AD-A193 384

(12)

ARO Report 88-1

DTIC FILE COPY

TRANSACTIONS OF THE FIFTH ARMY

CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING



DTIC
ELECTE
MAY 0 4 1988
S D
α 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 on

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

88 5 04 05

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

U. S. ARMY RESEARCH OFFICE

Report No. 88-1

March 1988

TRANSACTIONS OF THE FIFTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

QUALITY INSPECTED

Host

U. S. Military Academy
West Point, New York

15-18 June 1987

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 Army Mathematics Steering Committee (AMSC) sponsors annually the Army Conferences on Applied Mathematics and Computing. As the title indicates these meetings deal with the mathematics needed to understand the world around us. This is a contrast with core mathematics, which in the main, does not deal directly with events and objects of the physical world. Since very few of the papers presented at the four conferences held to date were in the field of pure mathematics, these meetings are rightly named. The U.S. Military Academy served as the host of the fifth meeting in this series, which was held at West Point, New York, on 15-18 June 1987. Colonel David Cameron served as Chairperson on Local Arrangements. He was assisted with this task by Majors David Arney and Scott Huxel. The members of the AMSC would like to thank these three individuals for all their efforts in coordinating the many details needed to conduct this successful scientific meeting.

The program of this years conference consisted of three parts, namely: (a) Contributed papers by Army, academic and other scientific personnel; (b) Three special sessions; and (c) Seven invited addresses. There were more than fifty contributed papers presented in the technical sessions. About half of these papers were contributed by scientists from ten Army installations. These presentations gave the attendees an opportunity to hear about scientific research being conducted within these laboratories. The topics for the special sessions were organized in three different areas, namely, stochastic analysis, solid modeling and CAD/CAM, and mathematical aspects of composites. For the invited speaker phase of the meeting, the Program Committee obtained the services of the following nationally known scientists to talk on topics of current interest to Army personnel:

SPEAKERS AND AFFILIATION	TITLE AND ADDRESSES
Professor David Munford URI Center on Intelligent Control Systems Harvard University	Some Mathematical Problems Arising from Computer Vision
Professor Roland Glowinski University of Houston	On the Numerical Solution of Time Dependent Problems in High Dimensions
Dr. Sukmar Chakravarthy Rockwell International Corporation	Unified Euler and Navier- Stokes Numerical Methods

Professor Robert Taylor University of California- Berkeley	Computation Mechanics: Today and Tomorrow
Professor Charles VanLoan Mathematical Sciences Institute Cornell University	Parallel Matrix Computations on Loosely Coupled Systems of Array Processors
Professor Anthony Jameson Princeton University	Computational Methods for Transonic Flow
Professor James Glimm Courant Institute	The Interaction on Nonlinear Waves

The success of the conference was due to many individuals, the active and enthusiastic members of the audience, the chairperson, and the many speakers. The members of the AMSC were pleased with the fact that most of the speakers were able to find time to prepare papers for the Transactions. These research articles will enable many persons that were not able to attend the symposium to profit by these contributions to the scientific literature.

TABLE OF CONTENTS

Title	Page
Foreword	iii
Table of Contents	v
Program	
Lie Transforms Applied To a Nonlinear Parametric Excitation Problem Jonathan L. Len, Richard H. Rand	1
Knowledge Representation and Planning Control in an Expert System for the Creative Design of Mechanisms David A. Hoeltzel, Wei-Hua Chieng, John Zissimides	29
Statistical Machine Learning for the Cognitive Selection of Nonlinear Programming Algorithms in Engineering Design Optimization D. A. Hoeltzel, W. H. Chieng	65
Toward a Nonequilibrium Thermodynamics of Two Phase Materials with Sharp Interface Morton E. Gurtin	80
The Shear-Lag Model for a Unidirectional Composite With Viscoelastic Matrix Dimitris C. Lagoudas, Chung-Yuen Hui, S. Leigh Phoenix	87
Element Level Elimination of Nonlinear Constraints in Total Lagrangian Finite Element Formulations A. R. Johnson and C. J. Quigley	111
Condition of the Finite Element Stiffness Matrix of Highly Irregular Triangular Grids I. Fried and A. R. Johnson	139
A Simple Analysis of Swage Autofrettage Process Peter C. T. Chen	149
Optimal Design of a Two-Way Conductor Gilbert Strang and Robert Kohn	161
On A Refined Nonlinear Theory of Laminated Composite Plates J. N. Reddy	183
Numerical Solution of Parabolic Problems in High Dimensions Edward Dean, Roland Glowinski, Chin-Hsien Li	207
Algorithms for Rational Spline Curves Klaus Hollig	287
Convexity-Preserving Quasi-interpolation and Interpolation by Box Spline Surfaces Charles K. Chui, Harvey Diamond, Louise A. Raphael	301

Knot Selection for Least Squares Aproximation Using Thin Plate Splines
John R. McMahon and Richard Franke..... 311

A Rapid, Backscatter Simulation Technique for Complex
B-Spline Target Models
Karl D. Reinig..... 331

An Algorithm for Processing Scanning Spectrometer Data
Joseph E. Zurlinden..... 345

The KP Equation - A Comparison To Laboratory Generated Biperiodic Waves
Norman W. Scheffner..... 355

Asymptotics Beyond All Orders
Harvey Segur..... 369

Computational Issues in Goal Programming Resource Allocation
Leon Medler..... 377

Design of a Feeling-Thinking Machine
Ray Scanlon and Mark Johnson..... 383

Polynominal Definition of Discrete Field Point of Map of
Diffusion Equation
William F. Donovan..... 395

The Numerical Simulation of Richtmyer-Meshkov Unstable
Interfaces: A Conference Report
John Grove..... 423

A Posteriori Error Estimation of Adaptive Finite Difference
Schemes for Hyperbolic Systems
David C. Arney, Rupak Biswas, Joseph E. Flaherty..... 437

An Adaptive Overlapping Local Grid Refinement Method for
Two-Dimensional Parabolic Systems
Peter K. Moore and Joseph E. Flaherty..... 459

Propagator Matrices in the Solution of EMP Problems
K. C. Heaton..... 475

Spline-Based Finite-Element Method for Solving
A Stefan's Problem in a Finite Domain - Formulation
Shunsuke Takagi..... 515

On Some Finite Element Error Estimates for Stress Intensity Factors
in Mode I Linear Elastic Fracture Problems
J. R. Whiteman and G. Goodsell..... 541

Some Issues of Numerical Integration and Penalty Relaxation
in Anisoparametric Shell Element Formulation
Alexander Tessler and Luciano Spiridigliozzi..... 549

A Block or Factorization Scheme for Loosely Coupled Systems of Array Processors
Charles Van Loan..... 567

Nonparametric Estimation from Queues Arising in Staggered Entry Clinical Trials
Michael J. Phelan and N. U. Prabhu..... 589

A Class of Diffusion-Type Probability Distributions
Siegfried H. Lehnigk..... 597

Column Movement Model Used to Support AMM
George B. McKinley..... 601

Influence of Reflected Shock Waves on a Hypersonic Shaped Charge Jet
H. W. Meyer and J. E. Danberg..... 617

Skew Grids and Irrotational Flow
Robert S. Bernard..... 631

A Randomness Property of m-Sequences
Harold Fredricksen and Gary Krahn..... 643

Thermodynamic Gauge Theory of Solids and Quantum Liquids with Internal Phase
Richard A. Weiss..... 649

Lagrangian Formulation of Relativistic Thermodynamics
Richard A. Weiss..... 697

Modelling of the Lean Flammability Limit in Flame Theory
Richard Y. Tam and G. S. S. Ludford..... 735

The Solution of the Type-Problem for N 4
Nam P. Bhatia and Walter O. Egerland..... 745

Neural Networks and the Symmetirc Group S_N
John L. Johnson..... 751

The Interaction of Nonlinear Hyperbolic Waves: A Conference Report
James Glimm..... 763

Stabilization of Ziegler's Pendulum by Means of the Method of Vibrational Control
G. L. Anderson, and I. G. Tadjbakhsh..... 787

List of Registrants..... 845

FIFTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

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

15-18 June 1987

AGENDA

Monday June 15, 1987

0800 - 1600 Registration - Mahon Hall Lobby

0800 - 0845 Opening Remarks - Mahon Auditorium

0845 - 0945 General Session I - Mahon Auditorium

Chairperson: Dennis M. Tracey, Materials Technology Laboratory, Watertown Massachusetts

Some Mathematical Problems Arising from Computer Vision

David Mumford, Harvard University
Cambridge, Massachusetts

0945 - 1015 Break

1015 - 1215 Technical Session 1 - Vibration and Structural Optimization
Mahon Auditorium

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

Nonlinear Parametric Excitation

J. Len and Richard H. Rand, Cornell University, Ithaca, New York

Stabilization of Ziegler's Double Pendulum by Means of the Method of Vibration Control

Gary L. Anderson and Iradj G. Tabjbakhsh, U.S. Army Research Office, Research Triangle Park, North Carolina

An Expert System for the Creative Design of Mechanisms

D. A. Hoeltzel, F. Freudenstein, W. H. Chieng, Columbia University, New York, New York

Monday (Continued)

Statistical Machine Learning for the Application of Nonlinear Programming Algorithms in Computer-Aided Mechanical Design

D. A. Hoeltzel, W. H. Chieng, Columbia University, New York, New York

Application of Symbolic Computing for Optimal Design of Large Mechanical Systems

Hashem Ashrafeiuon and Neel K. Mani, State University of New York, Buffalo, New York

Multigrid Analysis Applied to Transmission Dynamics

Albert F. Kascak, Lewis Research Center, Cleveland, Ohio

******

1015 - 1215 Technical Session 2 - Solid Mechanics I - Mahon Hall, Room 205

Chairperson: John Vasilakis, Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

Toward a Nonequilibrium Thermodynamics of Solidification

Morton E. Gurtin, Carnegie-Mellon University, Pittsburgh, Pennsylvania

Shear-Lag Model for a Composite with Viscoelastic Matrix

Dimitris C. Lagoudas, C. Hui, and S. Phoenix, Cornell University, Ithaca, New York

Element Level Elimination for Nonlinear Constraints in Total Lagangian Finite Element Formulations

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

Condition of the Finite Element Stiffness Matrix Generated from Highly Nonuniform Triangular Elements

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

Monday (Continued)

A Simple Analysis of Swage Autofrettage Process

Peter C. T. Chen, Benet Weapons Lab, Watervliet, New York

Modification of the KO Finite Difference Code for Internal State Variable Constitutive Models

Thomas Benton, U.S. Army Materials Technology Lab, Watertown, Massachusetts

1215 - 1330 Lunch

1330 - 1530 Special Session I - Mathematical Aspects of Composites - Mahon Auditorium

Chairperson: J. N. Reddy, Virginia Polytechnic Institute, Blacksburg, Virginia

Structural Optimization and Composite Material as in Optimal Designs

Gilbert Strang, Massachusetts Institute of Technology, Cambridge, Massachusetts

An Integrated Finite Element Analysis of Composite Structures

Jan L. Teply, Alcoa Technical Center, Alcoa Center, Pennsylvania and G. J. Dvorak, Rensselaer Polytechnic Institute, Troy, New York

A Finite Element Approach to Singularities in Composite Materials

Roshdy S. Barsoum, U.S. Army Materials Technology Lab, Watertown, Massachusetts

On Nonlinear Shear Deformation Theories of Composite Laminates

J. N. Reddy, Virginia Polytechnic Institute, Blacksburg, Virginia

1530 - 1600 Break

Monday (Continued)

1600 - 1700 General Session II - Mahon Auditorium

Chairperson: Stephen Wolff, National Science Foundation, Washington, DC

On the Numerical Solution of Time Dependent Problems in High Dimension. Applications

Roland Glowinski, University of Houston, Houston, Texas

Tuesday, 16 June 1987

0800 - 1600 Registration - Mahon Hall Lobby

0815 - 0955 Technical Session 3 - Approximations - Mahon Auditorium

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

High Accuracy of Geometric Hermite Interpolation

Klaus Hollig and Carl De Boor, University of Wisconsin, Madison, Wisconsin

On the Convexity of Bivariate Quadratic Spline Approximants

Charles Chui, Texas A&M, College Station, Texas; Harvey Diamond, West Virginia University, Morgantown, West Virginia; and Louise Raphael, National Science Foundation, Washington, DC

Knot Selection 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, California

A Rapid, Backscatter Simulation Techniques for Complex B-spline Target Models

Karl D. Reinig, Harry Diamond Lab, Adelphi, Maryland

An Algorithm for Processing Scanning Spectrometer Data

Joseph E. Zurlinden, U.S. Army White Sands Missile Range, White Sands Missile Range, New Mexico

Tuesday (Continued)

0815 - 0955 Technical Session 4 - Mahon Hall, Room 205

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

The KP Equation - A Comparison to Laboratory Generated Biperiodic Waves

Norman W. Scheffner, U.S. Army Engineer Waterways Experiment Station, Coastal Engineering Research Center, Vicksburg, Mississippi

Asymptotics Beyond All Orders

Harvey Segur and Martin D. Kruskal, ARAP Division of CRT, Inc., Princeton, New York

Ground Mobility Tactical Decision Aids on a Microcomputer

T. C. Falls, U.S. Army Engineer Waterways Experiment Station, Geotechnical Laboratory, Vicksburg, Mississippi

Computational Issues in Goal Programming Resource Allocation

Leon Medler, U.S. Army Belvoir Research, Development, and Engineering Center, Ft. Belvoir, Virginia

Designing A Feeling, Thinking Machine

M. Johnson and R. Scanlon, Benet Weapons Lab, Watervliet, New York

******

0815 - 0955 Poster Session - Mahon Hall, Room 207

Diffusion Equation Resolution by Difference Equation Regression, Part I

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

0955 - 1015 Break

Tuesday (Continued)

1015 - 1115 General Session III - Mahon Auditorium

Chairperson: Miles Miller, Chemical Research Development Center, Aberdeen Proving Ground, Maryland

Unified Euler and Navier-Stokes Numerical Methods

Sukmar Chakravarthy, Science Center, Rockwell International Corporation, Thousand Oaks, California

1115 - 1330 Tour/Lunch

Also Demo: Data Analysis/Mathematical Modeling
1130 - 1200

Computer Aided Geo. Analysis
1230 - 1300

1330 - 1530 Special Session II - Solid Modeling & CAD/CAM - Mahon Auditorium

Chairperson: Mark S. Shephard, Rensselaer Polytechnic Institute, Troy, New York

Analysis Model Generation and Control from Solid Models

Mark S. Shephard, Rensselaer Polytechnic Institute Troy, New York

Solid Modeling for Automated Tolerance Analysis

Joshua Turner, IBM, Poughkeepsie, New York

Trimmed Surface Algorithms for the Evaluation and Interrogation of Solid Boundary Representations

Rida T. Farouki, IBM Thomas J. Watson Research Center, Yorktown Heights, New York

The Use of Topology in Geometric Modeling Systems

Kevin Weiler, General Electric Corporate Research and Development Center, Schenectady, New York

1530 - 1600 Break

Tuesday (Continued)

1600 - 1700 General Session IV - Mahon Auditorium

Chairperson: COL David H. Cameron, U.S. Military Academy, West Point, New York

Computational Mechanics:Today and Tommorrow

Robert Taylor, University of California-Berkeley, Berkeley, California

**

1830 - 1900 Social Gathering - West Point Officer's Club

1900 - 2000 Banquet

2000 - 2030 Invited Speaker

Wednesday, 17 June 1987

0800 - 1600 Registration - Mahon Hall Lobby

0815 - 1015 Technical Session 5 - Numerical PDE - Mahon Auditorium

Chairperson: K. O'Neill, Cold Region Research and Engineering Laboratory, Hanover, New Hampshire

Front Tracking and Shock-Contact Interactions

John Grove, New York University, New York, New York

A Posteriori Error Estimation of Adaptive Finite Difference Schemes for Hyperbolic Systems

David C. Arney, U.S. Military Academy, West Point, New York, Rupak Biswas and Joseph E. Flaherty, Rensselaer Polytechnic Institute, Troy, New York

Local Refinement Finite Element Methods for Parabolic Systems

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

Wednesday (Continued)

Magnetic Resonance Coil Design in the Presence of Modifying Half-Space

J. F. Schenck and M. A. Hussain, General Electric Corporate Research and Development Center, Schenectady, New York

Electromagnetic Pulses Including the Calculating of the Magnetics as well as the Electric Fields Near Explosion Site (Propagator Matrices)

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

On a Comparison of Exact and Empirical Results of Convective Heat Transfer

Rao Yalamanchili, U.S. Army ARDEC, Picatinny Arsenal, New Jersey

******

0815 - 1015 Technical Session 6 - Solid Mechanics II - Mahon Hall, Room 205

Chairperson: S. C. Chu, Army Research and Development Command, Dover, New Jersery

Problems with the Solutions of Crack Problems of Elastic Composites

Ram Srivastav, State University of New York, Stony Brook, New York

Spline-Based Finite Element Method for Solving a Stefan's Problem in a Finite Domain

Shunsuke Takagi, Cold Regions Research and Engineering Lab, Corps of Engineers, Hanover, New Hampshire

Plasticity and Microstructural Damage

C. Freese, P. Perrone, D. Tracey, and P. Tsirigotis, U.S. Army Materials Technology Lab, Watertown, Massachusetts

Optimal Bounds and the G-Closure Problem for Two Dimensional Homogenized

Robert Lipton, Cornell University, Ithaca, New York

Wednesday (Continued)

Superconvergence in Finite Element Methods for Linear and Nonlinear Fracture

John Whiteman, Institute of Mathematics, University of Brunel, Uxbrdige, Middleses, United Kingdom

Some Issues of Numerical Integration and Penalty Relaxation in Anisoparametric Shell Element Formulation

A. Tessler and L. Spiridigliozzi, U.S. Army Materials Technology Lab, Watertown, Massachusetts

1015 - 1045 Break

1045 - 1145 General Session V - Mahon Auditorium

Chairperson: Colin E. Freese, Materials Technology Laboratory, Watertown, Massachusetts

Parallel Matrix Computations on Loosely Coupled Systems of Array Processors

Charles Van Loan, Cornell University, Ithaca, New York

1145 - 1300 Lunch

1300 - 1530 Special Session III - Stochastic Analysis - Mahon Auditorium

Chairperson: Gerald Andersen, U.S. Army Research Office, Research Triangle Park, NC

Stochastic Quantization

Sanjoy Mitter, Massachusetts Institute of Technology, Cambridge, Massachusetts

Stochastic Growth Models

Richard Durrett

Statistical Inference from an Infinite Server Queueing System

Narahari Prabhu

Wednesday (Continued)

The Analytical Approach to a Class of Diffusion Type Probability Distributions

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

Nonlinear Filtering of Discrete Parameter Point Processes

Gerald R. Andersen, U.S. Army Research Office, Research Triangle Park, North Carolina

1530 - 1600 Break

1600 - 1700 Technical Session 7 - Mahon Auditorium

Chairperson: Billy Z. Jenkins, U.S. Army Missile Command, Redstone Arsenal, Alabama

Column Movement Model Used to Support AMM

G. B. McKinley, U.S. Army Engineer Waterways Experiment Station, Geotechnical Laboratory, Vicksburg, Mississippi

A Study of the Aerodynamics of a Shaped Charge Jet

Hubert W. Meyer, Jr. and James E. Danbery, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Skew Grids and Irrotational Flow

Robert S. Bernard, U.S. Army Engineering Waterways Experimental Station, Hydraulics Lab, Vicksburg, Mississippi

******

1600 - 1700 Technical Session 8 - Mahon Hall, Room 205

Chairperson: COL James Kays, U.S. Military Academy, West Point, New York

A Randomness Property of m-Sequences

Gary W. Krahn, U.S. Military Academy, West Point, New York and Harold Fredricksen, Naval Postgraduate School, Monterey, California

Wednesday (Continued)

Thermodynamic Guage Theory of Solids and Quantum Liquids with Internal Phase

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

Lagrangian Formulation of Relativistic Thermodynamics

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

******

Thursday, 18 June 1987

0800 - 1100 Registration - Mahon Lobby

0830 - 0950 Technical Session 9 - Analysis - Mahon Auditorium

Chairperson: Raymond Sedney, U.S. Army Ballistic Research Lab, Aberdeen Proving Ground, Maryland

Modelling of the Lean Flammability Limit in Flame Theory

Richard Y. Tam, Purdue University, Indianapolis, Indiana

On Plastic Shear Instability at High Strain Rate

Timothy J. Burns, National Bureau of Standards, Washington, DC

The Solution of the Type-Problem for N = 4

Walter O. Egerland, U.S. Army Ballistic Research Lab, Aberdeen Proving Ground, Maryland

Neural Networks and the Symmetric Group S_N

John L. Johnson, U.S. Army Missile Command, Redstone Arsenal, Alabama

******

Thursday (Continued)

0830 - 0950 Technical Session 10 - Estimation and Filtering - Mahon Hall, Room 205

Chairperson: Robert E. Green, White Sands Missile Range, New Mexico

An Algorithm for Adaptive System Identification

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

Simulation of Sub-Optimal Kalman Filter Design for Target Tracking Application Using Computer Algebra

Radhakisan S. Baheti and Moayyed A. Hussian, General Electric Corporate Research & Development Center, Schenectady, New York

A New Approach in the Optimal Guidance of Tactical Missiles

R. V. Ramnath, Sparta Systems, Inc., Lexington, Maassachusetts

Shannon's Sampling Theorem and Control Theory

Charles E. Hall, Jr., U.S. Army Missile Command, Redstone Arsenal, Alabama

0950 - 1015 Break

1015 - 1215 General Session VI - Mahon Auditorium

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

Computational Methodology for Transonic Flow

Antony Jameson, Princeton University, Princeton, New Jersey

The Interactions of Nonlinear Waves

James Glimm, Courant Institute, New York University, New York, New York

1215 - 1230 Adjournment

Lie Transforms Applied To A Nonlinear Parametric Excitation Problem*

Jonathan L. Len
Center for Applied Mathematics

Richard H. Rand
Department of Theoretical and Applied Mechanics
Cornell University
Ithaca, New York 14853

Abstract

We use Lie transforms to approximate the Poincaré map of a weakly nonlinear periodic perturbation of the simple harmonic oscillator in order to study the stability of the trivial solution. Resonant frequencies, corresponding to nonremovable terms in the differential equation, are identified through $O(\epsilon^2)$. We show that detuning from resonance stabilizes the trivial solution when the perturbation contains no linear periodic terms. Finally, we study a typical bifurcation between two lowest-order resonant frequencies. A MACSYMA program which performs the Lie transform algorithm to arbitrary order is presented in the appendix with a sample run.

1 Introduction

In this paper we present some results concerning the stability of the trivial solution of the equation

$$\ddot{x} + \omega^2 x + \epsilon f(t,x) = 0 \tag{1}$$

where $f(t,x)$ is T-periodic in t, Taylor-Fourier expandable in x and t respectively, and $f(t,x)$ satisfies $f(t,0) \equiv 0$. The hamiltonian structure of eq.(1)

*This work was partially supported by NSF grant 85-09481 and by the Army Research Office through the Mathematical Sciences Institute, Cornell University, Ithaca, NY 14853

permits us to use *Lie transforms* to reduce the nonautonomous hamiltonian induced by eq.(1) to an autonomous one by means of a periodic canonical near-identity transformation. The resulting autonomous hamiltonian describes the Poincaré map in a neighborhood of the origin.

Analysis of the Poincaré map gives substantial information concerning the original equation. The presence of a periodic point in the Poincaré map implies the existence of a periodic orbit in the original equation. In particular, a periodic saddle point corresponds to a hyperbolic periodic orbit, and a periodic center corresponds to an elliptic periodic orbit.

We begin by describing the Lie transform algorithm as used in this work. We then present a theorem which defines the $O(\epsilon)$ and $O(\epsilon^2)$ resonances for the general case of eq.(1), and show that almost all higher order resonances are stable.

Next, we study the properties of the trivial solution of a simple equation of the type (1). We identify the $O(\epsilon)$ and $O(\epsilon^2)$ resonances, and characterize the stability of the trivial solution for all nonzero ω. Results of the Lie transform analysis are compared with numerically generated Poincaré maps.

Finally, we study a bifurcation between $O(\epsilon)$ resonances of cubic and quadratic nonlinearities. In this example, a 4π-periodic hyperbolic orbit becomes a 2π-periodic hyperbolic orbit through a sequence of bifurcations.

2 Results

We consider the general equation

$$\ddot{x} + \omega^2 x + \epsilon \sum_{\alpha} g_\alpha(t) x^{N_\alpha - 1} = 0 \tag{2}$$

where the $g_\alpha(t)$ are periodic and the N_α are positive integers. This equation was studied extensively in [1]. Here we summarize some results and refer the reader to [1] for additional information.

In canonical variables q and p, eq.(2) is generated by the hamiltonian

$$h(q,p,t) = \frac{p^2}{2} + \frac{\omega^2 q^2}{2} + \epsilon \sum_{\alpha} \frac{1}{N_\alpha} g_\alpha(t) q^{N_\alpha}. \tag{3}$$

The change of variables

$$\begin{aligned} q &= \sqrt{2J/\omega}\sin(\theta + \omega t), \\ p &= \sqrt{2J\omega}\cos(\theta + \omega t) \end{aligned} \tag{4}$$

reduces eq.(3) to the $O(\epsilon)$ hamiltonian

$$H(J,\theta,t) = \epsilon \sum_{\alpha} \frac{1}{N_\alpha} g_\alpha(t) \left(\frac{2J}{\omega}\right)^{N_\alpha/2} \sin^{N_\alpha}(\theta + \omega t) \equiv \epsilon H_1(J,\theta,t). \tag{5}$$

We will apply the Lie transform procedure to this $O(\epsilon)$ hamiltonian.

Definition 1 *Let*

$$w(t,x,\epsilon) = w_1(t,x) + \epsilon w_2(t,x) + \epsilon^2 w_3(t,x) + \cdots$$

be the Lie generating function defining a canonical transformation which reduces the hamiltonian (5) to an autonomous one. Then ω is a resonance at $O(\epsilon^n)$ if it is a pole of $w_n(t,x)$ but not of $w_k(t,x)$, for $1 \leq k \leq n-1$.

We denote by Ω_n the set of frequencies which are resonant at $O(\epsilon^n)$.

An equivalent definition of a resonant frequency may be formulated in terms of the near-identity transformation generated by periodic averaging.

For example, the $O(\epsilon^n)$ resonance for the linear Mathieu equation

$$\ddot{x} + \omega^2 x + \epsilon x \cos t = 0$$

is $\omega = n/2$, for $n \geq 1$. Resonant frequencies correspond to non-removable terms in the hamiltonian (with respect to Lie transforms) or in the differential equation (with respect to periodic averaging).

In order to show how to generate all $O(\epsilon)$ and $O(\epsilon^2)$ resonances for eq.(2), we introduce some notation. By assumption, each $g_\alpha(t)$ is periodic and may therefore be expanded in a Fourier series. Let

$$c_\mu^{(\alpha)} = \frac{1}{2\pi}\int_0^{2\pi} g_\alpha(t) e^{-i\mu t} dt$$

be the Fourier coefficients of $g_\alpha(t)$ for integer μ. Let M_α denote the set of frequencies of $g_\alpha(t)$, that is,

$$M_\alpha = \{\mu : c_\mu^{(\alpha)} \neq 0, \mu \in Z\}.$$

Then

$$g_\alpha(t) = \sum_{\mu \in M_\alpha} c_\mu^{(\alpha)} e^{i\mu t}.$$

As shown in [1], Ω_1 consists of all ω satisfying

$$\omega = \frac{\mu}{N_\alpha - 2\nu}, \qquad \begin{array}{c} \mu \in M_\alpha \\ 0 \leq \nu \leq N_\alpha. \end{array}$$

and Ω_2 consists of all ω satisfying

$$\omega = \frac{\mu + \gamma}{N_\alpha + N_\beta - 2(\nu + \delta)} \qquad \begin{array}{c} \mu \in M_\alpha \\ \gamma \in M_\beta \\ 0 \leq \nu \leq N_\alpha \\ 0 \leq \delta \leq N_\beta \\ \delta N_\alpha \neq \nu N_\beta \end{array}$$

which are not also included in Ω_1.

It is also shown in [1] that if $\omega \notin \Omega_1 \cup \Omega_2 \cup \{0\}$, then the resulting reduced hamiltonian is of the form

$$K = \epsilon f_1(J) + \epsilon^2 f_2(J) + \cdots$$

where f_1 and f_2 contain integer or half-integer powers of J. This implies that the origin of the Poincaré map is a center, and therefore the trivial solution is a stable elliptic orbit.

We show here that if the perturbation is strictly nonlinear then detuning creates a hyperbolic periodic orbit which traps the trivial solution, stabilizing it. We then analyze a bifurcation problem between periodic orbits near resonances showing how a 4π-periodic orbit bifurcates into a 2π-periodic orbit.

3 Lie Transforms

An important characteristic of autonomous hamiltonian systems is that the hamiltonian is constant along solutions of the system of differential equations. If the phase space has dimension two then the solutions are level curves of the hamiltonian. The reader is referred to [3] or [4] for a complete discussion of hamiltonian mechanics.

In this work we use Lie transforms to reduce eq.(3) to an autonomous hamiltonian, and then analyze the level curves of this autonomous hamiltonian to determine the behavior of solutions which have initial conditions close to the trivial solution. The implementation of the Lie transform algorithm which is presented here implicitly constructs a canonical change of coordinates which performs the reduction to an autonomous form. It is obvious that no autonomous canonical change of variables can make this reduction. Therefore the hamiltonian with respect to the new coordinates must be determined by means of a generating function or some equivalent method which takes into account the nonautonomous nature of the transformation. The Lie transform method is an efficient perturbation scheme which explicitly generates the functional form of the reduced hamiltonian under an implicitly defined canonical periodic near-identity transformation.

Let x and y denote the old and new coordinates, respectively. Let ϵ denote the perturbation parameter. Let H denote the hamiltonian with respect to the x coordinates, and let K denote the transformed hamiltonian. We assume that H and K may each be written as power series in ϵ,

$$H(t,x,\epsilon) = H_0(t,x) + \epsilon H_1(t,x) + \epsilon^2 H_2(t,x) + \cdots$$

and

$$K(t,x,\epsilon) = K_0(t,x) + \epsilon K_1(t,x) + \epsilon^2 K_2(t,x) + \cdots$$

The relation between x and y is defined implicitly in terms of a *Lie generating function* $w(t,x)$ as

$$\frac{\partial y_i}{\partial \epsilon} = \{x_i, w\} \tag{6}$$

where $\{\ ,\ \}$ is the *Poisson bracket* operator. For two-dimensional phase space, the Poisson bracket operator is

$$\{f,g\} = \frac{\partial f}{\partial x_1}\frac{\partial g}{\partial x_2} - \frac{\partial f}{\partial x_2}\frac{\partial g}{\partial x_1}.$$

In words, the new coordinate system evolves from the old one by means of a "hamiltonian flow" in the evolution quantity ϵ. See [3] for a complete discussion of this procedure. It is straightforward to show that the change of variables $x \rightarrow y$ defined by eq.(6) is canonical. This consists of showing that the fundamental Lagrange brackets are preserved under the transformation.

The reduced hamiltonian K is related to H by

$$\begin{aligned} K_0 &= H_0 \\ K_1 &= H_1 + \{w_1, H_1\} + \frac{\partial w_1}{\partial t} \\ K_2 &= H_2 + \frac{1}{2}\{w_1, K_1 + H_1\} + \frac{1}{2}\frac{\partial w_2}{\partial t} + \{w_2, H_0\} \\ &\vdots \end{aligned} \tag{7}$$

Although this sequence can be written in closed form to arbitrary order, we need it only through $O(\epsilon^2)$. See [2] for full details of this topic.

It is important to interpret eq.(7) correctly. The right-hand side of each equation is a function of x and t, and the Poisson brackets are computed with respect to the x coordinate system. The resulting function K is evaluated at $K(t,x)$. The x are dummy variables, and may be replaced by y to give the transformed hamiltonian.

The sequence (7) gives the transformed hamiltonian for an arbitrary generating function w. The trick is to choose successive w_i to make the corresponding K_i as simple as possible. This means choosing w_i at the ith step such that

$$\frac{\partial w_i}{\partial t} + \{w_i, H_0\}$$

removes as many terms as possible in the right-hand side of the ith equation in (7). While this operator is linear, it has a nontrivial kernel; therefore some terms may not be removable. In the context of periodic perturbations, this means that w_i cannot be chosen to make K_i autonomous directly. However, after all nonessential terms have been removed to desired order using Lie

transforms, a final canonical transformation of the form

$$\begin{aligned} J &\rightarrow I \\ \theta &\rightarrow \Phi + \alpha t \end{aligned}$$

for some scalar α can always be found which makes it autonomous.

The method may be simplified considerably by the following trick: Apply a canonical transformation which removes the $O(1)$ terms of the hamiltonian so that $H_0 \equiv 0$. Then all Poisson brackets in (7) involving H_0 vanish, and the terms which are removable are precisely the t-dependent ones. The appropriate choice of w_i is to take $-w_i$ as the t-antiderivative of the t-dependent terms. The resulting K is autonomous by construction. This modified Lie transform algorithm has been implemented in MACSYMA since the amount of algebra required to carry the perturbation scheme through even $O(\epsilon^2)$ is too daunting to compute by hand with any confidence. The program and sample runs are given in the appendix. For a further discussion on the use of computer algebra in perturbation schemes, see [1], [6], and [7].

While the simplification of the algorithm is important from the computer algebra point of view, it is perhaps more important for analytical purposes. This modified method was used to determine the $O(\epsilon)$ and $O(\epsilon^2)$ resonances of the general equation given previously.

In principle, this strategy may be used in any system where the $\epsilon = 0$ problem may be solved exactly. For example, a system of linear oscillators with weak nonlinear coupling may be studied using this simplification.

4 Determining the Resonant Frequencies

In this section we briefly describe the procedure by which resonances may be found using Lie transforms. For complete details, see [1].

We assume that the equation is of the form

$$\ddot{x} + \omega^2 x + \epsilon \sum_\alpha g_\alpha(t) x^{N_\alpha - 1} = 0.$$

In canonical variables q and p, this equation gives rise to the hamiltonian

$$h(q,p,t) = \frac{p^2}{2} + \frac{\omega^2 q^2}{2} + \epsilon \sum_\alpha \frac{1}{N_\alpha} g_\alpha(t) q^{N_\alpha}. \tag{8}$$

Our first step, as described at the end of the previous section, will be to perform a transformation of coordinates to a system in which the hamiltonian contains no $O(1)$ terms. The change of variables

$$\begin{aligned} q &= \sqrt{2J/\omega}\sin(\theta + \omega t), \\ p &= \sqrt{2J\omega}\cos(\theta + \omega t) \end{aligned} \tag{9}$$

reduces eq.(8) to the simplified hamiltonian

$$H(J,\theta,t) = \epsilon \sum_{\alpha} \frac{1}{N_\alpha} g_\alpha(t) (\frac{2J}{\omega})^{N_\alpha/2} \sin^{N_\alpha}(\theta + \omega t) \equiv \epsilon H_1(J,\theta,t). \qquad (10)$$

We now apply the Lie transform procedure to transform eq.(10) into an autonomous hamiltonian. Note that no $O(1)$ terms are present. As noted at the end of the previous section, this simplification permits us to compute the Lie generating function at each step by integration of exponentials.

