral component of the velocity). The `set-speed-constant` macro is applied to the speed defined as the velocity of the origin of NRB dotted with the unit-vector [NRB1] (i.e., the the forward component of the velocity). These expressions are always valid and do not require knowledge of how the speeds are currently named. For example, if we were to add more degrees of freedom to the model, such that the generalized speeds would be numbered differently, the generic descriptions in the listing of Figure 5 would still be correct.

The macro `add-point` is used to define a point called `front` at which the front tire force is applied. The macros `add-line-force` and `add-moment` are used to define tire forces and moments.

The arguments to the `add-line-force` macro are: (1) a symbol for the force object (to go in the *symbol* slot), (2) a name for the force (to go in the *name* slot), (3) an expression for the magnitude of the force (to go in the *exp* slot), (4) a point upon which the force acts (to go in the *point1* slot), (5) a direction associated with the force (to go in the *dir* slot), and (5) a point associated with a body from which the force is reacted (the point itself is not saved, however the body associated with the point goes into the *body2* slot). The *body1* slot is assigned the body associated with the point in *point1*, and the *point2* slot is NIL.

The most complicated of the above arguments is the one for the force magnitude. The expression involves the slip angle for a point, defined as the angle between the forward direction of the point, and the velocity vector of the point. The slip angle angle is defined for the front tires with the F-string

```
!"angle([nrbl], vel(front), [nrb3]) - steer"
```

The F-string is parsed (interpreted) by AUTOSIM as: derive an expression for the angle between the forward direction [nrbl] and the velocity of the point named front, `vel(front)`, with a sign defined by a positive angle corresponding to a clockwise sweep when viewed from an observer looking in the direction [nrb3], and then subtract `steer` from that angle. The expression obtained by AUTOSIM is

```
(STEER - (U(1) + L*U(2)) / SPEED)
```

Once the system is described to AUTOSIM, the equations of motion are derived by a function named `dynamics`, and the simulation code is generated with a function `write-sim`. Sections of Fortran code generated by AUTOSIM are shown in some following figures. Before the code removal, the equations of motion include 83 intermediate variables. A portion of the code is shown in Figure 9 as it appears before unused intermediate variables are removed.

<pre> C Equations of Motion C ----- C C      C(3) = COS(Q(3)) C      S(3) = SIN(Q(3)) C C Kinematical equations C C      Z(1) = SPEED*C(3) C      Z(2) = U(1)*S(3) C      QP(1) = (Z(1) -Z(2)) C      Z(3) = U(1)*C(3) C      Z(4) = SPEED*S(3) C      QP(2) = (Z(3) + Z(4)) C      QP(3) = U(2) C      QP(4) = U(3) </pre>	<pre> C C Dynamical equations C C      Z(5) = Q(4)*U(2) C      Z(6) = THETAR*U(2) C      Z(7) = (U(3) + Z(6)) C      Z(8) = U(3)*Z(5) C      Z(9) = U(3)*U(2) C      Z(10) = CE*Q(4) C      ... C      Z(80) = PC(26)*Z(76) C      Z(81) = PC(39)*Z(79) C      Z(82) = (-Z(70) + Z(80) + Z(81)) C      Z(83) = PC(56)*Z(82) C      UP(1) = -Z(76) C      UP(2) = -Z(79) C      UP(3) = Z(83) </pre>
---	---

Figure 9. Portions of simulation code before code removal.

After unused code is removed and intermediate variables that appear but once are eliminated, only five intermediate variables are retained. The listing of the Fortran assignment statements actually written into the simulation code are shown in Figure 10. The indexed-sym objects that are printed as the Fortran arrays PC and Z are renumbered, so the specific array elements in the listing of Figure 10 are not the same as those in Figure 9.

The listing in Figure 10 requires 24 multiply/divides, 18 add/subtracts, and 2 function evaluations to compute the derivatives. 37 constant expressions were identified and are precomputed. The code to compute them is listed in Figure 11.

Details of this analysis and the computational efficiency of AUTOSIM have been presented elsewhere, and the simplification techniques were shown to influence the computational efficiency by almost a factor of 50: Using only the simplification techniques numbered 1,2,3, and 10, a total of 878 multiplies, divides, and function evaluations are needed to compute the derivatives at each time step. The best efficiency occurred using all techniques except no. 7, in which case the number of multiplies, divides, and function evaluations was reduced to 19 [17].

## Summary

Automated modeling of multibody systems has usually offered convenience over the alternative of formulating equations of motion by hand and writing a specialized simulation code to solve them. However, there have been trade-offs. Numerical

```

C Equations of Motion
C -----
C
C      C(3) = COS(Q(3))
C      S(3) = SIN(Q(3))
C
C Kinematical equations
C
C      QP(1) = (SPEED*C(3) -U(1)*S(3))
C      QP(2) = (U(1)*C(3) + SPEED*S(3))
C      QP(3) = U(2)
C      QP(4) = U(3)
C
C Dynamical equations
C
C      Z(1) = PC(34)*U(2)
C      Z(2) = (PC(16)*Q(4) + CROLL*U(3) + H*Z(1))
C      Z(3) = (PC(14) + PC(38)*Q(4) -PC(13)*U(1) -PC(12)*U(2) +
C      &      PC(1)*Z(1) -PC(19)*Z(2))
C      Z(4) = PC(31)*(PC(2) -PC(8)*Q(4) -PC(9)*U(1) + PC(35)*U(2) + Z(1)
C      &      -PC(20)*Z(2) -PC(25)*Z(3))
C      Z(5) = PC(32)*(Z(3) -PC(27)*Z(4))
C      UP(1) = -Z(4)
C      UP(2) = -Z(5)
C      UP(3) = (-PC(33)*Z(2) + PC(39)*Z(4) + PC(40)*Z(5))

```

Figure 10. Listing of code generated to compute derivatives of state variables for example vehicle model.

PC(1) = (CE + H*THETAR)	PC(19) = PC(17)/PC(18)
PC(2) = CA1*STEER	PC(20) = PC(7)/PC(18)
PC(3) = L*CA1/SPEED	PC(21) = PC(17)*PC(19)
PC(4) = RBM*GEES	PC(22) = (PC(15) -PC(21))
PC(5) = NRBM*SPEED	PC(23) = PC(17)*PC(20)
PC(6) = RBM*(CE + H*THETAR)	PC(24) = (PC(6) -PC(23))
PC(7) = H*RBM	PC(25) = PC(24)/PC(22)
PC(8) = (-CA2*KRS2 + CG1*CCOEF1)	PC(26) = PC(7)*PC(19)
PC(9) = (CA1 + CA2)/SPEED	PC(27) = (PC(6) -PC(26))
PC(10) = (RBM + NRBM)	PC(28) = PC(7)*PC(20)
PC(11) = (CAM2*KRS2 -	PC(29) = PC(25)*PC(27)
L*CG1*CCOEF1)	PC(30) = (PC(10) -PC(28) -PC(29))
PC(12) = L*(CAM1 + L*CA1)/SPEED	PC(31) = 1.0/PC(30)
PC(13) = (CAM1 + CAM2 +	PC(32) = 1.0/PC(22)
L*CA1)/SPEED	PC(33) = 1.0/PC(18)
PC(14) = STEER*(CAM1 + L*CA1)	PC(34) = RBM*SPEED
PC(15) = (IZZR + IZZNR + RBM*(CE	PC(35) = (NRBM*SPEED -L*CA1/SPEED)
+ H*THETAR)**2)	PC(36) = H*RBM/PC(18)
PC(16) = (KROLL -H*RBM*GEES)	PC(37) = (IXZ + IXX*THETAR +
PC(17) = (IXZ + IXX*THETAR +	H*RBM*(CE +
H*RBM*(CE + H*THETAR))	H*THETAR))/PC(18)
PC(18) = (IXX + RBM*H**2)	

Figure 11. Listing of code generated to precompute constants for example vehicle model.



generalized simulation codes have been much less efficient, because many of the simplification methods routinely used by a human analyst are so specific to the system being analyzed that they cannot be incorporated into a generalized formulation. Also, some types of subcomponent models are difficult or impossible to include in the system description. Symbolic multibody programs have offered better efficiency than the generalized numerical codes, but they have not been capable of providing "complete" solutions. The user must still develop expressions for many active forces and moments by hand, and incorporate them correctly into the portion of the code generated automatically.

Methods have been presented for representing all of the components of a simulated multibody system in symbolic form on the computer. Data objects are defined for representing (1) symbolic algebraic expressions for vector/dyadic analyses, (2) physical components in a multibody system, and (3) program structures needed in a simulation code. A language called AUTOSIM has been written in Lisp to implement these methods. When all of these objects are available for computer manipulation, the same modeling and programming strategies employed by humans can be mimicked in computer software. An example vehicle handling model is used to illustrate how forces and moments basic to vehicle simulations are described in this language, and how the symbolic computation is combined with Kane's dynamics analysis formalism to generate a working simulation code for that model.

With these methods, an automated modeling capability now exists that combines the convenience of a "complete" solution (associated with a generalized simulation code) with the efficiency and specialization possible when simulation codes are developed by hand.

## Acknowledgements

The work described in this paper was funded by the U.S. Army Tank and Automotive Research Command (TACOM). The author is grateful for the help and suggestions of R. Wehage, C. Mousseau, and A. Stribersky in performing this work.

## References

1. Chace, M. A., "Methods and Experience in Computer Aided Design of Large-Displacement Mechanical Systems." *Computer Aided Analysis and Optimization of Mechanical System Dynamics*, E. J. Haug, ed., NATO ASI Series, Series F, Vol. 9, Springer-Verlag, Heidelberg, 1984, pp. 233-259.
2. Nkravesh, P. E., and I. S. Chung, "Application of Euler Parameters to the Dynamic Analysis of Three Dimensional Constrained Mechanical Systems." *ASME Journal of Mechanical Design*, Vol. 104, No. 4, Oct. 1982, pp. 785-791
3. Nielan, P.E. and Kane, T.R., "Symbolic Generation of Efficient Simulation/Control Routines for Multibody Systems." *Dynamics of Multibody Systems*. IUTAM/IFTOMM Symposium Udine 1985. Editors: G. Bianchi and W. Schiehlen, Springer Berlin Heidelberg 1986. pp. 153-164.
4. Krishnaswami P., Bhatti, M.A., "Symbolic Computing in Optimal Design of Dynamic Systems." Presented at the ASME Design Engineering Division Conference and Exhibit on Mechanical Vibration and Noise. Cincinnati, Ohio, Sept 10-13, 1985.

5. Kortüm, W., and Schiehlen, W. "General Purpose Vehicle System Dynamics Software Based on Multibody Formalisms." *Vehicle System Dynamics*, vol 14 pp. 229-263, June 1985.
6. Rosenthal, D. E., and M. A. Sherman, "Symbolic Multibody Equations via Kane's Method." AAS-Paper 83-303, AAS Publ. Office, San Diego, 1983.
7. Wittenburg, J. Wolz, U., "MESA VERDE: A Symbolic Program for Nonlinear Articulated-Rigid-Body Dynamics." Proceeding, ASME Design Eng. Div., Conference and Exhibition on Mechanical Vibration and Noise, Sept. 1985.
8. Nielan, P.E. "Efficient Computer Simulation of Motions of Multibody Systems." PhD Dissertation, Stanford University, Sept. 1986.
9. Kane, T.R., and Levinson, D.A., *Dynamics: Theory and Applications*. McGraw-Hill Book company, New York, 1985
10. Featherstone, W.R., "Robot Dynamics Algorithms." PhD Dissertation, University of Edinburgh, 1984.
11. Wehage, R.A. "Application of Matrix Partitioning and Recursive Projection to Order-n Solution of Constrained Equations of Motion." *Proceeding of the 20th Biennial ASME Mechanism Conference*, Orlando, Florida, 1988.
12. Bae, D.S., Haug, E.J., "A Recursive formulation for Constrained Mechanical System Dynamics: Part I, Open Loop Systems." *Mech Structures and Machines* 15 (3), 1987, pp 359-382.
13. Steele, G.L., *Common Lisp: The Language*. Digital Press, Bedford Mass., 1984.
14. Huston, R.L., and Passerello, C.E., "Multibody Structural Dynamics Including Translation Between the Bodies." *Computers and Structures*, Vol 12, pp 713-720.
15. Stribersky, A., Fancher, P.S., MacAdam, C.C., Sayers, M.W., "On Nonlinear Oscillations in Road Trains at High Forward Speeds." *Proceedings, 11th IAVSD Symposium on the Dynamics of Vehicles on Roads and Tracks, Kingston, Ontario, CANADA, August 21-25 1989.* (in press).
16. Segel, L. "Theoretical Prediction and Experimental Substantiation of the Response of the Automobile to Steering Control." *Proceedings, the Automobile Division of the Institution of Mechanical Engineers*, No. 7, 1956-57.
17. Sayers, M.W. "Automated Formulation of Efficient Vehicle Simulation codes by Symbolic Computation (AUTOSIM)." *Proceedings, 11th IAVSD Symposium on the Dynamics of Vehicles on Roads and Tracks, Kingston, Ontario, CANADA, August 21-25 1989.* (in press).

# EFFECT OF THERMAL SOFTENING ON THE RESPONSE OF SHEARING MOTIONS

Athanasios E. Tzavaras  
Department of Mathematics  
University of Wisconsin-Madison  
Madison, WI 53706

**ABSTRACT.** Shear instabilities in the form of shear bands are often observed when metals are deformed at high strain rates. According to one theory, shear band formation is attributed to a destabilizing feedback mechanism induced by thermal softening properties of materials. In this note we review some mathematical results on simple model problems, with the goal of assessing the effect of various contributing factors (like thermal softening, strain hardening and strain-rate sensitivity) on the response of shear motions.

## 1. INTRODUCTION.

One of the most striking manifestations of instability in solid mechanics is the localization of plastic deformation and the consequent formation of shear bands, observed during torsional tests, at high strain rates, on steels (e.g. [1], [6], [9], [10], [17]). According to one popular theory formation of shear bands is attributed to a thermoplastic instability mechanism that is induced by thermal softening properties of materials (i.e., the property that plastic flow is enhanced with temperature increase). The argument goes as follows: Non-uniform straining induces non-uniform heating, which, in turn, enhances the plastic flow at hotter regions and reduces it a colder regions. This creates a destabilizing feedback mechanism which tends to create localization of plastic deformation and formation of shear bands. To the above destabilizing mechanism there is opposition from internal dissipation and, possibly, from strain hardening properties of the material. Whether localization will occur depends on the relative weight of thermal softening, strain hardening and strain-rate sensitivity. The intent of this work is to elucidate the interplay of thermal softening and strain hardening in shearing deformations of strain-rate dependent materials and to provide quantitative criteria for stability as well as instability.

As a test problem we consider the adiabatic, plastic shearing of an infinite plate of unit thickness. The plate is subjected to either steady shearing or prescribed tractions at the boundaries. In a Cartesian coordinate system the infinite plate occupies the region between the planes  $x = 0$  and  $x = 1$ . The thermomechanical process is described by the velocity field in the shearing direction  $v(x, t)$ , the shear strain  $\gamma(x, t)$ , the shear stress  $\sigma(x, t)$ , the temperature  $\theta(x, t)$  and the heat flux  $q(x, t)$ . We assume that the referential density and the specific heat are constants, taken equal to one and that the elastic effects are negligible. Then the balance laws of momentum, energy and the kinematic compatibility relation read

$$v_t = \sigma_x \tag{1.1}$$

$$\theta_t = q_x + \sigma \gamma_t \tag{1.2}$$

$$\gamma_t = v_x \quad (1.3)$$

The above equations are supplemented with constitutive laws for the heat flux,

$$q = 0, \quad (1.4)$$

corresponding to the assumption that the process is adiabatic, and for the stress,

$$\sigma = f(\theta, \gamma, \gamma_t). \quad (1.5)$$

The constitutive assumption (1.5) is appropriate for a solid in the plastic region exhibiting thermal softening ( $f_\theta(\theta, \gamma, \gamma_t) < 0$ ), strain hardening ( $f_\gamma(\theta, \gamma, \gamma_t) > 0$ ) and strain rate sensitivity ( $f_{\gamma_t}(\theta, \gamma, \gamma_t) > 0$ ). At the present time there is not sufficient theoretical or experimental evidence to indicate precisely the form of the function  $f(\theta, \gamma, \gamma_t)$ . Several choices have been used in the literature. An example is the empirical power law (e.g. [6])

$$\sigma = \theta^\nu \gamma^m \gamma_t^n, \quad \nu < 0, m > 0, n > 0. \quad (1.6)$$

Several studies of (1.1) - (1.5) under various types of loading have appeared recently in the mathematical literature [2-4], [8], [13-15]. They mainly deal with special instances of the constitutive law (1.6). In this note we present a survey of these results with two objectives:

- (a) To lay out the stabilizing or destabilizing influence of the various factors associated with the problem (like thermal softening, strain hardening and strain-rate sensitivity) by studying completely the special case of the power law. This is done in Section 2.
- (b) To give some preliminary answers to the question: "How to define mathematically a shear band?" This question is undertaken in Section 3.

## 2. THE POWER LAW

We consider the system of partial differential equations

$$v_t = (\theta^\nu \gamma^m |v_x|^{n-1} v_x)_x \quad (2.1)$$

$$\theta_t = \theta^\nu \gamma^m |v_x|^{n+1} \quad (2.2)$$

$$\gamma_t = v_x \quad (2.3)$$

$$\sigma = \theta^\nu \gamma^m |\gamma_t|^{n-1} \gamma_t, \quad \nu < 0, n > 0, m \in \mathbb{R}, \quad (2.4)$$

where  $0 \leq x \leq 1, t > 0$ , together with boundary conditions

$$v(0, t) = 0, \quad v(1, t) = 1 \quad (BCV)$$

or

$$\sigma(0, t) = 0, \quad \sigma(1, t) = 1 \quad (BCS)$$

and initial conditions

$$v(x, 0) = v_0(x), \quad \theta(x, 0) = \theta_0(x), \quad \gamma(x, 0) = \gamma_0(x), \quad 0 \leq x \leq 1. \quad (2.5)$$

The system (2.1) – (2.3) consists of the degenerate parabolic equation (2.1) in  $v$ , coupled through the diffusion coefficient  $\theta^\nu \gamma^m$  with the hyperbolic equations (2.2) and (2.3). Equations (2.1) – (2.3) are based on the power law (2.4), which is appropriate for a material exhibiting thermal softening ( $\nu < 0$ ), strain hardening ( $m > 0$ ) or strain softening ( $m < 0$ ), and strain rate sensitivity. The shearing is caused by steady shearing at the boundary or by prescribed tractions at the boundaries in case (BCV) or (BCS) hold, respectively.

The relevant question is to study the behavior of solutions of the equations (2.1) – (2.3) together with (BCV) (or (BCS)) and (2.5) for different values of the parameters  $\nu, m$  and  $n$ . From the point of view of analysis, the key question is whether the diffusion coefficient in equation (2.1) tends to zero relatively slowly and in an “orderly” fashion, or whether nonuniformities develop and the material exhibits unstable response. A goal of the analysis is to provide quantitative criteria for stability and instability in terms of the parameters  $\nu, m$  and  $n$  and in the case of instability to determine whether shear bands form.

The system (2.1) – (2.3) is invariant under a group of scaling transformations (cf. [12]). If  $\{v(x, t), \theta(x, t), \gamma(x, t)\}$  is a solution of (2.1) – (2.3) on  $\mathbb{R} \times (0, \infty)$ , then the rescaled functions  $\{v_{(\lambda)}(x, t), \theta_{(\lambda)}(x, t), \gamma_{(\lambda)}(x, t)\}$  defined by

$$v_{(\lambda)}(x, t) = \lambda^{\frac{\delta}{\alpha}} v(\lambda x, \lambda^{-\frac{1}{\alpha}} t) \quad (2.6)$$

$$\theta_{(\lambda)}(x, t) = \lambda^{2\frac{\delta}{\alpha}} \theta(\lambda x, \lambda^{-\frac{1}{\alpha}} t) \quad (2.7)$$

$$\gamma_{(\lambda)}(x, t) = \lambda^{\frac{\alpha+\delta+1}{\alpha}} \gamma(\lambda x, \lambda^{-\frac{1}{\alpha}} t) \quad (2.8)$$

where  $\lambda > 0$  and  $\delta, \alpha$  are any constants with

$$2\nu\delta + m(\alpha + \delta + 1) + n(\alpha + \delta) + \alpha - \delta + 1 = 0 \quad (2.9)$$

is again a solution. It is shown in [12] that the scaling invariance induces self-similar solutions which blow up when  $\nu + m + n < 0$ . Although these self-similar solutions blow up at the boundary, they indicate the existence of a destabilizing mechanism induced by the variable diffusion coefficient in (2.1).

In what follows we describe some recent results towards studying the asymptotic behavior of (2.1) – (2.5).

(a) *Velocity Boundary Conditions* (BCV)

The system of equations (2.1) – (2.3) together with boundary conditions (BCV) admits the class of solutions

$$\bar{v}(x, t) = x \quad (2.10)$$

$$\bar{\gamma}(x, t) = t + \Gamma_0 \quad (2.11)$$

$$\bar{\Theta}(x, t) = \left\{ \Theta_0^{1-\nu} + \frac{1-\nu}{m+1} [(t + \Gamma_0)^{m+1} - \Gamma_0^{m+1}] \right\}^{\frac{1}{1-\nu}} \quad (2.12)$$

where  $\Gamma_0$  and  $\Theta_0$  are positive constants. These solutions represent uniform shearing. The relevant question is whether as  $t$  increases  $v_x(x, t)$ ,  $\theta(x, t)$  and  $\gamma(x, t)$  develop substantial nonuniformities or else, they approach the uniform shearing solutions as  $t \rightarrow \infty$ .

This question remains open in the general case. However, some answers have been given in the special cases  $m = 0$  and  $\nu = 0$ ; in each of these, one of the equations decouples from the rest. Specifically, it is known that if (i)  $\nu < 0, \nu + n < 0, m = 0$ , or if (ii)  $m < 0, m + n > 0, \nu = 0$ , the uniform the uniform shearing solutions are asymptotically stable ([8], [13], [14], [16]).

(b) *Stress boundary conditions* (BCS)

The main difficulty in the general case (i.e. both  $\nu$  and  $m$  nonzero) is that the diffusion coefficient  $\theta^\nu \gamma^m$  in the parabolic equation (2.1), governing the asymptotic distribution of  $v_x$ , does not have an a-priori trend of growth or decay, as was the situation in the special cases  $m = 0$  or  $\nu = 0$ . However, this difficulty can be circumvented in the case of stress boundary conditions.

For technical simplicity, we restrict attention to the case  $n = 1$ . System (2.1) - (2.3) is equivalent (cf. [15]) to the system of reaction diffusion equations:

$$\sigma_t = \theta^\nu \gamma^m \sigma_{xx} + \frac{\sigma^2}{\theta^\nu \gamma^{m+1}} (\nu \frac{\sigma \gamma}{\theta} + m) \quad (2.13)$$

$$\theta_t = \frac{\sigma^2}{\theta^\nu \gamma^m} \quad (2.14)$$

$$\gamma_t = \frac{\sigma}{\theta^\nu \gamma^m} \quad (2.15)$$

with boundary conditions (BCS) and initial conditions

$$\sigma(x, 0) = \sigma_0(x) \geq 0, \quad \theta(x, 0) = \theta_0(x) > 0, \quad \gamma(x, 0) = \gamma_0(x) > 0. \quad (2.16)$$

For these initial conditions, it turns out that  $\sigma(x, t) \geq 0, \theta(x, t) \geq \theta_0(x) > 0, \gamma(x, t) \geq \gamma_0(x) > 0, 0 \leq x \leq 1, t \geq 0$ , and the above initial-boundary value problem is well posed locally in time (in Schauder spaces).

Moreover, viewing (2.13) as a parabolic equation in  $\sigma$  with coefficients governed by (2.14), (2.15), one can use comparison principles for (2.13) and obtain a-priori bounds for  $\sigma(x, t)$  provided  $\nu + m < 0$ . These estimates open the way to showing that the parameter space can be separated into three distinct regions, namely,  $\nu + m < -1, -1 \leq \nu + m \leq -\frac{1}{2}$  and  $-\frac{1}{2} < \nu + m < 0$  across which the behavior of solutions changes drastically (cf. [15]).

1) In the region  $\nu + m < -1$  smooth solutions break down in finite time  $T^*$ , for any initial data. As  $t \rightarrow T^*$ ,  $\sup_{0 \leq x \leq 1} \gamma(x, t), \sup_{0 \leq x \leq 1} v_x(x, t)$  and  $\sup_{0 \leq x \leq 1} \theta(x, t)$  (the latter in case  $m > -1$ ) tend to  $\infty$ , in such a way that  $\sigma = \theta^\nu \gamma^m v_x$  remains bounded.

2) In the region  $-1 \leq \nu + m < 0$  smooth solutions exist globally in time for any initial data.

3) If in addition  $-\frac{1}{2} < \nu + m < 0$ , solutions stabilize as  $t \rightarrow \infty$  and asymptotically they behave as follows:

$$\sigma(x, t) = x + O(t^{-\frac{2(\nu+m)+1}{\nu+m+1}}) \quad (2.17)$$

$$\theta(x, t) = x\gamma(x, t) + O(t^{-\frac{\nu+m}{\nu+m+1}}) \quad (2.18)$$

$$v(x, t) = t + O(t^{-\frac{\nu+m}{\nu+m+1}}). \quad (2.19)$$

4) In the special case  $m = 0$ , in the in-between region  $-1 \leq \nu \leq -\frac{1}{2}$ , (3.17) - (3.19) are in general no longer satisfied. More precisely, given any  $\varepsilon > 0$ , there are initial data  $\sigma_0(x)$ ,  $\theta_0(x)$  and  $\gamma_0(x)$  such that  $|\sigma_0(x) - x| < \varepsilon$ ,  $0 \leq x \leq 1$ , but  $\sigma(x, t)$  does not converge to  $x$  as  $t \rightarrow \infty$ .

Finally, for  $m = 0$ ,  $\nu < -1$  there are solutions that blow up only at the boundary  $x = 1$  and look like shear bands. Currently, the above results are being extended to cover the general case  $n$  arbitrary [16].

(c) *Other boundary conditions.*

Charalambakis [2], [3] and Charalambakis and Houstis [4] consider (2.1) - (2.4), in the special case  $\nu = 0$  or  $m = 0$  in situations when the shearing is caused by other boundary conditions or by body forces; they establish asymptotic stability results in these cases, under restrictions on the range where the parameters vary.

### 3. SHEAR BANDS

In this Section we discuss the following question: "How to approach shear bands from an analysis viewpoint?" Traditionally, the formation of shear bands is associated with some type of development of nonuniformities in the field variables of the problem (e.g. [6], [9], [17]). It is however debatable whether a "slowly" evolving nonuniformity should be termed as a shear band. Thus we think that the aforementioned question has also interesting practical implications.

As a point of reference for this discussion we will use the simple model

$$v_t = (\mu(\theta)v_x)_x \quad (3.1)$$

$$\theta_t = \mu(\theta)v_x^2 \quad (3.2)$$

with boundary conditions (BCV) or (BCS) and initial conditions

$$\theta(x, 0) = \theta_0(x), v(x, 0) = v_0(x), 0 \leq x \leq 1 \quad (3.3)$$

with  $\theta_0(x) > 0$ ,  $v_{0,x}(x) > 0$ ,  $0 < x < 1$ . Recall that

$$\sigma = \mu(\theta)v_x \quad (3.4)$$

with  $\mu(\theta) > 0$ ,  $\mu'(\theta) < 0$ .

For this system in the case of stress boundary conditions (BCS), it turns out [15] that there is a unique classical solution defined on a maximal interval of existence  $[0, 1] \times [0, T^*)$ . Moreover, if  $T^* < \infty$

$$\lim_{t \uparrow T^*} \sup_{0 \leq x \leq 1} \theta(x, t) = \infty \quad (3.5)$$

and

$$\limsup_{t \uparrow T^*} \sup_{0 \leq x \leq 1} v_x(x, t) = \infty \quad (3.6)$$

In case  $\int_1^\infty \mu(\xi)d\xi = \infty$  then  $T^* = +\infty$ , while, in case  $\int_1^\infty \mu(\xi)d\xi < \infty$  then  $T^* < +\infty$ .

In the special case

$$\mu(\theta) = \theta^\nu \quad (3.7)$$

the parameter space  $\nu \leq 0$  is decomposed into three regions: (a)  $\nu < -1$  a region of blowup, (b)  $-1 \leq \nu \leq -\frac{1}{2}$  a region of global existence but where nonuniformities of the initial data persist and (c)  $-\frac{1}{2} < \nu \leq 0$  a region where dissipation is predominant and leads to stable response (cf. Section 2).

It can be seen by the relevant analysis [15] that the region  $\nu < -1$  is clearly associated with development of shear bands; also the region  $-1 \leq \nu \leq -\frac{1}{2}$  is associated with nonuniform response.

Further than the above type of response, there is a different more subtle type of response which may be associated with localization and formation of shear bands. This possibility was pointed out to the author by Dr. T. Wright. Namely, it is conceivable that large time results describe the predominance of dissipation as  $t \rightarrow \infty$ ; but, maybe at intermediate times large structures develop and then they get washed out because of the dissipation. This is another possible scenario that needs to be investigated.

To pursue the subtleties of this question one step further, consider the system (3.1) - (3.4) with velocity boundary conditions (BCV) in the special case  $\mu(\theta) = \theta^\nu$ ,  $\nu < 0$ . In this case the uniform shearing solutions  $(\bar{v}(x), \Theta(t))$  are

$$v(x) = x, \Theta(t) = [\Theta_0^{1-\nu} + (1-\nu)t]^{\frac{1}{1-\nu}},$$

corresponding to initial data  $(v_0(x), \theta_0(x)) = (x, \Theta_0)$ . It is shown in [8] that if  $[v(x, t), \theta(x, t)]$  is any solution of (3.1) - (3.4) and (BCV) then

$$v_x(x, t) = 1 + O(t^{-\frac{\nu+1}{1-\nu}}) \quad (3.8)$$

$$\frac{1}{1-\nu} \theta^{1-\nu}(x, t) = t + O(t^{1-\frac{\nu+1}{1-\nu}}) \quad (3.9)$$

as  $t \rightarrow \infty$ . Moreover, if  $(v_0(x), \theta_0(x))$  is a small perturbation of  $(x, \Theta_0)$ , for some  $\Theta_0 > 0$ , one obtains in [16] that

$$|\theta(x, t) - \Theta(t)| \leq \delta(1+t)^{\frac{1}{1-\nu}} \quad (3.10)$$

where  $\delta$  is of the order of magnitude of the initial perturbation. In other words the distance of the solution  $\theta(x, t)$  from the uniform shearing solution  $\Theta(t)$  is controlled by the rate of growth of  $\Theta(t)$  and the initial perturbation. Nevertheless, it is still possible that this difference grows but at a slower rate. The relevant question here is when do we call a time dependent solution asymptotically stable, if the perturbation grows at a slower rate than the basic solution or if the perturbation decays. Both answers are legitimate as far as mathematical definitions are concerned but they have different implications on when to call a process stable and, for this particular problem, on what to define as nonuniform response and shear band. Further analysis, as well as numerical experimentation on simple models, are needed in order to answer this question.

## References

1. T. Burns, "A Mechanism for Shear Band Formation in the High Strain Rate Torsion Test," Institute of Standards and Technology Report #89-4121, Gaithersburg, MD. 1989.



- 2 . N. Charalambakis, "Adiabatic Shearing Flow Caused by Time Dependent Inertial Force," *Quart. Appl. Math.* **XLII** (1984), pp. 275-280 .
- 3 . N. Charalambakis, "Time-Asymptotic Stability of Non-Newtonian Fluid or Plastic Solid," *Mech. Res. Comm.* **12** (1985), pp. 311-317.
- 4 . N. Charalambakis and E. Houstis, "Adiabatic Shearing of One-Dimensional Thermo-viscoelastic Flows Caused by Boundary and Inertial Forces," *Eng. Anal.* **2** (1985), pp. 205-210.
- 5 . K. Chueh, C. Conley, and J. Smoller , "Positively Invariant Regions for Systems of Nonlinear Diffusion Equations," *Indiana Univ. Math. J.* **26** (1977), pp. 372-411 .
- 6 . R. Clifton, J. Duffy, K. Hartley, and T. Shawki , "On Critical Conditions for Shear Band Formation at High Strain Rates ," *Metallurgica* **18** (1984), pp. 443-448 .
- 7 . C. Dafermos, "Contemporary Issues in Dynamic Behavior of Continuous Media," LCDS Lecture Notes No. 85-1 , Brown Univ., 1985.
- 8 . C. Dafermos and L. Hsiao, "Adiabatic Shearing of Incompressible Fluids with Temperature Dependent Viscosity," *Quart. Appl. Math.* **XLI** (1983), pp. 45-58.
- 9 . D. Drew and J. Flaherty, "Adaptive Finite Element Methods and the Numerical Solution of Shear Band Problems," in *Phase Transformations and Material Instabilities in Solids*, ed. M. Gurtin, Academic Press, New York, 1984.
- 10 . A. Molinari and R. Clifton, "Analytical Characterization of Shear Localization in Thermoviscoplastic Materials," *J. Appl. Mech.* **109** (1987).
- 11 . T. Shawki, *Ph.D. Thesis*, Brown Univ., Providence, 1985.
- 12 . A. Tzavaras, *Ph.D. Thesis*, Brown Univ., Providence, 1985.
- 13 . A. Tzavaras, "Shearing of Materials Exhibiting Thermal Softening or Temperature Dependent Viscosity," *Quart. Appl. Math.* **XLIV** (1986), pp. 1-12 .
- 14 . A. Tzavaras, "Plastic Shearing of Materials Exhibiting Strain Hardening or Strain Softening," *Arch. Rat. Mech. Anal.* **94** (1986), pp. 39-58 .
- 15 . A. Tzavaras. "Effect of Thermal Softening and Strain Hardening in Shearing of Strain Rate Dependent Materials," *Arch. Rat. Mech. Anal.* **99** (1987), pp. 349-374 .
- 16 . A. Tzavaras. "Interplay of Thermal Softening and Strain-Rate Sensitivity on the Response of Shearing Motions," , in preparation.
- 17 . T. Wright and J. Walter, "On Stress Collapse in Adiabatic Shear Bands." *J. Mech. Phys. Solids.* **35** (19-).

**ELASTIC-PLASTIC ANALYSIS OF A THICK-WALLED COMPOSITE TUBE  
SUBJECTED TO INTERNAL PRESSURE**

Peter C. T. Chen  
U.S. Army Armament Research, Development, and Engineering Center  
Close Combat Armaments Center  
Benet Laboratories  
Watervliet, NY 12189-4050

**ABSTRACT.** This paper presents an elastic-plastic analysis of a thick-walled composite tube subjected to internal pressure. The composite tube is constructed of a steel liner and a graphite-bismaleimide outer shell. Analytical expressions for stresses, strains, and displacements are derived for all cases where the structure is subjected to internal pressure. The loading ranges include elastic, elastic-plastic, and fully-plastic up to failure. Numerical results for the hoop strains in several composite tubes are presented.

**INTRODUCTION.** Organic composites have become familiar structural components in many applications that require high stiffness and low weight. A current problem in Army cannon design is to replace a portion of the steel wall thickness with an organic composite. The steel liner maintains the tube projectile interface and shields the composite from the extremely hot gases. The steel also has elastic properties in the radial direction that are better than the composites for transferring loads. The theoretical and experimental results for an organic composite-jacketed steel tube subjected to internal pressure in the elastic range were reported in a recent paper [1]. This paper presents an elastic-plastic analysis of the composite tube problem. The composite tube is constructed of a steel liner and a graphite-bismaleimide outer shell. Analytical expressions for stresses, strains, and displacements are derived for loading within and beyond the elastic range up to failure.

**ELASTIC RANGE.** Figure 1 shows a schematic of the composite tube problem. The composite tube consists of an inner steel "liner" and an outer composite "jacket." The steel liner of inside radius  $a$  and outer radius  $b$  is wrapped in the circumferential direction with a graphite-bismaleimide organic composite of outside radius  $c$ . The elastic material constants for the composite and the steel are given in Table 1.

TABLE 1. ELASTIC CONSTANTS OF COMPOSITE JACKET AND STEEL LINER

Elastic Constants for IM6/Bismaleimide, 55% F.V.R.		
$E_r = 1.126$ Mpsi	$\nu_{r\theta} = 0.01524$	$\nu_{\theta r} = 0.3155$
$E_\theta = 23.31$ Mpsi	$\nu_{\theta z} = 0.3155$	$\nu_{z\theta} = 0.01524$
$E_z = 1.126$ Mpsi	$\nu_{zr} = 0.3991$	$\nu_{rz} = 0.3911$
Elastic Constants for Steel		
$E = 30.0$ Mpsi	$\nu = 0.3$	

$$\frac{u_b}{b} = q \left[ k \alpha_{22} \frac{(c/b)^{2k} + 1}{(c/b)^{2k} - 1} - \alpha_{12} \right] \quad (11)$$

$$\frac{u_c}{c} = \frac{2q k \alpha_{22} (c/b)^{k-1}}{(c/b)^{2k} - 1} \quad (12)$$

**ELASTIC-PLASTIC RANGE.** When the internal pressure  $p$  is large enough, part of the steel liner will become plastic. Using Tresca's yield criterion, the associated flow rule, and assuming linear strain-hardening, the elastic-plastic solution based on Bland can be used [2,3]. Let  $\rho$  be the elastic-plastic interface. The solution can be written in the elastic portion ( $\rho \leq r \leq b$ ) as

$$\frac{E}{\sigma_0} \frac{u}{r} = \frac{1+\nu}{2} \frac{\rho^2}{r^2} + (1-\nu-2\nu^2) \left[ \frac{1}{2} \frac{\rho^2}{b^2} - \frac{q}{\sigma_0} \right] \quad (13)$$

$$\frac{\sigma_r}{\sigma_0} = \mp \frac{\rho^2}{r^2} + \frac{\rho^2}{b^2} - \frac{q}{\sigma_0} \quad (14)$$

$$\sigma_{\theta} \quad (15)$$

$$\frac{\sigma_z}{\sigma_0} = \nu \rho^2 / b^2 - 2\nu q / \sigma_0 \quad (16)$$

and in the plastic portion ( $a \leq r \leq \rho$ )

$$\frac{E}{\sigma_0} \frac{u}{r} = (1-\nu-2\nu^2) \frac{\sigma_r}{\sigma_0} + (1-\nu^2) \frac{\rho^2}{r^2} \quad (17)$$

$$\frac{\sigma_r}{\sigma_0} = \mp \frac{1}{2} (1-\eta\beta + \eta\beta \frac{\rho^2}{r^2}) + \frac{1}{2} \frac{\rho^2}{b^2} - (1-\eta\beta) \ln \frac{\rho}{r} - \frac{q}{\sigma_0} \quad (18)$$

$$\sigma_{\theta} / \sigma_0 \quad (19)$$

$$\frac{\sigma_z}{\sigma_0} = \nu \rho^2 / b^2 - 2\nu(1-\eta\beta) \ln \frac{\rho}{r} - 2\nu q / \sigma_0 \quad (20)$$

$$\bar{\epsilon}^p = \beta(\rho^2/r^2 - 1) \quad , \quad \eta\beta = \frac{m}{m + \frac{3}{4} \frac{(1-m)}{(1-\nu^2)}} \quad (21)$$

$$\eta = \frac{2}{\sqrt{3}} \frac{E}{\sigma_0} \frac{m}{1-m} \quad , \quad m = \frac{E_t}{E} \quad , \quad \sigma = \sigma_0(1 + \eta \bar{\epsilon}^p) \quad (22)$$

where  $\sigma_0$  is the initial tensile yield stress and  $E_t$  is the tangent modulus in the plastic range of the stress-strain curve.

Using Eqs. (11) and (13) and the requirement of displacement continuity at the interface, i.e.,  $u_{b-}$  (liner) =  $u_{b+}$  (jacket), we obtain the expression for the interface pressure  $q$  as

$$\frac{g_-}{\sigma_0} = \frac{(1-\nu^2)\rho^2/b^2}{(1+\nu)(1-2\nu) + E[\alpha_{22}k \frac{(c/b)^{2k} + 1}{(c/b)^{2k} - 1} - \alpha_{12}]} \quad (23)$$

Given any value of  $\rho$  in  $a \leq \rho \leq b$ , we can now determine  $q$ ,  $u$ , and all the stresses and strains in the tube. In particular, the expressions for the internal pressure, the displacements at the bore and the interface are

$$\frac{p_-}{\sigma_0} = \frac{g_-}{\sigma_0} + \frac{1}{2} \left(1 - \frac{\rho^2}{b^2}\right) + (1-\eta\beta) \ln \frac{\rho}{a} + \frac{1}{2} \eta\beta \left(\frac{\rho^2}{a^2} - 1\right) \quad (24)$$

$$\frac{E_-}{\sigma_0} \frac{u_a}{a} = - (1-\nu-2\nu^2) \frac{p_-}{\sigma_0} + (1-\nu^2)\rho^2/a^2 \quad (25)$$

$$\frac{E_-}{\sigma_0} \frac{u_b}{b} = (1-\nu^2) \frac{\rho^2}{b^2} - (1-\nu-2\nu^2) \frac{g_-}{\sigma_0} \quad (26)$$

By letting  $\rho = a$  and  $b$ , we can determine the lower limits  $p^*$ ,  $q^*$ ,  $u_a^*$ ,  $u_b^*$ ,  $u_c^*$ , and the upper limits  $p^{**}$ ,  $q^{**}$ ,  $u_a^{**}$ ,  $u_b^{**}$ ,  $u_c^{**}$ , respectively.

**FULLY-PLASTIC RANGE.** When the internal pressure  $p$  is further increased, i.e.,  $p > p^{**}$ , the steel liner will become fully-plastic. The composite jacket remains elastic as long as the failure pressure is not reached. In this section, a fully-plastic solution is derived here.

Subject to  $\sigma_\theta \geq \sigma_z \geq \sigma_r$ , Tresca's criterion states that yielding occurs when

$$\sigma_\theta - \sigma_r = \sigma \quad (27)$$

where  $\sigma$  is the yield stress. For a linear strain-hardening material,

$$\sigma = \sigma_0(1+\eta\bar{\epsilon}^p) \quad (28)$$

where  $\sigma_0$ ,  $\eta$ ,  $\bar{\epsilon}^p$ , are the initial yield stress, hardening parameter, and equivalent plastic strain, respectively. The associated flow rule states that, subject to  $\sigma_\theta > \sigma_z > \sigma_r$ ,

$$d\epsilon_\theta^p = -d\epsilon_r^p \geq 0 \quad \text{and} \quad d\epsilon_z^p = 0 \quad (29)$$

$d\epsilon_z^p$  is an increment of plastic strain and is defined by  $d\epsilon_z^p = d\epsilon_z - d\epsilon_z^e$ . Since  $d\epsilon_z^p = 0$ ,  $d\epsilon_z = d\epsilon_z^e$ , and therefore

$$\epsilon_z = \epsilon_z^e = (\sigma_z - \nu\sigma_r - \nu\sigma_\theta)/E \quad (30)$$

In the plane-strain case ( $\epsilon_z = 0$ ) and using the equation of equilibrium,

$$\sigma_\theta = \sigma_r + r\sigma_r' \quad \text{and} \quad \sigma_r' = d\sigma_r/dr \quad (31)$$

we have

$$\sigma_z = 2\nu\sigma_r + \nu r\sigma_r' \quad (32)$$

Since the dilatation is purely elastic

$$u' + u/r + \epsilon_z = E^{-1}(1-2\nu)(\sigma_r + \sigma_\theta + \sigma_z) \quad (33)$$

Substituting from Eqs. (31) and (32)

$$u' + u/r = E^{-1}(1-2\nu)(1+\nu)(2\sigma_r + r\sigma_r') \quad (34)$$

On integration,

$$ru = E^{-1}(1-2\nu)(1+\nu)r^2\sigma_r + \phi b^2 \quad (35)$$

where

$$\phi = u_b/b + (1-2\nu)(1+\nu)E^{-1}q \quad (36)$$

Using Hooke's law and Eqs. (27), (28), (31), and (32), we obtain

$$E\epsilon_\theta^e = (1-2\nu)(1+\nu)\sigma_r + (1-\nu^2)\sigma_o(1+\eta\bar{\epsilon}^P) \quad (37)$$

Substituting from Eq. (35) for  $\epsilon_\theta$  and from Eq. (37) for  $\epsilon_\theta^e$

$$\epsilon_\theta^P = \epsilon_\theta - \epsilon_\theta^e = \phi b^2/r^2 - E^{-1}(1-\nu^2)\sigma_o(1+\eta\bar{\epsilon}^P) \quad (38)$$

By Eq. (29) and the definition of equivalent plastic strain,

$$\bar{\epsilon}^P = \int d\bar{\epsilon}^P = \sqrt{\frac{2}{3}} \int \{(\epsilon_\theta^P)^2 + (\epsilon_r^P)^2\}^{1/2} = \frac{2}{\sqrt{3}} \epsilon_\theta^P \quad (39)$$

Combining Eqs. (38) and (39) leads to

$$\bar{\epsilon}^P = \frac{2}{\sqrt{3}} [\phi b^2/r^2 - (1-\nu^2)\sigma_o/E] / [1 + \frac{2}{\sqrt{3}} (1-\nu^2)\eta\sigma_o/E] \quad (40)$$

Substituting Eqs. (27) and (28) into Eq. (31) and integrating, we have

$$\sigma_r = -p + \sigma_o \ln(r/a) + \sigma_o \eta \int_a^r \bar{\epsilon}^P r^{-1} dr \quad (41)$$

Now using Eq. (40), an explicit expression for the radial stress is obtained

$$\sigma_r = -p + \sigma_o(1-\eta\beta) \ln\left(\frac{r}{a}\right) + \frac{1}{2} \frac{\eta\beta}{(1-\nu^2)} \left[\frac{b^2}{a^2} - \frac{b^2}{r^2}\right] E\phi \quad (42)$$

Using Eqs. (11) and (36) and the requirement of displacement continuity at the interface, i.e.,  $u_{b-}$  (liner) =  $u_{b+}$  (jacket), we get

$$\phi = [E\alpha_{22}k \frac{(c/b)^{2k} + 1}{(c/b)^{2k} - 1} - E\alpha_{12} + (1-2\nu)(1+\nu)]q/E \quad (43)$$

Evaluating  $\sigma_r$  at the interface from Eq. (42) and using Eq. (43), we obtain the relation between  $p$  and  $q$

$$p = \sigma_0(1-\eta\beta)\ln \frac{b}{a} + q\left\{1 + \frac{1}{2}\eta\beta\left(\frac{b^2}{a^2} - 1\right)\left[Ak \frac{(c/b)^{2k} + 1}{(c/b)^{2k} - 1} + B + 1\right]\right\} \quad (44)$$

It is interesting to point out that  $p$  is a linear function of  $q$ . Similarly, when evaluating  $u$  at the bore from Eq. (35), we obtain

$$u_a/a = -(1-2\nu)(1+\nu)P/E + \phi b^2/a^2 \quad (45)$$

which can also be expressed as a linear function of  $q$  with the aid of Eqs. (43) and (44). Since the relation between  $q$  and  $u_b$  is linear from Eq. (11),  $p$  and  $u_a$ , given by Eqs. (44) and (45), respectively, can be expressed as linear functions of  $u_b$ .

**NUMERICAL RESULTS.** Given any value of internal pressure, we can obtain numerical results for the stresses and strains in the radial and tangential directions and also for the displacement at any radial position in a composite tube. However, only those values at the bore, interface, and outside surface have been calculated. The organic composite material is considered to be elastic until brittle failure occurs at a maximum strain of 1.3 percent. The steel is assumed to be elastic-plastic, linear strain-hardening with  $\sigma_0 = 120$  Ksi,  $E_t = 0.04 E$ , and  $\sigma_u$  (ultimate stress) = 140 Ksi.

The relations between bore hoop strain and internal pressure are presented in Figures 2 and 3. Figure 2 shows the relations for four tubes of wall ratio 1.321 and Figure 3 for three tubes of wall ratio 1.546. The percentage of composite in each tube is defined by  $(c-b)/(c-a) \times 100$  percent. The relation between hoop strain and internal pressure is nonlinear in the elastic-plastic range and the two limits are indicated in the figures. The nonlinear range becomes smaller as the percentage of composite increases. For a given strain in the elastic range, the steel tube can resist larger pressure than the composite tube. However, for a large strain in the fully-plastic range, the composite tube can support larger pressure than the steel tube. This advantage in containing higher pressure seems very attractive for using composite tubes. It is also interesting to note that the nonlinear elastic-plastic range becomes larger as the wall ratio increases as shown in these two figures.

The numerical results for the hoop strains at the bore, interface, and outside surface of three composite tubes are shown in Figures 4, 5, and 6 as functions of internal pressure. The actual specimens were constructed [1] using steel liners with two thicknesses and the appropriate thickness of the composite circumferentially wound on the liner. The geometric dimensions ( $a, b, c$ ) for the

three composite tubes are (0.9, 1.0, 1.189), (0.9, 1.07, 1.189), (0.9, 1.07, 1.391). Figures 4, 5, and 6 show the numerical results for these tubes, respectively. The complete (including elastic, elastic-plastic, and fully-plastic) ranges of loadings up to failure pressure have been considered. Brittle failure of the composite material occurs at a maximum strain of 1.3 percent. The maximum values of internal pressure these three tubes can contain without failure are 56.9, 53.1, and 78.0 Ksi, respectively. In these figures we also show the limits of internal pressure in the elastic-plastic range, i.e., ( $p^*$ ,  $P^{**}$ ) = (20.48, 23.93), (23.06, 28.75), (27.47, 34.98 Ksi), respectively.

#### REFERENCES

1. M. D. Witherell and M. A. Scavullo, "An Investigation of Stresses and Strains in an Internally Pressurized, Composite-Jacketed, Steel Cylinder," ARDEC Technical Report ARCCB-TR-88042, Benet Laboratories, Watervliet, NY, November 1988.
2. D. R. Bland, "Elastoplastic Thick-Walled Tubes of Work-Hardening Material Subject to Internal and External Pressures and to Temperature Gradients," Journal of the Mechanics and Physics of Solids, Vol. 4, 1956, pp. 209-229.
3. P. C. T. Chen, "Stress and Deformation Analysis of Autofrettaged High Pressure Vessels," ASME Pressure Vessels and Piping, Vol. 110, July 1986, pp. 61-67.

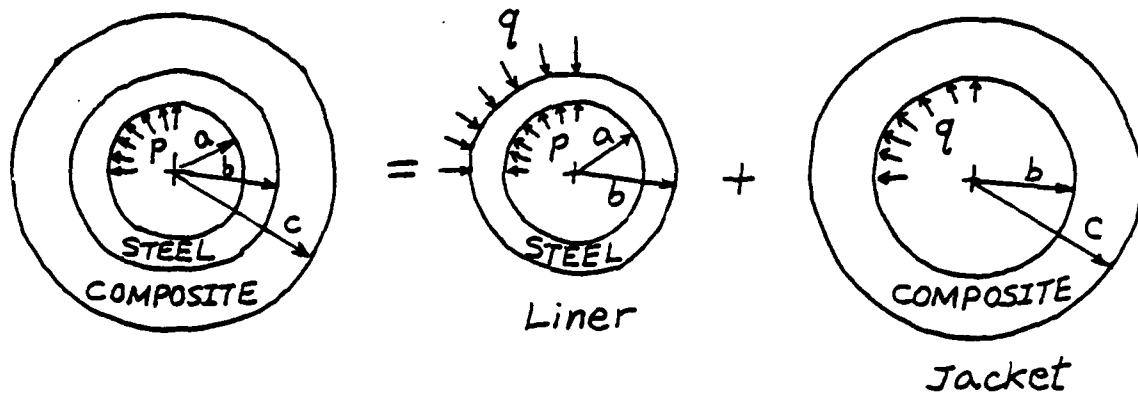


Figure 1. Schematic of a composite tube problem

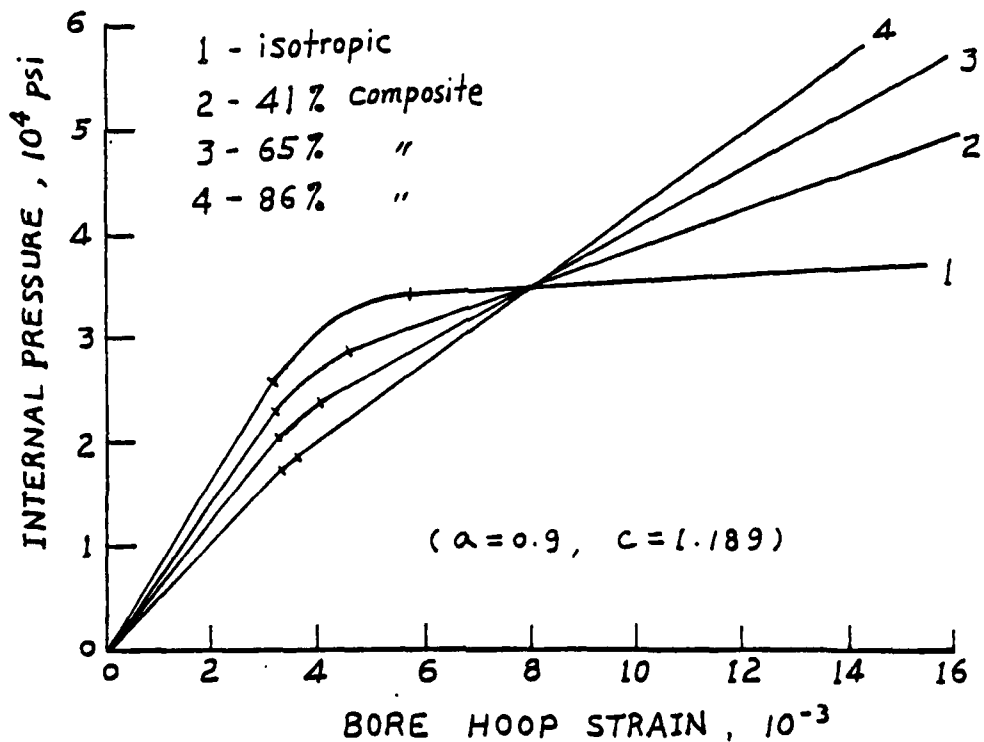


Figure 2. The relation between bore hoop strain and internal pressure for four tubes of wall ratio 1.321



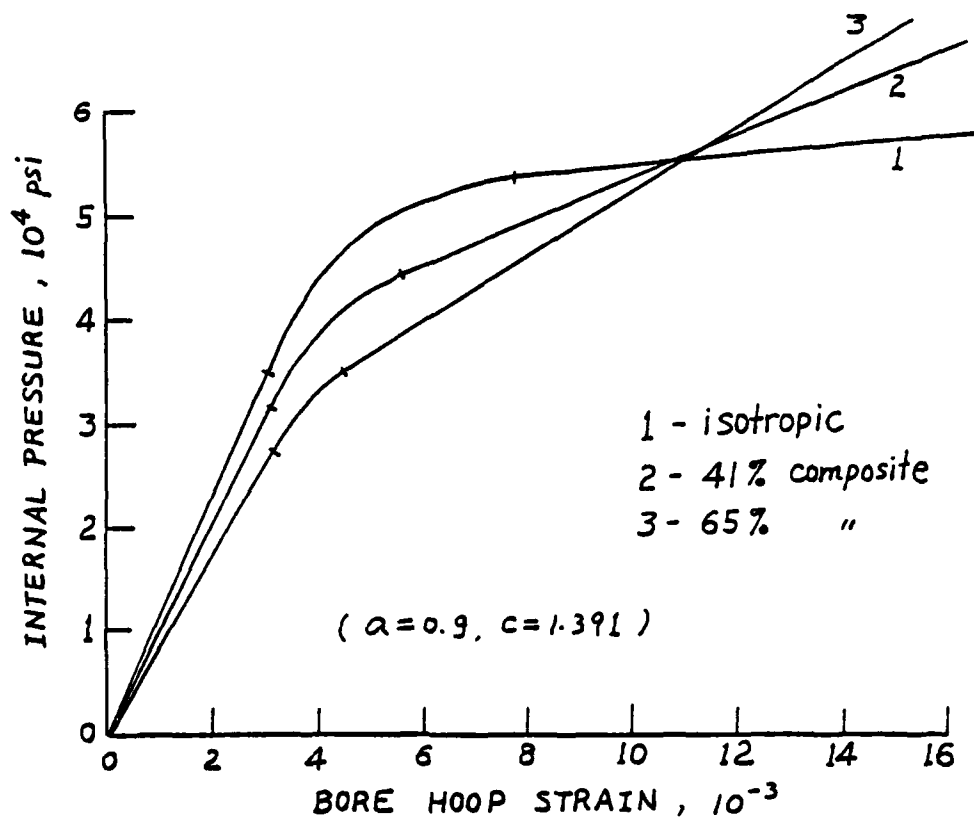


Figure 3. The relation between bore hoop strain and internal pressure for three tubes of wall ratio 1.546

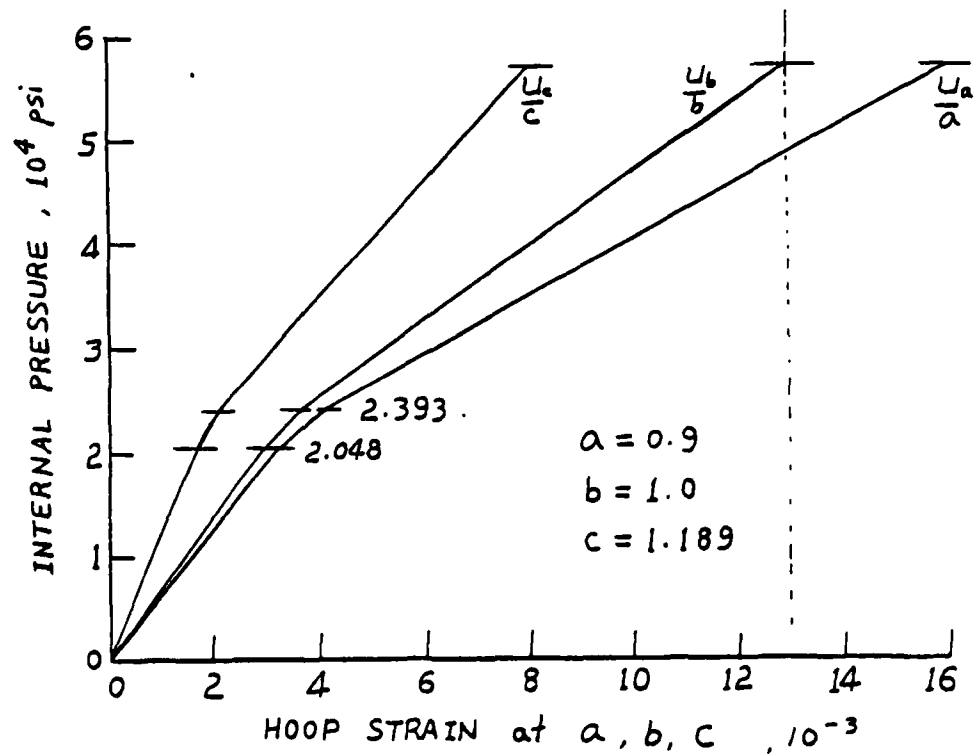


Figure 4. Hoop strains at the bore, interface, and outside surface as functions of internal pressure for a composite tube ( $\alpha = 0.9, b = 1.0, c = 1.189$ )

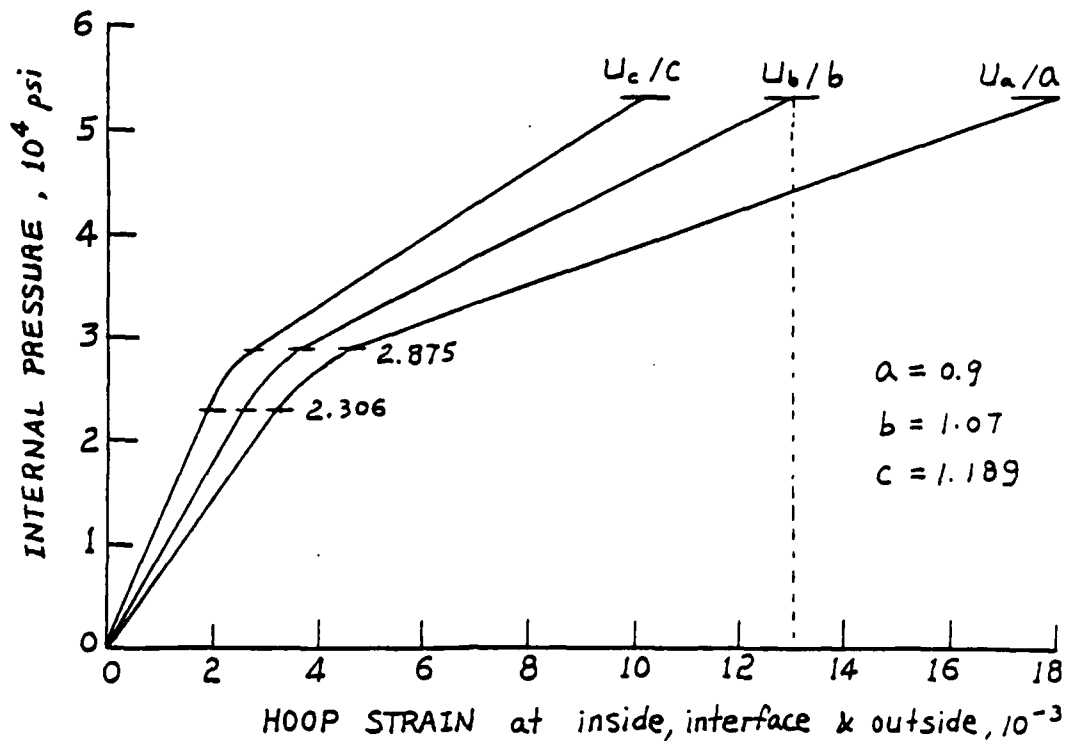


Figure 5. Hoop strains at the bore, interface, and outside surface as functions of internal pressure for a composite tube ( $a = 0.9$ ,  $b = 1.07$ ,  $c = 1.189$ )

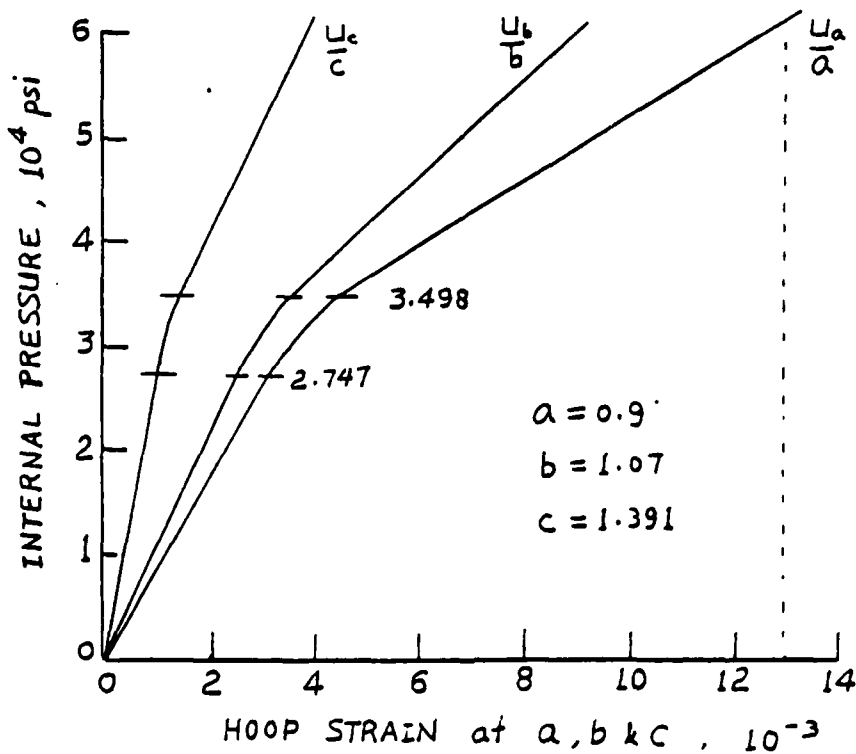


Figure 6. Hoop strains at the bore, interface, and outside surface as functions of internal pressure for a composite tube ( $a = 0.9$ ,  $b = 1.07$ ,  $c = 1.391$ )

## ULTRAFAST THERMODYNAMIC PROCESSES

Richard A. Weiss  
U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The conventional thermodynamic description of a rapid reversible process assumes that the process is adiabatic and no heat is exchanged between the thermodynamic system and the environment so that the entropy of the system remains constant. This paper suggests the possibility of processes that occur so fast that the magnitudes of both the entropy and internal energy of the system remain constant. For such a system there is an exchange of heat with the environment in the form of a change of the internal phases of the thermodynamic system. The thermodynamic equations for internal phase changing processes are developed, and a general procedure is developed for relating the temperature and density for a system undergoing an ultrafast process. The magnitude and internal phase angle of the pressure associated with an ultrafast thermodynamic process are calculated. The rapid processes that occur in supernovae may possibly be described by these calculations. Applications to the early stages of chemical reactions are suggested.

**1. INTRODUCTION.** Processes that occur very fast appear in both astrophysical and laboratory situations. For example, rapid nuclear processes occur in stars before and during supernova explosions.<sup>1-5</sup> These include electron capture by protons and the rapid capture of neutrons by atomic nuclei. In addition there are the processes associated with the core bounce and the subsequent generation of shock waves. Finally, associated with stellar core collapse is the generation of neutrinos which interact with the stellar atmosphere and often produce pressures that are sufficient to blow off the atmosphere.<sup>1-5</sup> These processes occur on very short time scales and the question of the adequacy of the adiabatic assumption of ordinary thermodynamics arises because the adiabatic process requires the internal energy to change and this may occur on a slower time scale than the short time scale of the physical process itself.

The description of the interaction of gravity waves with matter, as in the case of a laboratory gravitational wave detector, needs to account for the rapid distortion of atoms and molecules due to the rapid change of the curvature of spacetime.<sup>6,7</sup> A description of these ultrafast gravity wave interactions requires a description of a state equation for matter which includes parameters that determine the effects of gravity waves on the atomic structure of matter. Such a state equation has been developed for the real gases.<sup>8</sup> Again the question arises as to whether the adiabatic assumption is a valid description of the interaction of gravity waves with matter or whether something more sophisticated is required to describe this extremely fast process.

Rapid processes also occur in more conventional laboratory experiments. Consider the actual processes that occur during chemical reactions such as chemical bond breaking and formation which may occur on the femtosecond time scale.<sup>9-11</sup> Another example of an ultrafast process that may require reinterpretation is the

case of subpicosecond laser pulses interacting with matter.<sup>12-14</sup> A better understanding of the state equations of matter and of ultrafast thermodynamic processes are needed to describe these physical processes.

A theory of relativistic thermodynamic state equations has been developed in order to account for a difficulty with the state equation of matter at high densities, namely the fact that the state equation is not nearly as stiff as is predicted by conventional calculations.<sup>15</sup> The four dimensional Minkowski space-time of special relativity was introduced through the development of a relativistic trace equation, and specific solutions of this equation were developed for solids, quantum liquids and the real classical gases.<sup>15,16</sup> In order to have the Lie group  $e^{\pm j\phi}$  (and  $e^{\pm\phi}$ ) as the gauge group of relativistic thermodynamics, the concept of thermodynamic variables with internal phase angles was introduced.<sup>17,18</sup>

The trace equation for completely symmetric matter is given by<sup>15</sup>

$$U + T(dU/dT)_{PV} - 3V d/dV(PV)_U = U^a + T(dU^a/dT)_{p^aV} \quad (1)$$

where  $U$  and  $P$  = relativistic internal energy and pressure respectively,  $U^a$  and  $P^a$  = unrenormalized energy and pressure respectively, and  $T$  and  $V$  = temperature and volume respectively. The trace equation for matter whose thermodynamic functions have broken symmetries is given by

$$\bar{U} + T(d\bar{U}/dT)_{\bar{P}V} - 3V d/dV(\bar{P}V)_{\bar{U}} = U^a + T(dU^a/dT)_{p^aV} \quad (2)$$

where  $\bar{U}$  and  $\bar{P}$  = complex number representations of the renormalized internal energy and pressure respectively. Equation (2) can be further simplified by using the following form of the Gibbs-Helmholtz-Maxwell equation<sup>18</sup>

$$\partial\bar{U}/\partial V = T(\partial\bar{P}/\partial T)_V - \bar{P} \quad (3)$$

The complex numbers  $U$  and  $P$  that appear in equations (2) and (3) are written as<sup>18</sup>

$$\bar{U} = Ue^{j\theta_U} \quad (4)$$

$$\bar{P} = Pe^{j\theta_P} \quad (5)$$

where  $U$ ,  $P$ ,  $\theta_U$  and  $\theta_P$  are obtained from a solution of equations (2) and (3). In a similar fashion the complex number entropy is written as<sup>18</sup>

$$\bar{S} = Se^{j\theta_S} \quad (6)$$

As an illustrative example of the use of equations (1) and (2) they can be applied to real gases.

The pressure of an ordinary real gas is written as<sup>19</sup>

$$P^a = nR^a T(1 + B^a n + C^a n^2 + \dots) \quad (7)$$

where  $n$  = reciprocal volume, and  $R^a$ ,  $B^a$  and  $C^a$  = ordinary gas constant, second virial coefficient and third virial coefficient respectively. The corresponding

pressure for a symmetric relativistic real gas is written as<sup>8,15</sup>.

$$P = nRT(1 + Bn + Cn^2 + \dots) \quad (8)$$

where R, B and C = corresponding relativistic gas constant, second virial coefficient and third virial coefficient respectively which are given by<sup>8,15</sup>

$$R = R^a \quad (9)$$

$$B = B^a \quad (10)$$

$$C = C^a - 3B_a^2 \ln \psi^a \quad (11)$$

where

$$\psi^a = \frac{T}{T_R} \left| \frac{B^a(T)}{B^a(T_R)} \right|^{2/3} \quad (12)$$

where  $T_R$  = species dependent relativity temperature constant. For the case of a relativistic real gas with broken internal symmetry the pressure is written as<sup>17</sup>

$$\bar{P} = nRT(1 + Bn + \bar{C}n^2 + \dots) \quad (13)$$

where  $\bar{C}$  is obtained from a solution of equation (2) and is given by<sup>17</sup>

$$\bar{C} \sim C_a - 3B_a^2 \ln \psi^a e^{j\theta_f} \quad (14)$$

where  $\theta_f$  is given by the solution of a set of coupled differential equations.<sup>17</sup> Ultrafast thermodynamic processes, for which both the entropy and the magnitude of the internal energy are fixed, are only possible in systems like the real gases that have a parameter  $T_R$  which varies during the process. The parameter  $T_R$  depends on the species of atoms in the gas and therefore  $T_R$  changes for processes that alter the composition of the gas.

This paper considers thermodynamic processes that are sufficiently rapid to keep both the entropy and the magnitude of the internal energy constant or to keep the magnitudes of both the entropy and the internal energy fixed. Such processes change the internal phases of the entropy and internal energy, and the entropy and internal energy vectors essentially rotate (in internal space) but do not stretch. This is a special case of the general situation where the complex number thermodynamic functions rotate and stretch during thermodynamic processes.<sup>18</sup> A theory of ultrafast thermodynamic processes is developed and an expression for the pressure associated with these processes is derived.

**2. ULTRAFAST THERMODYNAMIC PROCESSES.** This section considers the thermodynamic equations that describe ultrafast processes occurring in matter that has internal phase angles associated with the thermodynamic functions. The general thermodynamic equations of matter and radiation with internal phase have already been developed in the literature.<sup>17,18</sup> The expression for the first law of thermodynamics for matter with internal phase is written as<sup>18</sup>

$$d\bar{Q} = Td\bar{S} = d\bar{U} + \bar{P}dV + \sum_{\alpha} \bar{M}_{\alpha} d\alpha \quad (15)$$

where  $\bar{Q}$  = complex number heat,  $\bar{M}_{\alpha}$  = set of generalized complex number forces and  $\alpha$  = set of generalized extensive variables.

For the special case where the change in entropy and internal energy are of the form of rotations it follows from equations (4) and (6) that<sup>18</sup>

$$d\bar{Q} = Td\bar{S} = jT\bar{S}d\theta_S \quad (16)$$

$$d\bar{U} = j\bar{U}d\theta_U \quad (17)$$

which combined with equation (15) gives the following result for an ultrafast process where S and U are both constant

$$jT\bar{S}d\theta_S = j\bar{U}d\theta_U + \bar{P}dV + \sum_{\alpha} \bar{M}_{\alpha} d\alpha \quad (18)$$

The generalized force  $\bar{M}_{\alpha}$  can be written as

$$\bar{M}_{\alpha} = M_{\alpha} e^{j\theta_{M\alpha}} \quad (19)$$

The real and imaginary parts of equation (18) can be written as

$$- TS \sin \theta_S d\theta_S = - U \sin \theta_U d\theta_U + P \cos \theta_P dV + \sum_{\alpha} M_{\alpha} \cos \theta_{M\alpha} d\alpha \quad (20)$$

$$+ TS \cos \theta_S d\theta_S = + U \cos \theta_U d\theta_U + P \sin \theta_P dV + \sum_{\alpha} M_{\alpha} \sin \theta_{M\alpha} d\alpha \quad (21)$$

For the special case of the relativistic real gas the generalized extensive variable is  $\alpha = T_R$  and the generalized force is  $\bar{M}_{\alpha} = \bar{S}_R$ , where  $\bar{S}_R$  is the complex number generalization of the scalar parameter  $S_R$  that appears in Reference 8. For the real gas equation (18) becomes

$$jT\bar{S}d\theta_S = j\bar{U}d\theta_U + \bar{P}dV + \bar{S}_R dT_R \quad (22)$$

where

$$\bar{S}_R = S_R e^{j\theta_{SR}} \quad (23)$$

is the following complex number generalization of the scalar result in Reference 8

$$\bar{S}_R = -1/2NRTn^2 (\partial\bar{C}/\partial T_R)_T \quad (24)$$

where

$$S_R = -1/2NRTn^2 [(\partial C/\partial T_R)^2 + (C\partial\theta_C/\partial T_R)^2]^{1/2} \quad (25)$$

and

$$\tan \theta_{SR} = \frac{C\partial\theta_C/\partial T_R}{\partial C/\partial T_R} \quad (26)$$

and where  $N$  = number of moles. The real and imaginary parts of equation (22) can be written as

$$- TS \sin \theta_S d\theta_S = - U \sin \theta_U d\theta_U + P \cos \theta_P dV + S_R \cos \theta_{SR} dT_R \quad (27)$$

$$+ TS \cos \theta_S d\theta_S = + U \cos \theta_U d\theta_U + P \sin \theta_P dV + S_R \sin \theta_{SR} dT_R \quad (28)$$

which are the thermodynamic equations for an ultrafast process in a real gas with both  $S$  and  $U$  held constant. If the process is truly adiabatic with  $\bar{S} =$  constant and  $dS = 0$  and  $d\theta_S = 0$  then the left hand sides of equations (27) and (28) must be set equal to zero. Note that in general  $U = U(T, V, T_R)$ ,  $\theta_U = \theta_U(T, V, T_R)$ ,  $S = S(T, V, T_R)$  and  $\theta_S = \theta_S(T, V, T_R)$ .

In order to utilize equations (20) and (21) it is necessary to evaluate the differentials  $d\theta_S$  and  $d\theta_U$  for the case of an ultrafast process with both  $U$  and  $S$  held constant. These are obtained from the following total derivatives

$$(d\theta_S/d\alpha)_{U,S} = \partial\theta_S/\partial\alpha + \partial\theta_S/\partial V(dV/d\alpha)_{U,S} + \partial\theta_S/\partial T(dT/d\alpha)_{U,S} \quad (29)$$

$$(d\theta_U/d\alpha)_{U,S} = \partial\theta_U/\partial\alpha + \partial\theta_U/\partial V(dV/d\alpha)_{U,S} + \partial\theta_U/\partial T(dT/d\alpha)_{U,S} \quad (30)$$

where  $(dV/d\alpha)_{U,S}$  and  $(dT/d\alpha)_{U,S}$  are obtained from the following two conditions which state that  $S$  and  $U$  are constant

$$\partial S/\partial\alpha + \partial S/\partial V(dV/d\alpha)_{U,S} + \partial S/\partial T(dT/d\alpha)_{U,S} = 0 \quad (31)$$

$$\partial U/\partial\alpha + \partial U/\partial V(dV/d\alpha)_{U,S} + \partial U/\partial T(dT/d\alpha)_{U,S} = 0 \quad (32)$$

In general  $S = S(\alpha, V, T)$ ,  $\theta_S = \theta_S(\alpha, V, T)$ ,  $U = U(\alpha, V, T)$  and  $\theta_U = \theta_U(\alpha, V, T)$ . From equations (31) and (32) it follows that

$$(dV/d\alpha)_{U,S} = \frac{(\partial U/\partial T)(\partial S/\partial\alpha) - (\partial S/\partial T)(\partial U/\partial\alpha)}{(\partial S/\partial T)(\partial U/\partial V) - (\partial U/\partial T)(\partial S/\partial V)} \quad (33)$$

$$(dT/d\alpha)_{U,S} = \frac{(\partial S/\partial V)(\partial U/\partial\alpha) - (\partial U/\partial V)(\partial S/\partial\alpha)}{(\partial S/\partial T)(\partial U/\partial V) - (\partial U/\partial T)(\partial S/\partial V)} \quad (34)$$

Eliminating  $d\alpha$  from equations (33) and (34) gives

$$(dT/dV)_{U,S} = \frac{(\partial S/\partial V)(\partial U/\partial\alpha) - (\partial U/\partial V)(\partial S/\partial\alpha)}{(\partial U/\partial T)(\partial S/\partial\alpha) - (\partial S/\partial T)(\partial U/\partial\alpha)} \quad (35)$$

Equation (35) relates  $T$  and  $V$  for the case of constant  $U$  and  $S$ . Only if  $\partial U/\partial\alpha \neq 0$  and  $\partial S/\partial\alpha \neq 0$  are  $T$  and  $V$  related. The derivative of the temperature with respect to the reciprocal volume at constant  $U$  and  $S$  is given by

$$n(dT/dn)_{U,S} = - V(dT/dV)_{U,S} \quad (36)$$

If  $\alpha = \alpha(V,T)$  is calculated from the condition  $U(\alpha,T,V) = \text{constant}$  then the constant  $S$  condition can be written as

$$\partial S/\partial V + \partial S/\partial T(dT/dV)_{U,S} = 0 \quad (37)$$

$$\partial S/\partial T + \partial S/\partial V(dV/dT)_{U,S} = 0 \quad (38)$$

Similarly if  $\alpha = \alpha(V,T)$  is eliminated by the condition  $S(\alpha,V,T) = \text{constant}$  then the constant  $U$  condition can be written as

$$\partial U/\partial V + \partial U/\partial T(dT/dV)_{U,S} = 0 \quad (39)$$

$$\partial U/\partial T + \partial U/\partial V(dV/dT)_{U,S} = 0 \quad (40)$$

Neglecting the  $d\alpha$  term in equations (20) and (21) gives

$$- TS \sin \theta_S d\theta_S \sim - U \sin \theta_U d\theta_U + P \cos \theta_P dV \quad (41)$$

$$+ TS \cos \theta_S d\theta_S \sim + U \cos \theta_U d\theta_U + P \sin \theta_P dV \quad (42)$$

Then

$$- P \cos \theta_P \sim TS \sin \theta_S (d\theta_S/dV)_{U,S} - U \sin \theta_U (d\theta_U/dV)_{U,S} \quad (43)$$

$$P \sin \theta_P \sim TS \cos \theta_S (d\theta_S/dV)_{U,S} - U \cos \theta_U (d\theta_U/dV)_{U,S} \quad (44)$$

Now assume that  $\theta_S \sim \theta_U$  in the trigonometric terms

$$- P \cos \theta_P \sim \sin \theta_U [TS(d\theta_S/dV)_{U,S} - U(d\theta_U/dV)_{U,S}] \quad (45)$$

$$P \sin \theta_P \sim \cos \theta_U [TS(d\theta_S/dV)_{U,S} - U(d\theta_U/dV)_{U,S}] \quad (46)$$

From equations (45) and (46) it follows that

$$\tan \theta_P \sim - \tan \theta_U \quad (47)$$

$$\theta_P \sim \theta_U + \pi/2 \quad (48)$$

$$\cos \theta_P \sim - \sin \theta_U \quad (49)$$

$$\sin \theta_P \sim \cos \theta_U \quad (50)$$

Combining equations (45) through (50) gives

$$P \sim TS(d\theta_S/dV)_{U,S} - U(d\theta_U/dV)_{U,S} \quad (51)$$



where

$$(d\theta_S/dV)_{U,S} = \partial\theta_S/\partial V + \partial\theta_S/\partial T(dT/dV)_{U,S} + \partial\theta_S/\partial\alpha(d\alpha/dV)_{U,S} \quad (52)$$

$$(d\theta_U/dV)_{U,S} = \partial\theta_U/\partial V + \partial\theta_U/\partial T(dT/dV)_{U,S} + \partial\theta_U/\partial\alpha(d\alpha/dV)_{U,S} \quad (53)$$

where  $(dT/dV)_{U,S}$  is given by equation (35) and  $(d\alpha/dV)_{U,S}$  is given by equation (33). Equations (48) and (51) give the pressure associated with an ultrafast thermodynamic process that has both U and S held constant.

The magnitude of the pressure given by equation (51) can be written in terms of the reciprocal volume  $n = 1/V$  as follows

$$\begin{aligned} P &\sim Un^2(d\theta_U/dn)_{U,S} - TSn^2(d\theta_S/dn)_{U,S} \\ &= En(d\theta_U/dn)_{U,S} - TSn(d\theta_S/dn)_{U,S} \end{aligned} \quad (54)$$

where the energy density  $E$  and the entropy density  $\$$  are given by

$$E = U/V = nU \quad (55)$$

$$\$ = S/V = nS \quad (56)$$

where U and S are constants in this paper. A further approximation for the pressure can be obtained by taking  $\theta_S \sim \theta_U$  in equations (51) and (54) with the result

$$\begin{aligned} P &\sim (TS - U)(d\theta_U/dV)_{U,S} \\ &= (E - T\$)n(d\theta_U/dn)_{U,S} \end{aligned} \quad (57)$$

Equation (57) has a proper  $T = 0$  limit.

Equation (54) is not an equation of state but rather gives the pressure for a thermodynamic process for which U and S are constants. Therefore  $P = P(T)$  or  $P = P(n)$  because V and T are related by equation (35). The total derivative  $dP/dn$  can be calculated from equation (54) as follows

$$\begin{aligned} (dP/dn)_{U,S} &= U[2n(d\theta_U/dn)_{U,S} + n^2(d^2\theta_U/dn^2)_{U,S}] \\ &\quad - TS[2n(d\theta_S/dn)_{U,S} + n^2(d^2\theta_S/dn^2)_{U,S}] \\ &\quad - S(dT/dn)_{U,S} n^2(d\theta_U/dn)_{U,S} \end{aligned} \quad (58)$$

Similarly

$$(dP/dT)_{U,S} = (dP/dn)_{U,S}(dn/dT)_{U,S} \quad (59)$$

where  $(dn/dT)_{U,S}$  is given by equations (35) and (36).

Consider now the case of an ultrafast adiabatic process in which the entropy  $\bar{S}$  and  $U$  remains fixed. For this case  $dS = 0$ ,  $d\theta_S = 0$  and  $dU = 0$ . Then equation (18) gives

$$j\bar{U}d\theta_U + \bar{P}dV + \sum_{\alpha} \bar{M}_{\alpha}d\alpha = 0 \quad (60)$$

Neglecting  $d\alpha$  gives

$$\bar{P} = -U(d\theta_U/dV)_{U,\bar{S}} e^{j(\pi/2+\theta_U)} \quad (61)$$

$$P = -U(d\theta_U/dV)_{U,\bar{S}} \quad (62)$$

$$\theta_P = \pi/2 + \theta_U \quad (63)$$

For this case  $\theta_U \neq \theta_S$  because  $\theta_S = \text{constant}$ . The derivative in equation (62) is obtained from equations (33), (35) and (53). For this case two parameters  $\alpha$  and  $\beta$  are required in equation (60) to evaluate the derivative in equation (62).

For the case of broken symmetry of space the first law of thermodynamics is written as

$$d\bar{Q} = d\bar{U} + \bar{P}|d\bar{V}| \quad (64)$$

where

$$|d\bar{V}| = \sec \beta_{V,V} dV \quad (65)$$

$$\tan \beta_{V,V} = V\partial\theta_V/\partial V \quad (66)$$

From equations (64) and (65) it follows that in order to obtain the basic equations of thermodynamics for broken symmetry space the substitution  $\bar{P} \rightarrow \bar{P} \sec \beta_{V,V}$  is made in the basic thermodynamics equations such as those given in Reference 18. For instance, the trace equation (2) becomes

$$\bar{U} + T(d\bar{U}/dT)_{\bar{P}V \sec \beta_{V,V}} - 3Vd/dV(\bar{P}V \sec \beta_{V,V})_{\bar{U}} = U^a + T(dU^a/dT)_{pav} \quad (67)$$

while equation (3) becomes

$$\partial\bar{U}/\partial V = T\partial/\partial T(\bar{P} \sec \beta_{V,V}) - \bar{P} \sec \beta_{V,V} \quad (68)$$

$$= (T\partial\bar{P}/\partial T - \bar{P})\sec \beta_{V,V} + \bar{P}T\partial/\partial T(\sec \beta_{V,V})$$

For  $\partial\theta_V/\partial T = 0$  equation (68) becomes

$$\cos \beta_{V,V} \partial\bar{U}/\partial V = T\partial\bar{P}/\partial T - \bar{P} \quad (69)$$

For  $T = 0$  equation (68) or (69) gives

$$\bar{P} = - \cos \beta_{V,V} \partial \bar{U} / \partial V \quad (70)$$

In general the broken symmetry of space lowers the calculated thermodynamic pressure. For example the broken internal symmetry of space requires equation (13) for the pressure of real gases with broken internal symmetry to be written as

$$\bar{P} = nRT \cos \beta_{V,V} (1 + Bn + \bar{C}n^2 + \dots) \quad (71)$$

For the case of an ultrafast thermodynamic process equation (64) becomes

$$jT\bar{S}d\theta_S = j\bar{U}d\theta_U + \bar{P} \sec \beta_{V,V} dV \quad (72)$$

and equations (57) and (62) become respectively

$$P \sim \cos \beta_{V,V} (E - T\bar{S})n(d\theta_U/dn)_{U,S} \quad (73)$$

$$P = - \cos \beta_{V,V} U(d\theta_U/dV)_{U,\bar{S}} \quad (74)$$

For radial symmetry  $\beta_{V,V} = \beta_{r,r}$  where  $r$  = radial coordinate. Because  $\beta_{r,r}$  is related to the internal phase angle  $\theta_r$  of the radial coordinate it follows that the laws of thermodynamics such as equations (67) through (74) depend on the internal phase structure of space. But the value of  $\theta_r$  on a macroscopic scale depends on gravity. For instance for the earth's surface  $\theta_r \sim 5.7^\circ$ . Therefore the calculations of thermodynamics must include the effects of gravity.

3. CONCLUSION. For systems, such as the real gas, with broken internal symmetries in the pressure, internal energy and entropy it is possible to have a thermodynamic process that occurs so fast as to keep the magnitudes of the entropy and internal energy constant. This is possible only for systems like the real gases which have a parameter (like the relativity temperature  $T_R$ ) that changes during the process. Such a process involves a change of structure of the molecules or atoms of the system as in the case of chemical or nuclear reactions.

#### ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

#### REFERENCES

1. Trimble, V., "Supernovae. Part I: The Events," Rev. Mod. Phys., Vol. 54, No. 4, Oct. 1982, p. 1183.
2. Fowler, W. A., "Experimental and Theoretical Astrophysics: The Quest for the Origin of the Elements," Rev. Mod. Phys., Vol. 56, No. 2, Apr. 1984, p. 149.
3. Woosley, S. E. and Phillips, M. M., "Supernova 1987A!," Science, Vol. 240, May 6, 1988, p. 750.

4. Bethe, H. A., "Nuclear Physics Needed for the Theory of Supernovae," *Ann. Rev. Nucl. Part. Sci.*, Vol. 38, 1988, p. 1.
5. Trimble, V., "1987A: The Greatest Supernova Since Kepler," *Rev. Mod. Phys.*, Vol. 60, No. 4, Oct. 1988, p. 859.
6. Thorne, K. S., "Gravitational Wave Research: Current Status and Future Prospects," *Rev. Mod. Phys.*, Vol. 52, No. 2, Part 1, April 1980.
7. Press, W. H. and Thorne, K. S., "Gravitational Wave Astronomy," *Ann. Rev. Astron. Astrophys.*, Vol. 10, 1972, p. 335.
8. Weiss, R. A., "Relativistic Wave Equations for Real Gases," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, ARO 87-1, May 27-30, 1986, p. 341.
9. Zewail, A., "Laser Femtochemistry," *Science*, Vol. 242, 23 Dec. 1988, p. 1645.
10. Peters, K. S. and Snyder, G. J., "Time-Resolved Photoacoustic Calorimetry: Probing the Energetics and Dynamics of Fast Chemical and Biochemical Reactions," *Science*, Vol. 241, 26 Aug. 1988, p. 1053.
11. Davis, W. C., "The Detonation of Explosives," *Scientific American*, Vol. 256, No. 5, May 1987, p. 106.
12. Yablonovitch, E., "Energy Conservation in the Picosecond and Subpicosecond Photoelectric Effect," *Phys. Rev. Lett.*, Vol. 60, No. 9, 29 Feb. 1988, p. 795.
13. Hicks, J. M., Urbach, L. E., Plummer, E. W. and Dai, H. L., "Can Pulsed Laser Excitation of Surfaces be Described by a Thermal Model?," *Phys. Rev. Lett.*, Vol. 61, 28 Nov. 1988, p. 2588.
14. Murnane, M. M., Kapteyn, H. C. and Falcone, R. W., "High-Density Plasmas Produced by Ultrafast Laser Pulses," *Phys. Rev. Lett.*, Vol. 62, 9 Jan. 1989, p. 155.
15. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.
16. Weiss, R. A., "Relativistic Wave Equations for Solids and Low Temperature Quantum Systems," Third Army Conference on Applied Mathematics and Computing, Georgia Institute of Technology, ARO 86-1, May 13-16, 1985, p. 717.
17. Weiss, R. A., "Relativistic Thermodynamics of Real Gases with Broken Internal Symmetry," Sixth Army Conference on Applied Mathematics and Computing, Univ. of Colorado, Boulder, ARO 89-1, 31 May-3 June, 1988, p. 203.
18. 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.
19. Hirschfelder, J. O., Curtiss, C. F. and Bird, R. B., Molecular Theory of Gases and Liquids, John Wiley, New York, 1954.

## THE INTERNAL PHASE STRUCTURE OF ATOMS

Richard A. Weiss  
U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The three dimensional Schrödinger equation for hydrogen-like atoms under pressure is solved and the spectra and eigenfunctions are calculated using the fact that in a pressure field the coordinates have internal phase angles. Because the coordinates have broken internal symmetries the energy eigenvalues are complex numbers whose real parts yield the measurable quantities that can be experimentally tested by examining the spectra of one-electron atoms under pressure. The magnetic, azimuthal and principal quantum numbers must be represented as complex numbers for hydrogen-like atoms under pressure. It is found that under pressure hydrogen-like atoms will exhibit a pressure dependent fine structure in which the energy levels of the valence electron depend on the magnetic quantum number as well as on the principal quantum number. The pressure dependence of the spectra of hydrogen-like atoms is determined. This research will have applications to stellar atmospheres and to gases at high pressures associated with conventional and nuclear explosions.

**1. INTRODUCTION.** The early development of atomic and nuclear physics made minimal use of gauge field theory because the only gauge field known was electromagnetism.<sup>1</sup> The importance of gauge fields was only fully realized in the past twenty-five years from a search for a unifying principle behind the four fundamental interactions.<sup>2</sup> Gauge theories are now important in many scientific and mathematics disciplines.<sup>3,4</sup> Recently a gauge theory of relativistic thermodynamics has been developed which suggests that the pressure in a bulk matter system has a broken internal symmetry and must be represented by a complex number.<sup>5,6</sup> A set of renormalization group equations has been developed which gives the recipe for calculating the magnitude and internal phase angle of the pressure as a function of temperature and density for an interacting bulk matter system.<sup>5,6</sup> This can be applied to gases, liquids or solids.

If the pressure has a broken internal symmetry then Euler's equations of motion suggest that space and time coordinates must also have broken internal symmetries and be treated as complex numbers.<sup>7</sup> Arguments from string theory also predict a complex number coordinate representation.<sup>8</sup> The complex number values of the space and time coordinates imply that the basic wave equations of classical and quantum physics must also contain these broken internal coordinate symmetries. For instance the Schrödinger and Dirac equations must be written as complex number coordinate equations whose eigenvalues and eigenfunctions have broken internal symmetries.<sup>9</sup> Therefore, indirectly, a gauge theory of relativistic thermodynamics predicts microscopic effects which affect the basic calculations of atomic physics and the structure of atoms. In fact, atoms located in a pressure field should exhibit an internal phase structure which depends on the magnitude of the ambient pressure.

The Bohr atom under zero external forces does not exhibit an internal phase

structure. This can be seen by writing the complex number generalization of the Bohr atom equations as follows<sup>10,11</sup>

$$\mu \bar{r}^2 \bar{\omega} = n \hbar \quad (1)$$

$$\mu \bar{\omega}^2 \bar{r} = Ze^2 / \bar{r}^2 \quad (2)$$

where the complex number radius and frequency are given by<sup>7</sup>

$$\bar{r} = r e^{j\theta_r} \quad (3)$$

$$\bar{\omega} = \omega e^{j\theta_\omega} \quad (4)$$

and where  $\mu$  = reduced mass of the electron,  $n$  = integer,  $\hbar = h/(2\pi)$ ,  $e$  = electron charge and  $Z$  = atomic number. Equations (1) and (2) can be written as

$$\mu r^2 \omega = n \hbar \quad (5)$$

$$\mu \omega^2 r = Ze^2 / r^2 \quad (6)$$

$$2\theta_r + \theta_\omega = 0 \quad (7)$$

$$3\theta_r + 2\theta_\omega = 0 \quad (8)$$

The solution of equations (5) and (6) yield the standard expressions for the radius and total energy of the Bohr hydrogen atom

$$r = n^2 \hbar^2 / (\mu Z e^2) \quad (9)$$

$$E_n = - \mu Z^2 e^4 / (2 \hbar^2 n^2) \quad (10)$$

while equations (7) and (8) give

$$\theta_r = 0 \quad (11)$$

$$\theta_\omega = 0 \quad (12)$$

Thus the internal phase structure vanishes for an isolated Bohr atom, and the Bohr radius and energy levels are real numbers. Because all of the quantities on the right hand side of the energy equation (10) are constants it is not possible for the energy levels to be pressure dependent.