To identify the resonances, we first compute w_1 for arbitrary ω. This gives a function similar in form to H_1 but whose coefficients are rational functions of ω. The poles of these coefficients, which correspond to non-removable terms in $H_1(J,\theta,t)$, are frequencies which are resonant at $O(\epsilon)$. Having identified the $O(\epsilon)$ resonances, we may then implicitly compute w_2 to identify possible $O(\epsilon^2)$ resonances.

We first introduce some notation. By assumption, each $g_\alpha(t)$ is periodic and may therefore be expanded in a Fourier series. Let

$$c_\mu^{(\alpha)} = \frac{1}{2\pi} \int_0^{2\pi} g_\alpha(t) e^{-i\mu t} dt$$

be the Fourier coefficients of $g_\alpha(t)$ for integer μ. Let M_α denote the set of frequencies of $g_\alpha(t)$, that is,

$$M_\alpha = \{\mu : c_\mu^{(\alpha)} \neq 0, \mu \in Z\}.$$

Then

$$g_\alpha(t) = \sum_{\mu \in M_\alpha} c_\mu^{(\alpha)} e^{i\mu t}.$$

Expanding the trigonometric functions with the binomial theorem and inserting the expansion for $g_\alpha(t)$ in eq.(10) gives

$$H_1(J,\theta,t) = \sum_{\alpha} \sum_{\mu \in M_\alpha} \sum_{\nu=0}^{N_\alpha} a_{\mu\nu}^{(\alpha)} e^{i\theta(2\nu - N_\alpha)} e^{it((2\nu - N_\alpha)\omega + \mu)} \qquad (11)$$

where

$$a_{\mu\nu}^{(\alpha)} = \frac{1}{N_\alpha} c_\mu^{(\alpha)} (\frac{J}{2\omega})^{N_\alpha/2} \binom{N_\alpha}{\nu} (-1)^{N_\alpha/2 - \nu}. \qquad (12)$$

Proceeding formally, w_1 is just the negative of the t-antiderivative of H_1:

$$w_1 = \sum_{\alpha} \sum_{\mu \in M_\alpha} \sum_{\nu=0}^{N_\alpha} \frac{-a_{\mu\nu}^{(\alpha)} e^{i\theta(2\nu - N_\alpha)} e^{it((2\nu - N_\alpha)\omega + \mu)}}{i((2\nu - N_\alpha)\omega + \mu)}.$$

This choice of w_1 makes K_1 the t-independent part of H_1.

The poles of w_1 are $\omega = 0$, which we shall ignore, and

$$\omega = \frac{\mu}{N_\alpha - 2\nu}, \qquad \begin{array}{c} \mu \in M_\alpha \\ 0 \le \nu \le N_\alpha. \end{array}$$

Let Ω_1 denote the set of $O(\epsilon)$ resonances. Let Ω_2 denote the set of poles of w_2 which are not in Ω_1. Then Ω_2 the set of $O(\epsilon^2)$ resonances. The equation defining w_2 is

$$\frac{\partial w_2}{\partial t} = 2K_2 - \{w_1, K_1\} - \{w_1, H_1\}. \tag{13}$$

It is sufficient to determine the possible exponents introduced in the right hand side of eq.(13) since the poles of w_2 correspond to roots of the exponents. Since K_1 and K_2 are autonomous by construction, the new resonances can come only from the term $\{w_1, H_1\}$. It is clear from the definition of $a_{\mu\nu}^{(\alpha)}$ that

$$\frac{\partial a_{\mu\nu}^{(\alpha)}}{\partial J} = \frac{N_\alpha}{2J} a_{\mu\nu}^{(\alpha)}.$$

Therefore,

$$\{w_1, H_1\} = \sum_{\alpha,\beta} \sum_{\substack{\mu \in M_\alpha \\ \gamma \in M_\beta}} \sum_{\nu=0}^{N_\alpha} \sum_{\delta=0}^{N_\beta} \frac{a_{\mu\nu}^{(\alpha)} a_{\gamma\delta}^{(\beta)} (\delta N_\alpha - \nu N_\beta)}{i((2\nu - N_\alpha)\omega + \mu)} e^{i\theta(2(\nu+\delta) - N_\alpha - N_\beta)}$$

$$e^{it((2(\nu+\delta) - N_\alpha - N_\beta)\omega + \mu + \gamma)} \tag{14}$$

If a frequency is a resonance, then it is a root of the t-dependent exponential. It is easily seen that Ω_2 consists of all ω which satisfy

$$\omega = \frac{\mu + \gamma}{N_\alpha + N_\beta - 2(\nu + \delta)} \qquad \begin{array}{c} \mu \in M_\alpha \\ \gamma \in M_\beta \\ 0 \le \nu \le N_\alpha \\ 0 \le \delta \le N_\beta \\ \delta N_\alpha \ne \nu N_\beta \end{array}$$

but which are not also included in Ω_1.

We conclude this section with some examples which demonstrate how to compute the resonant frequencies.

4.1 Examples

Example 1

$$\ddot{x} + \omega^2 x + \epsilon x \cos(t) = 0$$

For this example,

$$N_1 = 2, \qquad M_1 = \{-1, 1\}.$$

Ω_1 is generated by the numerators ± 1 and denominators $2 - 2\nu$ with $\nu = 0$ or $\nu = 2$. The only positive resonance is $\omega = 1/2$. For Ω_2, the possible numerators are $1 + 1$ and $-1 - 1$. The denominators are given by $2 + 2 - 2(\nu + \delta)$ where $\nu = 0, 1, 2$ and $\delta = 0, 1, 2$ with $\nu \neq \delta$. Therefore $\nu + \delta$ can take on the values 1,2, and 3, and consequently the allowed denominators are ± 2. The only frequency generated is $\omega = 1$. Therefore

$$\begin{aligned} \Omega_1 &= \{\frac{1}{2}\}, \\ \Omega_2 &= \{1\}. \end{aligned}$$

This agrees with the classical result for the Mathieu equation, which is that the $O(\epsilon^n)$ resonance is $n/2$.

Example 2

$$\ddot{x} + \omega^2 x + \epsilon x \cos(t) + \epsilon x^3 = 0$$

Then

$$N_1 = 2, \qquad M_1 = \{-1, 1\}.$$

$$N_2 = 4, \qquad M_2 = \{0\}.$$

Ω_1 is determined exactly as in the previous example since the set M_2 cannot contribute a nonzero frequency. (The $O(\epsilon)$ resonances can always be found by considering each term of the perturbation separately). For Ω_2, the resonance $\omega = 1$ is generated as in the previous example. The "mixing" of the sets M_1 and M_2 introduces the possible numerators $\pm 1 + 0$, with corresponding denominators $2 + 4 - 2(\nu + \delta)$ where $\nu = 0, 1, 2$ and $\delta = 0, 1, 2, 3, 4$. The forbidden pairs are $(\nu, \delta) = (0, 0)$, $(\nu, \delta) = (1, 2)$, and $(\nu, \delta) = (2, 4)$. The allowed values of $\nu + \delta$ are 1,2,3,4,and 5, giving allowed denominators ± 2 and ± 4, so the new resonance is $\omega = 1/4$:

$$\begin{aligned} \Omega_1 &= \{\frac{1}{2}\}, \\ \Omega_2 &= \{1, \frac{1}{4}\}. \end{aligned}$$

Example 3

$$\ddot{x} + \omega^2 x + \epsilon x^n \cos(t) = 0, \qquad n \text{ odd}$$

Here

$$N_1 = n+1, \qquad M_1 = \{-1, 1\}.$$

The resonant frequencies are

$$\begin{aligned} \Omega_1 &= \{\frac{1}{n+1}, \frac{1}{n-1}, \frac{1}{n-3}, \cdots, \frac{1}{2}\}, \\ \Omega_2 &= \{\frac{1}{n}, \frac{1}{n-2}, \frac{1}{n-4}, \cdots, 1\}. \end{aligned}$$

Example 4

$$\ddot{x} + \omega^2 x + \epsilon x^n \cos(t) = 0, \qquad n \text{ even}$$

For this example,

$$N_1 = n+1, \qquad M_1 = \{-1, 1\}.$$

The resonant frequencies are

$$\begin{aligned} \Omega_1 &= \{\frac{1}{n+1}, \frac{1}{n-1}, \frac{1}{n-3}, \cdots, 1\}, \\ \Omega_2 &= \{\frac{1}{n}, \frac{1}{n-2}, \frac{1}{n-4}, \cdots, \frac{1}{2}\}. \end{aligned}$$

Example 5

$$\ddot{x} + \omega^2 x + \epsilon x(\cos t + \cos 5t) + \epsilon x^3(1 + \cos 3t + \cos 7t) = 0$$

Then

$$\begin{aligned} N_1 &= 2 \quad M_1 = \{-1, 1, -5, 5\} \\ N_2 &= 4 \quad M_2 = \{0, -3, 3, -7, 7\}. \end{aligned}$$

The resonant frequencies are

$$\begin{aligned} \Omega_1 &= \{\frac{1}{2}, 1, \frac{7}{3}, \frac{5}{2}, 3, 7\}, \\ \Omega_2 &= \{\frac{1}{3}, \frac{2}{3}, \frac{3}{4}, \frac{4}{3}, \frac{3}{2}, \frac{5}{5}, \frac{7}{4}, 2, \frac{8}{3}, \frac{7}{2}, 4, 5, 6, 8, 12\}. \end{aligned}$$

Example 6

$$\ddot{x} + \omega^2 x + \epsilon x \cos t + \epsilon x^2 \cos 5t + \epsilon x^3 \cos 12t + \epsilon x^4(1 + \cos 22t) = 0$$

Here

$$\begin{aligned} N_1 &= 2 \quad M_1 = \{-1, 1\} \\ N_2 &= 3 \quad M_2 = \{-5, 5\} \\ N_3 &= 4 \quad M_3 = \{-12, 12\} \\ N_4 &= 5 \quad M_4 = \{0, -22, 22\}. \end{aligned}$$

The $O(\epsilon)$ resonant frequencies are

$$\Omega_1 = \{\frac{1}{2}, \frac{5}{3}, 3, \frac{22}{5}, 5, 6, \frac{22}{3}, 22\}.$$

The $O(\epsilon^2)$ resonances are

$$\begin{aligned}\Omega_2 \;=\; \{&34, 23, 21, 17, \frac{27}{2}, 12, \frac{34}{3}, 11, 10, \frac{17}{2}, \frac{23}{3}, 7\frac{34}{5}, \\ &\frac{27}{4}, \frac{13}{2}, \frac{17}{3}, \frac{11}{2}, \frac{34}{7}, \frac{23}{5}, \frac{9}{2}, \frac{17}{4}, \frac{21}{5}, 4, \frac{11}{3}, \frac{17}{5}, \frac{10}{3}, \\ &\frac{13}{4}, \frac{17}{6}, \frac{11}{4}, \frac{5}{2}, \frac{12}{5}, \frac{7}{3}, 2, \frac{12}{7}, \frac{10}{7}, \frac{7}{5}, \frac{4}{3}, \frac{5}{4}, 1, \frac{5}{6}, \frac{1}{3}, \frac{1}{5}\}.\end{aligned}$$

5 The Stability of the Trivial Solution Near Resonance

Having identified the resonant frequencies, we now study the behavior of solutions close to resonance. We first study the major qualitative difference between linear and nonlinear parametric excitation.

We consider equations of the form

$$\ddot{x} + \omega^2 x + \epsilon f(t, x) = 0 \tag{15}$$

where $f(t, x)$ is periodic in t and strictly nonlinear in x. The case when $f(t, x)$ contains terms which are linear in x with periodic coefficients has been studied previously [5].

Let ω_0 be a resonance, and take ω in eq.(15) to be

$$\omega^2 = \omega_0^2 + \epsilon\omega_1 + \epsilon^2\omega_2 + \cdots$$

Then eq.(15) becomes

$$\ddot{x} + \omega_0^2 x + \epsilon f(t, x) + (\epsilon\omega_1 + \epsilon^2\omega_2 + \cdots)x = 0.$$

Detuning from resonance introduces a linear t-independent perturbation. Since detuning at $O(\epsilon^n)$ introduces a term of the form $\epsilon^n J$ to H, it also contributes a term to K_n which is independent of θ and *linear* in J. Since a nonlinear term of order $O(x^m)$ in $f(t, x)$ contributes terms of order $O(J^{(m+1)/2})$ to K_1 and terms of higher order to subsequent K_i, the stabilizing effect of the detuning will dominate in a sufficiently small neighborhood of the origin. (This analysis requires that ϵ be held fixed, while J may be taken as small as necessary. For sufficiently small J the linear term dominates.) This implies that a strictly nonlinear periodic perturbation cannot cause the trivial solution to be unstable away from resonance.

6 The Effect of Detuning From Resonance

We demonstrate the effect of detuning from resonance on the equation

$$\ddot{x} + \omega^2 x + \epsilon x^3 \cos t = 0. \tag{16}$$

The resonances, as shown in a previous section, are

$$\begin{aligned} \Omega_1 &= \{\frac{1}{2}, \frac{1}{4}\}, \\ \Omega_2 &= \{1, \frac{1}{3}\}. \end{aligned}$$

The MACSYMA implementation of the Lie transform algorithm, which is listed in the appendix, shows that for $\omega^2 = \frac{1}{4} + \epsilon\omega_1 + \epsilon^2\omega_2$ the $O(\epsilon^2)$ the reduced hamiltonian is

$$\begin{aligned} K &= -\frac{3}{2}\epsilon^2 J^3 \cos 4\theta + 4\omega_1\epsilon^2 J^2 \cos 2\theta - \epsilon J^2 \cos 2\theta \\ &\quad -\frac{4}{3}\epsilon^2 J^3 + \omega_2\epsilon^2 J - \omega_1^2\epsilon^2 J + \omega_1\epsilon J. \end{aligned} \tag{17}$$

The fixed points satisfy

$$\begin{aligned} \frac{\partial K}{\partial J} &= 0, \\ \frac{\partial K}{\partial \theta} &= 0. \end{aligned}$$

Solving for fixed points gives the $O(1)$ pairs of fixed points

$$\begin{aligned} J &= \frac{\omega_1}{2} + \frac{7\omega_1^2 + 8\omega_2}{16}\epsilon - \frac{7\omega_1^3 + 8\omega_1\omega_2}{64}\epsilon^2 + \cdots \\ \theta &= \frac{\pi}{2}, \frac{3\pi}{2} \end{aligned}$$

for $\omega_1 < 0$ and

$$\begin{aligned} J &= \frac{\omega_1}{2} + \frac{7\omega_1^2 + 8\omega_2}{16}\epsilon - \frac{7\omega_1^3 + 8\omega_1\omega_2}{64}\epsilon^2 + \cdots \\ \theta &= 0, \pi \end{aligned}$$

for $\omega_1 > 0$. (Solutions which are $O(1/\epsilon)$ also exist, but we ignore them since we are interested in the behavior of the trivial solution a neighborhood of the origin. These fixed points indicate the presence of elliptic periodic orbits contained in the homoclinic loops of the $O(1)$ fixed points.)

We now classify the non-trivial fixed points by studying the hamiltonian in a neighborhood of the fixed points. Below resonance, for the fixed points at $\theta = \pi/2$ and $\theta = 3\pi/2$, the hamiltonian is

$$K = \frac{\epsilon}{16}\left((24\omega_2\epsilon + 17\omega_1^2\epsilon + 24\omega_1)\cos 2\theta - 8\omega_2\epsilon - 19\omega_1^2\epsilon - 8\omega_1\right) J.$$

Above resonance, for the fixed points $\theta = 0$ and $\theta = \pi$, the hamiltonian is

$$K = -\frac{\epsilon}{16}\left((24\omega_2\epsilon + 17\omega_1^2\epsilon + 24\omega_1)\cos 2\theta + 8\omega_2\epsilon + 19\omega_1^2\epsilon + 8\omega_1\right) J.$$

Both translated hamiltonians represent saddle points. As $\omega \to \frac{1}{2}^-$ the saddle points move in toward the origin along the lines $\cos 2\theta = -1$. At $\omega = \frac{1}{2}$, the origin is saddle-like. As ω increases from 1/2 the saddle points move out from the origin on the lines $\cos 2\theta = 1$. Figure 1 shows Poincaré maps below, at, and above resonance. The Poincaré maps were generated by integrating the second–order equation eq.(16). Figure 2 shows the level curves of the reduced hamiltonian (17), plotted on the same scale as the numerically generated Poincaré maps.

7 A Bifurcation Between Resonances

Finally, we consider a bifurcation between two $O(\epsilon)$ resonances. We will use $O(\epsilon)$ Lie transforms to study how the Poincaré map changes as the amplitudes of two perturbations change. The equation to study is

$$\ddot{x} + (1 + \epsilon\omega_1)x + \epsilon(sx^2\cos(t) + (1-s)x^3\cos(2t)) = 0$$

for $0 \le s \le 1$. When $s = 0$, the quadratic term is absent and the cubic term is resonant. When $s = 1$, the quadratic term is resonant and the cubic term is absent. When $0 < s < 1$ both terms are resonant. The interaction of the two resonances is of interest.

Without loss of generality, set $\omega_1 = 1$. The resulting reduced hamiltonian is

$$K_1 = \frac{1}{2}J + \frac{\sqrt{2}s}{4}J^{3/2}\sin\theta - \frac{1-s}{4}J^2\cos 2\theta. \tag{18}$$

Fixed points satisfy

$$\begin{aligned} \frac{\partial K_1}{\partial J} &= 0, \\ \frac{\partial K_1}{\partial \theta} &= 0 \end{aligned}$$

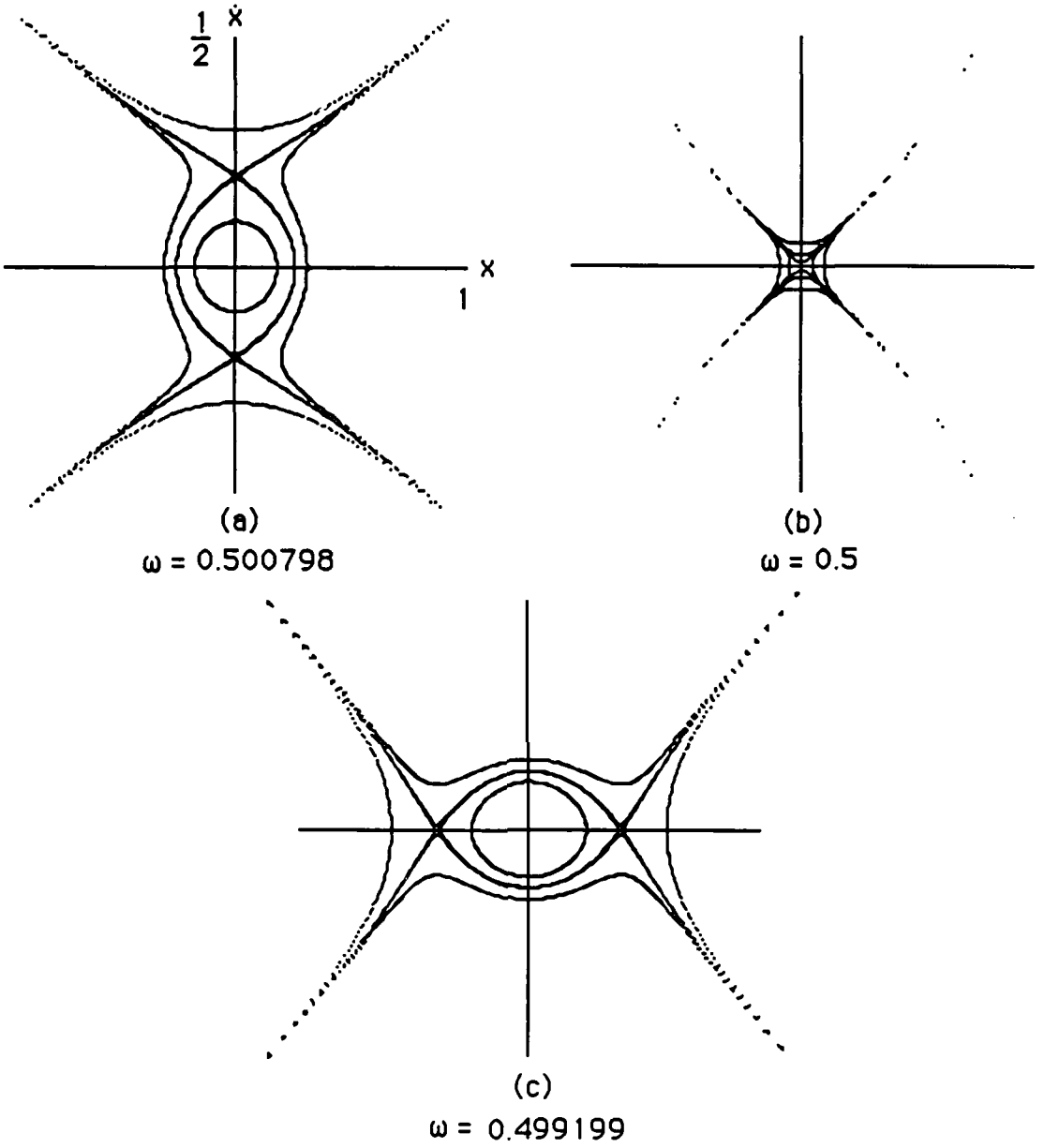


Figure 1
Numerically integrated Poincare maps (Σ:t=0 mod 2π)
of equation (16)

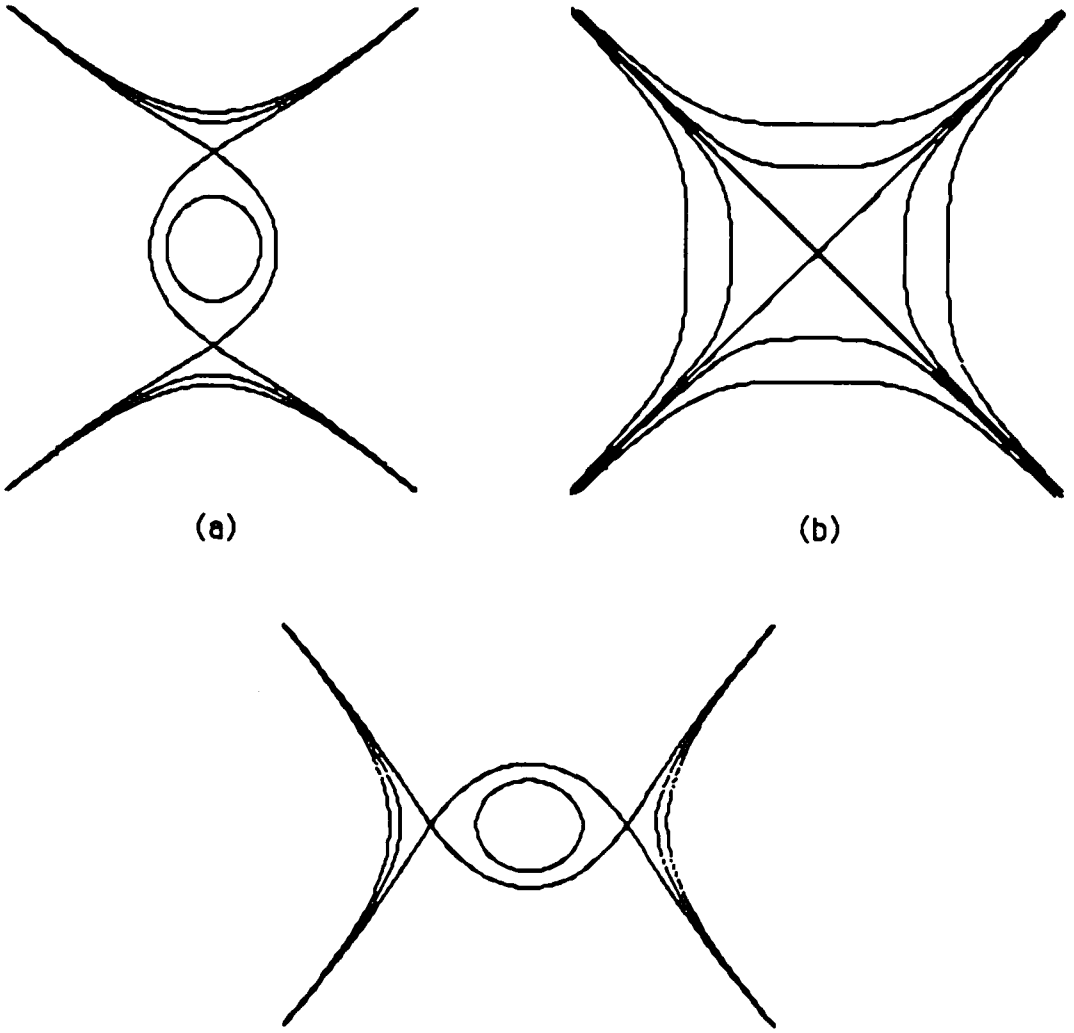


Figure 2

Level curves of the reduced $O(\varepsilon^2)$ hamiltonian (17) constructed by Lie transforms

which give the fixed points

$$\begin{aligned} \theta &= \frac{3\pi}{2}, \\ \sqrt{J} &= \frac{\sqrt{2}}{8(1-s)}(3s \pm \sqrt{9s^2+32s-32}) \end{aligned} \tag{19}$$

and

$$\begin{aligned} \sin\theta &= \frac{\sqrt{2}s}{4(s-1)\sqrt{J}}, \\ J &= \frac{8-8s-s^2}{8(1-s)^2}. \end{aligned} \tag{20}$$

The requirement that the radicand of eq.(19) be non-negative restricts s to the interval $0.8138 \leq s \leq 1$. The requirement that $|\sin\theta| \leq 1$ in eq.(20) restricts s to $0 \leq s \leq 0.828427$.

We now classify the stability of the fixed points. The stability of the critical point is characterized by the sign of the determinant of the Hessian $\mathbf{H}h$ evaluated at the fixed point. A rather lengthy computation (cf. [1]) shows that

$$\begin{aligned} \det(\mathbf{H}h(\mathbf{x}_0)) &= h_{qq}h_{pp} - (h_{qp})^2\,|_0 \\ &= h_{JJ}h_{\theta\theta} - (h_{J\theta})^2\,|_0. \end{aligned}$$

The critical point is a saddle if

$$h_{JJ}h_{\theta\theta} - (h_{J\theta})^2\,|_0 < 0$$

and is a center if

$$h_{JJ}h_{\theta\theta} - (h_{J\theta})^2\,|_0 > 0.$$

For the fixed points satisfying $\theta = 3\pi/2$,

$$(h_{JJ}h_{\theta\theta} - h_{J\theta}^2)|_0 < 0$$

gives the stability criterion

$$-\frac{J}{32}\left(16(1-s)^2J - 10\sqrt{2}s(1-s)\sqrt{J} + 3s^2\right) > 0$$

where, from eq.(19),

$$\sqrt{J} = \frac{\sqrt{2}}{8(1-s)}(3s \pm \sqrt{9s^2+32s-32}).$$

Substituting this into the inequality shows, after some algebra, that the limits of stable s are roots of the polynomial equation

$$s^4 + \frac{440}{21}s^3 + \frac{463}{9}s^2 - \frac{8576}{63}s + \frac{4096}{63} = 0.$$

The roots in the interval for which the fixed points exist are found to be $s = 0.818337$ (for the + root) and $s = 0.88871$ (for the − root). The stabilities of the fixed points for various s are listed in a table below.

For the other fixed points, the stability criterion is

$$-\frac{(1-s)^2J^2}{2} - \frac{(1-s)^2J^2}{2}\sin^2(2\theta) > 0$$

where

$$J = \frac{1}{1-s} - \frac{s^2}{8(1-s)^2}$$

and

$$\sqrt{J}\sin\theta = \frac{s\sqrt{2}}{4(s-1)}.$$

Inserting these relations gives the stability criterion

$$\frac{(s^2+4s-4)(s^2+8s-8)}{64(1-s)^2} < 0.$$

Since these fixed points fixed points exist for $0 \le s \le 0.828427$ and the inequality is not satisfied on this interval, these fixed points are always saddles. The behavior for various s is summarized below:

- At $s = 0$, saddles exist at $\sqrt{J} = 1$, $\theta = 0$ and $\sqrt{J} = 1$, $\theta = \pi$.
- As s increases, the saddles move into the left side of the plane and toward the horizontal axis.
- At $s = 0.8138$, a center appears at $\sqrt{J} = 2.31718$, $\theta = 3\pi/2$.
- As s increases, the centers separate, remaining on the horizontal axis.
- At $s = 0.818337$ the center farthest from the origin on the horizontal axis becomes a saddle.
- At $s = 2\sqrt{2} - 2$ the two saddles and the inner center coalesce and form a center.
- At $s = 0.88871$ the inner center becomes a saddle.
- As $s \to 1$ the inner saddle moves to $\sqrt{J} = \sqrt{8}/3$ and the outer one moves off to infinity.

Figure 3 shows the transitions for various values of s.

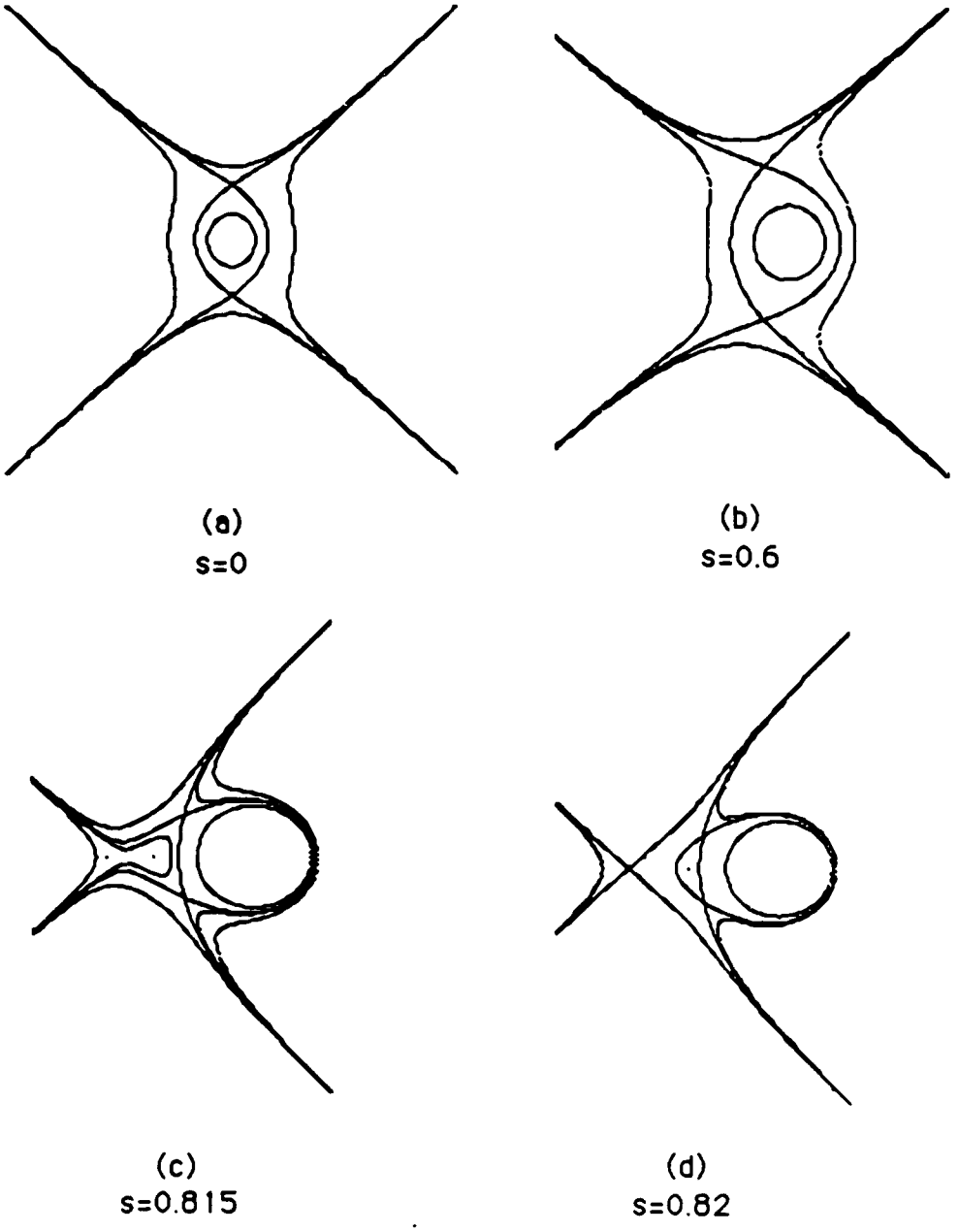
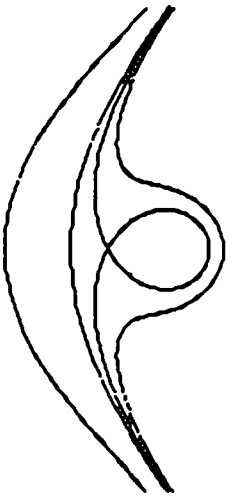
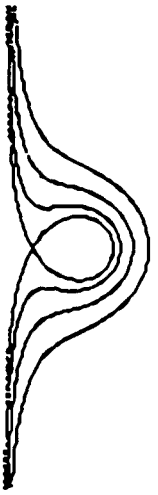


Figure 3

Level curves of the reduced hamiltonian (18) for indicated values of s.



(e)
s=0.9



(f)
s=1

Figure 3 (continued)

8 Appendix

The following MACSYMA program computes the lie transform of a weak perturbation of the simple harmonic oscillator.

```
/* Program to compute the lie transform near a resonance.*/
/*  If detuning is requested, the program will supply it */
/*in the form dw[i]*e^i. */
lie():=block
(
 kill(y,dw,n,j,dotran),
 assume(j>0),
 maperror:false,
 print(timedate()),
 trunc:read("Truncation order:"),
 om:read("Frequency"),
 f:read("Perturbation (use x, e, and t):"),
 detoon:read("Detune from resonance [y/n] ?"),
  if detoon = y then f:f+sum(e^i * dw[i],i,1,trunc)*x,
 dotran:read("Compute the co-ordinate
                              transformation [y/n] ?"),
 print("Equation to work with:"),
 print('diff(x,t,2) + om^2*x + f = 0),

/* Construct the Hamiltonian in complex */
/* slow-flow co-ordinates. */

 hh:map(pseudo_int_x,expand(exponentialize(f))),

/* Do the canonical change of co-ordinates to */
/* slow action-angle variables. */

 hh:ev(hh,x=%e^(%i*om*t)*q/(2*%i*om) - %e^(-%i*om*t)*p),
 hh:ev(hh,q=sqrt(2*j*om)*%e^(%i*th),
                  p=sqrt(2*j*om)/(2*%i*om)*%e^(-%i*th)),

/* Now taylor expand hh to order trunc */
/* and assign h[i] values. */
 tmp: expand(taylor(hh, e, 0, trunc)),
 for i from 0 thru trunc do
 (
  h[i] : coeff(tmp,e,i)
```

```
 ),


/* Initialize the new hamiltonian. */

 k[0] : h[0],

/* This loop does the transforms. */

 for n from 1 thru trunc do
 (
  print("Loop # ",n, "of ",trunc),
  temp: h[n] + sum(poisson(w[n-m],k[m]), m, 1, n-1)/n,
  temp: expand(temp + sum(m*inverse_evolution
                          (n-m, h[m]), m, 1, n-1)/n),
/* We don't need w[trunc] unless we are going to */
/* compute the net transformation. */
  if (dotran = y or n < trunc) then w[n]: getw(n,temp),
/* Cheat here.  w[n] was chosen to make k[n] the */
/* t-independent part of temp. */
  k[n]: map(nuke_t,temp)
 ),
/* The result is in new action-angle variables. */
/* Tell me what we got. */

 print(""),
 kk:sum(k[i]*e^i,i,0,trunc),
 kk:expand(rat(kk)),

/* Tell me all about the reduced hamiltonian. */

 print("The reduced hamiltonian in transformed
                              action-angle variables:"),
 realkk: expand(rat(realpart(kk))),
 print(realkk),

/* if requested, compute the co-ordinate transformation. */
 if dotran = y then
  block
   (

/* use the inverse evolution operator give the relation */
```

```
/* between old and new. */

   physical_j:sum(e^i*inverse_evolution(i,j),i,0,trunc),
   physical_th:sum(e^i*inverse_evolution(i,th),i,0,trunc),

   physical_j:expand(realpart(physical_j)),
   physical_th:expand(realpart(physical_th)),

/* Now tell me how big the transformation is: */

   print(""),
   print("The co-ordinate transformation has been
                                                computed."),
   print("length(physical_j)=",length(physical_j)),
   print("length(physical_th)=",length(physical_th))
   )
  else
   print("You told me not to compute the
                          co-ordinate transformation."),

/* Finished.  */

)$

/* Function to look like integration in x. */
/* This function is mapped. */
pseudo_int_x(f):=
(
 f*x/(hipow(f,x) + 1)
)$

/* function to compute poisson brackets in (th,j) space. */
poisson(f,g):=
(
 diff(f,th) * diff(g,j) - diff(f,j) * diff(g,th)
)$

/* function to nuke t-independent stuff. */
/* this function is mapped. */
nuke_no_t(f):=
(
 if freeof(t,f) then
```

```
  0
 else
  f
)$

/* function to nuke t-dependent stuff. */
/* this function is mapped. */
nuke_t(f):=
(
 if freeof(t,f) then
  f
 else
  0
)$

/* Function to compute generating function to */
/* nuke t-dependent terms.  This function is not mapped. */
getw(n,f):=
(
 [tmp],
 tmp:expand(-n*map(nuke_no_t,f)),
/* Factor the exponents in case an unspecified */
/* omega is given.  Note: lambda returns a list. */
 tmp:map(lambda([u],scanmap(factor,[u])),tmp),
 tmp:part(tmp,1),
 map(innegrate,tmp)
)$

/* Function to look like integration of complex */
/* exponential, hence the name. */
/* This function is mapped. */
 innegrate(f):=
(
 [nn,mm,tmp,z],
 matchdeclare([nn,mm],freeof(t)),

/* Define the pattern-matching rules for sines */
/* and cosines.  Note that the rules do not commute,*/
/* and ins must be performed before inc. */

 defrule(ins, sin(nn+mm*t),-cos(nn+mm*z)/mm),
 defrule(inc, cos(nn+mm*t),sin(nn+mm*t)/mm),
```

```
 tmp:expand(demoivre(f)),
 tmp:expand(applyb1(tmp,ins)),
 tmp:applyb1(map(nuke_no_t,tmp),inc)+map(nuke_t,tmp),
 tmp:ev(tmp,z=t),
 tmp:expand(exponentialize(tmp))
)$


/* Recursive function to compute kth term of */
/* inverse of evolution operator acting on h. */
inverse_evolution(k,h) :=
(
 if k = 0 then h
 else
 sum(poisson(w[k-m],inverse_evolution(m,h)), m, 0, k-1)/k
)$

/* Recursive function to compute kth term of evolution */
/* operator acting on h. Note that this function is not */
/* used by the program.  */
evolution(k,h) :=
(
 if k = 0 then h
 else
 -sum(evolution(m,poisson(w[k-m],h)),m,0,k-1)/k
)$
```

The following examples were run with the MACSYMA option "SHOWTIME:ALL" on a VAX 8500.

```
(c4) lie()$
Wed Jun 10 15:59:17 1987

Truncation order:
1;
Frequency
1/2;
Perturbation (use x, e, and t):
e*x^3*cos(t);
Detune from resonance [y/n] ?
y;
```

```
Compute the co-ordinate transformation [y/n] ?
y;

Equation to work with:

 2
d x             3             x
--- + e cos(t) x  + dw  e x + - = 0
  2                   1       4
dt

Loop #  1 of  1



The reduced hamiltonian in transformed action-angle
                                                variables:

             2
dw  e j - e j  cos(2 th)
  1

The co-ordinate transformation has been computed.

length(physical_j)= 5

length(physical_th)= 6

Totaltime= 52500 msec.  GCtime= 20116 msec.
```

Next, a run with a symbolic frequency to show how the resonant frequencies may be computed:

```
(c7) lie()$
Wed Jun 10 16:01:32 1987

Truncation order:
1;
Frequency
omega;
Perturbation (use x, e, and t):
e*x^3*cos(t);
Detune from resonance [y/n] ?
```

```
n;
Compute the co-ordinate transformation [y/n] ?
y;

Equation to work with:

 2
d x             3        2
--- + e cos(t) x  + omega  x = 0
  2
dt

Loop #  1 of  1



The reduced hamiltonian in transformed action-angle
                                              variables:

0



The co-ordinate transformation has been computed.

length(physical_j)= 5

length(physical_th)= 6

Totaltime= 72200 msec.  GCtime= 26500 msec.

(c8) factor(denom(rat(ev(w[1],t=0))));
Totaltime= 3250 msec.  GCtime= 1266 msec.

              2
(d8) 32 omega  (2 omega - 1) (2 omega + 1)
                  (4 omega - 1) (4 omega + 1)
```

The poles of the generating function at $O(\epsilon)$ are $\omega = 1/2$ and $\omega = 1/4$.

References

[1] J.L. Len,*Nonlinear Parametric Excitation With Averaging and Lie Transform Methods.* Ph.D. Thesis, Cornell University, 1987

[2] John R. Cary, *Lie Transform Perturbation Theory for Hamiltonian Systems.* Physics Reports (Review Section of Physics Letters) 79, No. 2(1981) 129-159

[3] Eugene J. Saletan and Alan H. Cromer, *Theoretical Mechanics.* John Wiley and Sons, 1971.

[4] Herbert Goldstein, *Classical Mechanics.* Addison-Wesley, Reading, Massachusetts, 1981.

[5] Leslie Anne Month and Richard H. Rand, *Bifurcation of 4:1 Subharmonics in the Nonlinear Mathieu Equation.* Mechanics Research Communications, Vol. 9(4), 233-240, 1982.

[6] Richard H. Rand, *Computer Algebra and Applications in Applied Mathematics.* Research Notes in Mathematics 94, Putnam, 1984.

[7] Richard H. Rand and Dieter Armbruster, *Perturbation Methods, Bifurcation Theory, and Computer Algebra.* Applied Mathematical Sciences, Volume 65, Springer-Verlag, New York, 1987.

Knowledge Representation and Planning Control in an Expert System for the Creative Design of Mechanisms

David A. Hoeltzel
Assistant Professor

Wei-Hua Chieng
Graduate Research Assistant

John Zissimides
Graduate Research Assistant

Laboratory for Intelligent Design
Department of Mechanical Engineering
Columbia University
New York, New York 10027

Abstract

An interactive system, referred to as MECXPERT {Mechanism Expert}, has been designed with the expressed purpose of assisting nonexpert design engineers in creating mechanisms for fulfilling specific motion-conversion and/or power-transmission requirements. The particular knowledge representation chosen for this application comprises a hybrid formulation of a rule-based production system with a frame-based approach. The underlying control strategy is based on a series of special-purpose, domain-specific operators whose function is to move from one problem space to another through various stages or "states" that comprise the mechanism design process.

The primary focus of this paper centers on the representation of knowledge and its control within an expert system for creative mechanism design. An overview summarizing the reasons for developing such an expert system is provided, and the formulation of a problem is discussed through an example taken from the design of a variable-stroke internal-combustion engine.

Introduction

The need for better and more nearly optimal and systematically designed mechanical devices in today's competitive world economy necessitates the development of expert systems. Capturing an expert's knowledge and heuristic skills in the performance of a domain specific task are the goals of an expert system. Such a system should assist less experienced engineers in producing better designs

Note: Bold, ***italicized*** works appear in the glossary in alphabetical order.

in a timely manner. Towards that end, an expert system for the creative design of mechanisms has been developed.

This paper discusses the manner in which knowledge is represented, manipulated and controlled in a mechanism design expert system, an important first step in the overall development of the system. Expandability, generality, system longevity and efficiency in expended effort were subjects of prime concern in developing the knowledge representation, with an eye toward long term committment to system improvement.

Historically there have been three approaches to the conceptual design of mechanisms: (1) the experience of a designer and/or layout draftsman, (2) the use of atlases or compendia of mechanisms, still the most widely used approach, and (3) the investigation of the kinematic structure of mechanisms. The second approach has been developed, most notably, by Jones et al. [1] and Artoboleskii [2] and makes for interesting and informative reading. The development of the third approach is more elusive, but holds remarkable promise for this most difficult phase of mechanical design, because of its systematic and unbiased nature. The expert system currently under development utilizes the later methodology as a basis for problem formulation and model development coupled with a heuristic approach for determining structure-function relationships for mechanisms as a basis for what we refer to as ***experience-based mechanism design.***

Few studies have made progress of any significance in developing expert systems for mechanism design. The work of Kota, Erdman and Riley [3,4], stands out as the most notable for applying expert system techniques to the design of dwell mechanisms. The authors reported on progress achieved on their system with an eye toward future development of a more general system whose purpose is to design mechanisms capable of generating straight lines, circular arcs, symmetric curves and parallel motion, in addition to dwell.

The overall process of ***creative mechanism design*** can be separated into a number of steps, some possessing considerable levels of difficulty and requiring significant long term mechanism design experience and intuition. These steps are depicted in the form of a flow chart in Figure 1. The creative design, i.e type and dimensional synthesis, of mechanisms is a complex task requiring ***deep domain knowledge*** as compared with the knowledge required to generate ***routine designs*** for fulfilling relatively simple or previously determined motion conversion requirements or to ***redesign***, through minor modifications, existing workable

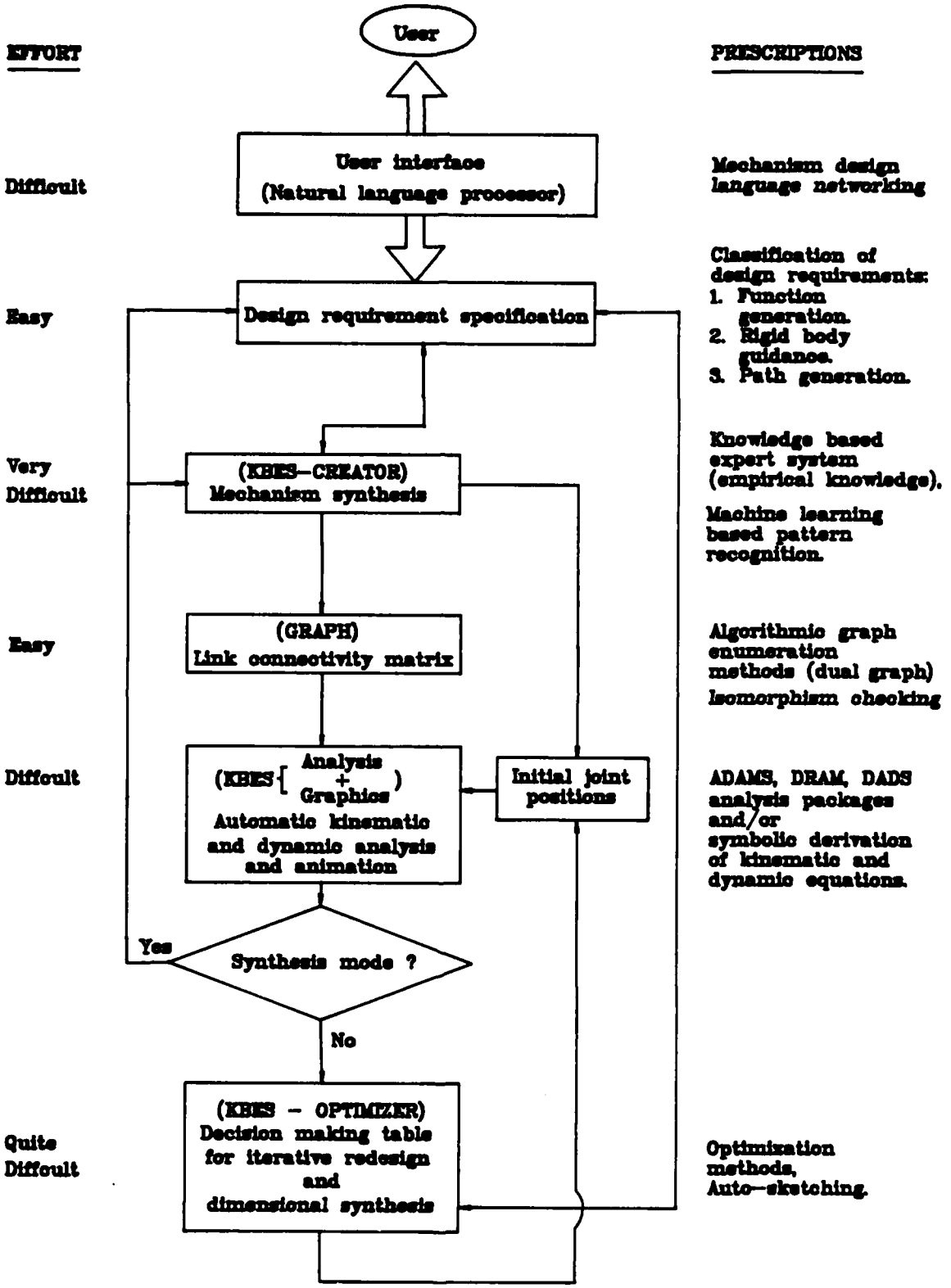


Figure 1. Overview of an expert system for creative mechanism design.

designs.

While the generation of numerical solutions corresponding to both the kinematics and dynamics of a known mechanism, as well as its animation represent relatively straightforward processes, creative mechanism design, in contrast, is extremely complex requiring, we believe, more of a heuristic approach particularly during the more conceptual phases of the mechanism design process.

The major obstacle to overcome in creative mechanism design centers around the determination of mechanism topologies (structures) for the fulfillment of specific design requirements, i.e. establishing a finite definable mapping between specified design requirements (functionality) and mechanism structure(s) capable of fulfilling the design requirements. Such a mapping will, in most cases, be one-to-many. Our work in this area centers around (1) the use of statistical machine learning for the cognitive recognition of characteristic motion patterns (functionality) associated with specific classes of mechanisms and the correlation of functionality (function generation, path generation, rigid body guidance) with structural features (links, joints and the manner in which they are connected) embodied in the mechanisms and (2) the development of a general vocabulary and "language" hierarchically structured consistent with the terminology of functional requirements and structural characteristics, through which a mechanism designer can establish a bi-directional channel of communication with the system in order to convey his functional requirements and receive feedback in a manner that is natural to mechanism design. This approach can be looked upon as a heuristic extension of the concept of the separation of kinematic structure and function conceived by Freudenstein [5].

Developing an Expert System for Mechanism Design

It is our contention that an expert system for mechanism design should act as an intelligent assistant and mentor, guiding the design engineer during the creative process of mechanism synthesis. Furthermore, the primary purpose of the system should be to fulfill user-specified predetermined motion-conversion or power-transmission requirements through the creation of an intelligent interactive environment with the mechanism design engineer.

With this idea in mind, the system under development has been fashioned around the concept of the separation of kinematic structure and function. Figure 2A depicts the essence of this subtle but important concept by means of an example. The functional requirements of the spatial slider crank mechanism (converts rotary motion into out-of-plane reciprocating motion) are provided as input-output

FUNCTION

Function generation

Rotary input ——→ Reciprocating output

STRUCTURE

Four links, four joints, input joint is R type, output joint is P type, one degree of freedom mechanism design.

MECHANISM

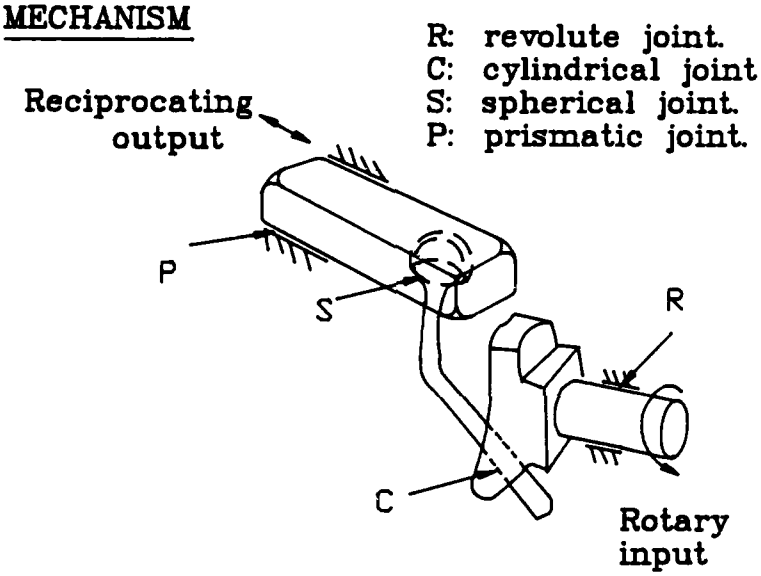


Figure 2a. Separation of kinematic structure from function in a spatial slider–crank mechanism.

GRAPH and ADJACENCY MATRIX

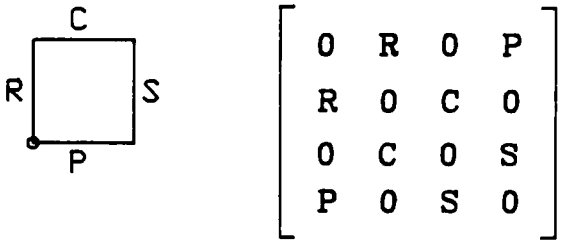


Figure 2b. Graph representation of the spatial slider–crank mechanism.

(functional) specifications by the user while the structural characteristics, i.e. those characteristics which will fulfill the functional requirements, are manifested in the actual physical embodiment of the mechanism, that is the number and type of links and joints and the manner in which they are interconnected.

The expert system system has been logically and hierarchically segmented into the following four subcomponents:

1. Specification of the desired kinematic structural characteristics and functional requirements.

2. Determination of the kinematic structure of all potentially useful mechanisms based on (1) the information provided in step 1 and (2) statistical machine learning for the cognitive matching of known kinematic topologies with the functional requirements the topologies are known to fulfill.

3. Screening of mechanisms according to their ability to fulfill both the functional and structural constraints.

4. Selection of the most favorable mechanism, i.e. the one(s) most nearly satisfying the constraints, for further development (analysis and animation).

As demonstrated by Dobrjanskyj and Freudenstein et al. [6] and Crossley [7], the kinematic structure of a mechanism can be conveniently, compactly and precisely represented, mathematically, using linear graph theory. Enumeration of the structure of mechanisms coupled with subsequent isomorphism checking for the elimination of duplicate mechanism structures using link connectivity matrices provides an efficient computational scheme for representing and sorting the kinematic structure of candidate mechanisms. Figure 2B depicts the graph representation and corresponding link connectivity matrix for a spatial slider crank mechanism. As previously mentioned, the correlation of kinematic structure(s) with predefined functional requirements, for large classes of mechanisms, represents the primary bottleneck to creative mechanism design and is the area in which much work remains to be done.

Remaining within the bounds of a limited domain is the natural and most logical course of action to be adopted in any new expert system development endeavor, particularly in a domain as complex as mechanism design. As a result of this, the time tested incremental approach to software design has been utilized [8]. In this approach a top-down building-block strategy is employed whereby modular pieces of the system are configured, keeping the overall system configuration in mind to avoid costly redevelopment, and incrementally tested to insure correct results. The system is presently limited to planar mechanisms having kinematic pairs with a

maximum of two degrees-of-freedom.

Finally, in designing a system for general applicability it is imperative that a test problem be selected that reflects the complexity of the domain (mechanism design) without including excessive detail or problem size which would unnecessarily and unavoidably complicate program verification and performance evaluation. For this reason a test problem fashioned around the design of a variable-stroke internal-combustion engine, which has been previously solved in detail by Freudenstein and Maki [9], has been selected. The designer's reasoning processes in making problem specific design decisions have been explicitly verbalized in their paper.

Software Implementation Issues

In its present state, the MECXPERT system has been implemented using the OPS5 production system programming language [10] embedded within the Knowledge Craft expert system development environment [11]. Programs developed in the OPS5 language are composed of data-sensitive unordered rules, where the data can be (1) ***instances*** of physical objects, (2) facts related to the domain of application and (3) conceptual objects (such as goals) related to the problem-solving strategy. The rules that constitute the program are composed of two parts. The first is the condition part and consists of data elements. The second part of a rule is the action part and is composed of instructions that change the current data configuration.

Program execution occurs in "cycles" in which each cycle consists of three actions: match rules, select matching rules and execute selected rule. A rule can be executed only if all the data elements in its condition part match the current data configuration. OPS5 provides two possible strategies, lexicographic ordering (LEX) and means-ends-analysis (MEA) for selecting the rule to be fired when more than one rule is applicable. In this case the MEA conflict resolution strategy was selected because it places additional emphasis on the recency of the ***working memory element*** that matches the first condition element of a rule. In this way, when the first condition element of a rule is a goal element, the system will not be distracted by a very recent working element that is not a goal (i.e. goal driven). Thus the data configuration changes after every cycle is completed, except the final one. For this reason the system can be said to use a ***data-driven inference strategy.***

In this system, a ***goal-driven inference strategy*** is inappropriate due to the fact that in the process of mechanism synthesis, the final mechanism topologies suitable for prescribed motion conversion or power transmission are not known apriori, but are to be uncovered through the interactive design process embedded within the expert

system.

The structure of the software has been developed in accordance with the requirements specified by the domain by developing data structures that insure the creation of a planning strategy capable of simulating mechanism design procedures and emulating human thought processes which occur during mechanism design as these would be performed within these procedures. The procedures use various types of knowledge to implement appropriate reasoning schemes. It is therefore of extreme importance to effectively represent knowledge and simulate planning, since these two functions determine the path that the design undergoes and whether or not all facets of the design process are properly taken into account.

Building a Model for Knowledge-Based Mechanism Design

In order to develop a formal description, i.e. model, of the mechanism synthesis problem it is necessary to:

1. Define a state space representation containing all the possible configurations of the relevant parts of the problem, without necessarily enumerating, in detail, all these states. In fact in mechanism design this represents an impractical task due to the NP-completeness nature of the problem, i.e. exponential time complexity growth rate. For example, the graphs corresponding to a planar six bar mechanism represent an upper limit of $O(10^4)$ unique kinematic structures, while those for a planar eight bar mechanism represent an upper limit of $O(10^{10})$ unique kinematic structures. This later number of possible mechanism structures is too large to undergo detailed development or in-depth evaluation given the present level of readily available engineering computing power. This presupposes that the problem is decomposable. We have found that it is.

2. Specify one or more states within the space representing possible situations from which the problem-solving process may start. These are the initial states.

3. Specify a set of rules describing the operations which permit movement through the space from its initial state to its goal state.

4. Specify acceptable solution or goal states to the problem. In mechanism design information of this nature would be provided to the system via user input in the form of (1) an input-output function to be generated, (2) a description of position and orientation of a rigid body to be guided, (3) a path to be generated through a finite number of points by a point on the coupler link of the mechanism or (4) as a power transmission or energy conversion requirement.

Each of the above listed parts of an overall system model, as they specifically relate to mechanism design, are discussed in the following sections.

Data Structures

Appropriately designed data structures are the means by which planning and knowledge representation can be effectively implemented in an expert system. The problem domain, in this case mechanism design, is represented or broken down, hierarchically, into ***problem-spaces*** **{PS}**. These {PS}'s represent states that the system can reside in and pass through in its effort to achieve its goal. Thus, the system can be imagined to emulate the mechanism designer's thought processes, where the current {PS} represents the issue or concept under consideration. After reaching a given state the system must choose the next state to which it will move. To achieve this a data structure element, referred to as a ***sub-problem*** **{SP}**, has been created to indicate to the current {PS} what the next available states, i.e. {PS}'s, will be. Therefore, within each {PS} there are {SP}'s which represent potential compatible {PS}'s to which the system can move. It should be emphasized that the term sub-problem is a relative one in the sense that its describes the next possible problem-space, {PS}, to which the system may move from the current problem-space, {PS}, thereby establishing a parent-child relationship between the two.

Operators, {OP}, whose function is to decide what the next {PS} is to be, based on the current status of the design, are present within all {PS}'s. A data structure, referred to as "***compatibility***", defines the compatible {PS}'s and {SP}'s and is created each time the system is initialized. In essence this data structure establishes the fixed (common for different design goals) graphical tree structure [12] that represents the domain of mechanism design (Figure 4), to the extent that it is represented in this model. This structural representation is possible since the design domain can be hierarchically subdivided.

The hierarchical nature of the mechanism design domain has be schematically illustrated in Figure 3. Details have been intentionally omitted at this point in the discussion in order to avoid confusion, however following sections will elaborate the details of the system structure specific to mechanism design. When the current {PS} is "1", the available {SP}'s will be "2", "3" and "4". If {SP} "3" is selected then the current {PS} becomes "3" and the subsequent {SP}'s are "7" and "8". Four different defining characteristics can be associated with each problem space. A {PS} is said to be "**complex**" when in order to be solved it has to be broken down into other {PS}'s (e.g. {PS}'s "1", "3", "4" and "9"). A {PS} is said to be "**simple**" when in order to be solved only a few predetermined steps (actions) are necessary (e.g.

{PS}'s "2", "7", "8", "10", "15", and "16"). Note that in Figure 3, circles are used to schematically represent steps rooted in simple {PS}'s. Therefore, a {PS} can be solved either by successfully executing a predefined number of steps (actions) such as eliminate prismatic (p-type) joints or increase the current number of independent loops, L_{ind}, by one, or by solving all the required {SP}'s (defined within each {PS} by an operator whose role is to evaluate the state of the current {PS}) that are rooted in the current {PS}. When a {PS} is successfully achieved, then its status is said to have been "**achieved**". When a {PS} has not been successfully achieved, because its rooted {SP}'s and/or steps are still being processed, its status is said to be "**pending**", otherwise its status is said to be "**failed**" (within each {PS} there is knowledge that is used by an operator, called the evaluate-state operator, to determine the status of the {PS}).

A {PS} is said to be "**fixed**" when the choice of {SP}'s is independent of the current design assignment specified by the user, and thus knowledge can be given to the system to reject all but a single {SP}. Finally, a {PS} is said to be "**probabilistic**" when the selection of a {SP} depends on its probability of success. The probabilistic {SP}'s are independent of one another and each carries a weighting factor that depends on problem specifications which are defined by the user during the "problem definition" (entry of input data) phase of mechanism design.

In Figures 3 and 4 probabilistic problem-spaces are denoted by rooted {SP}'s which are interconnected with their respective parent {PS}'s by dotted lines and fixed problem spaces are denoted by rooted {SP}'s that are interconnected with their respective parent {PS}'s by solid lines.

Making reference to Figure 3 it can be seen that when the current {PS} is "1", and if and only if {PS} "1" is fixed, then the following knowledge can be built (hard coded) into the system: If the status of all {SP}'s directly beneath {PS} "1" are pending then reject all {SP}'s except "2". If the status of "2" is achieved and the status of the remaining {SP}'s on the same level are pending, then select "4" by rejecting the other pending {SP}'s on this level, etc. This process of rejecting {SP}'s is realized through the use of a "reject" operator that is active in every {PS}. In the first case of the above example, the "reject" operator would reject {SP}'s "3" and "4" and in the second case it would reject {SP} "3". Each time control returns to a {PS} from a lower level {SP} (either because it has been achieved or failed) the status of {SP}'s corresponding to the current {PS} that have not been failed or achieved are set from "rejected" to "pending" so that they will be available when appropriate and necessary.

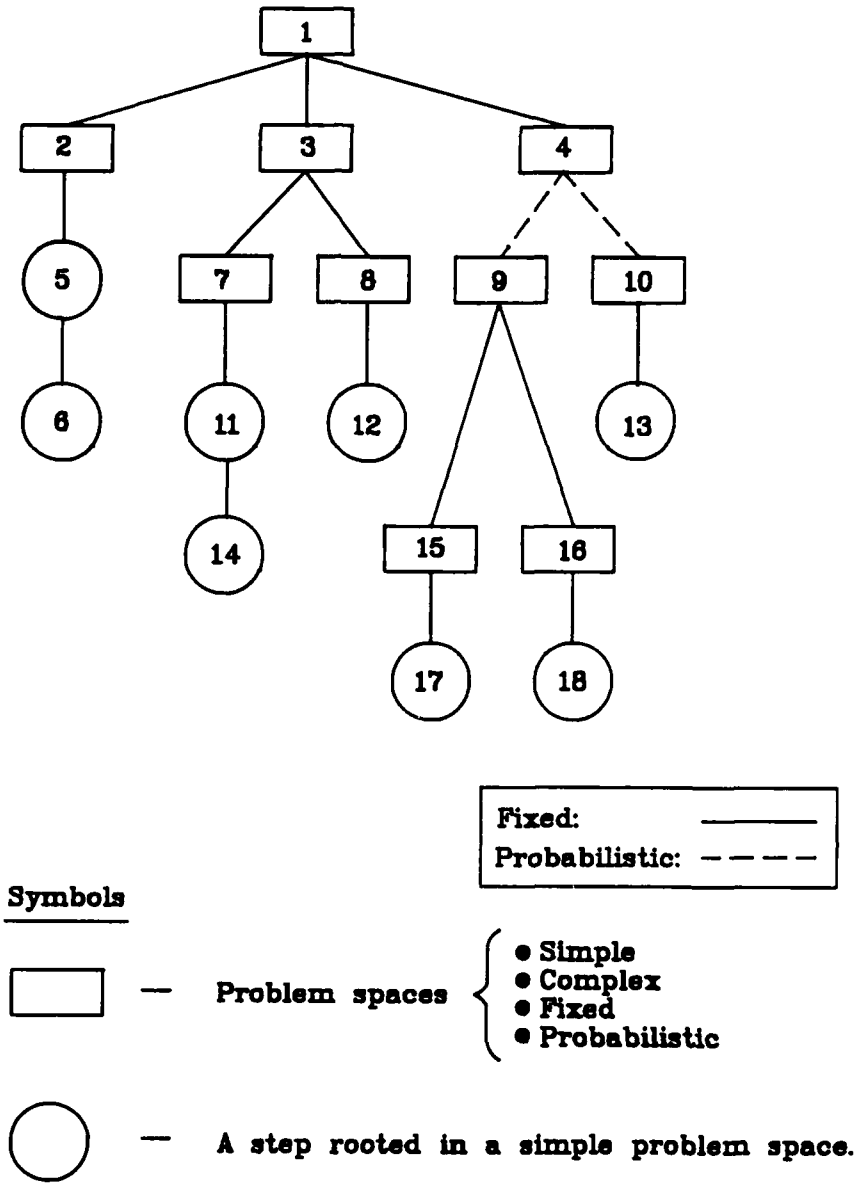


Figure 3. Hierarchical skeleton data structure upon which the knowledge representation and planning scheme have been based.

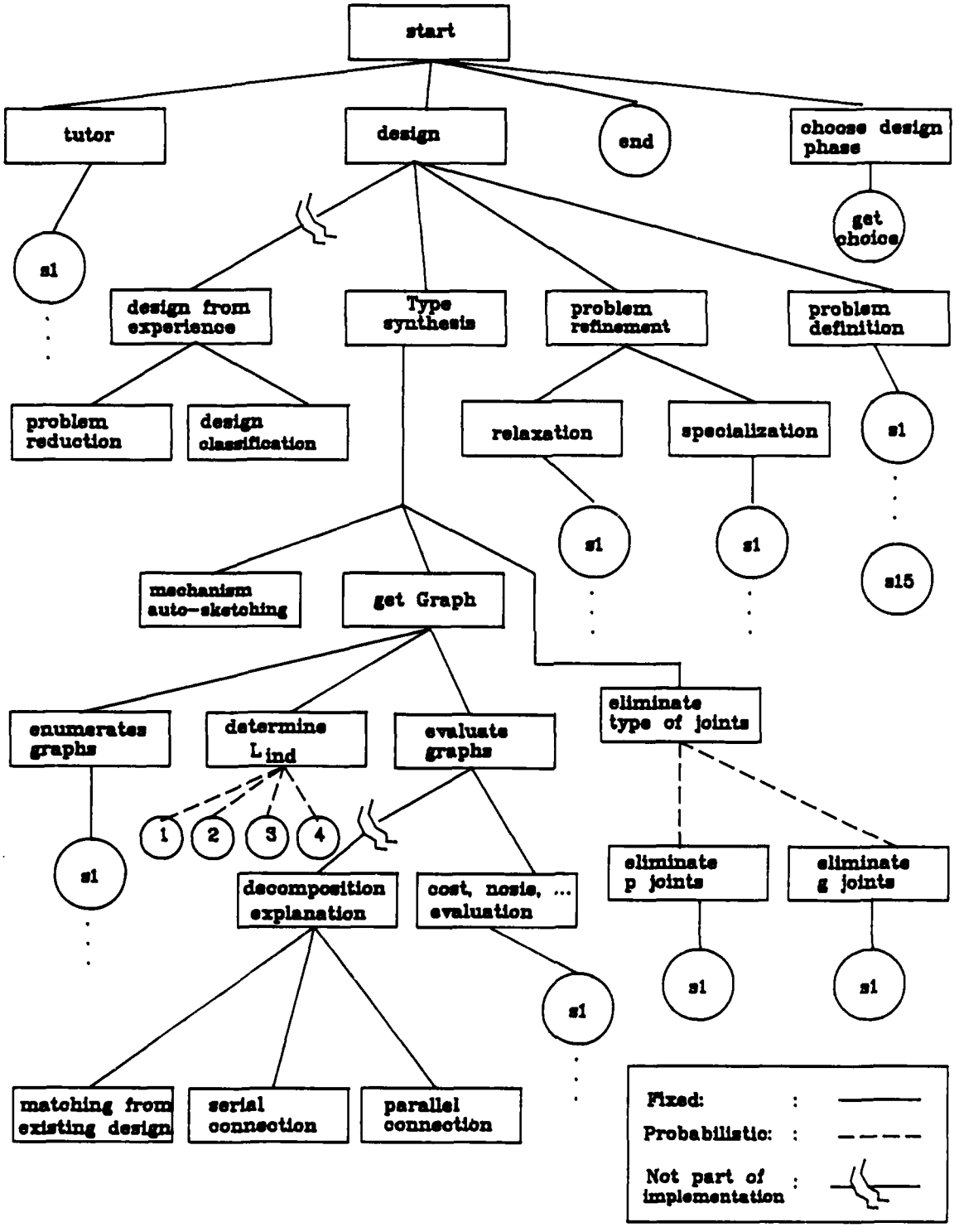


Figure 4. Hierarchically structured knowledge representation and planning scheme used for mechanism design.