Atoms subjected to pressure or other external forces have an internal phase structure. This can be seen in a simple fashion by noting that for the case of a Bohr electron under the influence of an external radial force equation (2) can be written as

$$\mu \bar{\omega}^2 \bar{r} = Ze^2 / \bar{r}^2 + \bar{F} \quad (13)$$

where  $\bar{F} = F e^{j\theta_F}$  = complex number external force acting on the electron. The external force can be transmitted to the electron by electrical forces from ad-

acent atoms and is ultimately related to the complex number pressure which can be written as

$$\bar{p} = p e^{j\theta_p} \quad (14)$$

where  $\theta_p$  = internal phase angle of the pressure. A more sophisticated approach to the problem of the Bohr atom in a pressure field is to solve Schrödinger's equation for this case. When this is done it is found that the presence of a broken symmetry pressure requires that the principal quantum number that appears in the Bohr quantization equation (1) must be a complex number so that

$$\mu \bar{r}^2 \bar{\omega} = \bar{n} \hbar \quad (15)$$

$$\bar{n} = n e^{j\theta_n} \quad (16)$$

where  $\bar{n}$  = complex number principal quantum number which is related to the broken symmetry of the azimuthal angle and ultimately to the complex number pressure (Section 5). For a hydrogen-like atom under pressure the quantization condition in equation (15) yields

$$\mu r^2 \omega = n \hbar \quad (17)$$

$$2\theta_r + \theta_\omega = \theta_n \quad (18)$$

while the force balance equation (13) gives

$$\mu \omega^2 r \cos(2\theta_\omega + \theta_r) = Ze^2/r^2 \cos(2\theta_r) + F \cos \theta_F \quad (19)$$

$$\mu \omega^2 r \sin(2\theta_\omega + \theta_r) = -Ze^2/r^2 \sin(2\theta_r) + F \sin \theta_F \quad (20)$$

Equations (17) through (20) are four scalar equations which can be solved for  $r$ ,  $\theta_r$ ,  $\omega$  and  $\theta_\omega$  in terms of  $n$ ,  $\theta_n$ ,  $F$  and  $\theta_F$  or ultimately in terms of  $P$  and  $\theta_p$ . For  $F \neq 0$  the solution of the Schrödinger equation becomes extremely difficult and therefore this is not considered in this paper. Instead it is assumed that  $F \sim 0$  so that  $\theta_r$  and  $\theta_\omega$  are obtained from equations (8) and (18) to be

$$\theta_r \sim 2\theta_n \quad \theta_\omega \sim -3\theta_n \quad (20A)$$

and the corresponding Schrödinger equation can also be solved.

This paper considers the solution of Schrödinger's equation for hydrogen-like atoms in a three dimensional space that exhibits broken internal symmetry. Section 2 gives the general form of Schrödinger's equation for a spherically symmetric potential with broken internal symmetries. Section 3 considers the azimuthal angle equation and introduces the complex magnetic quantum number. Section 4 treats the zenith angle equation which introduces the complex azimuthal quantum number, while Section 5 solves the radial equation which has a complex value of the principal quantum number. Section 6 presents the complex number wave functions for hydrogen-like atoms under pressure and Section 7 determines the complex number energy eigenvalues. Only bound states are considered in this paper.

## 2. SCHRÖDINGER'S EQUATION FOR HYDROGEN WITH BROKEN INTERNAL SYMMETRIES.

The three dimensional Schrödinger equation in spherical polar coordinates for an electron in a central force field in an atom whose space and time coordinates have broken internal symmetries is written as a generalization of the standard form of this equation as follows<sup>12-14</sup>

$$\frac{\partial^2 \bar{\psi}}{\partial \bar{r}^2} + \frac{2}{\bar{r}} \frac{\partial \bar{\psi}}{\partial \bar{r}} + \frac{1}{\bar{r}^2 \sin \bar{\psi}} \frac{\partial}{\partial \bar{\psi}} (\sin \bar{\psi} \frac{\partial \bar{\psi}}{\partial \bar{\psi}}) + \frac{1}{\bar{r}^2 \sin^2 \bar{\psi}} \frac{\partial^2 \bar{\psi}}{\partial \bar{\phi}^2} + \frac{8\pi^2 \mu}{h^2} [\bar{E} - \bar{V}(\bar{r})] \bar{\psi} = 0 \quad (21)$$

where the complex number spherical polar coordinates are written as

$$\bar{r} = r e^{j\theta_r} \quad (22)$$

$$\bar{\psi} = \psi e^{j\theta_\psi} \quad (23)$$

$$\bar{\phi} = \phi e^{j\theta_\phi} \quad (24)$$

where  $\bar{\psi}$  = complex number wave function,  $\bar{r}$  = complex number radial coordinate,  $\bar{\psi}$  = complex number zenith angle and  $\bar{\phi}$  = complex number azimuthal angle. The broken symmetry of the coordinates are described by the internal phase angles  $\theta_r$ ,  $\theta_\psi$  and  $\theta_\phi$  which are pressure dependent.<sup>7</sup> The measured values of the coordinates are given by the real parts of the complex number coordinates given in equations (22) through (24) and are written as<sup>7</sup>

$$r_m = r \cos \theta_r \quad (25)$$

$$\psi_m = \psi \cos \theta_\psi \quad (26)$$

$$\phi_m = \phi \cos \theta_\phi \quad (27)$$

The complex number potential is written as

$$\bar{V} = V e^{j\theta_V} \quad (28)$$

which for the Coulomb potential becomes

$$\bar{V} = -Ze^2/\bar{r} \quad (29)$$

$$V = -Ze^2/r \quad (30)$$

$$\theta_V = -\theta_r \quad (31)$$

If the potential were directly measurable the measured value would be given by<sup>7</sup>

$$V_m = V \cos \theta_V \quad (32)$$

$$= -Ze^2/r \cos \theta_r$$

Note the effect of the external pressure is assumed only to make the coordinates complex numbers and not to change the basic form of the Coulomb potential. The complex number total energy is written as



$$\bar{E} = E e^{j\theta_E} \quad (33)$$

while its measured value is given by

$$E_m = E \cos \theta_E \quad (34)$$

Both  $\bar{E}$  and  $E_m$  are determined in Section 7. The complex number wave function is written as

$$\bar{\Psi} = \Psi e^{j\theta_\Psi} \quad (35)$$

and will be determined in Section 5. The measured wave function is given by

$$\Psi_m = \Psi \cos \theta_\Psi \quad (36)$$

The complex number Schrödinger equation (21) can be separated into three component equations by following the standard recipe of writing the wave function as a product of three independent functions as follows<sup>12-14</sup>

$$\bar{\Psi} = \bar{R}(\bar{r})\bar{W}(\bar{\psi})\bar{\Phi}(\bar{\phi}) \quad (37)$$

where

$$\bar{R} = R e^{j\theta_R} = \text{radial wave function} \quad (38)$$

$$\bar{W} = W e^{j\theta_W} = \text{zenith angle wave function} \quad (39)$$

$$\bar{\Phi} = \Phi e^{j\theta_\Phi} = \text{azimuthal angle wave function} \quad (40)$$

Combining equations (37) through (40) gives

$$\Psi = RW\Phi \quad (41)$$

$$\theta_\Psi = \theta_R + \theta_W + \theta_\Phi \quad (42)$$

Placing equation (37) into equation (21) yields the following generalizations of the standard azimuthal angle equation, zenith angle equation and radial equation respectively<sup>12-14</sup>

$$d^2\bar{\Phi}/d\bar{\phi}^2 + \bar{M}^2\bar{\Phi} = 0 \quad (43)$$

$$1/\sin \bar{\psi} \, d/d\bar{\psi}(\sin \bar{\psi} \, d\bar{W}/d\bar{\psi}) + (\bar{\beta} - \bar{M}^2/\sin^2 \bar{\psi})\bar{W} = 0 \quad (44)$$

$$\bar{r}^2 \, d^2\bar{R}/d\bar{r}^2 + 2\bar{r} \, d\bar{R}/d\bar{r} + (\bar{k}^2\bar{r}^2 - \bar{\beta})\bar{R} = 0 \quad (45)$$

where

$$\bar{k}^2 = 2\nu/\hbar^2(\bar{E} - \bar{V}) \quad (46)$$

$$\bar{M} = M e^{j\theta_M} \quad (47)$$

$$\bar{\beta} = \beta e^{j\theta_\beta} \quad (48)$$

where  $\bar{M}$  = complex magnetic quantum number which will be determined in Section 3, and  $\bar{\beta}$  = complex number constant that will be determined in Section 4. The solution of the complex number Schrödinger equation for a hydrogen-like atom requires first the solution of equations (43) through (45) and the determination of the six functions  $R(r)$ ,  $\theta_R(r)$ ,  $W(\psi)$ ,  $\theta_W(\psi)$ ,  $\phi(\phi)$  and  $\theta_\phi(\phi)$ , and secondly the determination of the values of  $E$  and  $\theta_E$  which give the complex number energy eigenvalues.

3. THE AZIMUTHAL ANGLE EQUATION FOR A HYDROGEN-LIKE ATOM WITH BROKEN INTERNAL SYMMETRY. This section determines the solution of the complex number azimuthal equation and gives the magnitude and internal phase angle of the complex magnetic quantum number  $\bar{M}$ . The formal solution of equation (43) is written as<sup>12-14</sup>

$$\bar{\phi} = \bar{A}e^{i\bar{M}\phi} + \bar{B}e^{-i\bar{M}\phi} \quad (49)$$

It will now be shown that  $\bar{M}\bar{\phi}$  must be a real number if  $\bar{\phi}$  is to be symmetrical (unchanged) under a  $2\pi$  change of the value of  $\phi_m$ . In other words, because  $\phi_m$  given by equation (27) is the measured azimuthal angle, the wave function in equation (49) must be unchanged under  $\phi_m \rightarrow \phi_m + 2\pi$ , and this implies the reality of  $\bar{M}\bar{\phi}$ . The reality condition for  $\bar{M}\bar{\phi}$  can be written as

$$\overline{\bar{M}\bar{\phi}} = \bar{M}\bar{\phi} \quad (50)$$

$$\theta_{\bar{M}} + \theta_{\bar{\phi}} = 0 \quad (51)$$

where equations (24) and (47) were used. In order to verify this conclusion the complex numbers  $\bar{M}$  and  $\bar{\phi}$  are written in terms of their real and imaginary parts as follows

$$\bar{M} = M_R + jM_I = M(\cos \theta_M + j \sin \theta_M) \quad (52)$$

$$\bar{\phi} = \phi_R + j\phi_I = \phi(\cos \theta_\phi + j \sin \theta_\phi) \quad (53)$$

Using equations (52) and (53) allow  $\bar{M}\bar{\phi}$  to be written as

$$\bar{M}\bar{\phi} = M_R\phi_R - M_I\phi_I + j(M_I\phi_R + M_R\phi_I) \quad (54)$$

If the imaginary part of equation (54) is zero, then

$$\phi_I = -M_I\phi_R/M_R \quad (55)$$

and substituting equation (55) into equation (54) gives

$$\bar{M}\bar{\phi} = (M_R^2 + M_I^2)\phi_R/M_R = M^2\phi_R/M_R = M^2\phi_m/M_R \quad (56)$$

which shows that if  $\bar{M}\bar{\phi}$  is real it is also linear in  $\phi_m$  the measured value of  $\phi$  given by equation (27).

The linearity in  $\phi_m$  shown by equation (56) allows the possibility of having the azimuthal wave function given in equation (49) unchanged under  $\phi_m \rightarrow \phi_m + 2\pi$  by requiring

$$M^2/M_R = m \quad (57)$$

where  $m$  = standard magnetic quantum number which is a positive or negative integer. Combining equations (51), (52) and (57) gives

$$\begin{aligned} M &= m \cos \theta_M \\ &= m \cos \theta_\phi \end{aligned} \quad (58)$$

as the condition for symmetry under  $\phi_m \rightarrow \phi_m + 2\pi$ . The exponent terms in equation (49) can be written as

$$\bar{M}\phi = M\phi = m\phi \cos \theta_\phi = m\phi_R = m\phi_m \quad (59)$$

and therefore

$$\bar{\phi} = \bar{A}e^{1|m|\phi_m} + \bar{B}e^{-1|m|\phi_m} \quad (60)$$

The magnitude and internal phase angle of the complex magnetic quantum number  $\bar{M}$  are given by equations (58) and (51) respectively. The real and imaginary parts of  $\bar{M}$  are given by

$$M_R = m \cos^2 \theta_\phi \quad (61)$$

$$M_I = -m \sin \theta_\phi \cos \theta_\phi \quad (62)$$

$$\bar{M} = m \cos \theta_\phi e^{-j\theta_\phi} \quad (63)$$

The interesting thing about equation (58) is that  $M$  is not an integer but reduces to an integer for the symmetrical case of  $\theta_\phi = 0$ .

4. THE ZENITH ANGLE EQUATION FOR A HYDROGEN-LIKE ATOM WITH BROKEN INTERNAL SYMMETRY. This section solves the complex number zenith angle equation (44) and determines the complex number parameter  $\bar{\beta}$  that appears in this equation. This equation will be solved by a simple generalization of the standard technique used to solve the corresponding scalar form of this equation.<sup>12-14</sup> Define the complex number  $\bar{\xi}$  by

$$\bar{\xi} = \cos \bar{\psi} = \xi e^{j\theta_\xi} = C_\psi e^{-j\theta_{c\psi}} \quad (64)$$

so that<sup>7</sup>

$$\xi = C_\psi \quad (65)$$

$$\theta_\xi = -\theta_{c\psi} \quad (66)$$

$$C_\psi = [\cos^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)]^{1/2} \quad (67)$$

$$\tan \theta_{c\psi} = \tan(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \quad (68)$$

Then equation (44) can be written as

$$d/d\bar{\xi}[(1 - \bar{\xi}^2)d\bar{W}/d\bar{\xi}] + [\bar{\beta} - \bar{M}^2/(1 - \bar{\xi}^2)]\bar{W} = 0 \quad (69)$$

where  $\bar{M}$  is given by equation (63).

The solution of equation (69) follows from the standard procedure for the solution of the corresponding scalar equation by writing<sup>12-14</sup>

$$\bar{W} = (1 - \bar{\xi}^2)^{\bar{M}'/2} \bar{G}(\bar{\xi}) \quad (70)$$

where

$$\bar{M}' = |m| \cos \theta_\phi e^{-j\theta_\phi} \quad M' = |m| \cos \theta_\phi \quad (70A)$$

$$M'_R = |m| \cos^2 \theta_\phi \quad M'_I = -|m| \sin \theta_\phi \cos \theta_\phi \quad (70B)$$

which gives<sup>12-14</sup>

$$(1 - \bar{\xi}^2) d^2 \bar{G} / d\bar{\xi}^2 - 2(\bar{M}' + 1) \bar{\xi} d\bar{G} / d\bar{\xi} + [\bar{\beta} - \bar{M}'(\bar{M}' + 1)] \bar{G} = 0 \quad (71)$$

The solution of equation (71) is obtained by a power series expansion and gives the following generalization of the standard results for real numbers<sup>12-14</sup>

$$\bar{a}_{\sigma+2} / \bar{a}_\sigma = [(\sigma + \bar{M}')(\sigma + \bar{M}' + 1) - \bar{\beta}] / [(\sigma + 1)(\sigma + 2)] \quad (72)$$

where  $\bar{a}_\sigma$  and  $\bar{a}_{\sigma+2}$  = coefficients in a power series expansion of  $\bar{G}$ , and  $\sigma$  = integer. Equation (72) shows that the only way the power series breaks off at term  $\bar{\xi}^\nu$  is to have

$$\bar{\beta} = \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1) \quad (73)$$

where

$$\bar{\mathcal{L}} = \bar{M}' + \nu \quad (74)$$

where  $\nu$  = integer. Equations (73) and (74) are recognized as complex number generalizations of the standard scalar results.<sup>12-14</sup> The important thing about equations (72) through (74) is that  $\bar{\beta}$ ,  $\bar{M}'$  and  $\bar{\mathcal{L}}$  are complex numbers and not integers as in the standard case. The only integer requirement is that  $\bar{\mathcal{L}}$  and  $\bar{M}'$  differ by an integer as shown in equation (74).  $\bar{\mathcal{L}}$  is a complex number generalization of the standard integer azimuthal quantum number  $\ell$ .

The solution to equation (69) can then be written using equations (70), (72) and (73)

$$\begin{aligned} \bar{W} = & (1 - \bar{\xi}^2)^{\bar{M}'/2} c_0 \{ 1 + 1/2[\bar{M}'(\bar{M}' + 1) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)]\bar{\xi}^2 \\ & + 1/24[\bar{M}'(\bar{M}' + 1) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)][(\bar{M}' + 2)(\bar{M}' + 3) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)]\bar{\xi}^4 + \dots \} \\ & + (1 - \bar{\xi}^2)^{\bar{M}'/2} c_1 \{ \bar{\xi} + 1/2[(\bar{M}' + 1)(\bar{M}' + 2) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)]\bar{\xi}^3 \\ & + 1/120[(\bar{M}' + 1)(\bar{M}' + 2) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)][(\bar{M}' + 3)(\bar{M}' + 4) - \bar{\mathcal{L}}(\bar{\mathcal{L}} + 1)]\bar{\xi}^5 + \dots \} \end{aligned} \quad (75)$$

which clearly breaks off when equation (74) is satisfied for a series of integers  $\nu = 0, 1, 2, \dots$ . The solutions given in equation (75) are simple generalizations of the standard associated Legendre functions, so that formally

$$\bar{w} = \bar{P}_{\bar{L}}^{\bar{M}'}(\bar{\xi}) \quad (76)$$

Consider now the specific solutions corresponding to  $\nu = \ell - |m| = 0, 1, 2, \dots$ .

For  $\nu = 0$ ,  $\ell = |m|$ ,  $\bar{L} = \bar{M}'$

$$\bar{P}_{\bar{M}'}^{\bar{M}'}(\bar{\xi}) = (1 - \bar{\xi}^2)^{\bar{M}'/2} = \sin^{\bar{M}'} \bar{\psi} \quad (77)$$

For  $\nu = 1$ ,  $\ell = |m| + 1$ ,  $\bar{L} = \bar{M}' + 1$

$$\bar{P}_{\bar{M}'+1}^{\bar{M}'}(\bar{\xi}) = (1 - \bar{\xi}^2)^{\bar{M}'/2} \bar{\xi} = \sin^{\bar{M}'} \bar{\psi} \cos \bar{\psi} \quad (78)$$

For  $\nu = 2$ ,  $\ell = |m| + 2$ ,  $\bar{L} = \bar{M}' + 2$

$$\begin{aligned} \bar{P}_{\bar{M}'+2}^{\bar{M}'}(\bar{\xi}) &= (1 - \bar{\xi}^2)^{\bar{M}'/2} [(2\bar{M}' + 3)\bar{\xi}^2 - 1] \\ &= \sin^{\bar{M}'} \bar{\psi} [(2\bar{M}' + 3)\cos^2 \bar{\psi} - 1] \end{aligned} \quad (79)$$

For  $\nu = 3$ ,  $\ell = |m| + 3$ ,  $\bar{L} = \bar{M}' + 3$

$$\begin{aligned} \bar{P}_{\bar{M}'+3}^{\bar{M}'}(\bar{\xi}) &= (1 - \bar{\xi}^2)^{\bar{M}'/2} [(2\bar{M}' + 5)\bar{\xi}^3 - 3\bar{\xi}] \\ &= \sin^{\bar{M}'} \bar{\psi} [(2\bar{M}' + 5)\cos^3 \bar{\psi} - 3 \cos \bar{\psi}] \end{aligned} \quad (80)$$

The value of  $\bar{M}'$  that appears in equations (75) through (80) is given by equation (70A). In this way the solutions of the azimuthal equation for a hydrogen-like atom with broken internal symmetry are obtained. The solution given in equations (75) through (80) are given in Section 6 for specific atomic shells.

It is clear that the integer  $\nu$  in equation (74) must be given by  $\nu = \ell - |m|$  because equation (74) is also valid for the case of zero internal phase, and therefore

$$\begin{aligned} \bar{L} &= \bar{M}' + \ell - |m| \\ &= |m| \cos \theta_{\phi} e^{-j\theta_{\phi}} + \ell - |m| \\ &= \ell - |m| \sin^2 \theta_{\phi} - j|m| \sin \theta_{\phi} \cos \theta_{\phi} \end{aligned} \quad (81)$$

where  $\bar{M}'$  is given by equation (70A).

The complex azimuthal angular momentum quantum number  $\bar{L}$  can be written as

$$\bar{L} = \bar{L} e^{j\theta_{\phi}} \quad (82)$$

Combining equations (74), (81) and (82) gives

$$\bar{L} \cos \theta_{\bar{L}} = \bar{M}' \cos \theta_{\bar{M}'} + \ell - |m| \quad (83)$$

$$\bar{L} \sin \theta_{\bar{L}} = \bar{M}' \sin \theta_{\bar{M}'} \quad (84)$$

where  $M'$  and  $\theta_{M'}$  are given by equations (70A) and (51) respectively. Combining equations (51), (58), (83) and (84) gives

$$\begin{aligned}\tan \theta_{\mathcal{L}} &= (M' \sin \theta_{M'}) / (M' \cos \theta_{M'} + \ell - |m|) \\ &= - (|m| \sin \theta_{\phi} \cos \theta_{\phi}) / (\ell - |m| \sin^2 \theta_{\phi})\end{aligned}\quad (85)$$

$$\theta_{\mathcal{L}} \sim - |m| \theta_{\phi} / \ell \quad (85A)$$

$$\begin{aligned}\mathcal{L}^2 &= \ell^2 [1 - |m|/\ell (2 - |m|/\ell) \sin^2 \theta_{\phi}] \\ &= \ell^2 - |m| (2\ell - |m|) \sin^2 \theta_{\phi}\end{aligned}\quad (86)$$

Note that if  $\theta_{\phi} = 0$  then  $\theta_{\mathcal{L}} = 0$  and  $\mathcal{L} = \ell$ . Also if  $m = 0$  then  $\theta_{\mathcal{L}} = 0$  and  $\mathcal{L} = \ell$ . From equation (86) it follows that in general  $\mathcal{L} \leq \ell$ . The interesting point is that equation (86) shows that  $\mathcal{L}$  is not an integer and depends on the values of  $|m|$  as follows:

$$v = 0, \ell = |m|$$

$$\mathcal{L} = \ell \cos \theta_{\phi} = |m| \cos \theta_{\phi} \quad (87)$$

$$v = 1, \ell = |m| + 1$$

$$\begin{aligned}\mathcal{L}^2 &= (|m| + 1)^2 [1 - |m| (|m| + 2) / (|m| + 1)^2 \sin^2 \theta_{\phi}] \\ &= 1 + |m| (|m| + 2) \cos^2 \theta_{\phi} \\ &= \ell^2 - (\ell^2 - 1) \sin^2 \theta_{\phi}\end{aligned}\quad (88)$$

$$v = 2, \ell = |m| + 2$$

$$\begin{aligned}\mathcal{L}^2 &= (|m| + 2)^2 [1 - |m| (|m| + 4) / (|m| + 2)^2 \sin^2 \theta_{\phi}] \\ &= 4 + |m| (|m| + 4) \cos^2 \theta_{\phi} \\ &= \ell^2 - (\ell^2 - 4) \sin^2 \theta_{\phi}\end{aligned}\quad (89)$$

$$v = 3, \ell = |m| + 3$$

$$\begin{aligned}\mathcal{L}^2 &= (|m| + 3)^2 [1 - |m| (|m| + 6) / (|m| + 3)^2 \sin^2 \theta_{\phi}] \\ &= 9 + |m| (|m| + 6) \cos^2 \theta_{\phi} \\ &= \ell^2 - (\ell^2 - 9) \sin^2 \theta_{\phi}\end{aligned}\quad (90)$$

5. THE RADIAL EQUATION FOR HYDROGEN-LIKE ATOMS WITH BROKEN INTERNAL SYMMETRIES. In this section the complex number radial equation (45) is solved and the complex principal quantum number is introduced. Combining equations (45)

and (73) gives

$$\bar{r}^2 d^2 \bar{R} / d\bar{r}^2 + 2\bar{r} d\bar{R} / d\bar{r} + [\bar{k}^2 \bar{r}^2 - \bar{l}(\bar{l} + 1)] \bar{R} = 0 \quad (91)$$

where  $\bar{l}$  is given by equations (74) or (81) and  $\bar{k}$  is given by

$$\bar{k}^2 = 2\mu/\hbar^2 (\bar{E} + Ze^2/\bar{r}) \quad (92)$$

The solution to equation (91) can be found by a generalization of the standard method developed for the real number form of the radial equation.<sup>12-14</sup> A change of dependent and independent variables is made by writing

$$\bar{\rho} = 2\bar{a}_r = \rho e^{j\theta_\rho} \quad (93)$$

$$\bar{R} = \bar{\rho}^{\bar{l}} e^{-\bar{\rho}/2} \bar{L}(\bar{\rho}) = R e^{j\theta_R} \quad (94)$$

where

$$\bar{\alpha}^2 = -2\mu\bar{E}/\hbar^2 = \alpha^2 e^{2j\theta_\alpha} \quad (95)$$

Substituting equations (93) through (95) into equation (91) gives

$$\bar{\rho} d^2 \bar{L} / d\bar{\rho}^2 + [2(\bar{l} + 1) - \bar{\rho}] d\bar{L} / d\bar{\rho} + (\bar{n} - \bar{l} - 1) \bar{L} = 0 \quad (96)$$

where

$$\bar{n} = n e^{j\theta_n} = \mu Z e^2 / (\bar{\alpha} \hbar^2) = Z / (a_0 \bar{\alpha}) \quad (97)$$

$$a_0 = \hbar^2 / (\mu e^2)$$

where  $\bar{n}$  = complex principal quantum number and  $a_0$  = Bohr radius. From equation (95) it follows that

$$\alpha^2 = -2\mu E / \hbar^2 \quad (99)$$

$$2\theta_\alpha = \theta_E \quad (100)$$

while from equation (97) it follows that

$$n = \mu Z e^2 / (\alpha \hbar^2) = Z / (a_0 \alpha) \quad (101)$$

$$\theta_n = -\theta_\alpha = -\theta_E / 2 \quad (102)$$

Finally from equations (93) and (97) it follows that

$$\bar{\rho} = 2Z\bar{r} / (a_0 \bar{n}) \quad (103)$$

$$\rho = 2Zr / (a_0 n) \quad (104)$$

$$\theta_\rho = \theta_r + \theta_\alpha = \theta_r - \theta_n = \theta_r + \theta_E / 2 \quad (105)$$

Thus  $\theta_\rho$  introduces the radial coordinate internal phase angle  $\theta_r$ . The values of  $n$  and  $\theta_n$  will be determined later in this section.

The solution of equation (96) can be obtained by a power series whose terms have the form  $a_{\nu'} \rho^{\nu'}$  where  $a_{\nu'}$  are determined by the following generalization of the standard recursion formula<sup>12</sup>

$$(\bar{n} - \bar{l} - 1 - \nu')\bar{a}_{\nu'} + [2(\nu' + 1)(\bar{l} + 1) + \nu'(\nu' + 1)]\bar{a}_{\nu'+1} = 0 \quad (106)$$

where  $\nu' = \text{integer}$ . The condition that  $a_{N\rho^N}$  be the last non-zero term is that  $\bar{a}_{N+1} = 0$  so that the break off condition is

$$\bar{n} = \bar{l} + 1 + N \quad (107)$$

where  $N = \text{integer}$ . Equation (107) is the complex number generalization of the standard scalar result.<sup>12</sup> For a hydrogen-like atom with broken internal symmetries the principal quantum number is a complex number related to  $\bar{l}$  by equation (107). Equations (106) and (107) show that a break off solution to equation (96) is possible for complex principal and azimuthal quantum numbers provided that  $\bar{n} - \bar{l} = \text{integer}$ . Because equation (107) is also valid for zero internal phase angles it follows that

$$n = \ell + 1 + N \quad (108)$$

where  $n$  and  $\ell = \text{standard integer principal and azimuthal quantum numbers respectively}$ .

Combining equations (107) and (108) gives

$$\bar{n} = \bar{l} + n - \ell \quad (109)$$

where  $\bar{l}$  is given by equations (74) or (81). The real and imaginary parts of equation (109) give

$$n \cos \theta_n = \ell \cos \theta_\ell + n - \ell \quad (110)$$

$$n \sin \theta_n = \ell \sin \theta_\ell \quad (111)$$

Combining equations (110) and (111) gives

$$\tan \theta_n = (\ell \sin \theta_\ell) / (\ell \cos \theta_\ell + n - \ell) \quad (112)$$

$$n^2 = \ell^2 + 2\ell(n - \ell)\cos \theta_\ell + (n - \ell)^2 \quad (113)$$

where  $\theta_\ell$  and  $\ell$  are given by equations (85) and (86) respectively. Alternatively, equation (109) can be rewritten using equation (81) with the result

$$\bar{n} = \bar{M}' + n - |m| \quad (114)$$

$$= |m| \cos \theta_\phi e^{-j\theta_\phi} + n - |m|$$



where  $\bar{M}'$  is given by equation (70A). Taking the real and imaginary parts of equation (114) gives

$$\begin{aligned} \eta \cos \theta_\eta &= M' \cos \theta_{M'} + n - |m| & (115) \\ &= |m| \cos^2 \theta_\phi + n - |m| \\ &= n - |m| \sin^2 \theta_\phi \end{aligned}$$

$$\begin{aligned} \eta \sin \theta_\eta &= M' \sin \theta_{M'} & (116) \\ &= - |m| \sin \theta_\phi \cos \theta_\phi \end{aligned}$$

Combining equations (115) and (116) gives

$$\tan \theta_\eta = - (|m| \sin \theta_\phi \cos \theta_\phi) / (n - |m| \sin^2 \theta_\phi) \quad (117)$$

$$\begin{aligned} \eta^2 &= n^2 [1 - |m|/n(2 - |m|/n) \sin^2 \theta_\phi] & (118) \\ &= n^2 - |m|(2n - |m|) \sin^2 \theta_\phi \end{aligned}$$

Equations (117) and (118) give the internal phase angle and magnitude respectively of the complex principal quantum number for a hydrogen-like atom with broken internal symmetries. The values of  $\theta_\eta$  and  $\eta$  depend on both  $n$  and  $|m|$ . For  $m = 0$  it follows that  $\theta_\eta = 0$  and  $\eta = n$ . For the symmetric case with  $\theta_\phi = 0$  it follows that  $\theta_\eta = 0$  and  $\eta = n$ . From equation (118) it follows that for small  $\theta_\phi$

$$\eta \sim n [1 - 1/2 |m|/n(2 - |m|/n) \sin^2 \theta_\phi] \quad (119)$$

Equation (118) shows that  $\eta \leq n$ . Also, from equation (117) it follows that for small  $\theta_\phi$

$$\theta_\eta \sim - |m|/n \theta_\phi \quad (120)$$

while equations (102) and (105) show that

$$\theta_E \sim 2|m|/n \theta_\phi \quad (121)$$

$$\theta_\rho \sim \theta_r + |m|/n \theta_\phi \quad (122)$$

Equation (96) is a complex number generalization of the standard differential equation satisfied by associated Laguerre polynomials.<sup>12</sup> In fact the solution of equation (96) is a complex number associated Laguerre polynomial of degree  $\bar{\eta} - \bar{L} - 1 = N$  and order  $2\bar{L} + 1$ . Thus the degree is a real number (integer) but the order is a complex number. The solution of equation (96) can be written formally as<sup>12</sup>

$$\bar{L} = \bar{L} \frac{2\bar{L}+1}{\bar{\eta}+\bar{L}} (\bar{\rho}) \quad (123)$$

The argument  $\bar{\rho}$  is a complex number given by equation (93). The series solution

to equation (96) is obtained from the recursion relations in equation (106), so that

$$\begin{aligned} L_{\bar{n}+\bar{\ell}}^{-2\bar{\ell}+1}(\bar{\rho}) &= 1 - \frac{N}{2(\bar{\ell}+1)}\bar{\rho} + \frac{N(N-1)}{2(\bar{\ell}+1)[4(\bar{\ell}+1)+2]}\bar{\rho}^2 \\ &- \frac{N(N-1)(N-2)}{2(\bar{\ell}+1)[4(\bar{\ell}+1)+2][6(\bar{\ell}+1)+6]}\bar{\rho}^3 + \dots \end{aligned} \quad (124)$$

out to  $\bar{\rho}^N$ . Because the order of the polynomial  $2\bar{\ell}+1$  is a complex number it does not make sense to derive the complex number associated Laguerre polynomials in terms of derivatives of the Laguerre polynomials.<sup>12</sup> For the case of complex number order, equation (96) must be considered as the fundamental defining relation, and equation (124) is the basic solution. Also, the complex number associated Laguerre polynomials cannot be derived from a generating function as this requires the order to be an integer.<sup>12</sup> Finally, the complex number associated Laguerre polynomials can always be written as

$$L_{\bar{n}+\bar{\ell}}^{-2\bar{\ell}+1}(\bar{\rho}) = L_{\bar{n}+\bar{\ell}}^{2\bar{\ell}+1}(\rho, \theta_\rho, \ell, \theta_\ell) e^{j\theta_L} \quad (125)$$

where  $\theta_L$  = internal phase angle of the complex associated Laguerre polynomial.

The first few complex number associated Laguerre polynomials are obtained from equation (124) to be:

$$N = 0, \quad \bar{n} = \bar{\ell} + 1, \quad n = \ell + 1$$

$$L_{2\bar{\ell}+1}^{-2\bar{\ell}+1} = 1 \quad (126)$$

$$N = 1, \quad \bar{n} = \bar{\ell} + 2, \quad n = \ell + 2$$

$$L_{2\bar{\ell}+2}^{-2\bar{\ell}+1} = 1 - \frac{\bar{\rho}}{2(\bar{\ell}+1)} = \frac{2(\bar{\ell}+1) - \bar{\rho}}{2(\bar{\ell}+1)} \quad (127)$$

$$N = 2, \quad \bar{n} = \bar{\ell} + 3, \quad n = \ell + 3$$

$$\begin{aligned} L_{2\bar{\ell}+3}^{-2\bar{\ell}+1} &= 1 - \frac{2\bar{\rho}}{2(\bar{\ell}+1)} + \frac{2\bar{\rho}^2}{2(\bar{\ell}+1)[4(\bar{\ell}+1)+2]} \\ &= \frac{2(\bar{\ell}+1)(2\bar{\ell}+3) - 2(2\bar{\ell}+3)\bar{\rho} + \bar{\rho}^2}{2(\bar{\ell}+1)(2\bar{\ell}+3)} \end{aligned} \quad (128)$$

$$N = 3, \bar{n} = \bar{l} + 4, n = l + 4$$

$$\begin{aligned} \frac{\bar{l}^{-2\bar{l}+1}}{2\bar{l}+4} &= 1 - \frac{3\bar{\rho}}{2(\bar{l}+1)} + \frac{6\bar{\rho}^{-2}}{2(\bar{l}+1)[4(\bar{l}+1)+2]} \\ &\quad - \frac{6\bar{\rho}^{-3}}{2(\bar{l}+1)[4(\bar{l}+1)+2][6(\bar{l}+1)+6]} \\ &= \frac{4(\bar{l}+1)(2\bar{l}+3)(\bar{l}+2) - 6(2\bar{l}+3)(\bar{l}+2)\bar{\rho} + 6(\bar{l}+2)\bar{\rho}^2 - \bar{\rho}^3}{4(\bar{l}+1)(2\bar{l}+3)(\bar{l}+2)} \end{aligned} \quad (129)$$

The solution of the radial equation for hydrogen-like atoms with broken internal symmetries is then obtained from equations (94) and (123) to be

$$\bar{R} = \bar{c}\bar{\rho}^{\bar{l}} e^{-\bar{\rho}/2} L_{\frac{\bar{l}-2\bar{l}+1}{\bar{n}+\bar{l}}}^{-2\bar{l}+1}(\bar{\rho}) \quad (130)$$

which is a complex number generalization of the standard result.<sup>12</sup> The term  $\bar{\rho}^{\bar{l}}$  can be written explicitly as follows

$$\bar{\rho}^{\bar{l}} = e^{\bar{l} \ln \bar{\rho}} \quad (131)$$

where

$$\ln \bar{\rho} = \ln \rho + j\theta_\rho \quad (132)$$

Using equation (81) for  $\bar{l}$  gives

$$\bar{\rho}^{\bar{l}} = e^{A+jB} \quad (133)$$

where

$$A = (\ell - |m| \sin^2 \theta_\phi) \ln \rho + |m| \theta_\rho \sin \theta_\phi \cos \theta_\phi \quad (134)$$

$$B = -|m| \sin \theta_\phi \cos \theta_\phi \ln \rho + \theta_\rho (\ell - |m| \sin^2 \theta_\phi) \quad (135)$$

Note also that

$$e^{-\bar{\rho}/2} = e^{-\rho/2(\cos \theta_\rho + j \sin \theta_\rho)} \quad (136)$$

The quantities  $\rho$  and  $\theta_\rho$  that appear in equations (134) through (136) depend on  $\eta$  and  $\theta_\eta$  through equations (104) and (105), and in turn  $\eta$  and  $\theta_\eta$  depend on  $n$ ,  $|m|$  and  $\theta_\phi$  through equations (118) and (117) respectively. Combining equations (125), (130), (133) and (136) gives

$$\bar{R} = \bar{c}_e (A - \rho/2 \cos \theta_\rho)_e^{j(B - \rho/2 \sin \theta_\rho + \theta_L)} L_{\frac{2\bar{l}+1}{\bar{n}+\bar{l}}}^{-2\bar{l}+1}(\rho, \theta_\rho, \bar{l}, \theta_L) \quad (137)$$

The interesting thing about the complex number broken symmetry radial solution in equation (137) is that it depends on the magnetic quantum number  $|m|$  as well as on the principal quantum number  $n$ . In fact equations (104) and (119) show that

$$\rho = 2Zr/(a_0 n) \quad (138)$$

$$\sim 2Zr/(a_0 n) [1 + 1/2|m|/n(2 - |m|/n)\sin^2 \theta_\phi]$$

where the approximation is for small  $\theta_\phi$ . Also from equation (122)  $\theta_\rho$  introduces both  $\theta_r$  and  $\theta_\phi$ . Therefore for  $m \neq 0$  both  $\theta_r$  and  $\theta_\phi$  appear in the solution of the complex number radial equation. As the simplest case consider  $n = 1$ ,  $\ell = 0$  and  $m = 0$ . Then  $A = 0$ ,  $B = 0$  and  $\eta = 1$  and

$$\rho = 2Zr/a_0 \quad (139)$$

$$\theta_\rho = \theta_r \quad (140)$$

$$\bar{R} = C_0 e^{-Zr/a_0} \quad (141)$$

$$R_m = C_0 e^{-Zr/a_0} \cos \theta_r \cos(Zr/a_0 \sin \theta_r) \quad (142)$$

The wave functions for hydrogen-like atoms with broken internal symmetries are listed by atomic shells in Section 6.

#### 6. WAVE FUNCTIONS FOR A HYDROGEN-LIKE ATOM WITH BROKEN INTERNAL SYMMETRIES.

This section consists of a table of wave functions for the atomic shells of hydrogen-like atoms with broken internal symmetries. The broken internal symmetry is due basically to the broken internal symmetries of the coordinates which are described by  $\theta_r$ ,  $\theta_\psi$  and  $\theta_\phi$  as discussed in Sections 3 through 5. The complex magnetic quantum number given by equation (70A) appears frequently in the atomic wave functions and will be written as

$$\bar{M}' = |m|\bar{y} \quad M' = |m|y \quad (143)$$

where

$$\bar{y} = \cos \theta_\phi e^{-j\theta_\phi} = \cos^2 \theta_\phi - j \cos \theta_\phi \sin \theta_\phi \quad (144)$$

$$y = \cos \theta_\phi \quad (145)$$

The following is a table of complete wave functions for the K, L, M and N shells. This table is a generalization of the standard scalar results.<sup>1,2</sup>

K shell: n = 1

$$\ell = 0, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 0, L = 0, \bar{n} = 1, n = 1, \\ 2\bar{L} + 1 = 1, \bar{n} + \bar{L} = 1, \bar{\rho} = 2Z\bar{r}/a_0, N = 0$$

$$\bar{\psi}_{1s_0} = \bar{C}_{1s_0} e^{-\bar{\rho}/2} \quad (146)$$

L shell: n = 2

$$\ell = 0, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 0, L = 0, \bar{n} = 2, n = 2, \\ 2\bar{L} + 1 = 1, \bar{n} + \bar{L} = 2, \bar{\rho} = 2Z\bar{r}/(2a_0), N = 1$$

$$\bar{\psi}_{2s_0} = \bar{C}_{2s_0} 1/2(2 - \bar{\rho})e^{-\bar{\rho}/2} \quad (147)$$

$$\ell = 1, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 1, L = 1, \bar{n} = 2, n = 2, \\ 2\bar{L} + 1 = 3, \bar{n} + \bar{L} = 3, \bar{\rho} = 2Z\bar{r}/(2a_0), N = 0$$

$$\bar{\psi}_{2p_0} = \bar{C}_{2p_0} \bar{\rho} e^{-\bar{\rho}/2} \cos \bar{\psi} \quad (148)$$

$$\ell = 1, m = 1, \bar{M}' = \bar{y}, M' = y, \bar{L} = \bar{y}, L = y, \bar{n} = \bar{y} + 1, \\ n = 2(1 - 3/4 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 2\bar{y} + 1, \bar{n} + \bar{L} = 2\bar{y} + 1, \\ \bar{\rho} = 2Z\bar{r}/[(\bar{y} + 1)a_0], N = 0$$

$$\bar{\psi}_{2p_1} = \bar{C}_{2p_1} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \sin \bar{y} \bar{\psi} \cos(y\phi) \quad (149)$$

$\ell = 1, m = -1$ , same as for equation (149)

$$\bar{\psi}_{2p_{-1}} = \bar{C}_{2p_{-1}} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \sin \bar{y} \bar{\psi} \sin(y\phi) \quad (150)$$

M shell: n = 3

$$\begin{aligned} \ell = 0, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 0, L = 0, \bar{n} = 3, n = 3, \\ 2\bar{L} + 1 = 1, \bar{n} + \bar{L} = 3, \bar{\rho} = 2Z\bar{r}/(3a_0), N = 2 \end{aligned}$$

$$\bar{\Psi}_{3s_0} = \bar{C}_{3s_0} e^{-\bar{\rho}/2} \frac{1}{6}(6 - 6\bar{\rho} + \bar{\rho}^2) \quad (151)$$


---

$$\begin{aligned} \ell = 1, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 1, L = 1, \bar{n} = 3, n = 3, \\ 2\bar{L} + 1 = 3, \bar{n} + \bar{L} = 4, \bar{\rho} = 2Z\bar{r}/(3a_0), N = 1 \end{aligned}$$

$$\bar{\Psi}_{3p_0} = \bar{C}_{3p_0} e^{-\bar{\rho}/2} \bar{\rho}/4(4 - \bar{\rho}) \cos \bar{\psi} \quad (152)$$


---

$$\begin{aligned} \ell = 1, m = 1, \bar{M}' = \bar{y}, M' = y, \bar{L} = \bar{y}, L = y, \bar{n} = \bar{y} + 2, \\ n = 3(1 - 5/9 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 2\bar{y} + 1, \bar{n} + \bar{L} = 2\bar{y} + 2, \\ \bar{\rho} = 2Z\bar{r}/[(\bar{y} + 2)a_0], N = 1 \end{aligned}$$

$$\bar{\Psi}_{3p_1} = \bar{C}_{3p_1} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \frac{[2(\bar{y} + 1) - \bar{\rho}]}{2(\bar{y} + 1)} \sin^{\bar{y}} \bar{\psi} \cos(y\phi) \quad (153)$$


---

$\ell = 1, m = -1$ , same as for equation (153)

$$\bar{\Psi}_{3p_{-1}} = \bar{C}_{3p_{-1}} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \frac{[2(\bar{y} + 1) - \bar{\rho}]}{2(\bar{y} + 1)} \sin^{\bar{y}} \bar{\psi} \sin(y\phi) \quad (154)$$


---

$$\begin{aligned} \ell = 2, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 2, L = 2, \bar{n} = 3, n = 3, \\ 2\bar{L} + 1 = 5, \bar{n} + \bar{L} = 5, \bar{\rho} = 2Z\bar{r}/(3a_0), N = 0 \end{aligned}$$

$$\bar{\Psi}_{3d_0} = \bar{C}_{3d_0} e^{-\bar{\rho}/2} \bar{\rho}^2 (3 \cos^2 \bar{\psi} - 1) \quad (155)$$


---

$$\begin{aligned} \ell &= 2, m = 1, \bar{M}' = \bar{y}, M' = y, \bar{L} = \bar{y} + 1, L = 2(1 - 3/4 \sin^2 \theta_\phi)^{1/2}, \\ \bar{n} &= \bar{y} + 2, n = 3(1 - 5/9 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 2\bar{y} + 3, \bar{n} + \bar{L} = 2\bar{y} + 3, \\ \bar{\rho} &= 2Z\bar{r}/[(\bar{y} + 2)a_0], N = 0 \end{aligned}$$

$$\bar{\psi}_{3d_1} = \bar{C}_{3d_1} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+1} \sin^{\bar{y}} \bar{\psi} \cos \bar{\psi} \cos(y\phi) \quad (156)$$

---

$\ell = 2, m = -1$ , same as for equation (156)

$$\bar{\psi}_{3d_{-1}} = \bar{C}_{3d_{-1}} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+1} \sin^{\bar{y}} \bar{\psi} \cos \bar{\psi} \sin(y\phi) \quad (157)$$

---


$$\begin{aligned} \ell &= 2, m = 2, \bar{M}' = 2\bar{y}, M' = 2y, \bar{L} = 2\bar{y}, L = 2y, \bar{n} = 2\bar{y} + 1, \\ n &= 3(1 - 8/9 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 4\bar{y} + 1, \bar{n} + \bar{L} = 4\bar{y} + 1, \\ \bar{\rho} &= 2Z\bar{r}/[(2\bar{y} + 1)a_0], N = 0 \end{aligned}$$

$$\bar{\psi}_{3d_2} = \bar{C}_{3d_2} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}} \sin^{2\bar{y}} \bar{\psi} \cos(2y\phi) \quad (158)$$

---

$\ell = 2, m = -2$ , same as for equation (158)

$$\bar{\psi}_{3d_{-2}} = \bar{C}_{3d_{-2}} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}} \sin^{2\bar{y}} \bar{\psi} \sin(2y\phi) \quad (159)$$

---

N shell: n = 4

$$\begin{aligned} \ell &= 0, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 0, L = 0, \bar{n} = 4, n = 4, \\ 2\bar{L} + 1 &= 1, \bar{n} + \bar{L} = 4, \bar{\rho} = 2Z\bar{r}/(4a_0), N = 3 \end{aligned}$$

$$\bar{\psi}_{4s_0} = \bar{C}_{4s_0} e^{-\bar{\rho}/2} 1/24(24 - 36\bar{\rho} + 12\bar{\rho}^2 - \bar{\rho}^3) \quad (160)$$


---

$$\begin{aligned} \ell = 1, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 1, L = 1, \bar{n} = 4, n = 4, \\ 2\bar{L} + 1 = 3, \bar{n} + \bar{L} = 5, \bar{\rho} = 2Z\bar{r}/(4a_0), N = 2 \end{aligned}$$

$$\bar{\psi}_{4p_0} = \bar{C}_{4p_0} e^{-\bar{\rho}/2} \bar{\rho}/20(20 - 10\bar{\rho} + \bar{\rho}^2) \cos \bar{\psi} \quad (161)$$

$$\begin{aligned} \ell = 1, m = 1, \bar{M}' = \bar{y}, M' = y, \bar{L} = \bar{y}, L = y, \bar{n} = \bar{y} + 3, \\ n = 4(1 - 7/16 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 2\bar{y} + 1, \bar{n} + \bar{L} = 2\bar{y} + 3, \\ \bar{\rho} = 2Z\bar{r}/[(\bar{y} + 3)a_0], N = 2 \end{aligned}$$

$$\bar{\psi}_{4p_1} = \bar{C}_{4p_1} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \frac{[2(\bar{y} + 1)(2\bar{y} + 3) - 2(2\bar{y} + 3)\bar{\rho} + \bar{\rho}^2]}{2(\bar{y} + 1)(2\bar{y} + 3)} \sin \bar{y} \bar{\psi} \cos(y\phi) \quad (162)$$

$$\ell = 1, m = -1, \text{ same as for equation (162)}$$

$$\bar{\psi}_{4p_{-1}} = \bar{C}_{4p_{-1}} e^{-\bar{\rho}/2} \bar{\rho} \bar{y} \frac{[2(\bar{y} + 1)(2\bar{y} + 3) - 2(2\bar{y} + 3)\bar{\rho} + \bar{\rho}^2]}{2(\bar{y} + 1)(2\bar{y} + 3)} \sin \bar{y} \bar{\psi} \sin(y\phi) \quad (163)$$

$$\begin{aligned} \ell = 2, m = 0, \bar{M}' = 0, M' = 0, \bar{L} = 2, L = 2, \bar{n} = 4, n = 4, \\ 2\bar{L} + 1 = 5, \bar{n} + \bar{L} = 6, \bar{\rho} = 2Z\bar{r}/(4a_0), N = 1 \end{aligned}$$

$$\bar{\psi}_{4d_0} = \bar{C}_{4d_0} e^{-\bar{\rho}/2} \bar{\rho}^2 1/6(6 - \bar{\rho})(3 \cos^2 \bar{\psi} - 1) \quad (164)$$

$$\begin{aligned} \ell = 2, m = 1, \bar{M}' = \bar{y}, M' = y, \bar{L} = \bar{y} + 1, L = 2(1 - 3/4 \sin^2 \theta_\phi)^{1/2}, \\ \bar{n} = \bar{y} + 3, n = 4(1 - 7/16 \sin^2 \theta_\phi)^{1/2}, 2\bar{L} + 1 = 2\bar{y} + 3, \bar{n} + \bar{L} = 2\bar{y} + 4, \\ \bar{\rho} = 2Z\bar{r}/[(\bar{y} + 3)a_0], N = 1 \end{aligned}$$

$$\bar{\psi}_{4d_1} = \bar{C}_{4d_1} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+1} \frac{[2(\bar{y} + 2) - \bar{\rho}]}{2(\bar{y} + 2)} \sin \bar{y} \bar{\psi} \cos \bar{\psi} \cos(y\phi) \quad (165)$$



$\ell = 2$  ,  $m = -1$  , same as for equation (165)

$$\bar{\psi}_{4d_{-1}} = \bar{C}_{4d_{-1}} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+1} \frac{[2(\bar{y} + 2) - \bar{\rho}]}{2(\bar{y} + 2)} \sin^{\bar{y}} \bar{\psi} \cos \bar{\psi} \sin(y\phi) \quad (166)$$


---

$\ell = 2$  ,  $m = 2$  ,  $\bar{M}' = 2\bar{y}$  ,  $M' = 2y$  ,  $\bar{L} = 2\bar{y}$  ,  $L = 2y$  ,  $\bar{n} = 2\bar{y} + 2$  ,  
 $n = 4(1 - 3/4 \sin^2 \theta_\phi)^{1/2}$  ,  $2\bar{L} + 1 = 4\bar{y} + 1$  ,  $\bar{n} + \bar{L} = 4\bar{y} + 2$  ,  
 $\bar{\rho} = 2Z\bar{r}/[(2\bar{y} + 2)a_0]$  ,  $N = 1$

$$\bar{\psi}_{4d_2} = \bar{C}_{4d_2} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}} \frac{[2(2\bar{y} + 1) - \bar{\rho}]}{2(2\bar{y} + 1)} \sin^{2\bar{y}} \bar{\psi} \cos(2y\phi) \quad (167)$$


---

$\ell = 2$  ,  $m = -2$  , same as for equation (167)

$$\bar{\psi}_{4d_{-2}} = \bar{C}_{4d_{-2}} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}} \frac{[2(2\bar{y} + 1) - \bar{\rho}]}{2(2\bar{y} + 1)} \sin^{2\bar{y}} \bar{\psi} \sin(2y\phi) \quad (168)$$


---

$\ell = 3$  ,  $m = 0$  ,  $\bar{M}' = 0$  ,  $M' = 0$  ,  $\bar{L} = 3$  ,  $L = 3$  ,  $\bar{n} = 4$  ,  $n = 4$  ,  
 $2\bar{L} + 1 = 7$  ,  $\bar{n} + \bar{L} = 7$  ,  $\bar{\rho} = 2Z\bar{r}/(4a_0)$  ,  $N = 0$

$$\bar{\psi}_{4f_0} = \bar{C}_{4f_0} e^{-\bar{\rho}/2} \bar{\rho}^3 (5 \cos^3 \bar{\psi} - 3 \cos \bar{\psi}) \quad (169)$$


---

$\ell = 3$  ,  $m = 1$  ,  $\bar{M}' = \bar{y}$  ,  $M' = y$  ,  $\bar{L} = \bar{y} + 2$  ,  $L = 3(1 - 5/9 \sin^2 \theta_\phi)^{1/2}$  ,  
 $\bar{n} = \bar{y} + 3$  ,  $n = 4(1 - 7/16 \sin^2 \theta_\phi)^{1/2}$  ,  $2\bar{L} + 1 = 2\bar{y} + 5$  ,  $\bar{n} + \bar{L} = 2\bar{y} + 5$  ,  
 $\bar{\rho} = 2Z\bar{r}/[(\bar{y} + 3)a_0]$  ,  $N = 0$

$$\bar{\psi}_{4f_1} = \bar{C}_{4f_1} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+2} \sin^{\bar{y}} \bar{\psi} [(2\bar{y} + 3)\cos^2 \bar{\psi} - 1]\cos(y\phi) \quad (170)$$


---

$\ell = 3$  ,  $m = -1$  , same as for equation (170)

$$\bar{\psi}_{4f_{-1}} = \bar{C}_{4f_{-1}} e^{-\bar{\rho}/2} \bar{\rho}^{\bar{y}+2} \sin^{\bar{y}} \bar{\psi} [(2\bar{y} + 3)\cos^2 \bar{\psi} - 1] \sin(y\phi) \quad (171)$$

$\ell = 3$  ,  $m = 2$  ,  $\bar{M}' = 2\bar{y}$  ,  $M' = 2y$  ,  $\bar{L} = 2\bar{y} + 1$  ,  $L = 3(1 - 8/9 \sin^2 \theta_\phi)^{1/2}$  ,  
 $\bar{n} = 2\bar{y} + 2$  ,  $n = 4(1 - 3/4 \sin^2 \theta_\phi)^{1/2}$  ,  $2\bar{L} + 1 = 4\bar{y} + 3$  ,  $\bar{n} + \bar{L} = 4\bar{y} + 3$  ,  
 $\bar{\rho} = 2Z\bar{r}/[(2\bar{y} + 2)a_0]$  ,  $N = 0$

$$\bar{\psi}_{4f_2} = \bar{C}_{4f_2} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}+1} \sin^{2\bar{y}} \bar{\psi} \cos \bar{\psi} \cos(2y\phi) \quad (172)$$

$\ell = 3$  ,  $m = -2$  , same as for equation (172)

$$\bar{\psi}_{4f_{-2}} = \bar{C}_{4f_{-2}} e^{-\bar{\rho}/2} \bar{\rho}^{2\bar{y}+1} \sin^{2\bar{y}} \bar{\psi} \cos \bar{\psi} \sin(2y\phi) \quad (173)$$

$\ell = 3$  ,  $m = 3$  ,  $\bar{M}' = 3\bar{y}$  ,  $M' = 3y$  ,  $\bar{L} = 3\bar{y}$  ,  $L = 3y$  ,  $\bar{n} = 3\bar{y} + 1$  ,  
 $n = 4(1 - 15/16 \sin^2 \theta_\phi)^{1/2}$  ,  $2\bar{L} + 1 = 6\bar{y} + 1$  ,  $\bar{n} + \bar{L} = 6\bar{y} + 1$  ,  
 $\bar{\rho} = 2Z\bar{r}/[(3\bar{y} + 1)a_0]$  ,  $N = 0$

$$\bar{\psi}_{4f_3} = \bar{C}_{4f_3} e^{-\bar{\rho}/2} \bar{\rho}^{3\bar{y}} \sin^{3\bar{y}} \bar{\psi} \cos(3y\phi) \quad (174)$$

$\ell = 3$  ,  $m = -3$  , same as for equation (174)

$$\bar{\psi}_{4f_{-3}} = \bar{C}_{4f_{-3}} e^{-\bar{\rho}/2} \bar{\rho}^{3\bar{y}} \sin^{3\bar{y}} \bar{\psi} \sin(3y\phi) \quad (175)$$

The reader should be aware that the value of  $\bar{\rho}$  that appears in the above table depends on  $\bar{n}$  and is thus dependent on  $n$  and  $|m|$  .

7. ENERGY LEVELS AND RADII OF HYDROGEN-LIKE ATOMS WITH BROKEN INTERNAL SYMMETRY. This section calculates the measurable values of the energy levels and atomic radii of hydrogen-like atoms with broken internal symmetries. The complex number energy levels are obtained from equations (95) and (97) to be

$$\bar{E}_\eta = E_\eta e^{j\theta_{E\eta}} = -\hbar^2/(2\mu)\bar{a}^2 = -\mu Z^2 e^4/(2\hbar^2 \bar{\eta}^2) \quad (176)$$

where  $\bar{\eta}$  is given by equations (109) or (114). From equation (176) it follows that

$$E_\eta = -\mu Z^2 e^4/(2\hbar^2 \eta^2) \quad (177)$$

$$\sim -\mu Z^2 e^4/(2\hbar^2 \eta^2) [1 + |m|/\eta(2 - |m|/\eta)\sin^2 \theta_\phi]$$

$$\theta_{E\eta} = -2\theta_\eta \quad (178)$$

where  $\eta$  and  $\theta_\eta$  are given by equations (118) and (117) respectively. The measured energy levels are given by<sup>7</sup>

$$E_{\eta m} = E_\eta \cos \theta_{E\eta} \quad (179)$$

$$= -\mu Z^2 e^4/(2\hbar^2 \eta^2) \cos(2\theta_\eta)$$

From equation (116) it follows that

$$\sin \theta_\eta = -|m|/\eta \cos \theta_\phi \sin \theta_\phi \quad (180)$$

$$\cos(2\theta_\eta) = 1 - 2 \sin^2 \theta_\eta \quad (181)$$

$$= 1 - 2(|m|/\eta)^2 \cos^2 \theta_\phi \sin^2 \theta_\phi$$

$$\sim 1 - 2(|m|/\eta)^2 \cos^2 \theta_\phi \sin^2 \theta_\phi$$

Combining equations (177), (179) and (181) gives

$$E_{\eta m} \sim -\mu Z^2 e^4/(2\hbar^2 \eta^2) (1 + F \sin^2 \theta_\phi) \quad (182)$$

where

$$F = \{|m|/\eta[2 - |m|/\eta(1 + 2 \cos^2 \theta_\phi)]\} \quad (183)$$

$$\sim |m|/\eta(2 - 3|m|/\eta)$$

The values of F are given approximately for small  $\theta_\phi$  as follows

$$F = 0 \left\{ \begin{array}{l} m = 0 \text{ or } |m|/n = 2/3 \end{array} \right. \quad (184)$$

$$F > 0 \left\{ \begin{array}{l} |m|/n < 2/3 \end{array} \right. \quad (185)$$

$$F < 0 \left\{ \begin{array}{l} |m|/n > 2/3 \end{array} \right. \quad (186)$$

Equations (177), (179) and (182) reduce to the standard Bohr result in equation (10) for the case  $\theta_\phi = 0$ .

For the K and L shells  $|m|/n < 1/2$  so that  $F \geq 0$ . For the M and P shells, etc, it is possible to have the situation  $|m|/n = 2/3$  and  $F = 0$ . For the N, O, P, ... shells it is possible to have  $|m|/n > 2/3$  and  $F = < 0$ . In any case it is clear from equations (182) and (183) that the measured energy eigenvalues of hydrogen-like atoms with broken internal symmetry should depend on  $|m|$  as well as on the principal quantum number  $n$ . In addition, the asymmetry factor in equation (182) is  $F \sin^2 \theta_\phi$  where  $\theta_\phi$  depends on the pressure acting on the system of atoms. Thus the broken symmetry of the azimuthal angle destroys the degeneracy associated with the energy eigenvalues given by equation (10). The energy eigenvalues do not depend explicitly on  $\ell$  and therefore some remaining degeneracy still exists.

A transition from a state  $n', m'$  to the state  $n, m$  is according to equation (182) associated with the energy difference

$$\Delta E_{nm} = \mu Z^2 e^4 / (2\hbar^2) [1/n^2 - 1/n'^2 + (F/n^2 - F'/n'^2) \sin^2 \theta_\phi] \quad (187)$$

where

$$F/n^2 = |m|/n^3 [2 - |m|/n (1 + 2 \cos^2 \theta_\phi)] \quad (188)$$

$$F'/n'^2 = |m'|/n'^3 [2 - |m'|/n' (1 + 2 \cos^2 \theta_\phi)] \quad (189)$$

$$\begin{aligned} F/n^2 - F'/n'^2 &= 2(|m|/n^3 - |m'|/n'^3) - (|m|^2/n^4 - |m'|^2/n'^4) (1 + 2 \cos^2 \theta_\phi) \quad (190) \\ &\sim 2(|m|/n^3 - |m'|/n'^3) - 3(|m|^2/n^4 - |m'|^2/n'^4) \end{aligned}$$

It is the internal phase angle  $\theta_\phi$  of the azimuthal angle that introduces the magnetic quantum number into the energy eigenvalues given in equation (177) and in the formula for the transition energy given in equation (187).

The pressure variation of the energy eigenvalues will now be calculated. From equation (179) it follows that

$$\partial E_{nm} / \partial P = \mu Z^2 e^4 / (\hbar^2 n^2) [\cos(2\theta_\phi) 1/n \partial \eta / \partial P + \sin(2\theta_\phi) \partial \theta_\phi / \partial P] \quad (191)$$

The derivatives on the right hand side of equation (191) are easily evaluated. The value of  $\partial \eta / \partial P$  is obtained from equation (118) to be

$$1/n \partial n / \partial P = - \kappa_1 \partial \theta_\phi / \partial P \quad (192)$$

where

$$\kappa_1 = \frac{|m|/n(2 - |m|/n) \sin \theta_\phi \cos \theta_\phi}{1 - |m|/n(2 - |m|/n) \sin^2 \theta_\phi} \quad (193)$$

$$\sim |m|/n(2 - |m|/n) \sin \theta_\phi \cos \theta_\phi$$

The value of  $\partial \theta_\eta / \partial P$  can be obtained from equation (117) with the result

$$\partial \theta_\eta / \partial P = - \kappa_2 \partial \theta_\phi / \partial P \quad (194)$$

where

$$\kappa_2 = \frac{n|m| \cos(2\theta_\phi) + |m|^2 \sin^2 \theta_\phi}{n^2 - |m|(2n - |m|) \sin^2 \theta_\phi} \quad (195)$$

$$\sim |m|/n$$

Placing equations (192) and (194) into equation (191) gives

$$\partial E_{nm} / \partial P = - \mu Z^2 e^4 / (\hbar^2 n^2) [\kappa_1 \cos(2\theta_\eta) + \kappa_2 \sin(2\theta_\eta)] \partial \theta_\phi / \partial P \quad (196)$$

where from equations (115) and (116)

$$\cos(2\theta_\eta) \sim 1 - 2(|m|/n)^2 \cos^2 \theta_\phi \sin^2 \theta_\phi \quad (197)$$

$$\sim 1$$

$$\sin(2\theta_\eta) = - 2|m|/n^2 (n - |m| \sin^2 \theta_\phi) \sin \theta_\phi \cos \theta_\phi \quad (198)$$

$$\sim - 2|m|/n \sin \theta_\phi \cos \theta_\phi$$

$$\theta_\eta \sim - |m|/n \theta_\phi \quad (199)$$

for small  $\theta_\phi$ . Therefore combining equations (193) and (195) through (198) gives

$$\partial E_{nm} / \partial P \sim - \mu Z^2 e^4 / (\hbar^2 n^2) |m|/n(2 - 3|m|/n) \sin \theta_\phi \cos \theta_\phi \partial \theta_\phi / \partial P \quad (200)$$

The rate of change of the energy eigenstates with pressure can be positive or negative according to the value of  $|m|/n$ . In particular if  $\partial \theta_\phi / \partial P > 0$  equation (200) gives

$$\left. \begin{array}{l} \partial E_{\eta m} / \partial P \leq 0 \\ \partial E_{\eta m} / \partial P > 0 \end{array} \right\} \begin{array}{l} |m|/n \leq 2/3 \\ 2/3 < |m|/n < 1 \end{array} \quad (201)$$

$$\left. \begin{array}{l} \partial E_{\eta m} / \partial P \leq 0 \\ \partial E_{\eta m} / \partial P > 0 \end{array} \right\} \begin{array}{l} |m|/n \leq 2/3 \\ 2/3 < |m|/n < 1 \end{array} \quad (202)$$

For  $|m|/n < 2/3$  an increase in pressure will lead to a greater binding energy per electron while for  $|m|/n > 2/3$  an increase in pressure will produce less binding energy for the valence electron. These conclusions also follow directly from equations (182) and (183).

Finally, the simplest generalization of equation (9) for the radii of the Bohr orbits is given by

$$\bar{r}_n = \bar{n}^2 \hbar^2 / (\mu Z e^2) \quad (203)$$

and therefore combining equations (97), (118) and (203) gives

$$r_n = n^2 \hbar^2 / (\mu Z e^2) = n^2 \hbar^2 / (\mu Z e^2) [1 - |m|/n (2 - |m|/n) \sin^2 \theta_\phi] \quad (204)$$

$$\theta_{rn} = 2\theta_n \sim -2|m|/n \theta_\phi \quad (205)$$

where  $\theta_n$  is given by equations (117) and (120). The radii of the Bohr orbits of an atom with broken internal symmetry are pressure dependent through  $\theta_\phi(P)$ . The measured Bohr radii are given by

$$\begin{aligned} r_{nm} &= r_n \cos \theta_{rn} \\ &\sim n^2 \hbar^2 / (\mu Z e^2) (1 - G \sin^2 \theta_\phi) \end{aligned} \quad (206)$$

where

$$\begin{aligned} G &= |m|/n [2 + |m|/n \cos(2\theta_\phi)] \\ &\sim |m|/n (2 + |m|/n) \end{aligned} \quad (207)$$

where  $G \geq 0$ . From equations (206) and (207) it follows that the measured Bohr radii are pressure dependent and

$$\partial r_{nm} / \partial P \sim -2n^2 \hbar^2 / (\mu Z e^2) |m|/n (2 + |m|/n) \sin \theta_\phi \cos \theta_\phi \partial \theta_\phi / \partial P \quad (208)$$

Therefore if  $\partial \theta_\phi / \partial P > 0$  it follows that  $\partial r_{nm} / \partial P < 0$ . The analysis leading to equations (203) through (208) is simplistic because higher order terms associated with the azimuthal angular momentum  $\bar{L}$  need to be inserted into equation (203).<sup>12</sup> The results for the energy levels and atomic radii may possibly be tested using circular atoms for which  $|m| = \ell = n - 1$ .<sup>15</sup>

8. CONCLUSION. The pressure dependence of the energy levels and radii of hydrogen-like atoms can be determined by taking into account the broken symmetries of the coordinates of the electron and the nucleus. The broken symmetries of the azimuthal and zenith angles are due essentially to a broken symmetry pressure field. But the vacuum state also exhibits this broken symmetry.<sup>16</sup> The broken internal symmetries of the coordinates requires the magnetic, azimuthal and principal quantum numbers to be complex numbers which are associated with internal phase angles. The internal phase angles of the three quantum numbers are expected to be pressure dependent. Schrödinger's equation for a hydrogen-like atom can be solved with complex quantum numbers, and break off solutions to the azimuthal angle, zenith angle and radial equations can be obtained.

The broken symmetry of the azimuthal angle may be associated with a vector boson which can be called the "muthon". The muthon may have important physical effects in systems having large scale broken azimuthal symmetry such as perhaps in the layered copper oxide structures of high temperature superconductors where it may serve as the intermediary particle for the formation of Cooper pairs of electrons or holes. Alternatively, the muthon could also play a role as the component of dark matter in galaxies where large scale broken azimuthal symmetry may occur due to the broken internal symmetry of the gravitational field.

#### ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

#### REFERENCES

1. Itzykson, C. and Zuber, J. B., Quantum Field Theory, McGraw-Hill, New York, 1980.
2. Becher, P., Böhm, M. and Joos, H., Gauge Theories of Strong and Electroweak Interactions, John Wiley, New York, 1984.
3. Ryder, L. H., Quantum Field Theory, Cambridge University Press, 1985.
4. DeWit, B. and Smith, J., Field Theory in Particle Physics, North-Holland, New York, 1986.
5. Weiss, R. A., "Scale Invariant Equations for Relativistic Waves," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 87-1, 27-30 May 1986, p. 307.
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, New York, ARO 88-1, 15-18 June 1987, p. 649.
7. Weiss, R. A., "The Broken Symmetry of Space and Time in Bulk Matter and the Vacuum," Sixth Army Conference on Applied Mathematics and Computing, Boulder, Colorado, ARO 89-1, 31 May-3 June 1988, p. 317.
8. Witten, E., "Space-Time and Topological Orbifolds," Phys. Rev. Lett., Vol. 61, No. 6, 8 Aug 1988, p. 570.

9. Weiss, R. A., "Gauge Theory of Atomic Processes," Sixth Army Conference on Applied Mathematics and Computing, Boulder, Colorado, ARO 89-1, 31 May-3 June 1988, p. 223.
10. Richtmyer, F. K., Kennard, E. H. and Lauritsen, T., Introduction to Modern Physics, McGraw-Hill, New York, 1955.
11. Leighton, R. B., Principles of Modern Physics, McGraw-Hill, New York, 1959.
12. Pauling, L. and Wilson, E. B., Introduction to Quantum Mechanics, McGraw-Hill, New York, 1935.
13. Merzbacher, E., Quantum Mechanics, John Wiley, New York, 1961.
14. Powell, J. L. and Crasemann, B., Quantum Mechanics, Addison-Wesley, Reading, Massachusetts, 1961.
15. Hare, J., Gross, M. and Goy, P., "Circular Atoms Prepared by a New Method of Crossed Electric and Magnetic Fields," Phys. Rev. Lett., Vol. 61, No. 17, 24 Oct. 1988, p. 1938.
16. Weiss, R. A., "Maxwell's Equations with Broken Internal Symmetries," Sixth Army Conference on Applied Mathematics and Computing, Boulder, Colorado, ARO 89-1, 31 May-3 June 1988, p.271.



NEWTONIAN GRAVITY IN MATTER  
WITH BROKEN INTERNAL SYMMETRY

Richard A. Weiss  
U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The pressure field in matter is associated with a broken internal symmetry which manifests itself through the broken internal symmetry of space and time coordinates. This introduces an apparent non-Newtonian behaviour of gravity in matter. The effective Newtonian gravitational constant for a spherical body composed of matter with broken internal symmetry is calculated and determined to be a function of radial distance from the center of a planet or star. The gravity field of a rotating geometrically asymmetric planet composed of matter with broken internal symmetries is investigated. A theoretical analysis of Eötvös, mine shaft, borehole and tower gravity variation experiments is presented in terms of Newtonian gravity in matter with broken internal symmetry. It is found that the discrepancies from Newtonian gravity can be described by ordinary Newtonian gravity in matter with broken internal symmetry combined with the variation of atmospheric pressure down a mine shaft or borehole and up a tower. This research will affect the calculation of trajectories of missiles and projectiles in the earth's atmosphere, and will have applications to geophysics and astrophysics.

**1. INTRODUCTION.** Discrepancies between Newton's law of gravitation and the measured variation of gravity with distance and composition of the attracting bodies have been observed. These discrepancies appeared first in the measurements of the variation of the gravity force with depth in mine shafts.<sup>1-4</sup> These measurements indicate a larger value of the gravitational constant than is found from laboratory Eötvös experiments.<sup>1-4</sup> On the other hand, recent experiments on the variation of gravity up the length of a tower suggest a value of the gravitational constant which is less than that measured in the laboratory by Eötvös experiments.<sup>5</sup> Differences in behaviour from Newtonian gravity have also been reported for the Eötvös type of experiments and with beam balance experiments.<sup>6-15</sup> Other evidence for non-Newtonian behaviour has been presented from solar system and stellar system measurements.<sup>16-18</sup>

Attempts to explain these measured results by the introduction of new types of gravitational forces (the "fifth" and "sixth" forces) that have finite ranges of the order of hundreds or thousands of meters have been suggested.<sup>9-23</sup> These new forces would represent the effects of massive spin 0 and spin 1 supersymmetric partners to the ordinary massless spin 2 graviton that mediates Newtonian gravitation with its infinite range.<sup>19-23</sup> Much criticism of the reality of these finite range forces has been presented.<sup>24,25</sup> This is due in part to the difficulties of separating extraneous effects due to geological structure from the possible intrinsic non-Newtonian behaviour of gravity. In fact recent data from a borehole in the ice of a glacier in Greenland suggests that the gravitation constant is less than that measured by laboratory Eötvös experiments, and this disagrees with the results given in Reference 1-4 but

agrees with the observations in Reference 5. The state of both the experimental and theoretical situation is therefore uncertain.

This paper suggests an alternative explanation for the apparent non-Newtonian behaviour of gravity in the earth which is based on ordinary Newtonian gravitation and the broken symmetry of the thermodynamic and mechanical parameters of bulk matter such as pressure and internal energy.<sup>26,27</sup> Some results have already been obtained toward describing the apparent non-Newtonian behaviour of gravity in terms of the ordinary Newtonian gravity field in matter with broken internal symmetries.<sup>28</sup> This was done by showing that the space and time coordinates exhibit broken symmetries in matter where the pressure has a broken internal symmetry.<sup>28</sup> Section 2 introduces the relationship between Newtonian gravity and the broken internal symmetries of space and time, Section 3 deals with complex number coordinates and the measurement of space and time, Section 4 considers Newtonian gravity for rotating non-spherical masses composed of matter that induces broken symmetries in the pressure and coordinates, Section 5 presents a theory for the description of the Eötvös, mine shaft, borehole and tower experiments, and finally Section 6 gives a numerical calculation of the expected values of the internal phase angles of the radial and angular coordinates due to the earth's gravity field.

2. NEWTONIAN GRAVITY AND BROKEN INTERNAL SYMMETRIES. A gauge theory of relativistic thermodynamics has been developed which is based on a trace equation which for completely symmetrical matter or radiation is given by<sup>29</sup>

$$U + T \left( \frac{dU}{dT} \right)_{PV} - 3V \frac{d}{dV} (PV)_U = U^a + T \left( \frac{dU^a}{dT} \right)_{p^a V} \quad (1)$$

where  $U$  = relativistic internal energy,  $P$  = relativistic pressure,  $T$  = absolute temperature,  $V$  = volume of substance, and  $U^a$  and  $P^a$  = corresponding non-relativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic calculations. The temperature and volume are parameters for both the renormalized and unrenormalized systems. The trace equation for matter whose thermodynamic functions have broken internal symmetries is given by<sup>27</sup>

$$\bar{U} + T \left( \frac{d\bar{U}}{dT} \right)_{\bar{P}V} - 3V \frac{d}{dV} (\bar{P}V)_{\bar{U}} = U^a + T \left( \frac{dU^a}{dT} \right)_{p^a V} \quad (2)$$

where  $\bar{U}$  and  $\bar{P}$  are complex number representations of the renormalized internal energy and pressure respectively, and where  $T$  and  $V$  are the magnitudes of the complex number temperature and volume respectively. Equation (2) can be further simplified by using the following complex number form of the Gibbs-Helmholtz-Maxwell equation<sup>27</sup>

$$\frac{\partial \bar{U}}{\partial V} = T \frac{\partial \bar{P}}{\partial T} - \bar{P} \quad (3)$$

The complex numbers  $\bar{U}$  and  $\bar{P}$  that appear in equation (2) are written as<sup>27</sup>

$$\bar{U} = Ue^{j\theta_U} \quad (4)$$

$$\bar{P} = Pe^{j\theta_P} \quad (5)$$

where  $U$ ,  $P$ ,  $\theta_U$  and  $\theta_P$  can be obtained from a solution of equations (2) and (3). The temperature and volume parameters that appear in equation (2) are real numbers. However the temperature and volume themselves are complex numbers that are written as

$$\bar{T} = Te^{j\theta_T} \quad (6)$$

$$\bar{V} = Ve^{j\theta_V} \quad (7)$$

where  $T$  and  $V$  are the magnitudes of the temperature and volume and it is these quantities that appear in the trace equation (2). The measured thermodynamic quantities are given by<sup>28</sup>

$$U_m = U \cos \theta_U \quad (8)$$

$$P_m = P \cos \theta_P \quad (9)$$

$$V_m = V \cos \theta_V \quad (10)$$

$$T_m = T \cos \theta_T \quad (11)$$

The phase angles  $\theta_U$  and  $\theta_P$  are obtained from equations (2) and (3), while  $\theta_V$  and  $\theta_T$  are related to coordinate and velocity internal phase angles as will be shown later.

The determination of the space and time coordinate internal phase angles follows from the complex number Euler equations<sup>28</sup>

$$\partial \bar{v} / \partial \bar{t} = - \cos \beta_{r,r} \partial \bar{P} / \partial r + \rho \bar{F}_r \quad (12)$$

where the complex number external force (such as gravity) is written as

$$\bar{F}_r = F_r e^{j\theta_{Fr}} = - \partial \bar{w} / \partial \bar{r} \quad (13)$$

and where  $\bar{v}$  = complex number velocity,  $\bar{t}$  = complex number time and  $\bar{r}$  = complex number radial coordinate. The complex number velocity and space and time coordinates are written as<sup>28</sup>

$$\bar{r} = re^{j\theta_r} \quad (14)$$

$$\bar{t} = te^{j\theta_t} \quad (15)$$

$$\bar{v} = ve^{j\theta_v} \quad (16)$$

The measured values of the space and time coordinates and the particle velocity are given by<sup>28</sup>

$$r_m = r \cos \theta_r \quad (17)$$

$$t_m = t \cos \theta_t \quad (18)$$

$$v_m = v \cos \theta_v \quad (19)$$

For matter in equilibrium equation (12) becomes

$$\rho d\bar{v}/d\bar{t} = - \cos \beta_{r,r} \partial \bar{P} / \partial r + \rho \bar{F}_r = 0 \quad (20)$$

where<sup>28</sup>

$$d\bar{v}/d\bar{t} = [(dv/dt)^2 + (vd\theta_v/dt)^2]^{1/2} \cos \beta_{t,t} e^{j\phi_v} \quad (21)$$

$$\cos \beta_{r,r} \partial \bar{P} / \partial r = [(\partial P / \partial r)^2 + (P \partial \theta_p / \partial r)^2]^{1/2} \cos \beta_{r,r} e^{j\phi_p} \quad (22)$$

$$\phi_v = \theta_v + \beta_{v,t} - \theta_t - \beta_{t,t} \quad (23)$$

$$= \theta_r + \beta_{r,r} + \beta_{v,t} - 2(\theta_t + \beta_{t,t})$$

$$\phi_p = \theta_p + \beta_{p,r} \quad (24)$$

$$\tan \beta_{v,t} = \frac{vd\theta_v/dt}{dv/dt} \quad (25)$$

$$\tan \beta_{t,t} = t \partial \theta_t / \partial t \quad (26)$$

$$\tan \beta_{p,r} = \frac{P \partial \theta_p / \partial r}{\partial P / \partial r} \quad (27)$$

$$\tan \beta_{r,r} = r \partial \theta_r / \partial r \quad (28)$$

To obtain the second relation in equation (23) the following relationship is used<sup>28</sup>

$$\theta_v = \theta_r + \beta_{r,r} - \theta_t - \beta_{t,t} \quad (29)$$

Combining equations (21) through (29) shows that the equilibrium condition given by equation (20) is equivalent to

$$\phi_v = 0 \quad (30)$$

$$dv/dt = 0 \quad (31)$$

$$d\theta_v/dt = 0 \quad (32)$$

and<sup>28</sup>

$$\phi_p = \theta_{Fr} \quad (33A)$$

$$[(\partial P/\partial r)^2 + (P\partial\theta_p/\partial r)^2]^{1/2} \cos \beta_{r,r} = \rho F_r \quad (33B)$$

From equations (23) and (30) it follows that for equilibrium

$$0 = \theta_v + \beta_{v,t} - \theta_t - \beta_{t,t} \quad (34)$$

$$= \theta_r + \beta_{r,r} + \beta_{v,t} - 2(\theta_t + \beta_{t,t})$$

Neglecting the  $\beta$ 's in equation (34) gives the following approximation

$$\theta_v \sim \theta_t \sim \theta_r/2 \quad (35)$$

The relationship between  $\theta_r$  and  $\theta_p$  is obtained from equations (33A) and (33B). For gravity, equations (33A) and (33B) yield a set of coupled differential equations for  $P$ ,  $\theta_p$  and  $\theta_r$ .<sup>28</sup> An approximate solution of equation (33A) gives the following equation for matter in a gravity field<sup>28</sup>

$$\theta_r \sim -\theta_p \quad (36)$$

Then equations (35) and (36) give for a gravitating system

$$\theta_v \sim \theta_t \sim -\theta_p/2 \quad (37)$$

For a general system one has

$$\theta_r \sim -\sigma\theta_p \quad (38)$$

where  $\sigma$  = index that describes the state equation for matter. Equations (30) through (33) give the general conditions of equilibrium. For photons  $\theta_t = \theta_r$ , and the light speed has a zero internal phase angle.

From equation (36) and the relation  $\bar{V} \sim 4/3\pi\bar{r}^3$  it follows that the phase angle for the volume is given by

$$\theta_v \sim 3\theta_r \sim -3\theta_p \quad (40)$$

for a gravitating system. For a uniform system the volume is given by

$$\bar{v} = e^{j\theta_V} \frac{4}{3}\pi r^3 = e^{j\theta_V} \frac{4}{3}\pi r_a^3 \quad (41)$$

so that

$$V = V^a \quad (42)$$

The renormalized and unrenormalized scalar coordinates are parameters related by<sup>28</sup>

$$r = r^a \quad (43)$$

$$\psi = \psi^a \quad (44)$$

$$\phi = \phi^a \quad (45)$$

where  $r$  = magnitude of radial coordinate,  $\psi$  = magnitude of the complex number zenith angle and  $\phi$  = magnitude of the complex number azimuthal angle. Thus  $V$  and  $V^a$  are simply equivalent parameters in the trace equations (1) and (2).

The determination of the broken symmetry phase angle of the temperature  $\theta_T$  is determined from the energy equipartition theorem which can be written for a complex number particle velocity as

$$\bar{\epsilon} = \langle 1/2 m \bar{v}^2 \rangle = k\bar{T} \quad (46)$$

where  $\bar{\epsilon}$  = complex number average kinetic energy per particle and where  $m$  = particle mass and  $k$  = Boltzmann constant. The real and imaginary parts of equation (46) can be written as

$$\epsilon_R = m/2 \int_0^\infty v^2 \cos(2\theta_v) g(v) dv = kT \cos \theta_T \quad (47)$$

$$\epsilon_I = m/2 \int_0^\infty v^2 \sin(2\theta_v) g(v) dv = kT \sin \theta_T \quad (48)$$

where  $g(v)$  = renormalized molecular velocity distribution function. Then

$$\tan \theta_T = \epsilon_I / \epsilon_R \quad (49)$$

$$kT = (\epsilon_R^2 + \epsilon_I^2)^{1/2} \quad (50)$$

which are the equations of  $\theta_T$  and  $T$ . For a gravitating system  $\theta_v$  is given by equation (37), and for this case  $\theta_v$  is independent of the molecular speeds. Therefore for this case it follows from equation (37) and equations (46) through (48) that

$$\theta_T = 2\theta_v \sim -\theta_P \quad (51)$$

Because the magnitude of the temperature  $T$  must appear on both sides of the basic trace equation (1) and (2) it follows that

$$T = T^a \quad (52)$$

where

$$kT^a = m/2 \int_0^\infty v_a^2 g^a(v^a) dv^a \quad (53)$$

In fact equation (52) implies the validity of equation (51) and the relation  $g = g^a$ . Therefore  $T$  and  $T^a$  simply play the roles of equivalent parameters in the trace equations (1) and (2). The measured temperature is obtained from equations (11), (51) and (52) to be

$$T_m = T \cos \theta_T = T \cos \theta_P = T^a \cos \theta_P \quad (54)$$

In view of the complex number values of the volume and temperature it might be thought that the trace equation (2) should be written in the following completely asymmetric form

$$\bar{U} + \bar{T} \left( \frac{d\bar{U}}{d\bar{T}} \right)_{\bar{P}\bar{V}} - 3\bar{V} \frac{d}{d\bar{V}} (\bar{P}\bar{V})_{\bar{U}} = U^a + T^a \left( \frac{dU^a}{dT^a} \right)_{P^a V^a} \quad (55)$$

but this is not correct as can be seen by applying equation (55) to the real classical gases. The experimental fact that the first term of the virial expansion (the ideal gas) and the second virial coefficient must be unaffected by equation (55) requires that the temperature and volume terms that appear as complex numbers in the left hand side of equation (55) must in fact actually appear as real numbers equal to the magnitudes of the complex number volume and temperature as shown correctly in equations (2), (42) and (52).<sup>27</sup> The trace equation corresponds to a uniform pressure and energy density system so that equations (42) and (52) are implicitly assumed in equations (1) and (2).

For systems with nonuniform pressure fields, the determination of the internal phase angles of the coordinates generally involves the solution of coupled differential equations.<sup>28</sup> Thus for gravitating stars or planets the determination of  $\theta_r$  involves the solution of the following two equations<sup>28</sup>

$$\cos \beta_{r,r} \frac{1}{r} \frac{\partial}{\partial r} \left\{ \frac{r^2}{\rho} \frac{\partial P}{\partial r} \cos \beta_{r,r} \left[ 1 + P^2 \left( \frac{\partial \theta_P / \partial r}{\partial P / \partial r} \right)^2 \right]^{1/2} \right\} = -4\pi G \rho \quad (56)$$

$$\theta_P + \tan^{-1} \left( P \frac{\partial \theta_P / \partial r}{\partial P / \partial r} \right) = -2\theta_r + \pi \quad (57)$$

combined with the solution of the relativistic trace equation (2) which links

$\theta_p$  and  $P$  to density and temperature for gases, solids or Fermi liquids with internal phase. Equation (56) is the combined equation that arises from the following equilibrium equation<sup>28</sup>

$$\partial P / \partial r \cos \beta_{r,r} \left[ 1 + P^2 \left( \frac{\partial \theta_p / \partial r}{\partial P / \partial r} \right)^2 \right]^{1/2} = - GM_p / r^2 \quad (58)$$

and the relationship of mass and density (which will be treated in Section 3) given by<sup>28</sup>

$$\cos \beta_{r,r} \partial M / \partial r = 4\pi r^2 \rho \quad (59)$$

When the internal phase angles are set to zero the equilibrium equation (58) reduces to the standard result<sup>30</sup>

$$\partial P / \partial r = - GM_p / r^2 \quad (60)$$

The small gradient approximation to equation (57) is

$$\theta_p + P \frac{\partial \theta_p / \partial r}{\partial P / \partial r} = - 2\theta_r \quad (61)$$

which will be used in Section 5 for approximate solutions for  $\theta_r$ .

Newton's gravitational law can be written for spatial coordinates with broken internal symmetry as<sup>28</sup>

$$\bar{g} = - GM / \bar{r}^2 \quad (62)$$

where  $\bar{g}$  = complex number acceleration of gravity. The measured acceleration of gravity is given by the real value of equation (62) as follows<sup>28</sup>

$$g_m = - GM / r^2 \cos(2\theta_r) \quad (63)$$

Written in terms of the measured radial coordinate given by equation (17) gives<sup>28</sup>

$$g_m = - GM / r_m^2 \cos(2\theta_r) \cos^2 \theta_r \quad (64)$$

These formulas are valid for spherical masses. The conventional value of the acceleration of gravity is expressed in terms of the measured radial distance as<sup>28</sup>

$$g_c = - GM / r_m^2 = - GM / r^2 \cos^{-2} \theta_r \quad (65)$$

and therefore<sup>28</sup>



$$g_m - g_c = GM/r_m^2 [1 - \cos(2\theta_r) \cos^2 \theta_r] \quad (66)$$

$$\sim 3\theta_r^2 GM/r_m^2$$

The derivatives of  $g_m$  and  $g_c$  with respect to  $r$  are given by<sup>28</sup>

$$\partial g_m / \partial r = 2GM/r^3 [\cos(2\theta_r) + r\partial\theta_r/\partial r \sin(2\theta_r)] - 4\pi G\rho \cos(2\theta_r) \quad (67)$$

$$\partial g_c / \partial r = 2GM/(r^3 \cos^2 \theta_r) (1 - \tan \theta_r r\partial\theta_r/\partial r) - 4\pi G\rho/\cos^2 \theta_r \quad (68)$$

Then a parameter  $D$  can be defined given by<sup>28</sup>

$$D = \frac{\partial g_m / \partial r_m - \partial g_c / \partial r_m}{\partial g_c / \partial r_m} = \frac{\partial g_m / \partial r - \partial g_c / \partial r}{\partial g_c / \partial r} \quad (69)$$

$$= \frac{A + B}{C}$$

where

$$A = [\cos(2\theta_r) \cos^2 \theta_r - 1](1 - 2\pi r^3/M) \quad (70)$$

$$B = r\partial\theta_r/\partial r [\sin(2\theta_r) \cos^2 \theta_r + \tan \theta_r] \quad (71)$$

$$C = 1 - \tan \theta_r r\partial\theta_r/\partial r - 2\pi r^3/M \quad (72)$$

For small  $\theta_r$  the parameter  $D$  can be written as<sup>29</sup>

$$D = \div 3\theta_r^2(1 - \eta) \quad (73)$$

$$\eta = \frac{r/\theta_r \partial\theta_r/\partial r}{1 - 2\pi r^3/M} \quad (74)$$

From equation (64) it follows that for spherical bodies the Newtonian law of gravitation in space with broken internal symmetry requires a coordinate dependent effective gravitational constant given by

$$G_r = G \cos(2\theta_r) \cos^2 \theta_r \quad (75)$$

For small values of  $\theta_r$  equation (75) becomes

$$G_r = G(1 - 3\theta_r^2 + \dots) \quad (76)$$

Therefore for space with broken internal symmetry  $G_r < G$ , and  $G$  represents the ideal case of gravitation in totally symmetric space ( $\theta_r = 0$ ). For a homogeneous spherical planet or star the internal phase angle of the radial coordinate will be a function of the radial coordinate magnitude  $\theta_r = \theta_r(r)$  and is obtained as a solution to the coupled gravitational equilibrium (56) and (57). In general however  $\theta_r = \theta_r(r, \psi, \phi)$  for an inhomogeneous body such as the earth, and therefore  $G_r$  will depend on latitude and longitude. Equations (64) and (75) are valid for both the interior and exterior of a spherical planet.

From equation (75) it follows that

$$\partial G_r / \partial r = -GE\partial\theta_r / \partial r \quad (77)$$

$$\partial G_r / \partial \psi = -GE\partial\theta_r / \partial \psi \quad (78)$$

$$\partial G_r / \partial \phi = -GE\partial\theta_r / \partial \phi \quad (79)$$

where

$$E = \sin(2\theta_r)(4 \cos^2 \theta_r - 1) \quad (80)$$

For small  $\theta_r$ ,  $E \sim 6\theta_r$ . In Section 5 it is shown that  $\theta_r < 0$  and  $\partial\theta_r / \partial r > 0$  for idealized planets so that  $\partial G_r / \partial r > 0$ . From equations (77) through (79) it follows that

$$r/G_r \partial G_r / \partial r = -2Hr\partial\theta_r / \partial r \quad (81)$$

$$\psi/G_r \partial G_r / \partial \psi = -2H\psi\partial\theta_r / \partial \psi \quad (82)$$

$$\phi/G_r \partial G_r / \partial \phi = -2H\phi\partial\theta_r / \partial \phi$$

where

$$H = \tan(2\theta_r) + \tan \theta_r \quad (84)$$

For small  $\theta_r$ ,  $H \sim 3\theta_r$ .

It is the variation of the acceleration of gravity of the earth that is determined in gravity measurements, and it is important to have a measure of the difference between the rates of change of  $g_m$  and  $g_c$  with respect to radial distance. One such measure is given in equation (69). Another measure might consider the difference of the normalized rates of change of the acceleration of gravity as follows

$$D_2 = \frac{r/g_m \partial g_m / \partial r - r/g_c \partial g_c / \partial r}{r/g_c \partial g_c / \partial r} \quad (85)$$

$$= \frac{r_m/g_m \partial g_m / \partial r_m - r_m/g_c \partial g_c / \partial r_m}{r_m/g_c \partial g_c / \partial r_m}$$

From equation (63) it follows that

$$r/g_m \partial g_m / \partial r = -2[1 + \tan(2\theta_r) r \partial \theta_r / \partial r - 2\pi r^3 \rho / M] \quad (86)$$

while from equation (65) it follows that

$$r/g_c \partial g_c / \partial r = -2[1 - \tan \theta_r r \partial \theta_r / \partial r - 2\pi r^3 \rho / M] \quad (87)$$

Then

$$D_2 = \frac{H r \partial \theta_r / \partial r}{1 - \tan \theta_r r \partial \theta_r / \partial r - 2\pi r^3 \rho / M} \quad (88)$$

where H is given by equation (84). For small values of  $\theta_r$

$$D_2 \sim + 3\theta_r^2 \eta \quad (89)$$

where  $\eta$  is given by equation (74).