The primary function of the data structures described above is the computer implementation of a systematic methodology for mechanism design. The effectiveness of the expert system depends, to a large extent, on the manner in which explicit knowledge about the process of mechanism design can be built into the data structure.

Knowledge Representation Strategy

The corpus of knowledge contained within the expert system is discretized and partitioned into problem spaces, {PS}'s, through the use of operators, {OP}'s, possessing certain predefined ***knowledge roles.*** These operators are the means through which a planning strategy has been imparted to the system. The operators and their associated functions have been created as the means through which traversal, from one problem space to another, can systematically and consistently proceed within the system. Within each {PS} the operators and their associated knowledge roles are defined as follows (Figure 5).

Operator Definitions:

1. Propose operator:

   All {SP}'s applicable to the current problem-space, {PS}, have their status set to "pending". The status of these {SP}'s are determined by the propose operator whose function is to check all the {PS}'s for potentially compatible {SP}'s.

2. Evaluate constraints operator:

   In this stage only {PS}'s that are probabilistic exist. Under this operator the expert places knowledge that checks the user's input and the current status of the design. It assigns the appropriate ***weighting factors***, (wt), indicative of the contribution that each of the constraints should make to the rooted {SP}'s.

3. Assign probability of success operator:

   A probability of success, p(s), is assigned to each of the {SP}'s by averaging the weighting factors that have been determined under the evaluate constraints operator.

4. Evaluate state operator:

   When the current step of the data structure is "evaluate state", the system will attempt to match the current configuration of the data with condition elements that if present in working memory, would indicate failure or success of the current {PS}. If such a matching occurs then control, unless otherwise specified by the "***failure handler***", returns to the parent {PS}.

5. Reject operator:

   {SP}'s that have been proposed by the propose operator but which are forbidden due to the presence of an appropriate piece of knowledge are rejected by this operator. This operator is used in fixed {PS}'s to reject all but one {SP}.

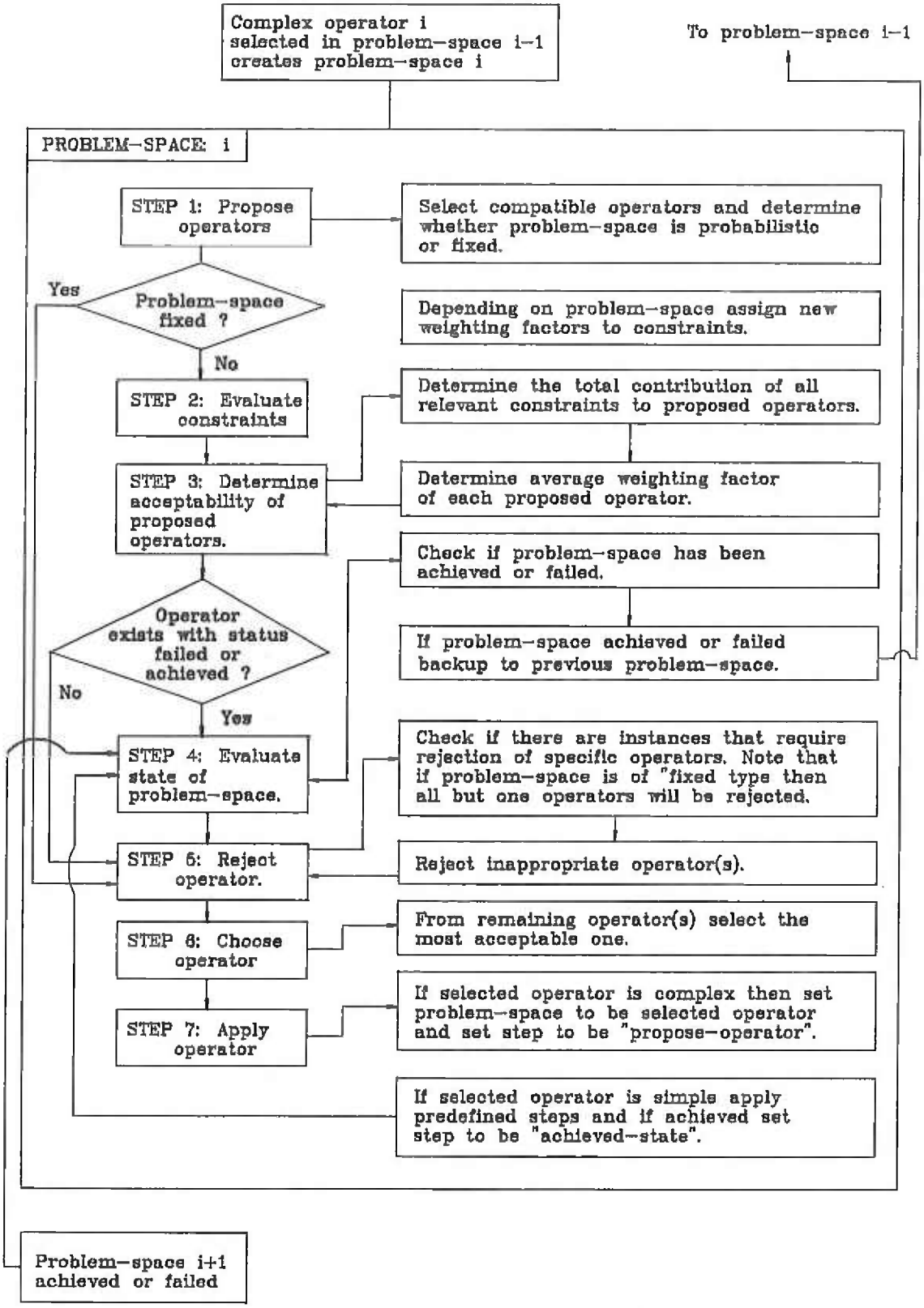


Figure 5. Knowledge roles (operators) defined within the i^{th} problem–space are partitioned into seven steps.

6. Choose operator:

   In probabilistic {PS}'s the {SP} with the highest probability of success (i.e. > cutoff value) is chosen. This is true for the eliminate-joints {PS} where more than one {SP} might have a weighting factor > cutoff value and thus more than one {SP} can be selected (more than one joint type rejected). In other probabilistic {PS}'s such as "get L_{ind}-min" only one {SP} must be selected. Therefore, after the one {SP} with the highest weighting factor is selected, applied and achieved, the evaluate-state operator should set the current {PS} to an achieved status, and control will then return to {PS} get-graph. In this example, what was described happens because when a {SP} is achieved the operator under control is evaluate-state. In fixed {PS}'s, the single {SP} with a pending status will be selected.

7. Apply operator:

   The chosen {SP} is applied by the apply operator. If it is complex then the {SP} becomes the current {PS}, otherwise the predefined steps of the appropriate simple {PS} are executed.

Figure 5 depicts the relationship of these operators within the ith problem space.

Attributes that are descriptive of the structure and function of mechanisms such as the number and types of links and joints, degrees-of-freedom, number of independent loops, etc., and knowledge that describes these quantities is represented in the form of hierarchical frames [13]. Frames make it possible to readily represent objects hierarchically and to simplify their communication control structure. The following is a frame-based knowledge representation of an atlas of graphs, written in the OPS5 language, corresponding to the structure of mechanisms. These representations can change and expand as the system grows.

```
(ATLAS   ^graph-id <a>                    ; This is an atlas of graphs
                                          ; each of which has a unique
         ^degree-of-freedom <b>)          ; id number <a> and a dof <b>
                                          ; associated with it.

(GRAPH   ^id  <a>                         ; Each graph has structural
                                          ; variables associated with it.
         ^#-of-independent-loops  <b>     ; Number of independent loops.
         ^ground-link     <d>             ; Indicates which link is grounded.
         ^input-link      <e>             ; Indicates which link is the input link.
         ^output-link     <f>             ; Indicates which link is the output link.
         ^#-of-P-joints   <p>             ; Number of prismatic joints.
         ^#-of-R-joints   <r>             ; Number of revolute joints.
         ^#-of-G-joints   <g>             ; Number of gear joints.
         ^status          <n>)            ; Can be pending, selected or rejected.
```

```
(ADJACENCY-MATRIX        ; Each graph, <a>, has an adjacency
                         ; matrix, <array-id>, associated with it.
        ^graph-id   <a>
        ^matrix     <array-id>)

(MECHANISM               ; The values of a mechanisms attributes,
                         ; given below, represent the current design
                         ; status. These values are used to determine
                         ; the graphs to be enumerated and
                         ; evaluated.
        ^dof    <a>      ; F is specified by the user-input to the
                         ; question, # of inputs and # of outputs.
        ^Lind   <b> )    ; This is defined by the knowledge found in
                         ; "get current Lind" (Figure 2).

(JOINTS     ^type       <a> ; Any of type R, P or G.
            ^max-#1     <b> ; A problem-space must be created at
                            ; the appropriate level which determines
                            ; the maximum number of <a> joints to
                            ; be used within any loop.
            ^max-#2     <c> ; Again, a {PS} must be created that will
                            ; have knowledge of the maximum number
                            ; of <a> joints to be used in the current
                            ; mechanism design.
            ^status     <d> ; Either rejected or pending, which is
                            ; determined in the eliminate-type of
                            ; joints problem space.
            ^wt         <e> ) ; The weighting value assigned to the {SP}
                              ; in the eliminate-type of joints problem-
                              ; space. To be used later (in case status is
                              ; pending) to help evaluate the graphs that
                              ; use joint <a>.
```

Note that letters in triangular brackets represent the values of variables associated with the structure of mechanisms. This data structure stores knowledge about a complex element (in this case mechanism structure) in a hierarchical format and communicates it through the use of an identification (id) number.

In addition to hierarchical frames, production rules serve as a second method of representing knowledge, a description of which has been provided under the section of this paper entitled Software Implementation Issues.

Systematic Planning for the Control of Knowledge

The strategy used to control knowledge is systematic in the sense that it guarantees the generation of a solution if one exists while avoiding the possibility of

repetituous computations or enumerations of mechanism structures [14] for efficiency.

The planning strategy incorporated within the system has been designed to address the following issues:

How is the next problem-space {PS} chosen when:

1. The current {PS} has to be broken down into {SP}'s in order to be solved.
2. The current {PS} has been solved successfully.
3. The current {PS} cannot be solved (i.e., is assigned a failed status).
4. A solution to the current {PS} is not acceptable at other levels of the knowledge representation hierarchy.

The first question refers to the concept of a complex {PS}. If, in addition to being complex, the current {PS} happens to be fixed then as already mentioned there must be predefined knowledge resident in the system to indicate what the next allowable {PS} will be (i.e., a predefined course of action). User specified input is assigned (1) weighting values over the range of values of 0 through 1 in decimal increments corresponding to their relative importance and (2) "degree of compatibility" values corresponding to how compatible they are considered to be with a given {PS} or {SP}.

If the current {PS} happens to be probabilistic, in addition to being complex, then selection of the next {PS} will depend on weighting factors associated with (carried by) the next level of {SP}'s. The overall probability of success of a given {SP} depends on (1) the values of weighting factors assigned by the user for input that is compatible with the {SP}, (2) the number of inputs that are compatible with the {SP}, (3) the degree of compatibility of the {SP} with the current {PS}, and (4) the probability of success, p(s), of the {SP} in the current {PS} (the last two compatibilities are defined by a domain expert). The possibility of dependence of a {SP} choice on the successes or failures of previous {SP}'s and/or {PS}'s is also taken into consideration. It can be seen that inputs provided by the user, representing specifications that must be satisfied up to a desired predefined degree of compatibility, are used to appropriately constrain or trace the path taken by the design process.

The following example demonstrates, in a simplified manner, how ***constraint propagation*** has been implemented. Referring to Figure 3, {PS} "4" is defined to be probabilistic. Inputs "a", "b", and "c" are defined to be compatible with {SP} "9", and inputs "a", "c", and "d" are defined to be compatible with {SP} "10". Furthermore, it

is assumed that the domain expert has assigned the following degree-of-compatibility values for inputs a, b and c with respect to {PS} "9": $doc_{a/9}$ = .7, $doc_{b/9}$ = .8, $doc_{c/9}$ = .9, and for inputs a, b and d with respect to {PS} "10": $doc_{a/10}$ = .8, $doc_{b/10}$ = .7 and $doc_{d/10}$ = .9. It is also assumed that the user has entered the following weighting factors for the inputs:

$$wt_a: 0.8,\ wt_b: 0.6,\ wt_c: 0.3, \text{ and } wt_d: 0.9$$

Based on the above degree of compatibility values and weighting factor values, the weighting factors for {SP}'s "9" and "10" will be respectively:

$$\{SP\}\ "9"\ : ((0.7 * 0.8) + (0.8 * 0.6) + (0.9 * 0.3)) / 3 = .4367$$
$$\{SP\}\ "10": ((0.8 * 0.8) + (0.7 * 0.6) + (0.9 * 0.9)) / 3 = .6233$$

Thus, in this case {SP} "10" would have been selected if there was no knowledge under step "8" (Figure 3) that would forbid the selection of {SP} "10". Note also that a "***cutoff value***" has been established for each probabilistic {PS}. In order for a {SP} to be selected it must acquire a composite weighting factor value that is higher than the cutoff value assigned to the current {PS}.

Finally, if the {PS} selected is simple and if it has been achieved then control returns to the parent {PS}. In the event of failure, if recovery is possible the ***failure handler*** will take over, otherwise control is returned to the parent {PS} and the failed {PS} will be assigned a "failed" status. As was previously discussed, in every {PS} there is knowledge about whether the {PS} has been achieved, failed or pending embedded within internal {SP}'s. The failure handler will only take over when the failure occurs at a simple {PS}. This is because only then is it possible to precisely recommend a specific plan of action for recovery. The failure handler, when activated, keeps track of the {PS} where failure has occured and when its role is completed. In this way, the design process will be able to resume at the point that it stopped. The failure handler, when activated, will go to the {PS} that is recommended within the simple {SP} where the failure has occured and a "parallel" process will take place until the design is re-established at a desired status. The control will then return to the failed simple {PS}.

Defining a Mechanism Design Problem Within MECXPERT (Problem Definition Phase)

The system, as previously discussed, is broadly based on the concept of the separation of kinematic structure from function. System functions have been subdivided into three major phases including (1) problem definition, (2) type synthesis, and (3) dimensional synthesis. Also included are utility modules for

automatic sketching with animation and kinematic and dynamic analyses (Figure 1). These utilities provide feedback to the system and designer as to the applicability of the heuristically chosen mechanism. If necessary an iterative design procedure can be instantiated or an alternative design may be chosen by making appropriate changes to the specified design specifications and constraints to be satisfied.

The system must first acquire knowledge about the problem through a knowledge acquisition facility. In this stage of operation the system attempts to acquire as much information as possible from the user so that the design can be constrained and the domain pruned. The only required information is the number of inputs and the number of outputs. However, from a practical standpoint additional information must be specified in order to narrow the number of alternative or candidate mechanisms to be further studied. The additional information is acquired by the system in a hierarchical manner so that only relevant questions need be asked.

Most questions require that a weighting factor be specified in the integer range of zero to one, in decimal increments, corresponding to a certainty factor. If the term "**explain**" is entered at any time during a user session, instead of the required answer, a ***help facility*** will provide a detailed explanation of the current question.

At this stage the system will attempt to narrow down the mechanism design search space (domain) as much as possible by identifying design goals that can be approached in a more specific way. This is necessary since the number of potential mechanisms for different design specifications obtained from the heuristic rules employed, for a general design case, would most likely be unreasonably large. This inefficiency is a result of the lack of knowledge about how the different specific domains and sub-domains (represented as {PS}'s in Figure 2) that constitute the general design domain relate to general concepts and of course computer-based limitations (memory and speed).

The following is a list of representative system querys requiring user input:

1. Enter the # of inputs.
2. Enter the # of outputs.
3. Enter the type of mechanism.
4. Enter the name of the function to be generated (ex. straight line motion).
5. (Questions that will specify the task of each output):
   a.Order of the path traced by each of the outputs.
   b.Output link(s) must be connected to a prismatic joint (wt. 0->1).
   c.Which outputs must be grounded (wt. 0->1).
6. Which input(s) must be grounded (wt. 0->1).

7. Which input(s) must be sliders (wt. 0->1).
8. Will there be a control or guidance function within the mechanism (wt. 0->1).
9. Enter the maximum number of links, ℓ_{max}.
10. Enter the minimum number of independent loops, $L_{ind,min}$.
11. Low cost design (wt. 0->1).
12. Reliable design (wt. 0->1).
13. Ease of manufacturability (wt. 0->1).
14. Speed of mechanism (wt. 0->1).
15. Load (wt. 0->1).

These inputs will be used to constrain the design domain by means of the constraint propagation method described earlier.

The following is a partial listing of the rules that will be used by the "evaluate-constraints" operator within the "graph-evaluation" problem space (Figure 4):

Rule-1. If the mechanism is a path generator then the output link must be a floating link.

Rule-2. If the mechanism is a function generator then the output link must be in contact with ground.

Rule-3. If there are more than two slider joints in any single loop then the topology is invalid.

Rule-4. If there is a need for a guidance or control loop then the output link should not belong to the loop that contains the input. This implies the need for $L_{ind,min} \geq 2$.

Rule-5. The total number of independent loops cannot be less than (the required number of links which are adjacent with the ground link) - 1.

Rule-6. For the purpose of simplifying the analysis phase of mechanism design, mechanisms containing at least one independent loop enclosed by {3 + (total number of dof's)} links should be selected for evaluation prior to those mechanisms which do not satisfy this rule.

These rules will assist in the process of pinning down the kinematic structural parameters during execution of the "get-graph" {PS}. Also, user input related to load, speed, noise level, cost, reliability and manufacturability considerations are used by the "eliminate-joint-types" {PS}, Figure 4, to reject or assign preferences for the different available joint-types.

When the problem definition (data input) phase has been completed the planning strategy imbedded within the MECXPERT system chooses the next design phase, either type synthesis or dimensional synthesis.

Type Synthesis Phase of Mechanism Design within MECXPERT

At this level two independent problem-spaces, {PS}'s, are available:

1. Modification of an existing design (Iterative Redesign):

This level assumes the existence of a known mechanism topology in order to fulfill the user specified design requirements. An iterative redesign procedure is initiated where changes can be made to structural characteristics (link lengths) of a known mechanism in order to move the existing design closer to the required design in an incremental fashion. After each change is made to the mechanism, animation and, if desired, dynamic analysis, are performed in order to assess the effect of the changes on nearing the desired mechanism functional requirements.

2. Systematic type synthesis:

In this problem space, {PS}, the system will first compute the number of links and joints for the simplest possible mechanism, i.e the one having the minimum allowable value for Lind. This is because the goal is to satisfy the functional requirements in the simplest way possible. It will then choose the appropriate non-isomorphic graphs of kinematic chains from an atlas stored in the database. Finally, all possible combinations for the ground link, the inputs and outputs and the types of joints will be systematically enumerated from the non-isomorphic graphs. After this, as shown in Figure 4, the next step will be "graph-evaluation". Heuristics will be used to assign a weighting factor to each of the graphs. Graphs with weighting factors greater than 0.5 (arbitrarily chosen, but tuning of this parameter may be required) will have a chance to continue on into the analysis phase, where the graphs will be examined in accordance with their priority as indicated by the weighting factors.

Dimensional Synthesis Phase of Mechanism Design within MECXPERT

The dimensional synthesis phase of MECXPERT has been subdivided into two problem-spaces, {PS}'s:

1. Automatic Sketching:

The graph representation restrains the link connectivity in mechanism design. However, a mechanism has to be uniquely defined not only by its link connectivity but also by its physical dimensions. The technique which applies default link lengths and orientations to the graph-to-mechanism conversion problem is usually referred to as the automatic sketching of mechanisms. In addition to the default dimensions, link lengths and orientations, default constraints associated with mechanism geometry

must also be specified. This includes (1) arbitrarily assigning a joint position to be coincident with the origin of the selected coordinate system and (2) arbitrarily assigning a horizontal link (this is usually the gound link). In accordance with the the number of degrees-of-freedom possessed by the mechansism, additional constraints must be specified, equal to the number of degree-of-freedom. These additional constraints are referred to as pseudo-constraints, and they determine the initial position of the mechanism. In order to satisfy all the constraints for automatic mechanism sketching, a Newton-Raphson iteration scheme has been adopted.

2. Mechanism Animation and Automatic Kinematic Analysis:

The concept of the loop closure equation, referred to as the Freudenstein equation, can be expanded for solving the kinematics of multiple loop mechanisms. As a result, by applying this new equation solving strategy, a computationally efficient divide and conquer algorithm has been developed to generate closed form solutions for 97% of all planar eight-link planar mechanisms and 56% of all ten-link mechanisms requiring only seconds of cpu time. This approach can greatly expedite the analysis phase of mechanism design. The remaining cases can be solved using traditional numerically-based techniques such as the Newton-Raphson method.

Demonstrative Example of Planning Operations within MECXPERT

Aspects of the MECXPERT system, related primarily to knowledge representation and planning, have been discussed in detail, while touching briefly on issues related to heuristically-based systematic type and dimensional synthesis and data input. Two {PS}'s will now be examined within the context of a specific problem in order to demonstrate how planning is actually carried out in the system.

Freudenstein and Maki [7], employed the method of separation of kinematic structure and function to develop a variable-stroke slider crank mechanism for the design of a new internal-combustion engine. The problem presented in their paper will be used to demonstrate the sequence of planning operations which can occur within the "Design" and "eliminate-type of joints" {PS}'s (Figure 6).

When the current {PS} becomes "Design", the system checks to determine whether the current {PS} is fixed or probabilistic. This, along with other information, is stored in the compatibility1 data structures that are created prior to the time the system enters the "start" {PS} phase, i.e. when the system is initialized. Next, the system checks the contents of the compatibility1 records (in the current {PS}, i.e the "Design" {PS}, there are three of them, (1) Problem definition, (2) Problem

When program starts:
1. Compatibilities establishment (Problem space to problem space).
2. Control passes to start problem space.

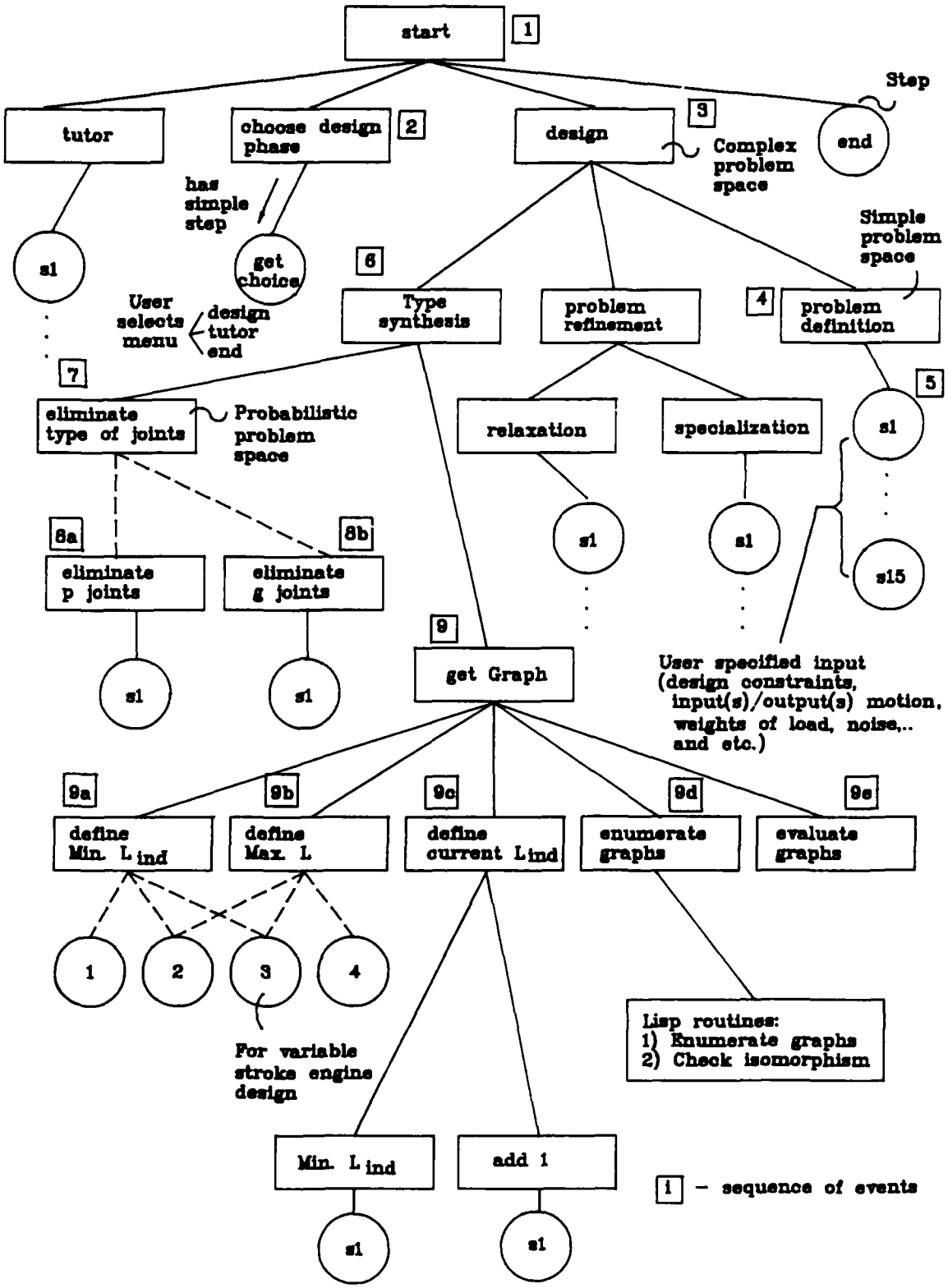


Figure 6. Sequence of events for Variable-Stroke engine mechanism design problem.

relaxation and (3) type synthesis) in order to propose compatible {SP}'s.

There are two types of compatibility elements shown below. Compatibility1 indicates compatible {SP}'s within a {PS} whether the {PS} is fixed or probabilistic and whether the {SP} is simple or complex. Compatibility2 is only used in probabilistic {PS}'s to indicate which constraints (associated with user input) are compatible with which {SP}'s in the current {PS}, the weighting value of that compatibility (assigned by rules based on the experts knowledge and user's input) and a cutoff value that indicates the minimum composite weighting factor value that a {SP} must have in order to be selected. Clearly, with only minor modification to the compatibility records it is possible to restructure the entire tree or to add new {PS}'s. This would have been a difficult task if a procedural language had been used to implement the system.

```
(compatibility1 ^PS Design          ^SP type-synthesis
                ^type complex       ^type1 fixed)
c Comments:
c Type-synthesis is a compatible subproblem-space of the problem-space "Design".
c Type-synthesis is a complex subproblem-space.
c Type-synthesis is a fixed subproblem-space.

(compatibility2 ^PS eliminate-type of joints    ^constraint speed
                ^SP eliminate-p joints ^wt <to-be-found-under-evaluate-
                constraints-operator>  ^cutoff <PS-dependent>)
c Comments:
c Speed is a constraint associated with the eliminate-type of joints
c problem-space.
c Eliminate-p joints is a compatible subproblem-space of the eliminate-
c type of joints problem space.
c A weighting factor value, <wt.>, associated with the eliminate-p
c joints problem-space determines if p-type joints should be eliminated
c from the mechanism design.
c A cutoff value, <cutoff>, indicates the minimum weighting factor value
c that the eliminate-p joints problem space must have in order to be
c selected (i.e., not rejected) for use in the design of the mechanism.

(compatibility1 ^PS eliminate-type of joints   ^SP eliminate-p joints
                ^type simple  ^type1 probabilistic)
c Comments:
c Eliminate-p joints is a compatible subproblem-space of the problem-space
c eliminate-type of joints.
c Eliminate-p joints is a simple subproblem-space.
c Eliminate-p joints is a probabilistic subproblem-space.
```

The three compatible {SP}'s corresponding to the "Design" {PS}, as shown in Figure 4, will then be created and set to a "pending" status. Since this {PS} is "fixed" the system will pass over steps 1 through 4 and jump to step 5, the reject operator (Figure 5), and will look for knowledge to reject all but a single {SP}. One of the rules that performs this function is given as follows:

```
(p reject-knowledge-in-design-1
   (goal ^step reject-operator  ^PS Design)
   (SP ^PS Design  ^name problem-definition
               ^status pending)
   (SP ^problem-space Design ^name {<> problem-definition}
                        ^status pending)
--->
   (modify 3 ^status rejected))
```

This rule states that when the current step is step 5, the reject-operator, in the "Design" {PS}, and when the "problem-definition" {SP} is "pending" then reject all the {SP}'s other than the "problem-definition" {SP}. After this step, the system will execute step 6, the choose operator, and choose the only available {SP}. Thus, the "problem-definition" {SP} is set to a "selected" status. Next, the system will move on to step 7, the apply operator, and apply the chosen {SP}. After the "problem definition" {PS} is executed, its status will change to "achieved". During the "problem-definition" phase the system will query the user and associate his answers with a data structure called constraint as follows:

```
(constraint ^name speed ^wt <a> ^status active)
c Comments:
c Speed is a constraint having both a weighting factor value, <a>, and a
c status associated with it.
```

After this {PS} has been executed different constraints will acquire an active status and weighting factor values. For the variable-stroke engine design, typical inputs would be:

1. High speed (wt. .9) and high loads (wt. .9)
2. Low noiseiness (wt. .8)
3. One input and output (wt. 1.0)
4. Rotary input (wt. 1.0) and slider output (wt. 1.0)
5. Control function within the mechanism (wt. 1.0)

   etc.

As soon as the system identifies an "achieved" {PS} it will backup to its parent {PS}. Thus, control will return to the "Design" {PS}. There the control will be assigned to step 4, the evaluate state operator (Figure 5), and the system will determine whether the built in evaluation knowledge for the current {PS} makes the current problem "achieved" or "failed". The system will also reset all the previously rejected {SP}'s to a status of pending. Since the successful completion of the "problem-definition" {PS} does not insure that the "Design" {PS} has been achieved, knowledge in step 5, the reject operator, will decide which {SP} will be rejected. The system at the current status will reject the "problem-relaxation" {SP} since it would be expected to select the "type-synthesis" {SP}. Once the "type-synthesis" {SP} is selected, it will become the current {PS} and the procedure shown in Figure 5 will be carried out by its {SP}'s. This procedure will set the "eliminate-type of joints" {SP} to be the current {PS}. Up to this point the status of the executed {PS}'s would be as follows:

Pending: Start, Design, type-synthesis, eliminate-type of joints.
Rejected: Tutor, problem-relaxation and get-graph.
Achieved: Choose phase, Problem-definition.

The "eliminate-type of joints" {PS} is probabilistic. After its two rooted {SP}'s are proposed, the operator in control will be "evaluate-constraints". Depending on user input, knowledge provided by a mechanism design expert will compute a weighting factor value that will be used by the next operator to compute the total degree of compatibility of each {SP}. Thus, the next operator will use the expert's assessment and the weighting factors assigned to the input, by the user, that are compatible with the current {SP} to compute the degree of compatibility of each of the {SP}'s. In a more general implementation of this system additional {SP}'s would exist such as reject cam joints, reject spherical joints, etc. For the design of a variable-stroke engine the last three {SP}'s would have the highest degrees-of-compatibility since for high speed and high load operating conditions, joints having surface contact rather than line contact are preferred and probably required.

Next, the system will check for the existence of any knowledge that would make the rejection of a {SP} necessary. For example, in this design case, both the "eliminate r-joints" and "eliminate p-joints" {SP}'s would be rejected. The eliminate-type of joint {PS} would continue to select and execute unrejected {SP}'s in order to make certain joint types available to lower level {SP}'s. When there are no longer any {SP}'s having weighting factor values greater than the cutoff value, the "eliminate-type of joints" {PS} will acquire a status of "achieved" and backup to the

"type synthesis" {PS}. In the "type synthesis" {PS}, the "get-graph" {SP} would then be selected. The current status check, relative to the last status check, of the executed {PS}'s would be as follows:

Pending: Type synthesis, get-graph.
Rejected: eliminate r-joints, eliminate p-joints.
Achieved: eliminate-type of joints.

The goal of the "get-graph" {PS} is to assign values to parameters that define the kinematic structure of the mechanism. These parameters are:

1. F, degree-of-freedom of the mechanism. This is determined by the number of inputs required to drive the mechanism as well as the number of required outputs, specific application requirements (i.e. how the mechanism is to be used) , the degree of complexity of the mechanism and whether or not the mechanism is required to be adjustable. This information is acquired from the user.
2. L_{ind}, number of independent loops in the mechanism. This variable is indicative of the degree of complexity of the mechanism. Its value is determined in the {PS} "define-current-Lind". The system will define, based on heuristics, a minimum and a maximum value for Lind, starting from the minimum value since simplicity is a desired property. In this design case, $L_{ind,min}$ must be ≥ 2, based on Rule-4 of the "evaluate-graph" {PS}, since a control loop is required to vary the stroke of the output link. The given design specifications require a value of $L_{ind} = 3$ in order to provide separate input, control and output loops. The maximum value for L_{ind} could, in general, be determined, for example, from cost and compactness limitations, as well as from input/output requirements.
3. f_i, the degree-of-freedom of relative motion permitted by the ith joint.
4. ℓ, the number of links.
5. j, the number of joints.
6. λ, the mobility of the space in which the mechanism operates. $\lambda = 3$ for general plane mechanisms, $\lambda = 6$ for spatial mechanisms.