3. MEASUREMENT AND GEOMETRY OF SPACE AND TIME. It has been assumed that the complex number space and time coordinates are Euclidian and that<sup>29</sup>

$$\bar{x}^2 + \bar{y}^2 = \bar{r}^2 \quad (90)$$

$$\sin^2 \bar{\phi} + \cos^2 \bar{\phi} = 1 \quad (91)$$

$$\tan \bar{\phi} = \bar{y} / \bar{x} \quad (92)$$

where

$$\bar{x} = x e^{j\theta_x} \quad (93)$$

$$\bar{y} = y e^{j\theta_y} \quad (94)$$

$$\bar{\phi} = \phi e^{j\theta_\phi} \quad (95)$$

$$\sin \bar{\phi} = S_\phi e^{j\theta_{s\phi}} \quad (96)$$

$$\cos \bar{\phi} = C_{\phi} e^{-j\theta_{c\phi}} \quad (97)$$

where  $\bar{\phi}$  = complex number azimuthal angle, and where<sup>28</sup>

$$S_{\phi} = [\sin^2(\phi \cos \theta_{\phi}) + \sinh^2(\phi \sin \theta_{\phi})]^{1/2} \quad (98)$$

$$C_{\phi} = [\cos^2(\phi \cos \theta_{\phi}) + \sinh^2(\phi \sin \theta_{\phi})]^{1/2} \quad (99)$$

$$\tan \theta_{s\phi} = \cot(\phi \cos \theta_{\phi}) \tanh(\phi \sin \theta_{\phi}) \quad (100)$$

$$\tan \theta_{c\phi} = \tan(\phi \cos \theta_{\phi}) \tanh(\phi \sin \theta_{\phi}) \quad (101)$$

The component equations of equation (90) determine  $r$  and  $\theta_r$  and are written as

$$x^2 \cos(2\theta_x) + y^2 \cos(2\theta_y) = r^2 \cos(2\theta_r) \quad (102)$$

$$x^2 \sin(2\theta_x) + y^2 \sin(2\theta_y) = r^2 \sin(2\theta_r) \quad (103)$$

while the component equations of equation (91) are

$$S_{\phi}^2 \cos(2\theta_{s\phi}) + C_{\phi}^2 \cos(2\theta_{c\phi}) = 1 \quad (104)$$

$$S_{\phi}^2 \sin(2\theta_{s\phi}) - C_{\phi}^2 \sin(2\theta_{c\phi}) = 0 \quad (105)$$

Equations (90) through (105) also give

$$S_{\phi}/C_{\phi} = y/x \quad (106)$$

$$\theta_{s\phi} + \theta_{c\phi} = \theta_y - \theta_x \quad (107)$$

$$S_{\phi}^2 = \sin(2\theta_{c\phi})/\sin[2(\theta_{c\phi} + \theta_{s\phi})] \quad (108)$$

$$C_{\phi}^2 = \sin(2\theta_{s\phi})/\sin[2(\theta_{c\phi} + \theta_{s\phi})] \quad (109)$$

The measured coordinates and angles are given by<sup>28</sup>

$$x_m = x \cos \theta_x \quad (110)$$

$$y_m = y \cos \theta_y \quad (111)$$

$$r_m = r \cos \theta_r \quad (112)$$

$$\phi_m = \phi \cos \theta_{\phi} \quad (113)$$

Substituting equations (110) through (112) into equations (102) and (103) shows that

$$x_m^2 + y_m^2 \neq r_m^2 \quad (114)$$

which indicates that in a broken symmetry system the measured coordinates are non-Euclidian.

#### A. Length of Curves

The length of a curve in complex number space is given by

$$\bar{L} = L e^{j\theta_L} = \int [\bar{r}^2 + (d\bar{r}/d\bar{\phi})^2]^{1/2} d\bar{\phi} \quad (115)$$

where

$$\frac{d\bar{r}}{d\bar{\phi}} = e^{j\phi_{r\phi}} \frac{\sec \beta_{r,r} dr}{\sec \beta_{\phi,\phi} d\phi} \quad (116)$$

$$\phi_{r\phi} = \theta_r + \beta_{r,r} - \theta_\phi - \beta_{\phi,\phi} \quad (117)$$

$$\sec \beta_{r,r} = [1 + (r\partial\theta_r/\partial r)^2]^{1/2} \quad (118)$$

$$\sec \beta_{\phi,\phi} = [1 + (\phi\partial\theta_\phi/\partial\phi)^2]^{1/2} \quad (119)$$

The measured length is given by

$$L_m = L \cos \theta_L \quad (120)$$

For a circle with  $\partial r/\partial\phi = 0$  the complex number length is

$$\bar{L} = r \int_0^{2\pi} e^{j(\theta_r + \theta_\phi + \beta_{\phi,\phi})} \sec \beta_{\phi,\phi} d\phi \quad (121)$$

If in addition  $\theta_\phi$  and  $\theta_r$  are independent of  $\phi$  equation (121) becomes

$$\bar{L} = 2\pi r e^{j(\theta_r + \theta_\phi)} \quad (122)$$

$$L = 2\pi r \quad (123)$$

$$\theta_L = \theta_r + \theta_\phi \quad (124)$$

The measured circumference is

$$L_m = L \cos \theta_L = 2\pi r \cos(\theta_r + \theta_\phi) \quad (125)$$

Note that if  $\theta_\phi = 0$

$$L = 2\pi r \quad (126)$$

$$\theta_L = \theta_r \quad (127)$$

$$L_m = 2\pi r \cos \theta_r = 2\pi r_m \quad (128)$$

which is the result obtained in Reference 28.

#### B. Area of a Plane Curve.

The area enclosed by a plane curve in complex number space is

$$\begin{aligned} \bar{A} &= \frac{1}{2} \int \bar{r}^2 d\bar{\phi} = \frac{1}{2} \int r^2 e^{j(2\theta_r + \theta_\phi)} (d\phi + j\phi d\theta_\phi) \\ &= \frac{1}{2} \int r^2 e^{j(2\theta_r + \theta_\phi + \beta_{\phi, \phi})} \sec \beta_{\phi, \phi} d\phi \end{aligned} \quad (129)$$

For a circle with  $r = \text{constant}$

$$\bar{A} = \frac{r^2}{2} \int e^{j(2\theta_r + \theta_\phi + \beta_{\phi, \phi})} \sec \beta_{\phi, \phi} d\phi \quad (130)$$

If  $\theta_r$  and  $\theta_\phi$  are constants

$$\bar{A} = A e^{j\theta_A} = \pi r^2 e^{j(2\theta_r + \theta_\phi)} \quad (131)$$

and therefore

$$A = \pi r^2 \quad (132)$$

$$\theta_A = 2\theta_r + \theta_\phi \quad (133)$$

$$A_m = \pi r^2 \cos(2\theta_r + \theta_\phi) = \pi r_m^2 \cos(2\theta_r + \theta_\phi) / \cos^2 \theta_r \quad (134)$$

where  $A_m = \text{measured area}$ . Finally, if  $\theta_\phi = 0$

$$A = \pi r^2 \quad (135)$$

$$\theta_A = 2\theta_r \quad (136)$$

$$A_m = \pi r^2 \cos(2\theta_r) = \pi r_m^2 \cos(2\theta_r) / \cos^2 \theta_r \quad (137)$$

which is the result obtained in Reference 28.

C. Area of Surface.

The area of a surface is given in complex number spherical coordinates as

$$\begin{aligned}\bar{S} &= \int \bar{r}^2 \sin \bar{\psi} d\bar{\psi} d\bar{\phi} & (138) \\ &= \int r^2 S_\psi e^{j(2\theta_r + \theta_{s\psi} + \theta_\psi + \theta_\phi)} (d\psi + j\psi d\theta_\psi) (d\phi + j\phi d\theta_\phi) \\ &= \int r^2 S_\psi \sec \beta_{\psi,\psi} \sec \beta_{\phi,\phi} e^{j\phi_S} d\psi d\phi\end{aligned}$$

where  $\bar{\psi}$  = complex number zenith angle given by

$$\bar{\psi} = \psi e^{j\theta_\psi} \quad (139)$$

and where

$$\sec \beta_{\psi,\psi} = [1 + (\psi \partial \theta_\psi / \partial \psi)^2]^{1/2} \quad (140)$$

$$\sec \beta_{\phi,\phi} = [1 + (\phi \partial \theta_\phi / \partial \phi)^2]^{1/2} \quad (141)$$

$$\phi_S = 2\theta_r + \theta_{s\psi} + \theta_\psi + \beta_{\psi,\psi} + \theta_\phi + \beta_{\phi,\phi} \quad (142)$$

$$S_\psi = [\sin^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)]^{1/2} \quad (143)$$

$$\tan \theta_{s\psi} = \cot(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \quad (144)$$

If  $\theta_\psi$  and  $\theta_\phi$  are constants then

$$\bar{S} = e^{j(\theta_{s\psi} + \theta_\psi + \theta_\phi)} \int r^2 S_\psi e^{j2\theta_r} d\psi d\phi \quad (145)$$

For  $r$  and  $\theta_r$  independent of angles

$$\bar{S} = e^{j(2\theta_r + \theta_{s\psi} + \theta_\psi + \theta_\phi)} 2\pi r^2 \int S_\psi d\psi \quad (146)$$

Directly from equation (138) it follows that if  $r$ ,  $\theta_r$ ,  $\theta_\psi$  and  $\theta_\phi$  are independent of angles (a crude approximation)

$$\bar{S} \sim 2\pi r^2 e^{j(2\theta_r + \theta_\phi)} [1 - \cos(\pi e^{j\theta_\psi})] \quad (147)$$

It is easy to show that equations (146) and (147) are equivalent. If  $\theta_\psi = 0$  it follows that

$$\bar{S} = 4\pi r^2 e^{j(2\theta_r + \theta_\phi)} \quad (148)$$

Finally, if  $\theta_\phi = 0$

$$\bar{S} = 4\pi r^2 e^{j2\theta_r} \quad (149)$$

$$S = 4\pi r^2 \quad (150)$$

$$\theta_S = 2\theta_r \quad (151)$$

$$S_m = 4\pi r^2 \cos(2\theta_r) \quad (152)$$

$$= 4\pi r_m^2 \cos(2\theta_r) / \cos^2 \theta_r$$

From equations (62) and (147) it follows that Gauss's law for spatial coordinates with broken internal symmetry is given by

$$\int \bar{g} d\bar{S} \sim -2\pi G M e^{j\theta_\phi} [1 - \cos(\pi e^{j\theta_\psi})] \quad (153)$$

assuming  $\theta_\phi$  and  $\theta_\psi$  are constants (which can only be a crude approximation).

#### D. Volume

The volume contained within a closed surface is written in complex number spherical polar coordinates as

$$\bar{V} = \int \bar{r}^2 \sin \bar{\psi} d\bar{\psi} d\bar{\phi} d\bar{r} \quad (154)$$

$$= \int r^2 S_\psi \sec \beta_{\psi,\psi} \sec \beta_{\phi,\phi} \sec \beta_{r,r} e^{j\phi_V} d\psi d\phi dr$$

where

$$\phi_V = 3\theta_r + \beta_{r,r} + \theta_{s\psi} + \theta_\psi + \beta_{\psi,\psi} + \theta_\phi + \beta_{\phi,\phi} \quad (155)$$

If  $\theta_\psi$  and  $\theta_\phi$  are constants

$$\bar{V} = e^{j(\theta_{s\psi} + \theta_\psi + \theta_\phi)} \int r^2 S_\psi \sec \beta_{r,r} e^{j(3\theta_r + \beta_{r,r})} d\psi d\phi dr \quad (156)$$

For  $\theta_\psi$  and  $\theta_\phi = 0$

$$\bar{V} = \int r^2 \sin \psi \sec \beta_{r,r} e^{j(3\theta_r + \beta_{r,r})} d\psi d\phi dr \quad (157)$$

For  $r$  and  $\theta_r$  independent of  $\psi$  and  $\phi$  (a sphere)

$$\bar{V} = 4\pi \int r^2 \sec \beta_{r,r} e^{j(3\theta_r + \beta_{r,r})} dr \quad (158)$$



The component parts of equation (158) are

$$\dot{V} \cos \theta_V = 4\pi \int r^2 \sec \beta_{r,r} \cos(3\theta_r + \beta_{r,r}) dr \quad (159)$$

$$V \sin \theta_V = 4\pi \int r^2 \sec \beta_{r,r} \sin(3\theta_r + \beta_{r,r}) dr \quad (160)$$

which determines  $V$  and  $\theta_V$  for a sphere in a gravitational field. For  $\theta_r = \text{constant}$

$$\bar{V} = e^{j3\theta_r} 4\pi/3r^3 = 4\pi/3\bar{r}^3 \quad (161)$$

$$V = 4\pi/3r^3 \quad (162)$$

$$\theta_V = 3\theta_r \quad (163)$$

$$V_m = 4\pi/3r^3 \cos(3\theta_r) = 4\pi/3r_m^3 \cos(3\theta_r)/\cos^3 \theta_r \quad (164)$$

#### E. Density

The rest mass of a body does not have an internal phase because it is invariant under the effects of the basic trace equation (2).<sup>29</sup> The instantaneous density is given by

$$\bar{\rho} = \rho e^{j\theta_\rho} = dM/d\bar{V} = \cos \beta_{V,V} dM/dV e^{-j(\theta_V + \beta_{V,V})} \quad (165)$$

where

$$\tan \beta_{V,V} = V d\theta_V/dV \quad (166)$$

Therefore

$$\rho = \cos \beta_{V,V} dM/dV \quad (167)$$

$$\theta_\rho = -\theta_V - \beta_{V,V} \quad (168)$$

Combining equations (165) and (154) gives the following results for the density

$$\rho = (r^2 S_\psi \sec \beta_{\psi,\psi} \sec \beta_{\phi,\phi} \sec \beta_{r,r})^{-1} \frac{dM}{d\psi d\phi dr} \quad (169)$$

$$\theta_\rho = -\phi_V \quad (170)$$

where  $\phi_V$  is given by equation (155). For radial symmetry equations (158) and (165) give

$$\rho \sim \frac{\cos \beta_{r,r}}{4\pi r^2} \frac{\partial M}{\partial r} \quad (171)$$

$$\theta_\rho \sim -3\theta_r - \beta_{r,r} \quad (172)$$

Equation (171) combined with the following stellar equilibrium equation for broken symmetry matter<sup>28</sup>

$$\begin{aligned} M &= -\frac{r^2 \partial P / \partial r}{G\rho} \cos \beta_{r,r} \left[ 1 + \left( P \frac{\partial \theta_P / \partial r}{\partial P / \partial r} \right)^2 \right]^{1/2} \\ &= -\frac{r^2 \partial P / \partial r}{G\rho} \cos \beta_{r,r} \sec \beta_{P,r} \end{aligned} \quad (173)$$

gives the combined stellar equilibrium (56) for ordinary stars. The angle  $\beta_{P,r}$  that appears in equation (173) is defined by

$$\tan \beta_{P,r} = P \frac{\partial \theta_P / \partial r}{\partial P / \partial r} \quad (174)$$

The measured density is given by

$$\rho_m = \rho \cos \theta_\rho \quad (175)$$

For a relativistic interacting system having a complex number internal energy  $\bar{U}$ , the mass is given by  $\bar{M} = \bar{U}/c^2$  and the instantaneous density is<sup>27</sup>

$$\bar{\rho}_r = c^{-2} d\bar{U}/|d\bar{V}| = d\bar{M}/|d\bar{V}| \quad (176)$$

Combining equations (158) and (176) gives

$$\rho_r \sim \frac{\cos \beta_{r,r}}{4\pi r^2} \left[ (\partial M / \partial r)^2 + (M \partial \theta_M / \partial r)^2 \right]^{1/2} \quad (177)$$

$$\theta_{\rho r} \sim \theta_U + \beta_{U,r} \quad (178)$$

where  $\theta_M = \theta_U$ , and  $\beta_{U,r}$  is given by

$$\tan \beta_{U,r} = U \frac{\partial \theta_U}{\partial r} / \frac{\partial U}{\partial r} = M \frac{\partial \theta_M}{\partial r} / \frac{\partial M}{\partial r} \quad (179)$$

If special relativity is included the pressure adds to the internal energy density, and the inertial mass density becomes<sup>31-33</sup>

$$\bar{\rho}_I = \bar{\rho}_r + \bar{P}/c^2 \quad (180)$$

while the gravitational mass density is

$$\bar{\rho}_G = \bar{\rho}_r + 3\bar{P}/c^2 \quad (181)$$

Equations (180) and (181) can be simplified by combining them with equation (158). It should be pointed out that the complex number values for coordinates is also suggested by string theory.<sup>34</sup>

#### 4. NEWTONIAN GRAVITY FOR NONSPHERICAL MASSES WITH BROKEN INTERNAL SYMMETRY.

##### A. Complex Number Gravitational Potential.

By analogy to the standard scalar form of the gravitational potential for a nonspherical body, the following expression for the complex number gravitational potential for a nonspherical body existing in space with broken internal symmetries is postulated<sup>35-39</sup>

$$\bar{V} = -GM/\bar{r} \left[ 1 - \sum_{n=2}^{\infty} \bar{I}_n (\bar{a}/\bar{r})^n \bar{P}_n(\cos \bar{\psi}) \right] \quad (182)$$

where  $\bar{V}$  = complex number potential,  $\bar{a}$  = complex number equatorial radius,  $\bar{r}$  = complex number radial coordinate of a point outside of the body,  $\bar{I}_n$  = complex number coefficients, and  $\bar{P}_n(\cos \bar{\psi})$  = complex number Legendre polynomials corresponding to the complex number zenith angle  $\bar{\psi}$ . The complex number quantities appearing in equation (182) can be written as

$$\bar{V} = V e^{j\theta_V} \quad (183)$$

$$\bar{a} = a e^{j\theta_a} \quad (184)$$

$$\bar{I}_n = I_n e^{j\theta_{I_n}} \quad (185)$$

$$\bar{P}_n = P_n e^{j\theta_{P_n}} \quad (186)$$

where, for instance,  $P_n$  and  $\theta_{P_n}$  = magnitude and phase angle of the complex number Legendre polynomials. The real and imaginary parts of equations (182) are given by

$$V \cos \theta_V = -\frac{GM}{r} \left[ \cos \theta_r - I_2 P_2 (a/r)^2 \cos(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 3\theta_r) - \dots \right] \quad (187)$$

$$V \sin \theta_V = \frac{GM}{r} \left[ \sin \theta_r + I_2 P_2 (a/r)^2 \sin(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 3\theta_r) + \dots \right] \quad (188)$$

Equations (187) and (188) can be used to determine  $V$  and  $\theta_V$ . For instance

$$\tan \theta_V = -\frac{\sin \theta_r + I_2 P_2 (a/r)^2 \sin(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 3\theta_r) + \dots}{\cos \theta_r - I_2 P_2 (a/r)^2 \cos(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 3\theta_r) - \dots} \quad (189)$$

while squaring and adding equations (187) and (188) gives  $V^2$ . In the limit  $a/r \rightarrow 0$  equations (187) and (188) become

$$V_m = V_R = V \cos \theta_V = -GM/r \cos \theta_r = -GM/r_m \cos^2 \theta_r \quad (190)$$

$$V_I = V \sin \theta_V = GM/r \sin \theta_r = GM/r_m \sin \theta_r \cos \theta_r \quad (191)$$

which corresponds to a point mass whose complex number potential is given by

$$\bar{V} = -GM/\bar{r} \quad (192)$$

The various terms in the gravitational potential will now be considered.

### B. Complex Number Legendre Polynomials.

The appearance of  $\bar{P}_n(\cos \bar{\psi})$  in equation (182) needs some explanation. Following the standard prescription for obtaining Legendre polynomials for scalar angles, the following generalizations to complex angles are given<sup>40</sup>

$$\bar{P}_0 = 1 \quad (193)$$

$$\bar{P}_1 = \cos \bar{\psi} \quad (194)$$

$$\bar{P}_2 = 1/2(3 \cos^2 \bar{\psi} - 1) \quad (195)$$

$$\bar{P}_3 = 1/2(5 \cos^3 \bar{\psi} - 3 \cos \bar{\psi}) \quad (196)$$

$$\bar{P}_4 = 1/8(35 \cos^4 \bar{\psi} - 30 \cos^2 \bar{\psi} + 3) \quad (197)$$

$$\bar{P}_5 = 1/8(63 \cos^5 \bar{\psi} - 70 \cos^3 \bar{\psi} + 15 \cos \bar{\psi}) \quad (198)$$

$$\bar{P}_6 = 1/16(231 \cos^6 \bar{\psi} - 315 \cos^4 \bar{\psi} + 105 \cos^2 \bar{\psi} - 5) \quad (199)$$

where<sup>28</sup>

$$\bar{\psi} = \psi e^{j\theta_\psi} \quad (200)$$

$$\cos \bar{\psi} = C_\psi e^{-j\theta_{c\psi}} \quad (201)$$

$$C_\psi = [\cos^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)]^{1/2} \quad (202)$$

$$\tan \theta_{c\psi} = \tan(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \quad (203)$$

From equations (186) and (194) it follows that

$$P_1 = C_\psi \quad (204)$$

$$\theta_{P1} = -\theta_{c\psi} \quad (205)$$

Equations (186) and (195) give

$$P_2 \cos \theta_{P2} = 1/2[3C_\psi^2 \cos(2\theta_{c\psi}) - 1] \quad (206)$$

$$P_2 \sin \theta_{P2} = -3/2 C_\psi^2 \sin(2\theta_{c\psi}) \quad (207)$$

which gives

$$P_2 = 1/2[9C_\psi^4 - 6C_\psi^2 \cos(2\theta_{c\psi}) + 1]^{1/2} \quad (208)$$

$$\tan \theta_{P2} = \frac{-3C_\psi^2 \sin(2\theta_{c\psi})}{3C_\psi^2 \cos(2\theta_{c\psi}) - 1} \quad (209)$$

From equation (186) and (196) it follows that

$$P_3 \cos \theta_{P3} = 1/2[5C_\psi^3 \cos(3\theta_{c\psi}) - 3C_\psi \cos \theta_{c\psi}] \quad (210)$$

$$P_3 \sin \theta_{P3} = 1/2[-5C_\psi^3 \sin(3\theta_{c\psi}) + 3C_\psi \sin \theta_{c\psi}] \quad (211)$$

which gives

$$P_3 = 1/2[25C_\psi^6 - 30C_\psi^4 \cos(2\theta_{c\psi}) + 9C_\psi^2]^{1/2} \quad (212)$$

$$\tan \theta_{P3} = \frac{-5C_\psi^3 \sin(3\theta_{c\psi}) + 3C_\psi \sin \theta_{c\psi}}{5C_\psi^3 \cos(3\theta_{c\psi}) - 3C_\psi \cos \theta_{c\psi}} \quad (213)$$

and so on for  $P_n$  and  $\theta_{Pn}$ .

The complex number Legendre polynomials can also be written in terms of the complementary angle  $\bar{\chi}$  which is defined by

$$\sin \bar{\chi} = \cos \bar{\psi} \quad (214)$$

$$\cos \bar{\chi} = \sin \bar{\psi} \quad (215)$$

where  $\bar{\chi}$  = complex number latitude and<sup>28</sup>

$$\sin \bar{\chi} = S_\chi e^{j\theta_s \chi} \quad (216)$$

$$\cos \bar{\chi} = C_\chi e^{-j\theta_c \chi} \quad (217)$$

and where

$$S_X = [\sin^2(\chi \cos \theta_X) + \sinh^2(\chi \sin \theta_X)]^{1/2} \quad (218)$$

$$C_X = [\cos^2(\chi \cos \theta_X) + \sinh^2(\chi \sin \theta_X)]^{1/2} \quad (219)$$

$$\tan \theta_{sX} = \cot(\chi \cos \theta_X) \tanh(\chi \sin \theta_X) \quad (220)$$

$$\tan \theta_{cX} = \tan(\chi \cos \theta_X) \tanh(\chi \sin \theta_X) \quad (221)$$

The defining relations given by equations (214) and (215) can be written as

$$S_X = C_\psi \quad (222)$$

$$C_X = S_\psi \quad (223)$$

$$\theta_{sX} = -\theta_{c\psi} \quad (224)$$

$$\theta_{s\psi} = -\theta_{cX} \quad (225)$$

which relate  $\chi$  and  $\theta_X$  to  $\psi$  and  $\theta_\psi$ .

Combining equation (214) with equations (193) through (199) gives

$$\bar{P}_0 = 1 \quad (226)$$

$$\bar{P}_1 = \sin \bar{\chi} \quad (227)$$

$$\bar{P}_2 = 1/2(3 \sin^2 \bar{\chi} - 1) \quad (228)$$

$$\bar{P}_3 = 1/2(5 \sin^3 \bar{\chi} - 3 \sin \bar{\chi}) \quad (229)$$

$$\bar{P}_4 = 1/8(35 \sin^4 \bar{\chi} - 30 \sin^2 \bar{\chi} + 3) \quad (230)$$

$$\bar{P}_5 = 1/8(63 \sin^5 \bar{\chi} - 70 \sin^3 \bar{\chi} + 15 \sin \bar{\chi}) \quad (231)$$

$$\bar{P}_6 = 1/16(231 \sin^6 \bar{\chi} - 315 \sin^4 \bar{\chi} + 105 \sin^2 \bar{\chi} - 5) \quad (232)$$

Note also that

$$P_1 = S_X \quad (233)$$

$$\theta_{P1} = \theta_{sX} \quad (234)$$

$$P_2 = 1/2[9S_X^4 - 6S_X^2 \cos(2\theta_{sX}) + 1]^{1/2} \quad (235)$$

$$\tan \theta_{P2} = \frac{3S_X^2 \sin(2\theta_{sX})}{3S_X^2 \cos(2\theta_{sX}) - 1} \quad (236)$$

$$P_3 = 1/2[25S_X^6 - 30S_X^4 \cos(2\theta_{sX}) + 9S_X^2]^{1/2} \quad (237)$$

$$\tan \theta_{P3} = \frac{5S_X^3 \sin(3\theta_{sX}) - 3S_X \sin \theta_{sX}}{5S_X^3 \cos(3\theta_{sX}) - 3S_X \cos \theta_{sX}} \quad (238)$$

Consider now the special case of the equator and the north pole. At the equator equations (193) through (238) give

$$\chi = 0 \quad \psi = \pi/2 \quad \theta_{\psi}(\pi/2) = 0 \quad (239)$$

$$\theta_{cX}(0) = 0 \quad \theta_{sX}(0) = \theta_X(0) \quad \theta_{s\psi}(\pi/2) = 0 \quad (240)$$

$$\theta_{c\psi}(\pi/2) = -\theta_X(0) \quad (241)$$

$$S_X(0) = 0 \quad C_X(0) = 1 \quad (242)$$

$$S_{\psi}(\pi/2) = 1 \quad C_{\psi}(\pi/2) = 0 \quad (243)$$

$$P_o = 1 \quad \theta_{Po} = 0 \quad \bar{P}_o = 1$$

$$P_1 = 0 \quad \theta_{P1} = \theta_X(0) \quad \bar{P}_1 = 0$$

$$P_2 = 1/2 \quad \theta_{P2} = \pi \quad \bar{P}_2 = -1/2 \quad (244)$$

$$P_3 = 0 \quad \theta_{P3} = \theta_X(0) \quad \bar{P}_3 = 0$$

$$P_4 = 3/8 \quad \theta_{P4} = 0 \quad \bar{P}_4 = 3/8$$

where  $\theta_X(0)$  = value of  $\theta_X$  at the equator.

At the north pole the following relationships are obtained from equations (193) through (238)

$$\chi = \pi/2 \quad \psi = 0 \quad \theta_X(\pi/2) = 0 \quad (245)$$

$$\theta_{c\psi}(0) = 0 \quad \theta_{s\psi}(0) = \theta_{\psi}(0) \quad \theta_{sX}(\pi/2) = 0 \quad (246)$$

$$\theta_{cX}(\pi/2) = -\theta_{\psi}(0) \quad (247)$$

$$S_{\psi}(0) = 0 \quad C_{\psi}(0) = 1 \quad (248)$$

$$S_X(\pi/2) = 1 \quad C_X(\pi/2) = 0 \quad (249)$$

$$\begin{array}{lll}
P_0 = 1 & \theta_{P0} = 0 & \bar{P}_0 = 1 \\
P_1 = 1 & \theta_{P1} = 0 & \bar{P}_1 = 1 \\
P_2 = 1 & \theta_{P2} = 0 & \bar{P}_2 = 1 \\
P_3 = 1 & \theta_{P3} = 0 & \bar{P}_3 = 1 \\
P_4 = 1 & \theta_{P4} = 0 & \bar{P}_4 = 1
\end{array} \tag{250}$$

where  $\theta_\psi(0)$  = value of  $\theta_\psi$  at the north pole.

Equations (214) and (215) are valid for complimentary angles. Combining these conditions with equations (218) through (221) shows that if  $\chi = 0$  then  $\psi = \pi/2$  and if  $\psi = 0$  then  $\chi = \pi/2$  so that

$$\theta_\psi(\pi/2) = 0 \tag{251}$$

$$\theta_\chi(\pi/2) = 0 \tag{252}$$

as indicated in equations (239) and (245). This shows that for complementary angles

$$\bar{\psi} + \bar{\chi} = \pi/2 \tag{253}$$

and the right angle  $\pi/2$  is not associated with an internal phase angle. The component parts of equation (253) are given by

$$\psi \cos \theta_\psi + \chi \cos \theta_\chi = \pi/2 \tag{254}$$

$$\psi \sin \theta_\psi + \chi \sin \theta_\chi = 0 \tag{255}$$

Equation (254) states that

$$\psi_m + \chi_m = \pi/2 \tag{256}$$

which states that the sum of the measured complementary angles is equal to  $\pi/2$ .

The angle  $\pi/2$  apparently is the only angle whose internal phase angle is zero, all other angles exhibit an internal phase. Consider an angle  $\bar{\psi}$  which is composed of two component parts  $\bar{\psi}_1$  and  $\bar{\psi}_2$  so that

$$\bar{\psi} = \bar{\psi}_1 + \bar{\psi}_2 \tag{257}$$

$$\psi \cos \theta_\psi = \psi_1 \cos \theta_{\psi_1} + \psi_2 \cos \theta_{\psi_2} \tag{258}$$

$$\psi \sin \theta_\psi = \psi_1 \sin \theta_{\psi_1} + \psi_2 \sin \theta_{\psi_2} \tag{259}$$



Equation (258) states that

$$\psi_m = \psi_{1m} + \psi_{2m} \quad (260)$$

which agrees with reality in that the measured total angle is the sum of each measured part. For the special case of  $\pi/2$ , equations (254) and (255) give

$$\psi = \pi/2 \sin \theta_\chi / (\cos \theta_\psi \sin \theta_\chi - \sin \theta_\psi \cos \theta_\chi) \quad (261)$$

$$\chi = -\pi/2 \sin \theta_\psi / (\cos \theta_\psi \sin \theta_\chi - \sin \theta_\psi \cos \theta_\chi) \quad (262)$$

Alternatively equations (254) and (255) can be written as

$$\cos \theta_\psi = (1 - \chi^2/\psi^2 \sin^2 \theta_\chi)^{1/2} \quad (263)$$

$$\psi(1 - \chi^2/\psi^2 \sin^2 \theta_\chi)^{1/2} + \chi \cos \theta_\chi = \pi/2 \quad (264)$$

Figure 1 shows the variation of  $\theta_\psi$  and  $S_\psi$ .

### C. Complex Number Gravity Potential Coefficients $\bar{I}_n$ .

The dimensionless complex number coefficients  $\bar{I}_n$  that appear in equation (182) describe the distribution of mass within a planet. Equation (182) shows that  $\bar{I}_0 = 1$  and  $\bar{I}_1 = 0$  because the origin of coordinates can be located at the center of mass of the planet. For practical calculations only the second order coefficient  $\bar{I}_2$  is retained. The value of  $\bar{I}_2$  is obtained as an obvious generalization of the standard scalar form for this coefficient as follows<sup>35</sup>

$$\bar{I}_2 = I_2 e^{j\theta_{I2}} = (\bar{C} - \bar{A}) / (Ma^2) \quad (265)$$

where  $\bar{C}$  = complex number moment of inertia about the polar axis, and  $\bar{A}$  = complex number moment of inertia about one of the transverse axes. If the z axis is taken to be the polar axis<sup>35</sup>

$$\bar{C} = Ce^{j\theta_C} = \int (\bar{x}^2 + \bar{y}^2) dM \quad (266)$$

$$\bar{A} = Ae^{j\theta_A} = \int (\bar{x}^2 + \bar{z}^2) dM = \int (\bar{y}^2 + \bar{z}^2) dM \quad (267)$$

The real and imaginary parts of equation (265) are

$$I_2 \cos \theta_{I2} = 1/(Ma^2) [C \cos(\theta_C - 2\theta_a) - A \cos(\theta_A - 2\theta_a)] \quad (268)$$

$$I_2 \sin \theta_{I2} = 1/(Ma^2) [C \sin(\theta_C - 2\theta_a) - A \sin(\theta_A - 2\theta_a)] \quad (269)$$

From equations (172) and (173) it follows that

$$I_2^2 = 1/(M^2 a^4) [C^2 + A^2 - 2AC \cos(\theta_C - \theta_A)] \quad (270)$$

$$\tan \theta_{I_2} = \frac{C \sin(\theta_C - 2\theta_a) - A \sin(\theta_A - 2\theta_a)}{C \cos(\theta_C - 2\theta_a) - A \cos(\theta_A - 2\theta_a)} \quad (271)$$

The measured values of  $I_2$ ,  $C$  and  $A$  are given by

$$I_{2m} = I_2 \cos \theta_{I_2} \quad (272)$$

$$C_m = C \cos \theta_C \quad (273)$$

$$A_m = A \cos \theta_A \quad (274)$$

#### D. Rotational Effects of a Gravitating Planet for Space and Time with Broken Internal Symmetries.

For a rotating planet (or star) the total potential also includes a rotational term, and is written as an obvious generalization to the standard scalar form as follows<sup>35-39</sup>

$$\bar{U} = -GM/\bar{r} + GM/\bar{r} \bar{I}_2 \bar{P}_2(\bar{a}/\bar{r})^2 - 1/2 \bar{r}^2 \bar{\omega}^2 \cos^2 \bar{\chi} \quad (275)$$

where  $\bar{U}$  = total complex number potential and  $\bar{\omega}$  = complex number angular speed. The complex number angular speed has already been considered in mechanical problems.<sup>28</sup> Combining equation (275) with equations (216), (217) and (228) gives

$$\bar{U} = -GM/\bar{r} + GM/\bar{r} \bar{I}_2 (\bar{a}/\bar{r})^2 - 1/2 (3S_\chi^2 e^{2j\theta_{S\chi}} - 1) - 1/2 \bar{r}^2 \bar{\omega}^2 C_\chi^2 e^{-2j\theta_{C\chi}} \quad (276)$$

The total potential can be evaluated at the equator  $\chi = 0$  and at the north pole  $\chi = \pi/2$  as follows

$$\bar{U}_0 = -GM/\bar{a} - GM\bar{I}_2/(2\bar{a}) - 1/2 \bar{a}^2 \bar{\omega}^2 \quad \text{equator} \quad (277)$$

$$\bar{U}_0 = -GM/\bar{c} + GM\bar{a}^2 \bar{I}_2/\bar{c}^3 \quad \text{north pole} \quad (278)$$

where  $S_\chi(0) = 0$  was used to obtain equation (277), and  $S_\chi(\pi/2) = 1$  and  $\theta_{S\chi}(\pi/2) = 0$  were used to obtain equation (278), and where  $\bar{c} = ce^{j\theta_c}$  = complex number radius at the poles. The potentials in equations (277) and (278) must have the same value for the geoid so that the complex number flattening is given by the following generalization of the standard scalar result<sup>35</sup>

$$\bar{f} = fe^{j\theta_f} = (\bar{a} - \bar{c})/\bar{a} = 3/2 \bar{I}_2 + \bar{\omega}^2 \bar{a}^3/(2GM) \quad (279)$$

where it is assumed that  $\bar{a} \sim \bar{c}$  to obtain this equation. From equation (279) it follows that

$$f \cos \theta_f = 3/2 I_2 \cos \theta_{I2} + \omega^2 a^3 / (2GM) \cos(2\theta_\omega + 3\theta_a) \quad (280)$$

$$f \sin \theta_f = 3/2 I_2 \sin \theta_{I2} + \omega^2 a^3 / (2GM) \sin(2\theta_\omega + 3\theta_a) \quad (281)$$

These equations determine  $f$  and  $\theta_f$ . The measured value of the flattening is  $f_m = f \cos \theta_f$ .

The real and imaginary parts of equation (275) are

$$U \cos \theta_U = -GM/r \cos \theta_r + GM I_2 P_2 a^2 / r^3 \cos(\theta_{I2} + \theta_{P2} + 2\theta_a - 3\theta_r) \quad (282)$$

$$- 1/2 r^2 \omega^2 C_\chi^2 \cos(2\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

$$U \sin \theta_U = GM/r \sin \theta_r + GM I_2 P_2 a^2 / r^3 \sin(\theta_{I2} + \theta_{P2} + 2\theta_a - 3\theta_r) \quad (283)$$

$$- 1/2 r^2 \omega^2 C_\chi^2 \sin(2\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

Equivalently equation (276) can be used to write

$$U \cos \theta_U = -GM/r \cos \theta_r + 3/2 GM I_2 a^2 S_\chi^2 / r^3 \cos(\theta_{I2} + 2\theta_{s\chi} + 2\theta_a - 3\theta_r) \quad (284)$$

$$- 1/2 GM I_2 a^2 / r^3 \cos(\theta_{I2} + 2\theta_a - 3\theta_r) - 1/2 r^2 \omega^2 C_\chi^2 \cos(2\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

$$U \sin \theta_U = GM/r \sin \theta_r + 3/2 GM I_2 a^2 S_\chi^2 / r^3 \sin(\theta_{I2} + 2\theta_{s\chi} + 2\theta_a - 3\theta_r) \quad (285)$$

$$- 1/2 GM I_2 a^2 / r^3 \sin(\theta_{I2} + 2\theta_a - 3\theta_r) - 1/2 r^2 \omega^2 C_\chi^2 \sin(2\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

#### E. Acceleration of Gravity

The acceleration of gravity of a rotating planet is obtained from equation (275) by

$$\bar{g} = g e^{j\theta} = -\partial \bar{U} / \partial \bar{r} \quad (286)$$

$$= -GM/\bar{r}^2 + 3GM/\bar{r}^2 (\bar{a}/\bar{r})^2 \bar{I}_2 \bar{P}_2 + \bar{r} \bar{\omega}^2 \cos^2 \bar{\chi}$$

The real and imaginary parts of equation (286) are

$$g \cos \theta_g = -GM/r^2 \cos(2\theta_r) + 3GMa^2/r^4 I_2 P_2 \cos(\theta_{I2} + \theta_{P2} + 2\theta_a - 4\theta_r) \quad (287)$$

$$+ r \omega^2 C_\chi^2 \cos(\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

$$g \sin \theta_g = GM/r^2 \sin(2\theta_r) + 3GMa^2/r^4 I_2 P_2 \sin(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 4\theta_r) \quad (288)$$

$$+ r\omega^2 C_\chi^2 \sin(\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

Equation (286) can also be written as

$$\bar{g} = -GM/\bar{r}^2 + 3GM/\bar{r}^2 (\bar{a}/\bar{r})^2 \bar{I}_2 \frac{1}{2}(3S_\chi^2 e^{2j\theta_{s\chi}} - 1) + \bar{r}\bar{\omega}^2 C_\chi^2 e^{-2j\theta_{c\chi}} \quad (289)$$

and therefore

$$g \cos \theta_g = -GM/r^2 \cos(2\theta_r) + 9/2 GMa^2/r^4 I_2 S_\chi^2 \cos(\theta_{I_2} + 2\theta_{s\chi} + 2\theta_a - 4\theta_r) \quad (290)$$

$$- 3/2 GMa^2/r^4 I_2 \cos(\theta_{I_2} + 2\theta_a - 4\theta_r) + r\omega^2 C_\chi^2 \cos(\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

$$g \sin \theta_g = GM/r^2 \sin(2\theta_r) + 9/2 GMa^2/r^4 I_2 S_\chi^2 \sin(\theta_{I_2} + 2\theta_{s\chi} + 2\theta_a - 4\theta_r) \quad (291)$$

$$- 3/2 GMa^2/r^4 I_2 \sin(\theta_{I_2} + 2\theta_a - 4\theta_r) + r\omega^2 C_\chi^2 \sin(\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

The measured acceleration of gravity is given by  $g \cos \theta_g$ . Equations (287) and (288) or (290) and (291) can be used to determine  $g$  and  $\theta_g$ . These equations can be written in terms of measured quantities by making the substitutions

$$r = r_m / \cos \theta_r \quad (292)$$

$$a = a_m / \cos \theta_a \quad (293)$$

$$c = c_m / \cos \theta_c \quad (294)$$

$$\chi = \chi_m / \cos \theta_\chi \quad (295)$$

$$\omega = \omega_m / \cos \theta_\omega \quad (296)$$

$$I_2 = I_{2m} / \cos \theta_{I_2} \quad (297)$$

$$f = f_m / \cos \theta_f \quad (298)$$

where  $r_m$  = measured radial coordinate,  $a_m$  = measured equatorial radius,  $c_m$  = measured polar radius,  $\chi_m$  = measured latitude,  $\omega_m$  = measured rotational speed,  $I_{2m}$  = measured mass distribution coefficients and  $f_m$  = measured flattening. Substituting equations (292) through (298) into equation (287) gives

$$g_m = -GM/r_m^2 \cos(2\theta_r) \cos^2 \theta_r \quad (299)$$

$$+ (3GMa_m^2 I_{2m} P_2 \cos^4 \theta_r) / (r_m^4 \cos^2 \theta_a \cos \theta_{I_2}) \cos(\theta_{I_2} + \theta_{P_2} + 2\theta_a - 4\theta_r)$$

$$+ (r_m \omega_m^2 C_\chi^2) / (\cos \theta_r \cos^2 \theta_\omega) \cos(\theta_r + 2\theta_\omega - 2\theta_{c\chi})$$

Equivalently, substituting equations (292) through (298) into equation (290) gives the measured acceleration of gravity as

$$\begin{aligned}
 g_m = & -GM/r_m^2 \cos(2\theta_r) \cos^2 \theta_r & (300) \\
 & + (9GMa_m^2 I_{2m} S_\chi^2 \cos^4 \theta_r) / (2r_m^4 \cos^2 \theta_a \cos \theta_{I2}) \cos(\theta_{I2} + 2\theta_{s\chi} + 2\theta_a - 4\theta_r) \\
 & - (3GMa_m^2 I_{2m} \cos^4 \theta_r) / (2r_m^4 \cos^2 \theta_a \cos \theta_{I2}) \cos(\theta_{I2} + 2\theta_a - 4\theta_r) \\
 & + (r_m \omega_m^2 C_\chi^2) / (\cos \theta_r \cos^2 \theta_\omega) \cos(\theta_r + 2\theta_\omega - 2\theta_\chi)
 \end{aligned}$$

From equations (218) and (219) it follows that  $S_\chi$  and  $C_\chi$  can be expressed in terms of the measured latitude by

$$S_\chi = [\sin^2 \chi_m + \sinh^2(\chi_m \tan \theta_\chi)]^{1/2} \quad (301)$$

$$C_\chi = [\cos^2 \chi_m + \sinh^2(\chi_m \tan \theta_\chi)]^{1/2} \quad (302)$$

For small  $\theta_\chi$  it follows that

$$S_\chi \sim \sin \chi_m \quad (303)$$

$$C_\chi \sim \cos \chi_m \quad (304)$$

From equations (220) and (221) it follows that

$$\tan \theta_{s\chi} = \cot \chi_m \tanh(\chi_m \tan \theta_\chi) \quad (305)$$

$$\tan \theta_{c\chi} = \tan \chi_m \tanh(\chi_m \tan \theta_\chi) \quad (306)$$

#### F. Acceleration of Gravity on the Geoid

The approximate shape of the geoid for a planet with broken internal symmetries is written as a simple generalization of the standard scalar expression as follows<sup>35</sup>

$$\bar{r} = \bar{a}(1 - \bar{f} \sin^2 \bar{\chi}) \quad (307)$$

From equation (307) it follows that

$$\bar{r}^{-2} = \bar{a}^{-2}(1 + 2\bar{f} \sin^2 \bar{\chi} + \dots) \quad (308)$$

Using equation (308) in the first term of equation (286) gives for the acceleration on the geoid

$$\bar{g} = -GM/\bar{a}^2(1 + 2\bar{f} \sin^2 \bar{\chi}) + 3GM/\bar{a}^2 \bar{J}_2 \bar{P}_2 + \bar{a}\bar{\omega}^2 \cos^2 \bar{\chi} \quad (309)$$

The real and imaginary parts of equation (309) are written as

$$g \cos \theta_g = - GM/a^2 \cos(2\theta_a) - 2GMfS_X^2/a^2 \cos(\theta_f + 2\theta_{sX} - 2\theta_a) \quad (310)$$

$$+ 3GMI_2P_2/a^2 \cos(\theta_{I2} + \theta_{P2} - 2\theta_a) + a\omega^2 C_X^2 \cos(\theta_a + 2\theta_\omega - 2\theta_{cX})$$

$$= - GM/a^2 \cos(2\theta_a) - 2GMfS_X^2/a^2 \cos(\theta_f + 2\theta_{sX} - 2\theta_a)$$

$$+ 9/2GMI_2S_X^2/a^2 \cos(\theta_{I2} + 2\theta_{sX} - 2\theta_a) - 3/2GMI_2/a^2 \cos(\theta_{I2} - 2\theta_a)$$

$$+ a\omega^2 C_X^2 \cos(\theta_a + 2\theta_\omega - 2\theta_{cX})$$

$$g \sin \theta_g = GM/a^2 \sin(2\theta_a) - 2GMfS_X^2/a^2 \sin(\theta_f + 2\theta_{sX} - 2\theta_a) \quad (311)$$

$$+ 3GMI_2P_2/a^2 \sin(\theta_{I2} + \theta_{P2} - 2\theta_a) + a\omega^2 C_X^2 \sin(\theta_a + 2\theta_\omega - 2\theta_{cX})$$

$$= GM/a^2 \sin(2\theta_a) - 2GMfS_X^2/a^2 \sin(\theta_f + 2\theta_{sX} - 2\theta_a)$$

$$+ 9/2GMI_2S_X^2/a^2 \sin(\theta_{I2} + 2\theta_{sX} - 2\theta_a) - 3/2GMI_2/a^2 \sin(\theta_{I2} - 2\theta_a)$$

$$+ a\omega^2 C_X^2 \sin(\theta_a + 2\theta_\omega - 2\theta_{cX})$$

Equations (310) and (311) can be written in terms of measured quantities by using equations (292) through (298). The measured acceleration on the geoid is given by equation (310).

The acceleration of gravity at the equator is obtained by using equation (309) with equations (239) through (244) that describe the equator, with the result that

$$\bar{g}_e = - GM/\bar{a}^2 - 3/2GMI_2/\bar{a}^2 + \bar{a}\bar{\omega}^2 \quad (312)$$

The real and imaginary parts of equation (312) give

$$g_e \cos \theta_{ge} = - GM/a^2 \cos(2\theta_a) - 3/2GMI_2/a^2 \cos(\theta_{I2} - 2\theta_a) + a\omega^2 \cos(\theta_a + 2\theta_\omega) \quad (313)$$

$$g_e \sin \theta_{ge} = GM/a^2 \sin(2\theta_a) - 3/2GMI_2/a^2 \sin(\theta_{I2} - 2\theta_a) + a\omega^2 \sin(\theta_a + 2\theta_\omega) \quad (314)$$

from which  $g_e$  and  $\theta_{ge}$  can be determined. Combining equations (279), (309) and (312) and neglecting higher order terms gives

$$\bar{g} = \bar{g}_e [1 + (5/2\bar{m} - \bar{f}) \sin^2 \bar{\chi}] \quad (315)$$

where

$$\bar{m} = me^{j\theta_m} = \frac{\omega^2 a^3}{GM} \quad (316)$$

and where  $\bar{m}$  is related to the complex number flattening  $\bar{f}$  by equation (279) which can be rewritten as

$$\bar{f} = 3/2\bar{I}_2 + \bar{m}/2 \quad (317)$$

Equation (315) is the complex number generalization of Clairut's equation.<sup>35</sup> Equation (315) can be written as

$$g \cos \theta_g = g_e \cos \theta_{ge} + 5/2g_e m S_\chi^2 \cos(\theta_{ge} + \theta_m + 2\theta_{s\chi}) - g_e f S_\chi^2 \cos(\theta_{ge} + \theta_f + 2\theta_{s\chi}) \quad (318)$$

$$g \sin \theta_g = g_e \sin \theta_{ge} + 5/2g_e m S_\chi^2 \sin(\theta_{ge} + \theta_m + 2\theta_{s\chi}) - g_e f S_\chi^2 \sin(\theta_{ge} + \theta_f + 2\theta_{s\chi}) \quad (319)$$

### G. Apparent Non-Newtonian Effects

The measured acceleration of gravity given by equation (299) has to be compared to the conventionally calculated acceleration of gravity in order to estimate the magnitude of the discrepancy. The conventionally calculated acceleration of gravity is just the scalar form of equation (286) in which the measured distances and angles appear as follows<sup>35</sup>

$$\begin{aligned} g^c &= -GM/r_m^2 + 3GMa_m^2/r_m^4 I_{2c} P_{2c} + r_m \omega_m^2 \cos^2 \chi_m \\ &= -GM/r_m^2 + 9/2GMa_m^2/r_m^4 I_{2c} \sin^2 \chi_m \\ &\quad - 3/2GMa_m^2/r_m^4 I_{2c} + r_m \omega_m^2 \cos^2 \chi_m \end{aligned} \quad (320)$$

where the conventionally calculated second order Legendre polynomial is written as<sup>29</sup>

$$P_{2c} = 1/2(3 \sin^2 \chi_m - 1) \quad (321)$$

The conventionally calculated mass distribution coefficient  $I_{2c}$  is similar to equation (265) in that<sup>35</sup>

$$I_{2c} = (C_c - A_c)/(Ma_m^2) \quad (322)$$

where the conventional moments of inertia are given by<sup>35</sup>

$$C_c = \int (x_m^2 + y_m^2) dM \quad (323)$$

$$A_c = \int (x_m^2 + z_m^2) dM = \int (y_m^2 + z_m^2) dM \quad (324)$$

Thus the conventional calculations are done using the measured coordinates of geodesy  $r_m$ ,  $x_m$ ,  $y_m$ ,  $z_m$  and  $\chi_m$ .

Comparing equation (320) with equation (299) gives

$$\begin{aligned}
g_m - g_c = & GM/r_m^2 [1 - \cos(2\theta_r) \cos^2 \theta_r] \quad (325) \\
& + \frac{3MGa^2}{r_m^4} \left[ \frac{\cos^4 \theta_r}{\cos^2 \theta_a} I_2 P_2 \cos(\theta_{I2} + \theta_{P2} + 2\theta_a - 4\theta_r) - I_{2c} P_{2c} \right] \\
& + r_m \omega^2 \left[ \frac{C_\chi^2}{\cos \theta_r \cos^2 \theta_\omega} \cos(\theta_r + 2\theta_\omega - 2\theta_{c\chi}) - \cos^2 \chi_m \right]
\end{aligned}$$

where  $P_2$  and  $\theta_{P2}$  are expressed in terms of the measured latitude  $\chi_m$  by using equations (235), (236), (301), (302) and (306). In a similar fashion combining equation (300) and (320) gives

$$\begin{aligned}
g_m - g_c = & GM/r_m^2 [1 - \cos(2\theta_r) \cos^2 \theta_r] \quad (326) \\
& + \frac{9GMa^2}{2r_m^4} \left[ \frac{\cos^4 \theta_r}{\cos^2 \theta_a} I_2 S_\chi^2 \cos(\theta_{I2} + 2\theta_{s\chi} + 2\theta_a - 4\theta_r) - I_{2c} \sin^2 \chi_m \right] \\
& - \frac{3GMa^2}{2r_m^4} \left[ \frac{\cos^4 \theta_r}{\cos^2 \theta_a} I_2 \cos(\theta_{I2} + 2\theta_a - 4\theta_r) - I_{2c} \right] \\
& + r_m \omega^2 \left[ \frac{C_\chi^2}{\cos \theta_r \cos^2 \theta_\omega} \cos(\theta_r + 2\theta_\omega - 2\theta_{c\chi}) - \cos^2 \chi_m \right]
\end{aligned}$$

The first term in equations (325) or (326) gives the dominant effect of the broken symmetry of space on the discrepancy between the measured and conventionally calculated values of the Newtonian acceleration of gravity. From these two equations the parameters  $D$  given by equation (69) and  $D_2$  given by equation (85) can be calculated.

5. MINE SHAFT, BOREHOLE, TOWER AND EÖTVÖS EXPERIMENTS. This section considers the apparent deviations from Newtonian gravity that have recently been reported in the literature.<sup>1-25</sup> These discrepancies have been found in laboratory Eötvös experiments where the validity of Newton's gravitation law is examined over short ranges for deviations from the inverse square law and to detect a possible dependence on the composition (baryon number) of the attracting masses.<sup>6-15</sup> Deviations from the inverse square law have also been found in the measurement of the acceleration of gravity over vertical distances of hundreds of meters in mine shaft, borehole and tower experiments.<sup>1-5, 24, 25</sup> An analysis of these apparent discrepancies is given in this section which is based on the broken symmetry of space that is induced by a pressure field.<sup>29</sup> A spherical earth assumption is made for the calculations done here so that the acceleration of gravity and the effective radial gravitational constant are given by equations (64) and (75) respectively in terms of the internal phase angle  $\theta_r$  of the radial coordinates.



A. Small Argument Approximation to the Equilibrium Equation for the Internal Phase Angles of the Radial Coordinates.

For the case where  $\theta_p$  varies slowly with radial distance, the following approximations can be written for equation (57)

$$\theta_p + P \frac{\partial \theta_p / \partial r}{\partial P / \partial r} = - 2\theta_r \quad (327)$$

The solution of equation (327) determines  $\theta_r(r)$  in terms of  $P(r)$  and  $\theta_p(r)$  and hence by equations (64) and (75) the acceleration of gravity and the effective gravitational constant are also obtained. As a simple example consider the case

$$P = P(0)e^{-\alpha r} \quad (328)$$

$$\theta_p = \theta_p(0)e^{-\beta r} \quad (329)$$

where  $P(0)$  and  $\theta_p(0)$  = pressure and its internal phase angle at the center of the earth. Combining equations (327) through (329) gives

$$\begin{aligned} \theta_r &= - 1/2(\alpha + \beta)/\alpha\theta_p(0)e^{-\beta r} \\ &= - 1/2(\alpha + \beta)/\alpha\theta_p \end{aligned} \quad (330)$$

For the center of the earth  $r = 0$  and equation (330) gives

$$\theta_r(0) = - 1/2(\alpha + \beta)/\alpha\theta_p(0) \quad (331)$$

At the earth's surface  $r = R$  and

$$\begin{aligned} \theta_r(R) &= - 1/2(\alpha + \beta)/\alpha\theta_p(0)e^{-\beta R} \\ &= - 1/2(\alpha + \beta)/\alpha\theta_p(R) \end{aligned} \quad (332)$$

where  $\theta_p(R)$  = internal phase angle of pressure at the earth's surface given by

$$\theta_p(R) = \theta_p(0)e^{-\beta R} \quad (333)$$

Note also that the pressure at the earth's surface is given by

$$P(R) = P(0)e^{-\alpha R} \quad (334)$$

Equations (333) and (334) can be used to evaluate  $\alpha$  and  $\beta$ .

For the case of a linear variation of the pressure and its internal phase angle of the form

$$P = P(0) - \alpha r \quad (335)$$

$$\theta_p = \theta_p(0) - \beta r \quad (336)$$

it follows from equation (327) that

$$\theta_r = - 1/2[\theta_p(0) + \beta/\alpha P(0)] + \beta r \quad (337)$$

The values of  $\alpha$  and  $\beta$  can be obtained by evaluating equations (335) and (336) at the earth's surface as follows

$$P(R) = P(0) - \alpha R \quad (338)$$

$$\theta_p(R) = \theta_p(0) - \beta R \quad (339)$$

At the center of the earth equation (337) gives

$$\theta_r(0) = - 1/2[\theta_p(0) + \beta/\alpha P(0)] \quad (340)$$

while for the surface of the earth

$$\theta_r(R) = - 1/2[\theta_p(0) + \beta/\alpha P(0)] + \beta R \quad (341)$$

Equations (330) and (337) show that in general  $\theta_r < 0$  within a gravitating body.

#### B. Theory of the Apparent Non-Newtonian Behaviour of Gravity in Mine Shaft, Borehole and Tower Experiments.

Measurements of the variation of the acceleration of gravity up the heights of a tower and down the depths of a mine shaft or borehole have indicated discrepancies with the inverse square law of Newtonian gravity. A possible explanation of these discrepancies has been given by assuming the validity of Newtonian gravitation in matter with broken internal symmetries.<sup>28</sup> The result is that the acceleration of gravity for a spherical earth is given by equation (64). In order to apply this equation to an analysis of mine shaft, borehole and tower gravity measurements it is first necessary to calculate the internal phase angle  $\theta_r$  from equation (327). Let the coordinates measured up a tower from the earth's surface be designated by  $h$ , so that the distance from the center of the earth to a point on the tower is given by

$$r = R + h \quad (342)$$

where  $R$  = magnitude of the earth's radius at the base of the tower. Equation (342) applies to a mine shaft or borehole if  $h < 0$ . Combining equations (327) and (342) gives

$$\theta_p + P \frac{\partial \theta_p / \partial h}{\partial P / \partial h} = - 2\theta_r \quad (343)$$

as the equation for determining  $\theta_r$ .

The magnitude of the atmospheric pressure at points on a tower, mine shaft or borehole can be written in its simplest form by the following linear equation

$$P = P(R) - \rho(R)g(R)h = \rho(R)g(R)(h^a - h) \quad (344)$$

where the equivalent height of the atmosphere is given by

$$h^a = P(R) / [\rho(R)g(R)] \quad (345)$$

and where  $P(R)$ ,  $\rho(R)$  and  $g(R)$  = magnitudes of the pressure, air density and acceleration of gravity respectively at the earth's surface. The measured values of these quantities are given by  $P_m(R) = P(R) \cos \theta_p(R)$ ,  $\rho_m(R) = \rho(R) \cos \theta_\rho(R)$  and  $g_m(R) = g(R) \cos \theta_g(R)$  respectively. The measured pressure is given by  $P_m = P \cos \theta_p$ . The internal phase angle of the pressure will be written in a form similar to equation (344) as follows

$$\theta_p = \theta_p(R) - nh \quad (346)$$

Equations (344) and (346) are the simplest equations that can be chosen to describe the variation with height (or depth) of the atmospheric pressure and its internal phase angle. Strictly speaking  $P$ ,  $\theta_p$  and  $\theta_r$  should be determined simultaneously from equations (56) and (57) and the renormalized state equation which is given by a solution of the complex number relativistic trace equation (2). Such a simultaneous solution is difficult to obtain. Equations (344) and (346) represent a crude solution to equations (2), (56) and (57). These assumed solutions will now be used to obtain  $\theta_r$  from equation (343). Combining equations (343), (344) and (346) gives

$$\begin{aligned} \theta_r &= -1/2\theta_p(R) - 1/2n(h^2 - 2h) \\ &= -1/2[\theta_p(R) + nh^2] + nh \end{aligned} \quad (347)$$

At the earth's surface

$$\theta_r(R) = -1/2[\theta_p(R) + nh^2] \quad (348)$$

Consider now the case where the pressure and its internal phase vary according to the following exponential forms

$$P = P(R)e^{-\delta h} \quad (349)$$

$$\theta_p = \theta_p(R)e^{-\kappa h} \quad (350)$$

Combining equations (343), (349) and (350) gives

$$\theta_r = -1/2(\delta + \kappa)/\delta\theta_p(R)e^{-\kappa h} \quad (351)$$

and the value at the earth's surface is

$$\theta_r(R) = -1/2(\delta + \kappa)/\delta\theta_p(R) \quad (352)$$

The values of  $\theta_r$  determine the apparent deviation of the acceleration of gravity from Newton's law of gravity as is shown in equations (64) and (75). Figures 2 and 3 show sketches of the variation of  $P$  and  $\theta_p$  for the solid earth, ocean and for the atmosphere in an air-filled mine shaft or borehole or adjacent to a tower. The expected variation of  $\theta_r$  in the solid earth, ocean and atmosphere is shown in Figure 4, while Figure 5 shows the corresponding variation of  $G_r$  as given by equation (75). Figure 5 shows that local measurements of the acceleration of gravity will yield values of  $G_r$  which are less than the value of  $G$ . The value of  $G$  is associated with the complete symmetry of time

and space, and as such it cannot be measured directly. The result  $G_r < G$  is due to the effects of the complex number atmospheric pressure in the case of mine shaft, borehole and tower experiments, and to the complex number water pressure for measurements of  $G_r$  carried out in the depths of the ocean. For measurements of the variation of  $G_r$  in a mine shaft, borehole or up a tower the characteristic range for the variation of  $G_r$  should be about 7 km because the atmospheric pressure decrease with height has a characteristic attenuation distance of about 7 km.<sup>43</sup>

Equation (75) and Figure 5 also show that were it possible to measure the variation of the acceleration of gravity with depth in solid rock the values for  $G_r$  would be less than those measured in an air-filled mine shaft or in the ocean. This is because  $P$ ,  $\theta_p$  and  $|\theta_r|$  are larger in rock than in the ocean or in an air-filled mine shaft at a corresponding depth. Measurements of  $G_r$  in the ocean should yield weaker gravity (smaller  $G_r$ ) than corresponding measurements in an air-filled mine shaft or borehole because values of  $|\theta_r|$  in the ocean are larger than their corresponding values in an air-filled mine shaft or borehole at the same depth (see equation (75) and Figure 4).

### C. Internal Phase Theory of the Eötvös Experiment and its Relationship to Mine Shaft, Borehole and Tower Experiments.

This part of the paper describes a theoretical analysis of the Eötvös experiment in terms of Newtonian gravity and the broken symmetry internal phase angles of the relevant coordinates of the experiment. The Eötvös experiment has been thoroughly described in the literature and only the briefest review is given in this paragraph.<sup>6-15</sup> This experiment measures the horizontal force of gravity between two spheres of material that are suspended in close proximity to each other. Conventional Newtonian theory predicts the measured gravity force to be dependent on the inverse square of the separation distance and on the product of the masses of the two spheres, but recent experiments suggest the possibility of composition dependent effects and deviations from the inverse square law.<sup>5-15</sup>

Consider now the Eötvös experiment from the perspective of the internal phase theory of coordinates. The two spheres can be oriented in any direction between the north-south direction (whose separation is described by a decrement of the zenith angle) or in the east-west direction (whose separation is then described by a decrement of the azimuthal angle). For the north-south orientation the complex number distance between the two spheres is written as

$$d\bar{\ell}_\psi = d\ell_\psi e^{j\theta\ell\psi} = \bar{r}d\bar{\psi} = \bar{r}e^{j\theta\psi}(d\psi + j\psi d\theta_\psi) \quad (353)$$

which gives

$$d\ell_\psi = r \sec \beta_{\psi,\psi} d\psi \quad (354)$$

$$\theta\ell\psi = \theta_r + \theta_\psi + \beta_{\psi,\psi} \quad (355)$$

where  $\bar{r} = \bar{R} + \bar{h}$  = complex number distance of the two spheres from the center of the earth,  $\bar{R}$  = complex number earth's radius at the position of the two spheres,  $\bar{h}$  = complex number distance above (or below) the earth's surface at which the

Eötvös experiment is conducted,  $d\bar{\psi}$  = complex number zenith angle separation of the two spheres situated in the north-south direction (longitudinal plane) and  $\beta_{\psi, \psi}$  is given by equation (140). The measured distance between the two spheres situated in the north-south orientation is given by

$$d\ell_{\psi m} = d\ell_{\psi} \cos \theta_{\ell\psi} \quad (356)$$

For the east-west orientation the complex number distance between the two spheres is

$$d\bar{\ell}_{\phi} = d\ell_{\phi} e^{j\theta_{\ell\phi}} = \bar{r} \sin \bar{\psi} d\bar{\phi} \quad (357)$$

where  $d\bar{\phi}$  = complex number azimuthal angle separation of the two spheres that are situated in a plane of latitude  $\bar{\chi} = \pi/2 - \bar{\psi}$ , and therefore

$$d\ell_{\phi} = r S_{\psi} \sec \beta_{\phi, \phi} d\phi \quad (358)$$

$$\theta_{\ell\phi} = \theta_r + \theta_{s\psi} + \theta_{\phi} + \beta_{\phi, \phi} \quad (359)$$

where  $S_{\psi}$  and  $\theta_{s\psi}$  are given by equations (143) and (144) respectively, and where  $\beta_{\phi, \phi}$  is given by equation (141). The measured distance between the two spheres situated in the east-west direction is given by

$$d\ell_{\phi m} = d\ell_{\phi} \cos \theta_{\ell\phi} \quad (360)$$

The gravitational force between the two spheres situated in the north-south direction is

$$\bar{F}_{\psi} = - Gm^2 / (d\bar{\ell}_{\psi})^2 = - Gm^2 / (d\ell_{\psi})^2 e^{-2j\theta_{\ell\psi}} \quad (361)$$

where  $m$  = mass of one sphere. The measured gravitational force between the two spheres in the north-south direction is

$$\begin{aligned} F_{\psi m} &= - Gm^2 / (d\ell_{\psi})^2 \cos(2\theta_{\ell\psi}) \\ &= - Gm^2 / (d\ell_{\psi m})^2 \cos(2\theta_{\ell\psi}) \cos^2 \theta_{\ell\psi} \end{aligned} \quad (362)$$

The conventional calculation of the Newtonian gravitational force between the two Eötvös spheres is given by

$$F_{\psi c} = - Gm^2 / (d\ell_{\psi m})^2 \quad (363)$$

Therefore the difference between the measured and conventionally predicted forces in the north-south direction is

$$\begin{aligned} \Delta F_{\psi} &= F_{\psi m} - F_{\psi c} \\ &= Gm^2 / (d\ell_{\psi m})^2 [1 - \cos^2 \theta_{\ell\psi} \cos(2\theta_{\ell\psi})] \\ &\sim 3\theta_{\ell\psi}^2 Gm^2 / (d\ell_{\psi m})^2 \end{aligned} \quad (364)$$

In a similar fashion the complex number gravitational force and the measured force between the two spheres situated in the east-west orientation are given respectively by

$$\bar{F}_\phi = -Gm^2/(d\bar{l}_\phi)^2 = -Gm^2/(dl_\phi)^2 e^{-2j\theta_{l\phi}} \quad (365)$$

$$\begin{aligned} F_{\phi m} &= -Gm^2/(dl_\phi)^2 \cos(2\theta_{l\phi}) \\ &= -Gm^2/(dl_{\phi m})^2 \cos(2\theta_{l\phi}) \cos^2 \theta_{l\phi} \end{aligned} \quad (366)$$

The conventionally calculated gravitational force is given by

$$F_{\phi c} = -Gm^2/(dl_{\phi m})^2 \quad (367)$$

and the difference between equations (366) and (367) is

$$\begin{aligned} \Delta F_\phi &= F_{\phi m} - F_{\phi c} \\ &= Gm^2/(dl_{\phi m})^2 [1 - \cos^2 \theta_{l\phi} \cos(2\theta_{l\phi})] \\ &\sim 3\theta_{l\phi}^2 Gm^2/(dl_{\phi m})^2 \end{aligned} \quad (368)$$

A measurement of the discrepancies between the measured and predicted values of the gravity force for the north-south and east-west orientations of the Eötvös experiment will give values of  $\theta_{l\psi}$  and  $\theta_{l\phi}$ .

From equations (75), (362) and (366) it follows that there are three effective gravitational constants each associated with a direction ( $r$ ,  $\psi$  or  $\phi$ ) of measurement of the gravitational force, so that

$$G_r = G \cos(2\theta_r) \cos^2 \theta_r \sim G(1 - 3\theta_r^2 + \dots) \quad (369)$$

$$G_\psi = G \cos(2\theta_{l\psi}) \cos^2 \theta_{l\psi} \sim G(1 - 3\theta_{l\psi}^2 + \dots) \quad (370)$$

$$G_\phi = G \cos(2\theta_{l\phi}) \cos^2 \theta_{l\phi} \sim G(1 - 3\theta_{l\phi}^2 + \dots) \quad (371)$$

Due to the internal phase structure of the coordinates the effective Newtonian gravitational constant has three distinct values along the three orthogonal directions at a point on the earth's surface. Because  $|\theta_\psi|$  and  $|\theta_\phi|$  are expected to be smaller than  $|\theta_r|$  it follows from equations (355) and (359) that for the same height (or depth)

$$\left. \begin{aligned} |\theta_{l\phi}| &< |\theta_{l\psi}| < |\theta_r| \\ G_\phi &> G_\psi > G_r \end{aligned} \right\} \theta_r < 0, \theta_\phi > 0, \theta_\psi > 0 \quad (372)$$

$$\left. \begin{array}{l} |\theta_{\ell\phi}| > |\theta_{\ell\psi}| > |\theta_r| \\ G_\phi < G_\psi < G_r \end{array} \right\} \theta_r < 0, \theta_\phi < 0, \theta_\psi < 0 \quad (373)$$

where for both cases  $\theta_{\ell\phi} < 0$  and  $\theta_{\ell\psi} < 0$ . According to this theory, depending on the signs of  $\theta_\phi$  and  $\theta_\psi$  the values of  $G_\psi$  and  $G_\phi$  measured by the Eötvös experiment can be greater or less than the value of  $G_r$  determined by gravimeter measurements in a mine shaft, borehole or tower experiment. References 1 through 3 suggest that equation (373) are the correct conditions while references 5 and 25 suggest that equation (372) gives the correct relationship between  $G_\phi$ ,  $G_\psi$  and  $G_r$ . The experimental results are not yet clear enough to decide between  $\theta_\phi > 0$  and  $\theta_\psi > 0$  or  $\theta_\phi < 0$  and  $\theta_\psi < 0$ .

On account of equations (369) through (371) it follows that  $G_\psi$  and  $G_\phi$  measured by an Eötvös experiment should have a similar variation with depth (or height) as does  $G_r$ . This is shown in Figure 5. The validity of equations (370) and (371) can possibly be tested by conducting Eötvös experiments in the depths of the ocean, down mine shafts, or up a tower in order to see if  $G_\psi$  and  $G_\phi$  vary in the same sense as  $G_r$ . For a tower or air-filled mine shaft Eötvös experiment the characteristic length over which  $G_\psi$  and  $G_\phi$  change should be about 7 km because this is the characteristic variation distance of the atmospheric pressure in the vertical direction.<sup>43</sup> The characteristic distance for the decrease of  $G_\psi$  and  $G_\phi$  with depth in the ocean (or solid earth if such experiments were possible) should be much larger than 7 km because the pressure changes in these cases are over hundreds and thousands of kilometers.<sup>35-40, 43-45</sup> Another possible test would be to perform the Eötvös experiment in a pressure chamber and measure the pressure dependence of  $G_\psi$  and  $G_\phi$  in order to verify that  $G_\psi$  and  $G_\phi$  are decreasing functions of the ambient pressure as suggested by equations (370) and (371). In any case, equations (369) through (371) show that the local measurements of gravity do not directly determine the Newtonian gravitational constant  $G$ . Approximate values of  $G$  can be determined directly from satellite or solar system measurements where the effects of ambient pressure are negligible, but in this case the values  $\theta_r^V$ ,  $\theta_\psi^V$  and  $\theta_\phi^V$  of the broken symmetry vacuum must be taken into consideration. Thus even for measurements in the vacuum  $G$  cannot be directly measured.

Consider the variation of the gravitational constant  $G_r$  given by equation (369) from which it follows that

$$G_r(R-h) - G_r(R) \sim 3G[\theta_r^2(R) - \theta_r^2(R-h)] \quad (374)$$

$$G_r(R+h) - G_r(R) \sim 3G[\theta_r^2(R) - \theta_r^2(R+h)] \quad (375)$$

A Taylor series expansion of  $\theta_r$  gives

$$\theta_r(R-h) = \theta_r(R) - h\partial\theta_r/\partial h + \dots \quad (376)$$

$$\theta_r(R+h) = \theta_r(R) + h\partial\theta_r/\partial h + \dots \quad (377)$$

Combining equations (374) through (377) gives for small  $h$

$$[G_r(R-h) - G_r(R)]/G \sim -s_r h \quad (378)$$

$$[G_r(R+h) - G_r(R)]/G \sim s_r h \quad (379)$$

where

$$s_r = 6|\theta_r(R)| \partial\theta_r/\partial r|_R > 0 \quad (380)$$

remembering that  $\theta_r < 0$ . Combining equations (378) and (379) gives

$$G_r(R-h) < G_r(R) < G_r(R+h) \quad (381)$$

as shown in Figure 5. Note that  $G_r(R)$  is not the value of the gravitational constant that is measured at the earth's surface by the Eötvös experiment. The value of the gravitational constant measured by the Eötvös experiment is given by  $G_\psi(R)$  or  $G_\phi(R)$ .

The Eötvös experiment can be done at various depths and heights. From equations (370) and (371) it follows that

$$G_\psi(R\pm h) - G_\psi(R) \sim 3G[\theta_{\ell\psi}^2(R) - \theta_{\ell\psi}^2(R\pm h)] \quad (382)$$

$$G_\phi(R\pm h) - G_\phi(R) \sim 3G[\theta_{\ell\phi}^2(R) - \theta_{\ell\phi}^2(R\pm h)] \quad (383)$$

A Taylor series is used to obtain

$$\theta_{\ell\psi}(R\pm h) = \theta_{\ell\psi}(R) \pm h\partial\theta_{\ell\psi}/\partial r|_R + \dots \quad (384)$$

$$\theta_{\ell\phi}(R\pm h) = \theta_{\ell\phi}(R) \pm h\partial\theta_{\ell\phi}/\partial r|_R + \dots \quad (385)$$

Combining equations (382) through (385) gives for small  $h$

$$[G_\psi(R\pm h) - G_\psi(R)]/G \sim \pm s_\psi h \quad (386)$$

$$[G_\phi(R\pm h) - G_\phi(R)]/G \sim \pm s_\phi h \quad (387)$$

where

$$s_\psi = 6|\theta_{\ell\psi}(R)| \partial\theta_{\ell\psi}/\partial r|_R > 0 \quad (388)$$

$$s_\phi = 6|\theta_{\ell\phi}(R)| \partial\theta_{\ell\phi}/\partial r|_R > 0 \quad (389)$$

because  $\theta_{\ell\psi} < 0$  and  $\theta_{\ell\phi} < 0$ . Therefore

$$G_\psi(R-h) < G_\psi(R) < G_\psi(R+h) \quad (390)$$

$$G_\phi(R-h) < G_\phi(R) < G_\phi(R+h) \quad (391)$$

Thus  $G_\psi$  and  $G_\phi$  are increasing functions of height.

The Eötvös experiments done at the same height but in the east-west and north-south directions give the following difference obtained from equations



(370) and (371)

$$\begin{aligned}
 (G_\psi - G_\phi)/G &\sim 3(\theta_{\ell\phi}^2 - \theta_{\ell\psi}^2) & (392) \\
 &= 3[(\theta_r + \theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi})^2 - (\theta_r + \theta_\psi + \beta_{\psi,\psi})^2] \\
 &\sim 6\theta_r(\theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi} - \theta_\psi - \beta_{\psi,\psi}) \\
 &\sim 6\theta_r\theta_\phi
 \end{aligned}$$

Therefore

$$G_\phi > G_\psi \quad \theta_\phi > 0 \quad (393)$$

$$G_\phi < G_\psi \quad \theta_\phi < 0 \quad (394)$$

The measurement of the east-west/north-south asymmetry will give the sign (and approximate magnitude if  $\theta_r$  is known) of the internal phase angle of the angular coordinates.

In general the Eötvös experiment is done at the earth's surface and determines  $G_\psi(R)$  and  $G_\phi(R)$ , while the vertical gravity measurements using gravimeters are done down mine shafts and boreholes or up towers and determine  $G_r(R\pm h)$ . Consider a comparison of  $G_r(R\pm h)$  and  $G_\psi(R)$  which can be obtained using equations (355), (369) and (370)

$$\begin{aligned}
 [G_r(R\pm h) - G_\psi(R)]/G &\sim 3[\theta_{\ell\psi}^2(R) - \theta_r^2(R\pm h)] & (395) \\
 &= 3\{[\theta_r(R) + \theta_\psi(R) + \beta_{\psi,\psi}(R)]^2 - \theta_r^2(R\pm h)\}
 \end{aligned}$$

Combining equations (376), (377) and (395) gives

$$\begin{aligned}
 [G_r(R\pm h) - G_\psi(R)]/G &\sim \pm s_r h + 6\theta_r(R)[\theta_\psi(R) + \beta_{\psi,\psi}(R)] + 3[\theta_\psi(R) + \beta_{\psi,\psi}(R)]^2 \\
 &\sim \pm s_r h + 6\theta_r(R)[\theta_\psi(R) + \beta_{\psi,\psi}(R)] & (396)
 \end{aligned}$$

From equation (396) it follows that

$$\begin{aligned}
 [G_r(R) - G_\psi(R)]/G &\sim 6\theta_r(R)[\theta_\psi(R) + \beta_{\psi,\psi}(R)] + 3[\theta_\psi(R) + \beta_{\psi,\psi}(R)]^2 & (397) \\
 &\sim 6\theta_r(R)[\theta_\psi(R) + \beta_{\psi,\psi}(R)]
 \end{aligned}$$

It follows from equations (381), (396) and (397) that

$$G_r(R-h) < G_r(R) < G_r(R+h) < G_\psi(R) < G \quad \left. \vphantom{G_r(R-h)} \right\} \quad \theta_\psi(R) > 0 \quad (398)$$

$$G_\psi(R) < G_r(R-h) < G_r(R) < G_r(R+h) < G \quad \left. \vphantom{G_\psi(R)} \right\} \quad \theta_\psi(R) < 0 \quad (399)$$

The inequalities in equations (398) and (399) hold only for small  $h$ .

Now consider the case of the east-west oriented Eötvös experiment. Combining equations (359), (369) and (371) gives

$$\begin{aligned} [G_r(R\pm h) - G_\phi(R)]/G &\sim 3[\theta_{\ell\phi}^2(R) - \theta_r^2(R\pm h)] & (400) \\ &= 3\{[\theta_r(R) + \theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)]^2 - \theta_r^2(R\pm h)\} \end{aligned}$$

Combining equations (376), (377) and (400) gives

$$\begin{aligned} [G_r(R\pm h) - G_\phi(R)]/G &\sim \pm s_r h + 6\theta_r(R)[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)] & (401) \\ &+ 3[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)]^2 \\ &\sim \pm s_r h + 6\theta_r(R)[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)] \end{aligned}$$

From equation (401) it follows that

$$\begin{aligned} [G_r(R) - G_\phi(R)]/G &\sim 6\theta_r(R)[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)] & (402) \\ &+ 3[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)]^2 \\ &\sim 6\theta_r(R)[\theta_{s\psi}(R) + \theta_\phi(R) + \beta_{\phi,\phi}(R)] \end{aligned}$$

From equations (381), (401) and (402) it follows that

$$G_r(R-h) < G_r(R) < G_r(R+h) < G_\phi(R) < G \quad \left. \vphantom{G_r(R-h)} \right\} \theta_\phi(R) > 0, \theta_\psi(R) > 0 \quad (403)$$

$$G_\phi(R) < G_r(R-h) < G_r(R) < G_r(R+h) < G \quad \left. \vphantom{G_\phi(R)} \right\} \theta_\phi(R) < 0, \theta_\psi(R) < 0 \quad (404)$$

The inequalities in equations (403) and (404) hold only for small h.

Inequalities (398) and (403) are supported by the experimental data in References 5, 25 and 49 and suggest that mine shaft, borehole and tower determinations of the gravitational constant will be less than the gravitational constant determined by an Eötvös experiment performed at the earth's surface. On the other hand, the inequalities (399) and (404) are supported by the experimental data given in References 1 through 3 and indicate that the gravitational constant determined from a mine shaft, borehole or tower experiment will be larger than the value of the gravitational constant obtained by an Eötvös experiment conducted at the earth's surface. Only one set of data can be correct. When the correct set of experimental data is finally determined the proper signs and approximate magnitudes of  $\theta_\phi$  and  $\theta_\psi$  will be fixed. Neither mine shaft, borehole, tower or Eötvös experiments directly measure the constant G.

If it is possible to conduct an Eötvös experiment at various depths in a mine shaft or at different heights up a tower, it becomes important to compare  $G_r$  with  $G_\psi$  and  $G_\phi$  at the same depth or height. From equations (369) through (371) it follows that

$$\begin{aligned}
(G_r - G_\psi)/G &\sim 3(\theta_{\ell\psi}^2 - \theta_r^2) & (405) \\
&= 3[(\theta_r + \theta_\psi + \beta_{\psi,\psi})^2 - \theta_r^2] \\
&= 3[2\theta_r(\theta_\psi + \beta_{\psi,\psi}) + (\theta_\psi + \beta_{\psi,\psi})^2] \\
&\sim 6\theta_r(\theta_\psi + \beta_{\psi,\psi}) \\
&\sim 12\theta_r\theta_\psi
\end{aligned}$$

In a similar fashion

$$\begin{aligned}
(G_r - G_\phi)/G &\sim 3(\theta_{\ell\phi}^2 - \theta_r^2) & (406) \\
&= 3[(\theta_r + \theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi})^2 - \theta_r^2] \\
&= 3[2\theta_r(\theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi}) + (\theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi})^2] \\
&\sim 6\theta_r(\theta_{s\psi} + \theta_\phi + \beta_{\phi,\phi}) \\
&\sim 18\theta_r\theta_\phi
\end{aligned}$$

The inequalities (372) and (373) can also be deduced from equations (405) and (406).

The variation of  $G_r$ ,  $G_\phi$  and  $G_\psi$  with depth in a mine shaft and borehole or with height up a tower is due to Newtonian gravity in broken symmetry space combined with the variation of the broken symmetry atmospheric pressure. The variation of the broken symmetry atmospheric pressure with radial distance induces a variation of  $\theta_r$ ,  $\theta_\phi$  and  $\theta_\psi$  with radial distance (Figure 4) and this determines the non-Newtonian variation of  $G_r$ ,  $G_\phi$  and  $G_\psi$  according to equations (369) through (371). This apparent non-Newtonian behaviour of gravity has been interpreted as being due to the existence of graviscalar (spin 0) and graviphoton (spin 1) component forces of gravity (the "fifth" and "sixth" forces).<sup>1-4, 13-25</sup> The hypothetical graviscalar is an attractive force while the hypothetical graviphoton mediates a repulsive force, and both are described by finite range Yukawa terms that are added to the ordinary Newtonian potential. But in fact these hypothetical forces are not required to describe the experimental results. The apparent non-Newtonian behaviour is due to ordinary Newtonian gravity in matter and space whose pressure and coordinate fields exhibit broken internal symmetries. The relative magnitudes of  $G_r(R\pm h)$  and  $G_\psi(R)$  or  $G_\phi(R)$  are not related to new gravitation forces but rather to the broken symmetry of pressure and spatial coordinates.

6. NUMERICAL VALUES OF THE INTERNAL PHASE ANGLES. This section determines numerical values of  $\theta_r$ ,  $\theta_\phi$ ,  $\theta_\psi$  and  $\theta_p$  within the atmosphere in the vicinity of the earth's surface. Two methods are used. The discrepancy between the measured and predicted values of the gravitational red shift of  $\gamma$ -rays in the Pound-Rebka-Snider experiment.<sup>32, 33, 46-48</sup> The second method is based on the measurement of the apparent departure of the force of gravity from Newtonian behaviour in mine shaft, borehole and tower experiments.<sup>1-3, 5, 25, 49</sup>

### A. Measurement of the Gravitational Red Shift

The experiments of Pound, Rebka and Snider measured the gravitational red shift of a  $\gamma$ -ray falling in the earth's gravitational field. The conventional expression for the red shift in frequency is given by<sup>32,33</sup>

$$z_c = (\Delta\nu/\nu)_c = [V(r_{2m}) - V(r_{1m})]/c^2 \quad (407)$$

where the conventionally calculated gravitational potential is written as<sup>28</sup>

$$V(r_m) = -GM/r_m \quad (408)$$

and therefore

$$z_c = GM/c^2(1/r_{1m} - 1/r_{2m}) \quad (409)$$

In this paper the theory of coordinates with internal phase requires a complex number red shift given by

$$\begin{aligned} \bar{z} &= ze^{j\theta} z = \Delta\bar{v}/\bar{v} = [\bar{v}(\bar{r}_2) - \bar{v}(\bar{r}_1)]/c^2 \\ &= GM/c^2(1/\bar{r}_1 - 1/\bar{r}_2) \end{aligned} \quad (410)$$

The measured gravitational red shift is given by the real part of equation (410)

$$z_m = z_R = z \cos \theta_z = GM/c^2(1/r_1 \cos \theta_{r1} - 1/r_2 \cos \theta_{r2}) \quad (411)$$

while the imaginary part of equation (430) is

$$z_I = z \sin \theta_z = -GM/c^2(1/r_1 \sin \theta_{r1} - 1/r_2 \sin \theta_{r2}) \quad (412)$$

where

$$\bar{r}_1 = r_1 e^{j\theta_{r1}} \quad (413)$$

$$\bar{r}_2 = r_2 e^{j\theta_{r2}} \quad (414)$$

Because  $r_{1m} = r_1 \cos \theta_{r1}$  and  $r_{2m} = r_2 \cos \theta_{r2}$  it follows from equation (411) that

$$z_m = GM/c^2(1/r_{1m} \cos^2 \theta_{r1} - 1/r_{2m} \cos^2 \theta_{r2}) \quad (415)$$

The difference between the measured and conventionally calculated gravitational red shift is obtained from equations (409) and (415) to be

$$\begin{aligned} z_c - z_m &= GM/c^2(1/r_{1m} \sin^2 \theta_{r1} - 1/r_{2m} \sin^2 \theta_{r2}) \\ &\sim GM/c^2(\theta_{r1}^2/r_{1m} - \theta_{r2}^2/r_{2m}) \\ &\sim \theta_r^2 z_c \end{aligned} \quad (416)$$

Therefore approximately

$$\theta_r^2 \sim (z_c - z_m)/z_c \sim 0.01 \quad (417)$$

where according to References 47 and 48 the fractional difference between the calculated and measured gravitational red shift is 1%. From equation (417) it follows that

$$\theta_r \sim -0.1 \text{ rad} = -5.7^\circ \quad (418)$$

In this way a value of  $\theta_r$  is obtained from the measurement of the gravitational red shift. This laboratory value of  $\theta_r$  is probably more accurate than the corresponding values of  $\theta_r$  that may possibly be obtained from measurements of the apparent non-Newtonian variation of gravity in mine shafts, boreholes and towers. The value of  $\theta_r$  given in equation (418) will be used to obtain values  $\theta_\phi$  and  $\theta_\psi$  in section B.

#### B. Analysis of the Apparent Non-Newtonian Gravity Measurements.

Conflicting experimental data have been presented for the values of the gravitational constant derived from measurements of the variation of the force of gravity with distance in mine shafts, boreholes and towers. According to References 1 through 3, the values of the gravitational constant derived from mine shaft gravity variations are larger than those derived from Eötvös experiments conducted at the earth's surface. On the other hand, References 25 and 49 indicates that borehole measurements in the ice of a glacier produce values of the gravitational constant that are smaller than the values of the gravitational constant derived from Eötvös experiments performed at the surface of the earth. In addition, Reference 5 indicates that the gravitational constant derived from gravity measurements on a tower is smaller than that measured by the Eötvös experiments at the earth's surface. The experimental results given in References 1 through 3 are in conflict with the experimental results of References 5, 25 and 49. Therefore the numerical calculations in this section are done for each situation. According to the theory of Newtonian gravity in matter and vacuum with broken internal symmetries the discrepancies between the measured and conventionally predicted values of the force of gravity in mine shaft, borehole, tower and Eötvös experiments are related to the values of  $\theta_r$ ,  $\theta_\psi$  and  $\theta_\phi$ .

Two cases will be examined in this section according to the relative magnitudes of the internal phase angles  $|\theta_r|$ ,  $|\theta_\psi|$  and  $|\theta_\phi|$ .

$$\text{Case 1: } |\theta_r(R)| \gg |\theta_\psi(R)| \text{ and } |\theta_r(R)| \gg |\theta_\phi(R)|$$

Combining equations (392), (396) and (401) gives for small  $h/R$

$$[G_\psi(R) - G_\phi(R)]/G \sim 6xy \quad (419)$$

$$[G_r(R+h) - G_\psi(R)]/G \sim 12xy \quad (420)$$

$$[G_r(R+h) - G_\phi(R)]/G \sim 18xy \quad (421)$$

where

$$x = \theta_r(R) < 0 \quad (422)$$

$$y = \theta_\psi(R) \sim \beta_{\psi,\psi}(R) \sim \theta_{s\psi}(R) \sim \theta_\phi(R) \sim \beta_{\phi,\phi}(R) \quad (423)$$

so that  $|x| \gg |y|$ . Because  $x < 0$  it is the sign of  $y$  that determines the signs of the expressions in equations (419) through (421). The expressions in equations (420) and (421) can be either positive or negative, so that they can describe either the experimental results of References 1 through 3 (which requires  $y < 0$ ) or the experimental results of References 5, 25 and 49 (which requires  $y > 0$ ). Because the value of  $x = \theta_r(R)$  is known from equation (418) the determination of any one of the three differences in equations (419) through (421) would immediately determine the value for  $y$ .

Consider first the determination of the gravitational constant from mine shafts which is about 0.6% larger than the value obtained from Eötvös experiments performed at the earth's surface.<sup>1-3</sup> Using the average of equations (420) and (421) (because of the unspecified orientation of the Eötvös experiments) yields

$$15xy = 15(-0.1)y = 0.006 \quad (424)$$

where  $x = -0.1$  was obtained from equation (418). Equation (424) gives

$$x = \theta_r(R) = -0.1 \text{ rad} = -5.7^\circ \quad (425)$$

$$y = \theta_\psi(R) = \theta_\phi(R) = -0.004 \text{ rad} = -0.23^\circ$$

From equation (419) the north-south/east-west asymmetry of the Eötvös experiment is given by

$$[G_\psi(R) - G_\phi(R)]/G \sim +0.0024 \quad (426)$$

Borehole data from a Greenland glacier shows that the derived gravitational constant is about 2.8% smaller than the value of the gravitational constant derived from Eötvös experiments done at the earth's surface.<sup>25,49</sup> Also, measurements of the variation of the gravity force up a tower gives results for the gravitational constant that are 2.0% smaller than that obtained from Eötvös experiments at the surface of the earth.<sup>5</sup> Therefore using the average of the results of References 5 and 49 with the average of equations (420) and (421) gives

$$15xy = 15(-0.1)y = -0.024 \quad (427)$$

where again  $x = -0.1$  was obtained from equation (418). Then equation (427) yields

$$x = \theta_r(R) = -0.1 \text{ rad} = -5.7^\circ \quad (428)$$

$$y = \theta_\psi(R) = \theta_\phi(R) = +0.016 \text{ rad} = +0.92^\circ$$

Equation (419) gives the north-south/east-west asymmetry of the Eötvös experiment to be

$$[G_\psi(R) - G_\phi(R)]/G \sim -0.0096 \quad (429)$$

An independent determination of the north-south/east-west asymmetry of the Eötvös experiment would immediately determine the signs and values of  $\theta_\psi$  (and  $\theta_\phi$ ).

$$\text{Case 2: } \theta_r(R) = \theta_\psi(R) = \theta_\phi(R) = \theta_{s\psi}(R) = \beta_{\psi,\psi}(R) = \beta_{\phi,\phi}(R)$$

Combining equations (392), (395) and (400) gives for small  $h/R$

$$[G_\psi(R) - G_\phi(R)]/G \sim 3[(4w)^2 - (3w)^2] = 21w^2 \quad (430)$$

$$[G_r(R\pm h) - G_\psi(R)]/G \sim 3[(3w)^2 - w^2] = 24w^2 \quad (431)$$

$$[G_r(R\pm h) - G_\phi(R)]/G \sim 3[(4w)^2 - w^2] = 45w^2 \quad (432)$$

$$\text{where } w = \theta_r(R) = \theta_\psi(R) = \theta_\phi(R) = \theta_{s\psi}(R) = \beta_{\psi,\psi}(R) = \beta_{\phi,\phi}(R) < 0 \quad (433)$$

Thus  $w < 0$  because  $\theta_r(R) < 0$ . The expressions in equations (431) and (432) are always positive and therefore Case 2 agrees only with the experimental data given in References 1 through 3. Using the average of equations (431) and (432) (because of the unspecified orientation of the Eötvös experiments) and the 0.006 positive fractional difference between the values of the gravitational constant measured in a mine shaft and by an Eötvös experiment performed at the surface of the earth given by References 1 through 3 yields

$$34.5w^2 = 0.006 \quad (434)$$

$$w = \theta_r(R) = \theta_\psi(R) = \theta_\phi(R) = -0.76^\circ$$

where  $w$  is given by equation (433). Equation (430) gives the north-south/east-west asymmetry of the Eötvös experiment as

$$[G_\psi(R) - G_\phi(R)]/G \sim 0.0037 \quad (435)$$

The predicted value  $\theta_r(R) = -0.76^\circ$  is much less in magnitude than the value  $\theta_r(R) = -5.7^\circ$  predicted by the Pound-Rebka-Snider experiment. Therefore Case 2 as represented in the assumption given in equation (433) may not be physically realistic.

7. CONCLUSION. Newtonian gravity in space and time with broken internal symmetries produces an apparent non-Newtonian behaviour of the acceleration of gravity, and the gravitational constant varies with the radial distance from the center of a planet. This is due to the fact that the pressure and coordinates in matter (and vacuum) exhibit broken symmetries that are represented by internal phase angles which vary with radial distance. The measured apparent non-Newtonian gravity effects are therefore due to the variation of the atmospheric pressure in mine shafts and boreholes and on towers, and this introduces an apparent 7 km finite range force component. New forces in addition to Newtonian gravitation are not required to explain the experimental observations. The values of the internal phase angles of the coordinates can be obtained from the Pound-Rebka-Snider gravitational red shift experiment, the measurements of the apparent non-Newtonian gravity field, and the Eötvös experiments. The internal phase angles of space and time will influence the basic calculations of astrophysics and geophysics.

#### ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

#### REFERENCES

1. 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.
2. Moore, G. I., Stacey, F. D., Tuck, G. J., Goodwin, B. D., Linthorne, N. P., Barton, M. A., Reid, D. M. and Agnew, G. D., "Determination of the Gravitational Constant at an Effective Mass Separation of 22m," Phys. Rev. D, Vol. 38, No. 4, 15 Aug. 1988, p. 1023.
3. Stacey, F. D., Tuck, G. J. and Moore, G. I., "Quantum Gravity: Observational Constraints on a Pair of Yukawa Terms," Phys. Rev. D, Vol. 36, No. 8, 15 Oct. 1987, p. 2374.
4. Bartlett, D. F. and Tew, W. L., "Comment on Quantum Gravity: Observational Constraints on a Pair of Yukawa Terms," Phys. Rev. D, Vol. 38, No. 12, 15 Dec. 1988, p. 3843.
5. Eckhardt, D. H., Jekeli, C., Lazarewicz, A. R., Romaides, A. J. and Sands, R. W., "Tower Gravity Experiment: Evidence for Non-Newtonian Gravity," Phys. Rev. Lett., Vol. 60, No. 25, 20 June 1988, p. 2567.
6. Fischbach, E., Sudarsky, D., Szafer, A., Talmadge, C. and Aronson, S. H., "Reanalysis of the Eötvös Experiment," Phys. Rev. Lett., Vol. 56, No. 1, 6 Jan. 1986, p. 3.
7. Thieberger, P., Phys. Rev. Lett., Vol. 58, 1987, p. 1066.
8. Stubbs, C. W., Adelberger, E. G., Raab, F. J., Gundlach, J. H., Heckel, B. R., McMurry, K. D., Swanson, H. E. and Watanabe, R., Phys. Rev. Lett., Vol. 58, 1987, p. 1070.
9. Adelberger, E. G., Stubbs, C. W., Rogers, W. F., Raab, F. J., Heckel, B. R., Gundlach, J. H., Swanson, H. E. and Watanabe, R., Phys. Rev. Lett., Vol. 59, 1987, p. 849.
10. Niebauer, T. M., McHugh, M. P. and Faller, J. E., Phys. Rev. Lett., Vol. 59, 1987, p. 609.
11. Boynton, P. E., Crosby, D., Ekstrom, P. and Szumilo, A., Phys. Rev. Lett., Vol. 59, 1987, p. 1385.
12. 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, No. 18, 2 May 1988, p. 1801.



13. Speake, C. C. and Quinn, T. J., "Search for a Short-Range Isospin-Coupling Component of the Fifth Force with Use of a Beam Balance," Phys. Rev. Lett., Vol. 61, No. 12, 19 Sep. 1988, p. 1340.
14. Fischbach, E., Kloor, H. T., Talmadge, C., Aronson, S. H. and Gillies, G. T., "Possibility of Shielding the Fifth Force," Phys. Rev. Lett., Vol. 60, 4 Jan. 1988, p. 74.
15. Cowsik, R., Krishnan, N., Tandon, S. N. and Unnikrishnan, C. S., "Limit on the Strength of Intermediate-Range Forces Coupling to Isospin," Phys. Rev. Lett., Vol. 61, 7 Nov. 1988, p. 2179.
16. Damour, T., Gibbons, G. W. and Taylor, J. H., "Limits on the Variability of G Using Binary-Pulsar Data," Phys. Rev. Lett., Vol. 61, No. 10, 5 Sep. 1988, p. 1151.
17. Talmadge, C., Berthias, J. P., Hellings, R. W. and Standish, E. M., "Model-Independent Constraints on Possible Modifications of Newtonian Gravity," Phys. Rev. Lett., Vol. 61, No. 10, 5 Sep. 1988, p. 1159.
18. Burgess, C. P. and Cloutier, J., "Astrophysical Evidence for a Weak New Force," Phys. Rev. D, Vol. 38, No. 10, 15 Nov. 1988, p. 2944.
19. Goldman, T., Hughes, R. J. and Nieto, M. M., Phys. Lett. B, Vol. 171, 1986, p. 217.
20. Nieto, M. M., Goldman, T. and Hughes, R. J., "Phenomenological Aspects of New Gravitational Forces. IV. New Terrestrial Experiments," Phys. Rev. D, Vol. 38, No. 10, 15 Nov. 1988, p. 2937.
21. Ander, M. E., Goldman, T., Hughes, R. J. and Nieto, M. M., "Possible Resolution of the Brookhaven and Washington Eötvös Experiments," Phys. Rev. Lett., Vol. 60, No. 13, 28 March 1988, p. 1225.
22. Niebauer, T. M., Faller, J. E. and Bender, P. L., "Comment on Possible Resolution of the Brookhaven and Washington Eötvös Experiments," Phys. Rev. Lett., Vol. 61, 7 Nov. 1988, p. 2272.
23. Stubbs, C. W., Adelberger, E. G. and Gregory, E. C., "Constraints of Proposed Spin-0 and Spin-1 Partners of the Graviton," Phys. Rev. Lett., Vol. 61, 21 Nov. 1988, p. 2409.
24. Schwarzschild, B., "From Mine Shafts to Cliff's - The 'Fifth Force' Remains Elusive," Physics Today, July 1988, p. 21.
25. Pool, R., "Was Newton Wrong?," Science, Vol. 241, 12 Aug. 1988, p. 789.
26. Weiss, R. A., "Scale Invariant Equations for Relativistic Waves," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 87-1, 27-30 May 1986, p. 307.

27. 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.
28. Weiss, R. A., "The Broken Symmetry of Space and Time in Bulk Matter and the Vacuum," Sixth Army Conference on Applied Mathematics and Computing, Boulder, Colorado, ARO 89-1, 31 May-3 June 1988, p. 317.
29. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.
30. Chandrasekhar, S., "On Stars, Their Evolution and Their Stability," Revs. Mod. Phys., Vol. 56, April 1984, p. 137.
31. Tolman, R. C., Relativity Thermodynamics and Cosmology, Oxford, Clarendon Press, London, 1934.
32. Misner, C. W., Thorne, K. S. and Wheeler, J. A., Gravitation, W. H. Freeman and Company, San Francisco, 1973.
33. Weinberg, S., Gravitation and Cosmology, John Wiley, New York, 1972.
34. Witten, E., "Space-Time and Topological Orbifolds," Phys. Rev. Lett., Vol. 61, No. 6, 8 Aug. 1988, p. 670.
35. Stacey, F. D., Physics of the Earth, John Wiley, New York, 1977.
36. Garland, G. D., The Earth's Shape and Gravity, Pergamon Press, New York, 1965.
37. Jeffreys, H., The Earth, Cambridge University Press, New York, 1962.
38. Spencer Jones, H., "Dimensions and Rotation," article in The Earth as a Planet, edited by Kuiper, G. P., University of Chicago Press, Chicago, 1954.
39. Officer, C. B., Introduction to Theoretical Geophysics, Springer-Verlag, New York, 1974.
40. King-Hele, D. G., "The Earth's Gravitational Potential, Deduced from the Orbits of Artificial Satellites," article in The Earth Today, edited by Jeffreys, H., Interscience, New York, 1961.
41. Murphy, G. M., Ordinary Differential Equations and Their Solutions, Van Nostrand, New York, 1960.
42. Petit Bois, G., Tables of Indefinite Integrals, Dover, New York, 1961.
43. Humphreys, W. J., Physics of the Air, McGraw-Hill, New York, 1929.
44. Gutenberg, B., Physics of the Earth's Interior, Academic Press, New York, 1959.

45. Bullard, E., "The Interior of the Earth," Article in The Earth as a Planet, edited by Kuiper, G., University of Chicago Press, Chicago, 1954.

46. Pound, R. V. and Rebka, G. A., "Apparent Weight of Photons," Phys. Rev. Lett., Vol. 4, 1960, p. 337.

47. Pound, R. V. and Snider, J. L., "Effect of Gravity on Nuclear Resonance," Phys. Rev. Lett., Vol. 13, 1964, p. 539.

48. Pound, R. V. and Snider, J. L., "Effect of Gravity on Gamma Radiation," Phys. Rev. B, Vol. 140, 1965, p. 788.

49. Ander, M. E., Zumberge, M. A., Lautzenhiser, T., Parker, R. L., Aiken, C.L., Gorman, M. R., Nieto, M. M., Cooper, A. P., Ferguson, J. F., Fisher, E., McMechan, G. A., Sasagawa, G., Stevenson, J. M., Backus, G., Chave, A. D., Greer, J., Hammer, P., Hansen, B. L., Hildebrand, J. A., Kelty, J. R., Sidles, C. and Wirtz, J., "Test of Newton's Inverse-Square Law in the Greenland Ice Cap," Phys. Rev. Lett., Vol. 62, No. 9, 27 Feb. 1989.

ERRATA: Reference 28

equation (384)  $-\cos \beta_{r,r} (\partial \bar{P} / \partial r + dt/dr \partial \bar{P} / \partial t) - \rho \partial \bar{W} / \partial \bar{r}$

equation (385)  $-\cos \beta_{\phi,\phi} (1/r \partial \bar{P} / \partial \phi + 1/r dt/d\phi \partial \bar{P} / \partial t) - \rho / \bar{r} \partial \bar{W} / \partial \bar{\phi}$

equation (389)  $-\cos \beta_{x,x} (\partial \bar{P} / \partial x + dt/dx \partial \bar{P} / \partial t) - \rho \partial \bar{W} / \partial \bar{x}$

equation (393)  $\cos \beta_{r,r} \partial \bar{P} / \partial r = D_p e^{j\phi_p}$

equation (396)  $\phi_p = \theta_p + \beta_{p,r}$

equation (408)  $\theta_p + \beta_{p,r} = \theta_w + \beta_{w,r} - \theta_r - \beta_{r,r} + \pi$

equation (425)  $\theta_p + \beta_{p,r} = -2\theta_r + \pi$

equation (426)  $\theta_p + \tan^{-1} \left( p \frac{\partial \theta_p / \partial r}{\partial \bar{P} / \partial r} \right) = -2\theta_r + \pi$

equation (428)  $\theta_p + \beta'_{p,r} = -2\theta_r$

equation (431)  $\cos \beta_{r,r} \partial \bar{P} / \partial r = -GM_0 / \bar{r}^2$

equation (451)  $\cos \beta_{r,r} \partial \rho / \partial r = -G_0^2 M / (\bar{K}_S \bar{r}^2) = -G_0 M / (\bar{v}_S^2 \bar{r}^2)$

equation (456)  $\theta_{vS} = -\theta_r$

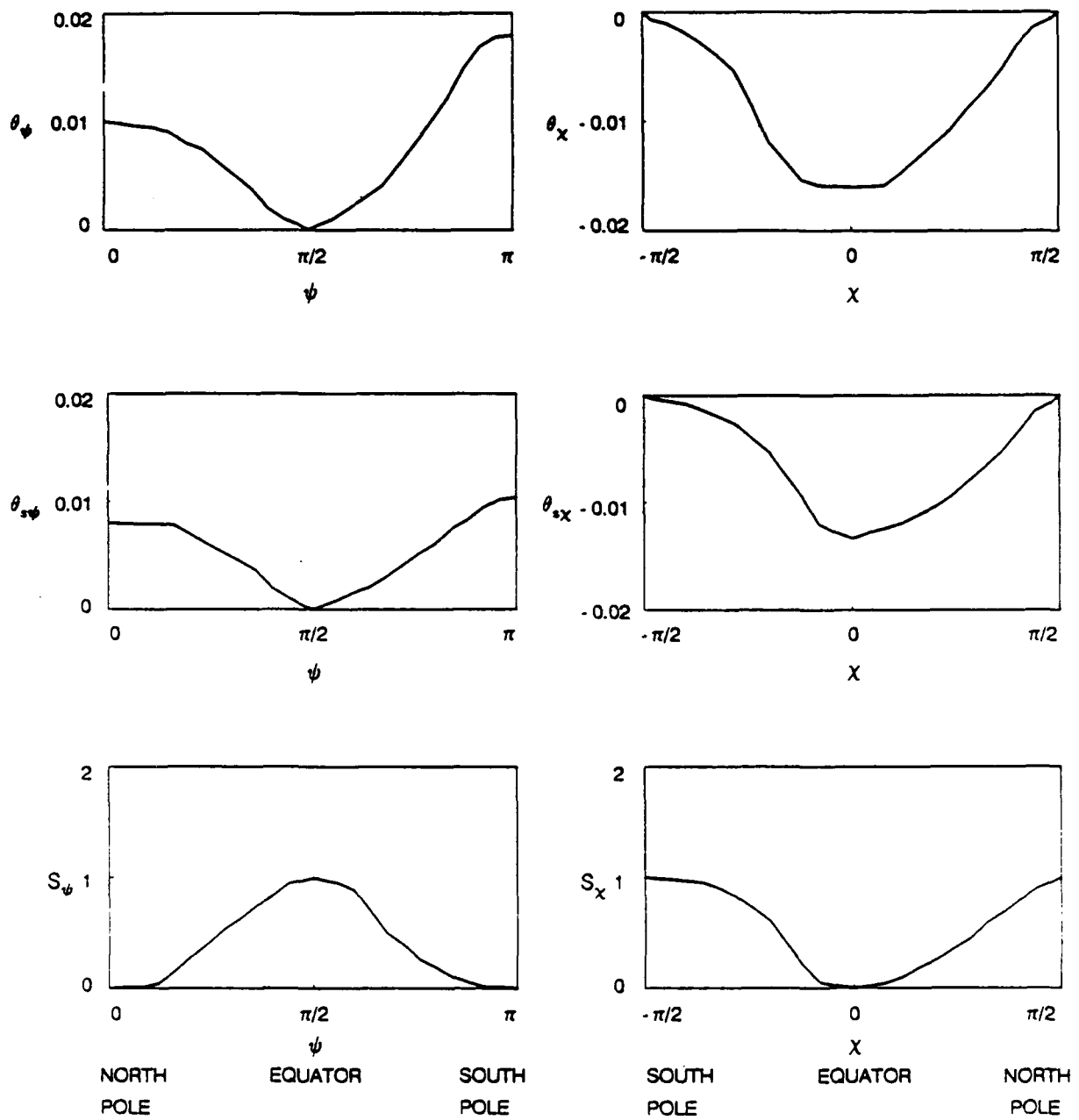


Figure 1. Sketch showing the dependence of  $\theta_\psi$ ,  $\theta_{s\psi}$  and  $S_\psi$  on zenith angle  $\psi$ , and the dependence of  $\theta_\chi$ ,  $\theta_{s\chi}$  and  $S_\chi$  on the latitude  $\chi$ . The experimental situation is not yet clear and it may be that  $\theta_\psi < 0$  and  $\theta_\chi > 0$ .

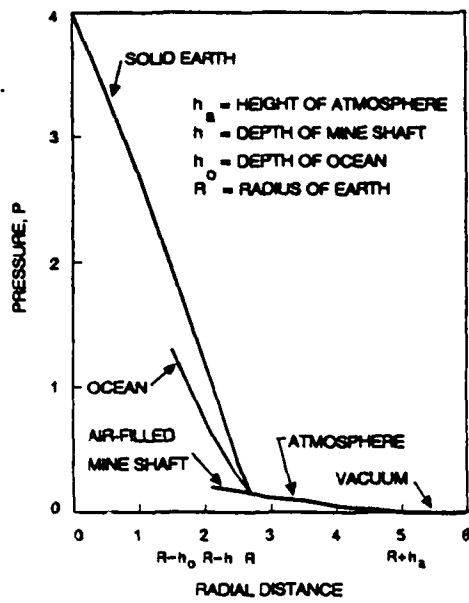


Figure 2. Sketch showing variation of pressure with radial distance for the atmosphere, ocean and solid earth (not to scale).

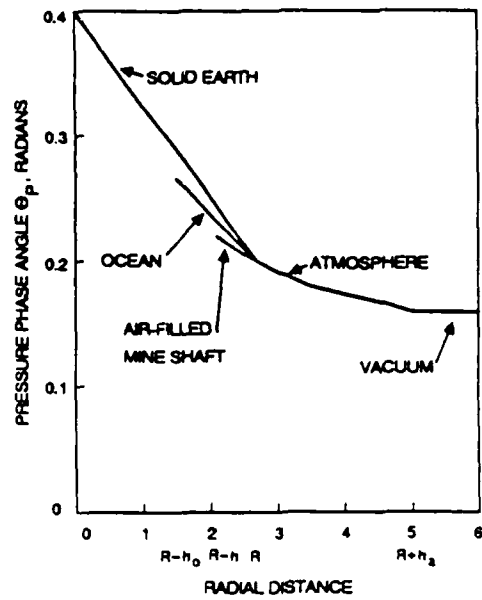


Figure 3. Sketch showing variation of the internal phase angle of the pressure with radial distance for the atmosphere, ocean and solid earth (not to scale).

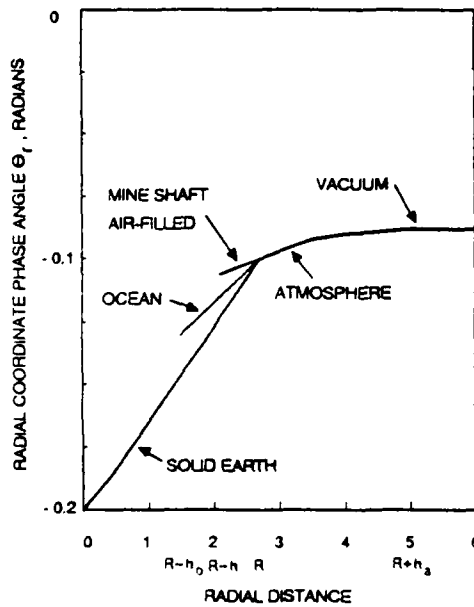


Figure 4. Sketch showing variation of the internal phase angle of the radial coordinate with radial distance for the atmosphere, ocean and solid earth (not to scale).

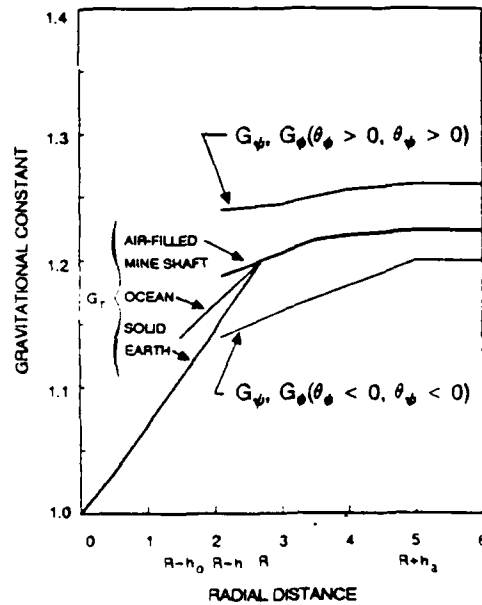


Figure 5. Sketch showing the variation of  $G_r$ ,  $G_\psi$  and  $G_\phi$  with radial distance. Two possible cases for  $G_\psi$  and  $G_\phi$  are shown (not to scale).

## WAVE PROPAGATION IN ASYMMETRIC MEDIA

Richard A. Weiss  
U. S. Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi 39180

**ABSTRACT.** The coordinates of space and time have broken internal symmetries for a region of spacetime located in a pressure field and perhaps even for the vacuum. Geometrical angles themselves have internal phase angles. A wave propagating in matter or the vacuum with broken internal symmetries will exhibit internal phase angles in its amplitude and dispersion characteristics. Cylindrical and spherical wave propagation in asymmetric matter is treated and the solution of the wave equation with broken internal symmetries is obtained. The observed periodicities of waves in measured time and measured space requires the propagation constants to be complex numbers but the phase must be a real number. A pressure field is associated with a broken internal symmetry, and therefore waves propagating in matter under pressure are expected to exhibit broken symmetry effects in the propagation parameters. Applications to acoustic and seismic waves are suggested.

**1. INTRODUCTION.** Matter and radiation exists within the continuum of spacetime, and it has been suggested that spacetime imprints measurable effects on the properties of bulk matter and radiation. These effects have been calculated by the development of a gauge theory of relativistic thermodynamics.<sup>1</sup> The effects of spacetime structure on matter and radiation occur in two ways, the first is by the effects of the Grüneisen parameter and bulk modulus which enter the relativistic trace equation as a requirement of gauge invariance.<sup>1</sup> The second way the metric of spacetime affects the state equation of matter and radiation is by requiring the thermodynamic functions such as pressure and internal energy to exhibit broken internal symmetries.<sup>2</sup> At the same time the coordinates of points located within matter, radiation or the vacuum also have broken internal symmetries and the internal phase angles of the coordinates must be determined simultaneously with the internal phase angles of the thermodynamic functions.<sup>3</sup>

All physical phenomena occurring within matter, radiation or the vacuum are affected by the broken symmetries of space and time. Electromagnetic and mechanical waves are expected to exhibit the effects of broken spacetime symmetry in both the wave amplitude and dispersion equation. The wave amplitude, wavelength and frequency are characterized by internal phase angles and therefore must be represented as complex numbers. The speed of sound and electromagnetic waves in matter must be represented as complex numbers, but the light speed in vacuum is a real number.

The broken symmetry of the pressure in matter or vacuum is derived from a relativistic trace equation.<sup>2</sup> In bulk matter or vacuum the space and time coordinates are complex numbers and are written as follows<sup>3</sup>

$$\bar{t} = te^{j\theta t} \quad (1)$$

for the time, while the cartesian space coordinates are

$$\bar{x} = xe^{j\theta x} \quad \bar{y} = ye^{j\theta y} \quad \bar{z} = ze^{j\theta z} \quad (2)$$

the cylindrical coordinates are

$$\bar{r} = re^{j\theta r} \quad \bar{\phi} = \phi e^{j\theta \phi} \quad \bar{z} = ze^{j\theta \psi} \quad (3)$$

and for spherical coordinates

$$\bar{r} = re^{j\theta r} \quad \bar{\phi} = \phi e^{j\theta \phi} \quad \bar{\psi} = \psi e^{j\theta \psi} \quad (4)$$

The measured time coordinate is just the real part of the complex number time<sup>3</sup>

$$t_m = t \cos \theta_t \quad (5)$$

while the measured space coordinates are given by

$$x_m = x \cos \theta_x \quad y_m = y \cos \theta_y \quad z_m = z \cos \theta_z \quad (6)$$

$$r_m = r \cos \theta_r \quad \phi_m = \phi \cos \theta_\phi \quad \psi_m = \psi \cos \theta_\psi \quad (7)$$

where  $t_m$  = measured time and  $x_m$ ,  $y_m$ ,  $z_m$ ,  $r_m$ ,  $\phi_m$  and  $\psi_m$  = measured space coordinates.

This paper considers the solution of a complex number wave equation whose space and time coordinates have broken internal symmetries. Section 2 considers the time dependence of periodic waves in broken symmetry matter where the periodicity occurs in the real part (the measured part) of the complex number time coordinate. Section 3 considers cylindrical waves with broken internal symmetry and it is shown that the azimuthal angle equation has a complex number separation constant. The real part of the complex number azimuthal angle has the  $0 \rightarrow 2\pi$  symmetry. The remaining coordinate equations also have complex number separation constants which are determined by the requirement that the wave periodicity occurs in the real parts of the complex number radial and z coordinates. Section 4 considers spherical wave propagation in asymmetric matter, and develops the equations describing the conditions of periodicity in the real parts of the radial, azimuthal and zenith angle coordinates.

**2. PERIODIC VIBRATIONS IN SPACETIME WITH BROKEN INTERNAL SYMMETRIES.** This section determines the relationship between the measured period and frequency from the experimental observation that waves and vibrations are periodic in measured space and time coordinates. The equation describing the time dependence of a periodic phenomena in space and time with broken internal symmetries is written as a generalization of the standard scalar equation<sup>4-7</sup>

$$d^2\bar{e}/d\bar{t}^2 + \bar{\omega}^2\bar{e} = 0 \quad (8)$$

whose solution is

$$\bar{e} = \bar{e}_0 e^{i\bar{\omega}\bar{t}} \quad (9)$$

where  $\bar{t}$  is given by equation (1) and where

$$\bar{\omega} = \omega e^{j\theta_\omega} = 2\pi/\bar{T} = 2\pi/T e^{-j\theta_T} \quad (10)$$

and therefore

$$\omega = 2\pi f = 2\pi/T \quad \theta_\omega = -\theta_T \quad (11)$$

where  $\bar{\omega}$  = complex number angular frequency whose magnitude and phase are  $\omega$  and  $\theta_\omega$  respectively, and  $\bar{T}$  = complex number period whose magnitude and phase are  $T$  and  $\theta_T$  respectively.

The requirement that equation (9) represents a periodic wave in measured time implies that  $\bar{\omega}\bar{t}$  is a real number. This can be seen by first writing  $\bar{\omega}$  and  $\bar{t}$  as complex numbers as follows

$$\bar{\omega} = \omega_R + j\omega_I = \omega(\cos \theta_\omega + j \sin \theta_\omega) \quad (12)$$

$$\bar{t} = t_R + jt_I = t(\cos \theta_t + j \sin \theta_t) \quad (13)$$

then

$$\bar{\omega}\bar{t} = \omega_R t_R - \omega_I t_I + j(\omega_I t_R + \omega_R t_I) \quad (14)$$

The reality condition gives

$$t_I = -\omega_I t_R / \omega_R \quad (15)$$

and therefore

$$\bar{\omega}\bar{t} = t_R \omega^2 / \omega_R = t_m \omega_m^2 / \omega_m = \omega t \quad (16)$$

where  $\omega_m = \omega_R =$  measured angular speed. Therefore the reality of the phase  $\bar{\omega}\bar{t}$  requires that the phase be linear in both the measured time  $t_m$  and the time magnitude  $t$ . The fact that the phase is linear in the measured time agrees with the experimental fact that vibrating systems are periodic in the measured time coordinate.

The measured period  $T_m = T_R = T \cos \theta_T$  is obtained from equations (12) and (16) to be

$$T_R \omega^2 / \omega_R = T_R \omega / \cos \theta_\omega = \omega T = 2\pi \quad (17)$$



and therefore

$$T_R = 2\pi/\omega \cos \theta_\omega = 2\pi/\omega_R \cos^2 \theta_\omega = 1/f_R \cos^2 \theta_\omega \quad (18)$$

where  $f_R = \omega_R/(2\pi)$  = real part of the frequency. Therefore the relationship between the measured period and the measured angular frequency is

$$T_m = 2\pi/\omega_m \cos^2 \theta_\omega = 2\pi/\omega \cos \theta_\omega = 1/f \cos \theta_\omega \quad (19)$$

or for the measured frequency and the measured period

$$f_m = 1/T_m \cos^2 \theta_\omega = 1/T_m \cos^2 \theta_T \quad (20)$$

Note that  $f_m \neq 1/T_m$ . From equations (10), (11) and (16) it follows that

$$\theta_\omega = \theta_{\bar{f}} = -\theta_T = -\theta_t \quad (21)$$

Periodicity requires the internal phase of the frequency to adjust itself so as to satisfy equation (21). Combining equations (11), (16), (19) and (21) gives

$$T = 1/f \quad (22)$$

$$t/f = t_m/f_m \quad (23)$$

$$tf = t/T = t_m/T_m \quad (24)$$

Thus the phase in equation (16) can be written as

$$\bar{\omega}t = \omega t = 2\pi t_m/T_m = \omega_m t_m / \cos^2 \theta_\omega \quad (25)$$

and

$$f_R/T_R = f_m/T_m = f^2 \quad (26)$$

$$t_m^2 / (f_m T_m) = t^2 \quad (27)$$

The phase  $\bar{\omega}t$  has a period  $T$  when expressed in terms of  $t$ , and a period  $T_m$  when expressed in terms of the measured time  $t_m$ . The general solution of equation (8) is

$$\begin{aligned} \bar{c} &= \bar{A}e^{i2\pi t_m/T_m} + \bar{B}e^{-i2\pi t_m/T_m} \\ &= \bar{A}e^{i2\pi t/T} + \bar{B}e^{-i2\pi t/T} \end{aligned} \quad (28)$$

The conclusions for broken symmetry space and time that  $f_m$  and  $T_m$  are related by equation (20) and that  $f_m \neq 1/T_m$  may possibly be experimentally verified if  $f_m$  and  $T_m$  can be independently measured for the same periodic phenomenon. Finally,

the conclusions of this section are based on the observed fact that periodic physical systems have definite periods in measured time.

3. CYLINDRICAL WAVES IN ASYMMETRIC MATTER. The wave equation for cylindrical waves in matter with broken internal symmetry is written as a generalization of the standard scalar wave equation as follows<sup>4-7</sup>

$$\partial^2 \bar{u} / \partial \bar{r}^2 + 1/\bar{r} \partial \bar{u} / \partial \bar{r} + 1/\bar{r}^2 \partial^2 \bar{u} / \partial \bar{\phi}^2 + \partial^2 \bar{u} / \partial \bar{z}^2 = 1/\bar{v}^2 \partial^2 \bar{u} / \partial \bar{t}^2 \quad (29)$$

where  $\bar{u}$  = wave amplitude with internal phase, and  $\bar{v}$  = complex number phase velocity. Equation (29) can be solved by the standard technique of separation of variables<sup>4-7</sup>

$$\bar{u} = \bar{Z}(\bar{z}) \bar{\Phi}(\bar{\phi}) \bar{R}(\bar{r}) \bar{U}(\bar{t}) \quad (30)$$

which gives the following simple complex number generalization of the standard equations for cylindrical waves<sup>4-7</sup>

$$d^2 \bar{\Phi} / d\bar{\phi}^2 + \bar{M}^2 \bar{\Phi} = 0 \quad (31)$$

$$d^2 \bar{Z} / d\bar{z}^2 + \bar{k}_z^2 \bar{Z} = 0 \quad (32)$$

$$\bar{r}^2 d^2 \bar{R} / d\bar{r}^2 + \bar{r} d\bar{R} / d\bar{r} + (\bar{k}_r^2 \bar{r}^2 - \bar{M}^2) \bar{R} = 0 \quad (33)$$

$$d^2 \bar{U} / d\bar{t}^2 + \bar{\omega}^2 \bar{U} = 0 \quad (34)$$

where  $\bar{M}$  = constant, and  $\bar{k}_z$  and  $\bar{k}_r$  are constants that are related by<sup>4-7</sup>

$$\bar{k}_z^2 = \bar{k}^2 - \bar{k}_r^2 \quad (35)$$

where  $\bar{k}$  is defined by

$$\bar{k} = \bar{\omega} / \bar{v} \quad (36)$$

Using

$$\bar{v} = v e^{j\theta_v} \quad \bar{k} = k e^{j\theta_k} \quad (37)$$

gives

$$k = \omega / v \quad \theta_k = \theta_\omega - \theta_v \quad (38)$$

which determines  $k$  and  $\theta_k$ . Writing  $\bar{k}_z$  and  $\bar{k}_r$  as

$$\bar{k}_z = k_z e^{j\theta_k z} = k_{zR} + j k_{zI} \quad (39)$$

$$\bar{k}_r = k_r e^{j\theta_k r} = k_{rR} + j k_{rI} \quad (40)$$

where  $k_{zR} = k_z \cos \theta_{kz}$ ,  $k_{zI} = k_z \sin \theta_{kz}$ ,  $k_{rR} = k_r \cos \theta_{kr}$  and  $k_{rI} = k_r \sin \theta_{kr}$ , allows equation (35) to be written as

$$k_z^2 \cos(2\theta_{kz}) = k^2 \cos(2\theta_k) - k_r^2 \cos(2\theta_{kr}) \quad (41)$$

$$k_z^2 \sin(2\theta_{kz}) = k^2 \sin(2\theta_k) - k_r^2 \sin(2\theta_{kr}) \quad (42)$$

Corresponding to the phase velocity given by equation (36) the complex number group velocity is given by

$$\bar{v}_g = v_g e^{j\theta_{vg}} = d\bar{\omega}/d\bar{k} \quad (43)$$

Therefore

$$v_g = \cos \beta_{k,k} [(d\omega/dk)^2 + (\omega d\theta_\omega/dk)^2]^{1/2} \quad (44)$$

$$\theta_{vg} = \theta_\omega + \beta_{\omega,\omega} - \theta_k - \beta_{k,k} = \theta_v + \beta_{\omega,\omega} - \beta_{k,k} \quad (45)$$

$$\tan \beta_{\omega,\omega} = \frac{\omega d\theta_\omega/dk}{d\omega/dk} = \omega d\theta_\omega/d\omega \quad (46)$$

$$\tan \beta_{k,k} = k d\theta_k/dk \quad (47)$$

The solution to equations (31) through (33) will now be considered.

#### A. Solution of $\bar{\phi}$ Equation.

Consider now the solution of equation (31) and the determination of the complex number constant  $\bar{M}$ . The solution of equation (31) can be written as

$$\bar{\phi} = \bar{A}e^{i\bar{M}\bar{\phi}} + \bar{B}e^{-i\bar{M}\bar{\phi}} \quad (48)$$

It will now be shown that  $\bar{M}\bar{\phi}$  must be a real number if  $\phi$  is to be a periodic function of the measured azimuthal angle  $\phi_m = \phi \cos \theta_\phi$ . Writing the complex number  $\bar{M}$  as

$$\bar{M} = Me^{j\theta_M} = M_R + jM_I \quad (49)$$

allows the phase  $\bar{M}\bar{\phi}$  in equation (48) to be written as

$$\bar{M}\bar{\phi} = M_R\phi_R - M_I\phi_I + j(M_I\phi_R + M_R\phi_I) \quad (50)$$

The reality of  $\bar{M}\bar{\phi}$  gives

$$M_I\phi_R + M_R\phi_I = 0 \quad (51)$$

and

$$\bar{M}\bar{\phi} = M\phi = \phi_R M^2 / M_R = \phi_m M^2 / M_R \quad (52)$$

For a periodicity of the form  $\bar{\phi}(\phi_R) = \bar{\phi}(\phi_R + 2\pi)$  it is required that

$$M^2 / M_R = m \quad (53)$$

where  $m =$  positive integer. From equation (49) it follows that

$$M_R = M \cos \theta_M = M \cos \theta_\phi \quad (54)$$

because  $\theta_M = -\theta_\phi$  from equation (52). Combining equations (53) and (54) gives

$$M = m \cos \theta_\phi \quad (55A)$$

$$M_R = m \cos^2 \theta_\phi \quad (55B)$$

$$M_I = -m \cos \theta_\phi \sin \theta_\phi \quad (55C)$$

as the condition for the function  $\bar{\phi}$  to be periodic in  $\phi_R$  with period  $2\pi$ . Therefore  $M$  is not an integer and equations (48) and (52) show that the wave amplitude is not periodic in the variable  $\phi$ . Equation (45) can be written as

$$\bar{\phi} = \bar{A}e^{im\phi_R} + \bar{B}e^{-im\phi_R} \quad (56)$$

which is periodic in  $\phi_R$ . Note that  $M\phi = m\phi_R$ , and that  $\bar{M}$  is a complex number in equation (31). Traditionally equation (31) accepts only integer values of the separation constant, but for waves in asymmetric space and time the separation constant is the complex number  $\bar{M}$ . The reality condition on the phase  $\bar{M}\bar{\phi}$  is

$$\theta_M = -\theta_\phi \quad (57)$$

which is the equation for evaluating the phase angle  $\theta_M$ . The internal phase angle of the magnetic quantum number must adjust itself in such a way that equation (57) is valid for periodic waves.

B. Solution of the  $\bar{Z}$  Equation.

Equation (32) has the following formal solution

$$\bar{Z} = \bar{C}e^{i\bar{k}_z \bar{z}} + \bar{D}e^{-i\bar{k}_z \bar{z}} \quad (58)$$

The exponent term in equation (58) can be written as

$$\bar{k}_z \bar{z} = k_{zR} z_R - k_{zI} z_I + j(k_{zI} z_R + k_{zR} z_I) \quad (59)$$

and the reality requirement for the exponent term in internal space gives

$$\bar{k}_z \bar{z} = k_z z = z_R k_z^2 / k_{zR} = z_R k_z / \cos \theta_{kz} \quad (60)$$

and which also gives  $\theta_{kz} = -\theta_z$ . If the waves propagate in the  $z$  direction

with a measured spatial wavelength  $L_{zm} = L_{zR} = L_z \cos \theta_z$  then equation (60) gives

$$L_{zR} k_z / \cos \theta_{kz} = 2\pi \quad (61)$$

for periodicity with wavelength  $L_{zR}$ . Therefore

$$k_z = 2\pi / L_{zR} \cos \theta_{kz} = 2\pi / L_z \quad (62)$$

$$k_{zR} = 2\pi / L_{zR} \cos^2 \theta_{kz} = 2\pi / L_z \cos \theta_{kz} = k_z \cos \theta_{kz} \quad (63)$$

The reality condition on  $\bar{k}_z \bar{z}$  shows that

$$\theta_{kz} = -\theta_z = -\theta_{Lz} \quad (64)$$

Equation (63) shows that  $k_{zR} \neq 2\pi / L_{zR}$ .

The solution given in equation (58) can be rewritten as

$$\begin{aligned} \bar{z} &= \bar{C} e^{ik_z z} + \bar{D} e^{-ik_z z} \\ &= \bar{C} e^{i2\pi z_R / L_{zR}} + \bar{D} e^{-i2\pi z_R / L_{zR}} \end{aligned} \quad (65)$$

The internal phase angle of the wave number must adjust itself to the local broken symmetry of spacetime such that equation (64) is satisfied for periodic waves. Equations (58) or (65) are the general solutions for plane waves in the  $z$  direction and equation (64) holds for  $-\infty < z < \infty$ . It is possible that  $\bar{k}_z$  is an imaginary number in real space so that  $\bar{k}_z = i\bar{k}_z$ , then the solution to equation (32) is attenuating in nature and given by

$$\bar{z} = \bar{C} e^{\bar{k}_z \bar{z}} + \bar{D} e^{-\bar{k}_z \bar{z}} \quad (66)$$

and apparently  $\bar{k}_z \bar{z}$  need not be a real number in internal space because there is no periodicity requirement in the  $z$  direction for this case.

### C. Solution of the $\bar{R}$ Equation.

The radial equation (33) is similar to the standard radial equation of vibration theory except that  $\bar{R}$ ,  $\bar{r}$ ,  $\bar{k}_r$  and  $\bar{M}$  are complex numbers. The formal solution to equation (33) can be written as a generalization of the standard result for scalar coordinates as follows<sup>4,7</sup>

$$\bar{R} = \bar{A} \bar{J}_{\bar{M}}(\bar{k}_r \bar{r}) + \bar{B} \bar{N}_{\bar{M}}(\bar{k}_r \bar{r}) \quad (67)$$

which represent standing waves, where  $\bar{J}_{\bar{M}}$  = complex number Bessel function and  $\bar{N}_{\bar{M}}$  = Neumann function of complex order  $\bar{M}$ . The progressive wave solutions to equation (33) are<sup>4,7</sup>

$$\bar{R} = \bar{A} \bar{H}_{\bar{M}}^{(1)}(\bar{k}_r \bar{r}) + \bar{B} \bar{H}_{\bar{M}}^{(2)}(\bar{k}_r \bar{r}) \quad (68)$$

where  $\bar{H}_{\bar{M}}^{(1)}$  and  $\bar{H}_{\bar{M}}^{(2)}$  = complex number Hankel functions of the first and second

kind of order  $\bar{M}$ . The asymptotic values of the Hankel functions are given by the following generalization of the standard results<sup>4,7</sup>

$$\bar{H}_{\bar{M}}^{(1)}(\bar{k}_r \bar{r}) \sim [2/(\pi \bar{k}_r \bar{r})]^{1/2} e^{i(\bar{k}_r \bar{r} - \bar{M}\pi/2 - \pi/4)} \quad (69)$$

$$\bar{H}_{\bar{M}}^{(2)}(\bar{k}_r \bar{r}) \sim [2/(\pi \bar{k}_r \bar{r})]^{1/2} e^{-i(\bar{k}_r \bar{r} - \bar{M}\pi/2 - \pi/4)} \quad (70)$$

If equations (69) and (70) represent in-going and out-going waves which have a periodicity in the measured radial coordinate  $r_R = r \cos \theta_r$  it follows that the phase  $\bar{k}_r \bar{r}$  must be a real number in the far field, with  $r \rightarrow \infty$ , so that

$$\bar{k}_r \bar{r} = k_r r = r_R k_r^2 / k_{rR} = r_R k_r / \cos \theta_{kr} \quad (71)$$

$$\theta_{kr} = -\theta_r (r = \infty) \quad (72)$$

where  $k_{rR} = k_r \cos \theta_{kr}$ .

Let the waves in the far field propagate in the  $r$  direction with a measured spatial wavelength  $L_{rR} = L_r \cos \theta_r$  where  $L_r =$  intrinsic spatial wavelength in the  $r$  direction. Then from equation (71)

$$L_{rR} k_r / \cos \theta_{kr} = 2\pi \quad (73)$$

$$k_r = 2\pi / L_{rR} \cos \theta_{kr} = 2\pi / L_r \quad (74)$$

$$k_{rR} = 2\pi / L_{rR} \cos^2 \theta_{kr} = 2\pi / L_r \cos \theta_{kr} = k_r \cos \theta_{kr} \quad (75)$$

so that  $k_{rR} \neq 2\pi / L_{rR}$ . Equations (71) through (75) hold only in the far field because only in this region is the concept of the wavelengths  $L_r$  and  $L_{rR}$  defined. The presence of periodic waves requires the broken symmetry of the wavelength to adjust itself so as to satisfy equation (72) at large distances from the source of the waves. In the far field of asymmetric waves equations (69) and (70) can be rewritten as

$$\bar{H}_{\bar{M}}^{(1)}(k_r r) \sim [2/(\pi k_r r)]^{1/2} e^{ik_r r} e^{-i\pi/2(\bar{M}+1/2)} \quad (76)$$

$$\bar{H}_{\bar{M}}^{(2)}(k_r r) \sim [2/(\pi k_r r)]^{1/2} e^{-ik_r r} e^{i\pi/2(\bar{M}+1/2)} \quad (77)$$

These solutions represent progressive waves.

As a special case consider the solution of standing waves in a vibrating membrane located in space and time with broken internal symmetries. The solution of the wave equation for this case is an obvious generalization of the standard results for scalar quantities<sup>6</sup>

$$\bar{u} = \bar{C} \bar{J}_{\bar{M}}(\bar{k}_r \bar{r}) \cos(\bar{M}\phi + \bar{a}) e^{i\bar{\omega}\bar{t}} \quad (78)$$

$$= \bar{C} \bar{J}_{\bar{M}}(\bar{k}_r \bar{r}) \cos(m\phi_m + \bar{a}) e^{i\omega_m t_m / \cos^2 \theta_\omega} \quad (79)$$

where for a membrane  $k_r = k$ . The small argument expansion of the Bessel function of order  $\bar{M}$  is given by the following generalization of the standard scalar

results

$$\bar{J}_{\bar{M}}(\bar{k}_r \bar{r}) = (\bar{k}_r \bar{r})^{\bar{M}} \left\{ 1 - \bar{k}_r^2 \bar{r}^2 / [4(\bar{M} + 1)] + \bar{k}_r^4 \bar{r}^4 / [32(\bar{M} + 1)(\bar{M} + 2)] - \dots \right\} \quad (80)$$

Using equations (49) and (80) gives

$$\bar{J}_{\bar{M}} = (k_r r)^x e^{-y+jw} \left\{ 1 - \bar{k}_r^2 \bar{r}^2 / [4(\bar{M} + 1)] + \bar{k}_r^4 \bar{r}^4 / [32(\bar{M} + 1)(\bar{M} + 2)] - \dots \right\} \quad (81)$$

where

$$x = M \cos \theta_M = m \cos^2 \theta_M \quad (82)$$

$$y = M(\theta_{kr} + \theta_r) \sin \theta_M \quad (83)$$

$$= m(\theta_{kr} + \theta_r) \sin \theta_M \cos \theta_M$$

$$w = m(\theta_{kr} + \theta_r) \cos^2 \theta_M + m \ln(k_r r) \cos \theta_M \sin \theta_M \quad (84)$$

The case  $\bar{M} = 0$  gives

$$\bar{J}_0 = 1 - 1/4 \bar{k}_r^2 \bar{r}^2 + 1/64 \bar{k}_r^4 \bar{r}^4 - \dots \quad (85)$$

In equation (78) and (79)  $\bar{k}_r \bar{r}$  is not a real number because the vibrations are not periodic in the radial direction.

4. SPHERICAL WAVES IN ASYMMETRIC MATTER. A simple generalization of the standard wave equation for spherical waves gives the following equation that describes spherical waves in space and time with broken internal symmetries<sup>4-7</sup>

$$\partial^2 \bar{u} / \partial \bar{r}^2 + 2/\bar{r} \partial \bar{u} / \partial \bar{r} + 1/(\bar{r}^2 \sin \bar{\psi}) \partial / \partial \bar{\psi} (\sin \bar{\psi} \partial \bar{u} / \partial \bar{\psi}) \quad (86)$$

$$+ 1/(\bar{r}^2 \sin^2 \bar{\psi}) \partial^2 \bar{u} / \partial \bar{\phi}^2 = 1/\bar{v}^2 \partial^2 \bar{u} / \partial \bar{t}^2$$

Separating the complex number wave amplitude as

$$\bar{u} = \bar{R}(\bar{r}) \bar{W}(\bar{\psi}) \bar{\Phi}(\bar{\phi}) \bar{U}(\bar{t}) \quad (87)$$

gives

$$(1 - \bar{u}^2) d^2 \bar{W} / d\bar{u}^2 - 2\bar{u} d\bar{W} / d\bar{u} + [\bar{L}(\bar{L} + 1) - \bar{M}^2 / (1 - \bar{u}^2)] \bar{W} = 0 \quad (88)$$

$$\bar{r}^2 d^2 \bar{R} / d\bar{r}^2 + 2\bar{r} d\bar{R} / d\bar{r} + [\bar{k}^2 \bar{r}^2 - \bar{L}(\bar{L} + 1)] \bar{R} = 0 \quad (89)$$

$$d^2 \bar{\Phi} / d\bar{\phi}^2 + \bar{M}^2 \bar{\Phi} = 0 \quad (90)$$

$$d^2 \bar{U} / d\bar{t}^2 + \bar{\omega}^2 \bar{U} = 0 \quad (91)$$

where  $\bar{u} = \cos \bar{\psi}$  and  $\bar{k} = \bar{L} / \bar{v}$ .

A. Solution of the  $\bar{W}$  Equation.

The solution of equation (88) can be obtained as a complex number generalization of the associated Legendre polynomials.<sup>7</sup> The complex number associated Legendre polynomials can be obtained from equation (88) by writing<sup>7</sup>

$$\bar{W} = (1 - \bar{\mu}^2)^{\bar{M}/2} \bar{F}(\bar{L}, \bar{M}, \bar{\mu}) \quad (92)$$

where

$$\bar{F} = \sum_{s=0}^{\infty} \bar{c}_s \bar{\mu}^s \quad (93)$$

by direct substitution one finds<sup>4-7</sup>

$$\bar{c}_2 = 1/2[\bar{M}(\bar{M} + 1) - \bar{L}(\bar{L} + 1)]\bar{c}_0 \quad (94)$$

$$\bar{c}_3 = 1/6[(\bar{M} + 1)(\bar{M} + 2) - \bar{L}(\bar{L} + 1)]\bar{c}_1 \quad (95)$$

$$\bar{c}_4 = 1/12[(\bar{M} + 2)(\bar{M} + 3) - \bar{L}(\bar{L} + 1)]\bar{c}_2 \quad (96)$$

$$\bar{c}_5 = 1/20[(\bar{M} + 3)(\bar{M} + 4) - \bar{L}(\bar{L} + 1)]\bar{c}_3 \quad (97)$$

The following is the complex number generalization of the standard scalar results<sup>4-7</sup>

$$\bar{c}_{\nu+2}/\bar{c}_\nu = [(\nu + \bar{M})(\nu + \bar{M} + 1) - \bar{L}(\bar{L} + 1)]/[(\nu + 1)(\nu + 2)] \quad (98)$$

where  $\nu = \text{integer}$ . Break off polynomial solutions can exist even when  $\bar{M}$  and  $\bar{L}$  are complex numbers provided that they are related by

$$\bar{L} = \bar{M} + \nu \quad (99)$$

where the integer  $\nu$  must have the value

$$\nu = \ell - m \quad (100)$$

where  $\ell$  and  $m = \text{integer separation constants}$ . Combining equations (99) and (100) gives

$$\bar{L} = \bar{M} + \ell - m \quad (101)$$

Equation (99) reduces to equation (100) for symmetric spacetime.

From equation (101) it follows that

$$L \cos \theta_L = M \cos \theta_M + \ell - m \quad (102)$$

$$L \sin \theta_L = M \sin \theta_M \quad (103)$$

From equations (102) and (103) it follows that



$$\tan \theta_L = (M \sin \theta_M) / (M \cos \theta_M + \ell - m) \quad (104)$$

$$\ell = \ell [1 - m/\ell(2 - m/\ell) \sin^2 \theta_M]^{1/2} \quad (105)$$

where from the analysis of the  $\phi$  equation given in Section 3 it follows that

$$M = m \cos \theta_M \quad (106)$$

$$\theta_M = -\theta_\phi \quad (107)$$

The complex number associated Legendre polynomial solutions can then be written as

$$\bar{W} = (1 - \bar{u}^2)^{\bar{M}/2} \quad \bar{L} = \bar{M} \quad (108)$$

$$\bar{W} = (1 - \bar{u}^2)^{\bar{M}/2} \bar{u} \quad \bar{L} = \bar{M} + 1 \quad (109)$$

$$\bar{W} = (1 - \bar{u}^2)^{\bar{M}/2} [(2\bar{M} + 3)\bar{u}^2 - 1] \quad \bar{L} = \bar{M} + 2 \quad (110)$$

$$\bar{W} = (1 - \bar{u}^2)^{\bar{M}/2} [(2\bar{M} + 5)\bar{u}^3 - 3\bar{u}] \quad \bar{L} = \bar{M} + 3 \quad (111)$$

Note that formally  $\bar{W}$  is given by the following associated Legendre polynomials

$$\bar{W} = \bar{P}_{\bar{L}}^{\bar{M}}(\bar{u}) \quad (112)$$

#### B. Solution of the $\bar{R}$ Equation

The solution of the complex number radial equation (89) can be obtained by formal analogy to the solution of the real number version of equation (89) and the result for standing waves in asymmetric matter is<sup>4-7</sup>

$$\bar{R} = \bar{A} \bar{j}_{\bar{L}}(\bar{k}\bar{r}) + \bar{B} \bar{n}_{\bar{L}}(\bar{k}\bar{r}) \quad (113)$$

while for progressive waves in asymmetric matter<sup>4-7</sup>

$$\bar{R} = \bar{A} \bar{h}_{\bar{L}}^{(1)}(\bar{k}\bar{r}) + \bar{B} \bar{h}_{\bar{L}}^{(2)}(\bar{k}\bar{r}) \quad (114)$$

where  $\bar{j}_{\bar{L}}(\bar{k}\bar{r})$  = complex number spherical Bessel function of order  $\bar{L}$ ,  $\bar{n}_{\bar{L}}(\bar{k}\bar{r})$  = complex number spherical Neumann function of order  $\bar{L}$ ,  $\bar{h}_{\bar{L}}^{(1)}(\bar{k}\bar{r})$  = complex number spherical Hankel function of first kind of order  $\bar{L}$  and  $\bar{h}_{\bar{L}}^{(2)}(\bar{k}\bar{r})$  = complex number spherical Hankel function of the second kind of order  $\bar{L}$ . These functions are defined as generalizations of the corresponding real valued functions as follows<sup>4-7</sup>

$$\bar{j}_{\bar{L}}(\bar{k}\bar{r}) = [\pi/(2\bar{k}\bar{r})]^{1/2} \bar{J}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r}) \quad (115)$$

$$\bar{n}_{\bar{L}}(\bar{k}\bar{r}) = [\pi/(2\bar{k}\bar{r})]^{1/2} \bar{N}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r}) \quad (116)$$

$$\bar{h}_{\bar{L}}^{(1)}(\bar{k}\bar{r}) = [\pi/(2\bar{k}\bar{r})]^{1/2} [\bar{J}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r}) + i\bar{N}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r})] \quad (117)$$

$$\bar{h}_{\bar{L}}^{(2)}(\bar{k}\bar{r}) = [\pi/(2\bar{k}\bar{r})]^{1/2} [\bar{J}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r}) - i\bar{N}_{\bar{L}+\frac{1}{2}}(\bar{k}\bar{r})] \quad (118)$$

The asymptotic expansions derived from equations (115) through (118) are<sup>4-7</sup>

$$\bar{j}_{\bar{L}}(\bar{k}\bar{r}) \rightarrow 1/(\bar{k}\bar{r}) \sin(\bar{k}\bar{r} - \bar{L}\pi/2) \quad (119)$$

$$\bar{n}_{\bar{L}}(\bar{k}\bar{r}) \rightarrow 1/(\bar{k}\bar{r}) \cos(\bar{k}\bar{r} - \bar{L}\pi/2) \quad (120)$$

$$\bar{h}_{\bar{L}}^{(1)}(\bar{k}\bar{r}) \rightarrow 1/(\bar{k}\bar{r}) e^{i[\bar{k}\bar{r} - (\bar{L}+1)\pi/2]} \quad (121)$$

$$\bar{h}_{\bar{L}}^{(2)}(\bar{k}\bar{r}) \rightarrow 1/(\bar{k}\bar{r}) e^{-i[\bar{k}\bar{r} - (\bar{L}+1)\pi/2]} \quad (122)$$

In order for equations (119) through (122) to describe periodic waves in the far field the following conditions must hold

$$\left. \begin{array}{l} \bar{k}\bar{r} = kr \\ \theta_k + \theta_r = 0 \end{array} \right\} r \rightarrow \infty \quad (123)$$

Thus for instance the replacement  $\bar{k}\bar{r} = kr$  can be made in the right hand sides of equations (119) through (122). Therefore the internal phase angle of spherical waves in the far field must adjust itself to the local broken symmetry or space such that

$$\theta_k = -\theta_r (r = \infty) \quad (124)$$

5. CONCLUSION. Waves propagating in matter or spacetime with broken symmetries will have complex number separation constants  $\bar{M}$  and  $\bar{L}$ . For spherical waves the separation constants must be related by  $\bar{L} - \bar{M} = \text{integer}$ . The observed periodicities of the waves in measured time and measured space requires complex number separation constants, wave numbers, frequencies and coordinates. However the quantities  $\omega t$ ,  $\bar{M}\phi$  and  $\bar{k}\bar{r}$  must be real numbers for periodic waves. Applications to seismic and acoustic waves are possible because the earth's gravity induces a broken symmetry in the coordinates.

#### ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

#### REFERENCES

1. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.
2. 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.
3. Weiss, R. A., "The Broken Symmetry of Space and Time in Bulk Matter and the Vacuum," Sixth Army Conference on Applied Mathematics and Computing, Univ. of Colorado, Boulder, ARO 89-1, 31 May-3 June, 1988, p. 317.
4. Morse, P. M. and Feshbach, H., Methods of Theoretical Physics, Vols. 1 and 2, McGraw-Hill, New York, 1953.
5. Morse, P. M., Vibration and Sound, McGraw-Hill, New York, 1948.
6. Lindsay, R. B., Mechanical Radiation, McGraw-Hill, New York, 1960.
7. Skudrzyk, E., The Foundations of Acoustics, Springer-Verlag, New York, 1971.

# ELASTIC DEFORMATION AND SLUG FLOW AS APPLICATIONS OF FRONT TRACKING

*X. Garaizar*<sup>1</sup>

Courant Institute of Mathematical Sciences  
251 Mercer Street  
New York, New York 10012

*J. Glimm*<sup>2,3</sup>

Chair, Department of Applied Mathematics and Statistics  
SUNY at Stony Brook  
Stony Brook, New York 11794-3600

*W. Guo*<sup>4</sup>

Courant Institute of Mathematical Sciences  
251 Mercer Street  
New York, New York 10012

## ABSTRACT

Two applications of the Front Tracking method form the basis of this paper.

A formulation of the small anisotropy hypothesis for nonlinear elastic deformation is given which is fully rotationally covariant and which is thermodynamically consistent in the sense that it is derived from a specific internal energy. An algorithm for the solution of the Riemann problem for nonlinear elasticity is presented. This algorithm uses Godunov type iterations. For the uniaxial deformations of an isotropic material, the Godunov iterations occur in one dimensional spaces, while in the general case, the iterations are at most in a two dimensional space.

Slug flow is studied in the context of Hele-Shaw cells. The transition from laminar to slug flow is the main object of study.

<sup>1</sup> Supported in part by the Air Force Office of Scientific Research AFOSR-88-0025

<sup>2</sup> Supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research,

U. S. Department of Energy, under contract DE-AC02-76ER03077

<sup>3</sup> Supported in part by the Army Research Office, grant DAAJ03-89-K-0017

<sup>4</sup> Supported in part by the National Sciences Foundation, grant DMS-8619856

## 1. Introduction.

The nonlinear deformations of an elastic body are described by a hyperbolic system of quasilinear equations. The constitutive relations close the system and can be characterized through the dependence of the specific internal energy  $W$  on the strain tensor  $E = \frac{1}{2}(F^t F - I)$ . For hyperelastic materials these relations are the stress-strain relations: in Eulerian coordinates,  $\sigma^{ij} = \rho f^{jk} \frac{\partial W}{\partial f_{ik}}$  [8, 9], where  $(\sigma^{ij})$  is the Cauchy stress tensor,  $f^i_j = \rho_0^{-1} F^i_j$  the Eulerian deformation gradient and  $\rho_0$  the material density.

For deformations small in shear, we model the specific internal energy in terms of the strain tensor  $E$  (or the deformation gradient  $F$ ) by a third order approximation on the effective shear strain  $\epsilon$  [3],

$$W(F, S) = \hat{W}(E, S) = W_I(\gamma, S) + \frac{1}{\rho_0} G_0(\gamma, S) \epsilon^2, \quad (1.1)$$

where  $S$  is the entropy,  $\gamma \equiv (1/3) \text{tr } E$  is the mean compressive strain and  $\epsilon^2 \equiv (\text{dev } E)_{ij} (\text{dev } E)_{ij}$ .  $G_0(\gamma, S)$  is the shear modulus at  $\epsilon = 0$  for the corresponding hydrostatic strain  $\gamma$  and  $W_I$  is a hydrostatic energy. Here,  $\gamma$ ,  $\epsilon^2$  and  $\omega^3 \equiv \det(\text{dev } E)$  form a complete set of invariants for the strain tensor  $E$ , with the third invariant  $\omega^3$  satisfying  $\omega^3 = O(\epsilon^3)$  [3].

The solution waves in elasticity are of three types: predominately longitudinal (or pressure), predominately transverse (or shear) and a thermo-contact. In the non-linear case, the shear waves split in two modes: radial and angular shear, while in the linear case the two shear waves speeds coincide. The elastic system expressed in Eulerian conserved variables, for a uniaxial deformation, is given in terms of the fundamental variables  $\rho$  (density),  $\vec{v}$  (velocity),  $\vec{\sigma}$  (Cauchy stress vector),  $\vec{f}$  (Eulerian deformation vector) and  $e = W(\vec{f}, S)$  (specific internal energy) [9],

$$\frac{\partial}{\partial t} \rho + \frac{\partial}{\partial x} (\rho v^1) = 0, \quad (1.2a)$$

$$\frac{\partial}{\partial t} (\rho f^i) + \frac{\partial}{\partial x} (\rho f^i v^1) = \frac{\partial}{\partial x} (v^i f^1) \quad i = 2, 3, \quad (1.2b)$$

$$\frac{\partial}{\partial t} (\rho v^i) + \frac{\partial}{\partial x} (\rho v^i v^1) = \frac{\partial}{\partial x} (\sigma^i) \quad i = 1, 2, 3, \quad (1.2c)$$

$$\frac{\partial}{\partial t} \left[ \rho \left( \frac{1}{2} v_i v^i + e \right) \right] + \frac{\partial}{\partial x} \left[ \rho \left( \frac{1}{2} v_i v^i + e \right) v^1 \right] = \frac{\partial}{\partial x} (v_i \sigma^i), \quad (1.2d)$$

Here  $\vec{f}$  and  $\vec{\sigma}$  are defined by  $f^i = f^{i_1}$  and  $\sigma^i = \sigma^{i_1}$ . The density satisfies the relation  $\rho = J^{-1} \rho_0 = (f^1)^{-1}$ , with  $J = \det F$ .

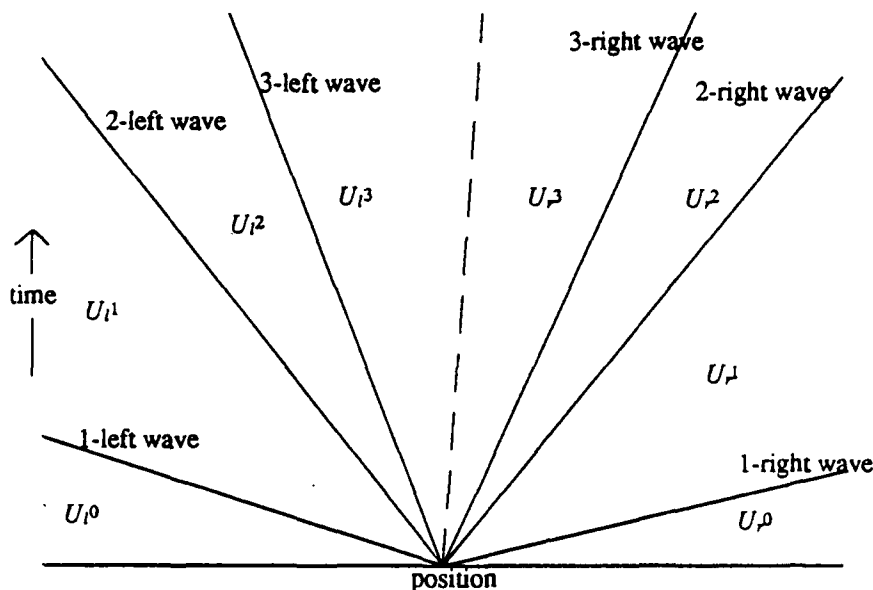


Fig. 1.1 The elastic Riemann solution

The elastic state is written as  $U = (\vec{\sigma}, \vec{v}, S)$ , with  $\vec{v} = (u, v, w)$  and  $\vec{\sigma} = (\sigma, \tau, \theta)$ , where  $\sigma = \sigma^1$ ,  $\sigma^2 = \tau \cos \theta$  and  $\sigma^3 = \tau \sin \theta$ . The elastic Riemann solution consists of eight constant states  $U_{i^\alpha}$ ,  $\alpha = l, r$ ,  $i = 0, 1, 2, 3$  separated by seven waves (see Figure 1). Of these waves, the middle one is a slip line of speed  $\lambda = u_l^3 = u_r^3$  and the remaining waves come in pairs. The fast waves are mainly longitudinal, in the sense that the change in the deformation vector  $\vec{f}$  occurs mainly in the direction of propagation of the waves. The other two pairs are mainly transverse. The slower of the transverse waves correspond to necking while the faster of the transverse waves are linearly degenerate and correspond to torque [4].

## 2. Numerical Solution

We relate the Riemann problem in elasticity to gas dynamics, where the longitudinal waves are pressure waves and the shear wave is a contact of zero speed. Using this perspective we employ a Godunov-type iteration [2,6] to solve the Riemann problem. Each step of the iteration is divided in two blocks: in the first block there are only pressure waves and in the second we solve both shear modes simultaneously. The importance of this decomposition is that it allows a reduction in the dimension of the space in which the Godunov iteration operates; in the most favorable case, each iteration occurs in a one dimensional space, while the general case (non isotropic or non uniaxial intermediate deformation) involves at most a two dimensional iteration. This is in contrast to the seven dimensions of the state space [4].

The first block consists of the 1-waves. In analogy to gas dynamics, we consider the shear and thermal waves as a "wide" slip line with "surface tension" that causes a prescribed jump in  $\sigma$  and  $u$ . This slip line lies between the two pressure waves. We solve the 1-waves for  $\sigma$  and  $u$ , while  $\tau$ ,  $\theta$ ,  $v$  and  $w$  are left free. The change of  $\sigma$  and  $u$  across this line (shear waves) is given by two parameters  $\Delta\sigma = \sigma_r^1 - \sigma_l^1$  and  $\Delta u = u_r^1 - u_l^1$ . The Riemann problem, formulated in this manner, consists only of two waves. We use a Godunov iteration method to evaluate  $\sigma_l^1$  and  $u_l^1$ . Finally the values of  $\sigma_\alpha^1$  determine the complete states  $U_\alpha^1$ , for  $\alpha = l, r$ .

The second block consists of shear and thermal waves. We solve these waves for the variables  $\tau$ ,  $\theta$ ,  $v$  and  $w$ , with the conditions

$$\theta_r^3 = \theta_l^3, \quad w_r^3 = w_l^3, \quad v_r^3 = v_l^3, \quad \tau_r^3 = \tau_l^3,$$

while  $\sigma$  and  $u$  are free variables. For isotropic materials and uniaxial left and right states, the degeneracy of the 2-waves allows us to solve these waves explicitly. We use a Godunov type iteration method for the remaining shear waves.

Proceeding in this fashion,  $\sigma$  and  $u$  will have a jump across the slip line and the difference will depend on the initial parameters  $\Delta\sigma$  and  $\Delta u$ . Thus, we define the function

$$F \begin{pmatrix} \Delta\sigma \\ \Delta u \end{pmatrix} = \begin{pmatrix} \sigma_r^3 - \sigma_l^3 \\ u_r^3 - u_l^3 \end{pmatrix}. \quad (2.1)$$

A solution to the Riemann problem will be given by the parameters  $\Delta\sigma$  and  $\Delta u$  such that  $F\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right] = 0$ . If we write

$$Q\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right] = \left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right] - F\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right], \quad (2.2)$$

we can construct the iteration. Given  $\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right]^{(0)}$ , let

$$\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right]^{(n+1)} = Q\left[\left[\begin{smallmatrix} \Delta\sigma \\ \Delta u \end{smallmatrix}\right]^{(n)}\right], \text{ for } n \geq 0. \quad (2.3)$$

From equation (2.3), the zeros of  $F$  correspond to fixed points of  $Q$ .

### 3. Front Tracking in Elasticity.

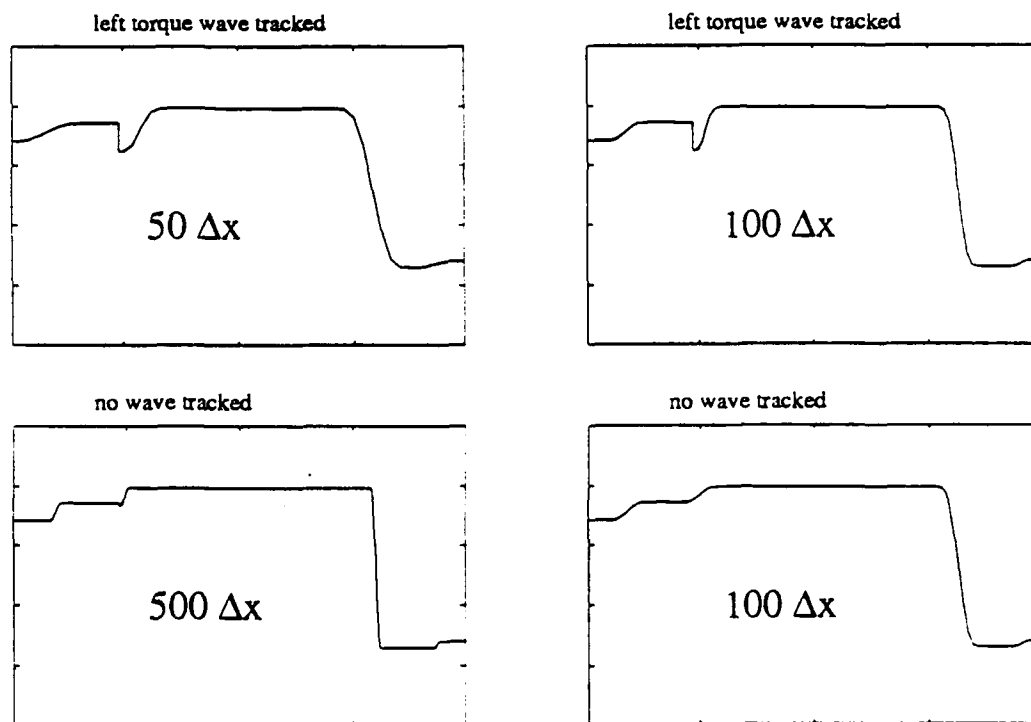


Fig. 3.1

In Figure 3.1 we compare four computational solutions for the same elastic Riemann problem. The main point of this comparison is to illustrate the advantage (or necessity) of tracking the torque wave. The variable shown is the conserved quantity  $\rho f^3$ , plotted at approximately the



same physical time  $t = 3.17$ . We use the Riemann solver described in the previous section in conjunction with a higher order Godunov scheme [7] in a form proposed and designed by I-L. Chern [1]. The top two frames used grids of 50 and 100 points respectively, and an algorithm which tracked the left 2-wave (torque). The bottom frames correspond to runs with grids of 500 and 1000 points, where no waves are tracked. We observe that the stiffness of the material produces two shear waves of similar and slightly different speeds. This, together with numerical diffusion on the linear wave, "hides" the intermediate state between the shear waves producing a single shear wave in the untracked computation. For finer grids (500 points) we observe a "ripple", where the torque wave should appear, which is still negligible compared with the size of the actual discontinuity (see Figure 3.2). Note that in Fig 3.2, 15 mesh cells are used within the ripple region, but the waves are still under resolved. Tracking of the linearly degenerate wave, on the other hand, forces the resolution of the shear waves independently, thus preserving the wave structure of the solution, regardless of the size of the grid.

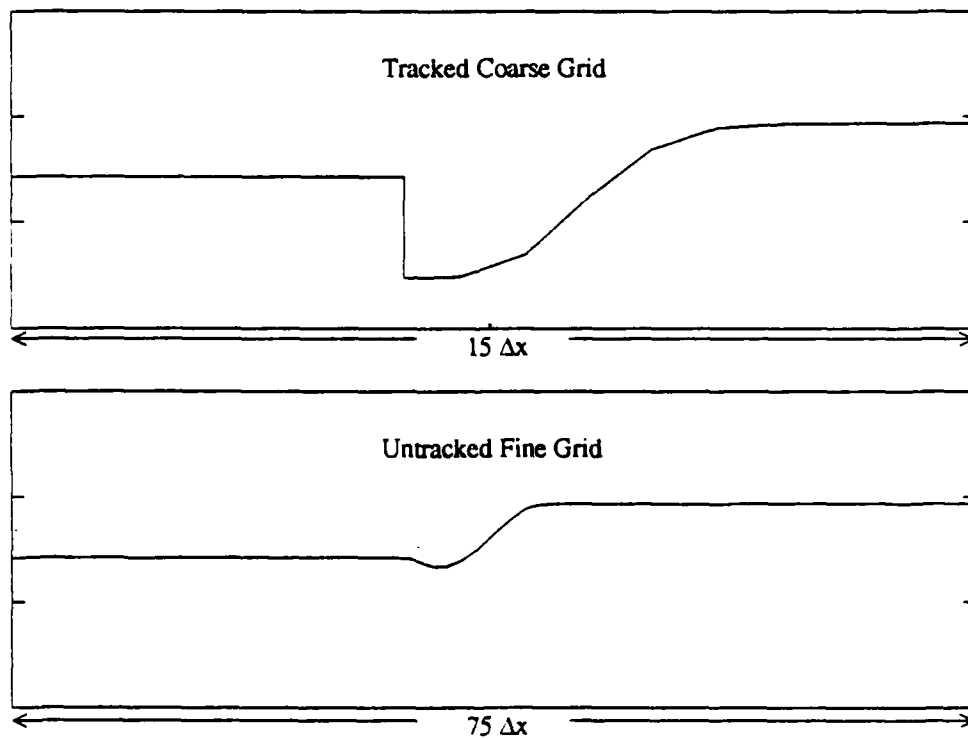


Fig. 3.2

#### 4. Umbilic Points

For isotropic materials, the dependence of  $W$  on  $\vec{f}$  is through  $f = f^1$  and  $g^2 = (f^2)^2 + (f^3)^2$ . In the previous section, for these materials, we have assumed that the eigenvalues associated to the nonlinear pressure and necking waves ( $\lambda_1$  and  $\lambda_3$  resp., near the reference state) and the linearly degenerate torque waves ( $\lambda_2$ ) are ordered as  $\lambda_1 \geq \lambda_2 \geq \lambda_3$ . A potential difficulty arises at states for which the two shear waves cross over or the eigenvalues associated to the nonlinear waves (pressure and necking) coincide. These states are called umbilic points. It follows from the symmetry of  $W(f, g^2)$  on  $g$ , that the line  $g = 0$  is a locus of umbilic points;  $\lambda_2 = \lambda_1$  or  $\lambda_3$ . These points are double (either  $\lambda_2 = \lambda_1$  or  $\lambda_2 = \lambda_3$ ) or triple ( $\lambda_1 = \lambda_2 = \lambda_3$ ). We refer to the double points as shear umbilic points and to the triple points as nonlinear umbilic points or simply umbilic points. Near the undeformed state, the pressure waves are always faster than the shear waves, but this ordering is reversed when crossing an umbilic point. Therefore the double points on the umbilic line  $g = 0$  correspond always to the two shear waves speeds coinciding, thus the name shear umbilic points. The shear umbilic points have a fairly simple mathematical structure. This is reflected, for example in the fact that the torque waves are linearly degenerate waves (contact discontinuities) across which only the angle  $\theta$  changes.

To study the occurrence of umbilic points, we have studied the small shear model of the specific internal energy for several common materials (aluminum, copper, lead, platinum, ...) and searched for the coinciding eigenvalues. We observe that the umbilic points occur only on the line  $g = 0$ , the ordering of the waves prevails and, furthermore, the points for which  $\lambda_1 = \lambda_3$  is satisfied, lie outside of the elastic region. They lie in the region of plastic compression, and are well within experimentally accessible limits.

We now describe the small shear constitutive law. In the formulation

$$W(\mathbf{F}, S) = W_I(\gamma, S) + \frac{1}{\rho_0} G_0(\gamma, S) \epsilon^2, \quad (4.1)$$

the hydrostatic energy  $W_I$  is given by a stiffened gamma law,

$$W_I(V, S) = \frac{1}{\Gamma-1} v^{-(\Gamma-1)} \exp\left(\frac{\Gamma-1}{R} S\right) + P_\infty V, \quad (4.2)$$

where  $V$  and  $\gamma$  are related by  $\left(\frac{V}{V_0}\right)^2 = (2\gamma + 1)^3$ . The shear modulus is taken from Steinberg-

Cochran-Guinan's formulation

$$G(V, S) = G_0 \left[ 1 + AP\eta^{-1/3} + B(T-300) \right] \quad (4.3)$$

( $G_0$ ,  $A$  and  $B$  are tabulated constants and  $\eta$  is the compression,  $\eta = \rho/\rho_0$ ). Here, the pressure  $P$  and temperature  $T$  are considered to be hydrostatic quantities; i.e.,  $P = -\partial(W_I)/\partial V$  and  $T = \partial(W_I)/\partial S$ .

To describe the elastic region, we consider Steinberg-Cochran-Guinan's expression [10] for the yield strength  $Y$ , in the Von Mises sense,

$$Y(V, S) = \frac{G}{G_0} Y_0 \left[ 1 + \beta(\epsilon + \epsilon_i) \right]^n \quad (4.4)$$

with the constraint  $Y_0 \left[ 1 + \beta(\epsilon + \epsilon_i) \right]^n \leq Y_{\max}$ .

The coefficients  $G_0$ ,  $A$ ,  $B$ ,  $Y_0$ ,  $Y_{\max}$ ,  $\beta$ ,  $n$  and  $\Gamma$  are obtained from Steinberg-Cochran-Guinan [10]. (See Table 1.)

We calculate the simple uniaxial compression needed to reach the umbilic point ( $\lambda_3 = \lambda_1$ ) at constant entropy. For the materials we consider (see Table 2), the umbilical points lie on the line  $g = 0$  and the compression values vary from 8% to 14.5%.

At the same time, the elastic region is determined by the effective shear stress  $\tau_e$ , from the inequality (Von Mises)  $3(\tau_e)^2 \leq 2Y^2$ . From the expression for the internal energy (1.1), we have

$$(\tau_e)^2 = 4\rho_0^2 \left[ \frac{\partial W}{\partial(\epsilon^2)} \right]^2 \epsilon^2 = 4G^2 \epsilon^2,$$

and therefore  $\epsilon^2 \leq \frac{1}{12} \left[ \frac{Y}{G} \right]^2 \leq \frac{1}{12} \left[ \frac{Y_{\max}}{G_0} \right]^2$ . This relation provides bounds for the elastic

uniaxial compression  $\eta \approx J^{-1} = (\det F)^{-1}$ . We see that the bounds are less than 1.3% in compression for a number of common metals (see Table 2). This shows that the umbilical point is outside the elastic region.

Material	<i>Al</i>	<i>Au</i>	<i>Cu</i>	<i>Pb</i>	<i>Pt</i>	<i>W</i>
$\rho_0$ (gr/cc)	2.702	18.88	8.92	11.3437	21.45	19.35
M (gr/(gr moles))	26.98	196.967	63.546	207.19	195.09	183.85
$P_{\infty}$ (kbar)	2.4970	2.3901	3.5005	1.36474	2.7417	2.6316
$G_0$ (Mbar)	.276	.28	.477	.086	.637	1.6
$A$ (Mbar <sup>-1</sup> )	6.5	3.8	2.8	11.6	2.5	.94
$-B$ (kK <sup>-1</sup> )	.62	.31	.38	1.16	.14	.14
$\Gamma$	2.97	3.99	3.02	3.74	3.74	2.67
$Y_0$ (Mbar)	.0029	.0002	.0012	.00008	.0003	.022
$Y_{\max}$ (Mbar)	.0068	.00225	.0064	.001	.0034	.04

Table 1. Material constants used to define the constitutive law for a number of common metals.

Material	Aluminum	Gold	Copper	Lead	Platinum	Tungsten
min. compr.	.987904	.996006	.993358	.994236	.997342	.987730
max. compr.	1.012551	1.004042	1.006777	1.005865	1.002679	1.012739
umbilic compr.	1.117324	1.127150	1.130208	1.080368	1.145728	1.144691
press. (kbars)	.8944	.3245	1.4207	.43235	1.621	1.0278

Table 2. The minimum and maximum uniaxial compression for the elastic region is tabulated for common metals along with the compression and pressure at the umbilic point. The main point of this table is that the umbilic point occurs in uniaxial compression at experimentally attainable pressures, within the plastic range, for a number of common metals.

## 5. Slug Flow

The instability of the interface of two viscous fluids has been extensively studied. The Taylor-Saffman problem is the most well known example, which concerns a finger growing on an interface between two fluids in a narrow, two dimensional channel; i.e., a Hele-Shaw cell. The equations for a two fluid Hele-Shaw flow can be written as

$$v = -\mu_i^{-1} \nabla P, \quad i = 1, 2, \quad (5.1)$$

$$\nabla \cdot v = 0,$$

where  $v$ ,  $\mu_i$  and  $P$  are velocity, viscosities and pressure. The indices  $i = 1, 2$  refer to the two fluids. In contrast to the classical Taylor-Saffman problem, we have studied the reversal of fingering stability caused by a strong transverse flow field, see Figure 5.1. Fingers growing into a strong transverse flow have qualitative features not present in the classical Taylor-Saffman instability, namely narrow fingers are produced in the Taylor-Saffman stable case, while wide fingers or nonfingering behavior occurs in the Taylor-Saffman unstable case see Figure 5.2.

We consider a geometry defined by a rectangle with no flow boundary conditions at the top and bottom and prescribed inflow boundary through the left edge, in the form of distinct channels of two distinct fluids, see Figure 5.1a.

The basic physical properties of Hele-Shaw flow, in the flow geometry of Figure 5.1a, are determined by three dimensionless parameters, namely the ratio of the inflow velocities  $V = v_2/v_1$ , mobility ratio  $M = \mu_2/\mu_1$  and width ratio  $W = l_2/(l_1 + l_2)$  of the two channels. When  $MV \neq 1$  in the two channels, the flow depicted in Figure 5.1a is not in equilibrium and fingering may result. To better understand the initiation of fingering, we write  $v = v' + (v - v')$  where uniform field  $v'$  satisfies  $MV' = 1$  and  $v - v'$  has zero outflow right boundary conditions. See Figure 5.1b. Two time scales are essentially important to flow patterns: the time  $t_s$ , which characterizes the transport of the finger down stream, and the pinch off time  $t_p$ , which is the characteristic time for a single finger growing to the height of the channel width. The ratio  $\beta$ , of the two time scales, is a function of  $V$ ,  $M$  and  $W$ .  $\beta$  can be used as an order parameter to classify the distinct flow regimes and the transition between them. When  $\beta \gg 1$ , pinch off occurs repeatedly, and turns a laminar flow to slug flow, see Figure 5.3.a. Even if the inflow boundary condition for Hele-Shaw flow is set to its equilibrium value, in a variable flow channel the instability

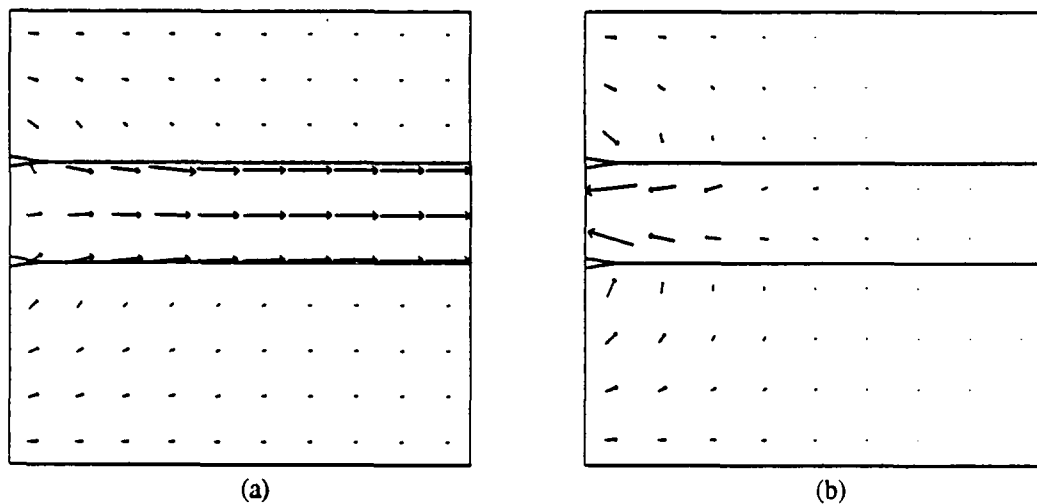
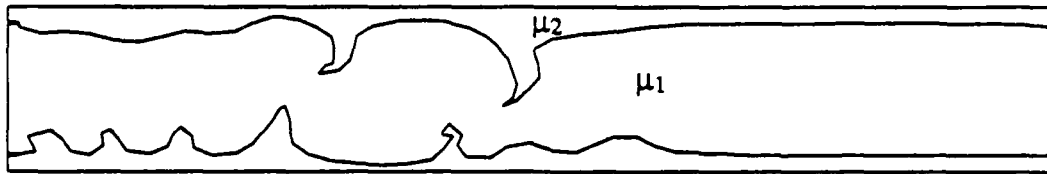


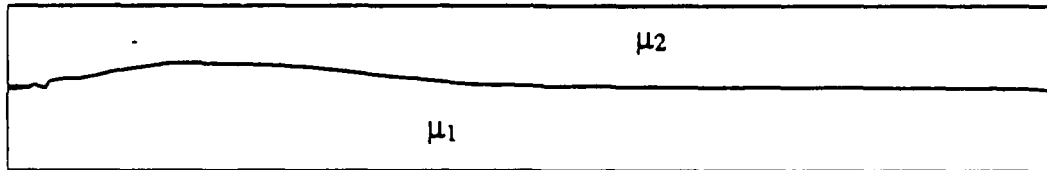
Figure 5.1. The central channel is solvent located between two layers of oil. (a) The velocity field of the total flow  $v$ . The injection rate on the right boundary satisfies  $MV = 1$ . The left boundary velocity ratio  $V$  is larger than the equilibrium value, i.e.  $MV > 1$ , which results in excess in flow of oil. (b) The velocity field of the nonuniform flow  $v - v'$ . There is no flow across the right boundary. The excess oil in flow produces the transverse field which drives the solvent flow backward at the inlet.

develops locally, which also can produce a laminar to slug flow transition, see Figure 5.3.b.

A detailed presentation of the results of this section will be given separately [5].

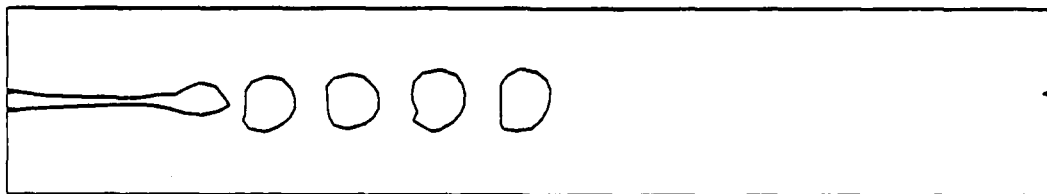


(a)

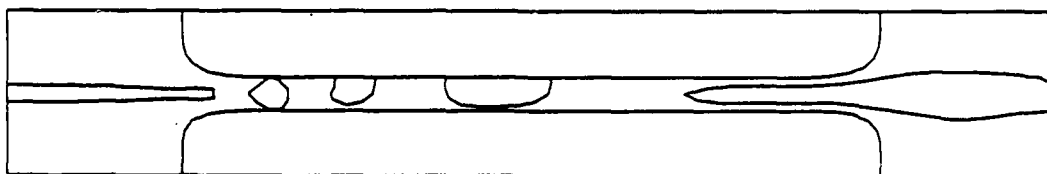


(b)

Figure 5.2. Finger growing on the interface of two viscous fluids, where  $\mu_2 > \mu_1$ . (a) Excess flow of fluid of  $\mu_2$ . Here  $V = 1$ ,  $M = 5$  and  $W = 0.4$ . (b) Excess flow of fluid of  $\mu_1$ . Here  $V = 1$ ,  $M = 40$  and  $W = 0.5$ . The mesh size for cases (a) and (b) are  $30 \times 10$  and  $41 \times 17$  for the hyperbolic equation,  $60 \times 30$  and  $70 \times 30$  for the elliptic equation.



(a)



(b)

Figure 5.3. (a) Consecutive pinch off turns a laminar flow into slug flow. Here  $V = 1$ ,  $M = 5$  and  $W = 0.92$ . The mesh size is  $42 \times 14$  for the hyperbolic equation,  $84 \times 30$  and for the elliptic equation. (b) The transition of laminar to slug flow in a channel of variable width.

## References

1. I-L. Chern, *Private Communication*, 1989.
2. A. Chorin, "Random Choice Solutions of Hyperbolic Systems," *J. Comp. Phys.*, vol. 22, pp. 517-533, 1976.
3. X. Garaizar, "The Small Anisotropy Formulation of Elastic Deformations," *Acta Applicandae Mathematicae*, vol. 14, pp. 259-268, 1989.
4. X. Garaizar, "Solution of a Riemann Problem for Elasticity," *In preparation*, 1989.
5. J. Glimm, W. Guo, and Q. Zhang, "The Laminar to Slug Flow Transition in Hele-Shaw Flow," *In preparation*, 1989.
6. S. K. Godunov, "Finite-difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics," *Mat. Sbornik*, vol. 47, pp. 271-306, 1959.
7. B. Van Leer, "Towards the Ultimate Conservative Difference Scheme. V. A Second-Order Sequel to Godunov's Method," *J. Comp. Phys.*, vol. 32, pp. 101-136, 1977.
8. J. E. Marsden and T. J. R. Hughes, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, N. J., 1983.
9. Bradley J. Plohr and D. Sharp, "A Conservative Eulerian Formulation of the Equations for Elastic Flow," *Adv. Appl. Math.*, vol. 9, pp. 481-499, 1988.
10. D. J. Steinberg, S. G. Cochran, and M. W. Guinan, "A Constitutive Model for metals applicable at high-strain rate," *J. Appl. Phys.*, vol. 51, pp. 1498-1504, 1980.



## NUMERICAL SOLUTION OF AN APERTURE ANTENNA INTEGRAL EQUATION

M.A. Hussain, Ben Noble, Wen-Tai Lin, and B. Becker  
General Electric Research and Development Center  
Schenectady, New York

### ABSTRACT

Synthesis of linear antenna arrays can be formulated in terms of an integral equation of the first kind by considering a linear array as radiation from a line source. This integral equation is a Fredholm equation of the first kind which is difficult to solve numerically by straightforward methods. The difficulty is overcome by exploiting the pattern theorem of T.T. Taylor, using an iterative procedure to refine Taylor's analytical solution. This numerical method can be used to tailor the beam so that a number of small sidelobes are of equal size and a few isolated nulls can be forced over certain regions. The method is illustrated for one and two dimensional apertures.

### INTRODUCTION

One definition of an optimum antenna is a design in which all the sidelobes are of equal level. However this is impossible since it would involve infinite energy, because of Parseval's theorem. The best one can hope for is a design in which a finite number of sidelobes are of equal level, and the level of the side sidelobes tend to zero at infinity.

Dolph [1] has derived the optimum current distribution for linear arrays with a finite number of elements using the properties of Tchebycheff polynomials. Van Der Maas [2] has obtained a simple asymptotic formula for the space factor when the number of radiating elements becomes very large. Using this formula and analytic properties of the space factor, Taylor [3] has shown that this optimum or ideal space factor is not realizable due to the singular behavior of the current distribution near the ends of the aperture. Taylor then gave a practical method to avoid such singularities, and obtain what are termed Taylor weights. This widely used method, provides practical weights for linear arrays.

The basic idea behind Taylor's method is that of bringing a selected number of zeros in the space factor closer to the center of the visible region and preserve the zeros in the far region at integer values. This is accomplished through the use of Woodward's synthesis procedure [4]. The one-dimensional space factor is then given analytically in the compact form. When this method is used, the sidelobes are no longer of equal magnitudes, as is the case with the Dolph-Tchebycheff method. Instead, they decay slowly from the designed sidelobe ratio.

In this paper, rather than use an analytical method, the solution of the integral equation is reduced to a set of algebraic equations, using the Woodward synthesis techniques. The illumination function (i.e. current distribution), as well as the pattern, is easily computed. The results are identical to those of Taylor. As in Taylor's result, however, the sidelobes are not of equal magnitude. We then iterate around the zeros to correct the problem. After only a few iterations the desired zeros, and sidelobes of equal amplitude, are obtained.

The problem is then formulated for specified levels of sidelobes, and the method illustrated by some numerical examples. The generalization to an arbitrary set of nulls or near nulls (i.e. low sidelobes) then becomes clear.

The integral equation approach to antenna synthesis assumes a continuous distribution of radiators. In practice this will be discretized. The method illustrated here will help understanding of the discretization procedure for periodic as well as aperiodic (unequally spaced) arrays, array thinning, and sub-arraying of large array synthesis.

## APERTURE INTEGRAL EQUATION

In this section we formulate the problem of the one dimensional aperture problem using the Hertzian vector potential [5]. Although the method is quite general, a few simplifying assumptions are made, namely: the one-dimensional approximation, the usual antenna approximation neglecting higher order terms, and finally considering a monopole rather than dipole or any other elemental characteristics of the antenna involved. The method can be extended to more general cases as desired.

Consider a linear oscillator source of dimensions as shown in figure 1. The Hertzian vector  $\bar{\pi}$  is given by [5] (note  $\pi_x = \pi_y = 0$ ).

$$\pi_z(\bar{r}, t) = \frac{i e^{-i \omega t}}{4 \pi \epsilon \omega} \int_{-\frac{l}{2}}^{\frac{l}{2}} \frac{u(\xi) e^{i k r}}{r} d\xi$$

where  $u(\xi)$  is the current distribution (illumination function) along the one-dimensional element. Using the antenna approximation, the above reduces to:

$$\pi_z(\bar{r}, t) = i \frac{e^{-i \omega t + i k R}}{4 \pi \epsilon \omega R} \int_{-\frac{l}{2}}^{\frac{l}{2}} u(\xi) e^{-i k \xi \cos \theta} d\xi \quad (1)$$

Using equation (1) the field can be computed as follows:

$$\begin{aligned} \bar{E} &= \nabla(\nabla \cdot \bar{\pi}) + k^2 \bar{\pi}, \\ \bar{H} &= -i \omega \epsilon \nabla \times \bar{\pi} \end{aligned} \quad (2)$$

One term approximation of pertinent quantities are then given by:

$$E_\theta = - \frac{i \mu_0 \omega e^{i k R - i \omega t} \sin(\theta)}{4 \pi R} \int_{-\frac{l}{2}}^{\frac{l}{2}} e^{-i k \xi \cos \theta} u(\xi) d\xi \quad (3)$$

$$H_\phi = \sqrt{\frac{\epsilon_0}{\mu_0}} E_\theta$$

Equation (3) illustrates the well-known principle of multiplication of elemental patterns and array factors [6]. We further simplify equation (3) to a monopole case.

$$E_\theta = - \frac{i \mu_0 \omega e^{i k R - i \omega t}}{4 \pi R} \int_{-\frac{l}{2}}^{\frac{l}{2}} e^{-i k \xi \cos(\theta)} u(\xi) d\xi \quad (4)$$

For a desired asymptotic electrical field, the above equation, (4), gives the integral equation for the current distribution on the line source. After normalization the integral equation

\* This procedure is similar to Huygen's source approximation.

becomes:

$$S(\nu) = k \int_{-a}^a e^{i k \nu x} G(x) dx \quad \nu = \cos \theta, \quad -a < x < a \quad (5)$$

where  $a = l/2$ ,  $k = 2\pi/\lambda$ , and

$$E_\theta = -\frac{i \mu_0 c e^{i k r - i \omega t}}{4 \pi r} S(\nu)$$

Equation (5) is the main integral equation to be worked with. It has been derived here in a simple manner. Alternatively, it may be derived by considering discrete isotropic elements and approaching the continuum limit [7].

#### NUMERICAL PROCEDURE

Consider a symmetric function,  $G(x)$ . Equation (5) then reduces to:

$$2k \int_0^a G(x) \cos\left(\frac{2\pi \cos\theta x}{\lambda}\right) dx = S(\cos\theta), \quad \text{For } 0 < \theta \leq \pi \quad (6)$$

where  $G(x)$  is related to the excitation or illumination function in the aperture and  $k = 2\pi/\lambda$ , with  $\lambda$  being the wavelength. Introducing  $p = \pi x/a$ ,  $u = 2a \cos\theta/\lambda$ ,  $g(p) = G(x)$ ,  $F(u) = \frac{\lambda}{2a} S(\nu)$ , with the visible range  $-2a/\lambda \leq u \leq 2a/\lambda$ , equation (6) becomes:

$$\int_0^\pi g(p) \cos(p u) dp = \frac{F(u)}{2}, \quad u \geq 0 \quad (7)$$

Once  $g(p)$  is known,  $F(u)$  is defined for all  $u$  by equation (7). In our case, however, we have to solve for  $g(p)$ . The only information we have is asymptotic zeros of  $F(u)$ . These are derived by Taylor [4],[8] and obtained by the asymptotic method in appendix A1. It is clear from appendix equation (A1) that to avoid singularities in the unknown function, we should place the asymptotic zeros at the integer points, i.e.  $\alpha = 0$  for the form of  $g(p)$  near the end points of the aperture, corresponding to a pedestal illumination. We further assume  $g(p)$  is infinitely differentiable, finite, and non-zero at  $p = \pi$ .