The general degree-of-freedom equation may be expressed as [5]:

$$F = \lambda(\ell - j - 1) + \sum_{i=1}^{j} f_i \qquad (1)$$

The number of independent loops is given by the equation [15]:

$$L_{ind} = 1 + j - \ell \qquad (2)$$

Equations (1) and (2) can be combined, into the following equation:

$$\sum_{i=1}^{j} f_i = F + \lambda L_{ind} \tag{3}$$

Based on equation (3) with $L_{ind} = 3$ (as previously discussed), $F = 1$, and $\lambda = 3$ (for a general plane mechanism) the sum of the degrees-of-freedom for all the joints can be calculated:

$$\sum_{i=1}^{j} f_i = 1 + 3 \star (3) = 10 \tag{4}$$

Since high load carrying capability was a specified design requirement, only revolute (R) and prismatic (P) joints, each having one degree-of-freedom ($f_i = 1$), can be included in the design. Based on this information, equation (4) yields a value of $j = 10$. Rearranging equation (2), the number of links can calculated as follows:

$$\ell = 1 + j - L_{ind} = 8 \tag{5}$$

In general, equations (2), (3), (4) and (5) can be used to determine values for j and ℓ, depending upon the values selected for the L_{ind}, F and λ structural parameters. Their values would be chosen, firstly, to achieve the simplest possible design, based on heuristic knowledge appropriate to their selection. Once values for ℓ and j are known, appropriate graphs can be enumerated (labeling the graphs in as many non-isomorphic way as possible) and different joint types can be assigned to the edges of the graph in a way that insures the satisfaction of equation (3). This has been implemented in a LISP routine (Figure 6). The next step involves the evaluation of the graphs in the "evaluate-graphs" {PS} and the assignment of an index to each of them indicative of the order in which they should be processed, i.e. studied in greater detail. Additional generic (problem independent) rules can be established to assist in the elimination of inappropriate kinematic structures thereby further pruning the size of the mechanism design space.

As an example of the output provided by the system to the user on the Symbolics 3640 AI workstation, figures 7A, 7B, 7C and 7D display an enumeration of the graphs and mechanism schematic diagrams of several eight-link planar kinematic chains corresponding to numbers 1, 2 and 3 in group 1 and number 9 in group 3 of those enumerated by Freudenstein and Maki [7] for the variable-stroke engine mechanism problem.

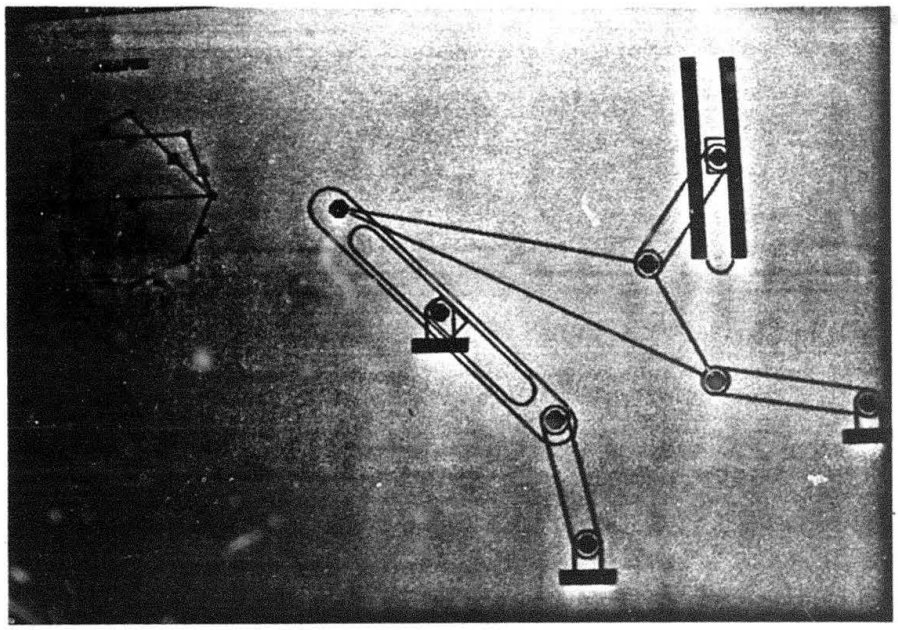


Figure 7A. : Graph enumeration and schematic drawing for an eight-link planar variable-stroke mechanism (group 1, number 1; Freudenstein and Maki, [7]).

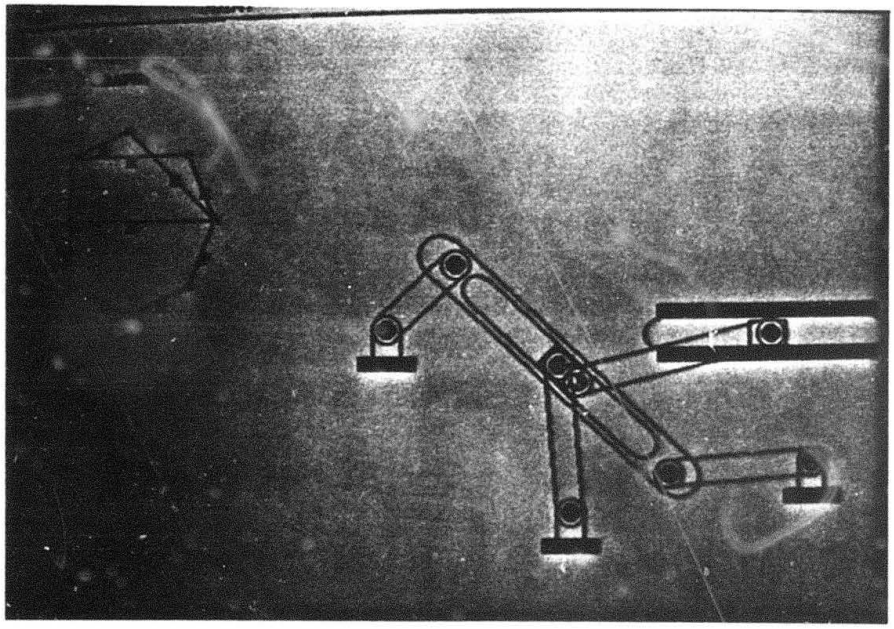


Figure 7B. : Graph enumeration and schematic drawing for an eight-link planar variable-stroke mechanism (group 1, number 2; Freudenstein and Maki [7]).

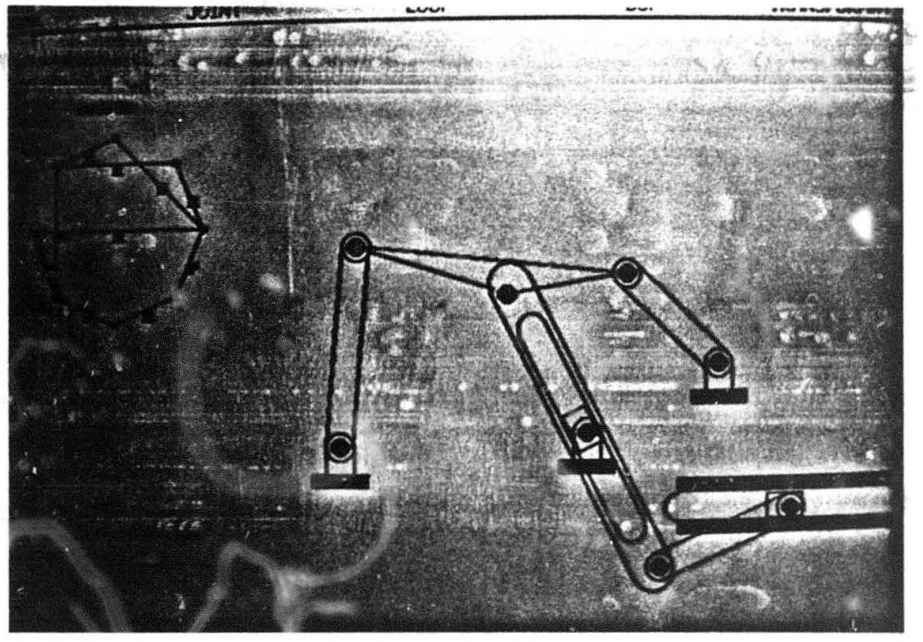


Figure 7C. : Graph enumeration and schematic drawing for an eight-link planar variable-stroke mechanism (group 1, number 3; Freudenstein and Maki [7]).

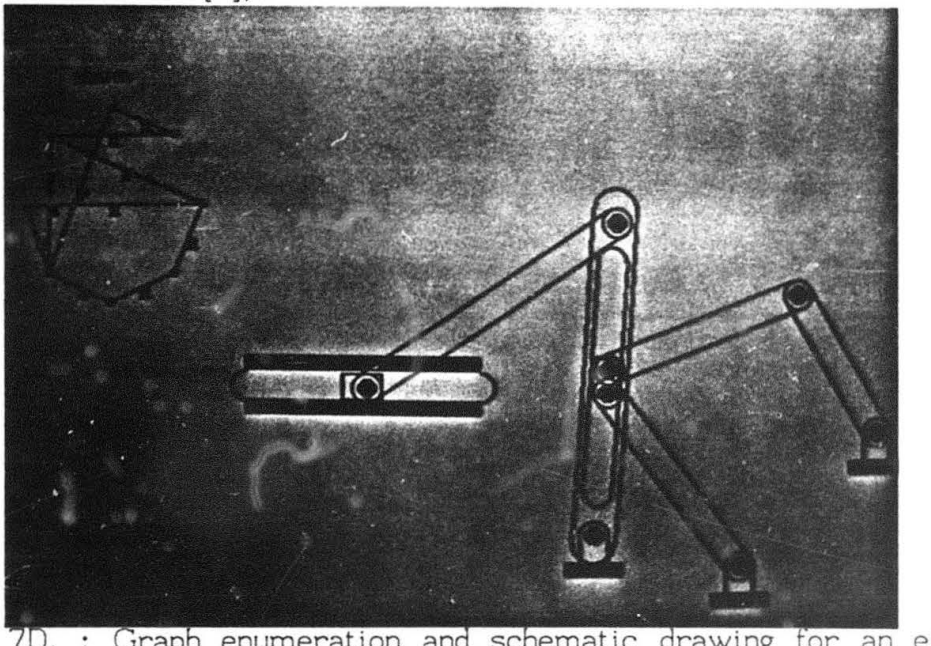


Figure 7D. : Graph enumeration and schematic drawing for an eight-link planar variable-stroke mechanism (group 3, number 9; Freudenstein and Maki [7]).

Eventually, kinematic and dynamic analyses would be undertaken for performance evaluation at a lower level of the design process.

Conclusions

A systematic methodology for representing knowledge and its control within an expert system for the creative design of mechanisms has been presented. Careful attention to the implementation of the control strategy for the manipulation of knowledge has been an important aspect of this research in anticipation of future growth of the MECXPERT system. The conceptual basis for the system relies on the separation of kinematic structure and function. An example based on the design of a variable-stroke engine mechanism serves to convey the manner in which information is imparted to and manipulated within the system in an effort to enumerate potentially viable mechanism designs.

Glossary of Terminology

Compatibility data structure:

Defines all the compatible {PS}'s and {SP}'s each time the system is initialized (started).

Constraint propagation:

The process of establishing the compatibility of interconnected elements, in this case {PS}'s and {SP}'s, within an expert system.

Creative mechanism design:

The process of solving a mechanism synthesis problem for which no prior, proven solution exists, based on the systematic separation of kinematic structure from function and employing heuristics where applicable for the selection of kinematic structural parameters in order to narrow the mechanism design search space.

Cutoff value:

In order for a probabilistic {PS} to be selected it must acquire a composite weighting factor value that is higher than the preset, expert defined, cutoff value assigned for that {PS}.

Data-driven inference strategy:

The search for new knowledge or information proceeds from known data to a final goal.

Deep domain knowledge:

Domain specific knowledge acquired over years of experience enabling an expert to solve difficult problems in that domain that cannot be solved by only analytical or numerical methods.

Experienced-based mechanism design:

The process of drawing upon knowledge concerning the structure-function relationships of mechanisms obtained from (1) mechanism design experts and (2) handbooks.

Failure handler:

Keeps track of a {PS} where failure has occured. When activated, the failure handler will go to the {PS} recommended by the {SP} where failure has occured and initiate a process parallel to one in which failure occured until the design is reset to a desired status.

Goal-driven inference strategy:

The search for new knowledge or information proceeds from the goal to be achieved, backwards, towards the known data.

Help facility:

A facility provided within MECXPERT which provides tutoring and advice to a user concerning the meaning and use of system commands. It can be initiated through the user specified system command word "explain".

Inference mechanism:

An interpreter that determines how to apply the rules in the knowledge base to infer new knowledge and the order in which these rules should be applied in an expert system.

Instance:

A variable whose value has been specified, i.e. instantiated.

Knowledge acquisition:

The process of acquiring knowledge about a specific area or domain (in this case mechanism design), from various sources, in order to bring this knowledge to bear on a narrow domain of difficult problems.

Knowledge base:

The collection of knowledge, typically in the form of facts and rules, about a specific domain (in this case mechanism design) to be used for decision making in an expert system.

Knowledge roles:

Knowledge and the action which it can impart are stored in data elements referred to as knowledge roles.

Mechanism synthesis:

The process of selecting the type, arrangement and number of links and joints in a mechanism for the purpose of fulfilling predetermined motion conversion or power transmission requirements.

Operators:

Data elements whose function permits the representation and control of knowledge.

Problem-space:

Represents the issue or concept currently under consideration. These are the states that the system can reside in and pass through in its effort to achieve its goal.

Problem space status:

The status of a problem space, {PS}, can take on one of three possible values: (1) Pending, (2) Achieved and (3) Failed. These are described in the text of the paper.

Routine design:

A design problem for which a proven solution methodology already exists and for which the design variables are known.

Redesign:

The process of changing an existing design, based on proven techniques, in order to comply with different design requirements.

Subproblem-space:

A subproblem-space, {SP}, represents the next available problem space.

Weighting factor:

Indicates the degree to which each of the constraints contributes to each rooted {SP}. The weighting factors are denoted as $wt_{i/j}$, the degree to which the ith constraint contributes to the jth {SP}.

Working memory element:

A data element that resides in the working memory portion of program memory during program execution.

Acknowledgements

The authors gratefully acknowledge the support provided by (1) the United States Army Research Office under grant DAAL03-86-K-0071, (2) the United States Army Research Office-Department of Defense University Instrumentation Program under grant DAAL03-86-6-0116, (3) the Carnegie Group through the use of the Knowledge Craft expert system development environment and (4) by a gift from the NCR Corporation. Also, the authors wish to express their appreciation to Professor Ferdinand Freudenstein for providing his expertise and knowledge in the area of mechanism design to the development of this expert system.

References

1. Jones, F.D., Horton, H.L., Newell, J.A. (Eds), Ingenious Mechanisms for Designers and Inventors, Industrial Press, New York, 1968.
2. Artobelevsky, I.I., Mechanisms in Modern Engineering Design, Vols. 1, 2- Parts I and II, 5 Part I, MIR Publishers, MOscow, 1986.
3. Kota, S., Erdman, A.G., Riley, D.R., Development of Knowledge Base for Designing Linkage-Type Dwell Mechanisms: Part 1-Theory, ASME Paper 86-DET-47, 1986.
4. Kota, S., Erdman, A.G., Riley, D.R., Development of Knowledge Base for Designing Linkage-Type Dwell Mechanisms: Part II-Application, ASME Paper 86-DET-48, 1986.
5. Freudenstein, F. Maki, E.R., The Creation of Mechanisms According to Kinematic Structure and Function, General Motors Research Laboratory Report GMR-3073, 1980.
6. Dobrjanskyj, L., Freudenstein, F., Some Applications of Graph Theory to the Structural Analysis of Mechanisms, ASME Trans., J. of Eng. for Industry, 89B, pp 153-158, 1967.
7. Crossley, F.R.E., The Permutation of Kinematic Chains of Eight members or Less from the graph-theoretic viewpoint, In Developments in Theoretical and Applied Mechanisms, Volume 2, Ed. W.A. Shaw, Pergamon Press, Oxford, pp. 467-486, 1965.
8. Hayes-Roth, F., Waterman, D.A., Lenat, D.B., Eds. Building Expert Systems, Addison-Wesley Publishers, 1983.
9. Freudenstein, F., Maki, E.R., Development of an Optimum Variable-Stroke Internal-Combustion Engine Mechanism from the Viewpoint of Kinematic Structure, ASME J. Mech., Trans., Auto. in Design, Vol. 105, pp 259-266, 1983.
10. Brownston, L., Farrell, R., Kant, E., Martin, N. Programming Expert Systems in OPS5: An Introduction to Rule-Based Programming, Addison-Wesley Pub., 1985.
11. Knowledge Craft Expert System Development Environment, The Carnegie Group, 1986.
12. Gibbons, A., Algorithmic Graph Theory, Cambridge University Press, 1985.

13. Waterman, D., A Guide to Expert Systems, Addison-Wesley Publishing Co., 73-77, 1986.
14. Nilsson, N., Problem-Solving Methods in Artificial Intelligence, McGraw-Hill, 1971.
15. Paul, B., A Unified Criterion for the Degree of Constraint of Plane Kinematic Chains, ASME J. of Applied Mechanics, Vol. 82, pp 196-200, 1960.

Statistical Machine Learning for the Cognitive Selection of Nonlinear Programming Algorithms in Engineering Design Optimization

D.A. Hoeltzel & W.H. Chieng

The Department of Mechanical Engineering
Columbia University
New York, New York 10027

Abstract

In order to overcome the problem of lack of generality in nonlinear programming (NLP) test problem formulation and to introduce the concept of cognitive NLP method switching, statistical machine learning has been applied to a sample data base of nonlinear programming problems. Reasonable conclusions have been drawn about an optimization problem type and a corresponding ***sequence*** of NLP solution algorithms, using statistical pattern recognition applied to ***local*** (***vs. global***) design information. A program, referred to as OPTDEX-OLDM, with the capability of learning from statistical pattern recognition is discussed. The statistical aspects and algorithmic optimization of the nonlinear programming problem are emphasized in this discussion. A clustering process has been performed on attributes assigned to the NLP problem sample data base, and an example which describes this statistical clustering process is discussed.

Introduction

Numerical optimization techniques, in the form of nonlinear programming (NLP) algorithms, have been applied extensively to ***critical*** structural design and analysis problems for more than 30 years [1], and to a lesser extend to mechanical design problems [2].

The nonlinear programming problem considered here takes the form,

$$
\begin{aligned}
&\text{Minimize: } F(X) && && \text{objective function}\\
&\text{Subject to:}\\
&\quad g_j(\bar{X}) \le 0 && j = 1, m && \text{inequality constraints}\\
&\quad h_k(\bar{X}) = 0 && k = 1, l && \text{equality constraints}\\
&\quad X_i^l \le X_i \le X_i^u && i = 1, n && \text{side constraints}
\end{aligned}
\qquad (1)
$$

$$
\text{where } \bar{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \text{ vector of design variables.}
$$

Numerical optimization provides a systematic, rational and directed approach to design decision making where previously, heavy reliance was placed on the experience and intuition of the designer in achieving an improved design. Due to complexities involved in the implementation of NLP algorithms, several researchers have undertaken performance analyses [3,4,5,], the purpose being to determine correlations among the design problem type, the numerical optimization method and the corresponding results. Based on such studies, it is anticipated that the novice user should be able to better understand the capabilities of existing optimization methods and furthermore, utilize them without the need to undertaken exhaustive programs for testing and learning. While in concept this appears to be a rational approach to ascertain the capabilities of a particular algorithm for a specific problem, in reality, Himmelblau [6] states that ***"a guarantee of convergence for an algorithm for special cases may offer little insight as regards satisfactory strategies for more complex problems"***.

An optimization process invariably involves a trade-off between reality (completing and understanding the search process) and economy (evaluating a limited number of test functions). A process referred to as statistical concept learning[1] is introduced to compensate for this trade-off. Based on a well organized data hierarchy, concept learning has been developed to eliminate unwanted knowledge which may occur due to noisy data[2] [7] and a scheme for generalization of the statistical results has been developed.

Method Switching Strategies in Nonlinear Optimization

Existing algorithms for nonlinear programming which have been surveyed [8,9] may converge to local optima which are not necessarily global optima. Many techniques for locating global optima, aside from knowing which method is the best first method have yet to be uncovered. Method switching strategies are based, by analogy, on the game of golf[3] rather than on the use of a one step optimization scheme. This method switching procedure is designed to be one level higher than the so called optimization strategy level [10] (monitors the numerical optimization

[1] Statistical concept learning: Learning about new concepts by using given statistical measurements.

[2] Noisy data: A small amount of data contradicting the conclusions which are agreed upon by a majority of the remaining data. In other words, data lying outside any of the defined cluster groups (Figure 1.).

[3] Game of golf analogy: The reason for method switching is in accordance with the local geographical design information at the numerical optimum, and is analogous to the reason for selecting an appropriate golf club, in the game of golf, to strike the the ball.

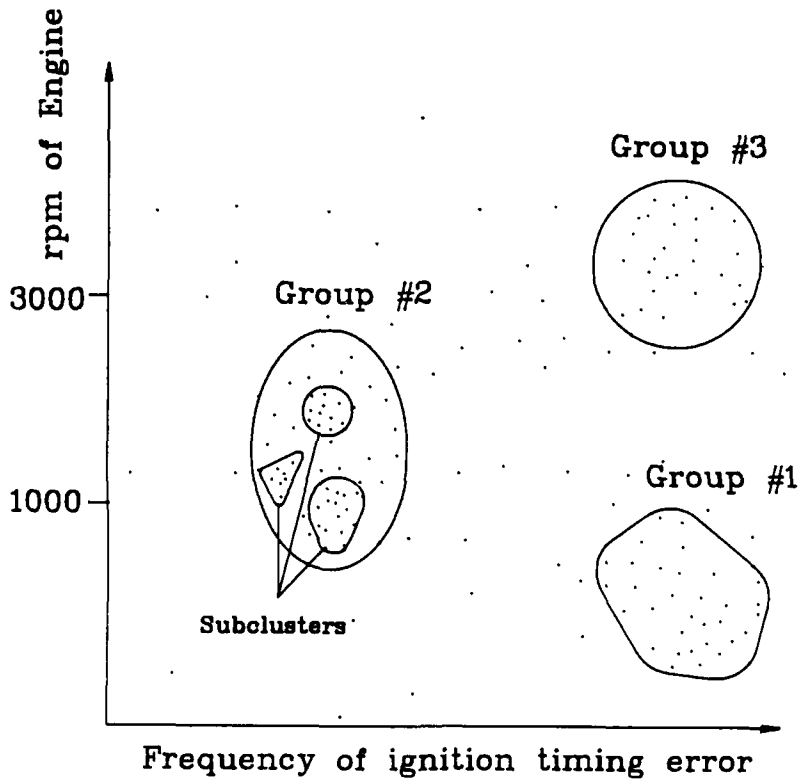


Figure 1. A clustering example.

1. Group #1 (low rpm) and Group #3 (high rpm) cause more ignition failures.
2. Those points which are not enclosed within any of the groups have been referred to as noisy data.

process) and switches suitable numerical method combinations according to local design data.

For example, in the following problem containing a single objective function, design variable and constraint:

Minimize the design objective function:	$\cos(x/1000 - 5.)$
Subject to the design constraint function:	$400. - x^2 \leq 0.$
with design variable bounds:	$0. \leq x \leq 7500.$

The following cases may possibly occur,

case 1: When $|x| >> 20$, the local information indicates that the design constraint is inactive.

case 2. When $x \simeq 5000$, the local information indicates that the objective function can be linearized to a polynomial of degree 2, which is $1 - (x/1000 - 5)^2/2$.

case 3. When $|x - 5000| < 10$, the local design information indicates that the objective function can be linearized to a polynomial of degree 4 by using an approximation of a Taylor series expansion.

case 4. When $x \geq 7500$, one more design constraint is added from the design bounds, which can be expressed as $x \leq 7500$.

This example demonstrates that local design information can change in various ways when the updated state of the design variables (position) is altered. Method switching strategies are based on this phenomenon and may be likened to a monitoring or blackboard[4] style decision making process. Method switching keeps track of the local optimization information and switches methods when the current method fails.

According to the schematic representation depicted in Figure 2, the first design starting point, P1, lies in an infeasible design region and is far away from the globally optimal point. A temporary goal may be expressed as "move the design into the feasible region as soon as possible" to increase the design efficiency. When the design "converges" at a local optimum, P2, current NLP methods fail to move away from this point. In accordance with the local information found in the vicinity of P2, the method switching manager pins down another temporary goal which may be stated as "find a feasible design with a smaller objective value". Method switching terminates when the convergence criteria have been satisfied. This is usually based on (1) a cpu time consumption limitation, (2) the number of algorithm iterations or (3) relative or absolute difference between successive values of the objective function.

[4]Blackboard architecture: A model in which all intermediate messages and results are displayed to the user and stored in a common area, called a blackboard.

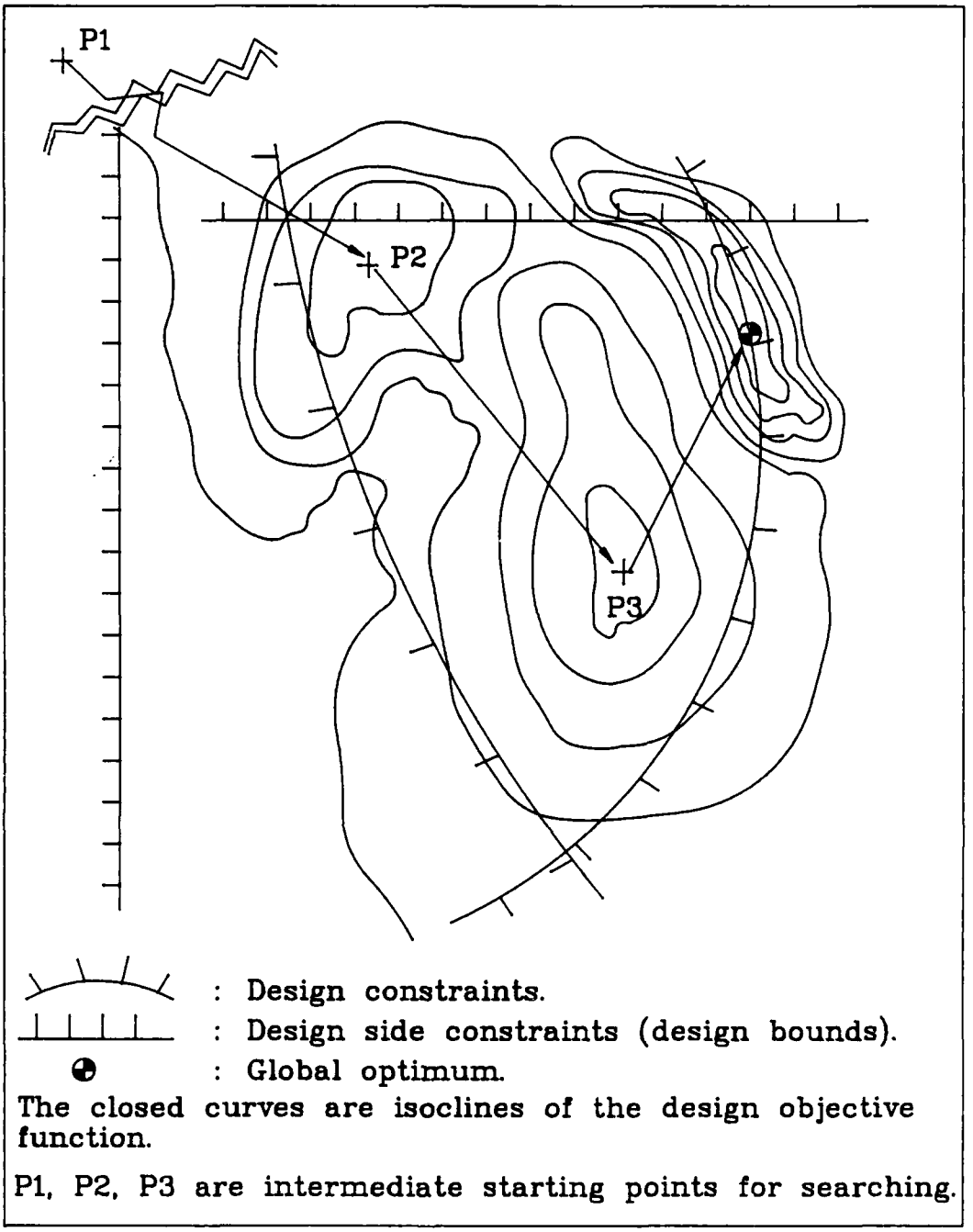


Figure 2. An example which demonstrates that local design information is different for different design starting points.

Sample Problem Testing

Fifteen different attributes have been chosen to characterize the test of a sample problem. The sample problems can be separated into three domains:

1. ***The Design Problem Type*** - contains 8 parameters, including the number of design variables, the number of total design constraints, the number of equality design constraints, the number of active inequality design constraints, the maximum (positive) order of testing polynomials, the minimum (negative) order of testing polynomials and the function evaluation cost for one design function evaluation.
2. ***The Choice of Nonlinear Programming method*** - contains 3 parameters, which according to the ADS numerical optimization library [10] are strategy, optimizer and one dimensional search method.
3. ***The Performance of the Result*** - contains 4 parameters, including the minimum objective value reachability, the design constraint violation condition and the maximum distance of search.

The set of test problems for the learning program have been produced by a random function generator (Figure 3), which randomly selects a problem type, and in accordance with the selected problem type generates the objective function and the design constraint equations. These polynomials can be thought of as local information in real world design problem formulations since many functions can be expressed in a Taylor series expansion. Nonlinearity, discontinuity and differentiability can be altered by appropriately adjusting the order of the polynomials.

After implementing these concepts using the ADS numerical optimization library, design problems have been tested by a number of method combinations, which have been randomly selected. The authors have generated approximately 10,000 samples with results using an IBM PC/AT microcomputer. These results have been subsequently analyzed, using statistical machine learning concepts incorporated within a program referred to as OPTDEX-OLDM {Optimum Design Expert-Optimization Level Design Manager}, on a Symbolics 3640 AI workstation.

Clustering and Associated Statistics

Every sample inherently has several attributes, which include the characteristics of the design problem type, the category of the nonlinear programming method and the corresponding result. All of these attributes are represented quantitatively and some of them are noisy, i.e. unreliable. To minimize the noise factor, a "variance" type of analysis [4] has been employed.

Clustering techniques [13, 14,15,16] are used to find groups of samples, whose common characteristics **have not been predefined.** The aim is to subdivide the

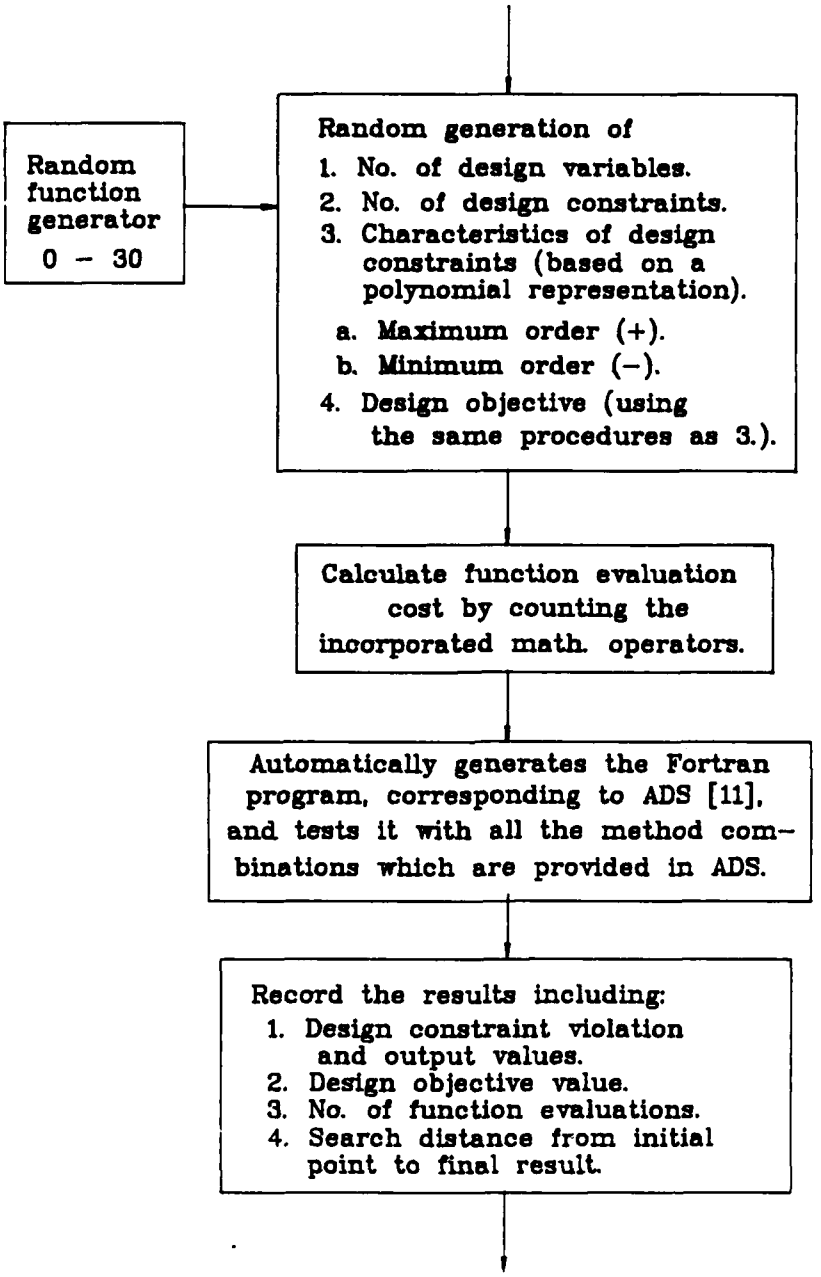


Figure 3. Flow control of random sample generation and testing.

available samples into a relatively small number of groups, based on the statistical behavior of the different attributes.