Following Woodward's synthesis [3] let  $g(p)$  have the Fourier expansion:

$$g(p) = \frac{D_0}{2} + \sum_{m=1}^{\infty} D_m \cos(m p) \quad (8)$$

for  $0 < p \leq \pi$  otherwise  $g(p) = 0$  for  $p > \pi$

Substituting equation (8) into equation (7) we have:

$$F(u) = 2\pi \left( \frac{D_0 \sin(\pi u)}{\pi u} + \sum_{m=1}^{\infty} D_m \left( \frac{\sin(\pi(u+m))}{\pi(u+m)} + \frac{\sin(\pi(u-m))}{\pi(u-m)} \right) \right) \quad (9)$$

From (9) we have:

$$D_m = \frac{F(m)}{2\pi} \quad (10)$$

Denoting  $F(m)$  by  $F_m$ , we have from equation (8) and (10):

$$g(p) = \frac{1}{4\pi} (F_0 + 2 \sum_{m=1}^{\infty} F_m \cos(m p)) \quad (11)$$

Suppose that we decide that our far field pattern  $F(u)$  should have zeros at  $u=m$  for  $m > M$ , then :\*

$$g(p) = \frac{1}{4\pi} (F_0 + 2 \sum_{m=1}^M F_m \cos(m p)) \quad (12)$$

This gives:

$$F(u) = a_0(u) F_0 + \sum_{m=1}^M F_m a_m(u) \quad (13)$$

where  $a_0(u) = \sin(\pi u) / \pi u$

$$a_m(u) = \frac{\sin(\pi(u+m))}{\pi(u+m)} + \frac{\sin(\pi(u-m))}{\pi(u-m)} \quad (14)$$

From (13) and (14) we have  $F(n) = 0$  for  $n > M$ , and hence the Taylor's synthesis theorem is satisfied for finite value of the illumination function.\*\*

We normalize so that  $F(0) = F_0 = 1$ . Then we have  $M$  unknowns,  $F_1, F_2, \dots, F_M$  at our disposal to control the size of the sidelobes and the null points. Also the knowledge of  $F_i$  gives  $g(p)$  from equation (12). We now set up  $M$  simultaneous linear algebraic equations for the determination of the  $F_i$ . It is convenient to discuss this in three stages.

1.  $M$  nulls are prescribed. Then the sidelobes will be completely determined.
2. The maximum sidelobe size is prescribed. We will determine the solution by assuming that the maximum value of each sidelobe equals this value, for the first  $M$  sidelobes. Then the position of the first  $M$  nulls will be determined.
3. The general case when we prescribe the position of  $k$  near-nulls and the maximum (equal) size of the  $M-k$  sidelobes. Then the remaining  $M-k$  near-nulls and the other  $k$  (maximum) sidelobe sites will be determined.

#### Case 1

$M$  nulls  $u_i$  are prescribed. The advantage of starting with this case is that we can check whether the proposed method is going to work by solving the Taylor problem illustrated in this paper. From (13), recalling that  $F_0 = 1$ , and setting  $u = u_i$  (known) for  $i = 1, \dots, M$  we find:

$$\sum_{m=1}^M a_m(u_i) F_m = -a_0(u_i) \quad (15)$$

These are  $M$  simultaneous linear equations for  $F_m$ . Now see how we can solve the Taylor

\*  $M$  corresponds to  $\bar{n} - 1$  in Taylor's paper.

\*\* From here on our method differs considerably from that of T. T. Taylor. Taylor introduces a single stretching parameter for shifting  $M+1$  zeros. On the other hand, we use these  $M$  degrees of freedom to form the desired pattern.

examples by this method, assuming that all we know are the positions of the nulls,  $u_i$ ,  $i=1, \dots, M$ . These are given by [4]:

$$u_i = \sigma \sqrt{\left(i - \frac{1}{2}\right)^2 + A^2} \quad i = 1, \dots, M \quad (16)$$

where  $\cosh(\pi A)$  is the sidelobe ratio, and  $\sigma$ , the dilatational factor is given by:

$$\sigma = \frac{M + 1}{\sqrt{\left(M + \frac{1}{2}\right)^2 + A^2}} \quad (17)$$

The procedure is to substitute (16) into (15) for the solution of  $F_m$  and recover  $F(u)$  and  $g(u)$  from equations (13) and (12), respectively.

The results are plotted in figures 2 and 3. From the figures it can be seen that these results agree with Taylor's results. It should be noted that we have used only the values of  $u_i$ , and none of the machinery of Taylor to obtain a closed-form expression. The method will work for any prescribed  $u_i$ ; however the success or accuracy will depend on how the  $u_i$  are spaced. The condition number of the matrix involved in (15) was of the order one and as can be seen, the sidelobes decay and are not of equal magnitudes.

We next describe how to insert a double null at some point  $u_k$ . To do this we use series expansion, from equation (15):

$$\sum_{m=1}^M a_m(u_k) F_m = -a_0(u_k) \quad (18)$$

$$\left(\sum_{m=1}^M a_m(u_k + \Delta)\right) F_m = -a_0(u_k + \Delta) \quad (19)$$

$$a_m(u_k + \Delta) \approx a_m(u_k) + \Delta a'_m(u_k) \quad (20)$$

Inserting equation (20) into (19) and using (18), we have:

$$\sum_{m=1}^M a'_m(u_k) F_m = -a'_0(u_k) \quad (21)$$

We now have two equations (18) and (21) and we can delete two  $u_i$  in the neighborhood of  $u_k$ . The results of such an example are shown in figures 4 and 5. As should be noted, the procedure leads to a disturbance in the sidelobes in the neighborhood. The null is quite broad, however, which may be a good feature if signal from a broad source has to be eliminated.

#### Case 2:

Make the first  $M$  sidelobes all the same size. The success of Dolph-Tchebycheff synthesis was due to equal sidelobe design. Taylor's modification, the stretching of the nulls, to reduce the singularities at the aperture edges, leads to unequal sidelobes. The following procedure is suggested by Taylor's example shown in figure 2. The maxima for the sidelobes occur approximately at the half-way points in between the nulls. A more exact analysis may not be necessary because of the iteration procedure. Suppose the Taylor nulls are at  $u_1, \dots, u_M$ . Iterate as follows:

Step 1. Compute the following approximation to the position of the maxima:

$$v_i = 1/2(u_i + u_{i+1}), \quad i = 1, \dots, M \quad (22)$$

Step 2. Suppose that the sidelobe level required is  $k$ . Solve the following equation derived from equation (13).

$$\alpha k = F_0 a_0(\nu_i) + \sum_{m=1}^M a_m(\nu_i) F_m \quad (23)$$

where  $\alpha=1$  for  $i$  even, and  $\alpha=-1$  for  $i$  odd.

Step 3. Reconstruct the radiation pattern and compute the new nulls,  $u_1^1, \dots, u_M^1$ .

Repeat step 2 with  $u_i^1$  in place of  $u_i$  and continue until convergence of  $u_i^{(n)}$

For the cases considered it was found that the convergence was extremely fast. The results for the radiation pattern and the illumination function are given in figure 6 and 7. Table 1 gives some results after about the fifth iteration, for design purposes. The diagrams show that the beam is a little sharper than that given by Taylor. This table gives the location of the first zero for the values of sidelobe levels. It can be seen that the beam width is narrower since the zeros have shifted to the center.

### Case 3:

We now present general case. It should be clear how to generalize. The  $k$  near-nulls give us  $k$  equations in a straightforward way. The other nulls will be determined by the iteration procedure of case 2, but one will only be able to make  $M-k$  sidelobes equal to a prescribed maximum value. Figure 8, 9, 10 show how regions can be made very close to zero by placing several consecutive near nulls.

### EXTENSION OF THE METHOD TO CERTAIN TWO-DIMENSIONAL CASES.

The Taylor pattern for the axially symmetric problem has been derived in [8] and [9]. The numerical method described above can be easily extended to this case.

The remainder of this section is devoted to a two dimensional problem with a rectangular aperture. Using spherical coordinates  $\theta, \phi$  and a rectangular illumination region of area  $A=4ab$ , consider the following generalization of equation (1):

$$\Pi(x,y) = \frac{e^{ikR}}{R} \iint_A J(\xi,\eta) e^{i k (\xi \sin \theta \cos \phi + \eta \sin \theta \sin \phi)} d\xi d\eta \quad (24)$$

where  $x = R \sin \theta \cos \phi$ ,  $y = R \sin \theta \sin \phi$  We assume that the current distribution can be represented in the product form as:

$$J(x,y) = J_1(x) J_2(y) \quad (25)$$

so that  $\Pi(x,y)$  must also be of product form. Using the same symmetrical case as before, we have:

$$\Pi = \frac{4e^{ikR}}{R} \int_0^a J_1(\xi) \cos\left(\frac{kx}{R}\xi\right) d\xi \int_0^b J_2(\eta) \cos\left(\frac{ky}{R}\eta\right) d\eta \quad (26)$$

Let  $\pi\xi = a\alpha$ ,  $\pi\eta = b\beta$ ,  $2a/\lambda = N/2$ ,  $2b/\lambda = M/2$ . Equation (26) becomes:

$$\Pi(x,y) = \frac{4e^{ikR}}{R} \frac{ab}{\pi^2} \int_0^\pi J_1(\alpha) \cos(u\alpha) d\alpha \int_0^\pi J_2(\beta) \cos(v\beta) d\beta \quad (27)$$

where  $u = N/2 \sin \theta \cos \phi$ ,  $v = M/2 \sin \theta \sin \phi$ . The above equation (27) compares with

equation (7) and we use the same procedure to solve this equation. Some of the results of the iteration procedure for the two dimensional cases are illustrated in figure 11 and 12.

As mentioned in the introduction array thinning and subarraying will require the discretization of the illumination function. For adaptive beamforming, of a very large array, the numerical computation of each element may become prohibitive and sub-arraying may become necessary. A simple method called product integration can be illustrated as follows. Consider the space factor,  $f(u)$ , with illumination function  $i(x)$  as:

$$f(u) = \int_{-\pi}^{\pi} e^{ikxu} i(x) dx \quad (28)$$

For discretization, approximate  $i(x)$  by a constant  $I_r$  in the  $r^{\text{th}}$  interval. Equation (28) then becomes:

$$f(u) = \sum_{r=1}^R \int_{x_r - \Delta x/2}^{x_r + \Delta x/2} I_r e^{ikxu} du = \sum_{r=1}^R I_r f_r(u) \quad (29)$$

where

$$f_r(u) = 2e^{ikx_r u} \frac{\sin(1/2kru)}{ku} \quad (30)$$

Using special values of  $I_r$  it can be easily shown that the above method gives some well known solutions (see for example [10]). In our case we use the solution to the illumination function given by equation (11) for the above procedure.

## CONCLUSION

To the best of our knowledge the method given in this paper has not been investigated in the literature. The literature of radiation pattern synthesis is extensive. A partial review of various methods can be found in a recent reference [11], where a least square approximation has been investigated. The method given in our paper, due to its simplicity, will have a wide application in large array synthesis as well as for adaptive beamforming problems.

## APPENDIX A

T.T. Taylor obtained the asymptotic behavior of zeros from a rigorous discussion of the behavior of an integral of the following form for large  $t$

$$F(t) = \int_0^{\pi} ((\pi^2 - u^2)^\alpha g(u) \cos(tu)) du \quad (A1)$$

where  $g(u)$  is a perfectly smooth function. His result can be obtained by evaluating the above integral exactly for  $g(u) = 1$  and then taking the asymptotic value of the integral. The point is that only the behavior at  $u = \pi$  is important in determining the distribution of zeros.

*Lemma:* If we take :  $g(u) = 1$  then the pattern function  $F(t)$  has the asymptotic form:

$$\lim_{z \rightarrow \infty} F(z) \approx \frac{\Gamma(1+\alpha)z^2}{\pi z^{1+\alpha}} \cos(\pi(z \cdot 1/2(1+\alpha))) \quad (A2)$$

for  $\text{real}(z) > 0$ .

Proof: Consider

$$\int_0^{\pi} \cos(ut) (\pi^2 - u^2)^{\alpha} du = F(t) \quad (\text{A3})$$

Using Poisson's integral formula [12]:

$$J_{\nu}(x) = \frac{2(x/2)^{\nu}}{\sqrt{\pi} \Gamma(\nu + 1/2)} \int_0^{\pi/2} \cos(x \cos \alpha) \sin(\alpha^{2\nu}) d\alpha \quad (\text{A4})$$

we have

$$F(z) = \frac{\Gamma(1 + \alpha) \pi^{1 + \alpha}}{(2^{\nu/2} - \alpha_2^{\nu/2} + \alpha)} J_{\alpha - 1/2}(\pi z) \quad (\text{A5})$$

Using the one term asymptotic formula we get:

$$J_{\nu}(z) \approx \sqrt{\frac{2}{\pi z}} \cos(z - \frac{\pi}{4}(1 + 2\nu)) \quad (\text{A6})$$

Then we have:

$$F(z) = \frac{\Gamma(1 + \alpha) z^{\alpha}}{\pi z^{1 + \alpha}} \cos(\pi(z - 1/2(1 + \alpha))) \quad (\text{A7})$$

Equation (A7) shows the asymptotic zeros.

## APPENDIX B

The integral equation for the determination of the illumination function can also be derived from an energy point of view using Poynting vectors. Energy methods are useful for incorporating effects due to various design requirements for large array synthesis. In this appendix we derive the form of the integral equation that would arise due to energy methods directly from the one given in this paper.

$$S(\nu) = k \int_{-a}^a e^{i k \nu x} g(x) dx \quad \nu = \cos(\theta), -a < x < a \quad (\text{B1})$$

Multiplying (B1) with  $e^{-ikx}$  and integrating with respect to  $\nu$  from -1 to 1 and changing the order of integration we get:

$$1/2 F_1(t) = \int_{-a}^a g(x) \frac{\sin(k(x-t))}{(x-t)} dx \quad (\text{B2})$$

where

$$F_1(t) = \int_{-1}^1 S(\nu) e^{-i k \nu t} d\nu$$

Even though (B2) has a symmetric kernel, having quite a different form from the one in equation (7), the eigenvalues are the same, as discussed by Slepian and Pollack [13]. The method



of synthesis using the analysis of reference [13] are extensively analysed in [14].

#### REFERENCES

1. Dolph, C. L., "A Current Distribution for Broadside Arrays which optimize the Relationship between Beamwidth and Side-lobe levels". Proc. IRE., vol 34, pp 335-348; June, 1946.
2. Van der Maas, G. J., "A Simplified Calculation for Dolph-Tchebycheff Arrays". Journal of Applied Physics, Vol. 25, no. 1, pp 121-124; January, 1954.
3. Woodward, P.M., "A Method of Calculating the Field Over a Plane Aperture Required to Produce a Given Polar Diagram", Journal of IEE vol 93, pp 155-1558, 1946.
4. Taylor, T. T., "Design of Line-source Antennas for Narrow Beamwidth and Low Sidelobes", IRE Trans. on Antennas and Propagation Vol. Ap-3 pp 16-28 January 1955.
5. Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill Book Company, 1941.
6. Ma, M. T., *Theory and Application of Antenna Arrays*, John-Wiley and Sons, p4, 1973.
7. Ishimaru, Akira, "Theory of Unequally Spaced Arrays", IRE Transaction on Antennas and Propagation, Nov 19 1962, pp 691-702.
8. Taylor, T. T., "Design of Circular Aperture for Narrow Beamwidth and Low Sidelobes", IRE Trans. on Antennas and Propagation Vol. AP-8 pp 17-22 January 1960.
9. Elliot, R. S., and Stern, A. J., "Shaped Pattern From a Continuous Planar Aperture Distribution, IEE Proc., Vol. 135, pt. H, No. 6 December 1988.
10. Steinberg, B. D., *Principles of Aperture and Array System Design* John-Wiley and Sons, p 62, NY, 1976.
11. Guy, R.F.E., "General Radiation Pattern Synthesis Techniques for Array Antennas of arbitrary Configuration and Element Type", IEE Proceedings, vol. 135, pp 241-248, August, 1988.
12. Magnus, W., and Oberhettinger, F., *Formulas and Theorems for the Functions of Mathematics*, p26, Chelsea Publishing Co., NY, NY, 1949.
13. Slepian, d., and Pollak, H.O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty -I", The Bell System Technical Journal, vol 40, pp 43-64, January, 1961.
14. Rhodes, D.R., *Synthesis of Planar Antenna Sources*, Clarendon Press, Oxford University Press, London, 1974.

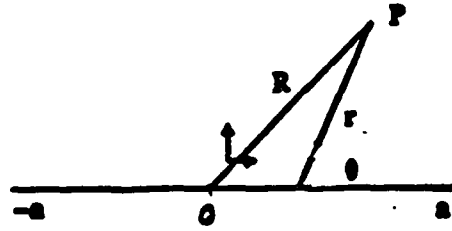


Figure 1. Line source.

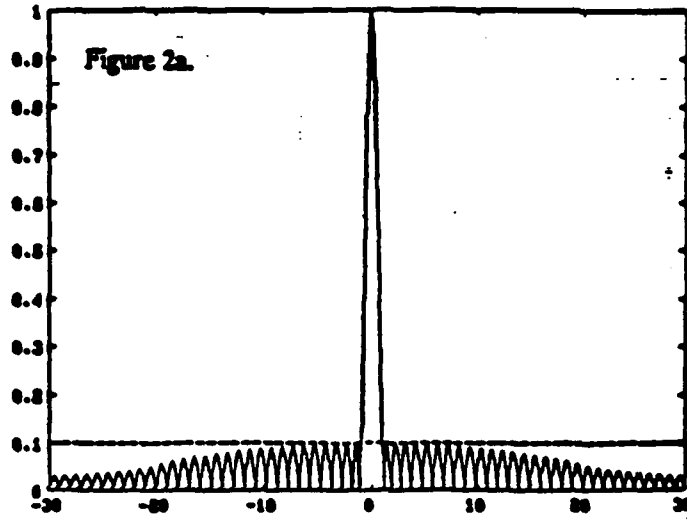


Figure 2a. Recovery of Taylor solution by numerical method.

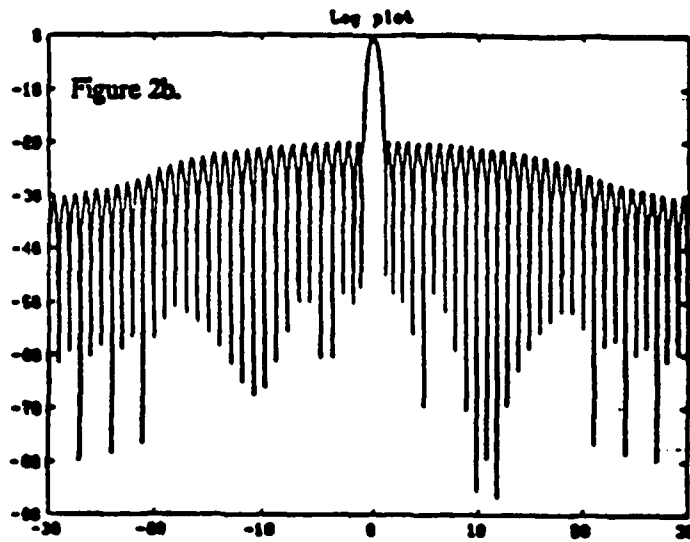


Figure 2b. Recovery of Taylor solution by numerical method.

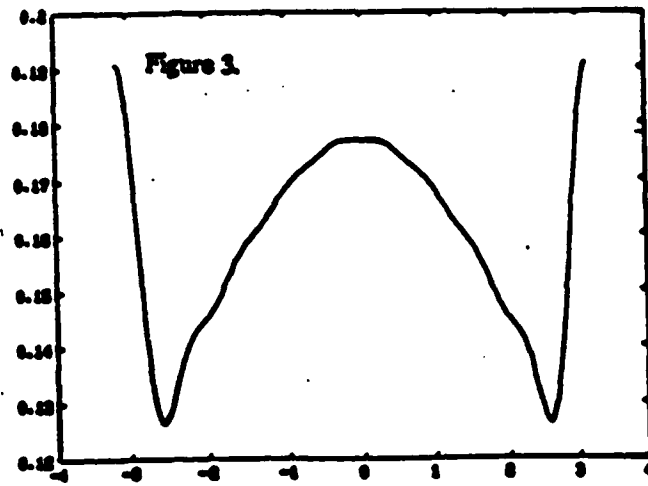


Figure 3. Illumination function for the Taylor problem.

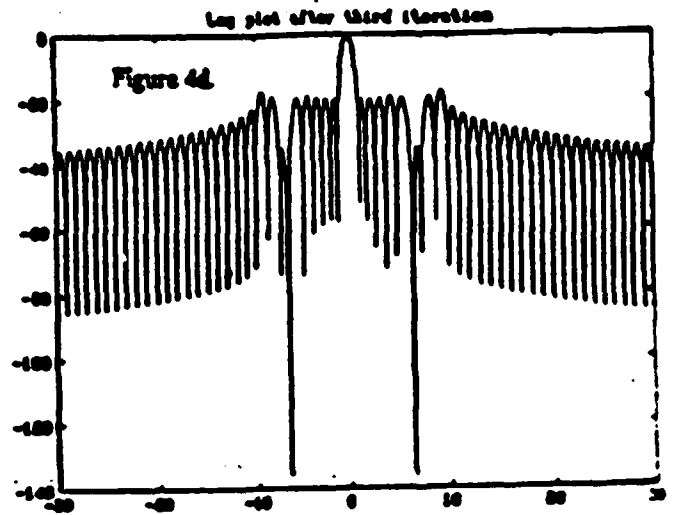
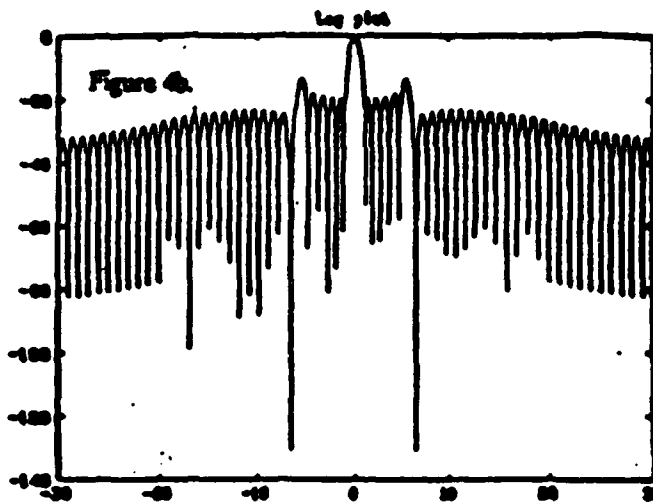
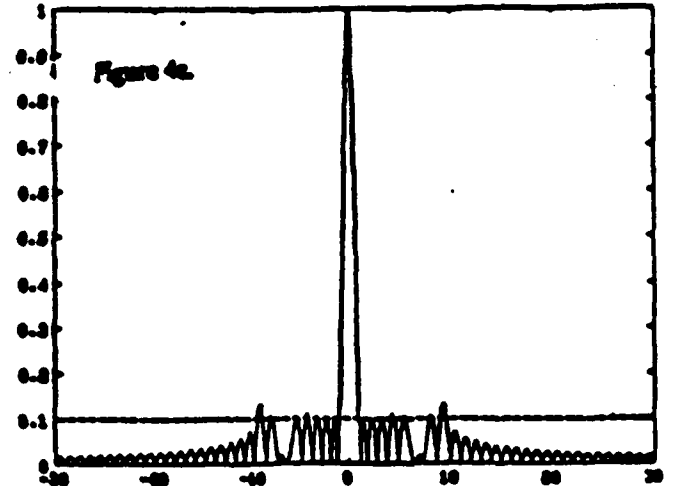
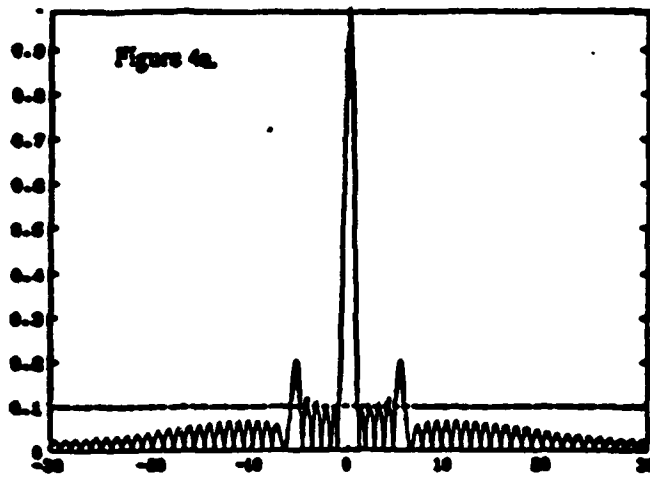


Figure 4 Spec factor for the case of a double null at 6.5

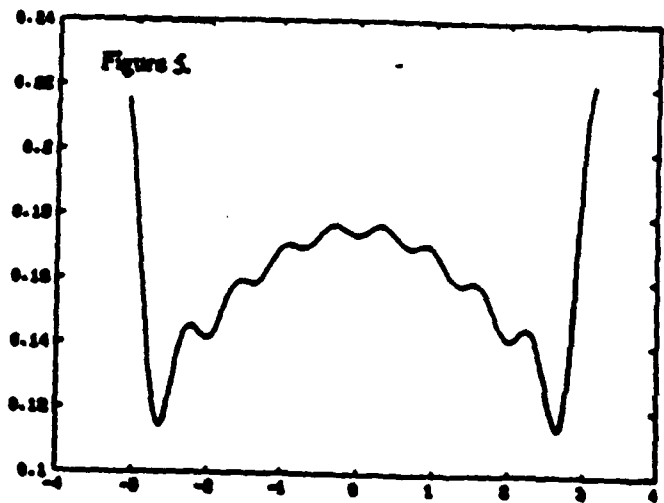


Figure 5. Illumination function for figure 4

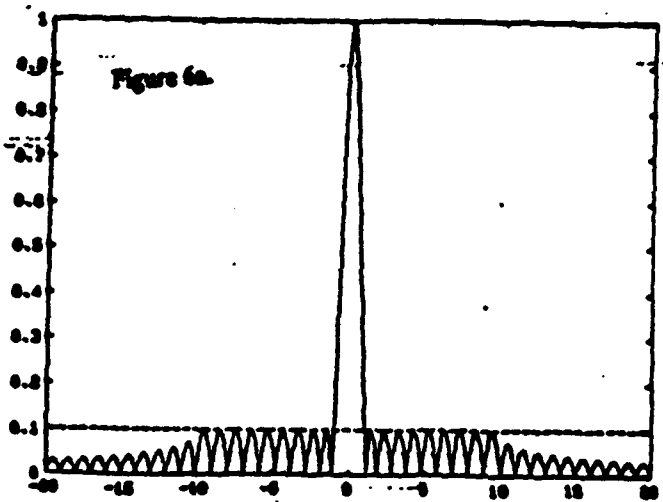


Figure 6a. Iterated Taylor solution with equal design sidelobes.

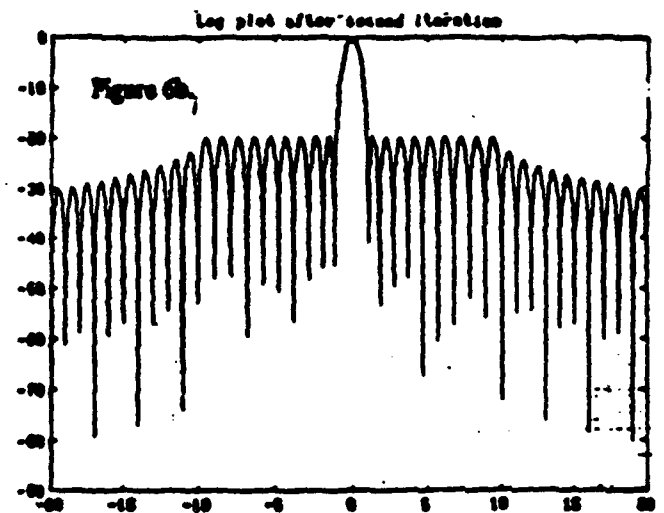


Figure 6b. Iterated Taylor solution with equal design sidelobes.

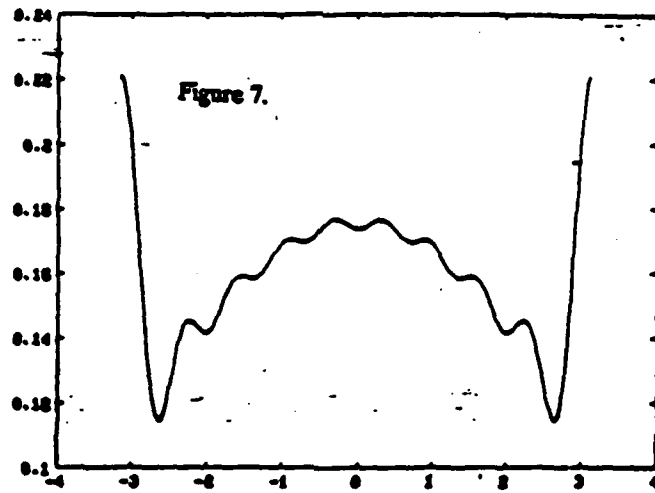


Figure 7. Illumination pattern for figure 6.

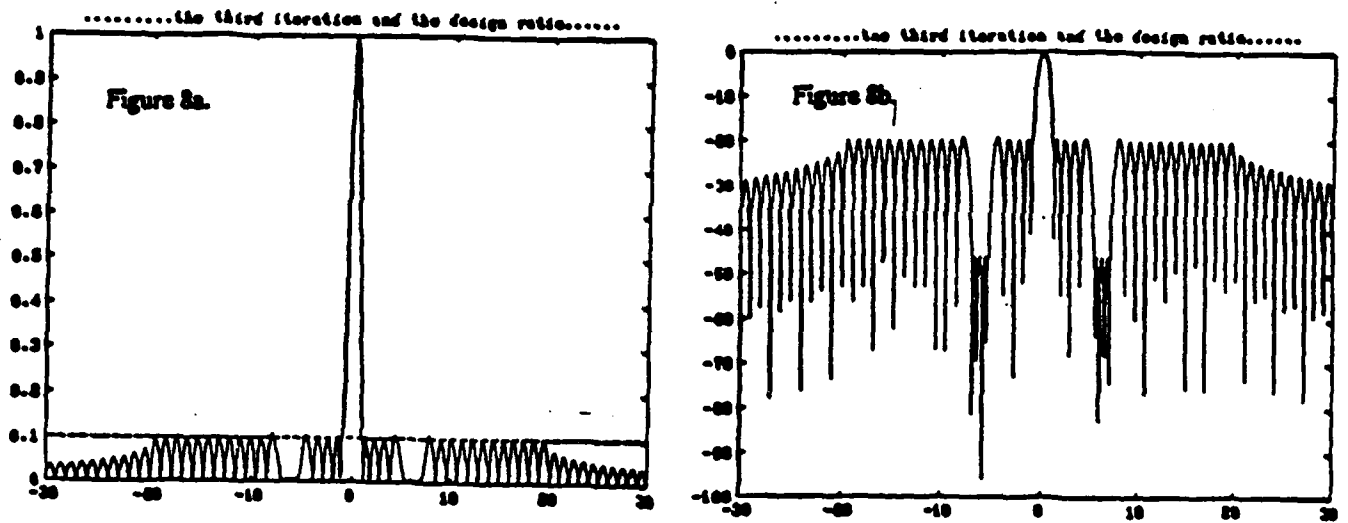


Figure 8. Iterated space factor for three near equal nulls.

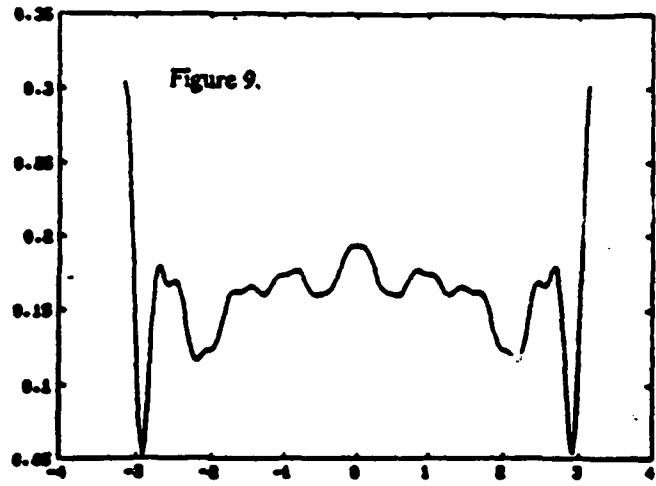


Figure 9. Illumination patterns for figure 8.

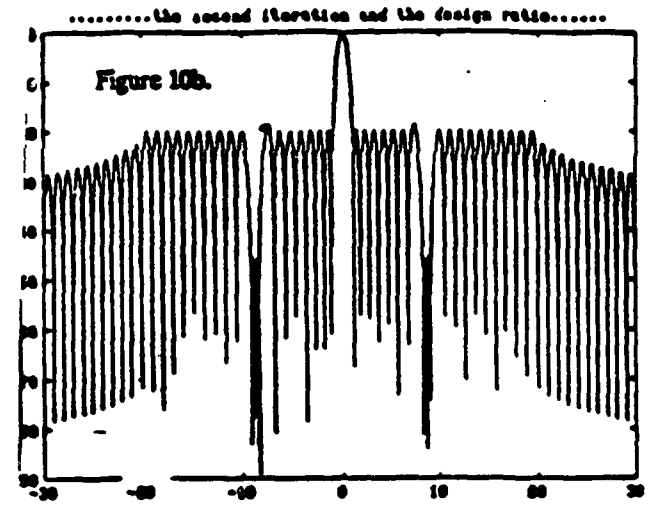
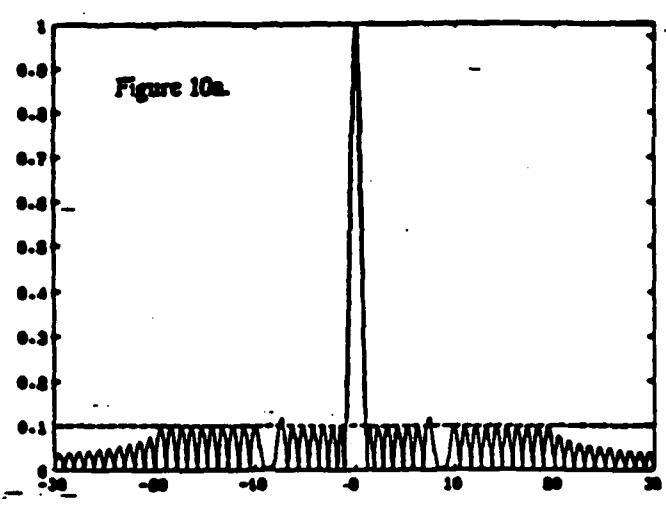


Figure 10. Iterated space factor for two near equal wells.

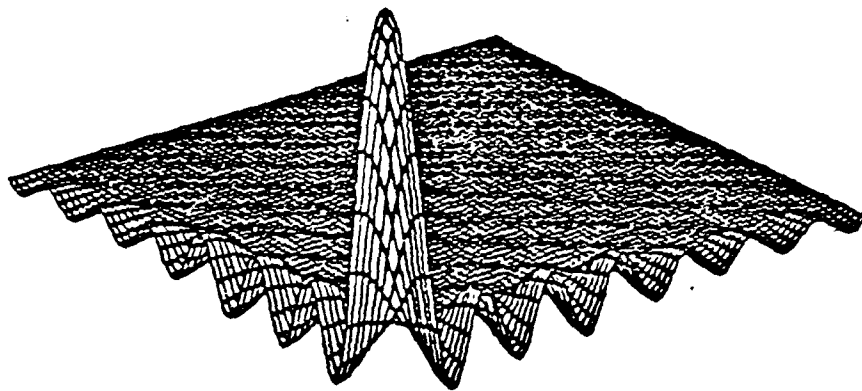


Figure 11.

Figure 11. Two dimensional space factor in  $u-v$  plane.

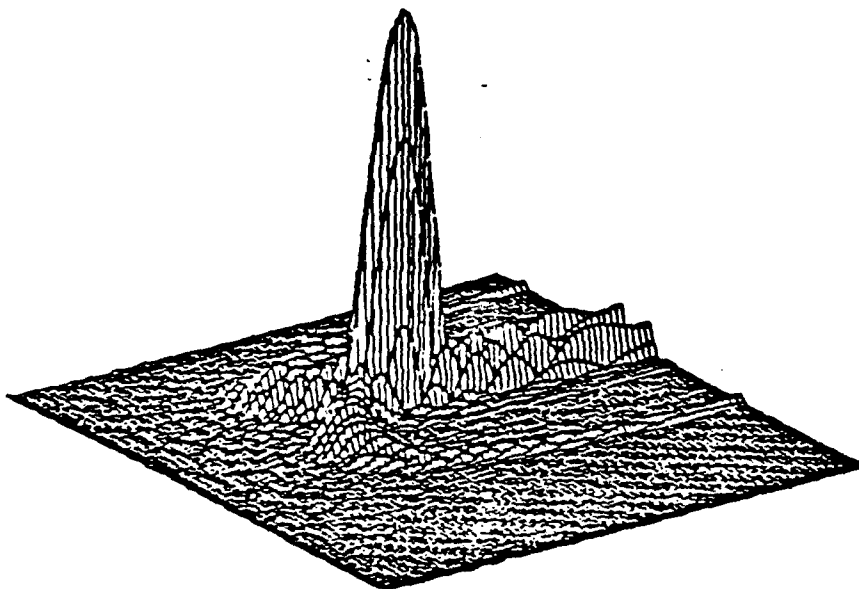


Figure 12.

Figure 12. Two dimensional space factor in  $\theta-\phi$  plane.

**OPTIMIZED TAYLOR ZEROES (Only 1st Shows)**

$\bar{n}$	dB		
	10	20	(TAYLOR) 40
5	.81558 (.84292)	1.13733 (1.16962)	1.8149 (1.8302)
6	.80656 (.82983)	1.12748 (1.15659)	1.8133 (1.8347)
7	.80021 (.82047)	1.12016 (1.14651)	1.8097 (1.8337)
8	.79551 (.81345)	1.114537 (1.138582)	1.8058 (1.8306)
9	.79187 (.80800)	1.110093 (1.132203)	1.8020 (1.8269)
10	.788993 (.803647)	1.1065 (1.1270)	1.7986 (1.8231)

Table 1. Values of first zeros obtained by numerical analysis  
with uniform sidelobes and compared with those of Taylor zeros.



# Computation of Leading Eigenspaces for Generalized Eigenvalue Problems

Abraham Kribus  
Sibley School of Mechanical and Aerospace Engineering  
Cornell University  
Ithaca, NY 14853

## Abstract

A method for computing leading eigenvalues (having the largest real part) and their eigenvectors for large generalized eigenvalue problems is presented. A linear fractional transformation is used to map a group of leading eigenvalues into dominant eigenvalues (having the largest modulus). The Dominant eigenvalues of the transformed problem are computed by Stewart's (1976) Simultaneous Iteration. Each iteration involves matrix-vector multiplication and the solution of a linear system, which can be done efficiently if the matrices involved are sparse or have some special structure. Convergence properties are similar to those of the inverse power iteration: the method requires an estimate for the region in the complex plane containing the desired eigenvalues, and converges rapidly when a good estimate is available. The amount of work is also comparable to that of the basic inverse iteration, which is significantly less than that required for full eigensolution. Examples from hydrodynamic stability demonstrate convergence rates, computation time and the ability to resolve simultaneously groups of leading eigenvalues.

## 1. Introduction

A generalized eigenvalue problem has the form:

$$A x = \sigma B x \quad (1.1)$$

where  $A, B \in \mathbb{C}^{n \times n}$  are general complex matrices. In many applications these matrices will have some useful structure, such as symmetry or sparsity.

Let the *Leading* Eigenvalues of (1.1) be those having the largest *real part*; the more common term, *Dominant* Eigenvalues, refers to those having the largest *modulus*. In some applications, only a few leading eigenvalues of (1.1) are sought; for example, in linear stability problems, the real part of  $\sigma$  is the growth rate, and the eigenvectors of the leading eigenvalues represent the most unstable modes.

Traditional methods for solving (1.1) usually involve finding all the eigenvalues, using the Q-Z algorithm (see IMSL or other numerical analysis libraries) and then sorting by the

real part. This involves  $O(n^3)$  work, where  $n$  is the order of the matrices, and becomes expensive or impractical for large  $n$ ; little or no advantage can be taken of sparsity or other structure of  $A$  and  $B$ .

Several methods exist for extracting selected eigenvalues and eigenvectors of standard eigenvalue problems, i.e. when  $B$  is invertible (see, for example, Golub and Van Loan, 1983.) Power and Lanczos methods compute the dominant eigenvalues; inverse iteration can find the eigenvalues closest to a given point in the complex plane and their eigenvectors. These are not directly applicable to the problem of computing the *leading* eigenvalues.

Recently, an integration method was proposed (Goldhirsch *et al.* 1987) for the leading eigenvalues of a standard eigenvalue problem. This method is simple and elegant; however, its convergence may become very slow (or, alternatively, the size of the reduced problem may become very large) if the separation of the eigenvalues is small. Another problem may arise if the problem is defective, i.e. a leading eigenvalue has generalized eigenvectors; in this case, the integration method may return inconclusive or inconsistent results.

## 2. The Dominance Mapping Method

This method attempts to address the problems of the form (1.1) which are not solved efficiently by the other methods mentioned above. It will work for singular  $A$  and  $B$ ; for defective problems; it will take full advantage of the structure of the matrices; and it allows some control over convergence rates. There are a few restrictions, however, which will be discussed below.

The eigenvalues in the complex  $\sigma$ -plane can be mapped to a  $\lambda$ -plane by the linear fractional transformation:

$$\sigma = \beta + \alpha \frac{\lambda - 1}{\lambda + 1} \quad (2.1)$$

where  $\alpha$  is a real positive, and  $\beta$  a complex, constant. The important effect of this linear fractional mapping is to map the half-plane to the left of  $\sigma = \beta$  to the inside of the unit circle in the  $\lambda$  plane, as seen in fig. 1. If  $m$  leading eigenvalues are required, and we select  $\beta$  such that:

$$\operatorname{Re}(\sigma_i) \begin{cases} > \operatorname{Re}(\beta) & i=1 \dots m \\ < \operatorname{Re}(\beta) & i=m+1 \dots n \end{cases}$$

then the corresponding  $m$  eigenvalues will be *dominant* in the  $\lambda$  plane:

$$|\lambda_i| \begin{cases} > 1 & i=1\dots m \\ < 1 & i=m+1\dots n \end{cases}$$

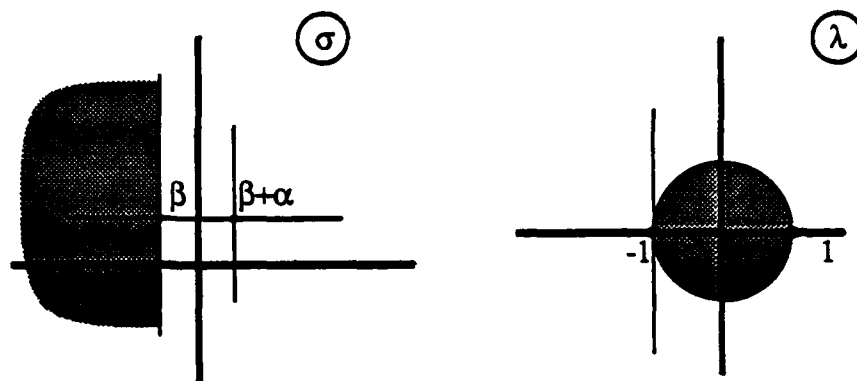


Figure 1: The Dominance Mapping (2.1)

The eigenvalue problem for  $\lambda$  is of standard form:

$$C u \equiv C_1^{-1} C_2 u = \lambda u \quad (2.2)$$

where:

$$C_1 = -[A - (\alpha + \beta) B]$$

$$C_2 = [A + (\alpha - \beta) B]$$

The problem of computing the leading eigenvalues of (1.1) becomes that of computing the *dominant eigenvalues* of (2.2); the methods mentioned in §1 can now be applied. We used Stewart's (1976) version of Simultaneous Iteration, which finds dominant invariant spaces of a general, non-hermitian, possibly defective  $C$ .

A transformation similar to (2.1) was proposed by Jennings (1977), in the context of converting a quadratic eigenvalue problem to standard form. Jennings (and no one else, to the best of the author's knowledge) has not made the second step of applying a dominant eigenvalue method to a transformed problem equivalent to (2.2).

The mapping constants  $\alpha$  and  $\beta$  allow the user some control over the rate of convergence and the order in which the leading eigenvalues emerge during the iteration. The user must have an estimate of where in the complex plane the leading eigenvalues reside;  $\beta$  is set to the left of this region. The point  $c = \beta + \alpha$  is a singular point of (2.1) which maps to infinity in the  $\lambda$  plane; eigenvalues close to  $c$  will map to very large

modulus in the  $\lambda$  plane, and will converge rapidly during the iteration of (2.2).  $\alpha$  should be set, therefore, so that  $c$  is near the center of the leading region or near the most important eigenvalue.

The following algorithm computes  $m$  leading eigenvalues and eigenvectors of (1.1), using the Dominance Mapping and Simultaneous Iterations:

1. estimate leading region; select  $\alpha, \beta$
2. perform L-U decomposition of  $C_1 = -[A - (\alpha + \beta) B]$ ;  
(use the structure of  $A$  and  $B$ )
3. select  $m$  initial vectors  $u^{(0)} \in \mathbb{C}^{n \times m}$
4. Simultaneous Iteration on  $C u = \lambda u$ :  
for each multiplication  $u^{(k+1)} = C u^{(k)}$  do:
  - 4.1 multiply:  $v = C_2 u^{(k)}$
  - 4.2 solve the system:  $C_1 u^{(k+1)} = v$
5. map converged  $\lambda_i \rightarrow \sigma_i$ .

### 3. Singularities in the Dominance Mapping

The algorithm of §2 may fail in two cases, corresponding to the two singularities of the mapping (2.1): the point  $\sigma=c$ , which maps to infinity in  $\lambda$ , and  $\lambda=-1$  which maps to infinity in  $\sigma$ .

When  $|c - \sigma_i| < \epsilon$  for some  $i \leq m$ , for a small (machine-dependent)  $\epsilon$ , then the matrix  $C$  will be ill-conditioned or numerically singular. This is easily remedied by a small change in  $\alpha$ , which does not significantly affect any other properties of the mapping.

When  $|\text{Im}(\sigma_i - c)| \gg 1$  for some  $i \leq m$ , the corresponding  $\lambda$ -eigenvalue is close to the singular point  $\lambda=-1$ . This implies that its separation from the subdominant eigenvalues inside the unit circle is small, often so small that convergence is impractical. Some improvement may result if we increase  $\alpha$ ; but this may decrease the modulus of other dominant  $\lambda$ -eigenvalues and slow down their convergence. In a case where leading eigenvalues are widely separated in the imaginary direction, it may be necessary therefore to restart the iteration with different  $\beta$  values to resolve separate clusters of leading  $\sigma$ -eigenvalues. An example of this situation appears in §5 below.

### 4. Example

The performance of the DM method can be demonstrated by observing the amount of work needed to resolve a fixed subset of leading eigenvalues, as the order of the problem

increases. The following example includes tridiagonal matrices of increasing size  $n$ , all having two leading eigenvalues:

$$\begin{aligned} \sigma_1 &= 1.2 & (4.1) \\ \sigma_2 &= 1.1 \\ \operatorname{Re}(\sigma_i) &\leq 1 \quad \text{for } i = 2 \dots n. \end{aligned}$$

Selecting  $\alpha = 0.3$ ,  $\beta = 1.0$  isolates  $\sigma_1, \sigma_2$ . The problem was solved first using the traditional QZ routines (IMSL), then using the DM method but treating the matrices as dense, and finally taking full advantage of the structure. The results are shown in Figure 2.

The savings in computing time relative to the full eigensolution can be significant: at  $n=100$ , only  $\frac{1}{5}$  of the work is necessary even without exploiting the band structure; the work is reduced by more than an order of magnitude when the structure is used.

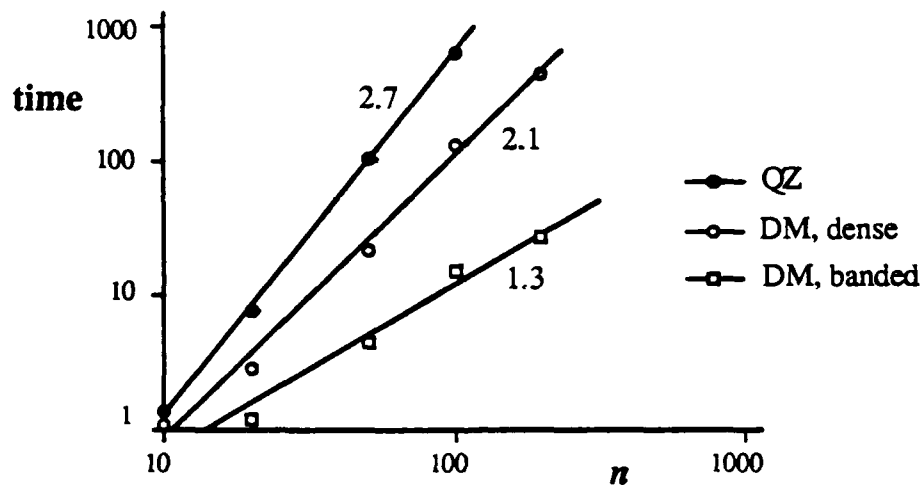


Figure 2: Work to resolve the two leading eigenvalues (4.1), using three solution algorithms

### 5. Application to the Orr-Sommerfeld Equation

The Orr-Sommerfeld equation:

$$(D^2 - \alpha^2)^2 \psi = i\alpha R [(U-c)(D^2 - \alpha^2)\psi - U''\psi] \quad (5.1)$$

describes the hydrodynamic stability of parallel shear flow (see, for example, Drazin and Reid 1981.) High accuracy eigenvalues were computed by Orszag (1971) for plane

Poiseuille flow with  $R=10000$  (Reynolds number) and  $\alpha=1$  (streamwise wavenumber.)  
 The location of the twelve least stable eigenvalues are shown in figure 3.

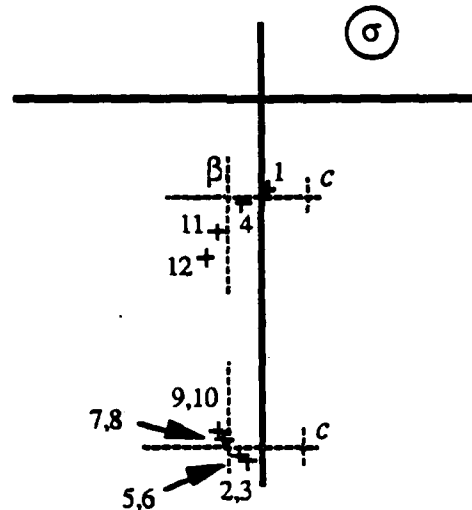
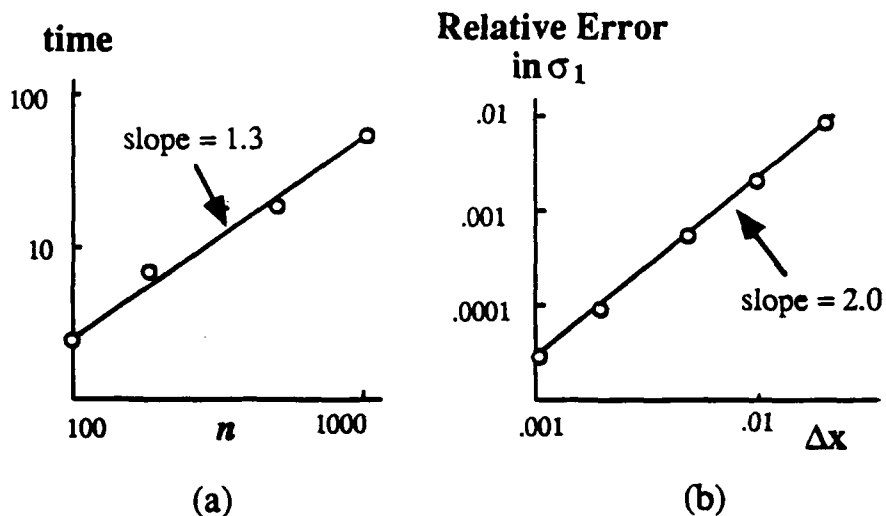


Figure 3: Poiseuille flow eigenvalues.  $R=10000$ ,  $\alpha=1.0$

Equation (5.1) is discretized using central differences (a spectral method may be more appropriate in this specific case, as in Orszag (1971), but the banded finite-difference matrix is a good example of candidate problems for the DM method.) The eigenvalue  $c$  is replaced by  $\sigma = -ic$ , to conform with the definitions of (1.1).

When  $\text{Im}(\beta) = \text{Im}(\sigma_1)$  (the upper dashed lines in fig. 3), eigenvalues 1 and 4 were the first to converge; 2, 3, 5 and 6 took longer to converge, since the imaginary part separation brought their  $\lambda$  counterparts close to the singular point  $\lambda = -1$ . For  $\text{Im}(\beta) = \text{Im}(\sigma_2)$  (the lower dashed lines in fig. 3), the order was reversed: first eigenvalues 2, 3, 5 and 6 and then 1 and 4. In both cases, the first group converged within 10 to 15 iterations, regardless of the number of grid points; the second group took much longer to converge.

The error associated with convergence of the  $\lambda$ -iteration was not significant in our computations. Using a stopping criterion of  $\|Cu - \lambda u\| \leq 10^{-3}$ , the leading  $\sigma$ -eigenvalues were converged to at least 5 digits. The discretization error of the finite-differencing (compared to Orszag's results) is proportional to  $\Delta x^2$ , as expected. The time to resolve the most unstable eigenvalue and its discretization error vs. the grid resolution are plotted in figure 4.



**Figure 4:** (a) the time to compute the most unstable eigenvalue of (5.1)  
 (b) discretization error vs. the grid interval  $\Delta x$

## 6. Conclusion

Using the Dominance Mapping and a Power Iteration method we can compute leading eigenvalues and eigenvectors of large generalized eigenvalue problems. This method can be more efficient than a full eigensolution even for general matrices, but is especially attractive when the matrices have a structure that can save work in the Gaussian elimination and matrix multiplication steps. The DM method can be applied to singular and defective problems that may cause failure or slow convergence in other methods.

Use of the DM method is restricted, however, to cases where an estimate for the leading eigenvalues is available. When this estimate shows a wide distribution of leading eigenvalues along the imaginary direction, several passes may be necessary with different mapping parameters to properly resolve all leading eigenvalues, as demonstrated for the Orr-Sommerfeld problem.

## Acknowledgment

This work was supported by the U. S. Army Research Office Mathematical Sciences Institute at Cornell University and by the Air Force Office of Scientific Research under contract AFOSR-87-0255.

## **References**

- Drazin, P.G. and Reid, W.H. (1981). *Hydrodynamic Stability*, Cambridge University Press, New York.
- Goldhirsch, I., Orszag, S.A., and Maulik, K. (1987). *An Efficient Method for Computing Leading Eigenvalues and Eigenvectors of Large Asymmetric Matrices*, J. Sci. Comput. 2(1), 33.
- Golub, G.H. and Van-Loan, C.F. (1983). *Matrix Computations*, The Johns Hopkins University Press, Baltimore.
- Jennings, A. (1977). *Matrix Methods for Engineers and Scientists*, John Wiley & Sons, Chichester.
- Orszag, S.A. (1971). *Accurate Solution of the Orr-Sommerfeld Stability Equation*, J. Fluid Mech. 50, 689-703.
- Stewart, G.W. (1976). *Simultaneous Iterations for Computing Invariant Subspaces of Non-Hermitian Matrices*, Numer. Math. 25, 123.



## APPLICATIONS OF FIBONACCI SEQUENCES AND TILING

Joseph Arkin, David C. Arney, Lee S. Dewald, and Charles Kennedy  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996-1786

**ABSTRACT** In the first part of this paper we develop some results for the recurring sequence

$$B(n) = a B(n-1) + b B(n-2),$$

$$\text{with } a \text{ and } b \neq 0, \quad B(0) = 1, \quad \text{and} \quad B(1) = 1,$$

and show a relationship between this sequence and the simple network of resistors known as a ladder network. Then, using certain values for the coefficients  $a$  and  $b$ , we show tiling of the plane using this general recurring sequence.

In the second part of the paper, using the Fibonacci recurring sequence and Fibonacci polynomials, we investigate the paths of light rays incident upon two stacked glass plates. We model the number of distinct paths of light rays, number of reflections of light rays, and number of crossings of the interface between the glass plates using both homogeneous and nonhomogeneous recurrence relations. Once again, we are able to tile the plane using these sequences.

### 1. Introduction

The Fibonacci numbers are a sequence of real numbers in which each successive value in the sequence is defined to be the sum of the previous two values. The first Fibonacci number is zero and the second is 1. The recurring sequence of Fibonacci numbers, then, can be described as

$$F(n) = F(n-1) + F(n-2) \quad \text{with} \quad F(0) = 0 \quad \text{and} \quad F(1) = 1. \quad (1)$$

The Fibonacci sequence exhibits many interesting mathematical properties. Some of these are:

- i) Every third Fibonacci number is even,

- ii) No two consecutive Fibonacci numbers have a common factor,
- iii) The sum of any ten consecutive Fibonacci numbers is always divisible by 11 [11].

Over the years, many physical and natural phenomena have been identified to exhibit properties which can be modeled with the Fibonacci sequence. One of the classic applications is a population model. Assume that a pair of a species is isolated and allowed to breed. Assume further that for each time period the original pair gives birth to a new pair. Each new pair matures and two time periods after birth gives birth to yet another new pair. If we assume no deaths, the number of pairs of the species at the beginning of each time period is a Fibonacci number.

Plant life exhibits many properties which can be modeled by Fibonacci numbers. The number of petals in the flowers of many plants are Fibonacci numbers. The number of spirals in the scale patterns of pine cones is usually a Fibonacci number [12].

The applications of the Fibonacci sequence are extensive in many diverse fields of mathematics. One area which is especially rich in the use of Fibonacci numbers is number theory. Fibonacci numbers are widely used in solutions to the problem of Diophantus [6,8]. Fibonacci numbers appear in the analysis of Aitken Acceleration, a numerical analysis procedure which speeds up the convergence of some sequences [13].

One of the better known applications of Fibonacci numbers is in the field of optimization of functions of a single variable. One very popular method for resolving the associated line search problem is the Fibonacci search method. Assuming only that a function  $g(x)$  is unimodal on an interval, the Fibonacci search is a method to successively select  $N < \infty$  measurement points so that one can determine the smallest possible region of uncertainty in which the modal value must lie. If  $d_1$  is the initial width of the interval of uncertainty, then after  $k \leq N$  measurements

$$d_k = \left( \frac{F(N - k + 1)}{F(N)} \right) d_1 \tag{2}$$

is the width of the interval where the integers;  $F(m)$ , are the Fibonacci numbers generated from the recursion in (1).

A practical use of the Fibonacci search was reported by Braverman [10] to locate the sample size that maximized the expected net gain from sampling in a Bayesian decision problem.

Some extensions of Fibonacci numbers and Fibonacci polynomials are related to certain classes of discrete probability distributions which have important applications in reliability theory. The Fibonacci numbers of order- $k$  are defined as

$$F_{(a)}^{(k)} = \begin{cases} F^{(k)}(n-1) + \dots + F^{(k)}(1) & 2 \leq n \leq k+1 \\ F^{(k)}(n-1) + \dots + F^{(k)}(n-k) & n \geq k+2 \end{cases} \quad (3)$$

with  $F^{(k)}(0) = 0$  and  $F^{(k)}(1) = 1$ .

Fibonacci polynomials of order- $k$  are likewise defined as

$$F_n^{(k)}(x) = \begin{cases} \sum_{i=1}^n x^{k-i} F_{n-i}^{(k)}(x) & 2 \leq n \leq k+1 \\ \sum_{i=1}^k x^{k-i} F_{n-i}^{(k)}(x) & n \geq k+2 \end{cases} \quad (4)$$

with  $F_0^{(k)}(x) = 0$  and  $F_1^{(k)}(x) = 1$ .

The Fibonacci polynomials are obtained from (4) by setting  $k = 2$ . In order to obtain the Fibonacci numbers of order  $k$ , set  $x = 1$  in (4).

Now the Fibonacci polynomials of order- $k$  have been generalized to a Fibonacci-Type polynomial of order- $k$  [15] as follows

$$f_n^{(k)}(x) = \begin{cases} x \left[ f_{n-1}^{(k)}(x) + \dots + f_1^{(k)}(x) \right] & 2 \leq n \leq k+1, \\ x \left[ f_{n-1}^{(k)}(x) + \dots + f_{n-k}^{(k)}(x) \right] & n \geq k+2. \end{cases} \quad (5)$$

It is polynomials of the form in (5) that occur in certain discrete density functions [15]. For example, let  $X$  be the random variable with the geometric distribution of order- $k$  and parameter  $p \in (0,1)$ . We have

$$P(X = n+k) = p^{n+k} f_{n+1}^{(k)} \left( \frac{1-p}{p} \right) \quad n \geq 0. \quad (6)$$

Likewise, let  $L_n$  be the random variable representing the length of the longest string of numbers in  $n \geq 1$  Bernoulli trials. We have

$$P(L_n \leq k) = \frac{p^{n+1}}{1-p} f_{n+2}^{(k+1)} \left( \frac{1-p}{p} \right) \quad 0 \leq k \leq n. \quad (7)$$

The reliability of a system may be increased without duplicating the system by using a "consecutive- $k$ -out-of- $n$  : F system." [16]. This is a system that fails if and only if  $1 \leq k \leq n$  consecutive components fail. This system was first introduced in connection with telecommunications and oil pipeline systems.

One simple result is cited. If the system consists of  $n$  components arrayed linearly, operate independently and identically with probability  $p$ , then the probability of failure is one minus the probability that in  $n$  trials there is no occurrence of a string of  $k$  failures:

$$P(F) = 1 - P(L_n \leq k-1) = 1 - \frac{(1-p)^{n+1}}{p} f_{n+2}^{(k)} \left( \frac{p}{1-p} \right) \quad 0 \leq k \leq n. \quad (8)$$

The Fibonacci numbers and polynomials also occur in the models and their solutions in the fields of engineering and applied physical science. In this paper we discuss two such applications.

For the first application we use the recurring sequence

$$B(n) = a B(n-1) + b B(n-2), \quad (9)$$

$$\text{with } a \text{ and } b \neq 0, \quad B(0) = 1, \quad \text{and} \quad B(1) = 1,$$

to model the simple network of resistors known as a ladder network. Then, using certain values for the coefficients  $a$  and  $b$ , we show a tiling of the plane using this general recurring sequence.

For the second application we use a recurring sequence similar to (1) and Fibonacci polynomials to investigate the paths of light rays incident upon two stacked glass plates. We model the number of distinct paths of light rays, number of reflections of light rays, and number of crossings of the interface between the glass plates using homogeneous and nonhomogeneous recurrence relations. Once again, we are able to tile the plane using these sequences to demonstrate a geometric interpretation.

## 2. Ladder-Network of Resistors

The ladder-network of resistors shown in Figure 1 is an important network in communication systems [2,9]. The following definitions are provided:

- a) The resistance through the resistor =  $R_1$ ,
- b) Voltage across the resistor =  $e_1$ ,
- c) The attenuation (input voltage/output voltage) =  $A$ ,
- d) the output impedance =  $z_0$ ,
- e) the input impedance =  $z_1$ .

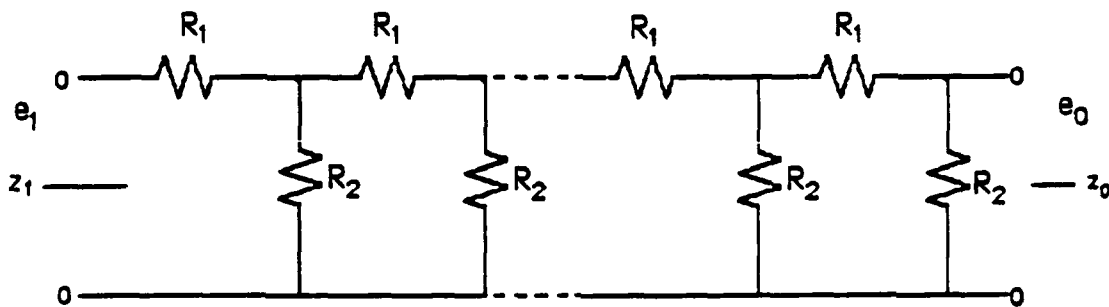


Figure 1: Schematic showing location of resistors in a ladder network.

A model of ladder-networks obtained from Kirchhoff's & Ohm's Laws was developed in [2,9,14]. The model for the attenuation is the generalized Fibonacci sequence

$$B_n = k_1 B_{n-1} + k_2 B_{n-2} \quad \text{with } B_0 = 1 \quad \text{and } B_1 = k_1. \quad (10)$$

The coefficients  $k_1$  and  $k_2$  depend on resistances  $R_1$  and  $R_2$ .

In this model,  $B_{2n}$  is the attenuation of the circuit with  $n$  pairs of  $R_1$  and  $R_2$  resistors.

First, we develop a useful result for solving our model.

The following relationship always holds for (10).

$$B_n^2 - B_{n+1} B_{n-1} = (-1)^n k_2^n. \quad (11)$$

where

$$B_0 = 1, B_1 = k_1, B_2 = k_1^2 + k_2, \dots \quad (\text{cf}[2,5]).$$

Proof

Another way of stating (10) is

$$\frac{1}{(1 - k_1 x - k_2 x^2)} = B_0 + B_1 x + B_2 x^2 + \dots \quad (12)$$

where  $x^2 - k_1 x - k_2$  is the auxiliary polynomial of the 2nd order homogeneous difference equation and therefore

$$x^2 = k_1 x + k_2. \quad (13)$$

Now, multiplying (13) through by  $x$ , we have

$$x^3 = k_1 x^2 + k_2 x$$

and replacing  $x^2$  with the values in (13) leads to

$$x^3 = (k_1^2 + k_2) x + k_1 k_2 \quad \text{or} \quad x^3 = B_2 x + k_1 k_2. \quad (14)$$

Then, continuing in the same way that we obtained (14), we have

$$x^{n+1} = B_n x + K \quad (\text{where } K \text{ is some constant}). \quad (15)$$

Solving for (13), we have the following two roots:

$$x_1 = \frac{1}{2} \left[ k_1 + \left( k_1^2 + 4k_2 \right)^{\frac{1}{2}} \right] \quad (16a)$$

and

$$x_1 = \frac{1}{2} \left[ k_1 - \left( k_1^2 + 4k_2 \right)^{\frac{1}{2}} \right]. \quad (16b)$$

In (15), we write

$$x_1^{n+1} = B_n x_1 + K \quad (17a)$$

and

$$x_2^{n+1} = B_n x_2 + K \quad (17b)$$

Subtracting (17b) from (17a) and combining terms, we have

$$B_n = \frac{(x_1^{n+1} - x_2^{n+1})}{(x_1 - x_2)}. \quad (18)$$

Substitution into (11) completes the proof for the cases  $x_1 \neq x_2$ .

When  $x_1 = x_2$ , then  $k_2 = \frac{-k_1^2}{4}$ . For this case we find  $B_n$  by inspection to be

$$B_n = (n+1) \left( \frac{k_1}{2} \right)^n. \quad (19)$$

Substitution of (19) into the left-hand side of (11) yields  $\left( \frac{k_1}{2} \right)^{2n}$  and completes the proof for that case.



Now, using (16a), (16b), and (18) or (19) the attenuation for any ladder-network circuit like that in Figure 1 can be explicitly determined.

The result in (11) for the generalized Fibonacci formula of (10) can be extended to higher-order recursive equations.

We now consider the generalized Tribonacci formula

$$T_n = k_1 T_{n-1} + k_2 T_{n-2} + k_3 T_{n-3} \quad (20)$$

where  $k_1$ ,  $k_2$ , and  $k_3$  are arbitrary constants,

$$T_0 = 1, T_1 = k_1, T_2 = k_1^2 + k_2, \quad \text{and} \quad T_3 = k_1^3 + 2 k_1 k_2 + k_3. \\ \dots \quad (\text{cf [ 5]}).$$

Using the same methods that we used to find (17a) and (17b) it is not difficult to show that

$$x_1^{n+2} = T_n x_1^2 + a x_1 + b \quad (21a)$$

$$x_2^{n+2} = T_n x_2^2 + a x_2 + b \quad (21b)$$

$$x_3^{n+2} = T_n x_3^2 + a x_3 + b \quad (21c)$$

where  $x_1$ ,  $x_2$ , and  $x_3$  are the three distinct roots of the auxiliary equation belonging to (20) and  $a$  and  $b$  are constants.

Now, subtracting (21b) from (21a) and (21c) from (21a), we are left with two equations. Solving the two equations we get:

$$T_n = \frac{E}{F} - \frac{G}{H} \quad (22)$$

where

$$E = \left( x_1^{n+2} - x_3^{n+2} \right), \quad (23a)$$

$$F = (x_1 - x_3)(x_3 - x_2) , \quad (23b)$$

$$G = (x_1^{n+2} - x_2^{n+2}) , \quad (23c)$$

$$H = (x_1 - x_2)(x_3 - x_2) . \quad (23d)$$

The following solutions for the three roots of the auxiliary polynomial

$$x^3 = k_1 x^2 + k_2 x + k_3 \quad (24)$$

of the third order homogeneous difference equation in (19) are found to be:

$$x_1 = A + B , \quad (25a)$$

$$x_2 = -\left(\frac{A+B}{2}\right) + \left(\frac{A-B}{2}\right)\sqrt{3}i , \quad (25b)$$

$$x_3 = -\left(\frac{A+B}{2}\right) - \left(\frac{A-B}{2}\right)\sqrt{3}i , \quad (25c)$$

with

$$A = \sqrt[3]{\frac{-C}{2} + \sqrt{\frac{C^2}{4} + \frac{D^3}{27}}} , \quad (26a)$$

$$B = -\sqrt[3]{\frac{C}{2} + \sqrt{\frac{C^2}{4} + \frac{D^3}{27}}} \quad (26b)$$

$$C = \frac{1}{27}(2k_1^3 - 9k_1k_2 + 27k_3) \quad (26c)$$

$$D = \frac{1}{3}(3k_2 - k_1^2) \quad (26d)$$

where

$$i = \sqrt{-1}, \quad \text{and} \quad n = 0, 1, 2, \dots$$

Therefore, an explicit formula can be obtained for the generalized Tribonacci equation using (26), (25), (23), and (22). Note when  $k_i = 1$  for  $i = 1, 2, 3$ , the value of (23) is the  $n$ th Tribonacci number.

Higher-order generalized formulas can also be solved explicitly when the roots of the equations (like (13) and (24)) are solved.

In order to demonstrate the geometry of this model in (10) with the simple values of  $k_1 = k_2 = 1$ , the values of  $B_n$  can be used to tile the plane by the arrangement in Figure 2. [3]. *Note that each tile is a square.* This method of arrangement is simple to follow. In each case, as each new tile is added to the region, a full rectangle results.

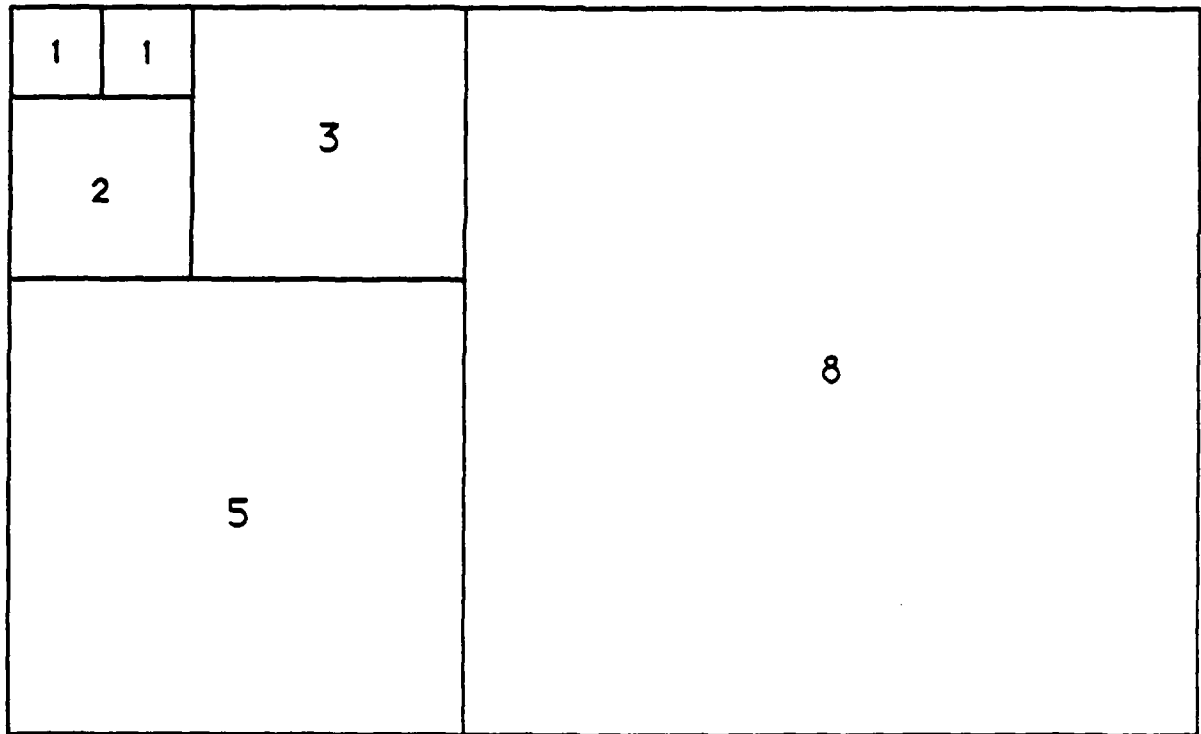


Figure 2 : Simple tiling of the plane using the *Fibonacci sequence*.

### 3. Paths of Light Rays

#### 3.1 Number of Distinct Paths with $n$ Reflections

The model for the number of distinct paths for light rays incident upon two stacked glass plates which have  $n$  reflections is the familiar Fibonacci sequence [7]. Several such possibilities of reflections are shown in Figure 3.

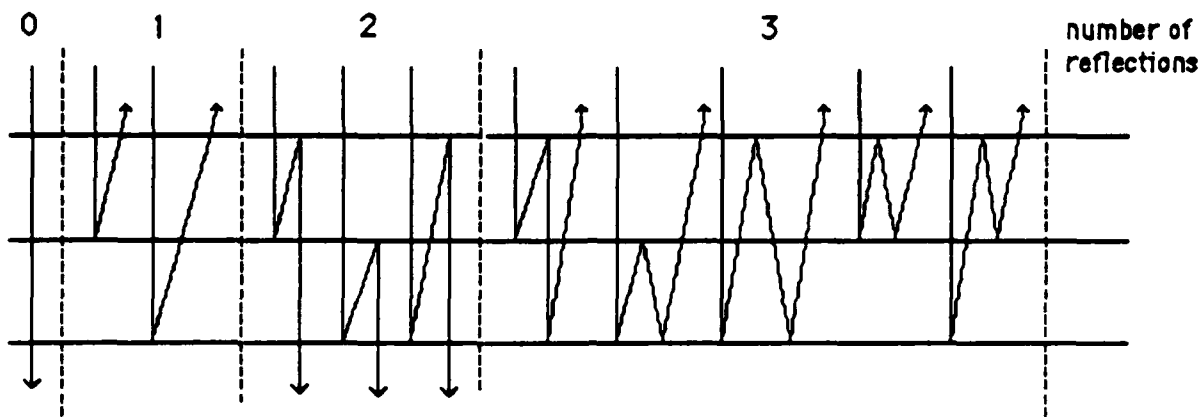


Figure 3: Portrayal of the possible scenarios with the given number of reflections

This model can be easily explained. Let  $P_n$  be the number of paths having  $n$  reflections. Clearly  $P_0 = 1$  and  $P_1 = 2$ .

Now, assume bundle A has  $P_{n-1}$  rays each with exactly  $n-1$  reflections while bundle B has  $P_n$  rays with exactly  $n$  reflections as shown in Figure 3 for  $n = 0, 1, 2, 3$ . There is no loss of generality regarding the parity of  $n$  as the following argument is valid if the diagram in Figure 4 is turned over top to bottom.

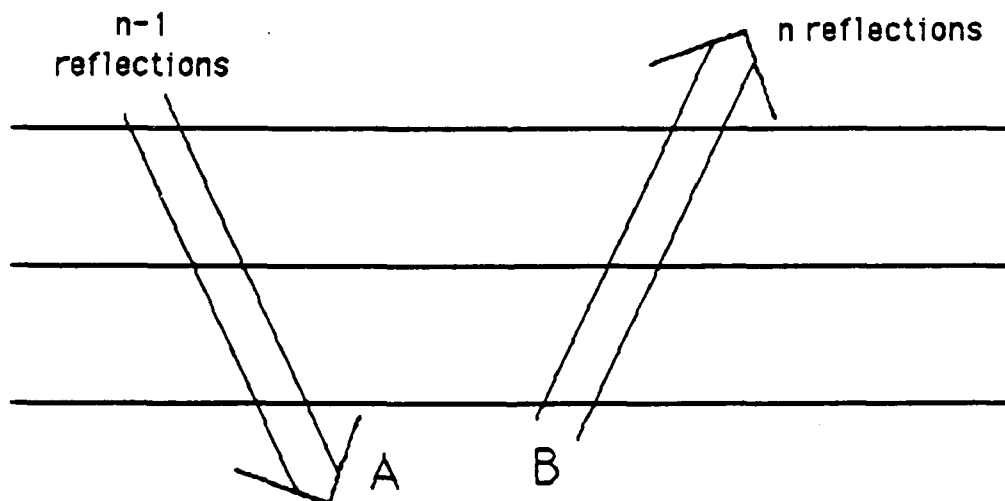


Figure 4: Portrayal of two bundles of rays with  $n-1$  and  $n$  reflections

$P_{n+1}$  is the number of distinct paths yielding  $n+1$  reflections. These all must come out of the glass plates in the opposite direction to those in bundle B and in the same direction as those in bundle A. Thus the rays of bundle B are reflected at their exit surface to get bundle B' while those of bundle A are reflected at their exit surface and then again at the interface in order to exit in the same direction to get bundle A'.

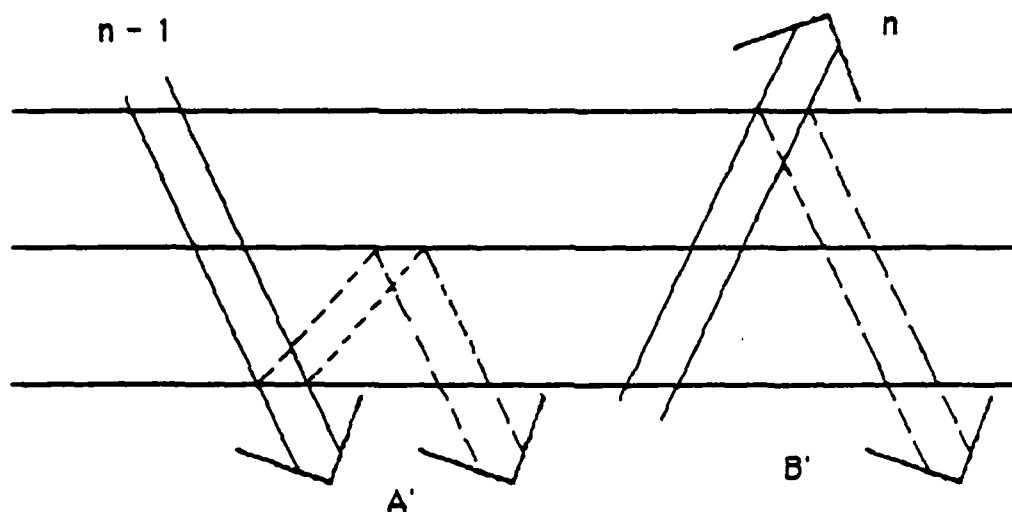


Figure 5: New reflections needed from bundles A and B to achieve  $n+1$  reflections

Since A contained all the distinct paths with  $n-1$  reflections and B contained all distinct paths with  $n$  reflections, the total number of paths with  $n+1$  reflections is determined by

$$P_{n+1} = P_n + P_{n-1}.$$

They are all distinct since each ray of bundle A' has a last reflection from the interface while each ray of bundle B' has a last reflection from an outside surface. Since the rays of A' and B' are each distinct within their bundles and are now distinct from paths in the other bundle, it follows that there is no duplication in bundle A' + B'.

### 3.2 The Crossing Numbers

Suppose as a further investigation we ask how many times,  $C_n$ , did the rays in the bundle with  $n$  reflections cross the interface between the two plates. Clearly, from Figure 3,  $C_0 = 1$ ,  $C_1 = 2$ ,  $C_2 = 5, \dots$ . From the

mechanics of propagation, shown in Figure 5, the crossings by bundle A is  $C_{n-1}$  and the crossings by bundle B is  $C_n$ . Bundle A' has the same number of crossings as bundle A, while bundle B' has all the crossings of B plus one extra crossing for each path in bundle B. Thus

$$C_{n+1} = (C_n + P_n) + C_{n-1} \quad (27)$$

where

$$C_0 = 1, \quad C_1 = 2, \quad C_2 = 5.$$

This sequence is related to the convolution of the Fibonacci sequence with itself or a higher order generating function. In performing this analysis we denote  $P_n$  using the familiar Fibonacci notation  $F_n$ , where  $P_n = F_{n+2}$ .

The recurrence relation for the Fibonacci numbers is written

$$F_{n+2} - F_{n+1} - F_n = 0$$

which is homogeneous, while the recurrence relation for the crossing numbers  $C_n$  is written as

$$C_n - C_{n-1} - C_{n-2} = F_{n+1}$$

which is nonhomogeneous.

The generating function for our model in equation (27) is

$$\frac{1}{(1 - x - x^2)^2}$$

which can also be used to tile the plane. Using the method from [4] to tile this sequence, the resulting tiling for the crossing numbers is shown in Figure 6.

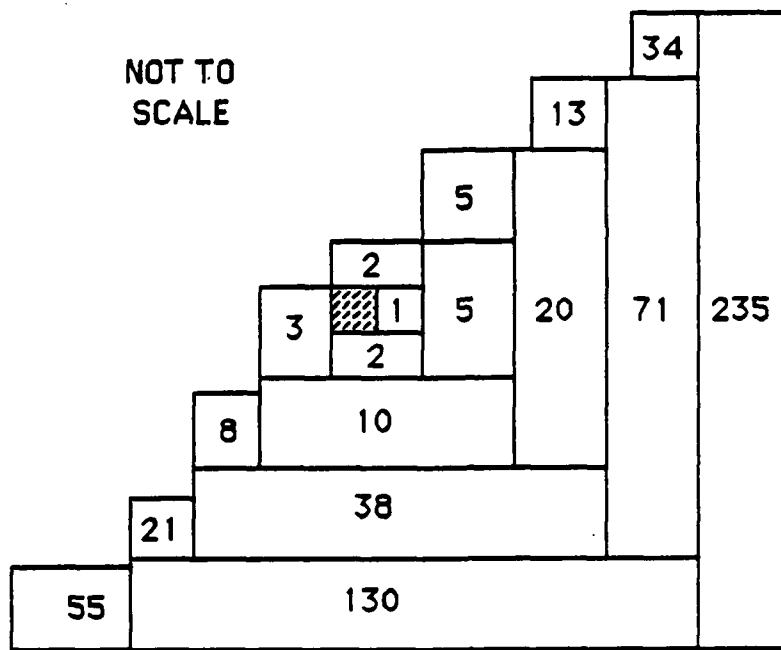
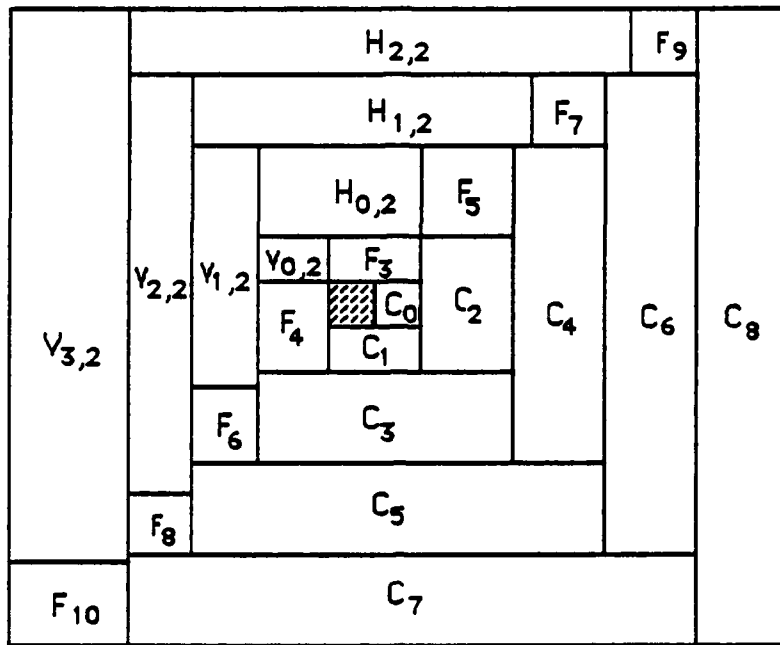


Figure 6: Tiling the plane using the values of  $C_n$  and  $F_n$ .

Under this construction, the tiling sequence starts with  $C_0=1$ ,  $C_1=2$ , and  $F_3=2$  and then continues in the manner shown in Figure 6.

Filler rectangles used to fill in the gaps left after placing the square tiles can also be determined. We redraw the tiling in Figure 6 using sequential notation rather than actual numbers to show these gap filling rectangles. We use the notation H and V to denote these filler rectangles that appear to be oriented horizontally and vertically, respectively.





NOT TO SCALE

Figure 7: Tiling of the plane showing filler rectangles.

One can see the following rectangle sizes from Figure 7 above:

$H_{0,2}$  is  $(C_3 - F_5)$  by  $F_5$

$H_{1,2}$  is  $(C_5 - F_7)$  by  $F_7$ .

$H_{2,2}$  is  $(C_7 - F_9)$  by  $F_9$ .

$V_{0,2}$  is  $F_4$  by  $(C_2 - F_4)$

$V_{1,2}$  is  $F_6$  by  $(C_4 - F_6)$ .

$V_{2,2}$  is  $F_8$  by  $(C_6 - F_8)$ .

$V_{3,2}$  is  $F_{10}$  by  $(C_8 - F_{10})$ .

In general,

$H_{1,n}$  is  $(C_{2n+3} - F_{2n+5})$  by  $F_{2n+5}$

and

$V_{1,n}$  is  $F_{2n+4}$  by  $(C_{2n+2} - F_{2n+4})$ .

### 3.3 Number of Reflections at the Interface

The model for the number of reflections at the interface,  $I_n$ , is also easily formed from a recursive sequence. Clearly, from Figure 3,  $I_0 = 0$ ,  $I_1 = 1$ ,  $I_2 = 2$ . In this case, this number of reflections is determined by

$$I_n = I_{n-1} + I_{n-2} + F_n, \quad \text{with} \quad I_0 = 0 \quad \text{and} \quad I_1 = 1.$$

This number pattern is generated from

$$\frac{x}{(1-x-x^2)^2}$$

and the tiling using  $I_n$  is shown in Figure 8.

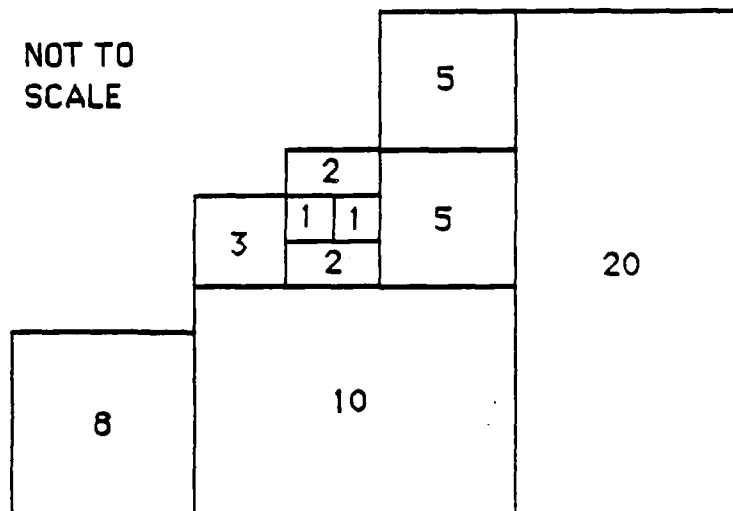


Figure 8: Tiling of the plane using values for  $I_n$

### 3.4 Number of Reflections at the Upper Surface

Many other reflection counts for this problem can be modeled with the recursive sequences and the plane tiled in similar ways. Our last example is the number of reflections at the upper surface,  $U_n$ . Here, the recurrence is mixed for odd and even numbers of  $n$  and is written as

$$U_{2n} = U_{2n-1} + U_{2n-2} + F_{2n+1}$$

or

$$U_{2n+1} = U_{2n} + U_{2n-1} + F_{2n+1} \quad \text{with } U_0 = 0, U_1 = 0.$$

The tiling of the plane using the  $U_n$  is shown in Figure 9.

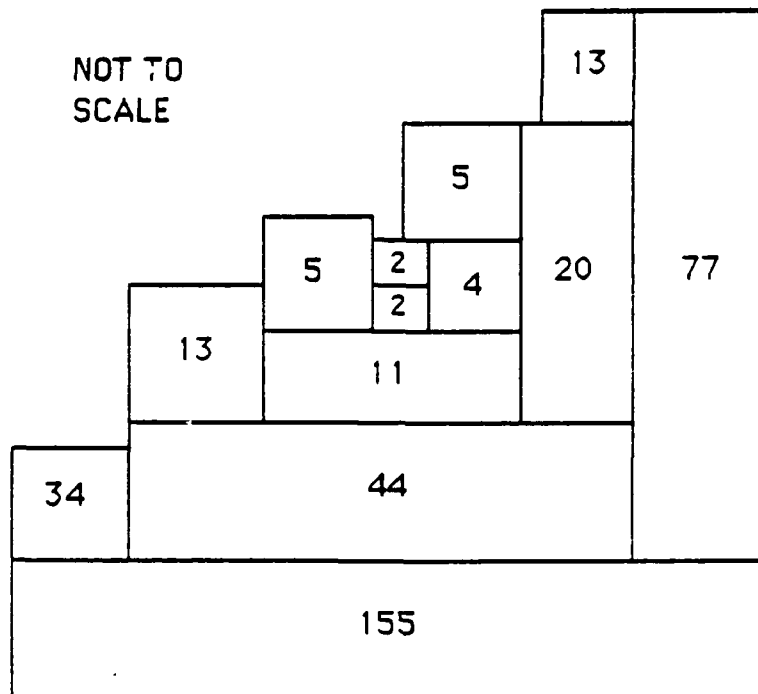


Figure 9: Tiling of the plane using values for  $U_n$ .

#### 4. Conclusion

While the Fibonacci-related sequences have been used to model numerous applications, their use in engineering and science has been limited. We have tried to show a few of the types of applications from science and engineering where their use is appropriate. We have also used the technique of tiling to provide a geometric flavor to these sequences and applications.

#### References

1. P. Anderson, "Fibonacci", Fibonacci Numbers and Their Applications, D. Reidel Publishing Co., 1986, pp. 1-8.
2. J. Arkin, "Ladder Network Analysis Using Polynomials", The Fibonacci Quarterly, Vol. 3, No. 2, April 1965, pp. 139-142.
3. J. Arkin, D. Arney, G. Bergum, S. Burr, and B. Porter, "Recurring-Sequence Tiling", to appear in The Fibonacci Quarterly.
4. J. Arkin, D. Arney, G. Bergum, S. Burr, and B. Porter, "Tiling the kth Power of a Power", submitted to The Fibonacci Quarterly.
5. J. Arkin, D. Arney, G. Bergum, S. Burr, and B. Porter, "Unique Fibonacci Formulas", to appear in The Fibonacci Quarterly.
6. J. Arkin and G. Bergum, "More On The Problem Of Diophantus", Applications Of Fibonacci Numbers, Kluwer Academic Publishers, 1986, pp. 177-181.
7. Arkin, J. and Hoggatt, Jr., V.E. "A Bouquet Of Convolutions". Washington State University Conference On Number Theory, March 1971, pp. 68-79.
8. G. Bergum and C. Long, "On A Problem Of Diophantus", Applications Of Fibonacci Numbers, Kluwer Academic Publishers, 1986, pp. 183-187.
9. M. Bicknell and V. E. Hoggatt, "A Primer For The Fibonacci Numbers: Part XIV", The Fibonacci Quarterly, Vol. 12, 1974, pp. 147-156.

10. M. J. Braverman and D. Toof, "An Application Of The Fibonacci Search Technique To Determine Optimal Sample Size In A Bayesian Decision Problem", A Collection Of Manuscripts Related To The Fibonacci Sequence, The Fibonacci Association, 1980, pp.137-145.
11. Garland, Trudi Hammel. Fascinating Fibonacci . Palo Alto: Dale Seymour Publications, 1987, pp.67-69.
11. K. Hirano, "Some Properties Of The Distributions Of Order k", Fibonacci Numbers and Their Applications, D. Reidel Publishing Co., 1986, pp. 43-53.
12. Hoggat, Jr., V. E. Fibonacci and Lucas Numbers . New York: Houghton Mifflin Co, 1969, pp.79,81.
13. J. McCabe and G. Phillips, "Fibonacci And Lucas Numbers And Aitken Acceleration", Fibonacci Numbers and Their Applications, D. Reidel Publishing Co., 1986, pp. 181-184.
14. B. R. Myers, "On Spanning Trees, Weighted Compositions, Fibonacci Numbers, And Resistor Networks", SIAM Review, Vol. 17, No. 3, July 1975, pp. 465-474.
15. A. Philipou, "Distributions And Fibonacci Polynomials Of Order k, Longest Runs, And Reliability Of Consecutive-k-Out-Of-n : F Systems", Fibonacci Numbers and Their Applications, D. Reidel Publishing Co., 1986, pp.203-227.

# APPROXIMATION AND INTERPOLATION FORMULAS FOR REAL-TIME APPLICATIONS

Charles K. Chui<sup>1</sup>

Department of Mathematics  
Texas A&M University  
College Station, TX 77843

and

Harvey Diamond<sup>2</sup>

Department of Mathematics  
West Virginia University  
Morgantown, WV 26506

**ABSTRACT.** A general scheme for constructing a compactly supported function that only requires finite (and relatively small) storage for the purpose of processing gridded discrete data in (near) real-time is presented. The attractive features are incoming data are used directly as filtering coefficients without matrix inversion and the optimal order of approximation is achieved while the data are being interpolated.