The clustering analysis involves the following concepts:

1. ***Scaling:*** Tranforms the real world value of each attribute into a machine understandable scale. This can be done by calculating the mean, μ, where

$$\mu = (1/m) \sum_{i=1}^{m} a_i \qquad (2.a)$$

and the standard deviation, σ, where

$$\sigma = \sqrt{E((A - \mu)^2)} \qquad (2.b)$$

where m = total number of data samples.
a_i = value of the i^{th} attribute.
A = random variable which can assume the value a_i.
E = expected value (statistical sense).

Various models may be chosen to represent the statistical distribution of the attributes. For example, if a Gaussian distribution is chosen, then 68% of samples will be distributed within one standard deviation about the mean, μ, and about 95% of the samples will be distributed within two standard deviations about the mean, μ. According to the mean, μ, and the standard deviation, σ, found for each variable, all the variables are normalized and digitized to a predefined scale. For the purposes of this research, 0 through 9 has been selected.

2. ***Non-hierarchical clustering:*** Non-hierarchical clustering is based on the optimization of a given grouping of objective functions, and represents the minimization of the sum of the variances within each group and the maximization of the sum of variances between groups.

$$\min_{C \in p(n, M)} \sum_{j=1, i \in C_j}^{n} \| a_i - a_j \|^2 \qquad (3)$$

and

$$\max_{C \in p(n, M)} \sum_{j=1}^{n} m_j \| \bar{a}_j - \bar{a} \|^2$$

where $C = (C_1, C_2, C_3, \ldots, C_n)$ and C_i represents the i^{th} cluster group.

$M = \{1, 2, 3, \ldots, m\}$; set of all samples.
$p(n,M)$ = set of all cluster groups C of M having length n.
n = number of cluster groups; $1 \leq n \leq m$.
$\bar{a}$ = the expected value of the total sample of attributes.
$\bar{a}_j$ = the expected value of C_j.
m_j = the number of samples in C_j.

Since the total scatter in a fixed sample size is constant [10], it is sufficient to minimize the sum of variances, W(n,M), within each group. Therefore, eq. (3) can be expressed as follows,

$$\min W(n,M) = \min_{C \in p(n,M)} \sum_{j=1}^{n} \sum_{i \in C_j} || a_i - a_j ||^2 \quad (4)$$

The necessary tools for the clustering process are described below.

To calculate the new mean value from two given groups:

$$\bar{a}_{p+q} = \frac{1}{m_p + m_q} (m_p \bar{a}_p + m_q \bar{a}_q) \quad (5)$$

and to calculate the objective value (sum of variances) of the two given groups:

$$W_{p+q} = W_p + W_q + m_p * [\bar{a}_p - \bar{a}_{p+q}] [\bar{a}_p - \bar{a}_{p+q}]^T + m_q * [\bar{a}_q - \bar{a}_{p+q}] [\bar{a}_q - \bar{a}_{p+q}]^T \quad (6)$$

3. ***Clustering strategy:*** Since the number of **all possible cluster group combinations** (total clustering) can become prohibitively large, it is imperative that a reduction in the number of clusters be attempted. For example, say m samples (attribute values) have to be clustered into less than or equal to n groups. This number of clusters is given by:

$$H(n,m) = \frac{1}{n!} \sum_{i=1}^{n} (-1)^{n-i} \binom{n}{i} i^m \quad (7)$$

For m = 1000 and n = 15, H (n,m) is greater than $1 * 10^{30}$. For this reseach, m=10,000 and $150 \leq n \leq 300$, therefore the clustering is not practically achievable. As a result, a special strategy has been employed to alleviate this problem. Instead of searching for total clustering, the OPTDEX-OLDM program starts from m samples and allows each single sample to be a group, i.e. n = m. The program then attempts to decrease the total number of cluster groups, during each clustering cycle, by one. During each cycle the program searches for any two groups from the current set which satisfies the criterion of equation (4). This clustering process terminates when the number of groups, denoted by n*, satisfies the following condition,

$$\min W(n^*, M) \geq W_{acceptable} \quad (8)$$

Based on this clustering strategy, H' (n,m), the reduced number of clusters are:

$$H'(n,m) \simeq \frac{m^2 - n^2}{2} \quad (9)$$

For the m = 1000 and n = 15 case, $H^1(n,m) \simeq 5E+05$. Athough noise may bias this type of clustering in the very early stages of processing, as previously predicted, when compared with the increased efficiency, of approximately $2 * 10^{12}$ times, it is an acceptable strategy. Flow control for this process is shown in Figure 4.

Explanation of The Statistical Results

An explanation facility [17] is an important feature which distinguishes artificial intelligence programs from usual programs. Its purpose is to present the computational results in the form of a natural language so that is comprehensible to a novice user. In addition, this capability forms the basis of incremental machine learning. A simple example that demonstrates how machine learning provides an explanation for a resulting cluster group follows:

Group 1. Number of members = 17

Attribute	Range	Mean	Variance
Nonlinearity	0-9	8	3.0
Strategy	0-9	2	0.2
Distance-of-Search	0-9	1	1.0

Response from the OLDM:

OLDM> I found that (as supported by 17 samples),

IF

Generally-speaking, the nonlinearity is very-high, and

Definitely, the strategy is the linear extended interior penalty function method.

Then

Most-likely, optimization searching will be very-local.

(Underlined explanations represent terminology derived from the statistical results).

Classification and Incremental Machine Learning

Automatic concept learning, implemented in the form of concept learning generalization[5], has been shown to be useful in interpreting and organizing large amounts of information about a domain [?] After performing the initial clustering from the test samples (ten thousands samples in this case) the OPTDEX-OLDM

[5]Concept learning generalization: The automatic generalization of a concept based on a sufficiently large number of agreements among specific case (non-general) concepts. In other words, expanding a concept to include a more general class of specific cases than previously included.

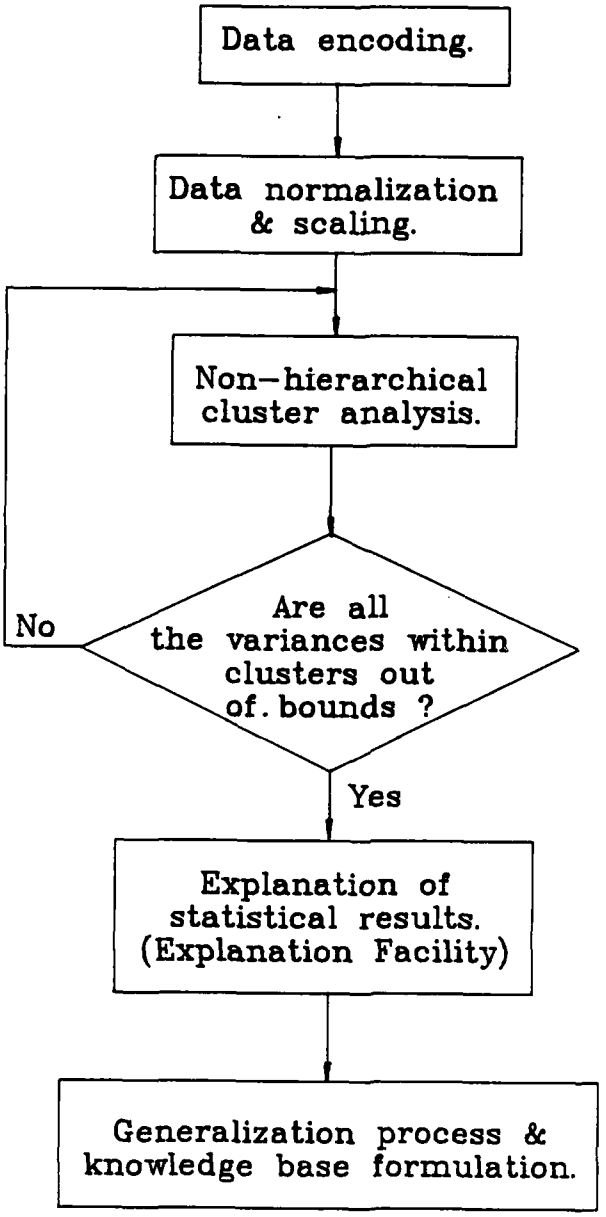


Figure 4. Flow control of cluster analysis.

program reaches approximately 500 conclusions. These conclusions may overlap one another, some of them may be redundant and they all have to be appropriately formatted into a rule-based expert system.

Creating a classification scheme is typically the first step in developing the heuristics (rules of thumb) for a collection of observations or phenomena. The goal of the classification scheme is to structure given observations into a hierarchy of meaningful categories [6]. The OLDM applies generalized-based memory to build up a hierarchy of conclusions. It actually constructs a connective network to derive conclusions in a canonical form. A detailed explanation of this process is provided by Lebowitz [7]. An important feature of the OLDM is its ability to manage contradictions between conclusions, referred to as noise, by simply counting the number of supporting members for each conclusion. For example, the following conclusions (non-generalized) have been drawn by the OLDM:

Conclusion 1. Supported by 19 members 1
If the Discontinuity is high and
the Optimizer-choice is Golden-section-method
then the objective value is less-minimized.

Conclusion 2. Supported by 25 members
If the Discontinuity is low and
the Optimizer-choice is Golden-section-method
then the objective value is less-minimized.

Conclusion 3. Supported by 4 members
If the Discontinuity is high and
the Optimizer-choice is Golden-section-method
then the objective value is minimized.

The generalized concept, drawn by the OLDM, based on these conclusions is:

OLDM> CONCEPT-008:
If the Discontinuity is high or low* and
comment: *<this result is based on the generalization of conclusions 1 & 2>
the Optimizer-choice is Golden-section-method
then the objective value is less-minimized.**
comment: **<the number of members supporting conclusion 1 is greater than
the number supporting conclusion 3>

Another important feature of the OLDM is its ability to perform on-line statistically incremental machine learning. The OLDM is an on-line consultant during

numerical optimization processing which has been incorporated within the ADS (Automated Design Synthesis) optimization library. According to the existing rules and local information from updated optimization searching, it chooses and switches methods combinations from ADS and feedsback the result of each applied rule. These feedbacks are always represented in a standardized format with 14 parameters as previously described. Each piece of standardized information can be treated as an additional test sample, a_e, clustered into a group, C_j which satisfies the following condition.

$$\min_{C \in p\,(n,\,M)} \frac{m_j}{m_j + 1} * [a_e - \bar{a}_j]\,[a_e - \bar{a}_j]^T \qquad (8)$$

During the incremental machine learning process, any of the existing cluster groups, say C_k, such that $W_k > W_{acceptable}$, has to be re-clustered by utilizing the procedures which have been discussed. After the re-clustering process has been completed, new concepts (conclusion) are born and/or old concepts die. This is referred to as the birth-and-death procedure for maintaining and renewing concepts in the knowledge base.

Conclusion

A new approach to design optimization, referred to as cognitive method switching, using nonlinear programming (NLP) algorithms applied sequentially, based on local design information, has been presented. Statistical evaluation with clustering of attributes associated with a randomly generated problem sample data base, containing over 10,000 samples, has led to the generation of guidelines for the application of NLP algorithms to design optimization problems. Continued expansion of the problem data base should permit more generalized guidelines to be obtained and thereby assist the nonexpert user in cognitively selecting an appropriate sequence of NLP algorithms for a specific design optimization problem.

Acknowledgements

This research was supported by grants from the IBM and NCR corporations and by the U.S. Army Research Office/DoD - University Research Instrumentation Program under grant DAAL03-86-6-0116.

References

1. Vanderplaats, G.N. Numerical Optimization Techiniques for Engineering Design With Applications, McGraw-Hill (1984).
2. Siddall, J.N. Optimal Engineering Design: Principles and Applications Marcel Dekker, Inc. (1982).
3. Schittkowski, K. Nonlinear Programming Codes, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag (Berlin), (1980).

4. Powell, M.J.D. editor, Nonlinear Optimization 1981, Academic Press (1982).

5. Sandgren, K.M. Ragsdell, "The Utility of Nonlinear Programming Algorithms: A Comparative Study - Part I & II" Journal of Mechanical Design ASME, (1980): pp 540-551.

6. Himmelblau, D., Applied Nonlinear Programming, McGraw-Hill (1972).

7. Lebowitz, M. "Concept Learning In A Rich Input Domain: Generalization-Based Memory" Machine Learning: An Artificial Intelligence Approach, Vol. II, Morgan Kaufmann (1986): pp 193-214.

8. Wilde, D., Globally Optimal Design, John Wiley & Sons, New York (1978).

9. McCormick, G.P., Nonlinear Programming: Theory, Algorithms, and Applications, John Wiley & Sons, New York (1983).

10. Vanderplaats, G.N., "ADS - Fortran Program for Automated Design Synthesis", NASA Contract Report 172460 (1984).

11. Spath, H. Cluster Dissection and Analysis, Ellis-Horwood Limited (1985).

12. Cochran, W.G., Sampling Techniques, John Wiley & Sons, New York, (1977).

13. Green, W.R., Computer-Aided Data Analysis - A Practical Guide, John Wiley & Sons Inc., New York, (1985), pp 185-202.

14. Diday, E. "Problem of Clustering and Recent Advances", in New Trends in Data Analysis and Applications, Janssen ed. (1983), pp 167-182.

15. Mandel, J., The Statistical Analysis of Experimental Data, Dover Pub., (1964).

16. Stepp, R.E., Michalski, R.S., "Concept Clustering Inventing Goal-Oriented Classifications of Structured Objects", Machine Learning: An Artificial Intelligence Approach, Vol. II, Morgan Kaufmann Pub., (1986), pp 193-214.

17. Hoeltzel, D.A., Chieng, W.H., An Adaptive, Generic Planning Model for Large Scale Integrated Engineering Design, 1st Eurographics Workshop on Intelligent Computer-Aided Design, Chapter 5, Springer-Verlag (1987).

TOWARD A NONEQUILIBRIUM THERMODYNAMICS OF TWO PHASE MATERIALS WITH SHARP INTERFACE

Morton E. Gurtin
Department of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT. This is a review of recent work of the author toward the development of a nonequilibrium thermodynamics of two-phase continua based on the first two laws in forms which contain interfacial contributions for energy and entropy. Topics discussed are: thermodynamic restrictions on constitutive equations; interface conditions; free-boundary problems for solidification and melting.

1. INTRODUCTION. The classical theory of Stefan, for the melting of a solid or the freezing of a liquid, is too simplistic to account for the myriad of phenomena which occur during solidification (an example being dendritic growth, in which simple shapes evolve to complicated tree-like structures).[1] Recent attempts to rectify this situation involve replacing the classical free-boundary condition,

$$\theta(\mathbf{x},t) = \theta_M \quad \text{on} \quad \mathcal{S}(t), \tag{1.1}$$

for the temperature $\theta(\mathbf{x},t)$ on the interface $\mathcal{S}(t)$, by a condition in which the mean curvature $H(\mathbf{x},t)$ and the normal velocity $V(\mathbf{x},t)$ of $\mathcal{S}(t)$ are allowed to influence the temperature:

$$\theta(\mathbf{x},t) = \theta_M - hH(\mathbf{x},t) - bV(\mathbf{x},t). \tag{1.2}$$[2]

Here θ_M, a constant, is the **transition temperature**, the temperature at which the bulk free energies of the solid and liquid coincide, while h and b are constants.

The relation (1.2) with $b = 0$ is usually derived by assuming that (at each time) the interface is in thermal equilibrium with the bulk material, and then linearizing the interfacial condition obtained as a consequence of Gibbs' criterion for stability. The complete relation (1.2) with $b = 0$ is generally justified on an ad hoc basis, since the presence of the normal velocity V precludes the use of equilibrium thermodynamics.

---

[1] Cf., e.g., Chalmers [1] and Delves [2] for discussions of these phenomena.

[2] For solidification problems, free-boundary conditions of this type, with $b = 0$, were introduced by Mullins and Sekerka [3], [4]; the term involving V was added by Voronkov [5], Seidensticker [6], and Tarshis and Tiller [7]. (See also the review articles by Sekerka [8], [9], [10], Chernov [11], Delves [2], and Langer [12].)

One expects that free-boundary conditions derived in this manner are a valid approximation in many situations. On the other hand, since the underlying physical problem involves a physical system out of - although possibly near to - equilibrium, it would seem advantageous to develop a nonequilibrium thermodynamics which yields, as consequences, appropriate free-boundary conditions for the interface between phases. This review discusses recent work [13][3] by the author toward the development of such a nonequilibrium thermodynamics.

2. BASIC CONCEPTS. The work [13] begins with the first two laws in forms which are appropriate to a continuum and which contain interfacial contributions for energy and entropy; but to avoid inessential complications, attention is restricted to nondeformable bodies in the absence of diffusion.

A fairly general constitutive theory for the interface is considered. The free energy f and entropy s are allowed to depend on the temperature θ and on the orientation of the interface through a dependence on its unit normal $\mathbf{m}$:

$$f = \hat{f}(\theta,\mathbf{m}), \qquad s = \hat{s}(\theta,\mathbf{m}). \tag{2.1}$$

(The dependence on $\mathbf{m}$ is included to model crystal growth.)

An essential requirement of the theory is that the *temperature depend on the kinematics of the interface*. In particular, a constitutive relation

$$\theta = \hat{\theta}(V,\mathbf{m},L) \tag{2.2}$$

giving the temperature as a function of the normal velocity V of the interface, the curvature tensor L for the interface, and the normal $\mathbf{m}$ is introduced.

One might expect that the motion of the interface (relative to the underlying material structure) induces a transfer of mechanical energy within the interface. To allow for this possibility, a tangential vector field $\mathbf{j}$ is introduced; for c an arbitrary subsurface of the interface $\mathcal{S}$,

$$-\int_{\partial c} \mathbf{j}\cdot\boldsymbol{\nu} \tag{2.3}$$

represents a flow of energy into c across ∂c. Here $\boldsymbol{\nu}$ (a tangential vector field) is the outward unit normal to the boundary curve ∂c. The vector field $\mathbf{j}$ is called the **accretive energy flux**, and the description of the interface is completed by adding a constitutive equation

$$\mathbf{j} = \hat{\mathbf{j}}(v,\mathbf{m},L); \tag{2.4}$$

interestingly, for an isotropic interface this flux vanishes identically.

---

[3] Based on the earlier study [14].

Triplets $(V,\mathbf{m},L)$ in the common domain of $\hat{\theta}$ and $\hat{\mathbf{j}}$ are called **states**, and states with $V = 0$, $L = 0$ are called **equilibrium states**.

It is not clear what thermodynamic restrictions ought to be placed on these constitutive assumptions, and it would seem appropriate to use the second law - in the manner of Coleman and Noll [15] - to derive such restrictions.[4] This is not as straightforward as it seems. For a rigid heat conductor the treatment of Coleman and Noll [15], developed for single-phase materials, is based on the hypothesis that the second law be satisfied in all processes generated - through the constitutive equations - by smooth temperature fields. Here, however, there is an additional degree of freedom, the evolution of the interface, and the constitutive restriction (2.2) does not allow for an arbitrary assignment of *both* the interface *and* the underlying temperature.

3. THERMODYNAMIC RESTRICTIONS ON CONSTITUTIVE EQUATIONS. Compatibility with the second law leads to the following constitutive restrictions:

(i) the free energy has the form

$$\hat{f}(\theta,\mathbf{m}) = f_0(\mathbf{m}) + f_1(\theta);$$

(ii) the entropy $\hat{s}(\theta,\mathbf{m}) = \hat{s}(\theta)$ is independent of $\mathbf{m}$ and determined by the free energy through the **entropy relation**

$$\hat{s}(\theta) = -\partial_\theta f_1(\theta);$$

(iii) the accretive energy flux $\hat{j}(V,\mathbf{m},L) = \hat{j}(V,\mathbf{m})$ is independent of L and linear in V:

$$\hat{j}(V,\mathbf{m}) = -V\hat{\xi}(\mathbf{m}); \tag{3.1}$$

(iv) $\hat{\xi}(\mathbf{m})$ is determined by the free energy through the **stress relation**

$$\hat{\xi}(\mathbf{m}) = -\partial_{\mathbf{m}} f_0(\mathbf{m});$$

(v) given any state $(V,\mathbf{m},L)$,

$$V\{[\psi(\theta)] - Hf - \partial_{\mathbf{m}}\xi(\mathbf{m})\cdot L\} \geq 0, \tag{3.2}$$

where $[\psi(\theta)]$ is the jump in bulk free-energy across the interface.

---

[4] Murdoch [16] has applied this procedure to interfaces which do not move relative to the underlying material.

Note that (3.1) reduces the energy flow (2.3) to

$$\int_{\partial c} V\hat{\xi}(\mathbf{m})\cdot\boldsymbol{\nu}.$$

This integral has an obvious interpretation as power expended on c, with $\hat{\xi}(\mathbf{m})\cdot\boldsymbol{\nu}$ a force[5] in the direction normal to the interface. For that reason $\hat{\xi}(\mathbf{m})$ is called the **accretive stress**.

One might object to the constitutive equations (2.1), (2.2), and (2.4), as they are not consistent with the principle of equipresence. Consider instead the system

$$\begin{aligned} f &= \tilde{f}(V,\mathbf{m},L), \qquad s = \tilde{s}(V,\mathbf{m},L), \\ \theta &= \hat{\theta}(V,\mathbf{m},L), \qquad \mathbf{j} = \hat{\mathbf{j}}(V,\mathbf{m},L). \end{aligned} \tag{3.3}$$

Near equilibrium this system is no more general than the original system (2.1), (2.2), *and* (2.4). Precisely, it is shown that, if (3.3) is compatible with thermodynamics, then there exist a neighborhood of equilibrium N and constitutive functions $\hat{f}(\theta,\mathbf{m})$ and $\hat{s}(\theta,\mathbf{m})$ such that

$$\tilde{f}(V,\mathbf{m},L) = \hat{f}(\hat{\theta}(V,\mathbf{m},L),\mathbf{m}), \qquad \tilde{s}(V,\mathbf{m},L) = \hat{s}(\hat{\theta}(V,\mathbf{m},L),\mathbf{m})$$

on N.

In classical theories of melting - in which the interface is devoid of structure - changes of phase occur at the transition temperature θ_M. Within the present theory a consequence of the inequality (3.2) is that the interface have temperature θ_M at equilibrium,

$$\hat{\theta}(V,\mathbf{m},L) = \theta_M \quad \text{whenever} \quad V = 0, \quad L = 0,$$

but that away from equilibrium this need not be so; in fact,

$$\theta = \theta_M - f(\mathbf{m})H - b(\mathbf{m})V + \operatorname{div}_\Delta \xi(\mathbf{m})$$

is the linear **approximation** to (2.2) near equilibrium. Here $f(\mathbf{m}) = \hat{f}(\theta_M,\mathbf{m})$ is the interfacial free energy at equilibrium, $\xi(\mathbf{m}) = \hat{\xi}(\mathbf{m})$ is the accretive

---

[5] Within a purely statical theory such a force was introduced by Cahn and Hoffman [17], [18], whose work pointed out the need for a term of this form in the energy equation when the interface is anisotropic. The vector $\hat{\xi}(\mathbf{m})$ is actually the tangential part of the vector used by Cahn and Hoffman.

stress, div_{Δ} is the surface divergence, $b(\mathbf{m})$ is an orientation-dependent constant, and we have chosen a scaling in which the latent heat ℓ satisfies $\ell = \theta_M$.

4. FREE BOUNDARY PROBLEMS. Approximate interface conditions are derived for a **weak interface**, that is, one in which the interfacial densities are small and the dependence on V and L weak. These interface conditions, when combined with the usual quasi-static heat equation in bulk, lead to the following system of partial differential equations and free-boundary conditions for the temperature difference $u = \theta - \theta_M$:

$$\begin{aligned} &\mathrm{div}\ \mathbf{q} = 0, && \mathbf{q} = -K_i \nabla u \quad \text{in } B_i, \\ &u = -f(\mathbf{m})H - b(\mathbf{m})V + \mathrm{div}_{\Delta}\xi(\mathbf{m}), && [\mathbf{q}]\cdot\mathbf{m} = \ell V \quad \text{on } \Delta. \end{aligned} \tag{4.1}$$

Here $\Delta = \Delta(t)$ is the interface; K_i is the conductivity tensor for phase i; B_i is the region of space occupied by phase i, $\mathbf{q}$ is the bulk heat flux; $[\mathbf{q}]$ is the jump in $\mathbf{q}$ across the interface.

Global growth conditions are found for the system (4.1). To state these succinctly, consider a bounded solid $B(t)$ in an infinite liquid melt, and write

$$F(\Delta) = \int_{\Delta} f(\mathbf{m})$$

for the total interfacial free-energy computed using the equilibrium values of the corresponding density. Then:

$$\mathrm{vol}(B)^{\cdot} = 0, \qquad F(\Delta)^{\cdot} \leq 0 \tag{4.2}$$

provided the liquid is thermally isolated at infinity, while

$$f(\Delta)^{\cdot} + u_{\infty}\, \mathrm{vol}(B)^{\cdot} \leq 0 \tag{4.3}$$

whenever the liquid is isothermal at infinity. Here u_{∞} is the (constant) far-field temperature-difference.

The results (4.2) and (4.3) motivate two variational problems:

(V1) minimize $F(\Delta)$ subject to $\mathrm{vol}(B) = \text{constant}$;

(V2) minimize $F(\Delta) + u_{\infty}\, \mathrm{vol}(B)$.

The problem (V1) and the problem (V2) with $u_{\infty} > 0$ are well posed. On the other hand, (V2) with $u_{\infty} < 0$ has no solution, as all minimizing sequences

have $\mathrm{vol}(B) \to \infty$. This is as expected: $u_\infty < 0$ corresponds to a solid in a supercooled liquid melt, and $\mathrm{vol}(B) \to \infty$ indicates the ultimate envelopment of the liquid by the more stable solid phase.

ACKNOWLEDGMENT. I gratefully acknowledge the support of this research by the Army Research Office and by the National Science Foundation.

REFERENCES

[1] Chalmers, B., *Principles of Solidification*, Wiley, New York (1964).

[2] Delves, R.R., Theory of interface instability. *Crystal Growth* (ed. B.R. Pamplin), Pergamon, Oxford (1974).

[3] Mullins, W.W. and R.F. Sekerka, Morphological stability of a particle growing by diffusion or heat flow, *J. Appl. Phys.* **34**, 323-329 (1963).

[4] Mullins, W.W. and R.F. Sekerka, Stability of a planar interface during solidification of a dilute binary alloy, *J. Appl. Phys.* **35**, 444-451 (1964).

[5] Voronkov, V.V., Conditions for formation of mosaic structure on a crystallization front, *Sov. Phys. Solid State* **6**, 2378-2381 (1965).

[6] Seidensticker, R.G., Stability considerations in temperature gradient zone melting, *Crystal Growth* (ed. H.S. Peiser), Pergamon, Oxford (1966).

[7] Tarshis, L.A. and W.A. Tiller, The effect of interface-attachment kinetics on the morphological stability of a planar interface during solidification, *Crystal Growth* (ed. H.S. Peiser), Pergamon, Oxford (1966).

[8] Sekerka, R.F., Morphological stability, *J. Crystal Growth* **3,4**, 71-81 (1968).

[9] Sekerka, R.F., Morphological stability, *Crystal Growth: an Introduction*, North-Holland, Amsterdam (1973).

[10] Sekerka, R.F., Morphological instabilities during phase transformations, *Phase Transformations and Material Instabilities in Solids*, (ed. M.E. Gurtin), Academic Press, New York (1984).

[11] Chernov, A.A., Theory of the stability of face forms of crystals, *Sov. Phys. Crystallog.* **16**, 734-753 (1972).

[12] Langer, J.S., Instabilities and pattern formation in crystal growth, *Rev. Mod. Phys.* **52**, 1-27 (1980).

[13] Gurtin, M.E., Toward a nonequilibrium thermodynamics of two phase materials, *Arch Rational Mech. Anal.* Forthcoming

[14] Gurtin, M.E., On the two-phase Stefan problem with interfacial energy and entropy, *Arch Rational Mech. Anal.* **96**, 199-241 (1986).

[15] Coleman, B.D. and W. Noll, The thermodynamics of elastic materials with heat conduction and viscosity, *Arch. Rational Mech. Anal.* **13**, 167-178 (1963).

[16] Murdoch, A.I., A thermodynamic theory of elastic material interfaces, *Q.J. Mech. Appl. Math.* **29**, 245-275 (1976).

[17] Hoffman, D.W. and J.W. Cahn, A vector thermodynamics for anisotropic surfaces - 1. Fundamentals and applications to plane surface junctions, *Surface Sci.* **31**, 368-388 (1972).

[18] Cahn, J.W. and D.W. Hoffman, A vector thermodynamics for anisotropic surfaces, *Act. Metall.* **22**, 1205-1214 (1974).

THE SHEAR-LAG MODEL FOR A UNIDIRECTIONAL COMPOSITE WITH VISCOELASTIC MATRIX

DIMITRIS C. LAGOUDAS
Mathematical Sciences Institute
Cornell University, Ithaca, NY 14853

CHUNG-YUEN HUI
Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

S. LEIGH PHOENIX
Sibley School of Mechanical and Aerospace Engineering
Cornell University, Ithaca, NY 14853

ABSTRACT

The shear-lag model is applied to a monolayer, unidirectional, fiber-reinforced composite loaded in tension. The monolayer contains an infinite number of parallel fibers, with an arbitrary number of them broken simultaneously. While the fibers are modelled as linearly elastic, a linear viscoelastic constitutive law is assumed for the matrix material. The time evolution of the overstress profiles in the fibers and matrix near breaks is determined. The time dependence of the effective load transfer length is also calculated. Explicit evaluations of the above quantities are given for a power-law creep compliance model, suitable for most epoxy thermosetting resins as matrix materials.

INTRODUCTION

The shear-lag model for a unidirectional composite was developed by HEDGEPETH (1961) as an attempt to describe the stress fields around broken fibers. It is a simplified micromechanics model for which closed form solutions can be obtained. In Hedgepeth's analysis the fibers are parallel, equally spaced and of infinite length. The monolayer includes an infinite number of fibers with a cluster of them broken (see Fig. 1) and is loaded by uniformly distributed tensile tractions in the direction of the fibers. Both fiber and matrix materials are assumed to be linearly elastic. The drastic simplification introduced by the shear-lag model is the decoupling between the mechanisms that respond to shear and normal stresses in the composite. It is thus assumed that the fibers alone bear the normal stresses along the fiber direction, while the matrix material acts only as a shear transfer mechanism that overloads the adjacent fibers in tension whenever a fiber breaks.

The method of influence coefficients was used for the solution of the above problem and the explicit evaluation of the overload coefficients of the intact fibers due to fiber breaks was given by HEDGEPETH (1961). Closed-form solutions in terms of Bessel and Weber functions for the overload and displacement fields of the fibers were reported by FICHTER (1969,1970), who also looked into the problem of more than one groups of breaks. A later work by HEDGEPETH and VAN DYKE (1967) incorporates an elastic-perfectly plastic model for the matrix material. In a subsequent work VAN DYKE and HEDGEPETH (1969) assumed that the matrix fails completely when a maximum shear stress is

reached. A modified version of Hedgepeth's shear-lag analysis was undertaken by ERINGEN and KIM (1974), who took into account the normal stresses in the matrix transversely to the direction of the fibers. Along the same lines was the analysis of GOREE and GROSS (1979) with the additional inclusion of longitudinal yielding and splitting of the matrix and later on an extension to the 3-D case (GOREE and GROSS, 1980). Comparisons of the predictions of the shear-lag model with 3-D finite element calculations were done by REEDY (1984). He found excellent agreement between the two methods for the fiber stress concentrations in a Kevlar/epoxy monolayer for load levels that do not cause matrix yielding.

In the present work we analyze the time response predicted by the shear-lag model of a unidirectional, monolayer composite with an infinite number of parallel fibers loaded in tension in the direction of the fibers, by assuming a time dependent consitutive model for the matrix material. We take the matrix to be linearly viscoelastic, and as a special case we investigate the consequences of a power-law, time dependent, creep compliance on the time evolution of the overstress profiles around broken fibers. Such a power-law creep compliance is commonly used to model the time response of epoxy thermosetting resins, which are often used as matrix material for non-metallic composites (POMEROY, 1978). A linear viscoelastic model for the matrix has previously been used by LIFSHITZ and ROTEM (1970) in their statistical theory of failure for composites, where Schapery's approximate technique was used to obtain the time-dependent solution of a shear-lag model that lumped all broken fibers into a single fiber.

In the first section the formulation of the shear-lag problem is presented for a unidirectional composite under tension with broken fibers and a linearly viscoelastic matrix. Also described is the method of solution which uses Laplace transforms and finite cosine transforms. In the second section a power-law creep compliance is assumed for the matrix, and explicit evaluations of the overloads in the adjacent intact fibers, the shear stresses in the matrix and the effective load transfer length are carried out.

1. FORMULATION OF THE SHEAR-LAG PROBLEM

The model of a thin, unidirectional laminate is shown in Fig. 1, where all fibers are identical and parallel to the X axis and have an equal center-line spacing H. The laminate is considered to be a two-dimensional infinite region with an infinite number of fibers, out of which (2N+1) neighboring fibers are broken along the Y axis at time T = 0. We are interested in calculating the subsequent stress fields near the breaks in the fibers and the matrix.

Both the X and Y axes are axes of symmetry for the laminate in terms of geometry and loading. The external loading is uniform tension applied in the direction of the fibers, which are taken to be the only tensile load carriers. This is a justifiable assumption for most non-metallic composites because the Young's modulus of the matrix is usually one or more orders of magnitude less than the axial Young's modulus of the fibers.

The thickness of the laminate B and the fiber spacing H are of the same order as the diameter of the fibers D, which is small compared to the length of the fibers L. If we take as a reference length unit the fiber diameter D, then $L \to \infty$. The width of the laminate becomes infinite in this length scale as well, as it consists of a large numbers of fibers. The infinite laminate model is therefore a good approximation to the real configuration of the composite, at least before extensive breaking of the fibers has taken place. If the clusters of breaks are not sufficiently far away from each other, their interactions should be taken into account. However, in the linear theory the

superposition principle can be applied, and the problem reduces again to the infinite domain problem with only one group of breaks (the (2N+1) broken fibers in our analysis).

The mechanism of the shear-lag model is a highly idelized one. In the absence of breaks the whole laminate is in a homogeneous stress state with the only nono-zero stresses being the constant normal stresses $4P_\infty/\pi D^2$ in the axial direction of the fibers. The load P_∞ is the constant tensile load applied to each fiber at an infinite distance from the breaks. The matrix material is normal stress free before any breaks occur. This is true if sufficient time has elapsed from the loading of the composite so that stress relaxation in the matrix has occured. Approximately the above is true for any time, since we have assumed that the fibers are much stiffer than the matrix in tension. As soon as one fiber breaks, the load of that fiber near the break is transferred to the neighboring fibers by means of shear forces, which are exerted on the matrix material through the fiber-matrix interface.