1. **INTRODUCTION.** Recently, there has been much interest in the problem of constructing univariate and multivariate approximation schemes by utilizing a single function  $\phi$ . The main ingredient in such construction processes consists of dilation (which we will also call scaling) and translation of the function  $\phi$ . Problems such as spline approximation and interpolation, realization of neural nodes in neural network structural analysis, synthesis via wavelets, and representation of surfaces by radial basis functions all fall into this category. In this paper, we are concerned with the problem of the construction of (near) real-time approximation and interpolation formulas by using  $\phi$ . More precisely, a compactly supported function  $\psi$  that can be evaluated at any time instant and space position efficiently will be constructed from scaling and translation of  $\phi$ , such that incoming discrete data samples can be used readily together with translates of  $\psi$  to give complete analog information with a minimal number of multiplications and additions, and that the representation guarantees optimal order of approximation provided by  $\phi$ . Since we seek a compactly supported  $\psi$ , we must start with a compactly supported function  $\phi$  which will be assumed to be piecewise continuous for the sake of convenience and feasibility in applications. Typically, in one variable,  $\phi$  is a  $B$ -spline, and in the multidimensional setting,  $\phi$  may be chosen as an appropriate linear combination of box splines in order to achieve the highest order of approximation and computational efficiency. By setting  $\phi$  to be the average of  $\phi(\mathbf{x})$  and  $\phi(-\mathbf{x})$ , if necessary, we may always assume that  $\phi$  is symmetric with respect to the origin.

We will call  $n = n(\phi) \in \mathbb{Z}_+$  the local order of approximation of  $\phi$  if it is the largest

---

<sup>1</sup> Supported by SDIO/IST managed by the U.S. Army Research Office under Contract No. DAAL03-87-K-0025

<sup>2</sup> Supported by DARPA under Contract No. MDA 972-88-C-0047

integer such that

$$\inf \left\| f - \sum_i a_i \phi \left( \frac{1}{h} (\cdot - ih) \right) \right\| = O(h^n), h > 0, \text{ for all } f \in C_0^n(\mathbb{R}^s).$$

Here,  $C_0^n = C_0^n(\mathbb{R}^s)$  denotes the class of all compactly supported  $n$  times continuously differentiable functions in  $s$  variables. We are given a set of discrete data  $\{f_i\}, i \in I, I \subset \mathbb{Z}^s$  (e.g.  $f_i = f(ih)$  or some partial derivatives, etc., of  $f$  at  $ih, f \in C_0^n(\mathbb{R}^s)$ ). The objective of this paper is to derive a real-time approximation-interpolation scheme to yield

$$s_h(f) = \sum_{i \in \mathbb{Z}^s} a_i \phi \left( \frac{1}{h} (\cdot - ih) \right)$$

that satisfies:

- (i)  $\|f - s_h(f)\| = O(h^n)$ , and
- (ii)  $(s_h(f) - f)(ih) = 0, i \in I$ ,

(or partial derivative versions, etc.) for all  $f \in C_0^n(\mathbb{R}^s)$ .

2. **RESULTS.** For convenience, we only state results on interpolation of function values. Our approach is to construct  $\tilde{\psi}$  which has *compact support* and can be easily "stored", such that

- (a)  $\tilde{\psi}(ih) = \delta_{i0}$ , and
- (b)  $\left\| f - \sum_{i \in I} f(ih) \tilde{\psi} \left( \frac{1}{h} (\cdot - ih) \right) \right\| = O(h^n)$

for all  $f \in C_0^n(\mathbb{R}^s)$ ; consequently, the above objectives (i) and (ii) are achieved. The index set  $I$  is assumed to contain 0 and to be homogeneous, in the sense that for any  $i \in I$ , we have  $I - i = I$ . In this regard we note that the following are equivalent conditions:

- (a)  $I - i = I$  for all  $i \in I$ .
- (b)  $I$  is a subgroup of  $\mathbb{Z}^s$
- (c)  $I$  is closed under addition and multiplication by  $-1$ .

We will assume as well that the quotient group  $\mathbb{Z}^s/I$ , the group of cosets generated by  $I$ , is finite; this will be the case if the elements of  $I$ , considered as elements of  $\mathbb{R}^s$ , span  $\mathbb{R}^s$ .

If  $I$  is not quite "full", then  $\tilde{\psi}$  can be constructed by using (scaled) translates of  $\phi$ ; but if  $I$  is quite "full" then a super space containing  $\phi$  has to be introduced.

For convenience, we will set  $h = 1$  and write  $s(f) = s_1(f)$ . Since the procedure is linear, the general result for arbitrary  $h > 0$  is attained by simply scaling. The first step is to construct an appropriate quasi-interpolation formula based on the given data. Details can be found in CD [2.3].

(1) **Quasi-interpolation.** We must construct an approximation of the form:

$$(Qf)(x) = \sum_{j \in I} \lambda_j f(\cdot + j) \phi(x - j)$$

where each  $\lambda_j$  is a linear functional, such that  $\lambda_j f(\cdot + j)$  involves only values of  $f(k)$ ,  $k \in I$ . Thus,  $\lambda_j$  must be expressed in terms of the values of  $f$  on the coset  $I - j$ . We must also choose  $\lambda_j$  such that  $Qp \equiv p$  for all  $p \in \pi_{n-1}^s$  (the space of all polynomials in  $s$  variables and with total degree at most  $n - 1$ ). Consequently, by scaling  $Qf$ , we have achieved (i).

We then define a function  $\psi$  by rewriting  $Qf$  as:

$$(Qf)(x) = \sum_{k \in I} f(k) \psi(x - k).$$

The basic technique to achieve our goal was introduced in our earlier work CD [3]:

Choose any  $\lambda$  such that

$$\sum_{j \in \mathbb{Z}^s} \lambda f(\cdot + j) \phi(x - j)$$

preserves  $\pi_{n-1}^s$ . Our favorite is  $\lambda$  is the one obtained by what we called the Neumann series approach in CD [1]. With this  $\lambda$ , we may now compute  $\lambda_j(p)$ . Then we may solve for  $\lambda_j$  by using

$$\lambda_j(p) = \lambda(p), \quad p \in \pi_{n-1}^s.$$

We show now that in general, we need to construct a  $\lambda_j$  for each element of the quotient space  $\mathbb{Z}^s/I$ , by picking a  $j$  from each coset. Consider the construction of a quasi-interpolant in the form

$$(Qf)(x) = \sum_{k \in I} f(k) \psi(x - k),$$

where  $\psi$  is a linear combination of translates of  $\phi$ , i.e.  $\psi(x) = \sum_{k \in \mathbb{Z}^s} c_k \phi(x - k)$ . Substituting into the expression for  $(Qf)(x)$ , we obtain

$$\begin{aligned} \sum_{k \in I} f(k) \psi(x - k) &= \sum_{k \in I} f(k) \sum_{j \in \mathbb{Z}^s} c_j \phi(x - k - j) = \sum_{j \in \mathbb{Z}^s} \sum_{k \in I} c_{j-k} f(k) \phi(x - j) \\ &= \sum_{j \in \mathbb{Z}^s} \sum_{k \in I} c_{j-k} f((k - j) + j) \phi(x - j). \end{aligned}$$

If we identify the coefficient of  $\phi(x - j)$  as  $\lambda_j f(\cdot + j)$  then  $\lambda_j$  is given by

$$\lambda_j f = \sum_{k \in I} c_{j-k} f(k - j) = \sum_{k \in I - j} c_{-k} f(k)$$



Now the index sets  $I - j$ , for  $j \in \mathbb{Z}^s$ , are precisely the cosets of  $I$  and hence for two different  $j$  are either identical or disjoint. We can therefore independently choose the  $c_k$  for  $k$  in each coset  $I - j$ , so as to satisfy the requirement that  $\lambda_j(p) = \lambda(p)$ ,  $p \in \pi_{n-1}^s$ . More precisely, we can now furnish the following algorithm for constructing  $\psi(x) = \sum_{k \in \mathbb{Z}^s} c_k \phi(x - k)$  such that  $(Qf)(x) = \sum_{k \in I} f(k) \psi(x - k)$  is a quasi-interpolant:

- 1) For each coset  $I - j$  calculate a  $\lambda_j$  in the form  $\lambda_j f = \sum_{k \in I - j} b_k^{(j)} f(k)$ .
- 2) Define  $c_k = \sum_{j \in \mathbb{Z}^s / I} b_{-k}^{(j)}$ , where the sum is over a set of coset representatives.
- 3) Define  $\psi(x) = \sum_{k \in \mathbb{Z}^s} c_k \phi(x - k)$ .

(2) **Choice of a basic cardinal interpolator.** Our second step is to construct  $\eta$ , by only using  $\phi$  if possible, or else by using a reasonable super space containing  $\phi$  such that

- (c)  $\eta(j) = \delta_{j0}$ ,  $j \in I$ , and
- (d) support ( $\eta$ ) is small.

It is clear from our assumption on the index set  $I$  that if  $\eta(j) = \delta_{j0}$ ,  $j \in I$ , then  $\eta(k - j) = \delta_{kj}$  for  $j, k \in I$ . One simple way of achieving (c) is for the support of  $\eta$  not to overlap with the other sample points  $j \in I$ . (We are not concerned with the approximation order at this stage.)

(3) **Construction of  $\tilde{\psi}$ .** From  $\psi$  and  $\eta$ , we may now construct our compactly supported cardinal interpolation function:

$$\begin{aligned} \tilde{\psi}(x) &= \sum_{i \in I} (\delta_{i0} - \psi(i)) \eta(x - i) + \psi(x) \\ &= \eta(x) + \psi(x) - \sum_{i \in I} \psi(i) \eta(x - i). \end{aligned}$$

Clearly,  $\tilde{\psi}(j) = \delta_{j0}$ . Note that since the index set  $I$  satisfies  $I - i = I$  for all  $i \in I$ , we may write

$$\begin{aligned} &\sum_k f(k) \tilde{\psi}(x - k) \\ &= \sum_k \left[ f(k) - \sum_j f(j) \psi(k - j) \right] \eta(x - k) \\ &+ \sum_k f(k) \psi(x - k), \end{aligned}$$

which is a noncommutative "blending operator", namely:

$$\mathcal{J}(f - Qf) + Qf = (\mathcal{J} \ominus Q)f$$

where  $\oplus$  denotes the Boolean sum of  $\mathcal{J}$  followed by  $Q$ . It is clear that  $\mathcal{J} \oplus Q \neq Q \oplus \mathcal{J}$  since  $\mathcal{J}f$  has lost too much information on  $f$  and  $Q$  is not an interpolation operator. Nonetheless, it is not difficult to show that  $\mathcal{J} \oplus Q$  not only preserves polynomials of highest degrees as  $Q$  does, it also provides the desirable interpolation property of  $\mathcal{J}$  (cf. CD [3]).

3. **EXAMPLES.** To illustrate our construction procedure, it is best to give several examples. In the following, we give three examples: the first reconstructs an old example due to Jenkin's, the second completes an example considered by DGM [4], and the third example utilizes a  $C^2$  quartic box spline to give a real-time interpolation formula that gives the fourth order of approximation, which is optimal.

(1)  **$C^2$ -quartic cardinal spline interpolation at  $\mathbf{Z}$ .** (Jenkins [5], cf. Schoenberg [6].)

Our procedure to construct Jenkins' basic cardinal interpolation function  $\tilde{\psi}$  is very simple:

Let  $\phi$  be the centered cubic  $B$ -spline with knots at  $\mathbf{Z}$ . Since  $I = \mathbf{Z}$ , we may choose

$$\lambda_j = \lambda, \quad j \in \mathbf{Z},$$

where  $\lambda f(\cdot) = -\frac{1}{6}f(-1) + \frac{4}{3}f(0) - \frac{1}{6}f(1)$ . This gives

$$\psi(x) = -\frac{1}{6}\phi(x+1) + \frac{4}{3}\phi(x) - \frac{1}{6}\phi(x-1).$$

The function  $\tilde{\psi}$  can now be constructed if we can find a  $C^2$ -quartic  $\eta$  such that

$$\begin{cases} \eta(0) = 1 & \text{and} \\ \text{support}(\eta) = [-1, 1]. \end{cases}$$

To do so, we simply set  $\eta(x) = \begin{cases} (1+3|x|)(1-|x|)^3, & |x| \leq 1 \\ 0, & \text{otherwise.} \end{cases}$

(2)  **$C^2$ -cubic cardinal spline interpolation at  $I = 2\mathbf{Z}$ .** (Partially worked out using a different method by Dahman, Goodman, Micchelli [4].)

Again, let  $\phi$  be the centered cubic  $B$ -spline with knots at  $\mathbf{Z}$ . Since the sample points are chosen to be  $2\mathbf{Z}$ , we may now choose  $\eta(x) = \frac{3}{2}\phi(x)$ , so that

$$\begin{cases} \eta(j) = \delta_{j0}, & j \in 2\mathbf{Z}, \quad \text{and} \\ \text{support}(\eta) = [-2, 2]. \end{cases}$$

Use any  $\lambda$  that induces a quasi-interpolant. Then

$$\lambda[1 \ x \ x^2 \ x^3] = \left[1 \ 0 \ -\frac{1}{3} \ 0\right].$$

Since only  $f(2k)$ ,  $k \in \mathbf{Z}$ , are used, we must construct at least two different  $\lambda_j$ . We consider even and odd integers; so

$$\lambda_e(f) = -\frac{1}{24}f(-2) + \frac{26}{24}f(0) - \frac{1}{24}f(2),$$

$$\lambda_o(f) = -\frac{1}{12}f(-3) + \frac{7}{12}f(-1) + \frac{7}{12}f(1) - \frac{1}{12}f(3).$$

With  $\lambda_j = \lambda_e$  for even  $j$ , and  $\lambda_j = \lambda_o$  for odd  $j$ , we have

$$\psi(x) = -\frac{1}{12}\phi(x+3) - \frac{1}{24}\phi(x+2) + \frac{7}{12}\phi(x+1)$$

$$+ \frac{26}{24}\phi(x) + \frac{7}{12}\phi(x-1) - \frac{1}{24}\phi(x-2) - \frac{1}{12}\phi(x-3).$$

Hence, we have

$$\tilde{\psi}(x) = \eta(x) + \psi(x) - \sum_{i \in 2Z} \psi(i)\eta(x-i)$$

$$= \frac{3}{2} \left\{ \frac{1}{72}\phi(x+4) - \frac{2}{36}\phi(x+2) + \frac{3}{36}\phi(x) - \frac{2}{36}\phi(x-2) \right.$$

$$\left. + \frac{1}{72}\phi(x-4) \right\} + \psi(x).$$

**(3) Cardinal interpolation at  $2Z^2$  by bivariate  $C^2$ -quartic spline on the 3-direction mesh.**

In the example, we let  $\phi$  be the box spline  $M_{222}$ . The order of approximation is  $n = 4$ . Since  $\phi(2k) = 0$  for all  $0 \neq k \in Z^2$  and support  $(\phi)$  is contained in  $[-2, 2]^2$ , we may use

$$\eta(\mathbf{x}) = 2\phi(\mathbf{x})$$

Using the "Neumann series" to produce  $\lambda$  cf. CD [1], we have

$$\lambda[1 \ x \ y \ x^2 \ xy \ y^2 \ x^3 \ x^2y \ xy^2 \ y^3]$$

$$= \left[ 1 \ 0 \ 0 \ -\frac{1}{3} \ -\frac{1}{6} \ -\frac{1}{3} \ 0 \ 0 \ 0 \ 0 \right].$$

Since only data at  $I = 2Z^2$  are to be used, we must have four different  $\lambda_j$ 's. We use  $\lambda^1, \lambda^2, \lambda^3, \lambda^4$  as described in Figure 1 with supports in  $I_1 = I = \{(\text{even}, \text{even})\}$ ,  $I_2 = I - (1, 0) = \{(\text{odd}, \text{even})\}$ ,  $I_3 = I - (0, 1) = \{(\text{even}, \text{odd})\}$ , and  $I_4 = I - (1, 1) = \{(\text{odd}, \text{odd})\}$ , respectively; so that

$$\sum_{\mathbf{j} \in Z^2} \lambda_{\mathbf{j}} f(\cdot + \mathbf{j}) \phi(\mathbf{x} - \mathbf{j})$$

$$= \sum_{k=1}^4 \sum_{\mathbf{j} \in I_k} \lambda^k f(\cdot + \mathbf{j}) \phi(\mathbf{x} - \mathbf{j}),$$

where for each  $k = 1, \dots, 4$ , and  $\mathbf{j} \in I_k$ ,  $\lambda^k f(\cdot + \mathbf{j})$  only involves evaluation at the even integers. The functions  $\psi$  and  $\tilde{\psi}$  are given in Figures 2 and 3, respectively, where the coefficients  $c_j$  of  $\phi(\mathbf{x} - \mathbf{j}) = M_{222}(\mathbf{x} - \mathbf{j})$  are shown.

### References

1. Chui, C.K. and Diamond, H., A natural formulation of quasi-interpolation by multivariate splines, Proc. Amer. Math. Soc. 99 (1987), 643-646.
2. Chui, C.K. and Diamond, H., Characterization of quasi-interpolation and application, Num. Math. To appear.
3. Chui, C.K. and Diamond, H., A general framework for local interpolation, CAT Report #190, 1989.
4. Dahmen, W., Goodman, T.N.T., and Micchelli, C.A., Compactly supported fundamental functions for spline interpolation, Num. Math. 52 (1988), 639-664.
5. Jenkins, W.A., Oscalatory interpolation: New derivation and formulae, Record Amer. Inst. Actuaries 15 (1926), 87.
6. Schoenberg, I.J., Contributions to the problem of approximation of equidistant data by analytic functions. Part A, Quart. Appl. Math. 4 (1946), 45-99.

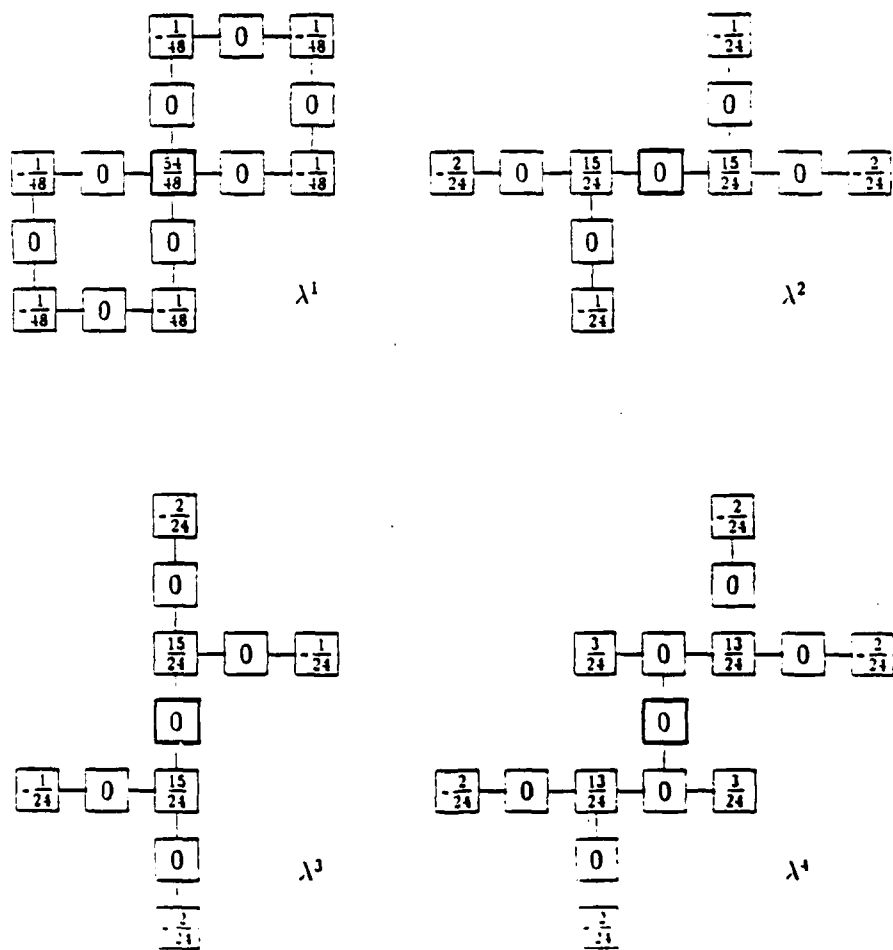


Figure 1

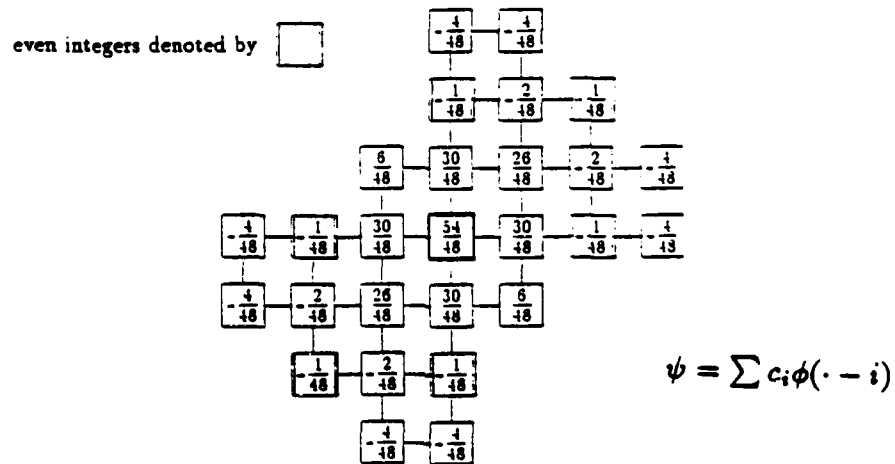


Figure 2 (Coefficients  $c_i$ )

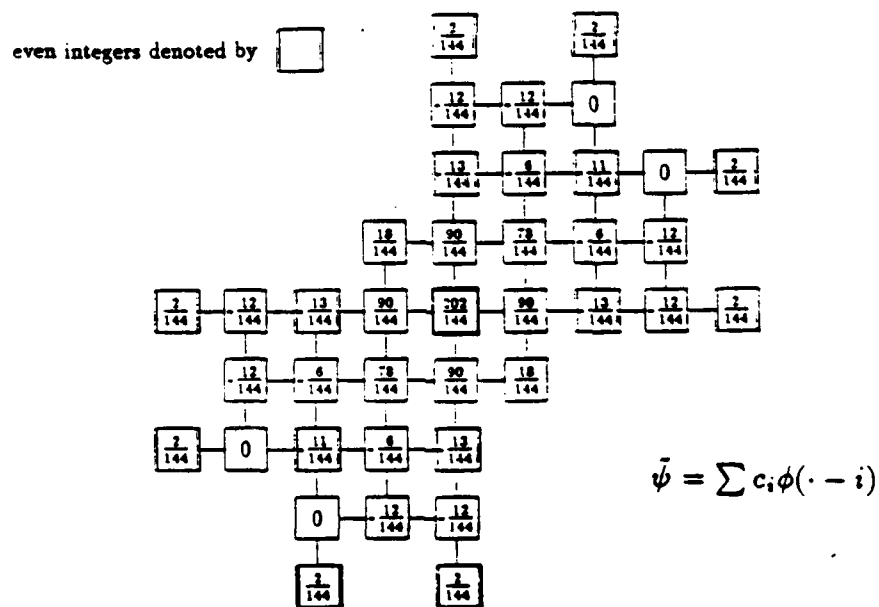


Figure 3 (Coefficients  $c_i$ )

AN ENHANCED KNOT SELECTION ALGORITHM FOR  
LEAST SQUARES APPROXIMATION USING THIN PLATE SPLINES

Major John R. McMahon  
Department of Mathematics  
United States Military Academy  
West Point, New York 10996

Richard Franke  
Department of Mathematics  
Naval Postgraduate School  
Monterey, California 93940

INTRODUCTION

The problem of fitting a surface to small sets of given data has been addressed in many different ways and several computer programs are currently available which enable one to deal with the problem effectively. Many of the methods available involve a global interpolation or approximation scheme and often involves solving a system of equations with an equivalent number of unknowns. For very large sets of data, the problem is computationally intractable. This consideration provides the motivation behind the development of a way to pare the problem down to a more manageable size.

We wish to construct a function  $F$  which approximately fits the data since we assume the data collection is subject to measurement error. We propose to use approximation by least squares Thin Plate Splines (TPS), where the surface function is constructed so as to minimize an error function subject to certain constraints. Solving the approximation problem will also involve as many equations as there are data points, but the number of unknowns will be significantly fewer. Part of the appeal of TPS approximation lies in the fact that it minimizes a certain linear functional, and in-

volves a linear combination of functions with no greater complexity than the natural logarithm of the distance function.

Approximation by least squares TPS is straightforward, once the coordinates  $(x_i, y_i)$  and  $(x_j, y_j)$  are known. We employ the TPS function

$$F(x, y) = \sum_{j=1}^K A_j d_j^2 \log(d_j) + ax + by + c$$

where  $d_j^2 = (x_i - x_j)^2 + (y_i - y_j)^2$ , and the coefficients  $A_j$ ,  $a$ ,  $b$  and  $c$  are chosen to minimize the error function

$$E = \sum_{i=1}^N \{ [F(x_i, y_i) - f_i] / s_i \}^2 .$$

The ordinates,  $f_i$ , may be subject to random errors, say with standard deviation,  $s_i$ , at the  $i^{\text{th}}$  data point. We model the plate under the point loads at the knot points (as opposed to the data points); therefore the constraint equations for the least squares TPS method, which may be thought of as 'equilibrium conditions' on the plate should be satisfied. Thus, the error function is minimized subject to the constraint equations:

$$\sum_{j=1}^K A_j = 0, \quad \sum_{j=1}^K A_j x_j = 0, \quad \sum_{j=1}^K A_j y_j = 0 .$$

We use LINPACK [1] subroutines to do the actual calculations.

Interpolation of scattered data by the method of TPS was developed from engineering considerations by Harder and Desmarais [2]. It can be thought of as a two dimensional generalization of the cubic spline, which models a thin beam under point loads subject to equilibrium constraints. The TPS function is derived from a differential equation which gives the deformation of an infinite, thin plate under the influence of point loads. A point load is applied at each data point so that the interpolating

surface can be constructed as a sum of fundamental solutions of the TPS equation.

In using the least squares TPS approximation method to fit the surface, a fewer number of basis functions than the number of given data points is employed. These basis functions are centered at a different, smaller set of points, which in analogy with the univariate case, we call the knots. Therefore, the problem at hand is one of selecting the knot points, and hence the basis functions.

Given a 'large' set of data points,  $(x_i, y_i, f_i)$ ,  $i = 1, \dots, N$ , we wish to find a smaller set of knot points,  $(x_j, y_j)$ ,  $j = 1, \dots, K$ , which will 'represent' the former reasonably well. This could be accomplished by choosing a subset of the original set, or by some process which produces a representative set. The ultimate goal is to approximate the surface from which the original data arose using the representative set. Hence, a surface fit to the large set and one fit to the representative set should be essentially the same.

#### ORIGINAL ALGORITHM

The original algorithm for solving the knot selection problem was developed based on the optimization of one function subject to a constraint formulated in terms of another function. Specifically, we sought to achieve an equal or as near-to-equal distribution of the data points amongst the knots. To do this, we move the knots around the domain of the fixed data points in search of an optimal configuration of knots which minimized the quantity

$$DSUM = \sum_{j=1}^K (N/K - N_j)^2$$



where  $N$  is the number of data points,  $K$  is the number of knot points and  $N_j$  is the actual number of data points belonging to the  $j^{\text{th}}$  knot. This objective function is the subject of later discussion.

A key advantage of this optimization is the natural heuristic which precipitates from it for moving the knots around the domain in search of a better configuration. This natural heuristic follows from the fact that for most configurations of knots, some knots will own more data points than others so that a simple mechanism for moving the knots around is realized by moving the knots owning the fewest points toward the knots owning the most.

We also sought to position each knot in such a way that the distances between the data points and their closest knot point was minimized. This was accomplished by minimizing the constraint function

$$GN^2 = \sum_{i=1}^N \text{MIN}_j [(x_i - x_j)^2 + (y_i - y_j)^2]$$

Thus our original algorithm would propose a certain configuration of knots, determine which data points belonged to which knot, move the knots to minimize the distances, and check the distribution of data points as a result of this movement. Then, based on how bad the distribution turned out, certain knots would be moved in accordance with the searching scheme, and the process would begin all over again.

However, the particular scheme we developed to search for the optimal configuration of knots left a lot to be desired in terms of the excessive computation time required. Hence, one objective of the research effort was to reduce the time spent searching for an optimal configuration of knots using a better searching scheme and any other means available. These two topics are also developed within this paper.

The constraint function  $GN^2$  leads naturally to a default Dirichlet

Tesselation, a partitioning of the plane with respect to the knot points (See figure 1). Thus, we say each data point belongs to some knot point according to the Dirichlet tile in which it lies. Differentiation of  $GN^2$  with respect to the  $x_j$  and  $y_j$  show that at the minimum, each knot point will occupy the centroid with respect to the data points inside that tile. The following theorem applies to this constraining algorithm.

Theorem: The function  $GN^2$  decreases with each iteration which involves movement of a knot point. See [3] for a proof.

#### OBJECTIVE FUNCTION

Previous work on this problem lacked sufficient consideration of the objective function upon which the optimization in the knot selection algorithm is based. Recall the function DSUM above defined as

$$DSUM = \sum_{j=1}^K (N/K - N_j)^2$$

where  $N_j$  is now the actual number of data points found in the  $j^{\text{th}}$  Dirichlet tile. It is a measure of the evenness of the distribution of the data points amongst the knot points; a smaller value is indicative of a better distribution. The minimization is justified in terms of the desire to have the Dirichlet tile for each knot contain the same or nearly the same number of data points. It is a continuous function in the sense that there is an infinite number of possible knot configurations, each corresponding to a value of the function. An analysis of the function is motivated by several factors, summarized in the following questions. What is the minimum value the function can assume? Under what circumstances can the minimum value be obtained and how feasible is obtaining the minimum value?

First, we consider the minimum value of the objective function.

Let  $N$ , the number of data points, be written as  $N = K m + n$ , where  $K$  is the number of knot points to be used and  $m$  and  $n$  are integers. Two cases must be investigated: I) For  $n = 0$ , and II) for  $n \neq 0$ . We shall refer to these even and nearly even distributions of data points as the Ideal distribution for the respective case of  $n$ .

Case I occurs when all  $K$  knots own the same number of data points; hence  $N_j = N/K$ . Thus, for  $n = 0$ , we have  $m = N/K$ , corresponding to an exactly even distribution of the data points, so that  $DSUM = 0$ . Case II occurs when  $K - n$  knots own  $m$  data points and  $n$  knots own  $m + 1$  data points. It is easy to verify that we are working with  $K - n + n = K$  knot points. Thus for  $n \neq 0$ ,  $N = K m + n$  or  $N/K = m + n/K$  so that  $DSUM = (K - n)(N/K - N_j)^2 + n (N/K - N_j)^2$

But the first  $K - n$  knots own  $N_j = m$  data points and the other  $n$  knots own  $N_j = m + 1$  data points so that with the substitution above, we have  $DSUM = (K - n)(m + n/K - m)^2 + n (m + n/K - m - 1)^2$   
Simplifying this expression yields  $DSUM = n - n^2/K$ .

Thus for the case where  $N = 5000$ ,  $K = 250$ , the minimum value of  $DSUM$  is 0; for the case where  $N = 1776$ ,  $K = 100$ , the minimum value of  $DSUM$  is  $76 - (5776/100) = 18.24$ .

In order to obtain some indication as to the feasibility of achieving the minimum value of  $DSUM$ , we look at the value of  $DSUM$  after an ever-so-slight perturbation as follows. Consider case I ( $n = 0$ ); the slightest perturbation from the ideal distribution occurs when there is one knot with  $m + 1$  data points,  $K - 2$  knots with  $m$  data points and one knot with  $m - 1$  data points. A quick check of the total number of knot points reveals  $K = 1 + K - 2 + 1$  and the total number data points  $N = m + 1 + m (K - 2) + m - 1 = m K$ .

Thus,

$$DSUM = 1 [N/K - (m + 1)]^2 + (K - 2)(N/K - m)^2 + 1 [N/K - (m - 1)]^2$$

Substituting for  $m = N/K$ , since  $n = 0$ , we have

$$DSUM = (m - m - 1)^2 + 0 + (m - m + 1)^2 = 2.$$

Thus the slightest perturbation from the ideal distribution of data points yields a DSUM value slightly larger than the optimal value.

Other slight perturbations for the example of  $N = 100$ ,  $K = 10$ , ( $n = 0$ ), such as 2 knots with 9 points, 6 knots with 10 points, and 2 knots with 11 points, or 2 knots with 9 points, 7 knots with 10 points, and 1 knot with 12 points, or 1 knot with 8 points, 8 knots with 10 points, and 1 knot with 12 points yield DSUM values of 4, 6, and 8, respectively.

Case II where  $n \neq 0$  is a bit more interesting since the slightest perturbation can take on several forms, each leading to the same DSUM value. We previously described the ideal distribution of this case as occurring with  $K - n$  knots owning  $m$  data points and  $n$  knots owning  $m + 1$  data points. A quick check of the number of data points reveals there are  $N = (K - n)m + n(m + 1) = Km + n$  data points. One of the slightest perturbations occurs when there is one knot with  $m + 2$  data points,  $n - 1$  knots with  $m + 1$  data points,  $K - n - 1$  knots with  $m$  data points, and one knot with  $m - 1$  data points. Thus,

$$DSUM = 1 (N/K - m - 2)^2 + (n - 1)(N/K - m - 1)^2 + (K - n - 1)(N/K - m)^2 + 1 (N/K - m + 1)^2$$

Substituting for  $N/K = m + n/K$ , we have

$$DSUM = (n/K - 2)^2 + (n - 1)(n/K - 1)^2 + (K - n - 1)(n/K)^2 + (n/K + 1)^2 = 4 + n - n^2/K \quad \text{upon simplification.}$$

The same result is obtained for other slight perturbations such as one knot with  $m - 1$  data points,  $K - n - 2$  knots with  $m$  data points, and  $n + 1$  knots with  $m + 1$  data points, OR one knot with  $m + 2$  data

points,  $n - 2$  knots with  $m + 1$  data points, and  $K - n + 1$  knots with  $m$  data points. We will take advantage of this knowledge about the objective function, DSUM, later as part of integrating some other techniques into the algorithm for speeding things up.

#### ALTERNATIVE KNOT MOVEMENT SCHEMES

Recall the natural heuristic mentioned earlier for moving the knots around the domain of the fixed data points in search of the optimal knot configuration. An essential task in exploiting it lies in identification of the knots owning the most and fewest number of data points. The knot movement schemes developed to search for an optimal knot configuration were based primarily on the idea of spreading the wealth of the knots owning the most data points by moving the knots owning the fewest data points toward the former. In developing these various schemes, we considered both ease of implementation and computational costs to be paramount.

As before in the Original scheme, the rationale for moving the knots around the plane is to tweak the current configuration to a sufficient degree so as to cause the Dirichlet tile boundaries to move in such a way that some of the data points will belong to a different knot point(s). This is followed by the usual settlement of the knots into the centroid locations of their respective tiles, such that the settlement will lead to a better configuration of knots in terms of the evenness of the distribution of the data points amongst them.

The original algorithm employs a symmetric scheme to conduct a confined but exhaustive (and correspondingly expensive) search for the optimal configuration of knots. As seen in the figure 2, the so-called low knot (that is, the one owning the fewest data points) is moved toward the high knot (the one owning the most data points). The movement

is done along the line which connects them and alternates between the low-towards-high and high-towards-low modes in a symmetrical fashion. The distance moved along this line connecting the two knots is a function of the iteration of the movement, up to a total of ten moves each, or until a better knot configuration is found. When no better configuration is found, the next pair of high-low knots is considered, if available. Once all possible combinations of high-low pairs have been considered, and no better configuration has been found, the search is ended.

The Outward Bound scheme (figure 3) is characterized by a move of a high knot away from the low knot along the line between the two extended beyond the high knot. Such a move is justified by the obvious vacuum created by such a move in the vicinity of the previous location of the high knot point. When such a move fails to lead to a better configuration, the distance moved along the same line is decreased as a function of the iteration number until the high knot settles back to its original location. Note that because this second and successive bounds are made closer to the concentration of data points, it is more likely that the new knot location will absorb some of the extra data points in the local vicinity. This is followed by a move of the low knot toward the high knot along the line connecting them in an effort to relieve some of the pressure near the high knot point. As before, when these moves fail to lead to a better configuration, the distance moved is decreased as a function of the iteration number until the low knot returns back to its original location.

An even better scheme evolved from the last one wherein the low knot is moved to coincide with the high knot (figure 4). This approach

takes advantage of the inherent subroutine contingency for handling the case wherein knots begin to coalesce. In order to preclude such coalescence, the one knot is immediately moved on top of the nearest DATA point, wherever it may be located. This method has the added advantage of moving a knot in a totally different direction from the line connecting the high-low pair of knots being considered. As seen in the figure 4, once splitting the monopoly of the high knot fails to lead to a better configuration, the low point is moved along the line connecting the two knots some decreasing distance between them. The high knot is also moved out along the same line extending beyond its current location in the direction opposite the low point location.

Finally, we considered the situation wherein the movement desired is ever-so-slight enough to nudge the Dirichlet tessellation into one of its neighboring configurations, one containing the optimal solution (figure 5). Thus, the distance moved or trial distance, became a function of the area of the domain of the data and the number of knots being used. This trial distance is increased as a function of the iteration up to a set amount until no better configuration was found. This tack was also used in conjunction with the monopoly-splitting approach mentioned above.

What we settled on after much testing with several different test data sets was a combination of several of these approaches as we shall see. It became apparent that more combinations of high-low pairs needed to be considered in any scheme employed. Thus, whenever fewer than five high-low pairs of knots were found to exist, more were generated by identifying the knot owning the second most data points and so on until at least five such combinations could be considered. We note that we could have also considered the knot(s) owning the second fewest number

of data points; however, such consideration is unwarranted since the best results are obtained by breaking up the monopoly of the knots owning the most (or second most, and so on, as the case may be) data points.

#### CELL METHOD

All of these approaches to the knot movement schemes involve the identification of the knots owning the most data points and those owning the fewest. This task follows from having determined which knot is closest to which data point. One obvious improvement needed to speed up any of the schemes is a way for determining the closest knot point for each data point without considering each pair of possibilities again and again. In other words, we needed to take advantage of the fact that not all points needed to be checked every time. The Cell Method was developed in a general sense for locating the closest knot point to a given data point by Renka [4]. Its employment involves the use of two subroutines, STORE2 and GETNP2. We will describe the general idea of the algorithm in terms of its application to the problem of knot selection.

The motivation behind the use of the cell method was simply to find a better, faster means of identifying the closest knot point for each of the data points. The original program took a brute force approach wherein the Euclidean distance between each of  $K$  knots and  $N$  data points pair was computed and compared to the others one at a time until the closest knot was found for each data point. Thus, a minimum of  $N \cdot K$  computations had to be made each time the subroutine was invoked to determine which data points belonged to which knot and move the knots to minimize the distances.

A simple example and sketch of the a situation offer the best



explanation of the method (figure 6). First consider the K knots which have variable locations while the N data points are fixed. The smallest rectangle containing the knots is found and partitioned into a 3 x 3 uniform grid of cells (by the STORE2 subroutine). Since not all the cells will contain knots, the indices of the knots contained in each cell are recorded. Now, for a given single data point, a call to GETNP2 is made to find the nearest data point. A search is begun in the cell containing the data point or in the cell to which it is closest. If a cell is empty with respect to the knots, it is not considered for obvious reasons. The distance between the data point and the first knot encountered in one of the cells is calculated and the search is confined to those cells within that distance of the data point under consideration. Thus, only the knots within those cells can be considered thereby reducing the scope of the search for the closest knot point.

This procedure must be followed for each data point in turn but the scope of the procedure is much reduced compared to the brute force approach; we estimate a savings of around 25% from the original computational effort required is achieved. The task of locating the closest knot point for each data point is performed by the MINORM subroutine which is called twice by the search subroutine, TWEAK. That is, each time a different configuration of knots is proposed within the TWEAK subroutine, MINORM is invoked twice, so that it is easy to appreciate the scope of the savings enjoyed by the use of the Renka Cell Method.

Additionally, the MINORM subroutine was enhanced in the sense that it could be applied in a more general setting wherein a prospective user could specify weights for each of the data points. One could think of

the weight as being the reciprocal of the standard deviation of the error associated with the data measurement at a given data point. Hence, instead of summing the number of data points in each Dirichlet tile, the weights associated with each of the data points within the tile are summed. With a relatively large weight at a given data point, one would be able to force the knot to be at, or very near, that data point. The rest of the knot selection algorithm works as before. However, before solving the least squares system, each equation must be scaled by the value of the corresponding weight.

#### SIMULATED ANNEALING

Besides incorporation of the cell method and improvement of the tweaking scheme for speeding up the knot selection process, it became apparent that another approach in the form of Simulated Annealing (SA) would prove useful. SA which is also known by other names such as Monte Carlo annealing, statistical cooling, probabilistic hill climbing, stochastic relaxation and probabilistic exchange algorithm, was independently developed and introduced by Kirkpatrick et al in 1982 and Cerny in 1985. The name comes from an analogy to the slow cooling of a solid until it reaches its low energy ground state as developed by Metropolis et al in 1953 [5]. Here, a Cost function assumes the role of energy, a control parameter is substituted for temperature, and knot configurations are analogous to states of the solid. The SA algorithm is a general approximation algorithm for solving a wide variety of combinatorial optimization problems such as the knot selection problem. It obtains near-optimal solutions based on some randomization techniques incorporated into an iterative improvement algorithm.

The application of an iterative improvement algorithm presupposes

the definition of configurations, a cost function and a mechanism for generating transitions from one configuration to another, all of which are present in the knot selection problem. The solutions obtained using SA have the additional advantage of being independent of the initial configuration of knots and usually lead to a solution near the *minimum*. The similarities between the original knot selection algorithm and the SA algorithm extend beyond the necessary overhead. As with the original knot selection algorithm, a configuration is given, followed by the generation of a sequence of configurations which are compared to the current configuration in terms of the evenness of the distribution of data points. When a neighboring configuration has a lower cost, the current configuration is replaced by the better one.

The randomization technique comes to bear in the event a better configuration is not found; here, small increases are permitted to occur in the cost function with a non-zero but decreasing probability. This Metropolis criterion, as it is called, is implemented by drawing random numbers from a uniform distribution on  $[0,1)$  and comparing them in turn to an exponentially decreasing probability of acceptance function defined as  $\exp(\Delta C_{ij}/c)$  where  $\Delta C_{ij}$  is the difference in the costs between the two competing configurations and  $c$  is the control parameter. Initially, the control parameter is given a high value so that as the algorithm is invoked, the values of  $c$  become smaller until virtually no deteriorations in the cost occurs and the algorithm terminates. Thus, the key ingredient of the SA algorithm lies in its occasional acceptance of a worse configuration early on in the search effort.

For our particular application of the SA algorithm, we were again concerned about ease of implementation and additional computational costs. As we shall see, neither of these concerns were warranted. As

a point of embarkation, say that after 20 iterations of the SA algorithm we wish to have the probability of accepting a worse configuration down to around .01. We can determine the value of the control parameter  $c$  using the Metropolis criterion; that is, for what value of  $c$  will the probability of acceptance approach .01 given an average difference in the cost function analyzed earlier for case I to be 5, after some 20 iterations. Thus, we solve for  $c$  in  $\exp(-5/c) = .01$ , which yields  $c = .92$ ; we approximate our control parameter value after 20 or so iterations to be around 1.0.

We continue with a determination of the initial value of the control parameter. Consider the simple 100 point/10 knot case ( $n = 0$ ) wherein we are initially willing to accept a worse configuration whose average cost is not greater than 10 (as compared to 2 for the slightest perturbation) with a probability of 0.5 in the first iteration. Thus, using the Metropolis criterion again, we solve for  $c$  in  $\exp(-10/c) = 0.5$  which yields  $c = 14$ . Using the same probability of acceptance for a 500 data point set where the average cost is no greater than 50 in the first iteration yields a control parameter value of 70. In general, we could express the initial value of the control parameter  $c$  as  $-N/[10 \log(.5)]$  where  $N$  is the number of data points and the probability of acceptance is initially 0.5.

Having bracketed the initial and final values of the control parameter using initial and final probabilities of acceptance of 0.5 and .01 respectively, we are now in a position to develop an expression to describe how the control parameter decreases as a function of the number of iterations through the SA algorithm. Let the recursive formula  $c_{i+1} = \alpha c_i$  describe the behavior of the control parameter. For

the 500 data point case, we have  $c_{20} = 1 = 1/70 c_0$  and  $c_1 = \alpha c_0$ . Since  $c_2 = \alpha c_1$ ,  $c_3 = \alpha c_2 = \alpha^2 c_1$ , and  $c_4 = \alpha c_3 = \alpha^3 c_1, \dots$ , we have  $c_{20} = \alpha^{19} c_1 = \alpha^{20} c_0$ . Thus, since  $c_{20} = 1 = 1/70 c_0 = \alpha^{20} c_0$ , we have  $\alpha^{20} = 1/70$  or  $\alpha = 0.80$ . Therefore the recursive formula is  $c_{i+1} = .8 c_i$  for the 500 point case. We can apply this recursive formula approach to our application of the SA algorithm as follows. Since  $c_I = \alpha^I c_0$ , we have  $\alpha^I = c_I/c_0 = 1/c_0$  for  $I = 20$  where  $I$  is the number of iterations. Thus, upon simplification, we have  $I \log(\alpha) = -\log(c_0)$  or, following exponentiation,  $\alpha = \exp[-\log(c_0)/I]$ .

We attempted one other modification to the SA method as part of its implementation in our knot selection application. Instead of decreasing the probability of acceptance as an exponential function of the control parameter and the difference in costs between the best current configuration and the proposed configuration, we made it a linear function of the same. We found that we obtained more earlier acceptances of worse configurations in this way which increased the likelihood that a better overall configuration would be found.

#### THE ENHANCED ALGORITHM AND ITS APPLICATION

We are now ready to outline the enhanced knot selection algorithm which is the subject of this paper. As before, we first identify the knots owning the most and fewest number of data points, making use of the cell method to accomplish the task efficiently. When less than 5 pairs of high-low combinations are found, the knot(s) owning the second most number of data points is identified and added to the search scheme. The knot moving scheme is then invoked tuned to the user's requirement for the degree of search to be carried out. As a minimum, the low knot is moved to coincide with the high knot, necessitating an immediate move

of one knot on top of the nearest data point. This is the monopoly splitting maneuver mentioned earlier. A greater degree of search involves subsequent moves of the low knot toward the high knot along the line connecting them once the monopoly splitting fails. Additional moves of the high point away from the low point along the line connecting them follow in accordance with the outward bound scheme. As part of any of these moves involving a worse configuration than the best one found to date, the simulated annealing method is triggered as previously described.

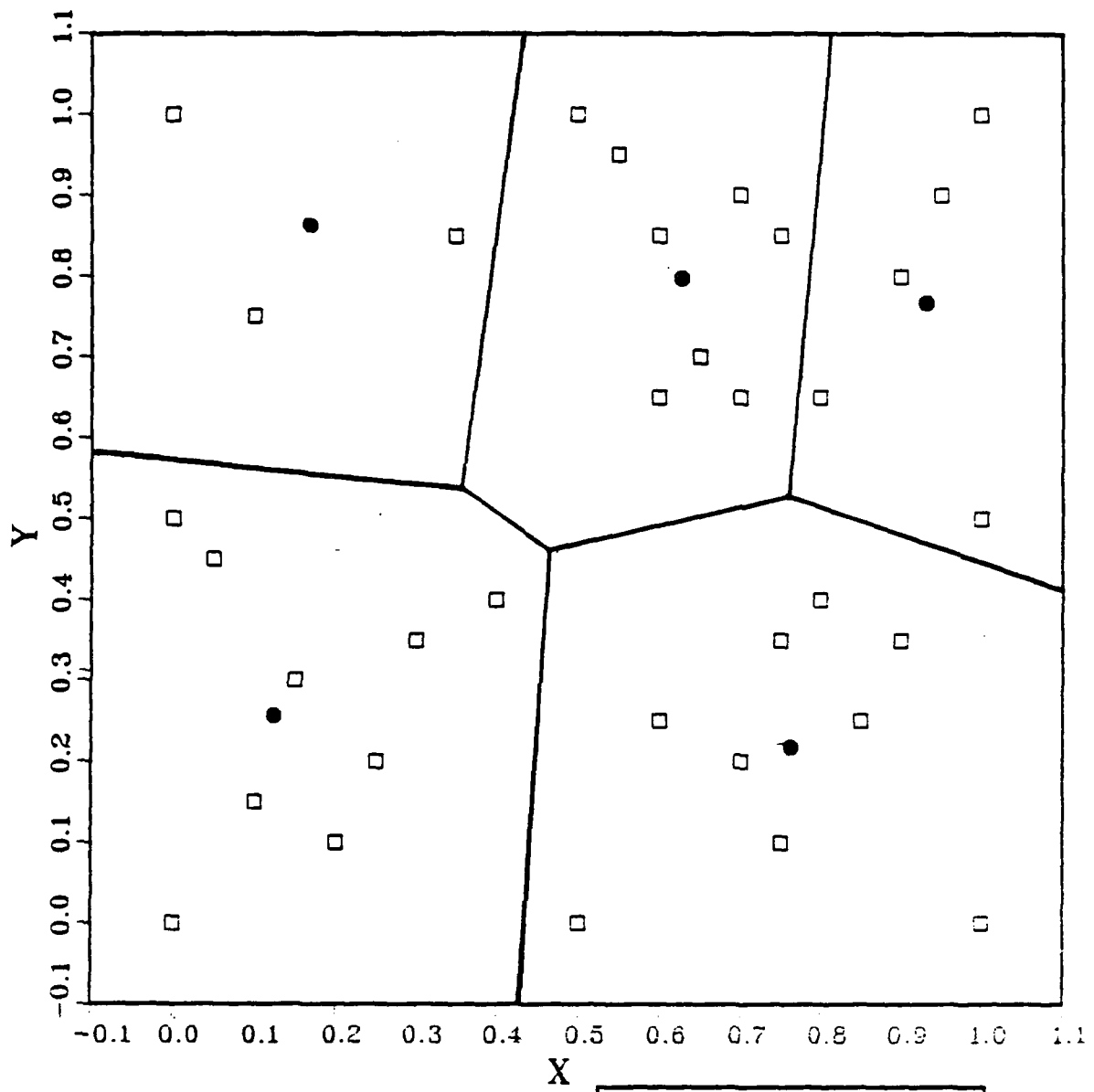
Another question that usually comes to mind has to do with how one might put this knot selection algorithm to use in conjunction with the surface evaluation using least squares thin plate splines. The program which we wrote for use here at the Academy and will publish for use by the scientific community in general incorporates several different options depending on what a prospective user might wish to do. The basic thrust of our effort has been to write a compact and efficient program to be used inside a larger user generated program written for some specific purpose. One call is made to a manager subroutine which identifies which option the user requires and which then sets up the necessary workspaces for efficient computation. As such, we envision the brunt of the computation time in search of an optimal knot configuration being accomplished as part of some preprocessing done by the user before any actual surface evaluation inside the user's program. Once the optimized knot locations have been identified, they can be used again and again within the larger code unless, of course the user generates more data as part of his particular methodology.

The first option sets the knot selection problem up, optimizes

the knot point locations, and solves for the least squares coefficients using the thin plate spline function. A user may provide his own initial guess for the knots or they can be generated in a quasi-gridded fashion automatically. Alternatively, the user may skip the knot point optimization altogether and provide his own optimized knots. This constitutes the extent of any pre-processing the user may wish to perform. However, given the parameters in the knot selection problem including the seed for the random number generator used with SA and the extent of search indicator, the user may wish to conduct further tests during the pre-processing phase in order to determine the best values of these parameters for his particular application. A user specified uniform grid of points is then used to construct a surface from the least squares coefficients found earlier. At this point, the user may wish to invoke the manager subroutine at regular or irregular time intervals in order to evaluate the surface using the least squares thin plate spline approximation method.

#### REFERENCES

1. Dongara, J. and others, LINPACK User's Guide, SIAM, 1979.
2. Harder, R.L. and Desmarais, R.N., "Interpolation Using Surface Splines", J.Aircraft, V.9, No. 2, February 1972.
3. McMahon, John R., "Knot Selection For Least Squares Approximation Using Thin Plate Splines", MS Thesis, Naval Postgraduate School, June, 1986.
4. Renka, R.J., "Multivariate interpolation of large sets of scattered data", ACM Trans. Math. Softw. 14,2 (June 1988), 139-148.
5. van Laarhoven, P. and Aarts, E., Simulated Annealing, March 1987.



**LEGEND**  
 □ = DATA POINTS  
 ● = KNOT POINTS

Figure 1. A Dirichlet Tessellation with 5 Tiles.



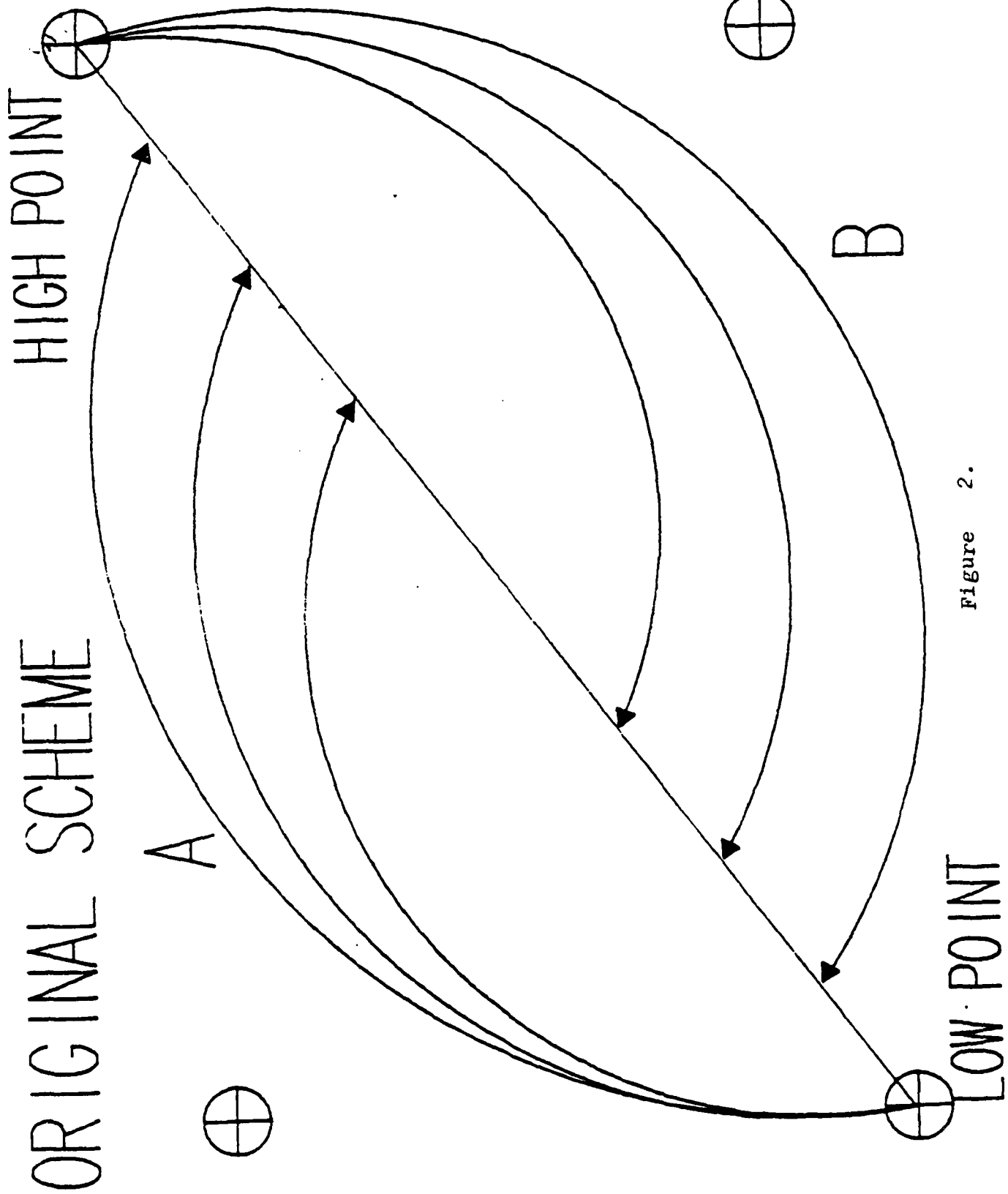
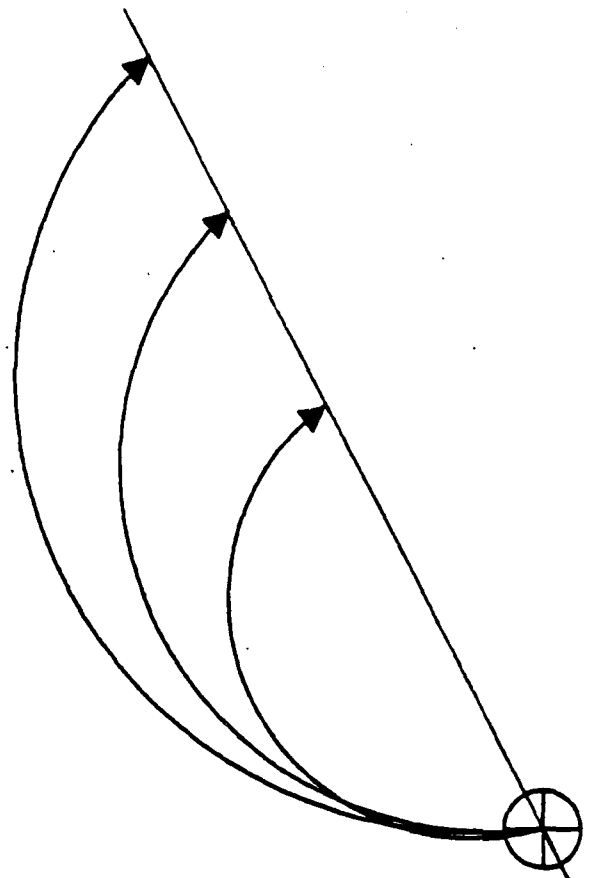


Figure 2.

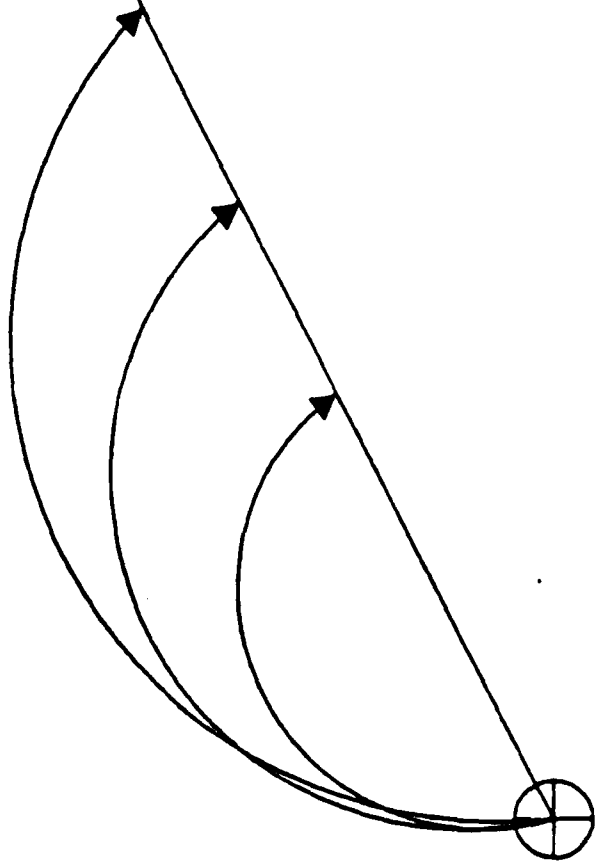
# OUTWARD BOUND SCHEME

A



HIGH POINT

B



LOW POINT

Figure 3.

# MONOPOLY SPLITTING SCHEME C

NEAREST

DATA POINT → 

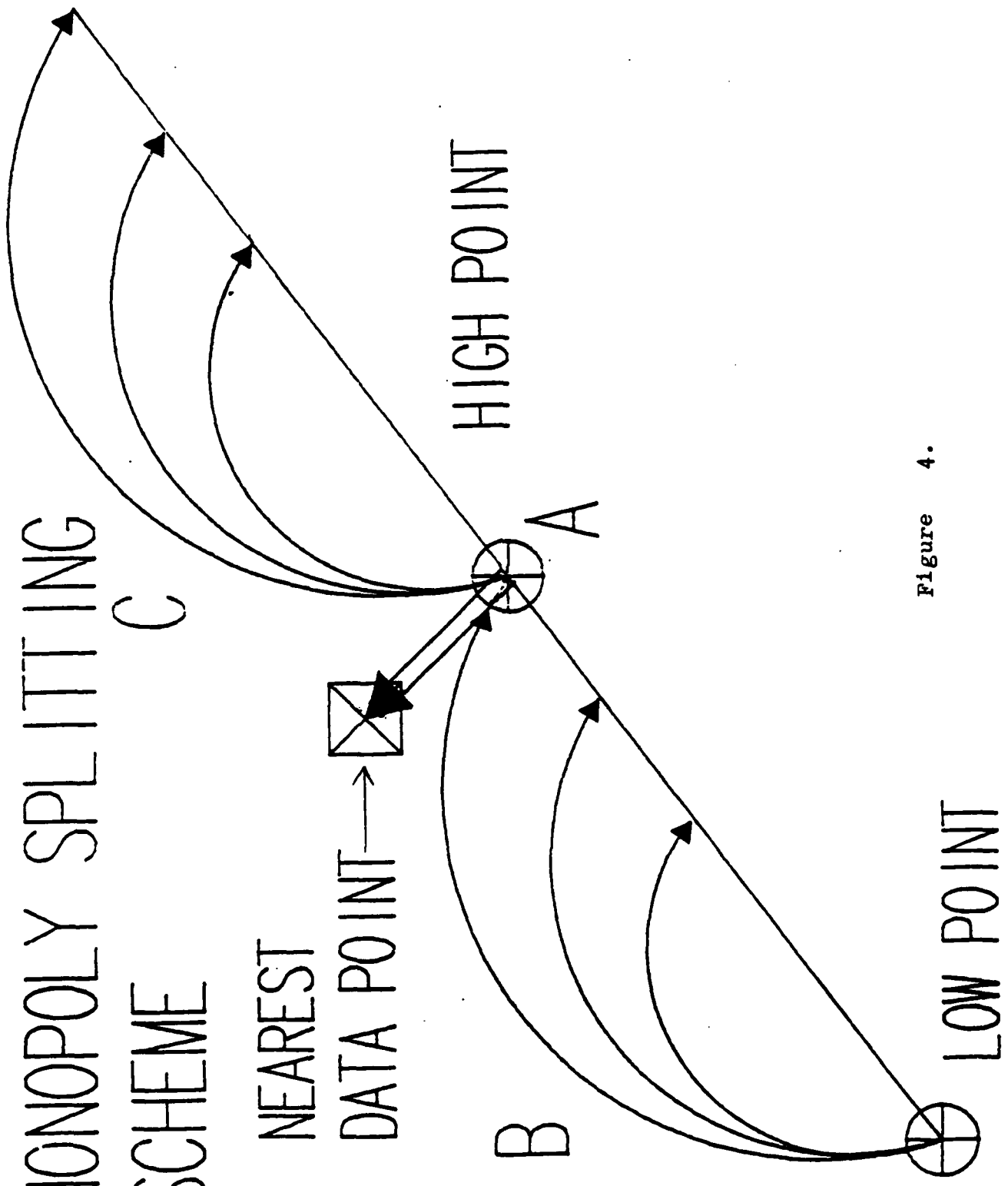


Figure 4.

# NEIGHBORLY NUDGE SCHEME

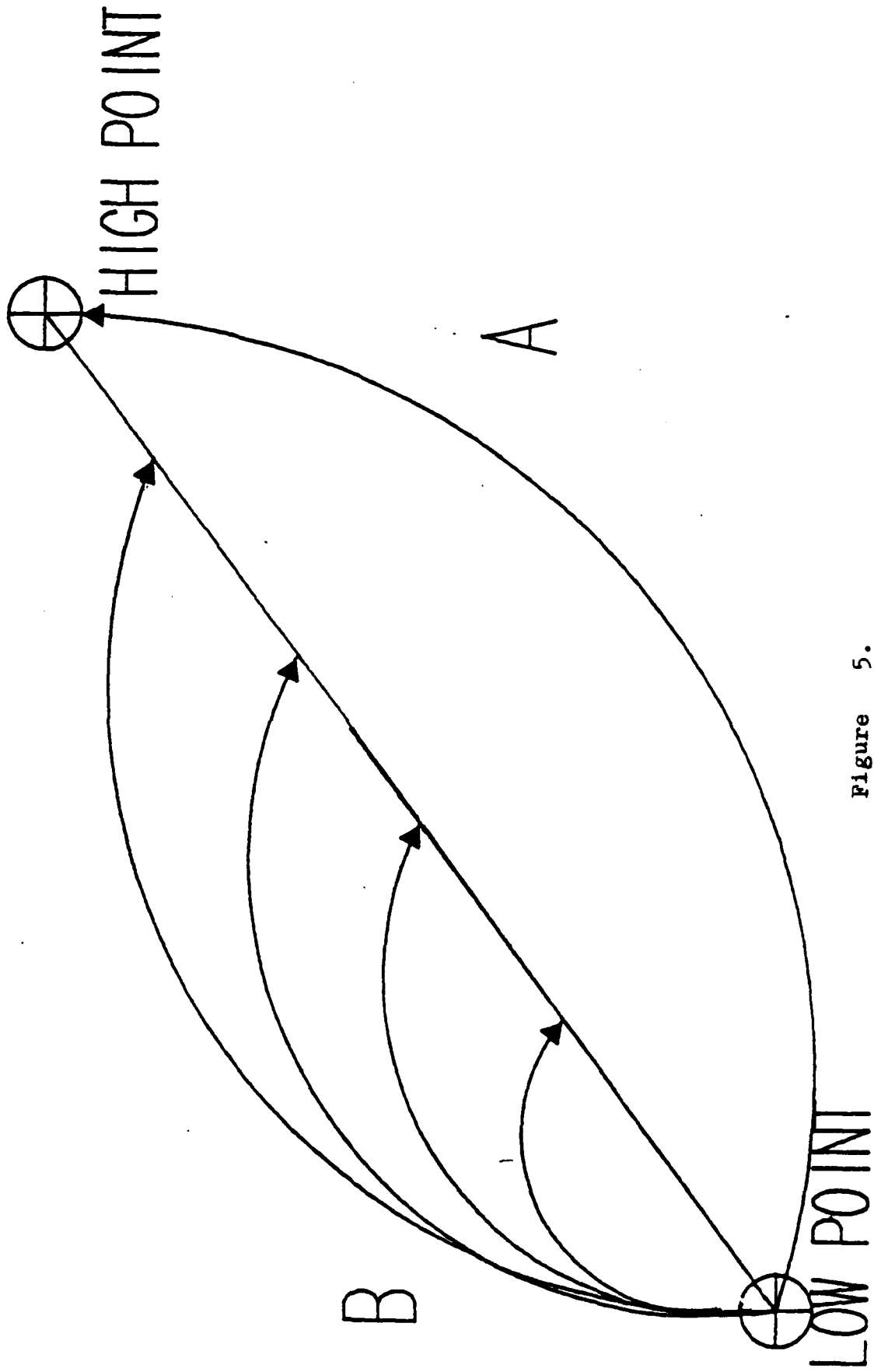
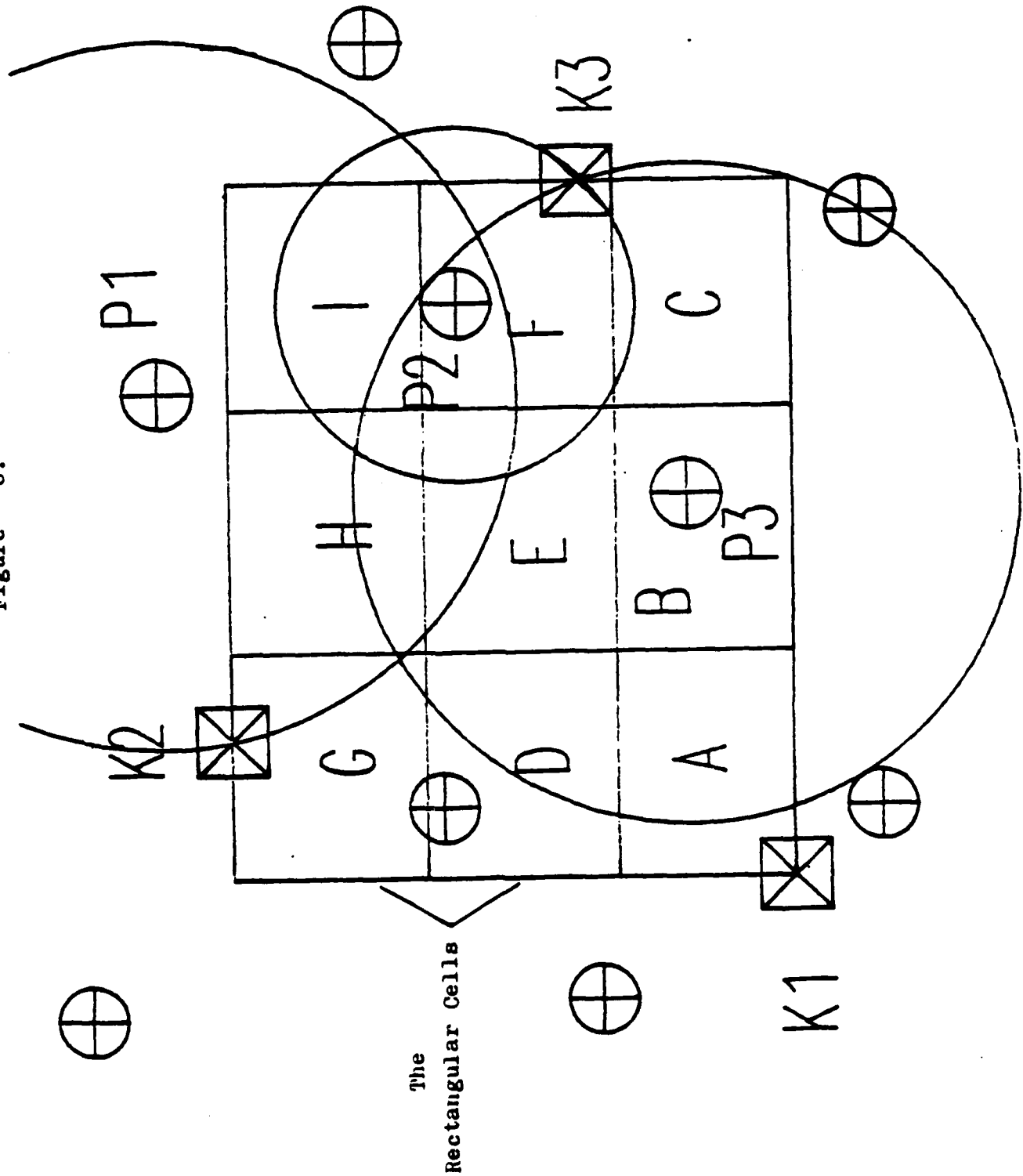


Figure 5.

Figure 6.



A MULTIVARIATE EXTENSION OF THE CRAMER-VON MISES

TEST FOR GAUSSIANTY

Kevin M. Beam  
Department of Mathematics  
United States Military Academy  
West Point, New York 10996-1786

Albert S. Paulson  
School of Management  
Rensselaer Polytechnic Institute  
Troy, New York 12180

TABLES

		<u>Page*</u>
2.1	Critical Values of $\omega_{n,p}^2$ ; $p = 1, \dots, 6$ ; $n=8, \dots, 120$ . . . . .	12
2.2	Critical Values of $\omega_{n,p}^2$ ; $p = 1, \dots, 12$ ; $n=600$ . . . . .	13
2.3	Moments of $\omega_{n,p}^2$ ; $p = 1, \dots, 6$ ; $n=8, \dots, 120$ . . . . .	14
2.4	Critical Values of $\max_p \omega_n^2$ ; $p = 1, \dots, 12$ ; $n=600$ . . . . .	15
2.5	Moments of $\max_p \omega_n^2$ ; $p = 1, \dots, 6$ ; $n=8, \dots, 120$ . . . . .	16
2.6	Comparison of Univariate Results . . . . .	17
2.7	Comparison of $\omega_{n,2}^2$ Using Pettitt's Transformation . . . . .	18
2.8	Power Study Table . . . . .	19
2.9	Johannsen's Bean Data . . . . .	24
2.10	Pig Data . . . . .	25
2.11	61 Points from a Paraboloid . . . . .	26
2.12	Iris Data . . . . .	27
2.13	Behavior of $\omega_{n,p}^2$ , $\max_p \omega_n^2$ and $\omega_{n \times p}^2$ in Examples 3 and 4 . . . . .	28
2.14	Behavior of $\omega_{n,p}^2$ , $\max_p \omega_n^2$ , $\omega_{n \times p}^2$ , and competitive statistics in Examples 3 and 4 . . . . .	29

\*To find the pages noted in this column use the set of numbers printed at the top of the pages,

FIGURES

	<u>Page</u>
2.1 Comparison of Critical Values; $n = 8, \dots, 12$ ; $\alpha = .90, .95, .99$ . . . . .	30
2.2 Frequency Graph of $\omega_{600,p}^2$ ; $p = 1, 2, 3$ . . . . .	31
2.3 Frequency Graph of $\omega_{600,p}^2$ ; $p = 1, \dots, 5$ . . . . .	32
2.4 Contour Graph of Johanssen's Bean Data . . . . .	33
2.5 Contour Graph of Pig Data .. . . .	34
2.6 Graph of Paraboloid Data . . . . .	35



## Introduction

There are several goodness-of-fit tests based on the empirical distribution function, e.d.f., for example the Kolmogorov-Smirnov, the Cramér-von Mises, the Anderson Darling, and Watson's test. The e.d.f. is defined by

$$F_n(x_i) = \begin{cases} 0 & , \quad x < x_{(1)} \\ \frac{i}{n} & , \quad x_{(1)} < x \leq x_{(i)} \\ 1 & , \quad x_{(n)} < x \end{cases}$$

where  $x_{(i)}$  is the  $i^{\text{th}}$  order statistic from a sample of  $n$  observations, or simply stated as the proportion of observations less than or equal to  $x$ . If the hypothesis is simple, that is  $F_0(x)$  is fully specified, then we have from the Strong Law of Large Numbers  $\lim_{n \rightarrow \infty} P\{F_n(x) = F_0(x)\} = 1$ . In a sense this

is the prosyllogism for all statistical theory. The well known distribution free univariate Cramér-von Mises test statistic

$$\omega^2 = n \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 d F_0(x) \quad (2.1)$$

where  $n$  is the sample size is based on a measure of divergence from this fundamental relationship. With the introduction of nuisance parameters, the composite  $\omega^2$  statistic is no longer distribution free, thus creating difficulties. Additionally, the extension of this test of fit to the multivariate arena compounds the difficulty. It is our intent to extend the composite Cramér-von Mises goodness-of-fit test statistic to  $p$ -dimensions. All results and power studies against alternatives are obtained through Monte-Carlo simulation.

### Historical Remarks

Cramér (1928) approached the basic problem of testing the hypothesis that a sample of  $n$  independent observations comes from a fully specified distribution by measuring the discrepancy between the e.d.f.,  $F_n(x)$  and the hypothesized d.f.,  $F_0(x)$  with the statistic

$$J_n = \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dx .$$

This was generalized by von Mises (1931) to

$$\omega^2 = \int_{-\infty}^{\infty} g(x)[F_n(x) - F_0(x)]^2 dx$$

where  $g(x)$  is a suitable weight function. Smirnov (1936) modified this to

$$W_n^2 = n \int_{-\infty}^{\infty} \psi[F_0(x)][F_n(x) - F_0(x)]^2 d F_0(x)$$

to yield a distribution free statistic. The special case when  $\psi = 1$ , (2.1), is commonly called the Cramér-von Mises best statistic.

Little is known about the exact distribution of the Cramér-von Mises test statistic, even in the univariate fully specified case. Several authors including Anderson and Darling (1952), Durbin and Knott (1972), and Knott (1974), have studied the distribution of the statistic in the simple univariate setting. These univariate results have been extended to include the composite hypothesis by Neuhaus (1973), Durbin, Knott and Taylor (1975), and Stephens (1976). Percentage points are given by Pearson and Hartley (1972). Multivariate extensions of the test statistic have been studied by Rosenblatt (1952), Dugue (1969), Durbin (1970), Kriuyakov, Martynov and Tyurm (1977), and Cotterill and Csörgö (1982) for the simple hypothesis. The multivariate composite setting has been investigated by Pettitt (1979) and Koziol (1982).

The most common technique used in these studies

...is to first find the decomposition of the integral operator associated with the covariance kernel of the process in terms of its eigenvalues and eigenvectors. Since the characteristic function of the functional may be expressed in terms of these eigenvalues, the requisite distribution may then be calculated by numerical inversion of the characteristic function

Koziol (1982)

### The Methodology.

The univariate Cramér-von Mises statistic (2.1) integrates to

$$\omega_n^2 = n \left\{ \frac{1}{3} + \frac{1}{n} \sum F_0^2(x_i) - \frac{1}{n^2} \sum (2i - 1)F_0(x_i) \right\} \quad (2.2)$$

where  $x_i$  is the  $i^{\text{th}}$  order statistic of  $n$  independent observations and summations are from  $i = 1$  to  $n$ . It can be shown that

$$\omega_n^2 = \frac{1}{12n} + \sum \left[ F_0(x_i) - \frac{2i - 1}{2n} \right]^2,$$

the usual computational form of the statistic. If one considers the multivariate case we have,

$$\begin{aligned} \omega_{n,p}^2 &= n \int_{R_p} \left[ \prod_{j=1}^p F_{n_j}(x_j) - \prod_{j=1}^p F_{0j}(x_j) \right] \prod_{j=1}^p dF_{0j}(x_j) \\ &= n \left\{ \left( \frac{1}{3} \right)^p - \left( \frac{1}{2} \right)^{p-1} \prod_{j=1}^p \left[ 1 - \frac{1}{n} \sum_{i=1}^n F_{0j}^2(x_i) \right] + \prod_{j=1}^p \left[ 1 - \frac{1}{n^2} \sum_{i=1}^n (2i - 1)F_{0j}(x_i) \right] \right\}. \end{aligned} \quad (2.3)$$

This is a direct extension of (2.2); if the  $F_j$ ,  $j = 1, 2, \dots, p$ , are orthogonal.

Thus let  $x_1, \dots, x_n$  be independent random  $p$ -dimensional vectors drawn from a gaussian population with d.f.  $F_\theta$ ,  $\theta = (\mu, D)$ , where  $\mu$  is an arbitrary  $p$ -dimensional vector and  $D$  is an arbitrary  $p \times p$  positive definite matrix. Define  $\psi(x, \theta) = (x - \mu)D^{-1/2}$ . If  $\theta = \theta_0$ , then the transformation  $X_i \rightarrow Y_i = \psi(x, \theta)$  yields a random sample of  $p$ -dimensional standard gaussian observations, i.e.,  $Y \sim N_p(0, I)$ .

In most situations,  $\theta_0$  is not known, and must be estimated from the sample. It is therefore of interest to determine critical values for assessing multivariate gaussianity with goodness-of-fit criteria when the parameter  $\theta$  is estimated. Let  $\hat{\theta}_n$  denote the maximum likelihood estimate of  $\theta_0$  based on  $x_1, \dots, x_n$ , that is,  $\hat{\theta}_n = (\bar{X}_n, \hat{D}_n)$  where  $\bar{X}_n = n^{-1} \sum X_i$  and  $\hat{D}_n = n^{-1} \sum (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T$ . Under the null hypothesis,  $Y' = (X - \bar{X})\hat{D}^{-1/2}$  is spherically normal, where  $\hat{D}^{-1/2} = P\Lambda^{-1/2}P^T$ ,  $\Lambda$  is a diagonal matrix having as its entries the eigenvalues of  $\hat{D}$  and  $P$  is a matrix of the associated eigenvectors of  $\hat{D}$ . Thus  $F'_0(y_1, \dots, y_p) = F'_{0,1}(y_1) \cdots F'_{0,p}(y_p)$ , and  $F'_n(y_1, \dots, y_p) = F'_{n,1}(y_1) \cdots F'_{n,p}(y_p) = \frac{\#y'_{1j} \leq y_1}{n} \cdots \frac{\#y'_{1p} \leq y_p}{n}$ , a direct extension of the univariate e.d.f., and (2.3) holds.

Monte Carlo simulations of (2.3) for gaussianity were performed. For all dimensions and sample sizes, 10,000 replications of the simulation were performed. While it is believed (2.3) approaches its asymptotic distribution very rapidly, the large number of replications insured convergence.

An alternative approach investigated is to orthogonalize the  $p$ -dimensional sample of size  $n$  as above and then consider these transformed observations as  $np$  univariate standard normal variates. This method was also studied through Monte-Carlo simulation and the associated statistic is denoted as  $\omega_{n \times p}^2$ . A final approach investigated is to consider the maximum of  $\omega_{n,1}^2$  over the margins. This statistic is noted as  $\max_p \omega_n^2$ .

All simulations were performed on a Prime 850 minicomputer using double precision. The random gaussian variates were obtained using NAG subroutines (Non-Linear Algorithms Group, Chapter G-05, 1983).

### Simulation Results.

Critical values from initial simulations for  $\omega_{n,p}^2$  of sample sizes 8,10, 12,15,20,24,30,40,60, and 120 for dimensions 1 through 6 are presented in Table 2.1.

Our univariate results are not in consonance with Pearson and Hartley (1972) who have for the composite case the following critical values where  $\alpha$  is the probability in the right tail. They indicate a relationship dependent

$$\omega^2 \quad \omega^2 \left(1 + \frac{0.5}{n}\right) \quad \frac{\alpha = .10}{0.104} \quad \frac{\alpha = .05}{0.126} \quad \frac{\alpha = .01}{0.178}$$

upon sample size. Our critical values do not follow the smooth curve as suggested by the above formula for  $n \leq 40$ . For larger sample sizes, our critical values are slightly smaller as shown in Figure 2.1. This may be due to the initialization of each simulation with the same value thus causing repetitive random numbers for differing sample sizes. For example, the first 5,000 samples of  $n = 30$  represent the 10,000 samples of  $n = 15$ .

Simulations of sample size 600 were not affected by various simulation initialization values and these percentage points were in close agreement with Pearson and Hartley. Thus  $n = 600$  was chosen as the sample size for the simulation producing the critical values for dimensions 1 through 12. Various percentage points and moments are given in Tables 2.2 and 2.3. From the close

agreement of the statistics' coefficients of skewness,  $\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}}$ , and

Kurtosis,  $\alpha_4 = \frac{\mu_4}{\mu_2^2}$ , where  $\mu_i$  is the  $i^{\text{th}}$  moment about the mean, we can see the

statistics are identically distributed with dimension acting as a scale

parameter. This is exemplified in Figure 2.2a through c showing frequency

graphs for  $p = 1, 2$ , and 3. Figure 2.3 depicts the statistics' frequencies for

$p = 1$  through 5 plotted against a common abscissa. From resultant moments and

graphs, the statistic  $\omega_{n,p}^2$  appears to be distributed as a non-central

Chi-Square random variable, as expected from the left hand side of (2.3).

The statistic  $\omega_{n \times p}^2$  was also investigated through simulations of dimension 2, 10, and 20. As expected, the distribution of  $\omega_{n \times p}^2$  matches that of  $\omega_{600,1}^2$ . The statistic  $\max_p \omega_n^2$  was investigated for sample size 600 for  $p = 1, \dots, 12$ . Results are given in Tables 2.4 and 2.5. Thus we've three approaches based upon the Cramér-von Mises statistic to test for multivariate gaussianity.

#### Comparison with Previous Investigations.

The results of our univariate simulations for sample size 600 match those of other investigators as given in Table 2.6. This validates our results for the univariate case.

In the case of the multivariate composite hypothesis, little has been achieved. Koziol (1982) considers the empirical process (2.3) as we do but uses the transformation  $Y' = (X - \bar{X})\hat{D}^{-1}(X - \bar{X})^t$ . Thus the  $Y'$  are asymptotic chi-squared random variables with  $p$  degrees of freedom.

Pettitt (1979) again uses the empirical process (2.7) but differs with the transformation  $Y' = A(\hat{D})(X - \bar{X})$  where  $A(\hat{D}) = \Lambda^{-1/2}P^t$ ,  $\Lambda$  and  $P$  defined above. He, as Kosiol, numerically inverts the characteristic function to obtain results. As Pettitt's transformation is different from ours a simulation of 20,000 replications of 600 bivariate samples was run using his methodology. A comparison of results is given in Table 2.5. The results are in agreement and thus validates our source code for our simulation for dimension  $p \geq 2$ .

Powers of the Tests,  $\omega_{n,p}^2$  and  $\omega_{n \times p}^2$ , for Gaussianity.

The powers of the tests for gaussianity were computed by Monte-Carlo simulation and are compared with several tests for multivariate Gaussianity. The power study included the multivariate skewness statistic of Mardia (1974)

$$b_{1p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \{(x_i - \hat{\mu})^t \hat{D}^{-1} (x_j - \hat{\mu})\}^3,$$

the multivariate Kurtosis statistic of Mardia (1974)

$$b_{2p} = \frac{1}{n^2} \sum_{i=1}^n \{(\mathbf{x}_i - \hat{\mu})^t \hat{D}^{-1} (\mathbf{x}_i - \mu)\}^2,$$

the Shapiro-Wilk statistic of Malkovich and Afifi (1973)

$$SW = \frac{[\sum_{j=1}^n 2_j u_j(j)]^2}{\hat{q}_m},$$

and the Anderson-Darling statistic of Paulson, et.al. (1986)

$$\hat{A}_{n,p}^2 = - \sum_{i=1}^n \frac{2_i - 1}{n} \{ \log G_p(\hat{q}_{(i)}) + \log[1 - G_p(\hat{q}_{(n+1-i)})] \} - n,$$

where  $G_p(x)$  is the distribution function of a chi-squared variate on  $p$  degrees of freedom; the  $\hat{q}_{(i)}$  are the in-ascending order  $\hat{q}_i$  given by

$$\hat{q}_j = (\mathbf{x}_j - \mu)^t \hat{D}^{-1} (\mathbf{x}_j - \mu);$$

$m$  is the index for which  $\hat{q}_j$  achieves its maximum, i.e.,

$$\hat{q}_m = \max_{1 \leq j \leq n} \hat{q}_j;$$

the  $u_j$  are the in-ascending order

$$u_j = (\mathbf{x}_m - \hat{\mu})^t \hat{D}^{-1} (\mathbf{x}_j - \mu),$$

and the  $a_j$  are the Shapiro Wilk constants tabulated in Shapiro (1980).

Table 2.8 provides powers (in percent) for the composite test of  $p$ -dimensional gaussianity for  $p = 1, 2$ , and  $5$ ,  $n = 20$  and  $50$ , and size of test =  $0.10$ , for  $\omega_{n,p}^2$  and  $\max_p \omega_n^2$ . All powers are based on 1,000 independent replications of the

test for gaussianity under the alternative listed. The powers of the competitive tests are taken from Paulson (1973). For  $p \geq 1$  we have provided the powers of six statistics for testing the hypothesis of gaussianity against the alternatives that the true distributions are  $p$ -variate chi-squared,  $p$ -variate  $t$ ,  $p$ -variate Dirichlet,  $p$ -variate log normal, and  $p$ -variate mixtures of gaussians. The definitions of these alternative distributions follow.

Let  $(x_{1l}, x_{2l}, \dots, x_{pl})^t$ ,  $l = 1, 2, \dots, r$ , be distributed as  $N_p(0, D)$  where  $R$  is a positive definite covariance matrix. If  $y_j = \sum_{l=1}^r x_{jl}^2$ ,  $j = 1, 2, \dots, p$ , then  $(y_1, y_2, \dots, y_p)^t$  are distributed as a  $p$ -variate chi-squared with correlation matrix  $D$  on  $r$  degrees of freedom. If  $W^2$  is independent of the  $x_{jl}$ 's, and  $t_{jl} = 1^{1/2} x_{jl}/W_l$ ,  $j = 1, 2, \dots, p$ , then  $(t_1, t_2, \dots, t_{pl})^t$  follows a  $p$ -variate  $t$  distribution on  $r$  degrees of freedom. If  $y_{jl} = \exp(x_{jl})$ ,  $j = 1, 2, \dots, p$ , then  $(y_{1l}, y_{2l}, \dots, y_{pl})^t$  are distributed as a  $p$ -variate lognormal. The vector  $y$  has the mixture of gaussian distributions  $bN_p(\mu, D) + (1 - b)N_p(\mu', D')$  if in the random sample of  $n$   $y$ 's,  $b_n$  of the  $y$ 's have distribution  $N_p(\mu, D)$  and  $(1 - b)_n$  of them have distribution  $N_p(\mu', D')$ , where  $0 \leq b \leq 1$  and  $b_n$  is an integer. Following Wilks (1962) let  $x_1, x_2, \dots, x_{p+1}$ , be independent random variables having gamma distribution  $G(v_1), G(v_2), \dots, G(v_{p+1})$ ,  $y_i = \frac{x_i}{x_1 + x_2 + \dots + x_{p+1}}$ ,  $i = 1, 2, \dots, p$ , where  $f(x; v) = \frac{x^{v-1} e^{-x}}{\Gamma(v)}$ ,  $x \geq 0$ , then  $(y_1, y_2, \dots, y_p)$  has the  $p$ -variate Dirichlet distribution  $D(v_1, v_2, \dots, v_p; v_{p+1})$ . When  $p = 1$  we have the Beta distribution.

Table 2.8 provides evidence that both  $\omega_{n,p}^2$  and  $\max_p \omega_n^2$  are excellent omnibus tests except for sample size  $n$  small and short tailed alternatives. The statistic  $\omega_{n \times p}^2$  was the least powerful of the three statistics considered and results are omitted. The performance of all tests improves with increasing samples size. The test  $\omega_{n,p}^2$  as a rule dominates  $\max_p \omega_n^2$  because of the inherent loss of information concerning the  $p$ -dimensional structure in the formulation of  $\max_p \omega_n^2$ . This is offset by the ability of  $\max_p \omega_n^2$  to indicate which margin(s) are in fact causing the non-gaussianity. Thus these statistics would be used in tandem. However, little is known about the nature of non-gaussianity upon rejection of the null hypothesis so we recommend the use of Mardia's  $b_1$  and  $b_2$  in conjunction with the proposed statistics.



While the Shapiro-Wilk statistic is a somewhat better omnibus statistic than  $b_{1p}$  or  $b_{2p}$ ,  $b_{1p}$  is naturally better for skewed alternatives such as the chi-squared family and  $b_{2p}$  is inherently better for the more peaked and longer tailed distributions such as the t-family. These results are consistent with the univariate case. Our results also indicate that the Anderson-Darling tests to have a marked loss of power as  $p$  increases while our proposed statistics, especially  $\omega_{n,p}^2$ , do not.

#### Some Examples and Discussions.

EXAMPLE 1. The length and breadth of 9,440 beans as measured by W. Johnsen and studied by Wicksell taken from Pretorius (1930) is given in Table 2.9 and contour graph in Figure 2.4. The data is considered in 4 manners. Let  $X$  = length and  $Y$  = breadth. Taken independently, the margin of  $X$  yields a  $\omega_{9440,1}^2$  statistic of 34.9934 and the margin of  $Y$  yields a  $\omega_{9440,1}^2$  statistic of 42.8855.

Taken as 9440 bivariate observations,  $\omega_{9440,2}^2 = 897.5406$  and  $\omega_{9440 \times 2}^2 = 2103.6814$ . All observed statistics are much greater than  $\max(\omega_{600,1}^2) = 0.3302$  and  $\max(\omega_{600,2}^2) = 0.1485$ . Clearly the margins and bivariate observations are non-gaussian.

EXAMPLE 2. Mardia (1970) gives the number of Müllerian glands as the right and left forelegs of 2,000 male pigs where  $x$  = the number of glands on the right legs and  $y$  = the number of glands on the left legs. The data is presented in Table 2.10 and contour graph in Figure 2.5. Proceeding as above, the margin of  $X$  yields a  $\omega_{2000,1}^2$  statistic of 5.727 and the margin of  $Y$  yields 5.5918. Taken as a bivariate sample,  $\omega_{2000,2}^2 = 2.2480$  and  $\omega_{2000 \times 2}^2 = 47.9995$ . Again, the data is clearly non-gaussian.

In the above two examples the obvious non-gaussianity of the observations does not allow other statistic,  $\omega_{n,p}^2$ , to be better than the other.

**EXAMPLE 3.** Granadesiken (1977, pp50-52) gives and discusses a set of 61 bivariate observations which are constructed by systematically taking points from the surface of a paraboloid and adding spherical gaussian noise to each. The data is given in Table 2.11 and a graph of X, Y, and Z versus observation number, 1-61, is presented in Figure 2.6. From Figure 2.6 we see no non-linearity and may be led to believe the data is in fact 3 dimensional gaussian. The margins of X, Y, and Z yield  $\omega^2_{61,i}$  statistics of 0.0239, 0.0833, and 0.1092 indicating X and Y are gaussian while Z is not with p-value 0.0845. Taken as trivariate observations,  $\omega^2_{61,3} = 2.2875$  with associated p value of 0,  $\max(\omega^2_{600,3} = .0624)$ , and  $\omega^2_{61 \times 3} = 0.0654$  with associated p value of 0.3287. We see here an example where the  $\omega^2_{n,p}$  statistic outperforms the statistic  $\omega^2_{n \times p}$ . As in the power-study above, we have lost the 3-dimensional structure of the data and allowed the gaussian margins, X and Y to influence the statistic  $\omega^2_{61 \times 3}$ . The performance of these statistics is compared with the competitors in Table 2.14. We see  $\omega^2_{n,p}$  outperforms all others.

**EXAMPLE 4.** The Iris data, Table 2.12, of Fisher (1936) has been extensively studied and is used to evaluate clustering algorithms (Nicholson, 1982). In particular, the iris versicolor and iris virginica groups are very difficult to separate. The reason for this difficulty is indicated by the results listed in Table 2.13. The margins of sepal and petal length, SL and PL, for iris versicolor and iris virginica are shown to be gaussian with the same p-values for each margin. For these two varieties, the margins sepal and petal width, SW and PW, are shown to be non-gaussian. The extent of the departures from gaussianity is nearly the same for the two varieties' margins. The ability to separate iris setosa from the other

two varieties is due to the non-gaussianity of iris setosa's margins of PL and SW and the acceptance of gaussianity of the margins SL and SW. Gnanadeslom (1977, (pp218-222) finds observation numbers 16 and 42 to be unusual in the iris setosa data set. This explains the large observed statistics for iris setosa's margins PL and specifically PW.

While we cannot conclude directly that there are two populations represented by the combined sets we are able to conclude that a 4 variate gaussian model is not adequate for the data sets taken independently not as a whole.

P	$\alpha$	SAMPLE SIZE									
		8	10	12	15	20	24	30	40	60	120
1 x10	.10	0.972	0.994	0.985	1.007	1.020	1.015	1.022	1.026	1.015	1.019
	.05	1.181	1.191	1.194	1.226	1.236	1.224	1.216	1.240	1.236	1.223
	.01	1.672	1.649	1.707	1.713	1.705	1.753	1.767	1.727	1.747	1.798
2 x100	.10	5.932	5.933	5.957	5.971	6.084	6.009	6.147	6.076	6.099	6.141
	.05	6.829	6.862	6.937	6.874	6.974	6.956	7.103	7.103	6.990	7.047
	.01	8.814	9.176	9.150	9.131	9.136	9.124	9.322	9.302	9.339	9.124
3 x100	.10	2.819	2.847	2.838	2.871	2.861	2.857	2.900	2.898	2.876	2.863
	.05	3.186	3.243	3.208	3.255	3.277	3.259	3.244	3.276	3.272	3.218
	.01	4.153	4.170	4.192	4.164	4.132	4.095	4.096	4.089	4.072	4.103
4 x100	.10	1.235	1.240	1.242	1.244	1.246	1.251	1.234	1.230	1.226	1.229
	.05	1.403	1.398	1.397	1.399	1.406	1.414	1.378	1.367	1.362	1.369
	.01	1.789	1.787	1.783	1.778	1.737	1.734	1.701	1.705	1.701	1.680
5 x1000	.10	5.143	5.186	5.092	5.049	5.093	5.012	5.023	5.010	4.996	5.005
	.05	5.839	5.792	5.784	5.719	5.682	5.589	5.578	5.553	5.559	5.540
	.01	7.641	7.390	7.633	7.219	7.000	6.930	6.795	6.784	6.764	6.822
6 x1000	.10	2.076	2.045	2.047	2.020	1.997	1.990	1.995	1.980	1.978	1.951
	.05	2.348	2.304	2.317	2.256	2.227	2.228	2.208	2.208	2.195	2.159
	.01	3.078	2.980	3.005	2.890	2.791	2.771	2.743	2.699	2.659	2.601

TABLE 2.1 Critical Values of  $\omega_{n,p}^2$ ;  $p = 1, \dots, 6$ ;  $n = 8, \dots, 120$

DIMENSION	ALPHA						
	.20	.15	.10	.05	.025	.01	.001
1 ( $\times 10^1$ )	0.8197	0.9111	1.0406	1.2555	1.4895	1.7984	2.0319
2 ( $\times 10^1$ )	0.5138	0.5563	0.6099	0.6997	0.8054	0.9215	1.0105
3 ( $\times 10^2$ )	2.4726	2.6642	2.8982	3.2456	3.6272	4.0366	4.4276
4 ( $\times 10^2$ )	1.0815	1.1484	1.2366	1.3806	1.5243	1.7115	1.8329
5 ( $\times 10^3$ )	4.4361	4.6809	5.0047	5.5123	6.0138	6.5131	6.8977
6 ( $\times 10^3$ )	1.7547	1.8388	1.9599	2.1637	2.3519	2.5731	2.7378
7 ( $\times 10^3$ )	0.6771	0.7092	0.7526	0.8205	0.8949	0.9837	1.0284
8 ( $\times 10^4$ )	2.5490	2.6700	2.8200	3.0672	3.3050	3.6115	3.8463
9 ( $\times 10^4$ )	0.9514	0.9911	1.0399	1.1259	1.2078	1.3060	1.3848
10 ( $\times 10^5$ )	3.4812	3.6211	3.8172	4.1167	4.4562	4.8273	5.1069
11 ( $\times 10^5$ )	1.2718	1.3226	1.3867	1.4903	1.5909	1.7311	1.8343
12 ( $\times 10^6$ )	4.5909	4.7583	4.9943	5.4018	5.7826	6.2861	6.5937

TABLE 2.2 Critical Values of  $\omega_{n,p}^2$ ;  $p = 1, \dots, 12$ ;  $n = 600$

DIMENSION	MOMENTS			
	MEAN	VARIANCE	SKEWNESS	KURTOSIS
1	$5.9772(\times 10^2)$	$1.1764(\times 10^3)$	1.7410	7.7783
2	$3.9730(\times 10^2)$	$2.6254(\times 10^4)$	1.2546	5.3986
3	$1.9819(\times 10^2)$	$4.5358(\times 10^5)$	1.0597	4.6504
4	$8.8850(\times 10^3)$	$6.9824(\times 10^6)$	0.9942	4.5859
5	$3.6752(\times 10^3)$	$9.8374(\times 10^7)$	0.9095	4.5397
6	$1.4759(\times 10^3)$	$1.3755(\times 10^7)$	0.9979	5.0181
7	$5.7300(\times 10^4)$	$1.8452(\times 10^8)$	0.8758	4.3227
8	$2.1786(\times 10^4)$	$2.3771(\times 10^9)$	0.8592	4.4465
9	$8.1670(\times 10^5)$	$2.9314(\times 10^{10})$	0.7513	4.1196
10	$3.0218(\times 10^5)$	$3.7224(\times 10^{11})$	0.8196	4.2651
11	$1.1089(\times 10^5)$	$4.5496(\times 10^{12})$	0.8138	4.5588
12	$4.0240(\times 10^6)$	$5.7521(\times 10^{13})$	0.9013	4.8506

TABLE 2.3 Moments of  $\omega_{n,p}^2$ ;  $p = 1, \dots, 12$ ;  $n = 600$

DIMENSION	ALPHA						
	.20	.15	.10	.05	.025	.01	.005
1	.08196	.09110	.10405	.12555	.14894	.17983	.20319
2	.10180	.11183	.12555	.14799	.17321	.20212	.22654
3	.11499	.12505	.13840	.16161	.18494	.21204	.23462
4	.12299	.13350	.14674	.16986	.19156	.22948	.25401
5	.13123	.14181	.15638	.17833	.20025	.23273	.25500
6	.13641	.14668	.16175	.18599	.20688	.24088	.26610
7	.14053	.15081	.16511	.18740	.21220	.24385	.26815
8	.14609	.15598	.17195	.19474	.21934	.24812	.27404
9	.15011	.16027	.17397	.19823	.22126	.25107	.27147
10	.15280	.16295	.17754	.20038	.22291	.25119	.27394
11	.15587	.16643	.18014	.20390	.22637	.25679	.27577
12	.15719	.16723	.18201	.20572	.22999	.25864	.27739

TABLE 2.4 Critical Values of  $\max_p \omega_p^2$ ;  $p = 1, \dots, 12$ ;  $n = 600$

DIMENSION	MOMENTS			
	MEAN	VARIANCE	SKEWNESS	KURTOSIS
1	0.05977	0.00117	1.74100	7.77820
2	0.07731	0.00138	1.58188	7.07833
3	0.08820	0.00146	1.39058	5.98084
4	0.09684	0.00150	1.43042	6.52263
5	0.10344	0.00158	1.30936	5.79542
6	0.10881	0.00162	1.39810	6.57393
7	0.11269	0.00159	1.31411	5.87041
8	0.11752	0.00168	1.32513	5.99709
9	0.12123	0.00169	1.32798	6.28711
10	0.12354	0.00164	1.21282	5.31291
11	0.12732	0.00164	1.28152	6.09602
12	0.12880	0.00163	1.18409	5.05303

TABLE 2.5 Moments of  $\max_p \omega_n^2$ ;  $p = 1, \dots, 12$ ;  $n = 600$



	$\omega_{600,1}^2$	<u>D,K&amp;T</u>	<u>P&amp;H</u>	<u>S</u>
(1 - $\alpha$ ) = .01	.01669	.01651		
.05	.02231	.02228		
.10	.02647	.02638		
.20	.03277	.03269		
.50	.05125	.05087		
.80	.08197	.08114		
.85	.09111		.091	
.90	.10406	.10354	.104	
.95	.12555	.12602	.126	
.975	.14894		.148	
.99	.17990	.17878	.178	
$\mu$	.05977			.0595
$\sigma^2$	.001176			.00117
$10^3\mu_3$	.07025			.0709
$10^4\mu_4$	.10764			.1116
$\alpha_3$	1.7140			1.780
$\alpha_4$	7.7783			8.186

D,K&T: Durbin, Knott and Taylor (1975)

S: Stephens (1976)

P&H: Pearson and Hartley (1972)

TABLE 2.6 Univariate  $\omega_{n,p}^2$  Comparisons

<u><math>\alpha</math></u>	<u>Pettitt</u>	<u>Monte Carlo</u>
.20	.075	.068
.15	.080	.076
.10	.088	.087
.05	.100	.106
.025	.112	.127
.01	.128	.157
.005	.140	.184

TABLE 2.7 Comparison of  $\omega_{n,2}^2$  Using Pettitt's Transformation

Powers of  $\omega_{n,p}^2$ ,  $\max_p \omega_n^2$ , Mardia's  $b_{1p}$  and  $b_{2p}$ , Shapiro Wilk (SW), Anderson Darling (AD), and Kolmogorov-Smirnov (KS) tests for gaussianity at significance level  $\alpha = .10$ ;  $p = 1,2,5$ ; and  $n = 20,50$

(a)  $p = 1, n = 20$

ALTERNATIVE	STATISTIC					
	$\omega_{n,1}^2$	$b_{1p}$	$b_{2p}$	SW	AD	KS
$\chi^2(1)$	96.4	90.0	6.0	99.2	99.2	92.5
$\chi^2(2)$	82.5	69.1	44.4	91.4	90.3	73.5
$\chi^2(4)$	54.7	47.6	31.7	65.8	64.0	41.5
$\chi^2(6)$	40.7	36.8	25.2	50.9	52.0	34.0
$\chi^2(10)$	27.5	25.2	20.6	36.6	38.0	25.5
$\chi^2(14)$	21.4	20.1	16.8	28.9	29.0	19.0
$t(1)$	89.9	78.7	89.3	87.7	94.4	90.6
$t(3)$	36.9	36.7	41.8	41.8	44.0	33.5
$t(5)$	21.6	21.4	26.2	24.8	29.5	21.0
$t(7)$	16.8	17.9	18.1	20.0	22.5	15.0
$t(9)$	13.9	16.9	16.8	17.4	19.7	14.5
Lognormal	93.2	87.1	66.5	96.5	96.0	85.7
Logistic	13.9	17.4	19.5	17.7	19.5	15.0
Beta(1,1)	28.6	6.5	46.5	34.3	37.0	13.5
(1,3)	52.6	31.5	22.4	63.7	63.0	38.0
(1,5)	61.0	48.6	29.1	76.1	75.5	53.5
(2,1)	38.8	12.0	20.9	48.8	45.3	30.0
(5,1)	63.0	50.1	27.6	80.1	73.8	52.5
.8N(0,1)+.2N(0,4)	9.6	25.9	26.0	22.7	30.0	16.5
.9N(0,1)+.1N(0,4)	8.7	19.7	19.9	19.1	25.5	13.0
.7N(0,1)+.3N(1,4)	7.9	28.3	23.8	28.8	35.0	24.0

TABLE 2.8 Power Study Table

(b)  $p = 1, n = 50$ 

ALTERNATIVE	STATISTIC					
	$\omega^2_{n,1}$	$b_{1p}$	$b_{2p}$	SW	AD	KS
$\chi^2(1)$	100	99.9	92.9	100	100	100
$\chi^2(2)$	99.4	98.3	72.6	100	100	98.8
$\chi^2(4)$	89.7	89.5	52.8	97.3	93.9	80.2
$\chi^2(6)$	74.9	77.6	43.3	88.9	85.0	65.5
$\chi^2(10)$	54.3	57.5	28.8	72.4	66.0	48.0
$\chi^2(14)$	41.3	45.1	26.5	58.2	50.0	34.7
$t(1)$	100	91.2	99.3	99.7	99.8	99.6
$t(3)$	65.7	57.3	73.2	65.3	70.0	59.7
$t(5)$	34.6	32.9	46.6	36.1	42.5	31.0
$t(7)$	21.9	29.2	34.9	23.6	29.5	20.0
$t(9)$	18.5	21.6	25.1	18.2	23.5	15.0
Lognormal	100	99.9	96.1	100	100	99.3
Logistic	21.8	18.8	25.8	20.4	22.6	15.1
Beta(1,1)	61.4	9.1	92.2	94.4	77.0	42.0
(1,3)	90.0	71.6	19.4	99.8	95.2	80.0
(1,5)	96.4	92.0	37.7	99.9	98.7	91.1
(2,1)	73.6	31.2	33.1	96.4	85.5	58.5
(5,1)	96.4	91.3	36.7	99.8	99.0	89.7
.8N(0,1)+.2N(0,4)	91.2	28.8	42.9	24.5	34.3	23.0
.9N(0,1)+.1N(0,4)	73.0	29.3	34.7	22.4	26.8	16.0
.7N(0,1)+.3N(1,4)	94.6	46.1	42.4	48.7	54.0	39.5

TABLE 2.8 Power Study Table (Cont)

(c)  $p = 2, 5$  and  $n = 20$ 

ALTERNATIVE	STATISTICS						
	$\omega_{n,p}^2$	$\max_p \omega_n^2$	$b_{1p}$	$b_{2p}$	SW	AD	
$p = 2$	$\chi^2(1)$	96.9	95.2	95.0	77.1	96.8	72.9
	$\chi^2(2)$	85.3	77.6	77.2	53.9	83.6	46.7
	$\chi^2(4)$	61.9	52.5	52.4	35.8	60.4	21.8
	$\chi^2(6)$	46.0	40.4	35.6	26.8	46.0	15.3
	$\chi^2(10)$	34.7	28.1	25.3	21.6	34.3	14.6
	$\chi^2(14)$	28.5	25.5	20.9	18.6	26.0	11.1
	$t(1)$	100	100	94.8	97.5	95.6	96.8
	$t(3)$	93.3	92.1	52.2	58.1	60.3	51.3
	$t(5)$	73.9	70.3	33.7	35.8	37.0	28.1
	$t(7)$	56.3	53.6	22.7	24.8	32.0	18.3
	$t(9)$	41.1	36.8	17.6	18.2	24.8	16.5
Lognormal	95.7	100	94.6	79.1	95.3	74.5	
Dirichlet(1,1,1)	26.2	22.3	77.4	55.3	82.8	45.3	
(1,2,3)	32.4	31.5	65.9	42.3	75.1	34.9	
(2,1,2)	35.8	33.0	63.9	44.0	74.5	33.4	
(5,1,5)	61.4	49.3	57.3	38.5	69.1	24.3	
(5,1,1)	32.6	35.0	54.9	36.6	64.3	25.4	
$p = 5$	$\chi^2(1)$	82.6	71.9	98.2	89.3	92.4	67.6
	$\chi^2(2)$	56.0	42.3	81.3	63.7	74.2	35.3
	$\chi^2(4)$	34.1	23.8	49.2	39.0	48.8	12.3
	$\chi^2(6)$	24.4	16.1	34.0	25.4	38.6	9.4
	$\chi^2(10)$	16.0	13.3	20.7	17.4	26.0	8.0
	$\chi^2(14)$	15.7	11.5	17.6	16.7	24.7	8.5
	$t(1)$	100	100	99.9	99.9	99.2	99.3
	$t(3)$	100	86.5	79.4	84.3	76.6	56.6
	$t(5)$	64.6	55.1	53.9	58.7	54.1	29.9
	$t(7)$	42.8	36.1	38.7	43.2	40.9	15.6
	$t(9)$	34.4	80.9	31.2	34.5	34.7	12.6
Lognormal	87.6	80.4	98.8	94.3	95.3	77.1	
Dirichlet(1,1,1,1,1)	18.9	15.6	82.6	64.9	72.8	35.2	
(1,2,1,2,1,2)	19.3	15.9	71.6	54.9	62.1	23.8	
(5,1,1,5,1,5)	22.1	19.0	62.4	48.8	57.4	18.6	
(2,1,1,2,1,1)	19.0	14.8	72.3	54.9	67.9	28.7	

TABLE 2.8 Power Study Table (Cont)

(d)  $p = 2, 5$  and  $n = 50$ 

ALTERNATIVE	STATISTICS						
	$\omega^2_{n,p}$	$\max_p \omega^2_n$	$b_{1p}$	$b_{2p}$	SW	AD	
$p = 2$	$\chi^2(1)$	100	99.8	100	97.7	100	99.8
	$\chi^2(2)$	98.9	98.7	99.8	88.5	99.5	88.3
	$\chi^2(4)$	92.5	88.0	97.0	64.8	93.8	56.2
	$\chi^2(6)$	82.7	76.7	86.8	51.9	88.2	37.3
	$\chi^2(10)$	64.7	55.6	64.5	34.3	70.9	20.1
	$\chi^2(14)$	57.4	46.8	54.1	30.9	55.9	17.7
	$t(1)$	100	100	99.5	100	100	100
	$t(3)$	100	100	77.2	92.5	82.5	91.7
	$t(5)$	97.5	95.7	50.3	68.6	57.1	64.1
	$t(7)$	88.5	84.6	34.4	50.2	42.4	42.6
	$t(9)$	75.7	71.4	27.4	40.3	34.9	17.7
Lognormal	100	100	100	99.1	100	99.5	
Dirichlet(1,1,1)	59.1	52.0	100	89.0	99.6	87.9	
(1,2,3)	52.0	50.0	94.4	79.8	98.0	72.0	
(2,1,2)	74.8	73.0	99.2	78.0	98.6	73.9	
(5,1,5)	96.8	93.6	98.1	70.6	93.5	53.8	
(5,1,1)	71.5	74.4	97.1	69.1	94.4	57.0	
$p = 5$	$\chi^2(1)$	99.8	98.9	100	100	99.9	99.8
	$\chi^2(2)$	94.2	87.7	100	98.0	98.7	94.4
	$\chi^2(4)$	74.5	62.1	98.5	77.7	88.6	66.0
	$\chi^2(6)$	56.2	43.9	92.0	63.9	76.2	42.1
	$\chi^2(10)$	33.8	26.7	71.1	42.8	55.8	24.5
	$\chi^2(14)$	28.3	21.8	54.4	31.4	45.2	15.3
	$t(1)$	100	100	100	100	100	100
	$t(3)$	100	100	98.4	99.7	97.5	99.5
	$t(5)$	96.2	92.6	86.1	95.4	85.4	91.3
	$t(7)$	85.4	74.7	70.1	85.5	71.6	73.0
	$t(9)$	66.7	54.2	61.4	73.6	58.3	54.7
Lognormal	99.9	99.5	100	100	100	100	
Dirichlet(1,1,1,1,1,1)	38.4	34.4	100	97.0	98.7	95.6	
(1,2,1,2,1,2)	37.4	33.7	100	92.4	96.9	87.2	
(5,1,1,5,1,5)	49.5	45.2	99.7	88.7	94.7	76.5	
(2,1,1,2,1,1)	42.0	38.2	100	93.0	96.3	87.9	

TABLE 2.8 Power Study Table (Cont)

Powers of  $\omega_{50,p}^2$ ,  $\max_p \omega_{50}^2$ , and Mardia's  $b_{1p}$  and  $b_{2p}$  tests for gaussianity on various mixtures.

(e)  $p = 2, 5$  and  $n = 50$

ALTERNATIVE	STATISTICS			
	$\omega_{n,p}^2$	$\max_p \omega_n^2$	$b_{1p}$	$b_{2p}$
$p = 2$				
.8N(0,I) + .2N(0,4I)	33.1	30.1	46.9	64.1
.9N(0,I) + .1N(0,4I)	24.7	21.5	47.9	38.9
.7N(0,I) + .3N(1,4I)	64.9	54.0	67.1	69.4
.7N(0,I) + .3N(2,4I)	85.3	79.5	88.1	71.6
.7N(0,I) + .3N(3,4I)	92.1	88.5	95.2	66.4
.5N(0,A) + .5N(0,B)	14.4	13.4	17.3	21.3

where

$$A = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -.5 \\ -.5 & 1 \end{bmatrix}$$

$p = 5$				
.8N(0,I) + .2N(0,4I)	45.4	32.7	87.8	95.3
.9N(0,I) + .1N(0,4I)	27.8	21.1	73.9	81.1
.7N(0,I) + .3N(1,4I)	75.3	58.6	97.1	98.6
.7N(0,I) + .3N(2,4I)	88.5	77.0	98.6	98.4
.7N(0,I) + .3N(3,4I)	91.2	83.6	99.3	98.8
.5N(0,C) + .5N(0,D)	11.1	8.6	26.8	33.5

where

$$C = \begin{bmatrix} 1 & .5 & .5 & .5 & .5 \\ .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & -.5 & .5 & -.5 & .5 \\ -.5 & 1 & -.5 & .5 & -.5 \\ .5 & -.5 & 1 & -.5 & .5 \\ -.5 & .5 & -.5 & 1 & -.5 \\ .5 & -.5 & .5 & -.5 & 1 \end{bmatrix}$$

TABLE 2.8 Power Study Table (Cont)

Observed frequencies of the number of beans of length X and  
breadth Y measured in millimeters.

Y	X															TOTALS	
	17	16.5	16	15.5	15	14.5	14	13.5	13	12.5	12	11.5	11	10.5	10		9.5
9.125	0	2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	5
8.875	4	8	17	19	0	0	0	0	0	0	0	0	0	0	0	0	48
8.625	2	23	101	156	93	23	2	0	0	0	0	0	0	0	0	0	400
8.375	0	18	105	494	574	227	56	9	0	0	0	0	0	0	0	0	1483
8.125	0	4	44	375	956	913	362	73	12	3	0	0	0	0	0	0	2742
7.875	0	0	7	81	385	871	794	330	89	19	3	0	0	0	0	0	2579
7.625	0	0	1	4	65	236	469	361	175	55	27	4	0	0	0	0	1397
7.375	0	0	0	0	6	23	91	137	124	78	37	22	11	0	1	0	530
7.125	0	0	0	0	0	1	13	18	28	35	25	32	11	6	1	0	170
6.875	0	0	0	0	0	0	0	1	9	8	21	12	13	7	1	0	72
6.625	0	0	0	0	0	0	0	0	0	0	2	0	1	4	3	0	10
6.375	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	4
TOTALS	6	55	275	1129	2082	2294	1787	929	437	199	115	70	36	18	7	1	9440

TABLE 2.9 Johannsen's Bean Data



Observed frequencies of Müllerian glands on the right (X)  
and left (Y) forelegs of 2,000 male pigs.

Y	0	1	2	3	4	X 5	6	7	8	9	10	TOTALS
	8	4	2	0	0	0	0	0	0	0	0	14
0	5	151	65	14	5	1	0	0	0	0	0	241
1	2	58	154	88	27	7	0	0	0	0	0	336
2	0	9	96	173	119	24	8	1	0	0	0	430
3	0	3	28	128	153	92	16	8	1	0	0	429
4	0	0	7	28	77	101	58	20	3	1	0	295
5	0	0	1	6	26	52	48	13	5	3	0	159
6	0	0	0	0	3	11	16	17	3	3	0	53
7	0	0	0	0	1	9	7	9	2	2	0	30
8	0	0	0	0	0	0	0	5	2	2	1	10
9	0	0	0	0	0	0	2	0	0	1	0	3
10	15	225	353	437	411	297	155	78	16	12	1	2000
TOTALS	30	450	706	874	822	594	310	156	32	24	2	4000

TABLE 2.10 Pig Data

61 points taken from a paraboloid with added spherical gaussian noise

X	Y	Z	X	Y	Z
-2.732	6.557	25.507	-3.452	2.948	25.591
-5.264	5.253	24.200	-7.261	6.959	26.789
-5.103	5.986	26.446	-2.370	3.617	25.510
-3.335	5.888	23.947	-4.181	4.530	29.118
-5.420	5.607	25.321	-2.360	3.916	24.879
-3.261	7.697	27.479	-5.297	5.802	29.073
-4.607	6.651	26.518	-1.585	2.524	26.954
-4.236	4.220	24.416	-3.267	4.402	28.899
-4.947	5.363	26.918	-1.187	3.257	26.100
-2.189	5.881	26.282	-2.095	6.931	27.269
-2.193	5.953	26.962	-4.800	3.339	27.011
-4.838	5.909	25.196	-5.602	5.322	28.759
-3.448	5.610	27.489	-1.478	1.644	26.057
-0.990	5.391	25.667	-5.151	4.481	27.583
-6.116	6.326	30.189	-0.694	3.408	24.997
-2.175	4.645	25.613	-5.687	4.766	29.640
-5.849	6.876	26.070	-1.733	3.932	26.198
0.162	5.521	25.027	-6.154	4.932	29.631
-5.360	5.494	28.675	-3.823	3.784	25.123
-1.740	4.070	27.311	-2.588	4.923	28.343
-2.975	6.716	27.999	-3.237	3.648	26.249
-4.220	3.853	26.396	-5.740	4.537	30.277
-6.306	4.573	25.715	-0.709	1.542	27.240
-1.972	5.615	24.900	-6.568	5.335	29.631
-4.497	5.314	27.978	-1.669	1.501	25.413
-2.005	3.352	24.599	-7.690	4.578	30.863
-3.809	5.421	28.794	0.837	1.271	25.303
-2.081	3.795	25.542	-5.832	7.020	28.915
-4.907	7.120	27.449	-0.405	3.669	27.587
-0.742	2.800	26.394	-3.019	3.752	29.665
-2.750	2.233	27.669			

TABLE 2.11 Gnanedesikan Data Set

Iris Setosa				Iris Versicolor				Iris Virginica			
Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width	Sepal Length	Sepal Width	Petal Length	Petal Width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.3	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

TABLE 2.12 Iris Data

## a. Gnanedesikan Data Set

		$\omega^2_{n,p}$	p-value	$\max_p \omega^2_n$	p-value	$\omega^2_{n \times p}$	p-value
margin	X	.083	>.15	---	---	---	---
	Y	.051	>.15	---	---	---	---
	Z	.109	.08	---	---	---	---
	trivariate	.024	>.15	.136	.11	.065	>.15
b. Iris Data							
setosa	margin	SL	.072	>.15	---	---	---
		SW	.075	>.15	---	---	---
		PL	.190	.07	---	---	---
		PW	.977	0	---	---	---
		quadrivariate	.008	>.15	.124	>.15	.059 >.15
versicolor	margin	SL	.057	>.15	---	---	---
		SW	.103	.10	---	---	---
		PL	.010	>.15	---	---	---
		PW	.152	.02	---	---	---
		quadrivariate	.016	.01	.191	.03	.081 >.15
virginica	margin	SL	.089	>.15	---	---	---
		SW	.108	.08	---	---	---
		PL	.086	>.15	---	---	---
		PW	.118	.06	---	---	---
		quadrivariate	.006	>.15	.060	>.15	.050 >.15
all iris	margin	SL	.127	.05	---	---	---
		SW	.181	.01	---	---	---
		PL	1.222	0	---	---	---
		PW	.722	0	---	---	---
		quadrivariate	.019	0	.210	.02	.088 >.15
versicolor	margin	SL	.066	>.15	---	---	---
plus		SW	.158	.02	---	---	---
virginica		PL	.047	>.15	---	---	---
		PW	.243	0	---	---	---
		quadrivariate	.018	0	.198	.02	.078 >.15

TABLE 2.13. Behavior of  $\omega^2_{n,p}$ ,  $\max_p \omega^2_n$  and  $\omega^2_{n \times p}$  in Examples 3 and 4

## a. Gnanedesikan Data Set

Statistic	Value	p-value
$\omega_{n,p}^2$	0.024	>.15
$\max_p \omega_n^2$	0.14	.11
$b_{1p}$	1.19	>.15
$b_{2p}$	12.15	>.15
SW	0.98	.02
AD	1.46	.07

## b. Iris Data

setosa	$\omega_{n,p}^2$	0.008	>.15
	$\max_p \omega_n^2$	0.124	>.15
	$b_{1p}$	3.08	.12
	$b_{2p}$	0.97	.03
	SW	1.08	>.15
	AD	0.17	.08
versicolor	$\omega_{n,p}^2$	0.16	.01
	$\max_p \omega_n^2$	0.191	.03
	$b_{1p}$	2.91	>.15
	$b_{2p}$	22.60	>.15
	SW	0.95	>.15
	AD	0.36	>.15
virginica	$\omega_{n,p}^2$	0.006	>.15
	$\max_p \omega_n^2$	0.060	>.15
	$b_{1p}$	3.15	.11
	$b_{2p}$	24.30	>.15
	SW	0.94	>.15
	AD	0.42	>.15

Table 2.14. Behavior of  $\omega_{n,p}^2$ ,  $\max_p \omega_n^2$ ,  $\omega_{p \times n}^2$ , and Competitive Statistics in Examples 3 and 4

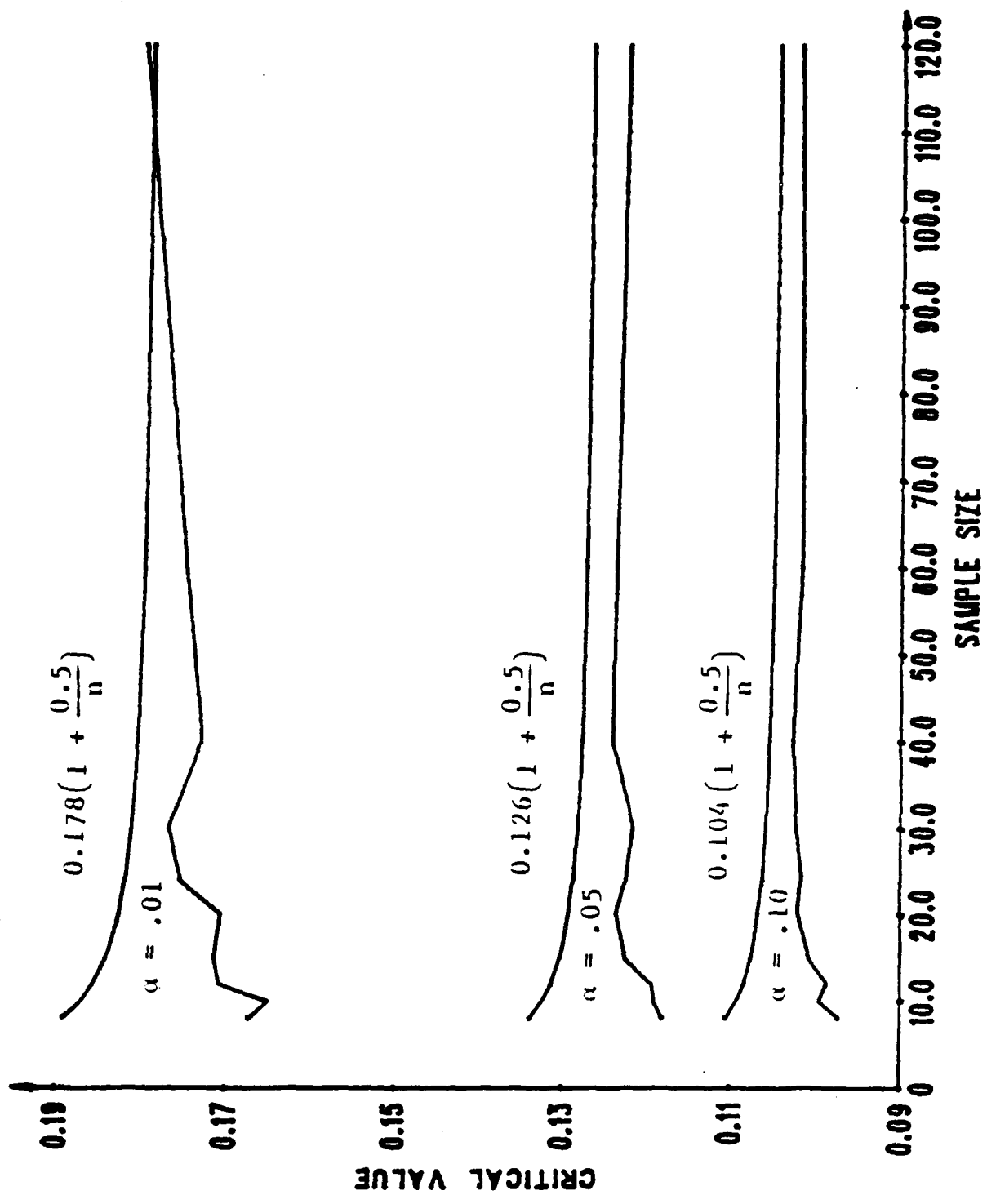


FIGURE 2.1 Comparison of Critical Values;  $n = 8, \dots, 12$ ;  $p = 1$ ;  $\alpha = .10, .05, .01$

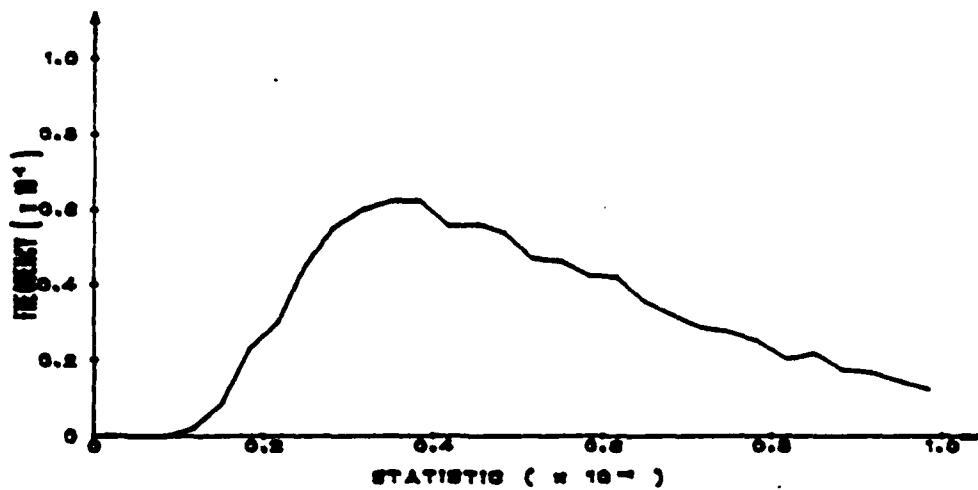
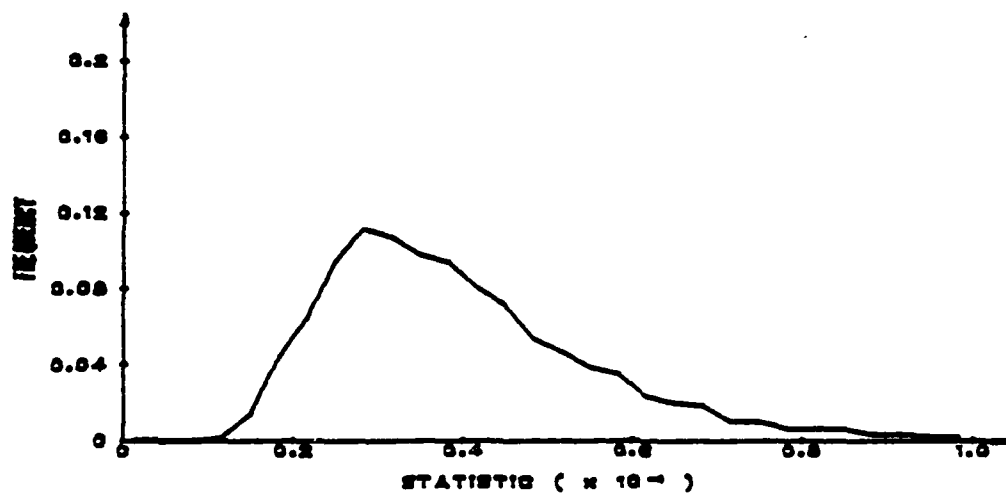
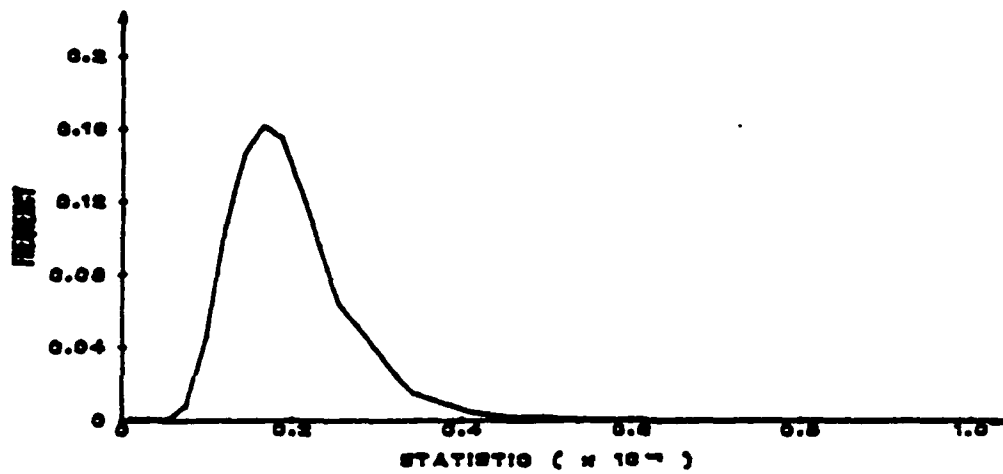


FIGURE 2.2 Frequency Graph of  $\omega_{600,p}^2$ ;  $p = 1, 2, 3$

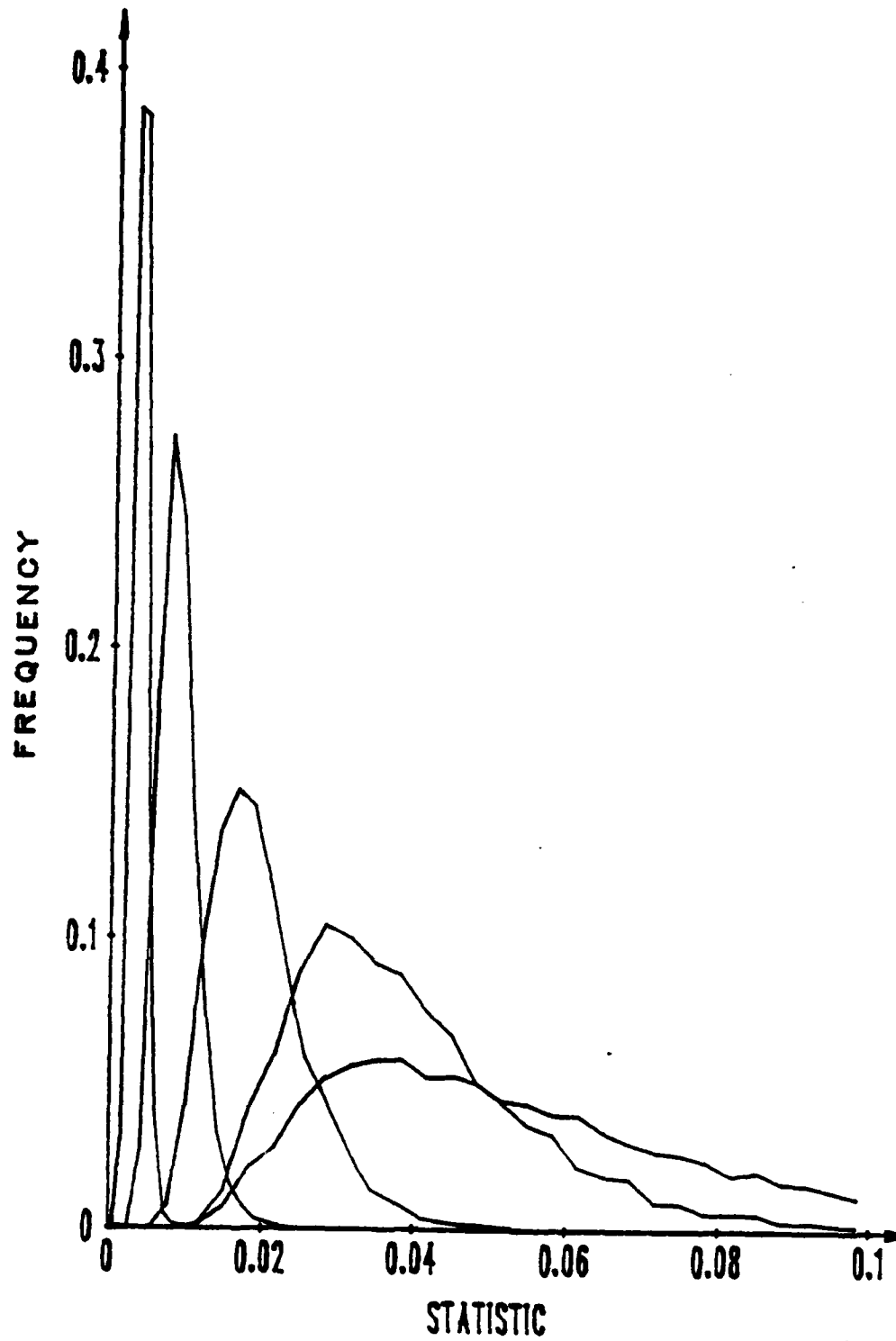


FIGURE 2.3 Frequency Graph of  $w_{600,p}^2$ ;  $p = 1, \dots, 5$



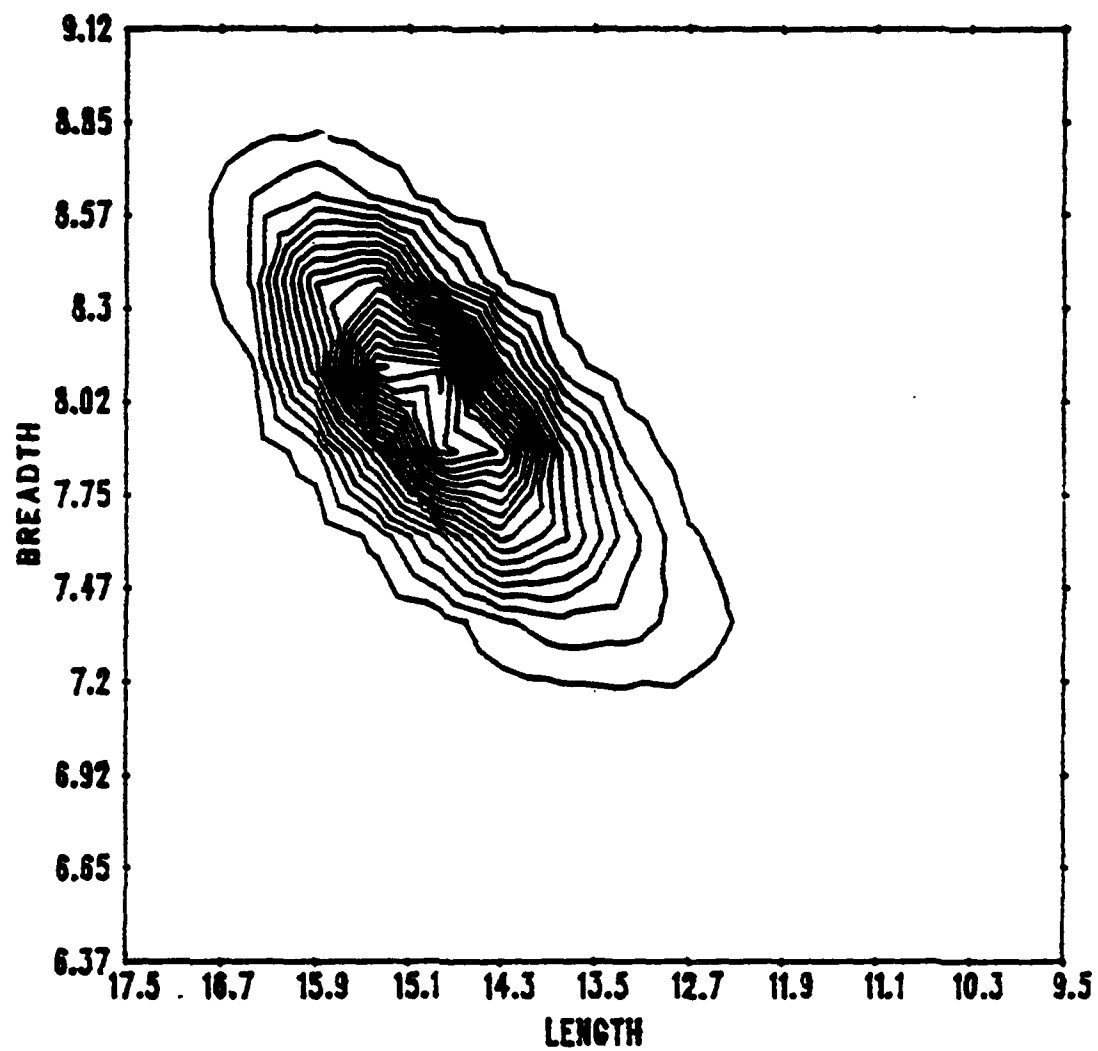


FIGURE 2.4 Contour Graph of Johanssen's Bean Data

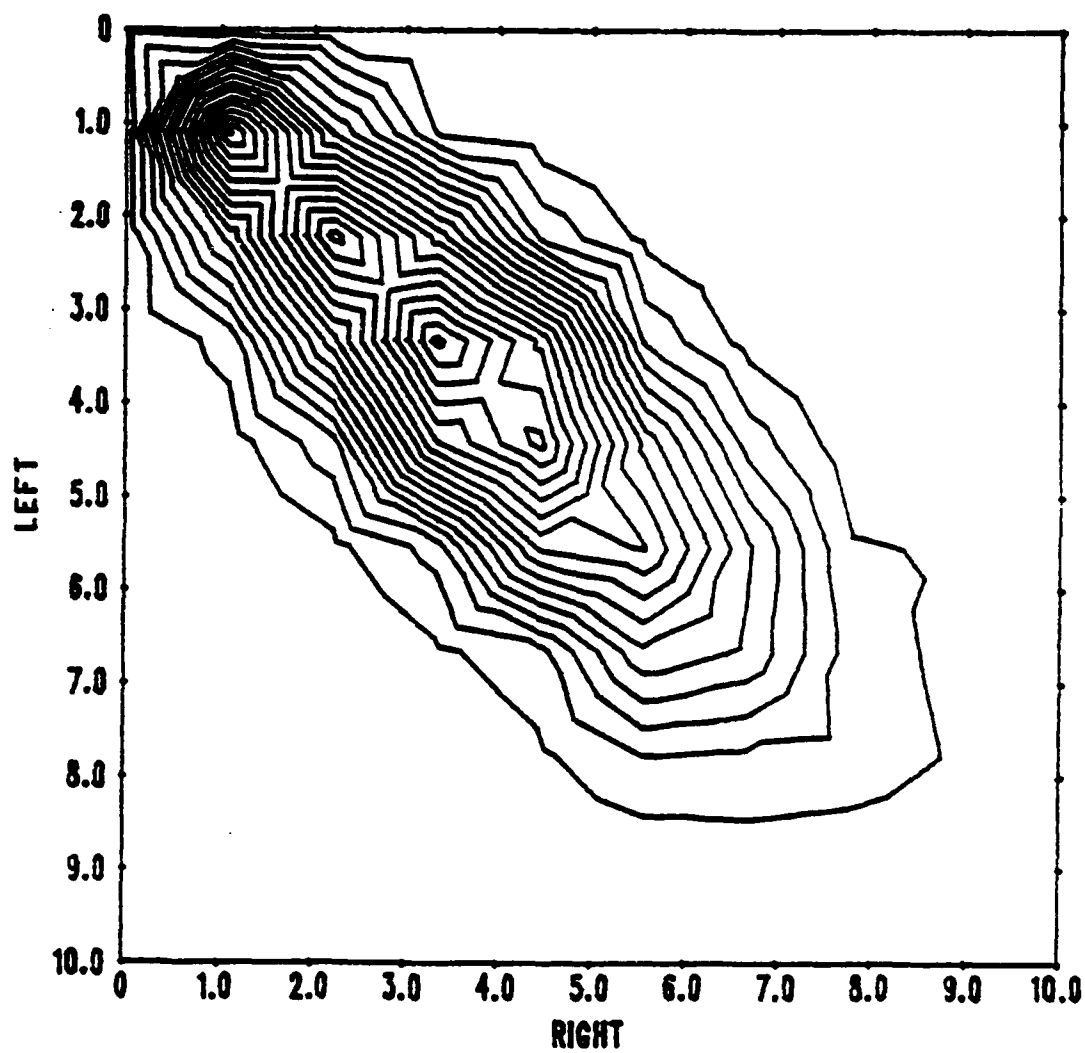


FIGURE 2.5 Contour Graph of Pig Data

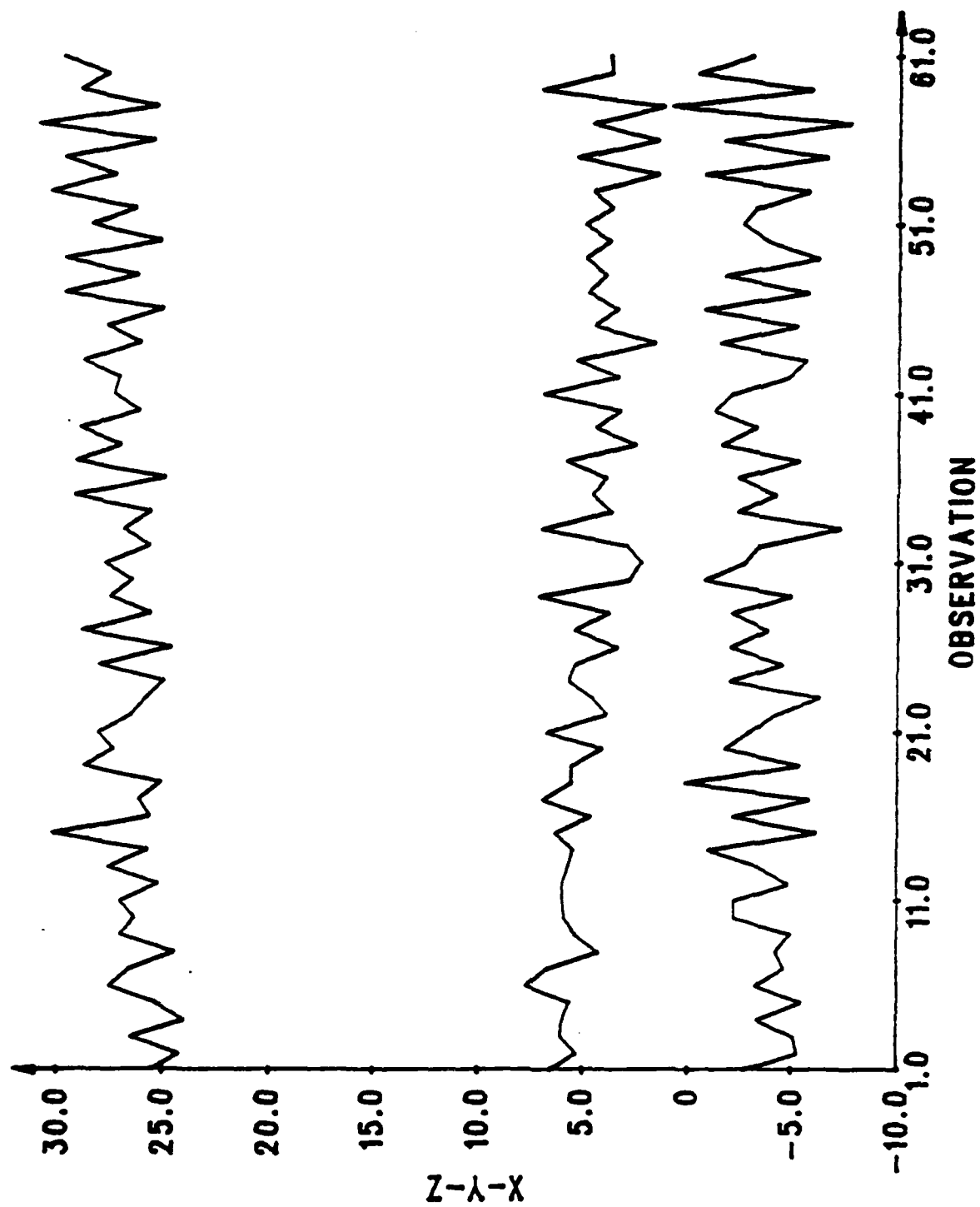


FIGURE 2.6 Graph of Paraboloid Data

## BIBLIOGRAPHY

- Anderson, T. W., and D. A. Darling. "Asymptotic Theory of Certain Goodness-of-Fit Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics*, Vol. 23, 1952, pp193-212.
- Cotterill, D. S. and M. Csörgö. "On the Limiting Distribution of and Critical Values for the Multivariate Cramér-von Mises Statistic" *The Annals of Statistics*, Vol. 10, 1982, pp233-244.
- Cramér, Harold. "On the Composition of Elementary Errors. Second Paper: Statistical Applications," *Skand. Aktuartidskr*, Vol. II, 1928, pp141-180.
- Dugue, D. "Characteristic Functions of Random Variables Connected with Brownian Motion and of the von Mises Multidimensional  $\omega_n^2$ " *Multivariate Analysis*, ed. P. R. Krishnaiah, Academic Press, New York, 1969, pp289-301.
- Durbin, J. "Asymptotic Distributions of Some Statistics Based on the Bivariate Sample Distribution Function" *Nonparametric Techniques in Statistical Inference* ed. N. L. Puri, Cambridge University Press, Cambridge, 1970, pp435-451.
- Durbin, J. and M. Knott. "Components of Cramér-von Mises Statistic I". *J. R. Statist. Soc. B*, Vol 34, 1972, pp290-307.
- Durbin, J., M. Knott, and C. C. Taylor. "Components of Cramér-von Mises Statistics II". *J. B. Statist. Soc., Ser B*, Vol. 37, 1975, pp216-237.
- Knott, M. "The Distribution of the Cramér-von Mises Statistic for Small Sample Sizes" *J. R. Statist. Soc. B*, Vol. 36, 1974, pp430-438.

- Koziol, James A. "A Class of Invariant Procedures for Assessing Multivariate Normality" *Biometric*, Vol. 69, pp423-427.
- Krlvyakov, E. N., G.V. Martynov, and Yu. N. Tyurin (translated by Richard A. Silverman). "On the Distribution of the  $\omega^2$  Statistics in the Multidimensional Case" *Theory of Probability and its Applications*, Vol. 22, 1977, pp406-410.
- Mardia, K. V. "Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies" *The Indian Journal of Statistics*, Vol. 36, Series B, Pt 2, 1974, pp115-128.
- Malkovich, J. F. and Afifi. "On Tests for Multivariate Normality" *Journal of the American Statistical Association*, Vol 68, No. 34, 1973, pp846-872.
- Neuhaus, G. "Asymptotic Properties of the Crámer-von Mises-Statistic When Parameters are Estimated." *Proceedings of the Prague Symposium of Asymptotic Statistic*, Vol II, ed. J. Jajek, Prague, 1973, pp257-297.
- Pearson, E. S., and H. D. Hartley, ed. *Biometrika Tables for Statisticians*, Vol 2, 2nd ed. Cambridge University Press, New York, 1972.
- Pettitt, A. N. "Testing for Bivariate Normality Using the Empirical Distribution Function" *Communications in Statistics, Theory and Methodology*, Vol A8, 1979, pp699-712.
- Rosenblatt, M. "Limit Theorems Associated with Variats of the von Mises Statistics" *Annals of Mathematical Statistics*, Vol. 23, 1952, pp617-623.
- Smirnov, N. V. "Sur la distribution de  $\omega^2$  (Critérium de M.R.v.Mises)," *C. R. Acad. Sci., Paris*, Vol 202, 1936, pp449-452.
- Stephens, M. A. "Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters." *The Annals of Statistics*, Vol. 4, 1976, pp357-369.
- von Mises, M. R. *Wahrscheinlichkeitsrechnung*, Wein, Leipzig, 1931.

A GENERALIZED HARMONIC BALANCE METHOD FOR FORCE  
NONLINEAR OSCILLATIONS -  
NUMERICAL SOLUTION FORMULATION AND RESULTS

B. Noble  
Brunell University, Uxbridge, UK

M. A. Hussain  
General Electric Corporate R&D Center, Schenectady, NY

J. J. Wu  
US Army Research Office, Research Triangle Park, NC

ABSTRACT

This paper deals with numerical solution formulations in conjunction with a generalized harmonic balance method, and, computational results of several specific examples in forced nonlinear vibrations. In a previous paper, approximate equations were derived using this harmonic balance method. Main results obtained in that earlier paper will be summarized here. An efficient formulation for numerical solutions is then described. The initial conditions needed in the generalized harmonic balance method can be derived from given initial conditions and such a relation is also derived here. Finally, several specific examples have been worked out. The numerical results include phase diagrams, evolution of various harmonics and comparisons between the present harmonic balance solutions and those obtained by integrating the original differential equation. Although only subharmonic cases are treated in the present paper, the formulation should apply also to superharmonic solutions.

## 1. INTRODUCTION

This paper is a sequel to one published in the Proceedings of the 1988 Army Conference on Applied Mathematics and Computing[1]. In that paper, equations were derived using a generalized harmonic balance method (G.H.B.) for problems associated with forced nonlinear oscillations where the nonlinearity has a polynomial form with dependence on a small parameter  $\epsilon$ . The method of harmonic balance proceeds by substituting a series of periodic functions into the given equation and then equate appropriate coefficients of same harmonics to zero. This approach may lead to erroneous results if carried out simply in a straightforward fashion, as noted by Nayfeh (see reference [2], for example). Unfortunately, the main alternative method, namely, multiple scaling, involves considerably laborous algebra, elimination of secular terms and solutions of differential equations at intermediate steps, and then reconstitutions of the multiple scaling results to obtain the equation governing the evolution of the amplitude and phase relation of the oscillation problem. We have shown in [1], however, that, by using only the simple part of the multiple scaling method to give the form of the solution, and then, using a generalized harmonic balance method, we can obtain the desired end-equation directly, avoiding much of the laborous algebra involved in multiple scaling, e.g., solving the intermediate differential equations and reconstitution.

The present paper deals with the numerical solution of the equations derived in [1]. In Section 2, main results derived in [1] will be summarized. An efficient formulation for numerical solutions is presented in Section 3. The initial conditions needed in the generalized harmonic balance method can be derived from given initial conditions. This relation is derived in Section 4. Finally in Section 5, several specific examples have been worked out. The numerical results include phase diagrams, evolution of various harmonics and comparisons between the G.H.B. solutions and those obtained by integrating the original differential equation. Although only subharmonic cases are treated in the present paper, the formulation should apply also to superharmonic solutions.

## 2. A BRIEF SUMMARY OF PREVIOUS RESULTS

Some of the key equations and results from the previous paper[1] are given here for easy reference. The nonlinear ordinary differential equation of interest is

$$\begin{aligned} d^2u/dt^2 + u + 2\epsilon u(du/dt) + \epsilon\alpha_1 u^2 + \epsilon^2\alpha_2 u^3 + \epsilon\alpha_3 (du/dt)^2 \\ + \epsilon^2\alpha_4 u(du/dt)^2 = 2f\cos(\Omega t) \end{aligned} \quad (1)$$

where  $u(t)$  is the unknown function  $\mu$  and  $\alpha_k$ ,  $k=2,3,4$  and  $5$ , are given constants,  $\varepsilon$  is the small parameter mentioned earlier; and,  $f$  and  $\Omega$  pertain to the magnitude and frequency of the forcing function. This equation has been treated previously by Nayfeh using the method of multiple scales[3,4].

In the subharmonic case,

$$\Omega = 2 + \varepsilon\sigma \quad (2)$$

where  $\sigma$  is a given detuning parameter.

It was shown in [1] that the solution of  $u(t)$  in (1) can be written in the following form,

$$u = \varepsilon U_0 + [(U_1 A + U_2 A^2) + \varepsilon(U_3 A^3 + U_4 A^4) + c.c.] \quad (3)$$

where the terms of order higher than one in  $\varepsilon$  have been neglected,

$$A = e^{it} \quad (4)$$

and c.c. stands for the complex conjugate; also,  $U_k$ ,  $k=0,1,\dots,4$ , are slowly varying functions compared with  $A$  of (4) in the sense that, while  $dA/dt$  is of order unity,  $dU_k/dt$  is of  $O(\varepsilon)$ .

The equations needed to obtain  $U_k$ 's were derived in [1] and they are,

$$\begin{aligned} & 2i(dU_1/dt + \varepsilon\mu U_1) - (2/3)\varepsilon(\alpha_2 + 2\alpha_4) f S \bar{U}_1 \\ & + \varepsilon^2 \{ [-\mu^2 + (2/9)f^2(3\alpha_3 + 4\alpha_5) - (1/18)f^2(5\alpha_2^2 + 12\alpha_2\alpha_4 - 12\alpha_4^2)] U_1 \\ & + (1/3)(9\alpha_3 + 3\alpha_5 - 10\alpha_2^2 - 10\alpha_2\alpha_4 - 4\alpha_4^2) U_1^2 \bar{U}_1 \\ & - (4/9)i\mu(2\alpha_2 + \alpha_4) f S \bar{U}_1 + (1/9)\sigma(11\alpha_2 + 16\alpha_4) f S \bar{U}_1 \} = 0 \end{aligned} \quad (5)$$

$$U_2 = -(1/3) f S + (1/3)\varepsilon[(\alpha_2 - \alpha_4) U_1^2 - (4/3)(i\mu - \sigma) f S] \quad (6)$$

$$U_0 = -2(\alpha_2 + \alpha_4) \bar{U}_1 U_1 - 2(\alpha_2 + 4\alpha_4) \bar{U}_2 U_2 \quad (7)$$

$$U_3 = (1/4)(\alpha_2 - 2\alpha_4) U_1 U_2 \quad (8)$$

$$U_4 = (1/15)(\alpha_2 - 4\alpha_4) U_2^2 \quad (9)$$

where a bar above a variable denotes its complex conjugate, and,

$$S = e^{i\Omega t} \quad (10)$$



### 3. SOLUTION FORMULATION FOR THE GENERALIZED HARMONIC BALANCE METHOD

We first simplify (5) by introducing the following constants,

$$\begin{aligned}
 c_1 &= -2f(\alpha_2 + 2\alpha_4)/3 \\
 c_2 &= 2f^2(3\alpha_3 + 4\alpha_5)/9 \\
 &\quad -f^2(5\alpha_2^2 + 12\alpha_2\alpha_4 - 12\alpha_4^2)/18 \\
 c_3 &= (9\alpha_3 + 3\alpha_5 - 10\alpha_2^2 - 10\alpha_2\alpha_4 - 4\alpha_4^2)/3 \\
 c_4 &= -4\mu f(2\alpha_2 + \alpha_4)/9 \\
 c_5 &= \sigma f(11\alpha_2 + 16\alpha_4)/9
 \end{aligned} \tag{11}$$

Equation (5) can now be written as

$$\begin{aligned}
 2i(dU_1/dt + \epsilon\mu U_1) + \epsilon c_1 S \bar{U}_1 \\
 + \epsilon^2 [c_2 U_1 + c_3 U_1^2 \bar{U}_1 + (ic_4 + c_5) S \bar{U}_1] = 0
 \end{aligned} \tag{12}$$

It is observed that (12) is an unsatisfactory form for numerical work since the variable  $S$  in the equation, defined in (10), is time-dependent. Due to the fact that the form of the equation

involve only terms of the form  $U_1$ ,  $S \bar{U}_1$ , and  $U_1^2 \bar{U}_1$  linearly, we can convert (12) into a differential equation with constant coefficient by setting

$$U_k = V_k S^{k/2} \tag{13}$$

Then

$$dU_k/dt = (dV_k/dt + i k \epsilon \sigma V_k / 2) S^{k/2} \tag{14}$$

and (12) becomes

$$\begin{aligned}
 2i(dV_1/dt + i\epsilon\sigma V_1/2 + \epsilon\mu V_1) + \epsilon c_1 S \bar{V}_1 \\
 + \epsilon^2 [c_2 V_1 + c_3 V_1^2 \bar{V}_1 + (ic_4 + c_5) S \bar{V}_1] = 0
 \end{aligned} \tag{15}$$

In terms of  $V_k$ , equation (6)-(9) become

$$V_2 = -(1/3)fS + (1/3)\epsilon[(\alpha_2 - \alpha_4)V_1^2 - (4/3)(i\mu - \sigma)fS] \tag{16}$$

$$V_0 = -2(\alpha_2 + \alpha_4)V_1\bar{V}_1 - 2(\alpha_2 + 4\alpha_4)V_2\bar{V}_2 \quad (17)$$

$$V_3 = (1/4)(\alpha_2 - 2\alpha_4)V_1V_2 \quad (18)$$

$$V_4 = (1/15)(\alpha_2 - 4\alpha_4)V_2^2 \quad (19)$$

Also equation (3) becomes

$$u(t) = \varepsilon V_0 + [(V_1B + V_2B^2) + \varepsilon(V_3B^3 + V_4B^4) + c.c.] \quad (20)$$

where

$$B = e^{i\phi} = \cos\phi + i\sin\phi, \quad \phi = (1 + \varepsilon\sigma/2)t \quad (21)$$

It will be convenient to use real functions to carry out computations. To do this, introduce

$$V_k = V_{kR} + iV_{kI} \quad (22)$$

Equations (15)-(19) become

$$\begin{aligned} 2dV_{1R}/dt &= \varepsilon[-2\mu V_{1R} + (\sigma + C_1)V_{1I}] \\ &+ \varepsilon^2[-C_4V_{1R} - (C_2 - C_5)V_{1I} - C_3(V_{1R}^2 + V_{1I}^2)V_{1I}] \end{aligned} \quad (23)$$

$$\begin{aligned} 2dV_{1I}/dt &= \varepsilon[-2\mu V_{1I} - (\sigma - C_1)V_{1R}] \\ &+ \varepsilon^2[+C_4V_{1I} + (C_2 + C_5)V_{1R} + C_3(V_{1R}^2 + V_{1I}^2)V_{1R}] \end{aligned}$$

$$3V_{2R} = -f + \varepsilon[-4(\sigma V_{2R} + \mu V_{2I}) + (\alpha_2 - \alpha_4)(V_{1R}^2 - V_{1I}^2)] \quad (24)$$

$$3V_{2I} = +\varepsilon[-4(\sigma V_{2I} - \mu V_{2R}) + 2(\alpha_2 - \alpha_4)(V_{1R}V_{1I})]$$

$$V_0 = -2(\alpha_2 + \alpha_4)(V_{1R}^2 + V_{1I}^2) - 2(\alpha_2 + 4\alpha_4)(V_{2R}^2 + V_{2I}^2) \quad (25)$$

$$V_{3R} = (\alpha_2 - 2\alpha_4)(V_{1R}V_{2R} - V_{1I}V_{2I})/4 \quad (26)$$

$$V_{3I} = (\alpha_2 - 2\alpha_4)(V_{1R}V_{2I} + V_{1I}V_{2R})/4$$

$$\begin{aligned}
 V_{4R} &= (\alpha_2 - 4\alpha_4)(V_{2R}^2 - V_{2I}^2)/15 \\
 V_{4I} &= 2(\alpha_2 - 4\alpha_4)V_{2R}V_{2I}/15
 \end{aligned}
 \tag{27}$$

The procedure now is to solve (23). Then substitute the resulting  $V_{1R}$ ,  $V_{1I}$  in (24), using the fact that on the right hand side we need only a zero order approximation to  $V_{2R}$  and  $V_{2I}$ , namely  $V_{2R} = -f/3$ ,  $V_{2I} = 0$ . This gives a first order approximation for  $V_{2R}$ ,  $V_{2I}$ . These values can then be substituted in (25)-(27) to give  $V_0$ ,  $V_{3R}$ ,  $V_{3I}$ ,  $V_{4R}$  and  $V_{4I}$ .

This procedure requires values for  $V_{1R}$ ,  $V_{1I}$  at  $t=0$  as initial conditions to start the numerical integration of (23). This process is considered in the next section.

To recover the solution  $u(t)$  of the original equation (1), we substitute (21), (22) in (20) to obtain

$$\begin{aligned}
 u(t) &= \epsilon V_0 + 2[V_{1R} \cos(\phi) - V_{1I} \sin(\phi) + V_{2R} \cos(2\phi) - V_{2I} \sin(2\phi)] \\
 &\quad + 2\epsilon[V_{3R} \cos(3\phi) - V_{3I} \sin(3\phi) + V_{4R} \cos(4\phi) - V_{4I} \sin(4\phi)]
 \end{aligned}
 \tag{28}$$

$$\begin{aligned}
 v(t) &= du(t)/dt \\
 &= -2\{ (1 + \epsilon\sigma/2)[V_{1R} \sin(\phi) - V_{1I} \cos(\phi)] \\
 &\quad + 2(1 + \epsilon\sigma/2)[V_{2R} \sin(2\phi) - V_{2I} \cos(2\phi)] \\
 &\quad + 3\epsilon[V_{3R} \sin(3\phi) - V_{3I} \cos(3\phi)] \\
 &\quad + 4\epsilon[V_{4R} \sin(4\phi) - V_{4I} \cos(4\phi)] \} \\
 &\quad + 2\{ (dV_{1R}/dt) \cos(\phi) - (dV_{1I}/dt) \sin(\phi) \}
 \end{aligned}
 \tag{29}$$

#### 4. INITIAL CONDITIONS FOR $V_{1R}(t)$ , $V_{1I}(t)$

In this section, the symbols  $V_0$ ,  $V_{1R}$ ,  $V_{1I}$ , ..., etc. will refer to the values of these quantities at  $t=0$ . The initial conditions for the original equation (1) are given as

$$u(0) = u_0, \quad v(0) = v_0
 \tag{30}$$

At  $t=0$ ,  $\phi=0$  in (28), (29). On solving the resulting equations for

$V_{1R}$ ,  $V_{1I}$ , we obtain

$$V_{1R} = [u_0 - 2V_{2R} - 2\varepsilon(V_0/2 + V_{3R} + V_{4R})]$$

$$V_{1I} = \{-[v_0 + 4cV_{2I} + 2\varepsilon(3V_{3I} + 4V_{4I})] + dV_{1R}/dt\}/c \quad (31)$$

where  $c = 1 + \varepsilon\sigma/2$ . We then use the following iterative procedure:

STEP 1: Drop terms of first order in  $\varepsilon$  in the right hand side of (24) and (31) and obtain

$$V_{2R} = -f/3, \quad V_{2I} = 0.$$

$$V_{1R} = u_0/2 + 2f/3, \quad V_{1I} = -v_0/2$$

STEP 2: Using these values in the right hand side of (23) and dropping terms of second order in  $\varepsilon$  gives

$$dV_{1R}/dt = \varepsilon[-2\mu(u_0/2 + 2f/3) - v_0(\sigma + c)]$$

STEP 3: Using (25)-(27), calculate  $V_0$ ,  $V_3$  and  $V_4$  with  $V_1$  and  $V_2$  obtained in STEP 1.

STEP 4: Substitute in the right hand side of (24) the values of  $V_1$  and  $V_2$  obtained in STEP 1 to obtain a new value of  $V_2$  which now is of first order in  $\varepsilon$ .

STEP 5: Finally, in (31), substitute  $V_2$  obtained in STEP 4,  $V_0$ ,  $V_3$  and  $V_4$  obtained in STEP 3, and  $dV_1/dt$  obtained in STEP 2 to calculate the new initial value of  $V_1$ , which is now of first order in  $\varepsilon$ .

Obviously the above method for obtaining initial conditions for  $V_{1R}(t)$ ,  $V_{1I}(t)$ , correct to order  $\varepsilon$ , is not unique. The question of obtaining the "best" choice of initial conditions requires further investigation.

## 5. NUMERICAL EXAMPLES

Following the procedure described in Section 4.3 and 4, several examples have been worked out. The three sets of parameters selected are:

### DATA SET I:

$$\alpha_2=0.1; \alpha_3=0.1; \alpha_4=0.1; \alpha_5=0.1;$$
$$\epsilon=0.1; \mu=0.1; f=1.0; \sigma=1.0$$

### DATA SET II:

$$\alpha_2=1.0; \alpha_3=0.5; \alpha_4=1.0; \alpha_5=0.5;$$
$$\epsilon=0.01; \mu=0.5; f=1.0; \sigma=1.0$$

and,

### DATA SET III:

$$\alpha_2=1.0; \alpha_3=0.5; \alpha_4=1.0; \alpha_5=0.5;$$
$$\epsilon=0.1; \mu=0.5; f=1.0; \sigma=1.0$$

The initial conditions for all the examples are taken as

$$u(0)=u_0; \quad (du/dt)_{t=0}=v_0$$

Note that the only difference between DATA SET II and DATA SET III is in the value of  $\epsilon$  which is 0.01 for SET II but is increased to 0.1 for SET III.

For DATA SET I, Figure I(a) is the evolution curve for  $V_0$ , which varies from its minimum value of -0.444 and settles down to a constant of about -0.083 for large  $t$  (greater than  $t=500$ , say). Since  $V_0$  is real, the phase angle is always zero. Figure I(b-1) shows the magnitude  $|V_1|$  of  $V_1$  (which is complex, as are all other  $V_k$ ,  $k=2,3$  and 4) and it varies from a maximum value of 0.949 and diminishes to about 0.003 at  $t=600$ . Since  $V_1$  represents a subharmonic motion to the problem, the fact that  $|V_1|$  diminishes to zero for large  $t$  indicates that there is no subharmonic vibrations in the steady state solution of the problem. Figure I(b-2) shows the phase angle  $\beta_1$  of  $V_1$  and it varies almost linearly with respect of time. The discontinuities simply reflect the fact that  $\beta_1$  changes sign, from  $-\pi$  to  $\pi$ , at those preselected angles so that  $\beta_1$  remains within the range of  $(-\pi, \pi)$ . Figures I(c-1,2) show the

magnitude  $|V_3|$  and the phase angle  $\beta_3$  of  $V_3$  respectively. It is noted that  $|V_3|$  is much smaller than  $|V_1|$ , with  $|V_3|_{\max} = 0.007$ , but,  $\beta_3$  behaves very much in the same way as  $\beta_1$ . The magnitudes and phase angles of  $V_2$  and  $V_4$  turn out to be constants, as can be observed also from equations (24) and (27) in Section 4, with

$$|V_2| = 0.289, \quad \beta_2 = -3.126 \text{ rads.}$$

$$|V_4| = 0.002, \quad \beta_4 = -3.111 \text{ rads.}$$

Figures I(d-1,2) show phase diagrams, i.e.,  $v(t)$  vs.  $u(t)$  as the parameter  $t$  varies from 0 to a value sufficiently large so that a steady state has almost been reached. In the case of Data Set I, this corresponds to a time  $t$  of about 600. The left hand side, Figure I(d-1), is the result by a reconstitution, using equations (28) and (29), from the  $V_k$ 's as given above, and, on the right hand side, Figure I(d-2), by integrating directly the equation (1). Figures I(e-1,2) show time evolutions of  $u(t)$  for a relatively short period from  $t=0$  to  $t=100$ . Again, on the left hand side, Figure I(e-1) shows the result by G.H.B. and reconstitution; and on the right hand side, by integrating directly the original differential equation. A comparison between Figures I(d-1) and I(d-2) and that between Figure I(e-1) and Figure I(e-2) indicate excellent agreement of results by using G.H.B. and by integrating the original differential equation directly.

Similar results are presented in figures II(a)-II(e) and III(a)-III(e) for Data Sets II and III. It is observed, however, that subharmonic vibrations exist in these two data sets as  $|V_1|=3.406$  for Set II and  $|V_1|=0.818$  and for Set III (Figures II(b-1) and III(b-1) respectively) and both stay constant as  $t$  becomes quite large.

Comparing Figure III(d-1) with III(d-2), also III(e-1) with III(e-2), it is observed that the difference of results between the two methods is more pronounced for Data Set III than for the other cases.

**ACKNOWLEDGEMENTS:** The original differential equation (1) was solved numerically by using a Runge-Katta-Fehlberg scheme[5]. The solution curves were made using a plotter routine in the Dynamical Software package[6].

## REFERENCES

[1] M. A. Hussain, B. Noble and J. J. Wu, Using Macsyma in a Generalized Harmonic Balance Method for a Problem of Forced Nonlinear Oscillation, Proc. Sixth Army Conference on Applied Mathematics and Computing (held 31 May - 3 June 1988, Univ. of Colorado, Boulder, Colorado), 1989, pp.713-732.

[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, Springer-Verlag, 1985, pp.238-314.

[5] G. E. Forsythe, M. A. Malcolm and C. B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, 1977, p.129.

[6] W. M. Shaffer, G. L. Truty and S. L. Fulmer, Dynamical Software, User's Manual and Introduction to Chaotic Systems, Dynamical Systems, Inc., 1988, p. 4.1.

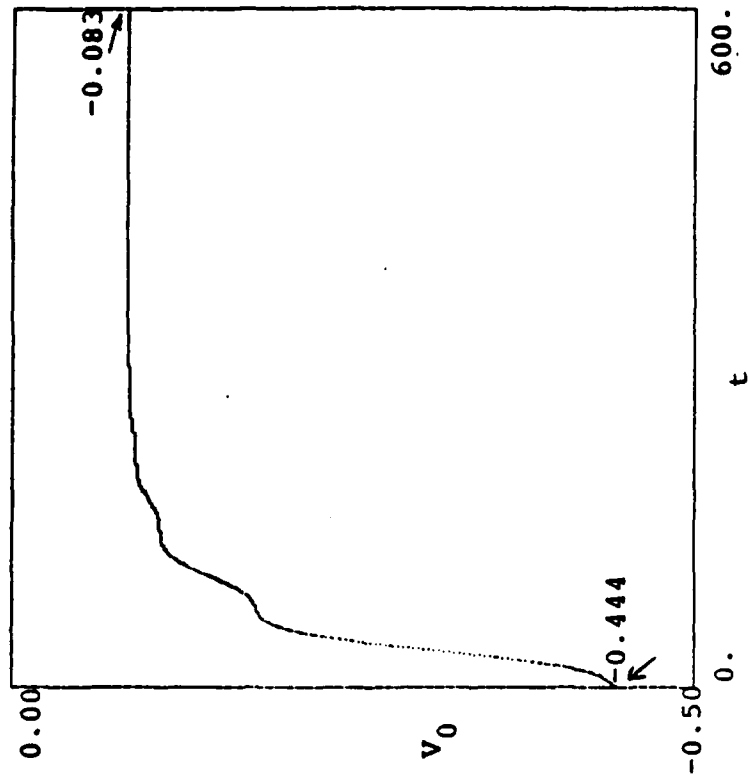


Figure I(a)



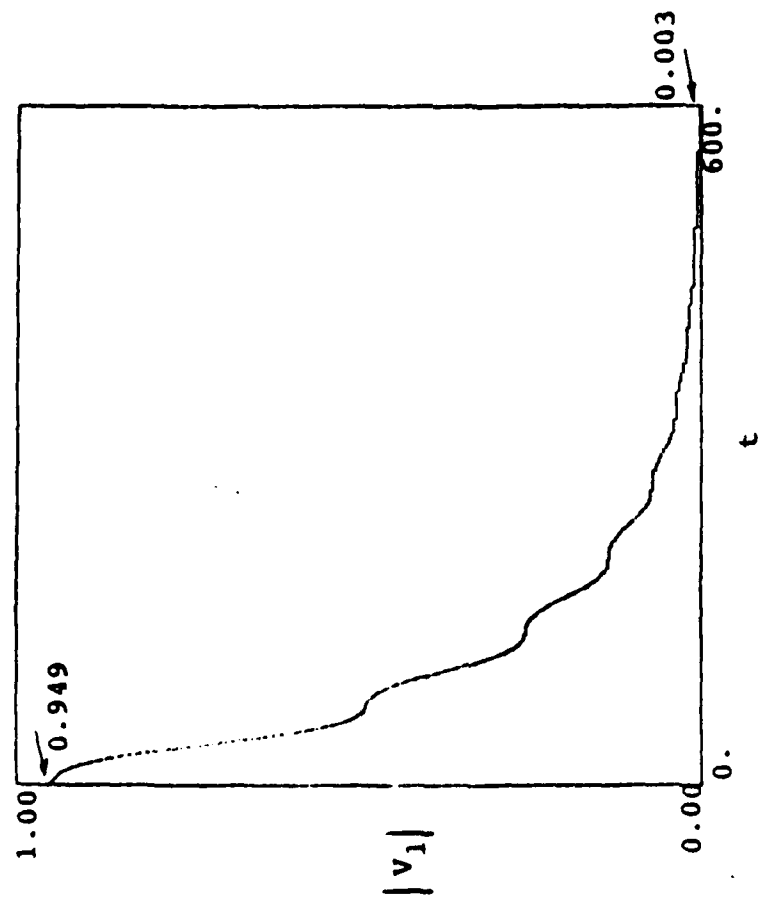


Figure I (b-1)

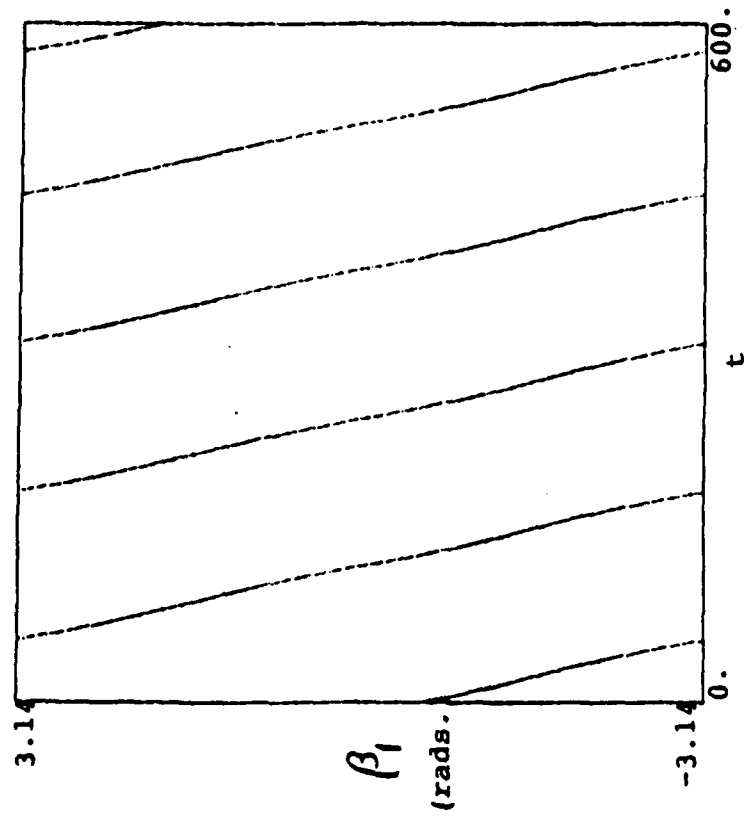


Figure I (b-2)

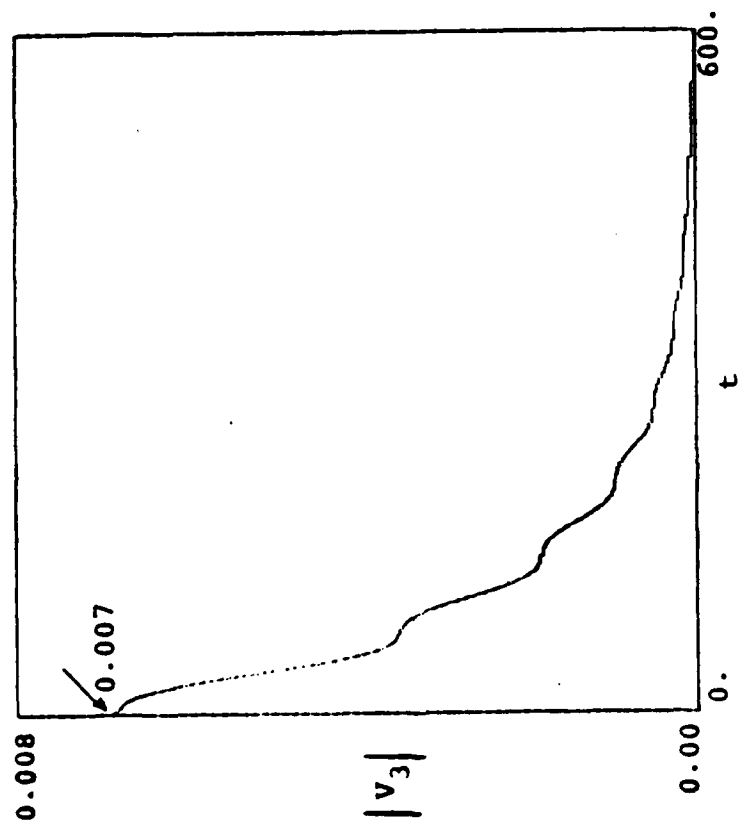


Figure I (c-1)

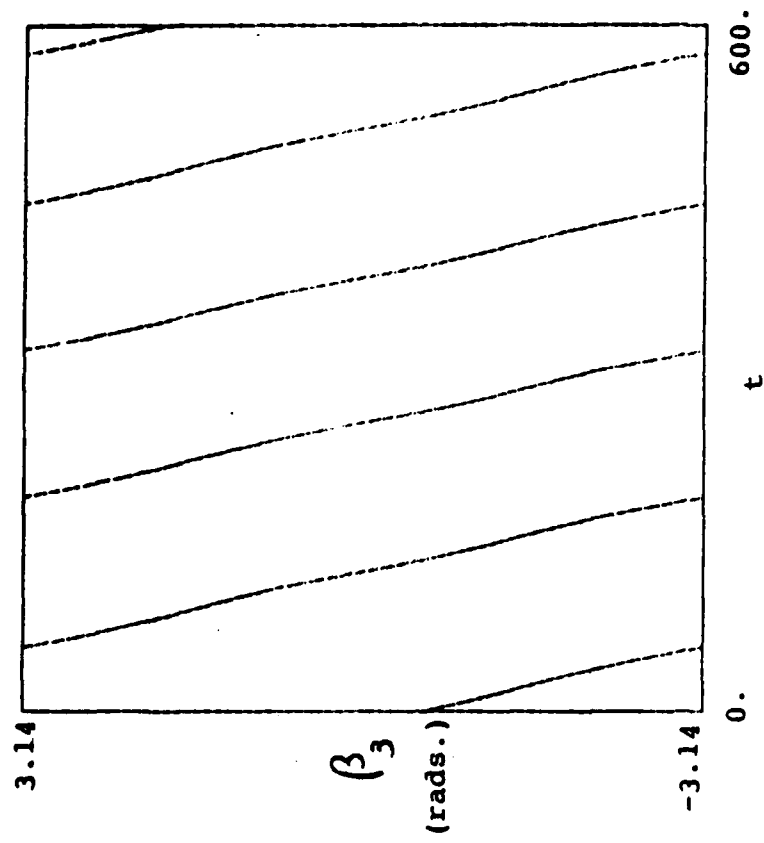


Figure I (c-2)

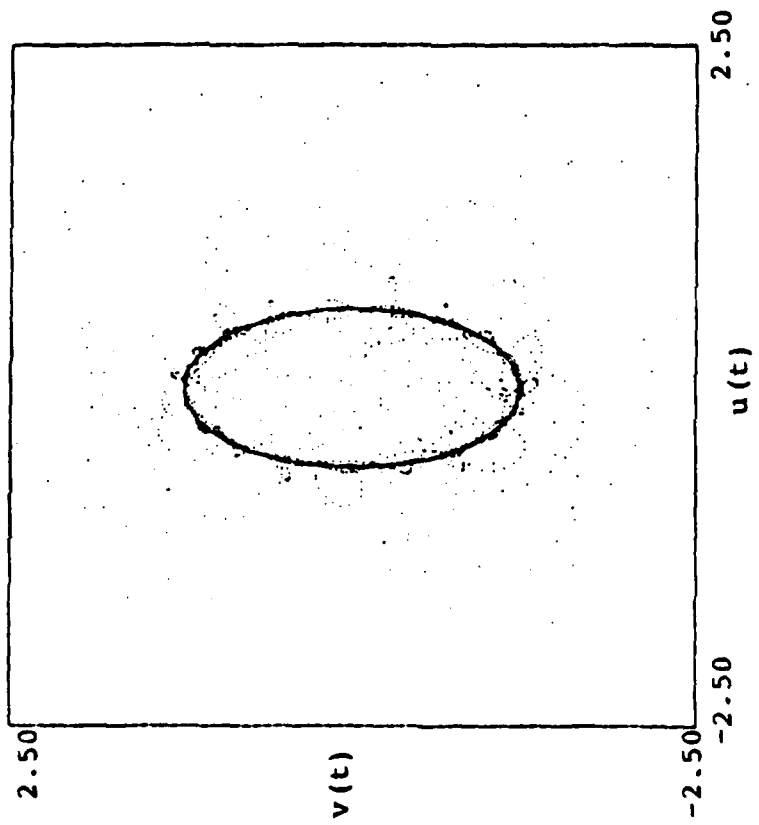


Figure I(d-1)

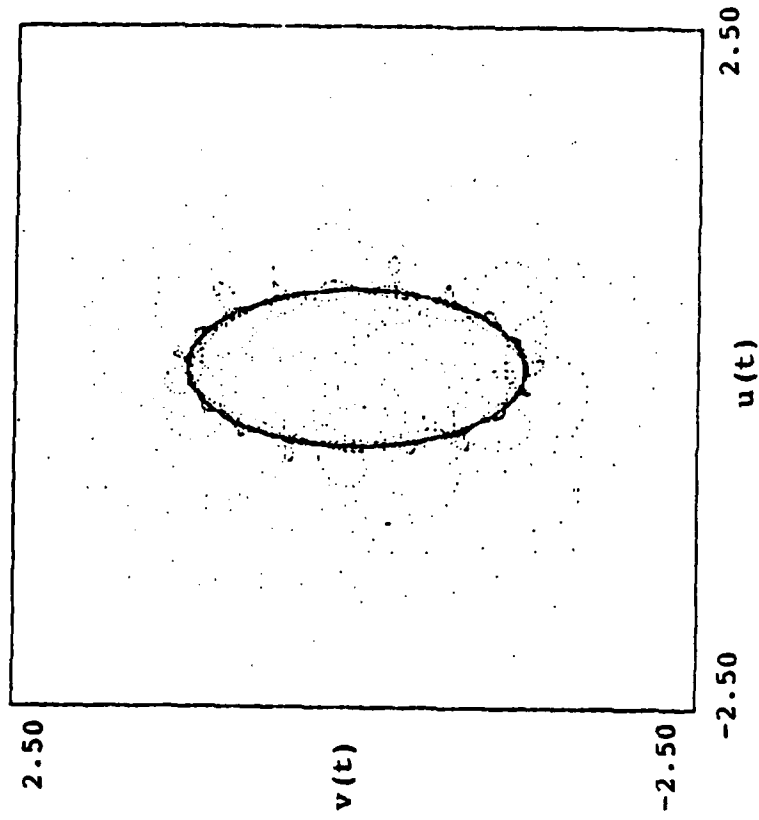


Figure I(d-2)

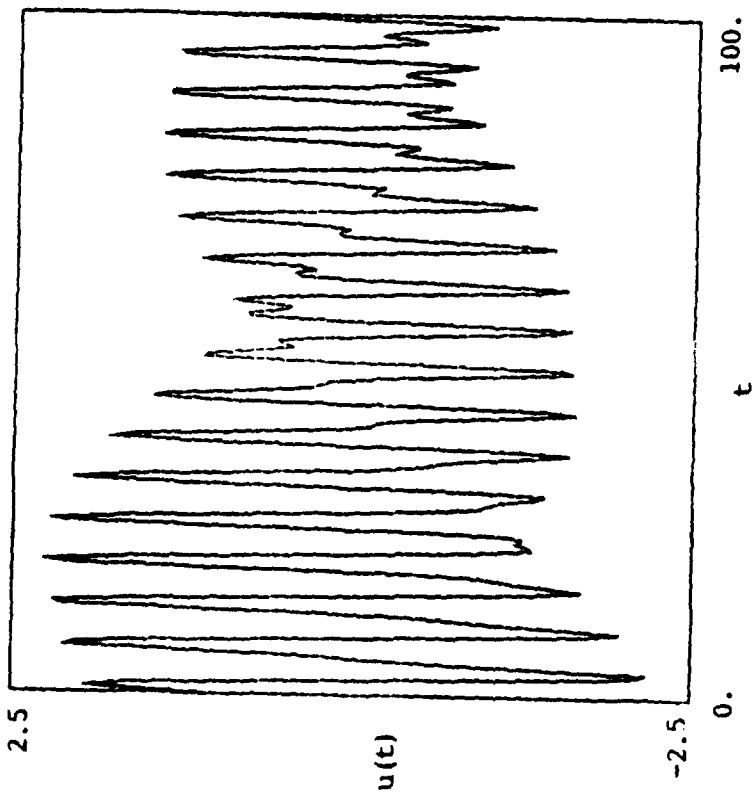


Figure I(e-1)

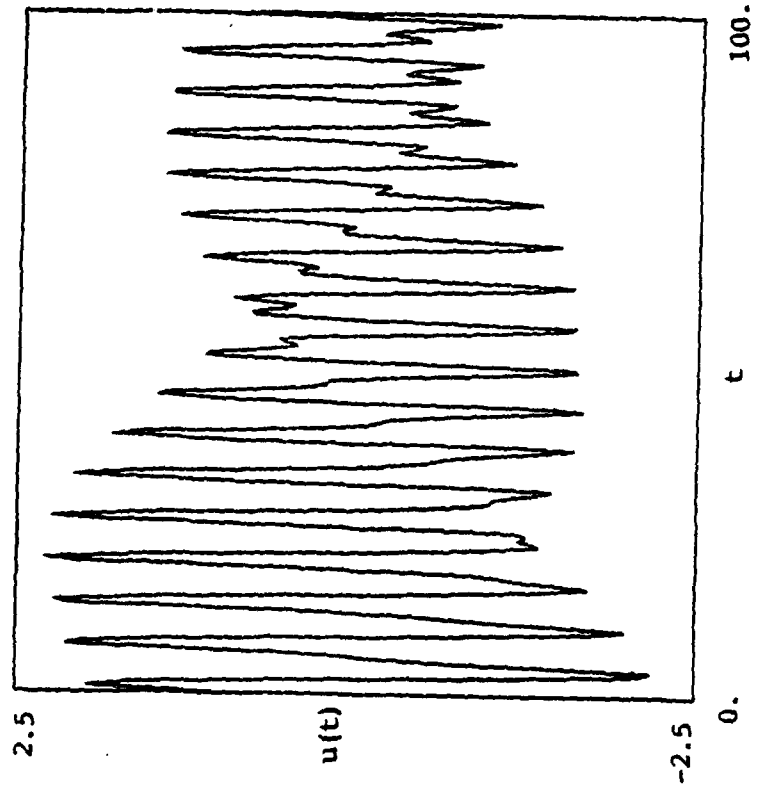


Figure I(e-2)