A free body diagram of an infinitesimal portion of the n^{th} fiber together with its surrounding matrix is shown in Fig. 2. Even though the fibers are cylindrical and the stress fields in the laminate are inherently three-dimensional, we simplify the problem by first assuming constant normal stresses in all cross-sections perpendicular to the fiber axis. We then assume constant shear stresses in the matrix in the XZ plane in the Z direction, and in the Y direction between two neighboring fibers. To justify the last assumption we introduce an effective width H_f of the matrix layer between two neighboring fibers, such that the product (BH_f) gives the matrix cross-sectional area $(BH - \pi D^2/4)$ between these fibers. It is obvious from the above that the effective width H_f must be equal to $(H - \pi D^2/4B)$. If B is substantially larger than D, the requirement of constant shear stresses in the Z direction is not valid any more, and an effective thickness B_f has to be introduced. As a first approximation we can choose $B_f = D$, in wich case $H_f = H - \pi D/4$. The assumptions about the effective width and the effective thickness require the notion of an effective shear modulus for the matrix, to be determined by experiments. The effective shear modulus will in general be different for different cross-sectional geometries of the fibers and different ratios B/D. Detailed discussion on the selection of H_f and B_f is given by REEDY (1984). Further simplifications introduced by the shear-lag model concern the normal stresses in the matrix in the X direction, which are neglected for reasons mentioned earlier. The normal stresses in the matrix in the Y direction are assumed to remain constant throughout the effective width of the matrix. Any out-of-plane stresses in the fibers and in the matrix are neglected as well, as the problem is assumed to be two-dimensional in the above introduced effective configuration.

By taking into account the above simplifications and in the absence of inertial forces, equilibrium of forces in the X and Y directions (see Fig. 2) results in the following equations:

$$\frac{\partial P_n}{\partial X} + B_f(\tau_{n+1} - \tau_n) = 0 \quad , \quad -\infty < n < \infty \quad , \tag{1}$$

$$\Sigma^m_{n+1} - \Sigma^m_n + \frac{H_f}{2}\left(\frac{\partial \mathcal{T}_{n+1}}{\partial X} + \frac{\partial \mathcal{T}_n}{\partial X}\right) = 0 \quad , \quad -\infty < n < \infty \quad , \tag{2}$$

where P_n is the normal load in the n^{th} fiber, $\mathcal{T}_n$ is the shear stress in the matrix between the n^{th} and $(n-1)^{th}$ fibers, and Σ^m_n is the normal stress in the matrix between the n^{th} and $(n-1)^{th}$ fibers in the Y direction. Eqn (1) implies that the variation of the normal load along a fiber is due to the difference in the shear stresses applied by the matrix on both sides of that fiber. Eqn (2) implies that the dependence of the matrix shear stress on the distance from the breaks results in non-zero normal stresses in the matrix in the Y direction. These normal stresses are maximum near the breaks, where we expect the largest variation in the shear stress, and they might be important in the analysis of the fiber-matrix interface, for example in the case of debonding. Note that equilibrium of moments does not hold in the infinitesimal element of Fig. 2, as a result of neglecting the shear stresses in the fiber cross-sections in the Y direction, unless we assume that the ratio D/H_f is very small. Since D is of the same order as H_f for most applications, we propose the use of a correction factor that restores balance of moments by replacing $\mathcal{T}_n$ with $H\mathcal{T}_n/H_f$, $-\infty < n < \infty$, in eqn (2).

Upon specifying constitutive relations for the matrix and fibers, the above set of equations becomes field differential-difference equations for the determination of the displacement fields U_n and V_n of the n^{th} fiber along the X and Y directions, respectively, as functions of position X and time T. In the present work we assume that the fibers are linearly elastic, namely

$$P_n = AE \frac{\partial U_n}{\partial X} \quad , \tag{3}$$

where A is the fiber cross-sectional area and E is the axial Young's modulus of the fibers. The matrix material is taken to be linearly viscoelastic in shear, that is

$$\mathcal{T}_n = \int_{-\infty}^{T} G(T-S) \frac{\partial \gamma_n(X,S)}{\partial S} dS \quad , \tag{4}$$

where G(T) is the relaxation modulus and $\gamma_n(X,T)$ is the shear strain in the matrix. In order to decouple the system of eqns (1) and (2) in U_n and V_n, we approximately take $\gamma_n \cong (U_n - U_{n-1})/H_f$ by neglecting the term $\partial_X(V_{n-1} + V_n)/2$ (ERINGEN and KIM, 1974), in which case (4) reduces to

$$\mathcal{T}_n = \frac{1}{H_f}\left[\int_{-\infty}^{T} G(T-S) \frac{\partial U_n}{\partial S} dS - \int_{-\infty}^{T} G(T-S) \frac{\partial U_{n-1}}{\partial S} dS\right] \quad . \tag{5}$$

We nondimensionalize the time variable by dividing T by some characteristic time T_0 of the matrix material, to be found by creep experiments, so that $t \equiv T/T_0$. We also define a normalized relaxation modulus

$\mathcal{G}(t) \equiv G(tT_0)/G_0$, where G_0 is the instantaneous elastic shear modulus of the matrix material. (In this work lower case letters and script letters, except for the script letter $\mathcal{T}_n$ used to denote dimensional shear stress, denote dimensionless quantities, while upper case letters stand for dimensional quantities.)

If we introduce the integral operator $\mathcal{J}$ so that its action on a function $f(\zeta)$ is given by

$$\mathcal{J}\langle f\rangle \equiv \int_{-\infty}^{t} \mathcal{G}(t-\zeta)\,\frac{\partial f}{\partial \zeta}\,d\zeta \quad , \tag{6}$$

substitution of (3) and (5) into (1), upon using (6), yields second order differential-difference equations for the determination of U_n, namely

$$\frac{AEH_f}{G_0B_f}\,\frac{\partial^2 U_n}{\partial X^2} + \mathcal{J}\langle U_{n+1} - 2U_n + U_{n-1}\rangle = 0 \quad , \quad -\infty < n < \infty \quad . \tag{7}$$

If the solution to (7) can be found, substitution of U_n into (5) yields the shear stresses $\mathcal{T}_n$ and hence eqns (2) can be solved for V_n. V_n can be easily determined if a linearly elastic constitutive model is selected for the normal stresses in the matrix perpendicular to the fiber direction, i.e., $\Sigma_n^m = E_m(V_n - V_{n-1})/H_f$ (E_m is the effective Young's modulus for the matrix). If we use a linear viscoelastic model for the normal stresses, the Laplace transform method can be used to render eqns (2) algebraic in $\bar{V}_n$, the Laplace transformed displacement V_n. The decoupling of the vertical and horizontal displacements allows us to consider only eqns (7) in our solution procedure.

X and U_n are normalized so that the field equations and the boundary and initial conditions are independent of the material parameters. If we select $x \equiv X/X_0 \equiv X/\sqrt{AEH_f/G_0B_f}$ and $u_n(x,t) \equiv U_n(X,T)/\sqrt{P_\infty^2 H_f/G_0 AEB_f}$, eqns (7) become

$$\frac{\partial^2 u_n}{\partial x^2} + \mathcal{J}\langle u_{n+1} - 2u_n + u_{n-1}\rangle = 0 \quad , \quad -\infty < n < \infty \quad . \tag{8}$$

The boundary conditions are given by

$$\frac{\partial u_n}{\partial x} = 1 \quad , \quad -\infty < n < \infty \quad , \quad x \to \infty \quad , \quad t > 0 \quad , \tag{9a}$$

$$\frac{\partial u_n}{\partial x} = 0 \quad , \quad -N \le n \le N \quad , \quad x = 0 \quad , \quad t > 0 \quad , \tag{9b}$$

$$u_n = 0 \quad , \quad -\infty < n < -N \quad , \quad N < n < \infty \quad , \quad x = 0 \quad , \quad t > 0 \quad , \tag{9c}$$

while the initial conditions are

$$u_n = x \quad , \quad -\infty < n < \infty \quad , \quad x \geq 0 \quad , \quad t = 0 \quad . \tag{10}$$

In order to avoid unbounded displacement fields in the analysis, we perform the transformation

$$w_n = u_n - x \quad , \tag{11}$$

which, after its substitution into eqns (8), (9) and (10), results in the following field equations, boundary and intital conditions:

$$\frac{\partial^2 w_n}{\partial x^2} + \mathcal{P}\langle w_{n+1} - 2w_n + w_{n-1}\rangle = 0 \quad , \quad -\infty < n < \infty \quad , \tag{12}$$

$$\frac{\partial w_n}{\partial x} = 0 \quad , \quad -\infty < n < \infty \quad , \quad x \to \infty \quad , \quad t > 0 \quad , \tag{13a}$$

$$\frac{\partial w_n}{\partial x} = -1 \quad , \quad -N \leq n \leq N \quad , \quad x = 0 \quad , \quad t > 0 \quad , \tag{13b}$$

$$w_n = 0 \quad , \quad -\infty < n < -N \quad , \quad N < n < \infty \quad , \quad x = 0 \quad , \quad t > 0 \quad , \tag{13c}$$

$$w_n = 0 \quad , \quad -\infty < n < \infty \quad , \quad x \geq 0 \quad , \quad t = 0 \quad . \tag{14}$$

Notice that the field equations remain unchanged in form. This is because the transformation (11) is a time independent translation. The change in the boundary conditions has altered the original problem into a new one, in which there are no loads at infinity and there are only compressive loads applied on the broken fibers suddenly at $t = 0$, which open up the breaks as t grows.

The above equations can be solved by using Laplace transforms. The Laplace transform of $\mathcal{P}\langle w_n\rangle$ is given by the convolution law $L(\mathcal{P}\langle w_n\rangle) = s\bar{\mathcal{G}}(s)\,\bar{w}_n(x,s)$, where $\bar{\mathcal{G}}(s)$ and $\bar{w}_n(x,s)$ are the Laplace transforms of $\mathcal{G}(t)$ and $w_n(x,t)$, respectively. The Laplace transforms of (12) and (13), upon using (14), become

$$\frac{\partial^2 \bar{w}_n(x,s)}{\partial x^2} + s\bar{\mathcal{G}}(s)[\bar{w}_{n+1}(x,s) - 2\bar{w}_n(x,s) + \bar{w}_{n-1}(x,s)] = 0, \quad -\infty < n < \infty \quad , \tag{15}$$

$$\frac{\partial \bar{w}_n}{\partial x} = 0 \quad , \quad -\infty < n < \infty \quad , \quad x \to \infty \quad , \tag{16}$$

$$\frac{\partial \bar{w}_n}{\partial x} = -\frac{1}{s} \quad , \quad -N \leq n \leq N \quad , \quad x = 0 \quad , \tag{17a}$$

$$\bar{w}_n = 0 \quad , \quad -\infty < n < -N \quad , \quad N < n < \infty \quad , \quad x = 0 \quad . \tag{17b}$$

We have thus transformed the original viscoelastic problem into an elastic shear-lag problem (correspondence principle, CHRISTENSEN, 1982). We will follow here the methodology presented by ERINGEN and KIM (1974) and used also by GOREE and GROSS (1979) for the solution of the elastic shear-lag problem, which is a dual integral equations technique. However, one can also use the influence function technique developed by HEDGEPETH (1961).

We reduce eqns (15) to a single differential equation by introducing the finite cosine transform (CHURCHILL, 1972). Define $\bar{w}$ by

$$\bar{w} = \frac{\bar{w}_0}{\pi} + \frac{2}{\pi} \sum_{n=1}^{\infty} \bar{w}_n \cos(n\theta) \quad , \quad 0 < \theta < \pi \quad , \tag{18a}$$

with the inversion formula given by

$$\bar{w}_n = \int_0^{\pi} \bar{w} \cos(n\theta) \, d\theta \quad . \tag{18b}$$

where $\bar{w} \equiv \bar{w}(x,s,\theta)$, $\bar{w}_n \equiv \bar{w}_n(x,s)$. By summing eqns (15) with n running from $-\infty$ to ∞, after having multiplied them by $\cos(n\theta)$, and by taking into account the symmetry $\bar{w}_n(x,s) = \bar{w}_{-n}(x,s)$, it is found that $\bar{w}$ satisfies

$$\frac{\partial^2 \bar{w}}{\partial x^2} - 4s\bar{g}(s) \sin^2(\theta/2) \, \bar{w} = 0 \quad . \tag{19}$$

The resulting simplification in the field equations has shifted the difficulty into the boundary conditions, which turn out to be integral equations, namely

$$\frac{\partial \bar{w}}{\partial x} = 0 \quad , \quad x \to \infty \quad , \tag{20}$$

$$\int_0^{\pi} \frac{\partial \bar{w}}{\partial x} \cos(n\theta) \, d\theta = -\frac{1}{s} \quad , \quad 0 \leq n \leq N \quad , \quad x = 0 \quad , \tag{21a}$$

$$\int_0^{\pi} \bar{w} \cos(n\theta) \, d\theta = 0 \quad , \quad N < n < \infty \quad , \quad x = 0 \quad . \tag{21b}$$

A solution to (19) that satisfies the boundary condition (20) is given by

$$\bar{w} = f(s,\theta) \exp[-2\sin(\theta/2) \, x\sqrt{s\bar{g}(s)} \,] \quad , \tag{22}$$

for some $f(s,\theta)$. Substitution of (22) into (21a) and (21b) yields the conditions

$$\int_0^{\pi} f(s,\theta)\,\sin(\theta/2)\,\cos(n\theta)\,d\theta = \frac{1}{2s\sqrt{s\bar{\mathcal{G}}(s)}} \quad , \quad 0 \le n \le N \quad , \tag{23a}$$

$$\int_0^{\pi} f(s,\theta)\,\cos(n\theta)\,d\theta = 0 \quad , \quad N < n < \infty \quad , \tag{23b}$$

for $f(s,\theta)$. By letting $f(s,\theta) \equiv \sum_{m=0}^{N} a_m(s)\,\cos(m\theta)$, the conditions (23a) for the broken fibers reduce to

$$\sum_{m=0}^{N} a_m(s) \int_0^{\pi} \sin(\theta/2)\cos(n\theta)\cos(m\theta)d\theta = \frac{1}{2s\sqrt{s\bar{\mathcal{G}}(s)}} \quad , \quad 0 \le n \le N \quad , \tag{24}$$

while conditions (23b) for the unbroken fibers are satisfied identically. The complete satisfaction of the boundary conditions reduces then to the solution of the algebraic sustem (24) of (N+1) equations, for the determination of the (N+1) unknown functions $a_m(s)$, $m = 0,1,2,\ldots,N$. The solution to the transformed problem is found by substituting $\bar{w}$ from (22) into (18b) and is given by the following expression:

$$\bar{w}_n(x,s) = \sum_{m=0}^{N} a_m(s) \int_0^{\pi} \exp[-2\sin(\theta/2)\sqrt{s\bar{\mathcal{G}}(s)}\,x]\,\cos(m\theta)\cos(n\theta)d\theta \quad . \tag{25}$$

The inversion of the Laplace transforms of $\bar{w}_n$ will result in $w_n(x,t)$. The difficulty of the inversion will mainly depend on the selection of the constitutive model (i.e., $\bar{\mathcal{G}}(s)$) for the viscoelastic matrix.

A clarifying remark regarding the number of broken fibers is mentioned at this point. We have assumed that the number of breaks is an odd integer, namely 2N+1, and as a consequence we have used the finite cosine transform (18), taking into account the symmetry of $\bar{w}_n$ about x axis. We could easily model any number of breaks by using the finite exponential transform (CHURCHILL, 1972), which is given by

$$\bar{w}(x,s,\theta) = \frac{1}{2\pi} \sum_{n=-\infty}^{n=+\infty} \bar{w}_n(x,s)\,\exp(in\theta) \quad . \tag{18c}$$

$$\bar{w}_n(x,s) = \int_{-\pi}^{\pi} \bar{w}(x,s,\theta)\,\exp(-in\theta)\,d\theta \quad . \tag{18d}$$

and reduces to the finite cosine transform whenever $\bar{w}_n = \bar{w}_{-n}$, or $\bar{w}$ is symmetric in θ. The only change in the previous analysis is that now $f(s,\theta) = \sum_{m=-M}^{N} a_m(s) \cdot \exp(-im\theta)$, where the total number of breaks is (M+N+1) and the algebraic system

(24) involves (M+N+1) unknown functions $a_m(s)$.

The important quantities in the analysis of shear-lag models are the overloads in the fibers near breaks and the shear stresses in the matrix. The nondimensional loads in the fibers, defined by $p_n(x,t) \equiv P_n(xX_0, tT_0)/P_\infty$, can be found by substituting $w_n(x,t)$ from (25) into (3) upon using (11), and they are given by

$$p_n(x,t) = \frac{\partial w_n(x,t)}{\partial x} + 1 \quad , \quad n \geq 0 \quad . \tag{26}$$

The normalized shear stresses $\tau_n(x,t) \equiv \mathcal{T}_n(xX_0, tT_0)/\sqrt{P_\infty^2 G_0/AEB_f H_f}$ between the n^{th} and the $(n-1)^{th}$ fibers are evaluated by substitution of $w_n(x,\tau)$ into (5) (which upon using (11) yields the normalization), and they are given by

$$\tau_n(x,t) = \int_0^t \mathcal{G}(\tau-\zeta) \frac{\partial(w_n - w_{n-1})}{\partial\zeta} d\zeta \quad , \quad n \geq 1 \quad . \tag{27}$$

Another useful quantity, especially for statistical models of failure of composites (PHOENIX and TIERNEY, 1983), is the effective load transfer length L_f, which for present purposes is defined as the distance from the breaks in the x direction, within which the overload of the first unbroken fiber has dropped to zero. Since in the shear-lag model the load P_{N+1} of the first intact fiber actually descends to values below P_∞ before it decays exponentially to P_∞ as $x \to \infty$, we define L_f as the distance from the breaks at which P_{N+1} crosses P_∞. In this case L_f or equivalently the normalized effective load transfer length $l_f \equiv L_f/\sqrt{AEH_f/G_0B_f}$ must satisfy the conditions

$$p_{N+1}(l_f,t) = 1 \quad , \text{ or } \quad \frac{\partial w_{N+1}(l_f,t)}{\partial x} = 0 \quad . \tag{28}$$

In general l_f will depend on time because p_{N+1} depends on time. The so defined l_f becomes a characteristic length for the whole laminate for a given number of breaks (2N+1).

We summarize the results of this section by giving explicit evaluations for the various quantities. If we define $b_m \equiv a_m(s)\ 2s\sqrt{s\bar{\mathcal{G}}(s)}$, then b_m are determined by solving the algebraic system

$$\sum_{m=0}^{N} b_m \int_0^\pi \sin(\theta/2)\cos(n\theta)\cos(m\theta)d\theta = 1 \quad , \quad 0 \leq n \leq N \quad , \tag{29}$$

which is independent of s. Eqns (25),(26) and (27) reduce to

$$w_n(x,t) = \sum_{m=0}^{N} b_m \int_0^{\pi} L^{-1}\{\exp[-2\sin(\theta/2)\sqrt{s\bar{\mathcal{G}}(s)}\, x]/(2s\sqrt{s\bar{\mathcal{G}}(s)})\}$$

$$\cdot \cos(m\theta)\cos(n\theta)d\theta \quad , \tag{30}$$

$$p_n(x,t) = 1 - \sum_{m=0}^{N} b_m \int_0^{\pi} L^{-1}\{\exp[-2\sin(\theta/2)\sqrt{s\bar{\mathcal{G}}(s)}\, x]/s\}$$

$$\cdot \sin(\theta/2)\cos(m\theta)\cos(n\theta)d\theta \quad , \tag{31}$$

$$\tau_n(x,t) = \sum_{m=0}^{N} b_m \int_0^{\pi} L^{-1}\{\exp[-2\sin(\theta/2)\sqrt{s\bar{\mathcal{G}}(s)}\, x]\ \sqrt{s\bar{\mathcal{G}}(s)}/2s\}$$

$$\cdot \cos(m\theta)[\cos(n\theta) - \cos((n-1)\theta)]d\theta \quad , \tag{32}$$

where

$$f(t) = L^{-1}[\bar{f}(s)] \equiv \frac{1}{2\pi i} \lim_{\beta\to\infty} \int_{\gamma-i\beta}^{\gamma+i\beta} \exp(ts)\bar{f}(s)ds \quad , \quad t > 0 \ . \tag{33}$$

2. POWER-LAW CREEP COMPLIANCE MODEL FOR THE MATRIX MATERIAL

A useful model that describes closely the viscoelastic properties of commercially used matrix materials (epoxy thermosetting resins) is a power-law creep compliance model that can be expressed in the form

$$J(T) = J_0\ [1 + (\frac{T}{T_0})^{\alpha}\ ] = J_0(1 + t^{\alpha}) \equiv J_0\ \mathcal{J}(t) \quad . \tag{34}$$

Here J_0 characterizes the instantaneous elastic response of the matrix material under loading and T_0 and α are material constants that describe the creep behavior under dead loading. The characteristic time T_0 is the time required for the initial displacement to be doubled, while the exponent α is usually much smaller than unity. The limit $\alpha \to 0$ corresponds to the elastic case, while $\alpha \to 1$ gives a linear time dependence which is equivalent to the Maxwell viscoelastic model. The connection between the relaxation modulus $\mathcal{G}(t)$ and the creep compliance $\mathcal{J}(t)$ is expressed through the Laplace transformed quantities (CHRISTENSEN, 1982) by the well-known formula

$$\bar{\mathcal{G}}(s)\ \bar{\mathcal{J}}(s)\ s^2 = 1 \quad , \tag{35}$$

if $G_0 = 1/J_0$. From (34) and (35) the Laplace transform of the relaxation modulus is found to be

$$s\ \bar{\mathcal{G}}(s) = \frac{s^{\alpha}}{s^{\alpha} + \Gamma(\alpha+1)} \quad . \tag{36}$$

By inserting (36) into (30),(31) and (32) it is possible to obtain explicit evaluations for w_n, p_n and τ_n in terms of x and t for different values of α. The inversion of the Laplace transforms has been done by contour integration. We will only report the solution here for the fiber loads and the shear stresses, while the displacement fields can be obtained by integrating (26). The fiber loads and the shear stresses are found to be

$$p_n(x,t) = 1 - \sum_{m=0}^{N} b_m \int_0^{\pi} h(x,t,\theta)\cos(m\theta)\cos(n\theta)\sin(\theta/2)\,d\theta \quad . \tag{37}$$

$$\tau_n(x,t) = \sum_{m=0}^{N} b_m \int_0^{\pi} g(x,t,\theta)\cos(m\theta)[\cos(n\theta) - \cos(n-1)\theta)]d\theta \quad . \tag{38}$$

where the functions $h(x,t,\theta)$ and $g(x,t,\theta)$ are given by

$$h(x,t,\theta) = 1 - \frac{1}{\pi}\int_0^{\infty} \exp(-tr)\,\exp[-\lambda\sqrt{\frac{r^{\alpha}}{\rho}}\cos(\frac{\alpha\pi-\varphi}{2})]\,\sin[\lambda\sqrt{\frac{r^{\alpha}}{\rho}} \cdot \sin(\frac{\alpha\pi-\varphi}{2})]\frac{dr}{r} \quad , \tag{39}$$

$$g(x,t,\theta) = -\frac{1}{2\pi}\int_0^{\infty} \exp(-tr)\,\exp[-\lambda\sqrt{\frac{r^{\alpha}}{\rho}}\cos(\frac{\alpha\pi-\varphi}{2})]\,\sin[\lambda\sqrt{\frac{r^{\alpha}}{\rho}}\sin(\frac{\alpha\pi-\varphi}{2}) - (\frac{\alpha\pi-\varphi}{2})]\sqrt{\frac{r^{\alpha}}{\rho}}\frac{dr}{r} \quad . \tag{40}$$

The quantities λ, ρ and φ have the following evaluations:

$$\lambda = 2\sin(\theta/2)\,x \quad , \tag{41}$$

$$\rho = \sqrt{[r^{\alpha}\cos(\alpha\pi) + \Gamma(\alpha+1)]^2 + [r^{\alpha}\sin(\alpha\pi)]^2} \quad , \tag{42}$$

$$\varphi = \tan^{-1}\left[\frac{r^{\alpha}\sin(\alpha\pi)}{r^{\alpha}\cos(\alpha\pi) + \Gamma(\alpha+1)}\right] \quad , \quad 0 < \varphi < \pi \quad . \tag{43}$$

Numerical integration of the above formulae has been carried out for both p_n and τ_n, even though they are related through (1). The reason for this is that p_n is usually the quantity of primary interest and the numerical evaluaton of τ_n from p_n involves differentiaton which should be avoided. Numerical integration has been done by using a midpoint Romberg integration technique, with an appropriate change of variables at the singular points of the integrands. The results are plotted in Figs 3, 4, 5 and 6, for one and three

broken fibers and for the first and second intact fibers for various times (α = 0.1 for all cases).

From Figs 3 and 4 we notice that at x = 0 we recover the overload coefficients ($P_n(x=0,t)/P_\infty \equiv p_n(x=0,t)$) in accordance with the elastic solution of HEDGEPETH (1961). The overload coefficient of the first intact fiber in a laminate with N neighboring breaks as calculated by Hedgepeth is given by

$$K_N = \frac{4 \cdot 6 \cdot 8 \cdot \cdots \cdot (2N+4)}{3 \cdot 5 \cdot 7 \cdot \cdots \cdot (2N+3)} \quad , \quad 0 \leq N \leq \infty \quad . \tag{44}$$

The above formula holds for the viscoelastic case as well because the overall static equilibrium of the composite is not affected by the viscoelastic properties of the matrix material. This is because the matrix material cannot sustain normal loads in the x direction and there is no stress relaxation in the fibers as they are assumed to be elastic. Therefore, the excess load caused by the simultaneous breaks has to be shared by the neighboring intact fibers and only the stress distributions are affected by the viscoelastic properties of the matrix material.

Several observations can be drawn from Figs 3 and 4. The slope of the stress distribution in the fibers decreases as time increases, resulting in a growth of the effective load transfer length l_f with time (Figs 3a, 4a). The overload undershoots and actually becomes negative before it decays to zero as $x \to \infty$ for the intact fibers. Global equilibrium of the composite in the x direction implies that $\Sigma[p_n(x,t) - 1] = 0$, with summation extending to all fibers. Since the negative overloads in the broken fibers grow with time, as a result of the shear stress relaxation in the matrix, the positive overloads in the intact fibers increase with time for fixed x, so that global equilibrium is satisfied (see Figs 3 and 4). This implies that the probability of failure for the intact fibers near breaks increases with time. The length over which this increased probability occurs also grows with time, this being the effective load transfer length l_f.

The relaxation of the shear stresses in the matrix can be seen in Figs 5 and 6. The shear-lag model gives inaccurate results for the shear stresses near the breaks (whithin one or two fiber diameters). The shear stresses in the matrix should go to zero at the break points and this is clearly violated according to the numerical results in Figs 5a and 6a. Modifications, like for example the correction in the calculation of the shear strain introduced by ERINGEN and KIM (1974), are consistent with continuum mechanics but in reality they are not accurate either. The reason for this is that debonding in the fiber-matrix interface near the breaks usually occurs due to the high stress concentrations there. This changes drastically the geometry in a small neighborhood around the breaks, and leads to additional plastic deformations in the matrix. Nevertheless, finite element results for the elastic shear-lag model (REEDY, 1984) indicate that the shear-lag model predicts correctly the stress concentrations in the intact fibers. Even though it is an approximate model, the shear-lag model for the viscoelastic case unravels the trend in the time dependence of the stress fields near broken fibers. Note that since the fibers are much stiffer than the matrix ($\cong$100), the region in which the stresses are perturbed due to fiber breaks is 50 or more fiber diameters, while the shear-lag analysis fails to predict correctly the shear stresses in a small region of one or two fiber diameters away from the breaks.

ACKNOWLEDGEMENTS

The support of the U. S. Army Research Office through the Mathematical Sciences Institute of Cornell University is gratefully acknowledged.

REFERENCES

1. CARRIER, G.F and M. KROOK and C.E. PEARSON, Functions of a Complex Variable, McGraw-Hill, New York (1966).
2. CHRISTENSEN, R.M., Theory of Viscoelasticity, Academic Press, New York (1971).
3. CHURCHILL, R. V., Operational Mathematics, 2nd Edn, McGraw-Hill, Tokyo (1972).
4. ERINGEN, A.C. and B.S. KIM, Stress concentration in filamentary composites with broken fibers, Letters in Applied and Engineering Sciences 2, 69-89 (1974).
5. FICHTER, W.B., Stress concentration around broken filaments in a filament-stiffened sheet, NASA TN D-5453, Langley Research Center (1969).
6. FICHTER, W.B., Stress concentrations in filament-stiffened sheets of finite length, NASA tn D-5947, Langley Research Center (1970).
7. GOREE, J.G. and R.S. GROSS, Analysis of a unidirectional composite containing broken fibers and matrix damage, Engineering Fracture Mechanics 13, 563-578 (1979).
8. GOREE, J.G. and R.S. GROSS, Stresses in a three-dimensional unidirectional composite containing broken fibers, Engineering Fracture Mechanics 13, 395-405 (1980).
9. HEDGEPETH, J., Stress concentrations in filamentary structures, NASA TN D-882, Langley Research Center (1961).
10. HEDGEPETH, J. and P. VAN DYKE, Local stress concentrations in imperfect filamentary composite materials, J. Composite Materials 1, 294-309 (1967).
11. LIFSHITZ, J.M. and A. ROTEM, Time-dependend longitudinal strength of unidirectional fibrous composites, Fibre Science and Technology 3, 1-20 (1970).
12. PHOENIX, S.L. and L-J. TIERNEY, A statistical model for the time dependent failure of unidirectional composite materials under local elastic load-sharing among fibers, Engineering Fracture Mechanics 18, 193-215 (1983).
13. POMEROY, C.D. (Editor), Creep of engineering materials, J. Strain Analysis Monograph, I. Mech. E. (1978)
14. REEDY, E.D., Jr., Fiber stresses in a cracked monolayer: comparison of shear-lag and 3-D finite element predictions, J. Composite Materials 18, 595-607 (1984).
15. VAN DYKE, P. and J.M. HEDGEPETH, Stress concentrations from single-filament failures in composite materials, Textile Research Journal 39, 618-626 (1969).

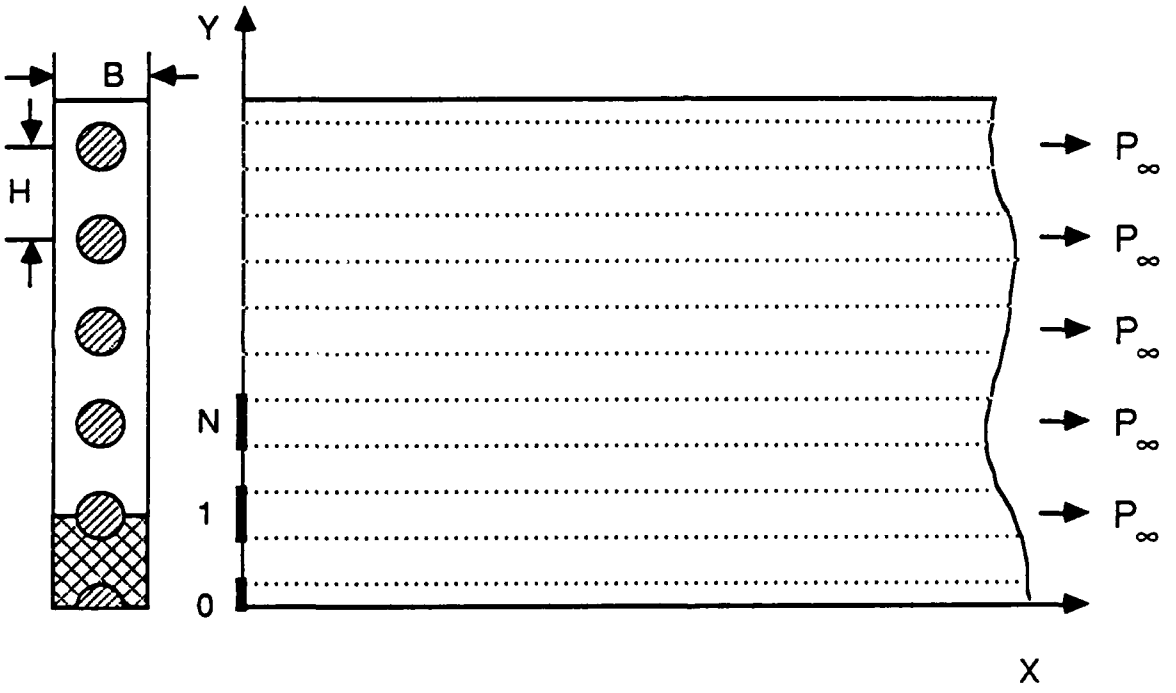


Fig. 1. A unidirectional composite with an infinite number of parallel fibers loaded in tension uniformly, with (2N+1) broken fibers along Y axis.

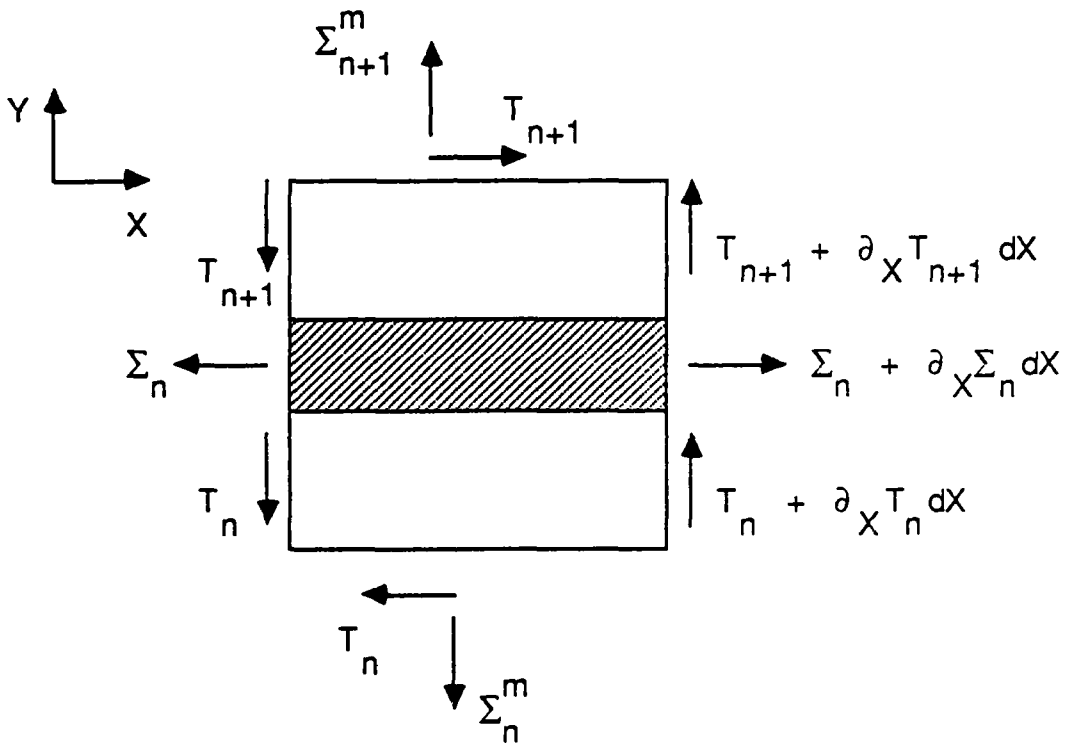


Fig. 2. Equilibrium of an infinitesimal element of the n^{th} fiber with its surrounding matrix.