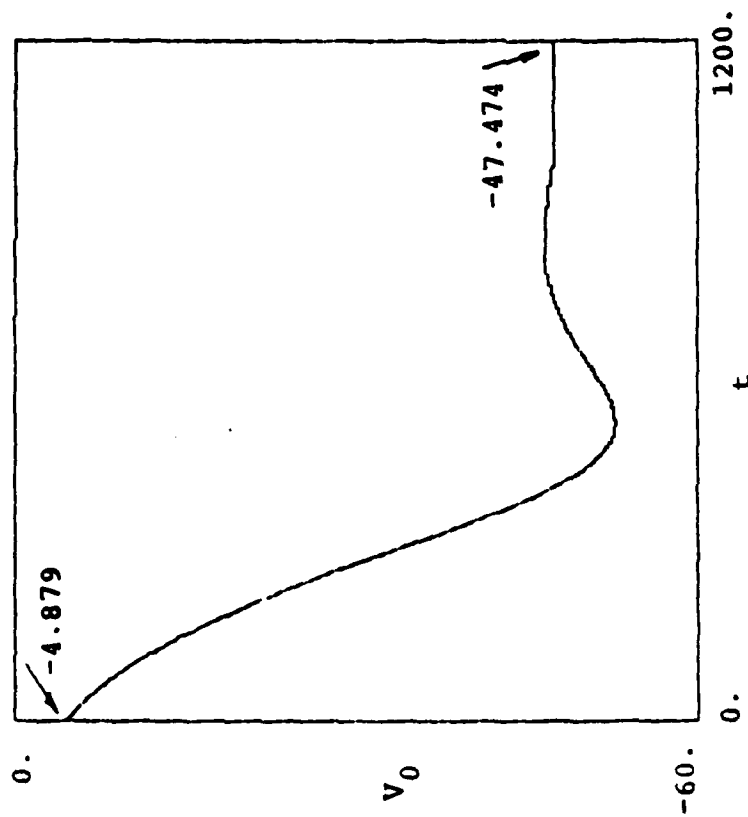


Figure II (a)

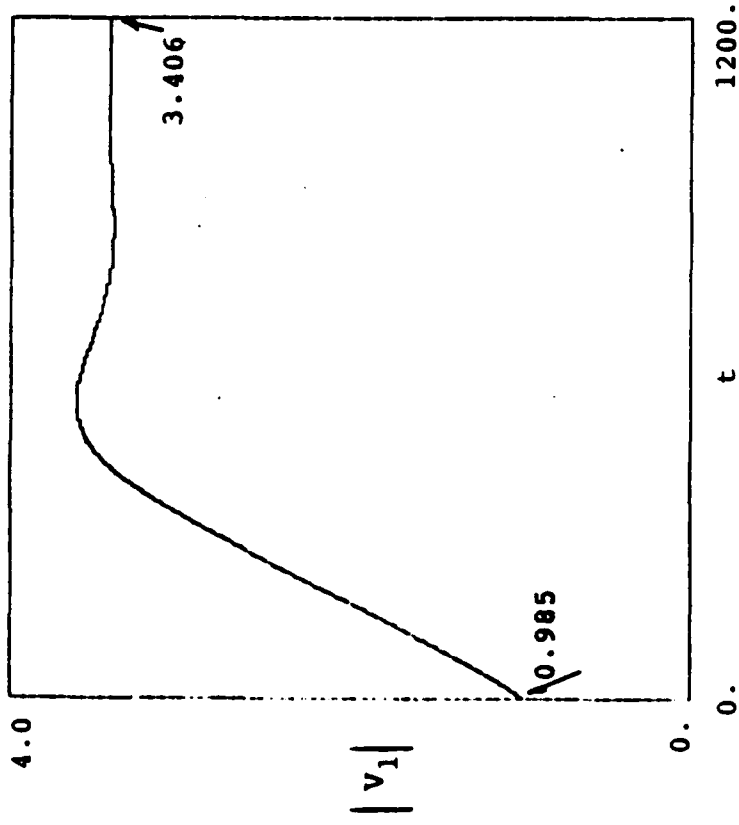


Figure 11(b-1)

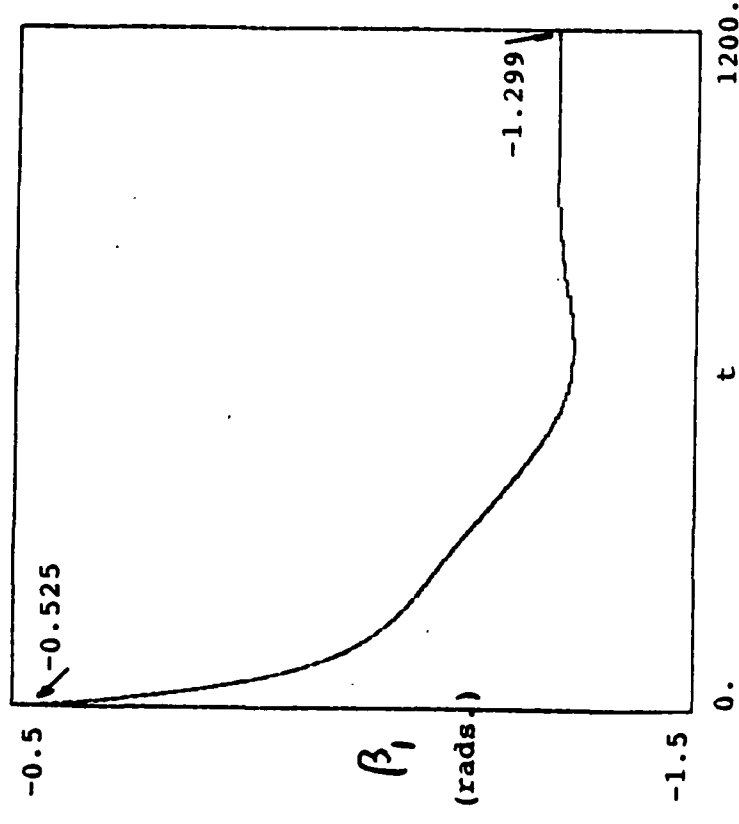


Figure 11(b-2)

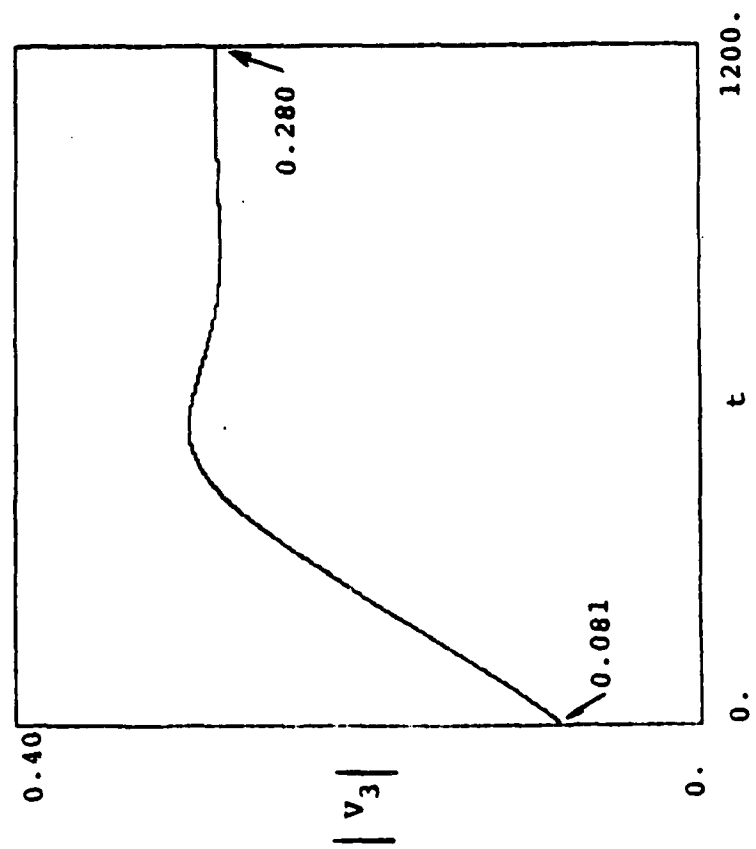


Figure II(c-1)

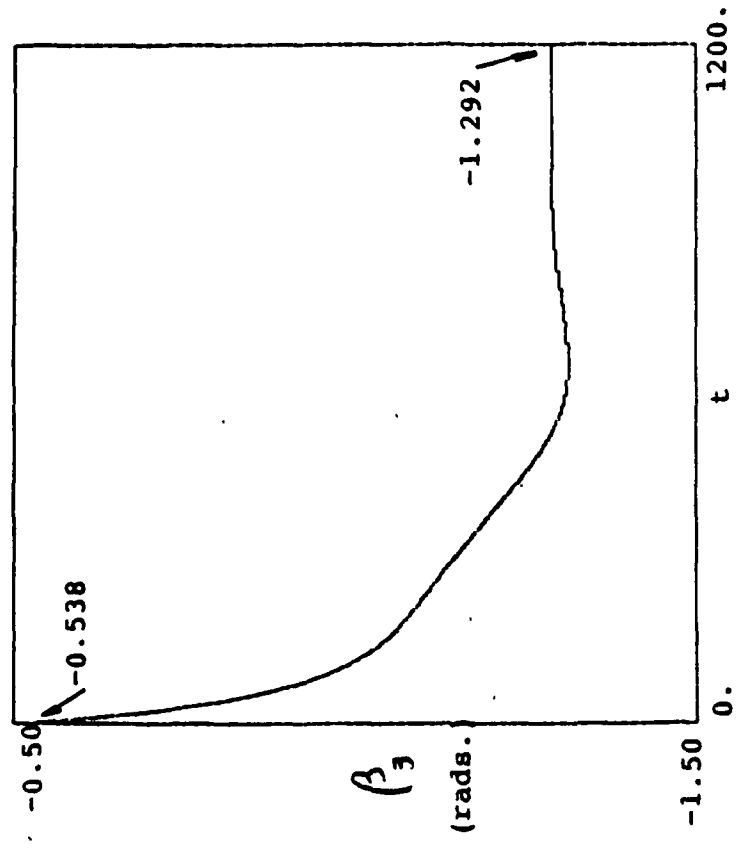


Figure II(c-2)

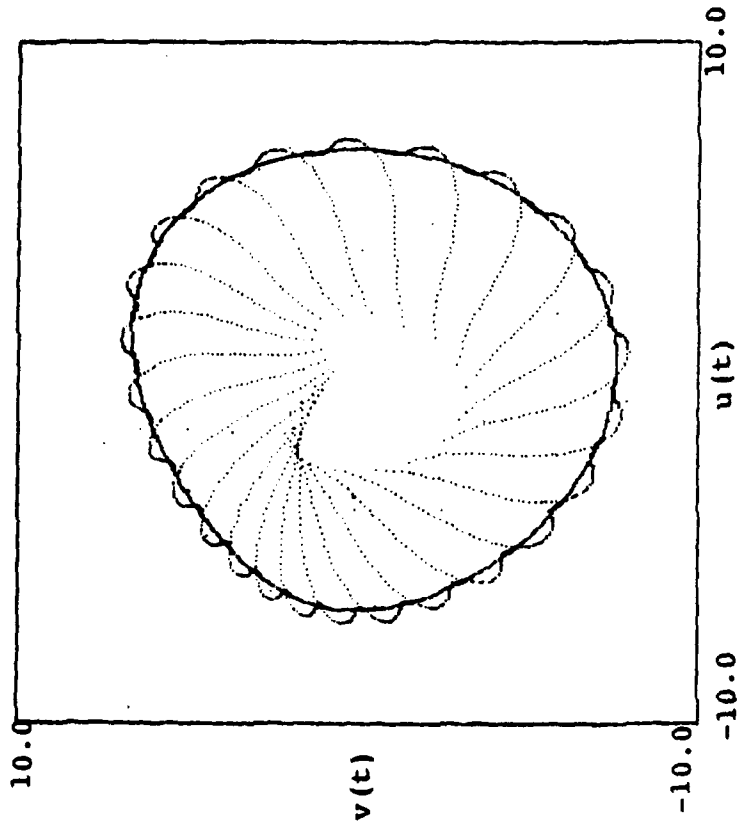


Figure II(d-2)

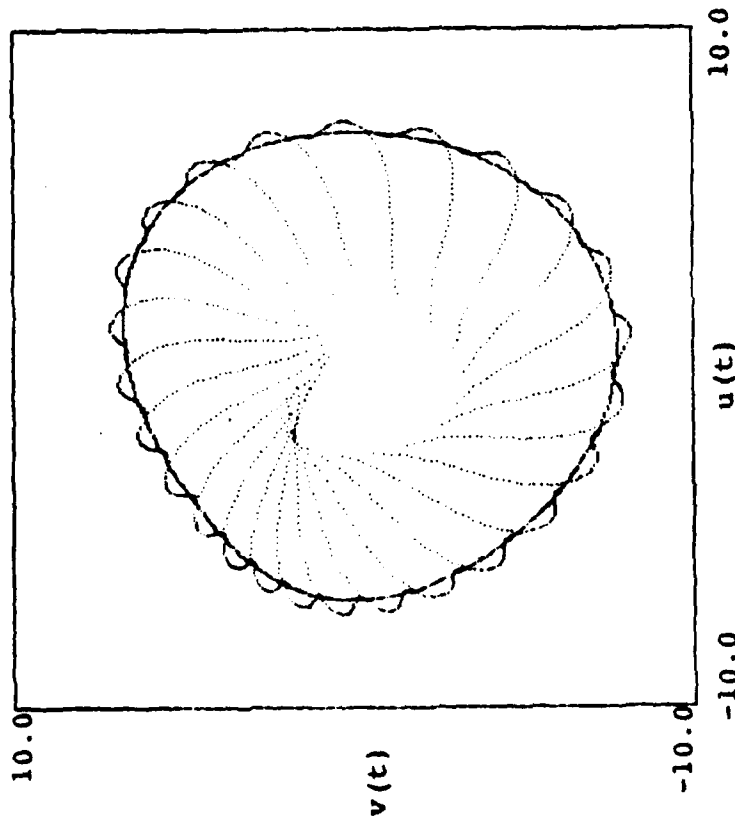


Figure II(d-1)



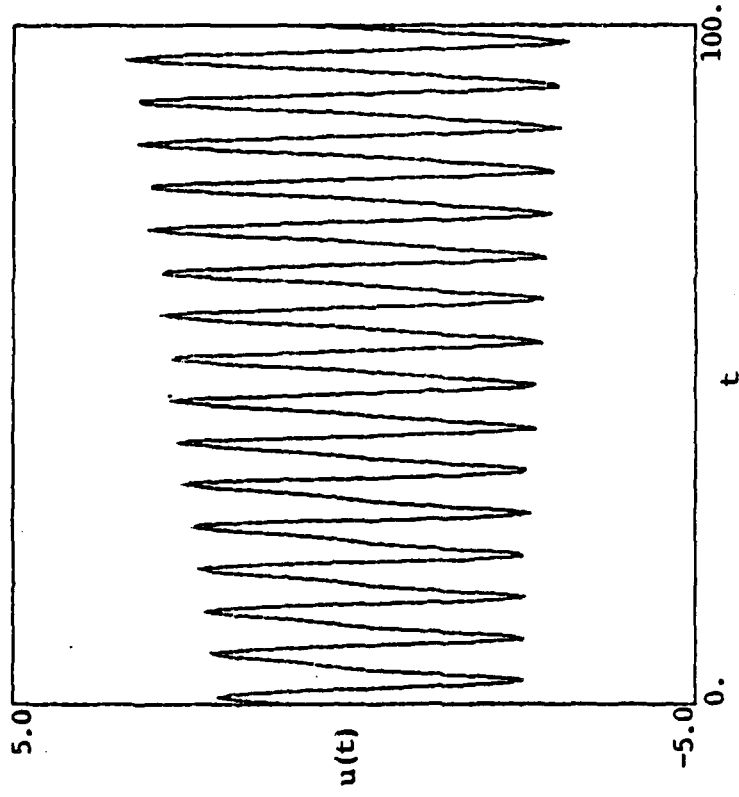


Figure II (e-2)

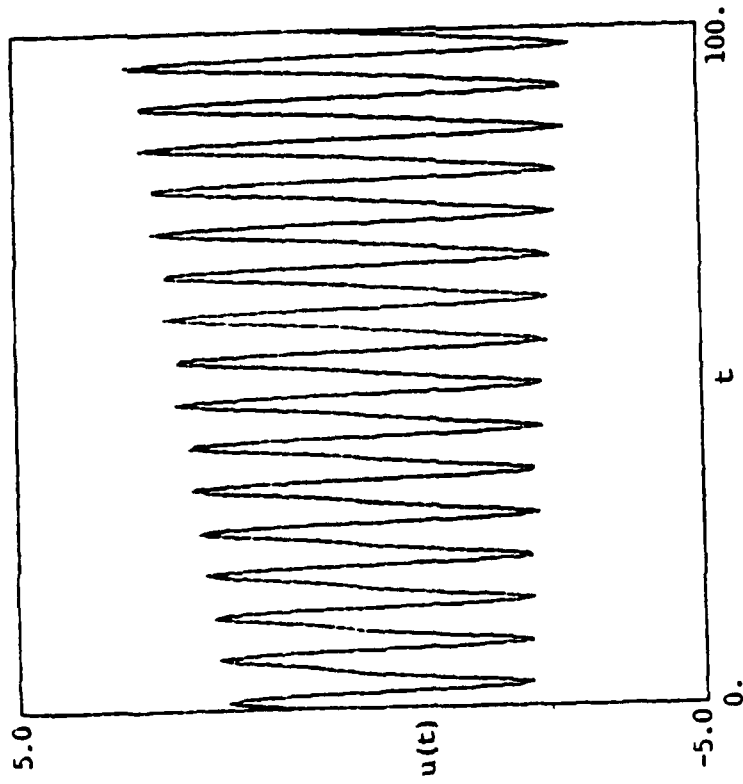


Figure II (e-1)

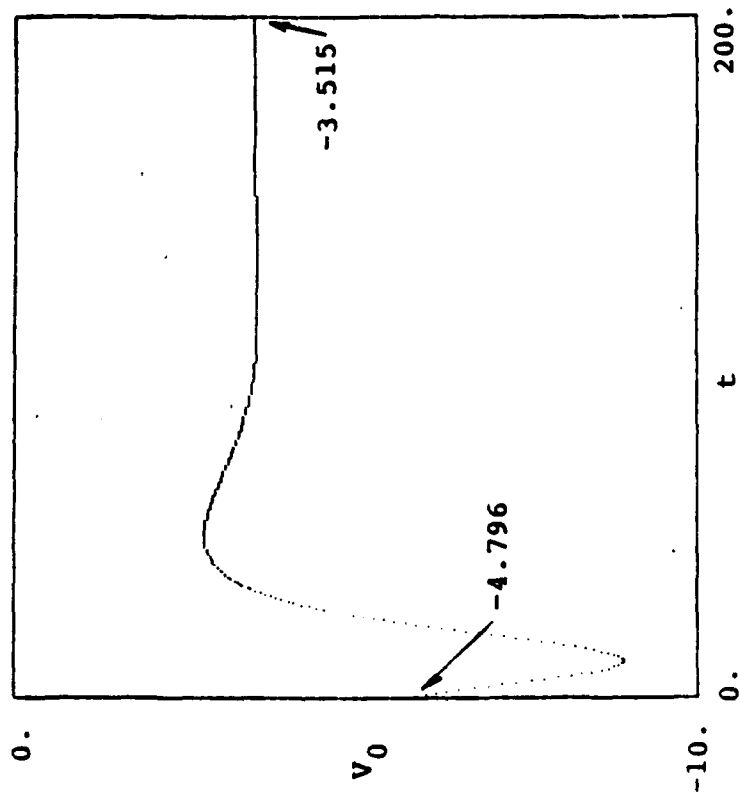


Figure III (a)

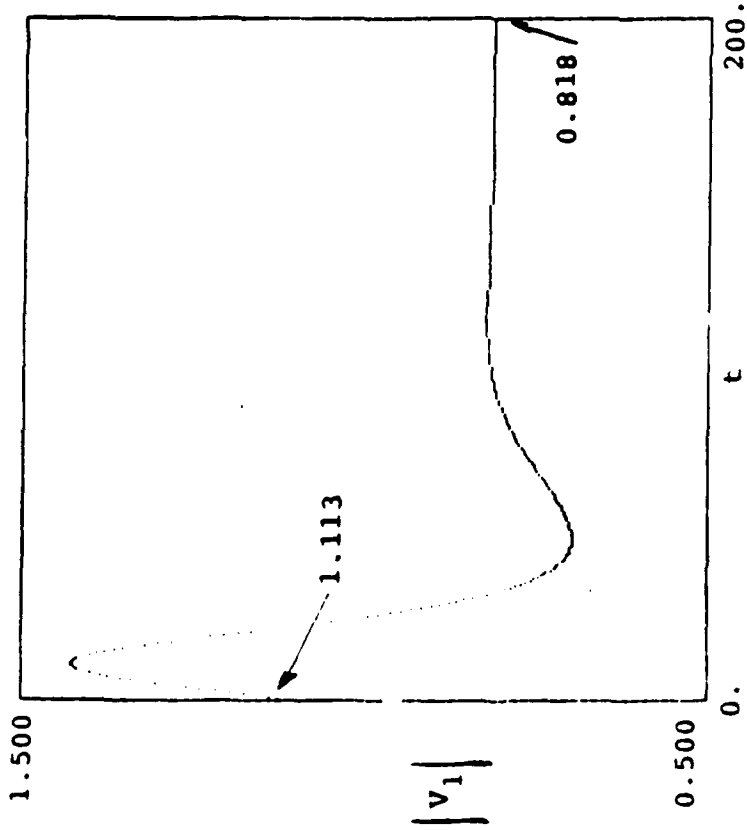


Figure III (b-1)

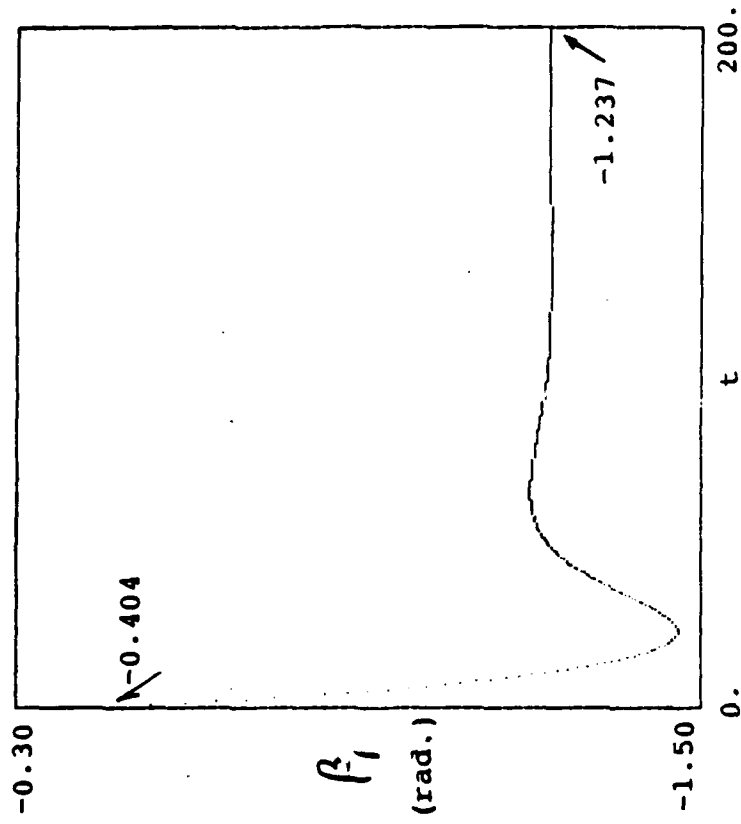


Figure III (b-2)

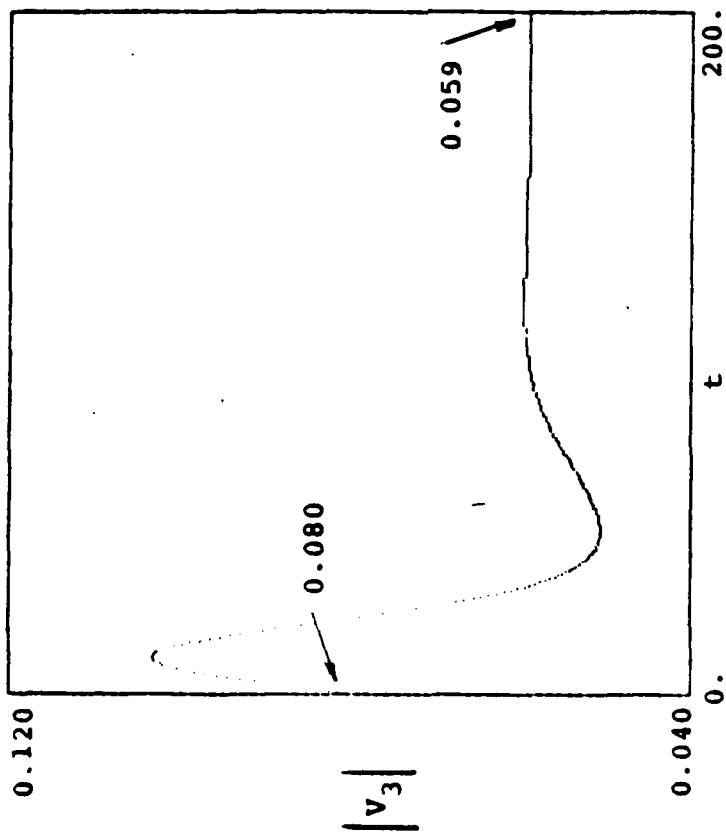


Figure III(c-1)

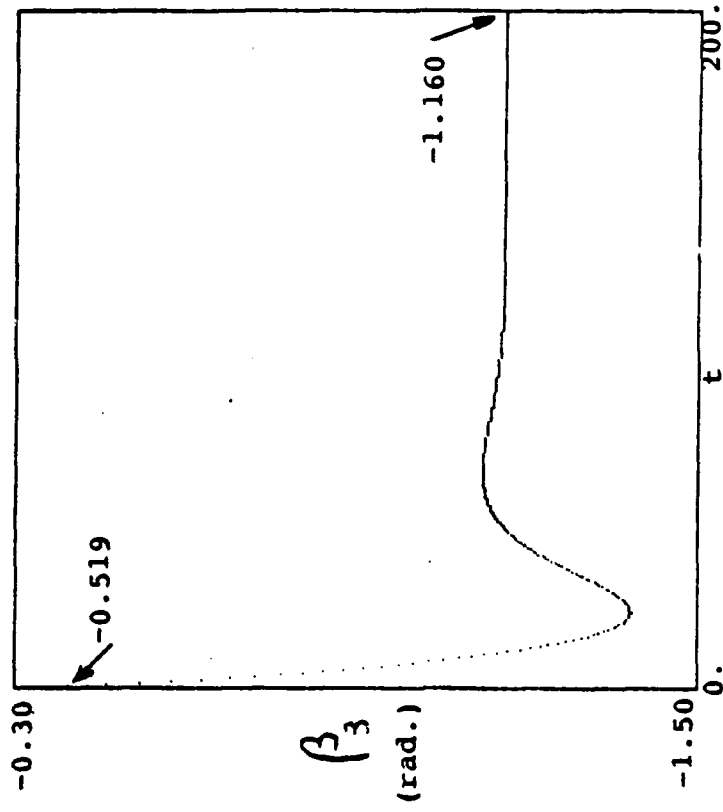


Figure III(c-2)

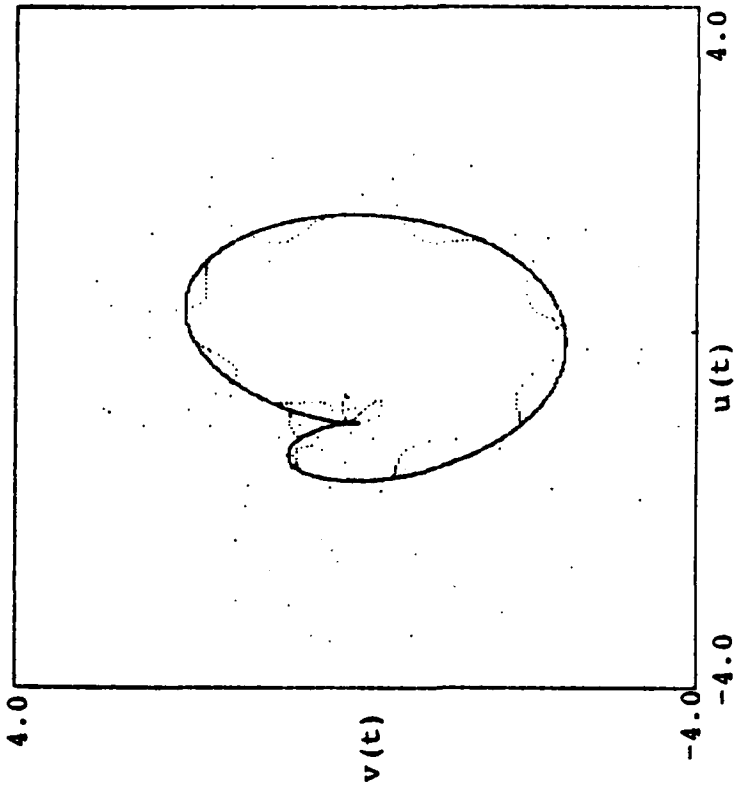


Figure III(d-2)

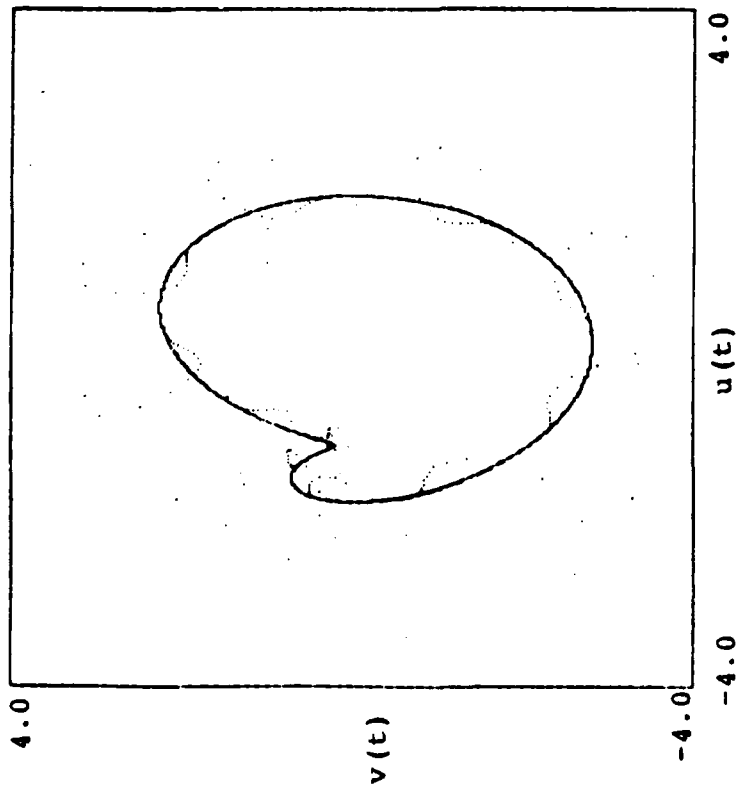


Figure III(d-1)

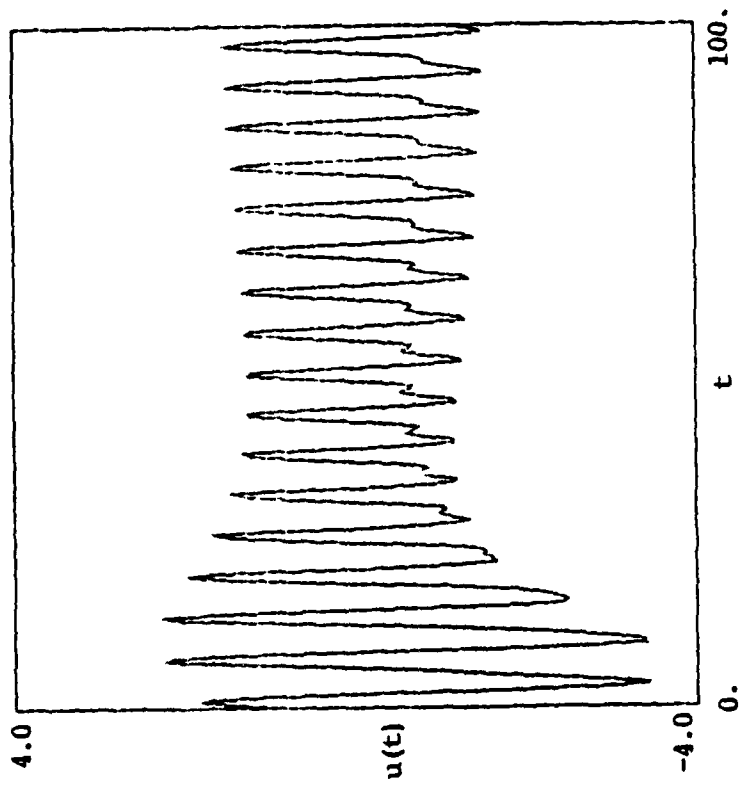


Figure III (e-2)

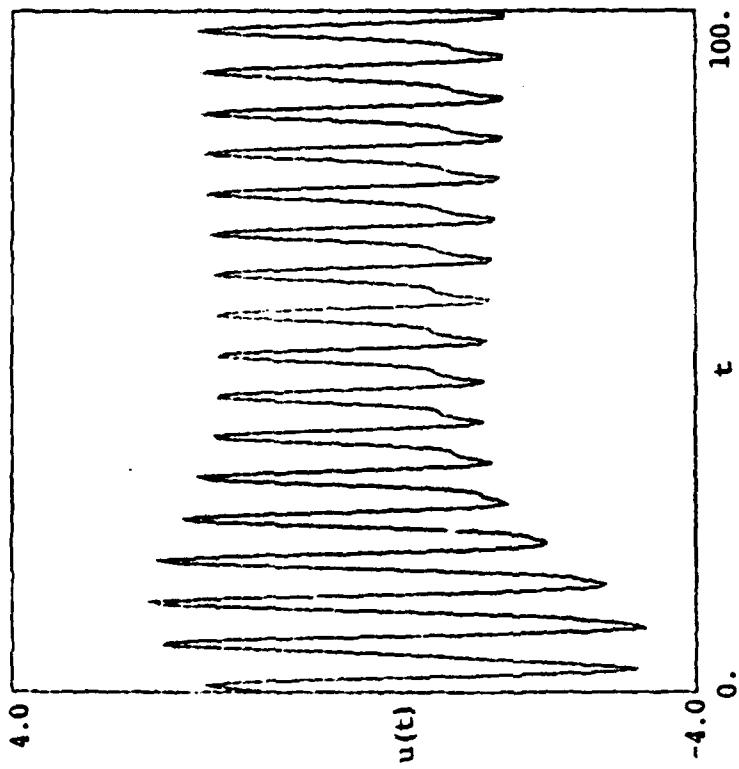


Figure III (e-1)

## Optimized Annulus-based Point-in-Region Inclusion Testing for d Dimensions

T. M. Cronin  
CECOM Center for Signals Warfare  
Warrenton, VA 22186-5100

**ABSTRACT.** Previous research into metrical inclusion testing for closed planar boundaries of general complexity resulted in a data structure called the inner annulus, which in this paper is shown to be the union of the set of unit normal vectors which point into the interior of a boundary. The annulus approach to the point-in-region problem circumvents the infinite precision requirement of the winding number approach, and also avoids the counting dilemma which has plagued a general implementation of the parity algorithm. The inner annulus, together with the boundary itself, serve as arguments to a function which compares distance from a query point. If the query point is nearer the annulus, the query point is inside the boundary; otherwise it is outside. Although the previous research presented the annulus as a declarative data structure, the resultant memory requirements were prohibitive for asymptotic boundaries. This paper presents an optimized algorithm which minimally encodes the inclusion information at each coordinate of a boundary. The inclusion information is independent of a query point, and position-insensitive to boundary translation. A preprocessing algorithm assures that the boundary is oriented in a counterclockwise fashion (so that by convention the interior is always to the left). The inward-pointing unit normal vector attached to a boundary element may be computed during preprocessing, or alternatively computed at run time with a procedural query. If preprocessed, it is shown that for each boundary coordinate, three bits are necessary and sufficient to represent the instruction to attach the vector; it is suggested that these three bits may be represented with an opcode at the coordinate itself. Hence, at run time, when the closest boundary element is computed, the opcode inclusion instruction is fetched and decoded along with it. This approach achieves a performance of  $O[\log n]$  query time for simultaneous closest point testing and metrical inclusion testing, with a storage requirement of  $O[n]$ , and preprocessing complexity  $O[n * \log n]$ . In the planar case, the storage constant multiplier of  $n$  is a negligible 1.09, using a standard word size of 32 bits. The technique is shown to be extensible to closed three dimensional surfaces, which are compositely defined as stacks of planar boundaries. The problem then becomes one of locating the nearest boundary in the stack, at which time planar logic is applicable. The final portion of the paper introduces an inductive argument to extend the technique into an arbitrary number of dimensions, and it is proven that the annulus attachment opcode consumes  $\log_2(3^d - 1)$  bits, where  $d$  is the number of dimensions.

## 1.0 INTRODUCTION.

The point inclusion problem (i.e., deciding whether point  $p$  is contained within boundary  $\beta$ ) is well-studied. One recent text contends that "The problem of locating a point in a subdivision of the plane or in a cell complex in a higher-dimensional space is one of the oldest and best understood problems in computational geometry"[E2]. Nevertheless, a fast deterministic algorithm has continued to evade a fully successful implementation. Although several elegant theoretical techniques are described in the literature, none has been successfully implemented for boundaries of general complexity. Previous attempts at fully successful implementations of inclusion testing have failed, due chiefly to one of two oversights: a) a digital computer is limited by finite precision arithmetic; b) the process of detecting boundary crossings is a non-trivial process.

### 1.1 Statement of the Problem.

Given a point and a closed digital boundary containing  $n$  coordinates, implement a deterministic, fast algorithm to discern whether or not the point is inside the boundary. By *deterministic*, it is meant that the solution is always correct, and not subject to round-off error due to finite precision arithmetic. By *fast*, it is meant that the technique's query time is a polynomial function of  $n$ , preferably convergent upon  $O[\log n]$ . In addition, the following problematic conditions must be accommodated by the inclusion testing process: 1) areal collapse due to low resolution of the digitizing process; 2) self-intersecting (non-simple) boundaries; 3) multiply-connected sets (Fig. 1).

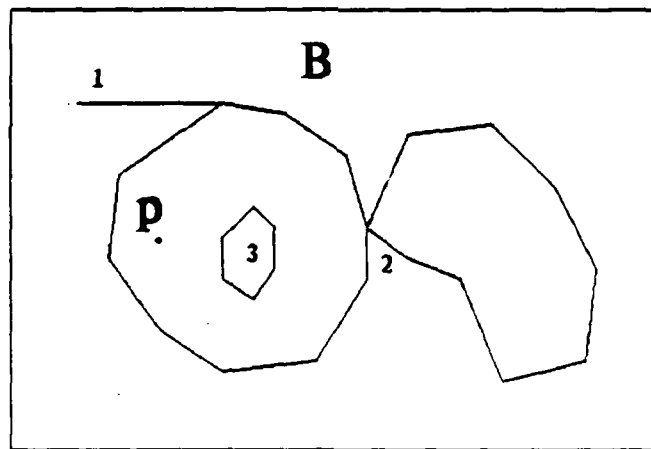


Figure 1. Planar boundaries may exhibit a variety of problematic conditions.

### 1.2 Previous Approaches to the Point Inclusion Problem.

It is not the intention of this paper to provide a historical perspective of the boundary inclusion testing problem, or for that matter, the point-in-polygon problem, as the planar case is called. Suffice it to say that for boundaries of general complexity, no fast deterministic implementation is documented in the literature. Three popular techniques are briefly discussed here; they are the *parity algorithm*, the *winding number*, and *refined triangulation*.

#### 1.2.1 The Parity Algorithm.

**Description.** The technique approaches the problem topologically, using the Jordan Curve Theorem. It proceeds by drawing a line from a query point through a boundary, while counting the



number of "crossings". The query point is inside the boundary if the count is odd; otherwise it is outside.

**Barriers to Implementation.** The technique may be deceived by degenerate tangent conditions which are perceived as crossings, and vice versa. Corrective measures such as vertex perturbation cause a prohibitive lag on algorithm performance, and still do not guarantee a deterministic decision [E1]. One researcher is dubious that a fully successful implementation can ever exist, due to the inherent sensitivity of line intersection algorithms to finite precision floating point arithmetic [F1].

### 1.2.2 The Winding Number Approach.

**Description.** The technique is analytic in nature. Based on Cauchy's Theorem, the integral of an analytic function about a query point is computed, and if zero, the point is judged to be outside the boundary; otherwise the integral must be a multiple of  $2\pi$ , and the point is judged to be inside [G1].

**Barriers to Implementation.** Roundoff error occasionally results in an incorrect inclusion decision, because a zero-sum integral, although theoretically possible, is not feasible on finite precision machines. Also, the technique exhibits inferior runtime complexity for two reasons: it uses floating-point trigonometric functions which are compute-intensive, and it must access each boundary element when accumulating the integral.

### 1.2.3 The Method of Refined Triangulation.

**Description.** This method, due to Kirkpatrick [K1], proceeds by triangulating a planar subdivision incrementally into bounded regions of finer granularity. A search is performed to determine if the query point resides in one of the triangulated subdivisions.

**Barriers to Implementation.** Although the algorithm is of  $O[\log n]$  time complexity, the question of whether the constants suffice for real time repetitive queries remains open [E2]. The best constants achieved to date are based on results obtained by analysis of the Four Color Theorem, and it is not clear that they facilitate a fast implementation.



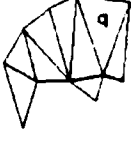
	Parity Algorithm	Winding Number	Refined Triangulation
			
Approach	Topological	Analytical	Graph-Theoretic
Theorem	Jordan Curve Theorem	Cauchy's Theorem	Four Color Theorem
Inside if	# crossings odd	$\sum \text{ang} \neq 0$	in a triangle
Complexity	Probabilistic	Probabilistic	Deterministic
Problems	Roundoff Error	Roundoff Error	Slow Search

Table 1. Point-in-polygon algorithms.

## 2. THE INNER ANNULUS: A PROXIMITY-BASED APPROACH TO INCLUSION TESTING.

In a digital domain, a boundary may be represented as a linked list of coordinates, with the head contiguous with the tail. If the boundary is oriented counterclockwise, then the left-handed limit of the boundary is on the interior. If a counterclockwise traversal of the list is performed, the set of discrete points to the left may be collected into another list called the *inner annulus* [C2]. Boundary inclusion testing for a query point is performed by comparing the distance to the nearest boundary point with that to the nearest annulus point; if the distance to the annulus is smaller, the query point is on the interior. Since the technique is metrical, it provides distance and direction to the boundary along with the inclusion decision. It is peculiar that inclusion testing thus reduces to a special kind of proximity testing.

### 2.1 Adopting a Convention to Assure a Unilateral Interior.

If the orientation of a boundary is assured prior to run time, an automated inclusion testing process can exploit knowledge of a unilateral interior during boundary traversal. A left-handed convention is adopted to achieve the search space reduction. By left-handed, it is meant that the boundary is oriented counterclockwise, to assure the interior is to the left. However, if the boundary is multiply-connected, the boundary of any hole it contains must be oriented in a clockwise direction, because the interior of the hole is outside the boundary [R1]. In Figure 2, a continuous boundary is represented by the solid line, and its inner annulus by the dashed line. The digital boundary is represented by black squares, and the annulus by white squares.

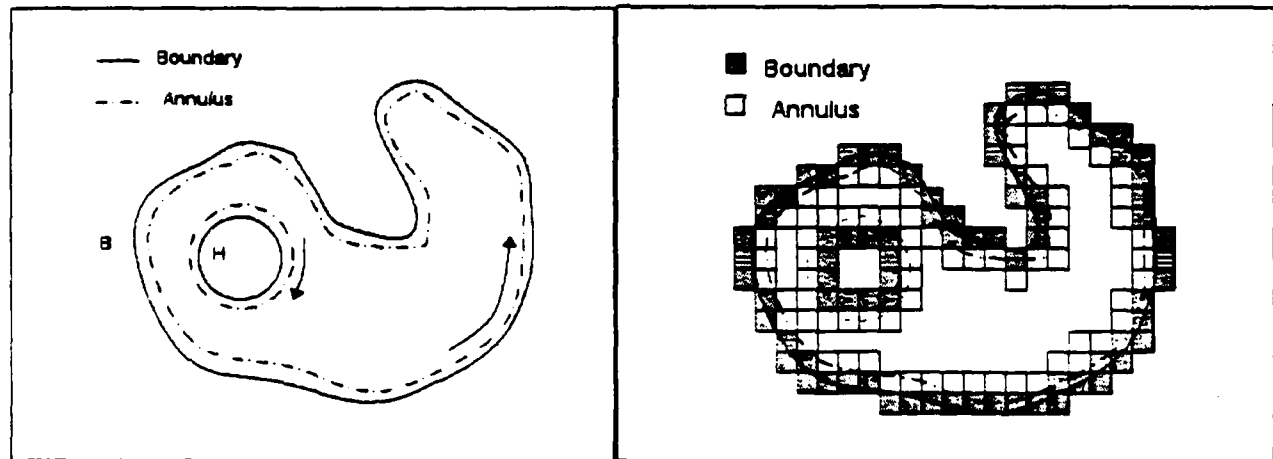


Figure 2. Continuous and digital versions of the inner annulus.

### 2.2 Automated Counterclockwise Orientation of a Digital Boundary.

An algorithm which automates the counterclockwise orientation of a boundary is published elsewhere [C3]. It is described only in passing here. The logic is as follows: a boundary's list of coordinates is searched sequentially and a coordinate with maximal abscissa is obtained, along with its predecessor and successor coordinates. The difference between the ordinates of the maximal-abscissa coordinate and its predecessor are computed, as well as the respective difference between the ordinates of the successor and maximal-abscissa coordinate. If either difference is less than zero, the boundary is oriented clockwise; otherwise it is counterclockwise. It is a simple matter to reverse the boundary list if the orientation is opposite that desired.

## 2.3 The Inner Annulus Technique Accommodates Problematic Boundary Conditions.

### 2.3.1 Multiply-connected Sets.

If a multiply-connected boundary contains a single hole, then the hole's boundary may be oriented in a clockwise fashion using the algorithm described above. This step is necessary because the interior of the outer boundary is actually exterior to the boundary of the hole.

If the hole itself contains another, the inner hole must be oriented counterclockwise, since its interior is also the interior of the outermost boundary. The general rule is as follows: let a boundary be oriented counterclockwise. Orient any hole it directly contains in a clockwise fashion. If this hole is multiply-connected, then any hole it contains must be oriented counterclockwise, etc. Continue until no multiple connectedness remains.

### 2.3.2 Counterclockwise Orientation of Non-Simple Boundaries.

A self-intersecting boundary is called *non-simple*. In Figure 3, the boundary on the left is non-simple. It may be oriented in a counterclockwise direction with the following retracing operation. Find a point with maximum abscissa and assure that its predecessor and successor are in counterclockwise order (if not then the boundary list must be reversed). Starting from that point, traverse the boundary in the direction of the successor, and proceed to the right at self-crossing areas. Continue collecting points until the predecessor is encountered. The points collected constitute a *simple* boundary, which is ordered counterclockwise. This linear-time algorithm may be invoked offline in a preprocessing step.

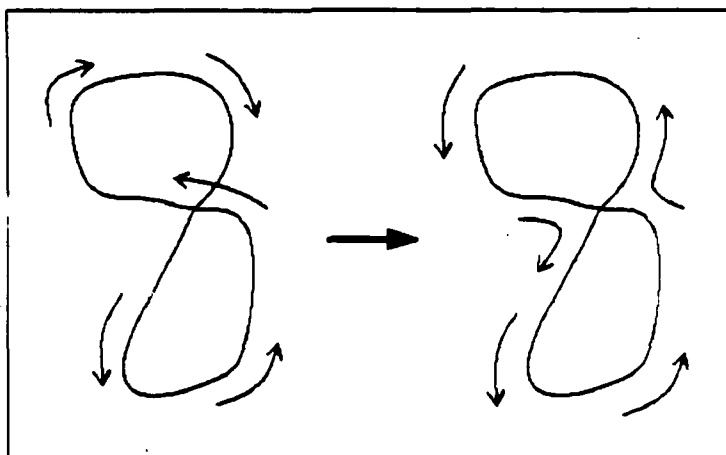


Figure 3. Retracing a non-simple boundary to obtain a simple one.

### 2.3.3 Areal Collapse due to Poor Digital Resolution.

If a boundary contains a region which possesses less width than the resolution of a digitizing process, the area is collapsed into a linear stub during digitization. In Figure 4, the closed boundary on the left exhibits a small convex region at its lower right. During digitization, resolution error causes the region to collapse (points 2-3). When the boundary is traversed in a counterclockwise direction, the ordered sequence of points  $\{0-1-2-3-3-2-1-4\}$  is visited. The problematic area may be detected with a preprocessing step which traverses the boundary in counterclockwise order, while looking for strings of duplicate boundary coordinates, where the second occurrence of the string is encountered in reverse order. Such duplicate coordinates of the boundary may be tagged by turning on a parity bit in the

upper portion of the word used to represent the coordinate. This concept is further developed in section 2.8.

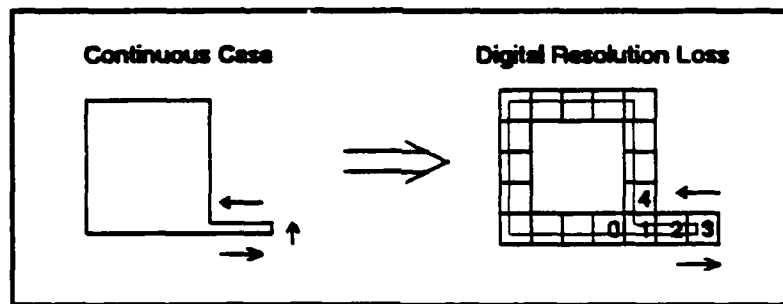


Figure 4. Area lost to digitization.

#### 2.4 The Relationship of the Inner Annulus of a Boundary to the Set of Normals to the Boundary.

A point in the plane may be connected diagonally (d-connected) or non-diagonally (4-connected) [R1]. Therefore, in a planar application, there are eight ways for an annulus point to be attached to a boundary point. This section demonstrates that the inner annulus is actually comprised of the set of unit normal vectors which point into the interior of a boundary.

##### 2.4.1 The Combinatorics of Local Boundary Behavior.

Since during preprocessing the annulus technique assures that a closed boundary does not self-intersect, the behavior of any local boundary section may be explicitly described. A boundary 3-tuple is a set of three counterclockwise-oriented boundary points called respectively the predecessor, the center, and the successor. The predecessor and center points of a 3-tuple may be connected in any of eight ways, whereas the center and successor may subsequently be connected in only five ways, producing a total of forty combinations. However, an annulus element can be attached to a point in the plane in only eight ways, since any planar point has exactly eight neighbors. Therefore, it is necessary to discover a many-to-one mapping which produces a range of eight states from a domain of forty. The mapping is obtained by observing the magnitudes of the differences between the abscissas and ordinates of contiguous elements of the 3-tuple.

An annulus element is actually a digital representation of the unit vector to the left of center. The logic which produces the unit vector is a function of three arguments: the predecessor, the center, and the successor coordinates. Since this 3-tuple is ordered counterclockwise, the inward-pointing unit vector is to the left, and is orthogonal to the direction of the 3-tuple. Note that if the order of the 3-tuple is reversed (i.e., changed to successor, center, predecessor), the same function produces a unit vector to the right. In fact, the function is utilized in this very manner to implement the algorithm for non-simple boundaries described above in section 2.3.2.

For elaboration of the annulus element / inside unit vector equivalence, refer to Figure 5. The black boxes represent boundary 3-tuple behaviors, and the associated white box is the annulus element produced for that specific behavior. The arrows on the far right are a legend which depict the direction of the predecessor to the center box, while the numbers in the far left column reflect the direction. For example, the up arrow represents the vector "01", which corresponds to: no change in x; increment y by 1. The numbers to the left of each icon depict the direction from the middle box to the successor. Referring to the upper left icon, the boundary's directional behavior is encoded by the string "0101", which represents two consecutive northerly directions. The numbers at the top of each icon are the

inner annulus attachment instruction. Therefore "-10" decodes to: "attach the annulus element by decrementing the x-value by 1 and leaving the y-value alone".

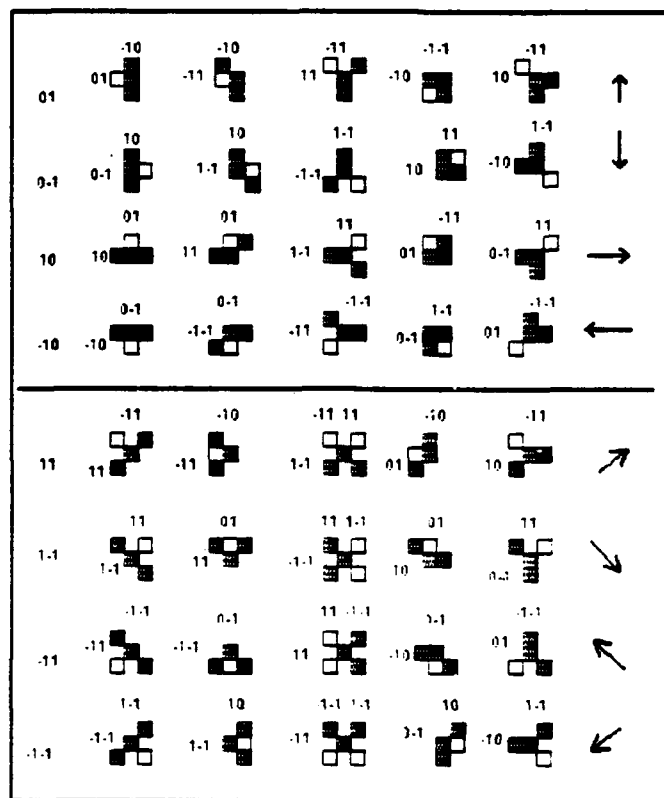


Figure 5. The annulus element is attached to the left of a boundary 3-tuple.

Eight of the forty behaviors are linear combinations of the other thirty-two. This set includes the fourth elements of the first four rows, and the third elements of the last four rows. These eight may be safely discarded because their conditions are duplicated by other behaviors. Thus we are required to develop a 32-to-8 mapping. We proceed in a backward-chaining fashion, retrospectively looking backwards from the eight output states to the various combinations of abscissa and ordinate differences which produce them.

Note that several behaviors produce the same output. For example, the instruction "-10" is produced by the first and second boundary behaviors in the first row, and by the second and fourth behaviors in the fifth row. Thus, four different boundary behaviors all generate the "-10" annulus attachment instruction, which dictates that the annulus element be attached at the left of center of the 3-tuple. These behaviors may therefore be combined into the system of conditional clauses represented in the table at the upper left of Figure 6. Note that the commonality for the conditional test lies in the fact that the ordinate differences are both equal to one, for all four behaviors. In this spirit we continue, and map the remaining twenty-eight behaviors of Figure 5 into the tables depicted in Figure 6. This explicit mapping constitutes the formal design specification for an algorithm, which we now develop in detail.

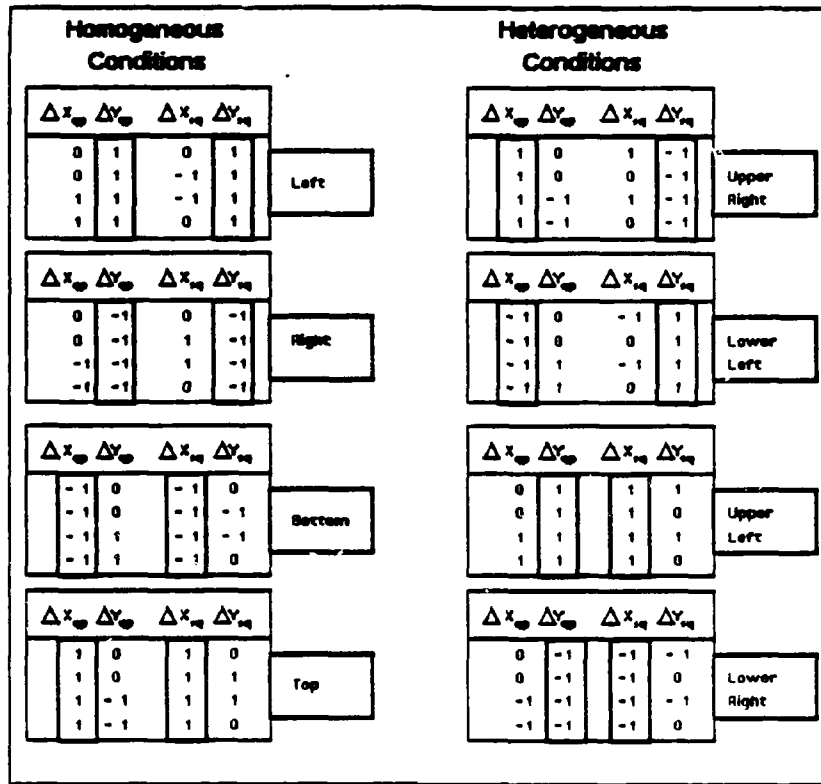


Figure 6. The compact mapping which relates boundary behavior to annulus vector attachment.

### 2.5 The Formal Design Specification for Computing the Inner Annulus Vector.

A query point may be considered as a query vector, drawn from the origin to the query point. When a normal vector  $N$  is drawn through a boundary from a query vector, it intersects the boundary first at some point  $p$ . But also normal to the boundary at  $p$  is some annulus element  $a$ , as specified at Figure 5. The vector drawn from the query point to the boundary is called the boundary vector, and the vector from the query point to the annulus element is called the annulus vector. The annulus vector is collinear with the boundary vector; both lie on the normal, and the magnitude of the annulus vector relative to the boundary vector may be used to perform an inclusion decision. This concept is formalized below.

Let  $r$  be a query vector and  $q=(x_q,y_q)$  be the boundary vector nearest to  $r$  on closed, counterclockwise-oriented boundary  $\beta$ . Let  $p(q) = (x_p,y_p)$  be the predecessor of  $q$  in  $\beta$ , and  $s(q) = (x_s,y_s)$  be the successor of  $q$  in  $\beta$ . Let  $\{i, j\}$  be a basis set of unit vectors as conventionally defined for the plane. Then the following logic provides the equations for computing the annulus vector  $a_p$ :

Let	$\Delta x_p = x_q - x_p;$	$\Delta y_p = y_q - y_p;$	$\Delta x_s = x_s - x_p;$	$\Delta y_s = y_s - y_p.$		
IF	c1.	$\Delta y_p = 1$	and	$\Delta x_s = 1$	THEN	$a_p = q - i + j$
ELSE IF	c2.	$\Delta y_p = -1$	and	$\Delta x_s = -1$	THEN	$a_p = q + i - j$
ELSE IF	c3.	$\Delta x_p = 1$	and	$\Delta y_s = -1$	THEN	$a_p = q + i + j$
ELSE IF	c4.	$\Delta x_p = -1$	and	$\Delta y_s = 1$	THEN	$a_p = q - i - j$

ELSE IF	c5.	$\Delta y_p = 1$	and	$\Delta y_s = 1$	THEN	$a_p = q - i$
ELSE IF	c6.	$\Delta y_p = -1$	and	$\Delta y_s = -1$	THEN	$a_p = q + i$
ELSE IF	c7.	$\Delta x_p = -1$	and	$\Delta x_s = -1$	THEN	$a_p = q - j$
ELSE IF	c8.	$\Delta x_p = 1$	and	$\Delta x_s = 1$	THEN	$a_p = q + j$

The ordering of the test is important. Note that conditions (c1) - (c4) are heterogeneous; i.e., a mix of abscissa and ordinate differences are involved. It is crucial that these conditional clauses be tested before homogeneous tests (c5) - (c8). This is because an injective mapping is not guaranteed unless the heterogeneous tests are triggered first. For example, note that the homogeneous test which generates an annulus element on the "right" specifies that respective ordinate differences be equal to -1. Suppose that this test was performed before any of the others. Referring back to Figure 6, notice that the heterogeneous tests for "upper right" and "upper left" contain two clauses which satisfy the condition for "right", which would result in an erroneous annulus attachment. Thus, the heterogeneous tests must be sequenced before the homogeneous tests to guarantee one-to-oneness.

## 2.6 The Distance to the Annulus Vector as a Measure of Inclusion.

The inner annulus technique relies on a comparison of proximity information to arrive at an inclusion decision. For computational efficiency, the distance metric implemented is the  $d_4$  distance, also variously known as the Manhattan distance, or the city-block distance [R1]. This distance metric avoids the multiplication and radical operations inherent to the Euclidean metric, and may be efficiently implemented with integer arithmetic.

**Definition 2.6.1** The  $d_4$  distance.

Let  $p = (x_1, y_1)$  and  $q = (x_2, y_2)$  be two points in the plane. Then the  $d_4$  distance from  $p$  to  $q$ , denoted  $d_4(p, q)$  is defined to be:

$$d_4(p, q) = |x_1 - x_2| + |y_1 - y_2|.$$

Because the  $d_4$  distance between two points equals the Euclidean distance only when either the respective ordinates or abscissas of the two points are themselves equal, one must take care to devise a proximity test which produces the same inclusion decision as the true Euclidean metric.

**Definition 2.6.2** Trichotomy of Metrical Inclusion.

Let  $q$  be a query vector,  $\beta$  a closed boundary, and  $p$  the nearest boundary vector to  $q$ . Let  $a_p$  be the annulus vector attached to  $p$ .

- Point  $q$  is said to be *on* boundary  $\beta$  if and only if  $d_4(q, p) = 0$
- else Point  $q$  is said to be *inside* boundary  $\beta$  if and only if  $d_4(q, a_p) \leq d_4(q, p)$
- else Point  $q$  is said to be *outside* boundary  $\beta$  if and only if  $d_4(q, a_p) > d_4(q, p)$ .

**Example.** Is the coordinate  $q = (50, 50)$  inside a boundary with closest point  $p = (100, 100)$ ; where the predecessor of  $p$  is  $(101, 99)$ , and the successor of  $p$  is  $(101, 101)$ ?

**Solution.**  $\Delta x_p = 100 - 101 = -1$ ;  $\Delta y_p = 100 - 99 = 1$ ;  $\Delta x_s = 101 - 100 = 1$ ;  $\Delta y_s = 101 - 100 = 1$ .

Since both  $\Delta y_p$  and  $\Delta x_s$  are equal to 1, clause c1 is satisfied, producing the annulus element  $a_p = (100 - 1, 100 + 1) = (99, 101)$ . The  $d_4$  distance from  $q$  to  $a_p$  is 100, as is the  $d_4$  distance from  $q$  to  $p$ . Since  $d_4(q, a_p) < d_4(q, p)$ ,  $q$  is inside the boundary.

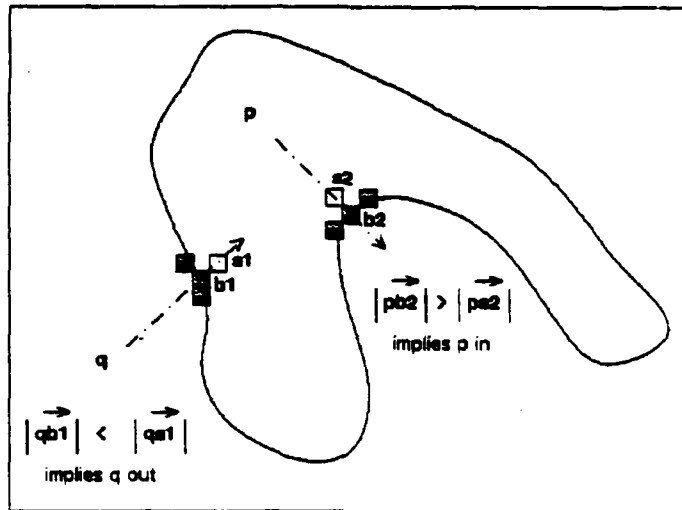


Figure 7. The magnitudes of the boundary and annulus vectors decide inclusion.

Theorem 2.6. The Boundary Vector and Corresponding Annulus Vector are Linearly Dependent.

Let  $q$  be a query point and  $\beta$  be the boundary of a closed simple curve. Let  $N = \overline{qp}$  be the normal vector drawn from  $q$  through the boundary, and let  $p$  be the point of intersection. Let  $T$  be the tangent through  $p$ , and  $a$  be the annulus element attached to  $p$ . Let the annulus vector be denoted by  $\overline{qa}$ , and let the boundary vector be denoted by  $\overline{qp}$ . Then  $\overline{qa}$  and  $\overline{qp}$  are linearly independent.

Proof:  $\overline{qp} \perp T$  by definition, and  $\overline{qa} \perp T$  by construction. Since in the plane there is only one line drawn through a point orthogonal to a given line,  $\overline{qp}$  and  $\overline{qa}$  are collinear. But collinear vectors are linearly dependent, which means that there exist  $\rho_1$  and  $\rho_2$ , not both zero, such that:

$$\rho_1 * \overline{qa} + \rho_2 * \overline{qp} = 0 \Rightarrow$$

$$\overline{qa} = -(\rho_2 / \rho_1) * \overline{qp}, \rho_1 \neq 0.$$

If  $q$  is on the interior of  $\beta$ , then by definition  $\|\overline{qa}\| < \|\overline{qp}\| \Rightarrow -(\rho_2 / \rho_1) < 1 \Rightarrow \rho_1 > -\rho_2$ . Conversely, if  $q$  is outside  $\beta$ , a similar argument may be used to show that  $\rho_1 < -\rho_2$ .

## 2.7 An Annulus Attachment Opcode.

In the last section it was demonstrated that in the plane there are eight ways in which to attach an annulus vector to a boundary vector. Since eight states may be minimally encoded with three bits, an annulus encoding algorithm is optimized if and only if it utilizes three bits to store the annulus attachment instruction. The figure below demonstrates a candidate opcode convention to store the instruction. For example, the opcode "111" decodes to the instruction "attach the annulus unit vector at the upper left of the coordinate". At query time, the opcode is used to compute the abscissa and



ordinate of the attached inner annulus vector. Respective distances from the query point to the nearest boundary point and the annulus vector are then compared to arrive at the inclusion decision.

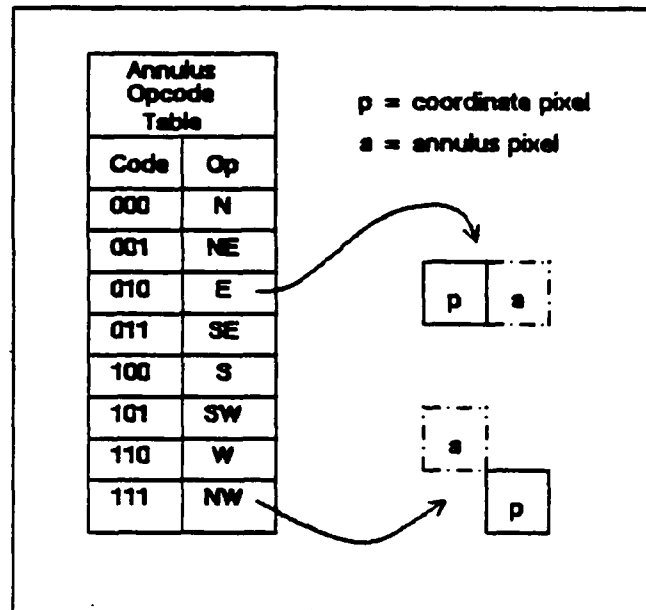


Figure 8. A three bit annulus opcode.

### 2.8 A Point-in-Polygon Computing Machine.

Since the planar annulus opcode consumes only three bits of information, it is feasible to embed it within a boundary coordinate, with a small corresponding loss in the number of coordinates expressible. One way to accomplish this within a computer word is depicted below. This scheme accommodates 16000 possible abscissa values and 16000 possible ordinate values, packed in a 32-bit word along with the annulus opcode.

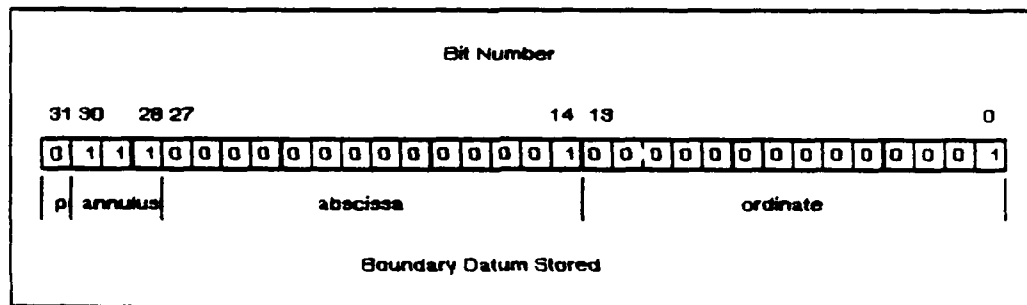


Figure 9. Packing the inclusion information into a coordinate word.

The embedding process may be performed during preprocessing with a single pass over the boundary. The ordinate of a boundary coordinate is stored in bits 0-13, and the abscissa in bits 14-27. The annulus opcode is precomputed and written into bits 28-30. Bit 31 is used to handle boundaries which have collapsed due to inferior resolution during the digitization process. A degeneracy exists at a boundary point if and only if bit 31 is set to 1, which indicates that the annulus element should be set equal to the boundary element at that point. In the example of Figure 9, the annulus opcode "111"

instructs that an inner annulus element be attached at the upper left of boundary coordinate (1, 1), at coordinate (0, 2).

At run time, masks may be used to decode the annulus opcode and coordinate information. In octal notation the annulus opcode mask is 1600000000. It is feasible to dedicate a decoding register in hardware for the unmasking process. This register could be coupled in a pipeline with another chip possessing arithmetic logic which computes the respective distances from the query point to the annulus element and the boundary element. The final pipeline stage might be a comparator register to render the minimum.

## 2.9 The Computational Complexity of the Planar Annulus Encoding Technique.

For domains with modest real world complexity, linear search is adequate to find the closest point (and encoded annulus element) to a query point. This is the case for most real world applications which are displayed on a single CRT screen. However, when boundaries become asymptotic, or when the map data becomes overwhelmingly dense, as on some topographic maps with extreme elevation changes, more efficient data structures and algorithms are recommended. Under such conditions, the author suggests the following approach.

*Preprocessing.*  $O[n * \log n]$ . The annulus technique begins by locating the nearest boundary point to a query point. The Voronoi diagram is an efficient representation scheme for proximity processing [E1, P1]. The Voronoi diagram for a set of  $n$  points can be constructed in preprocessing time  $O[n * \log n]$ .

*Storage.*  $O[n]$ . Since the annulus opcode utilizes three bits, the constant multiplier of  $n$  is  $(w + 3) / w$ , where  $w$  is the number of bits in the word used to encode a coordinate. In the plane, the storage requirement is  $1.09 * n$ , assuming a 32-bit word size, with both abscissa and ordinate packed into the same word. However, if the encoding scheme discussed in section 2.8 is utilized, the annulus opcode is stored in the same word as a coordinate, which reduces the storage requirement to exactly  $n$ . Of course, in this case, there is a corresponding loss in the number of coordinates expressible. Further improvements to achieve superlinear storage could be made if one elected to represent a boundary with a polynomial or polygonal approximation. The annulus-based approach readily accommodates data compression schemes, and the query time would improve due to a smaller search space; the price paid is the error introduced by the approximation scheme.

*Query Time.*  $O[\log n]$ . The closest boundary point to a query point can be obtained in  $O[\log n]$  time, using the preconstructed Voronoi diagram. Simultaneously, the annulus attachment opcode is fetched along with it, packed in the upper portion of the coordinate. Negligible constant time is required to compare the two distances from the query point. Thus the query time complexity is  $O[\log n]$ .

## 3. AN EXTENSION TO THREE DIMENSIONS.

The extension of the planar annulus technique to three dimensions is straightforward. A three-dimensional object may be conceptualized as a stack of planar boundaries, each one pixel in height. Testing for inclusion within the solid is equivalent to locating the (nearest point on) the nearest planar boundary, along with the annulus element attached to it; the respective distances are then compared to arrive at the inclusion decision.

Figure 10 illustrates the annulus technique for a sphere and another object modeled as a stack of planar boundaries. In the case of the sphere, query point  $p$  is nearest to some point on the equator of  $S$ . But the equator is a planar object, so it has an inner annulus. The planar annulus logic is applied to arrive at an inclusion decision. For the modeled object, query point  $p$  is nearest to  $q$ , which is an element of the highest boundary in the stack. But  $q$  is attached to annulus element  $a$ , which has been

precomputed as a function of local boundary behavior about  $q$ . Since the magnitude of  $q$  is less than that of  $a$ , it is decided that point  $p$  is on the exterior.

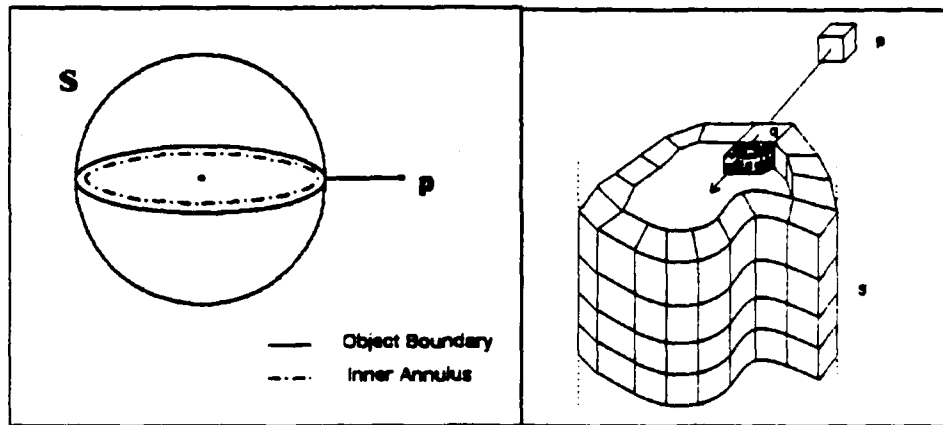


Figure 10. The extension of the annulus technique to three dimensional objects.

### 3.1 The Complexity of the Annulus Technique in Three Dimensions.

The closest point of a planar boundary to a query point can be found in  $O[\log m]$  query time, where  $m$  is the length of the boundary. This operation must be performed  $p$  times, where  $p$  is the number of boundaries contained in the stack which comprise the solid. Therefore, the point inclusion time complexity for a three-dimensional solid is  $\log m_1 + \log m_2 + \dots + \log m_p =$

$$\log_2 \prod_{i=1}^p m_i$$

The optimal time complexity is an open research issue. If a three-dimensional solid is represented as a cell bounded by a complex of  $n$  intersecting planes, then it has been shown independently by two researchers [C1, E3] that the three-dimensional point inclusion problem can be solved in  $O[\log^2 n]$  query time, with a storage requirement of  $O[n^3]$ .

## 4.0 GENERALIZATION TO HIGHER DIMENSIONS.

In this section an attempt is made to generalize the annulus-based inclusion testing technique to an arbitrary number of dimensions. Since the technique seeks to attach an annulus vector to the boundary point nearest a query point, it is necessary to know how many neighbors a boundary point possesses in  $d$  dimensions, to derive the number of bits required to encode the annulus attachment instruction. With this goal in mind, we proceed to develop a set of five axioms which describe a methodology to inductively construct a new dimension from a previous one. The first four axioms closely parallel the Peano axioms for the natural numbers. The fifth axiom diverges from Peano when we postulate the construction of a new dimension.

### 4.1 Axioms of Inductive Dimensionality.

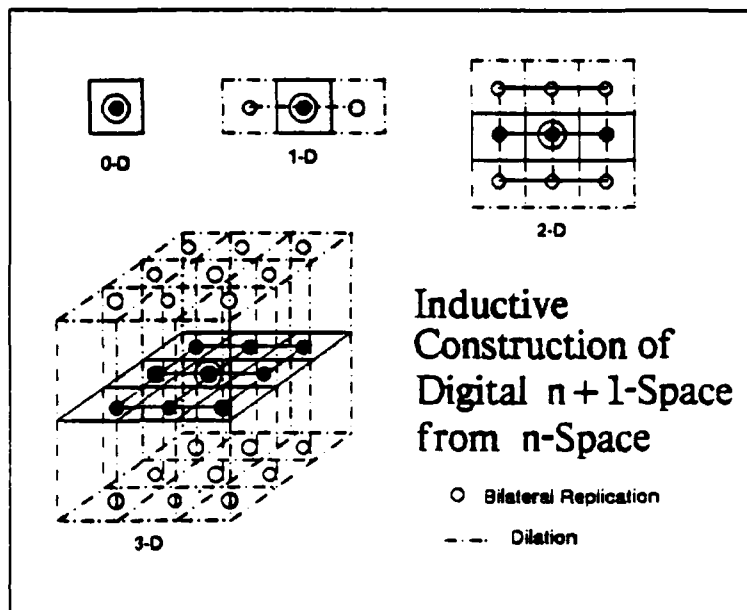
Axiom D1. A point has dimension 0, and is without axis.

Axiom D2. Every dimension  $D$  has a unique successor dimension  $D + 1$ , which has a unique axis.  $D$  is said to be the predecessor dimension of  $D + 1$ .

**Axiom D3.** A point is not a successor dimension of any dimension.

**Axiom D4.** Distinct dimensions have distinct successor dimensions.

**Axiom D5. Bilateral Dilation.** Dimension  $D + 1$  is constructed by interposing a hyperplane from dimension  $D$  between two other hyperplanes from  $D$ , and dilating the structure along the axis of  $D + 1$ .



**Figure 11. Building higher dimensions.**

The fifth axiom permits us to perform two operations (replication and dilation) on an object in a lower dimension to produce an object in a new dimension. Figure 11 illustrates successive applications of the operations to produce the first dimension from a point source; the plane from a line source; and a cube from a planar source. We can conceive of the operation of bilateral replication in the third dimension to get to the fourth (a cube is interposed between two others), but we cannot visualize the axis along which to dilate the composite, because our spatial world is restricted to three dimensions. However, if we subscribe to the axioms, we derive the following results.

**4.2 Neighbor Theorem (number of Digital Neighbors in  $d$ -Space).** In  $d$ -space, the number of digital points neighboring a reference point is  $3^d - 1$ .

**Proof (induction):**

**Step 1.** If  $d = 1$ , the space is linear, and the number of neighbors to a point is  $3 - 1 = 2$ .

**Step 2.** Assume that in  $k$ -space, the number of neighbors of a reference point is  $s_k = 3^k - 1$ .

**Step 3.** Prove that in  $(k + 1)$ -space, the number of neighbors  $s_{(k + 1)} = 3^{(k + 1)} - 1$ .

Since  $s_k = 3^k - 1$ , then  $3^k$  is the sum total of the reference point plus its neighbors. Axiom D5 (bilateral dilation) applied to this set creates  $3 \cdot 3^k$  points in dimension  $k + 1$ , of which one is a reference point and the remaining  $3 \cdot 3^k - 1$  are its neighbors. But  $3 \cdot 3^k - 1 = 3(k + 1) - 1$ . QED.

4.3 Inner Annulus Opcode Storage Theorem (bit length of the annulus opcode in  $d$  dimensions). The inner annulus opcode to effect an inclusion decision concerning a query point in  $d$  dimensions may be encoded in  $\log_2[3^d - 1]$  bits.

Proof. Theorem 4.2 asserts that a reference point in  $d$ -space has  $3^d - 1$  neighbors. But these are precisely the number of ways in which an inner annulus element may be attached to a reference point. The number of bits necessary to encode  $3^d - 1$  digital neighbors is  $\log_2[3^d - 1]$ . QED.

4.4 Opcode Storage Approximation Corollary (annulus opcode bit length approximation). The number of bits necessary to encode the inner annulus opcode in  $d$  dimensions is bounded from above by  $1.6d$ .

Proof. From Theorem 4.3, the number of bits is exactly  $\log_2[3^d - 1]$ .

But  $\log_2[3^d - 1] < \log_2[3^d] = d \cdot \log_2 3 < 1.6d$ . QED.

4.5 The Storage Requirement of the Annulus Encoding Technique in Higher Dimensions.

Given a word length of  $w$  bits, with  $\kappa$  bits used to encode the annulus opcode. The storage requirement is  $n + (\kappa/w) \cdot n = ((w + \kappa)/w) \cdot n$ , which is clearly  $O[n]$ . Application of Theorems 4.2 and 4.3 produce Table 2, which depicts the growth of the annulus opcode and storage constant in higher dimensions.

$d = \text{dimension}$	$n = \text{neighbors}$	$\kappa = \text{opcode bits}$	$\lambda = \text{storage inclusion constant}$
1 *	2	1.00	1.03
2	8	3.00	1.09
3	26	4.63	1.14
4	80	6.24	1.20
5	242	7.89	1.25
6	728	9.42	1.29
...	...	...	...
$d$	$3^d - 1$	$m - 1 + (3^d - 1) / [2^m]$ **	$(w + \kappa) / w$

\* Inclusion is not an issue in one dimension. However, since the annulus is by convention to the left of a boundary, the technique is useful for deciding upon which side of a line a query point lies.

\*\*  $[2^m]$  is a step function, which is defined by the minimum  $m$  such that  $2^m > 3^d - 1$ . But  $2^m > 3^d - 1 \Leftrightarrow m = \min \{ i \in \mathbb{I}; i > \log(3^d - 1) / \log 2 \}$ .

Table 2. The number of annulus opcode bits required to decide inclusion in higher dimensions.

## 5.0 CONCLUSIONS.

An optimized point-in-polygon algorithm with a negligible query time constant multiplier of  $\log n$  has been presented for planar boundaries. As a bonus, the magnitude and direction of the normal vector from the query point are returned along with the inclusion decision. The algorithm is based on a topological structure called the inner annulus, which is demonstrated to be the union of the set of unit normal vectors which point into the interior of a boundary. The algorithm operates by comparing the respective distances of the query point from the boundary and annulus vectors; a smaller magnitude for the annulus vector implies inclusion. The technique pointedly circumvents the finite precision problems which plague implementation of other point-in-polygon algorithms such as the parity algorithm and the winding number approach. It has been shown that for a planar subdivision, an opcode to attach the annulus vector to a boundary coordinate may be precomputed and encoded in the upper three bits of the coordinate. Hence, at run time, when the closest point on the boundary is computed, the instruction to compute the annulus element is fetched along with it. It has been demonstrated that the technique is extensible to higher dimensional objects. An axiomatic treatment of an inclusion decision in  $d$  dimensions has been presented, and an inductive argument has demonstrated that the number of opcode bits required to store the inclusion information is precisely  $\log_2(3^d - 1)$  bits.

## ACKNOWLEDGMENTS

The research benefited from discussions with the following individuals: Gerald Andersen, Richard Antony, Christopher Bogart, Nevin Bryant, Jagdish Chandra, Douglas Chubb, Herbert Edelsbrunner, Ray Freeman, Thomas Garvey, Thomas Logan, James Mulligan, John Robertson, Azriel Rosenfeld, Robert Sedgewick, and Nick Sizemore. Special thanks to Robert Connelly for assistance in proving the truth of a conjecture asserting the length of the inner annulus, and to Cathy Lamanna for an efficient ADA implementation of the optimized planar inclusion technique.

## BIBLIOGRAPHY

- [C1] Chazelle, B. "How to Search in History", *Information and Control* 64, 1985.
- [C2] Cronin, T.M. "Annulus-based Inclusion Testing for Multiply-Connected Sets", *Transactions of the Sixth Army Conference on Applied Mathematics and Computing*, ARO Report 89-1, Research Triangle Park NC, February 1989.
- [C3] Cronin, T.M. "Allocating Sensor Envelope Patterns to a Map Partitioned by Territorial Contours", *Proceedings of the US Army Symposium on Artificial Intelligence Research for Exploitation of the Battlefield Environment*, El Paso TX, November 1988.
- [E1] Edelsbrunner, H. and E. P. Mücke, "Simulation of Simplicity: A Technique to Cope with Degenerate Cases in Geometric Algorithms", *Proceedings of the Fourth Annual Symposium on Computational Geometry*, Urbana IL, June 1988.
- [E2] Edelsbrunner, H. Algorithms in Combinatorial Geometry, Springer-Verlag, Berlin Germany, 1987.
- [E3] Edelsbrunner, H., J. O'Rourke, and R. Seidel. "Constructing Arrangements of Lines and Hyperplanes with Applications", *SIAM Journal Comput.* 15, 1986.

[F1] Forrest, A.R. "Computational Geometry in Practice", in Fundamental Algorithms for Computer Graphics, ed. R.A. Earnshaw, Springer-Verlag, Berlin Germany, 1985.

[F2] Fulks, W. Advanced Calculus, Second Edition, John Wiley & Sons, New York NY, 1969.

[G1] Griesel, A., and J. Gillis, Ed. "A Collection of Area of Interest Algorithms", Jet Propulsion Laboratory Technical Report JPL D-171, Pasadena CA, July 1985.

[K1] Kirkpatrick, D.G. "Optimal Search in Planar Subdivisions", *SIAM Journal Comput.* 12, 1983.

[P1] Preparata, F. P., and M. I. Shamos. Computational Geometry: An Introduction. Springer-Verlag, New York NY, 1985.

[R1] Rosenfeld, A., and A.C. Kak. Digital Picture Processing, Vol. 2, Second Edition, Academic Press Inc., Orlando FL, 1982.

[S1] Sedgewick, R. Algorithms, Addison-Wesley, Reading MA, 1983.

[S2] Surany, A.P. "A Simple Algorithm for Determining whether a Point Resides within an Arbitrarily Shaped Polygon", in Fundamental Algorithms for Computer Graphics, ed. R.A. Earnshaw, Springer-Verlag, Berlin Germany, 1985.

LIST OF CONFERENCE ATTENDEES

SEVENTH ANNUAL ARMY CONFERENCE ON APPLIED MATHEMATICS AND  
COMPUTING

6-9 JUNE 1989

NAME

Gerald R. Andersen  
Army Research Office  
Research Triangle Park, NC 27709

Dr. Joseph Arkin  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

LTC Chris Arney  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

Romesh C. Batra  
Department of Mechanical Engineering  
University of Missouri-Rolla  
Rolla, MO 65401-0249

CPT Kevin Beam  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

Paul Broome  
Ballistic Research Laboratory  
Aberdeen Proving Ground, MD 21005

David A. Caughey  
Cornell University  
Ithaca, NY 14853

Dr. Jagdish Chandra  
Army Research Office  
P.O. Box 12211  
Research Triangle Park, NC 27709-2211



Peter C. T. Chen  
Watervliet Arsenal  
Watervliet, NY 12189-4050

Charles K. Chui  
Department of Mathematics  
Texas A&M University  
College Station, TX 77843

Michael A. Cipollo  
Benet Labs  
Watervliet Arsenal  
Watervliet, NY 12189

A. Cohen  
University of Paris-Dauphine

Norman P. Coleman, Jr.  
HQ, ARDEC  
Picatinny Arsenal, NJ

Terence M. Cronin  
US Army CECOM Center for Signals Warfare  
ATTN: AMSEL-RD-SE-TRI  
Vint Hill Farms Station  
Warrenton, MA 22186

Mark H. A. Davis  
Imperial College of Science & Technology  
London, England

Paul Davis  
Worcester Polytechnic Institute  
Worcester, MA

LTC Lee S. Dewald  
Department of Mathematics  
West Point, NY 10996

Donald A. Drew  
207 Worth St.  
Ithaca, NY 14850

Kie-Bum Eom  
Department of Electrical & Computer  
Engineering  
Syracuse University  
Syracuse, NY 13041

Aaron Das Gupta  
USA Ballistic Research Lab  
Aberdeen Proving Ground

Carl de Boer  
CMS  
610 Walnut  
Madison, WI 53205

Howard Elman  
Department of Computer Science  
University of Maryland  
College Park, MD 20742

Joseph E. Flaherty  
Department of Computer Science  
RPI  
Troy, NY 12180

COL Frank Giordano  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

James Glimm  
Department of Applied Mathematics  
SUNY at Stony Brook  
Stony Brook, NY

John W. Grove  
Department of Applied Math & Statistics  
SUNY at Stony Brook  
Stony Brook, NY 11794

Morton E. Gurtin  
Department of Mathematics  
Carnegie Mellon University  
Pittsburgh, PA 15213

William Hager  
Department of Mathematics  
University of Florida

Charles E. Hall, Sr.  
Research Directorate, RDEC  
US Army Missile Command

S. Hanagud  
Georgia Institute of Technology  
Atlanta, GA

Christian G. Hempel  
Cornell University  
Ithaca, NY 14850

M. A. Hussain  
General Electric  
PO Box 8  
Schenectady, NY 12301

William Jackson  
TACOM

Richard James  
Dept of Aerospace Engineering and  
Mechanics  
107 Ackerman Hall  
University of Minnesota  
Mpls, MN 55455

Arthur Johnson  
US Army Materials Tech. Lab  
Watertown, MA

Lennart Johnsson  
Yale University  
New Haven, Connecticut

R. L. Kashyap  
Department of Electrical Engineering  
Purdue University, IN

Robert V. Kohn  
Courant Institute  
New York University

MAJ Gary Krahn  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

Abraham Kribus  
M & AE  
Cornell University  
Ithaca, NY 14853

Dimitris Lagoudas  
Department of Civil Engineering  
RPI  
Troy, NY 12180

Siegfried H. Lehnigk  
U.S. Army Missile Command  
Redstone Arsenal, AL 35898-5248

Anatoly S. Libgaber  
Department of Mathematics  
University of Illinois at Chicago  
Box 4348  
Chicago, IL 60680

James Lipton  
Department of Mathematics  
Cornell University  
Ithaca, NY 14853

Ray Ludwig  
Department of Computer Science  
RPI  
Troy, NY

Alex Mahalov  
Center for Applied Mathematics  
Cornell University  
Ithaca, NY 14853

MAJ John McMahon  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

Roger Multer  
US Army Corps of Engineer Waterway  
Experiment Station  
Vicksburg, Mississippi 39180

CPT Joseph Myers  
Department of Mathematics  
United States Military Academy  
West Point, NY 10996

Anil Nerode  
Cornell University  
Ithaca, NY 14853

John A. Nohel  
Center for the Mathematical Sciences  
University of Wisconsin-Madison, Wisconsin

J. Tinsley Oden  
ASE/EM Department  
University of Texas at Austin  
Austin, TX 78712

Martin Otter  
Deutsche Forschungs u. Versuchsanstalt  
für Luft- und Raumfahrt e. V.  
Institut für Dynamik der Flugsysteme  
Oberpfaffenhofen  
8031 Wessling/Obb. FRG

Tom Pence  
Dept of Metallurgy, Mechanics,  
& Material Science  
Michigan State University, Michigan

Louis Piscitelle  
US Army Natick RD&E Center  
Natick, MA

Claudia Quigley  
Army Materials Technology  
Watertown, MA 02172-0001

Phoebus Rosakis  
Department of Theoretical & Applied  
Mechanics  
Cornell University  
Thurston Hall  
Ithaca, NY 14853

Rouben Rostamian  
University of Maryland

Mike Sayers  
University of Michigan  
Ann Arbor, Michigan

Ray D. Scanlon  
Watervliet Arsenal  
Watervliet, NY 12189

Ahmed A. Shabana  
Department of Mechanical Engineering  
P. O. Box 4348  
University of Illinois in Chicago  
Chicago, Illinois 60680

L. S. Shieh  
Department of Electrical Engineering  
University of Houston

B. D. Sivazlian  
Industrial Systems Engineering  
University of Florida  
Gainesville, Florida

Royce Soanes  
Watervliet Arsenal

Ram P. Srivastav  
State University of New York  
Stoney Brook, NY 11794-3600

Mike Steele  
Princeton University

Iradj G. Tadjbakhsh  
Department of Civil Engineering  
RPI  
Troy, NY 12181

B. A. Taylor  
Mathematics Department  
University of Michigan  
Ann Arbor, MI 48109

Lee Taylor  
1838 Cedar Drive  
Severn, Maryland 21144

Thomas C. T. Ting  
CEMM Department (M/C 246)  
University of Illinois  
Box 4348  
Chicago, IL 60680

Dennis M. Tracey  
Army Materials Technology Laboratory  
Watertown, MA 02172

Victor Trutzer  
Department of Mathematics  
University of Lowell  
One University Avenue  
Lowell, MA 01854

Athanasios Tzavaras  
Department of Mathematics  
University of Wisconsin, Madison WI

CPT Robert B. Underwood  
USAEWES ATTN: CEWES-GM  
Vicksburg, Mississippi 39181-0631

John D. Vasilakis  
Benet Laboratories  
SMCAR-CCB-RA (BLDG 115)  
Watervliet, NY 122189-4050

John Walter  
Ballistic Research Laboratory  
Aberdeen Proving Ground, MD 21005

Roger A. Wehage  
US Army Tank-Automotive Command  
System Simulation & Technology Division  
Warren, MI 48397-5000

Richard Weiss  
US Army Engineer Waterways Experiment Station  
Vicksburg, Mississippi

Duminda Wijesekera  
Cornell University

Alan S. Willisky  
MIT, Lab for Information & Decision Systems  
Cambridge, Massachusetts 02139

Wing S. Wong  
AT&T Labs  
Holmdel, NJ 07733

Julian Wu  
ARO, Mathematical Science Division  
Research Triangle Park, NC 27709-2211

Patrick Xavier  
Department of Computer Science  
Cornell University

R. Yalamanchili  
SMCAR-CCR-EM (Bldg 65N)  
Picatinny Arsenal, NJ 07806

Stephen Yau  
Department of Mathematics  
University of Illinois at Chicago  
P. O. Box 4348  
Chicago, IL 60680

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

## 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			3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release: Distribution unlimited			
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE			4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARO Report 90-1			
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)			
8c. ADDRESS (City, State, and ZIP Code)			9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER			
10. SOURCE OF FUNDING NUMBERS			PROGRAM ELEMENT NO.			
			PROJECT NO.			
			TASK NO.			
			WORK UNIT ACCESSION NO.			
11. TITLE (Include Security Classification) Transactions of the Seventh Army Conference on Applied Mathematics and Computing						
12. PERSONAL AUTHOR(S)						
13a. TYPE OF REPORT Technical Report		13b. TIME COVERED FROM Jan 89 TO Feb 90		14. DATE OF REPORT (Year, Month, Day) 1990 February		15. PAGE COUNT 888
16. SUPPLEMENTARY NOTATION						
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)			
FIELD	GROUP	SUB-GROUP	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 Seventh 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 <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> 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	