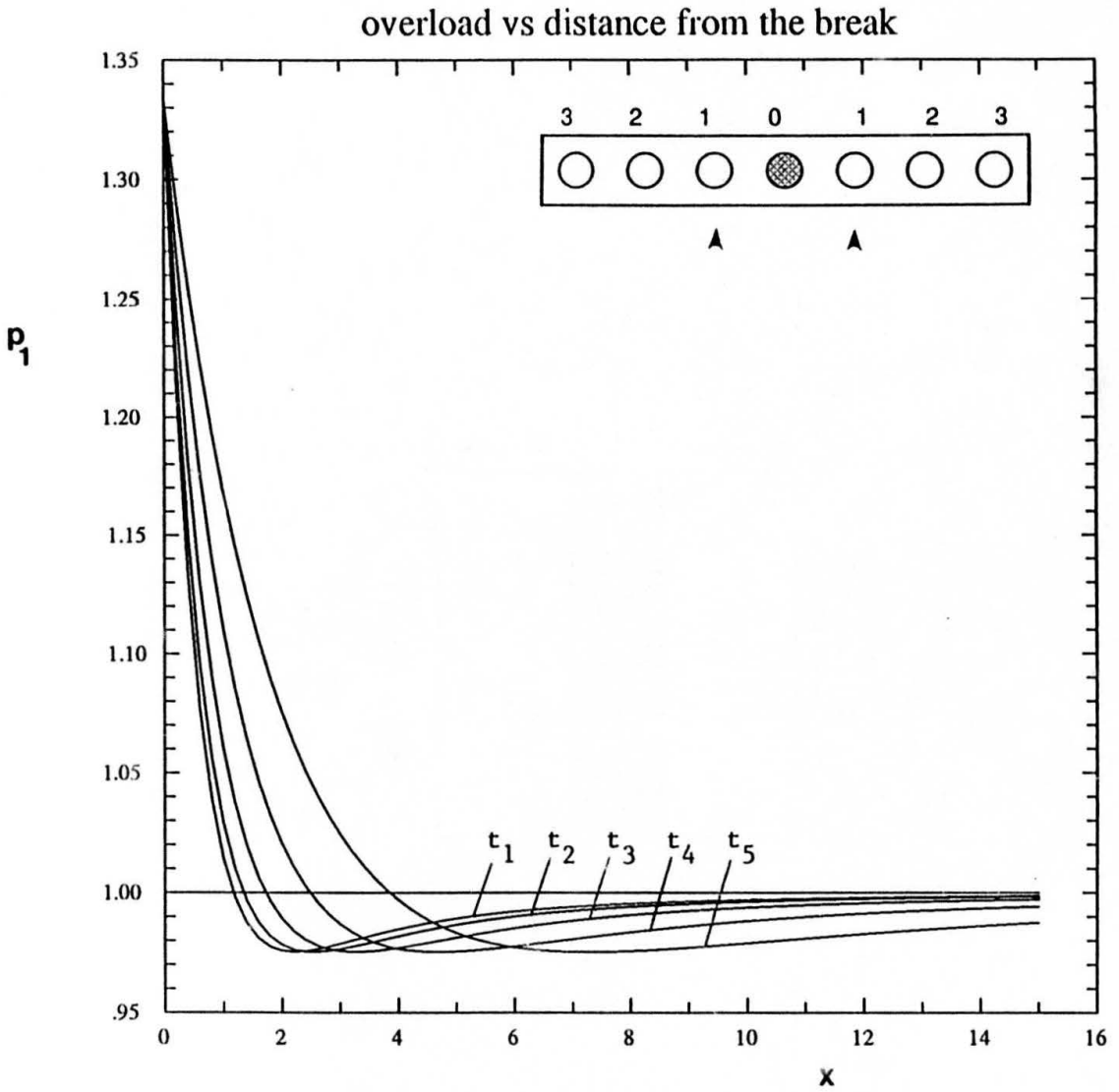


Fig. 3a. Load profile of the first (adjacent) intact fiber at various times; one break. ($t_1 = 3.059\text{x}10^{-7}$, $t_2 = 6.738\text{x}10^{-3}$, $t_3 = 1.484\text{x}10^{2}$, $t_4 = 3.269\text{x}10^{6}$, $t_5 = 7.200\text{x}10^{10}$, $\alpha = 0.1$).

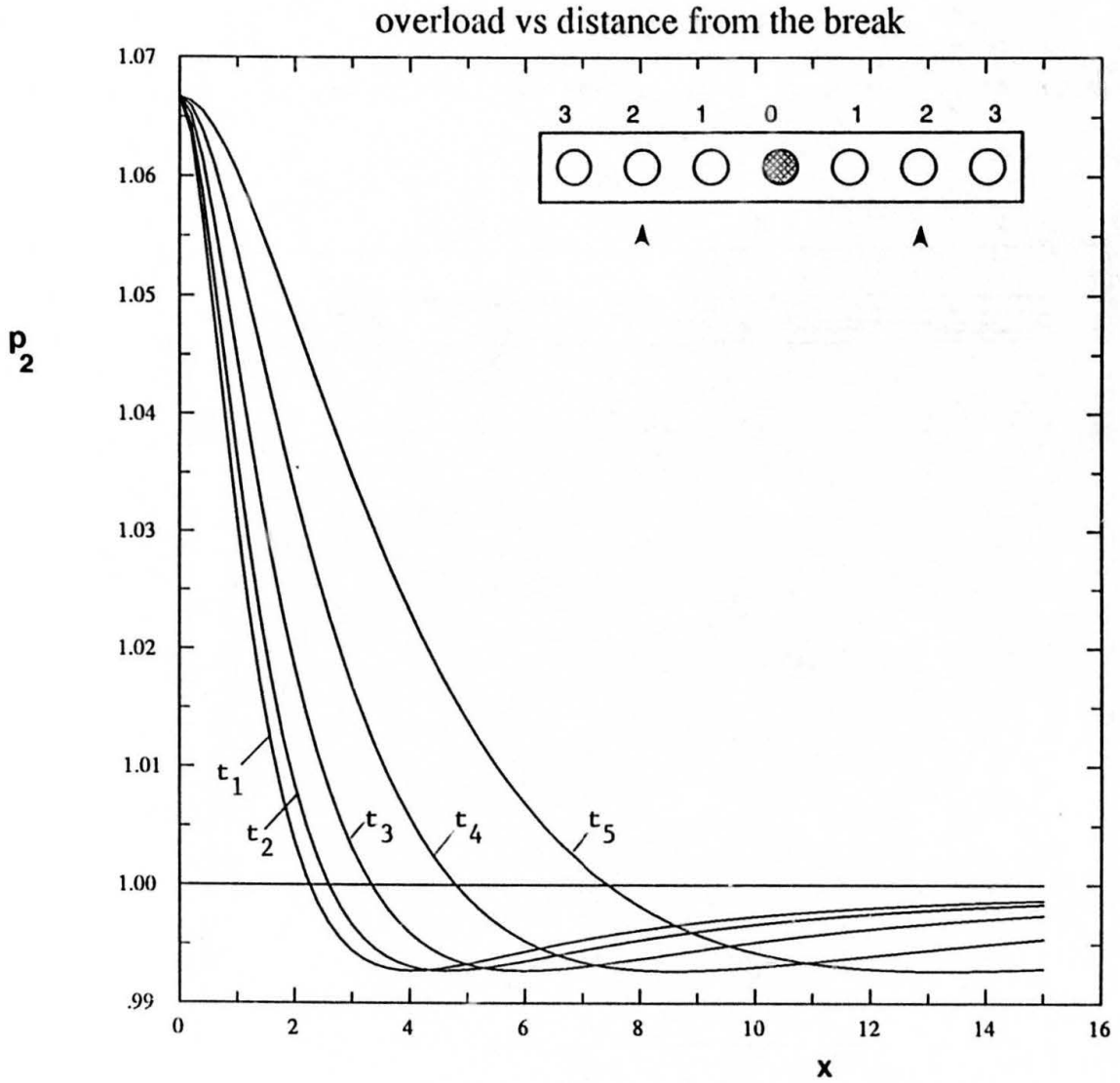


Fig. 3b. Load profile of the second intact fiber at various times; one break.

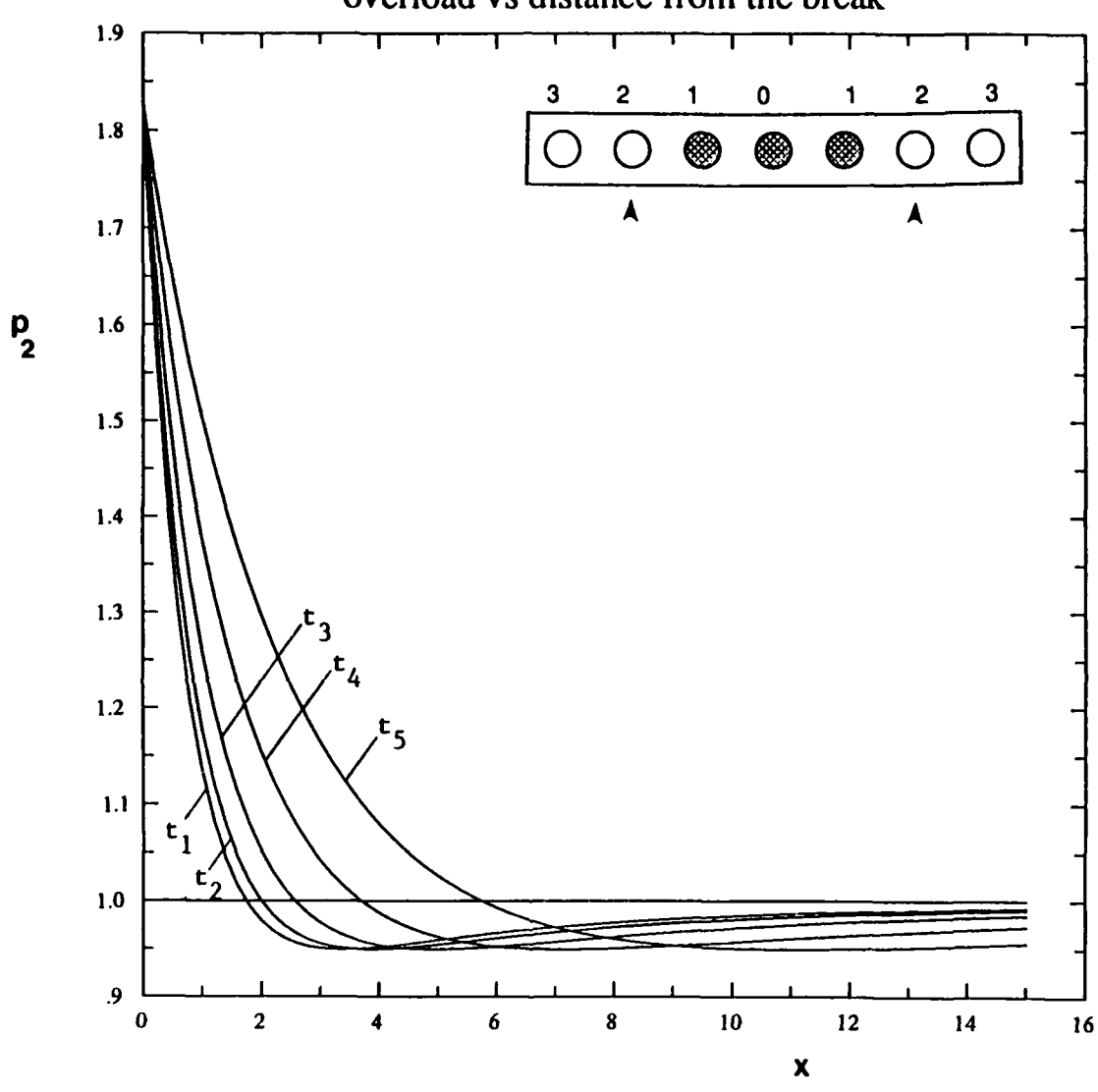


Fig. 4a. Load profile of the first intact fiber for various times; three breaks.

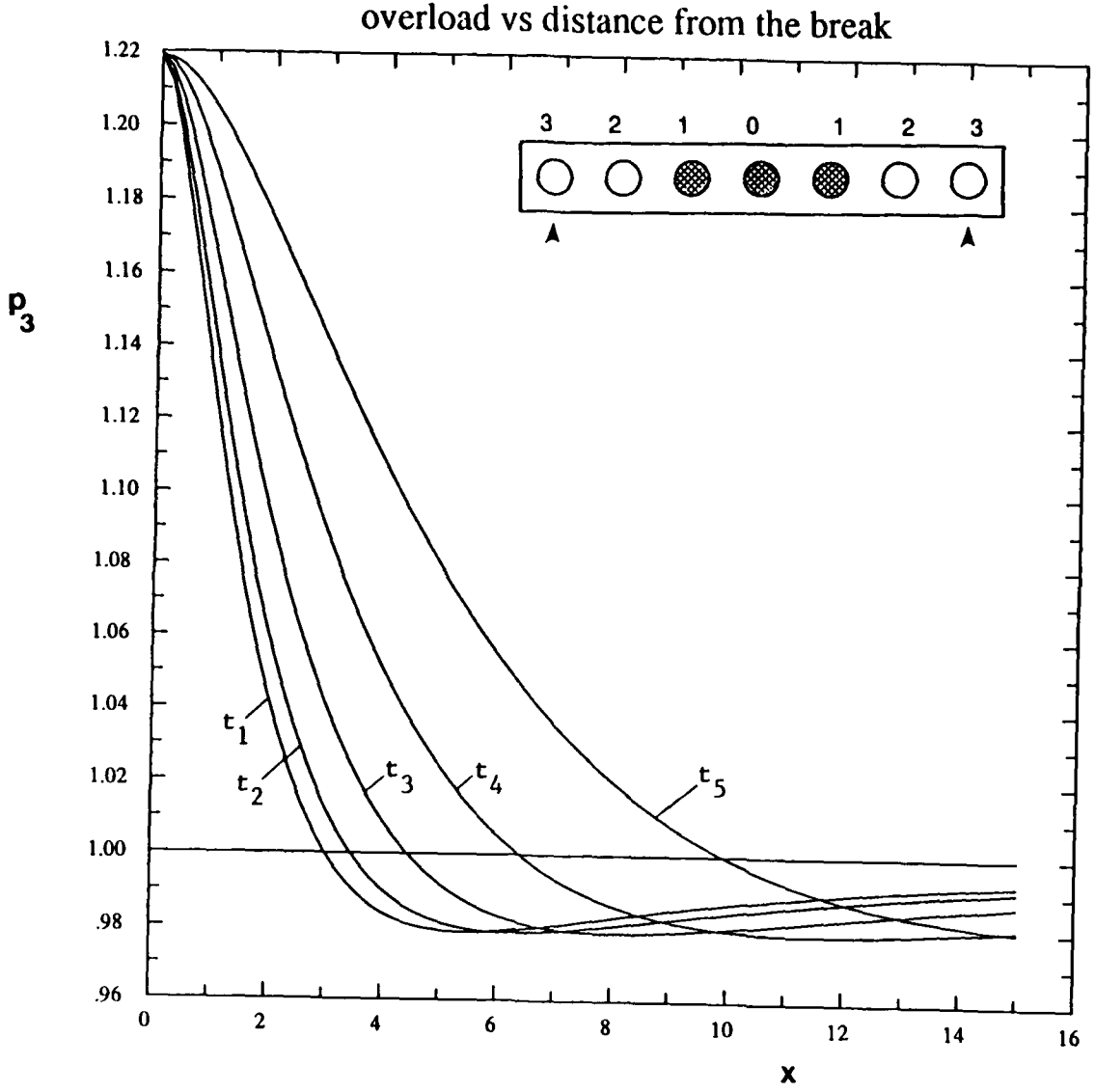


Fig. 4b. Load profile of the second intact fiber for various times; three breaks.

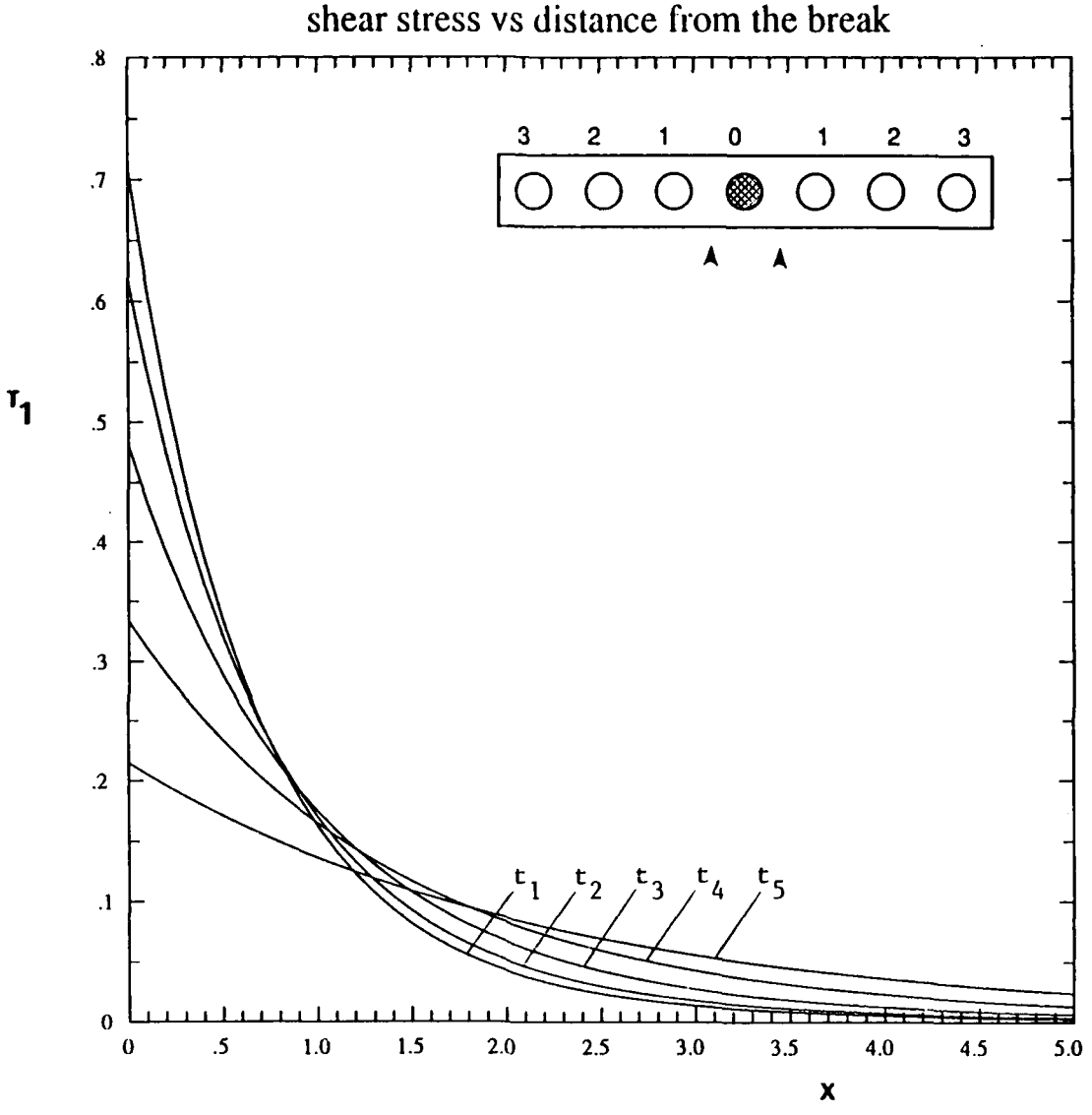


Fig. 5a. Shear stress in matrix between the broken and the first unbroken fibers for various times; one break.

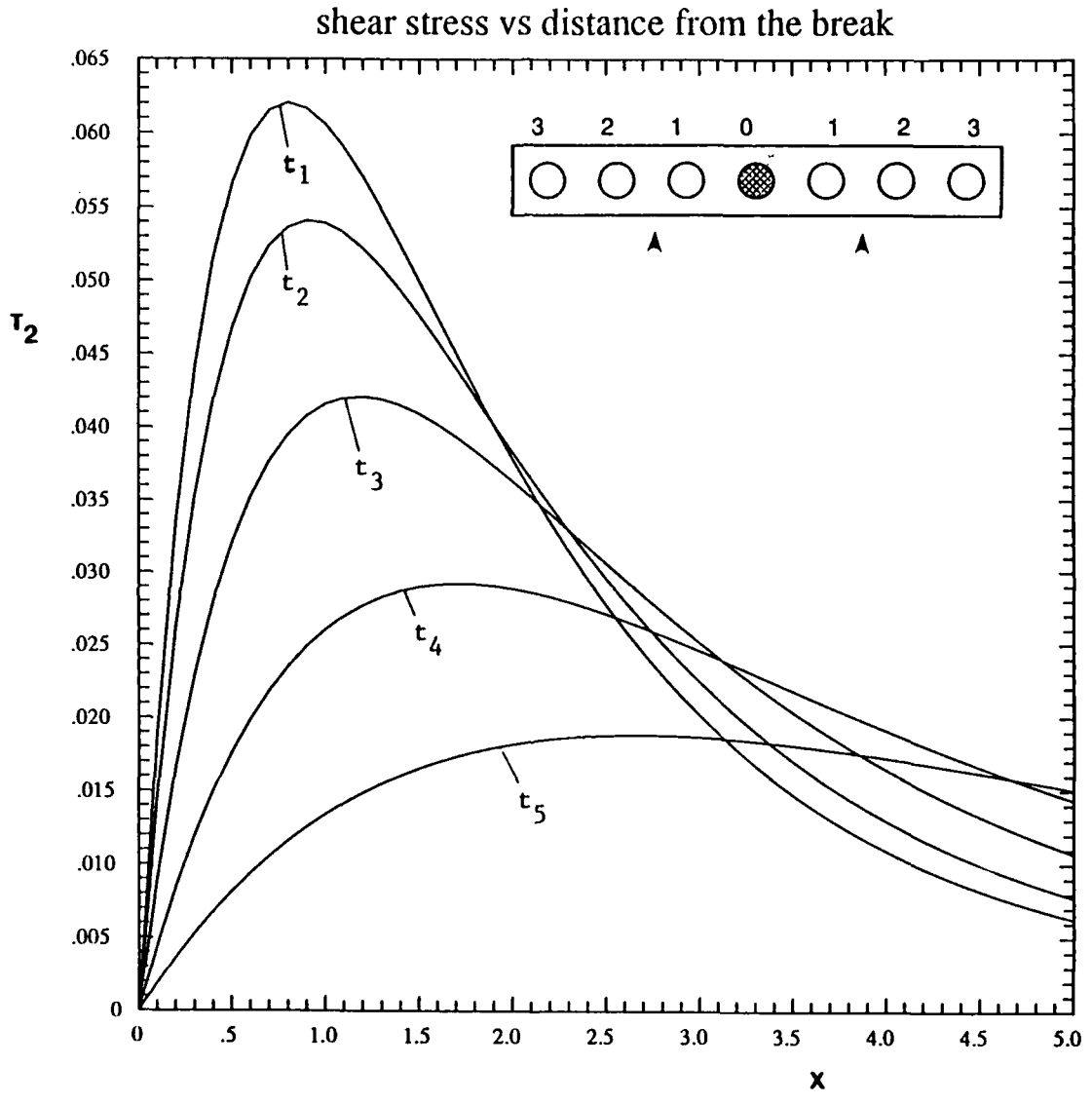


Fig. 5b. Shear stress in matrix between the first and second intact fibers for various times; one break.

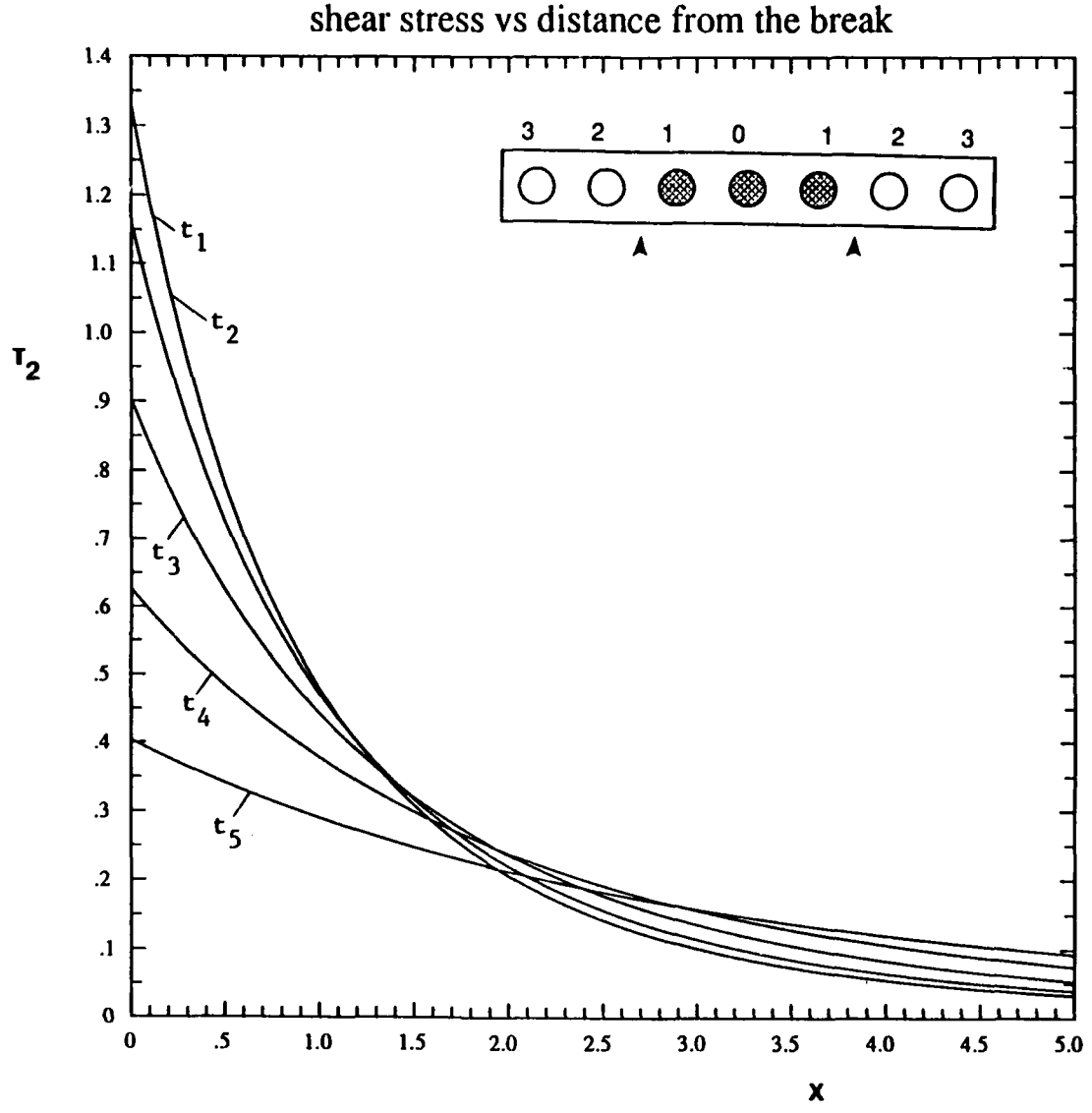


Fig. 6a. Shear stress in matrix between the last broken and the first unbroken fibers for various times; three breaks.

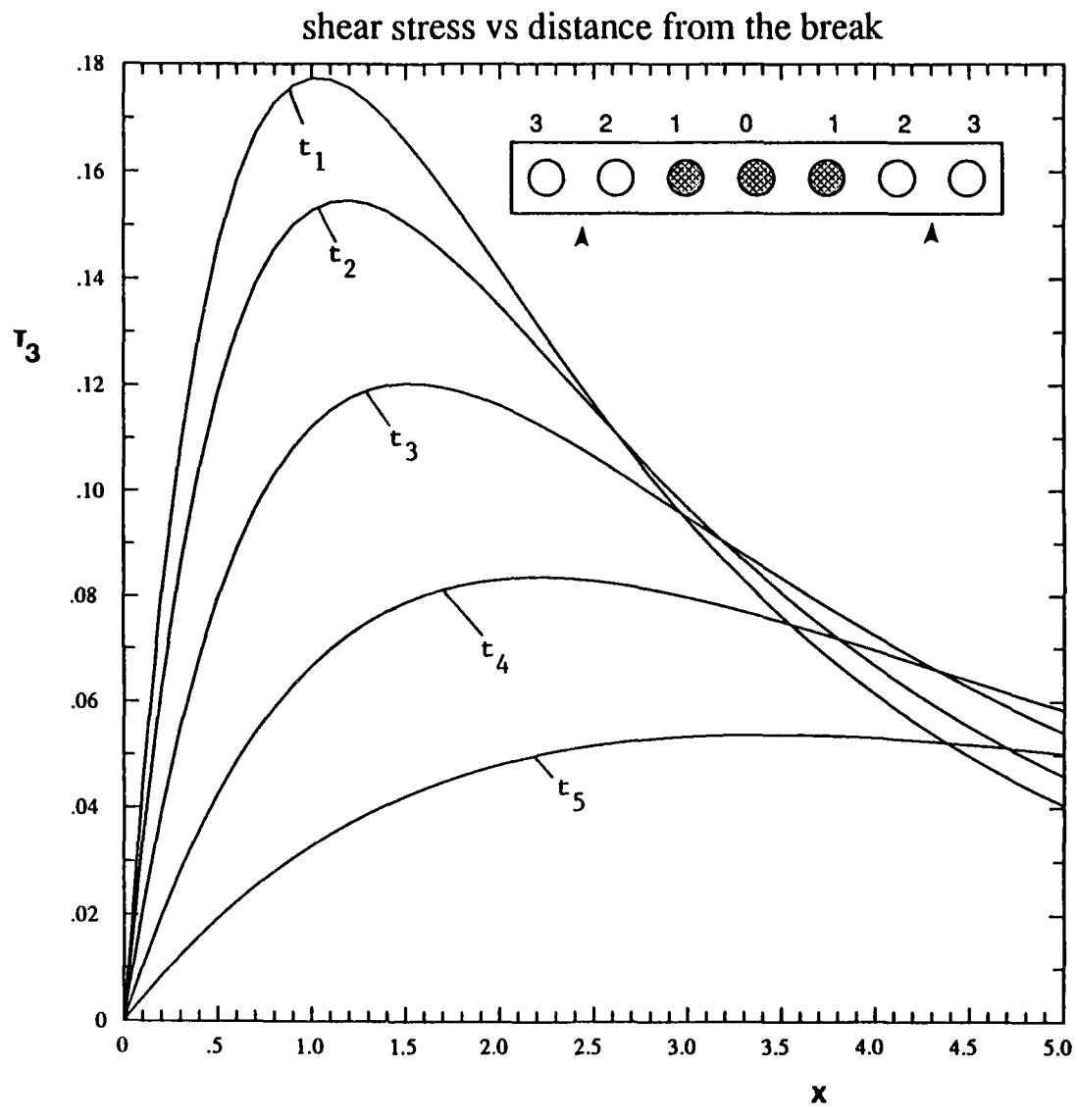


Fig. 6b. Shear stress in matrix between the first and the second intact fibers for various times; three breaks.

ELEMENT LEVEL ELIMINATION OF NONLINEAR CONSTRAINTS IN TOTAL LAGRANGIAN FINITE ELEMENT FORMULATIONS

A. R. Johnson and C. J. Quigley

Mechanics and Structures Division

U.S. Army Materials Technology Laboratory

Watertown, MA 02172-0001 USA

SUMMARY

Nonlinear constraints in elastic finite deformation theory can be enforced by an iterative element level variable elimination method which takes advantage of the finite element discretization. A Lagrangian potential energy method is used and load steps are taken small enough so that the potential energy is nearly quadratic when expanded as a function of displacement increments. The Newton - Raphson method is used to find minimal locations. Element gradient and tangent matricies are computed and modified to be consistent with an incremental representation of the nonlinear constraint. This iterative variable elimination method is used to determine the solutions to the bending of an elastica around an ellipse for aspect ratios of 0.75, 1.00 and 1.50. Two exterior methods are also used to solve these problems for comparison. The Lagrange multiplier method (ABAQUS code) and a penalty method are used. The results obtained using the element level elimination method are compared to the results obtained using ABAQUS and the penalty method.

INTRODUCTION

Many problems in solid and fluid mechanics involve finding either a minimum or stationary value of an energy functional such that an additional constraint equation remains valid. Frictionless contact problems which involve large elastic deformations and curved rigid contact surfaces are considered here. Applications include contact between long thin metal or paper items being passed through channels and rigid smooth surfaced indentors penetrating rubberlike solids. When these problems are formulated using the finite element method the enforcement of the constraint equations (description of the contact surface) is usually the cause of difficulties. The minimization problem is often modified by attaching the constraint equations using either the Lagrange multiplier technique or a penalty method. When the minimization problem is quadratic in the displacement variables and the constraint equations are linear, elimination methods are often used. This suggests that when the nonlinear minimization problem can be made nearly quadratic an elimination method may be possible. An appropriate representation of the nonlinear constraints is necessary which will allow variables to be eliminated from the minimization problem and approximately incorporate the constraint. We briefly describe the nonlinear minimization problem associated with the large deformations of a cantilever beam, the 'elastica'. The minimization problem is then modified by attaching the constraint that the elastica bend around a rigid frictionless elliptical surface. General methods for solving this constrained minimization problem are reviewed to provide background and to provide methods to compare to the element level elimination method proposed here.

Finite element formulations for large deformations of beams exist in several forms[1-5]. We selected Fried's[3] formulation since it is presented as a nonlinear minimization problem in terms of configuration variables. Contact surfaces can be described in terms of these variables. Also, the nonlinear 'B23' element in the ABAQUS[5] code can model the same problem allowing for independent comparisons. Analytical solutions are not available. The nonlinear programming problem which we are concerned with can be presented as follows:

$$\text{Minimize: } f(\{u\}) \; ; \; \{u\} \subset R^N \tag{1}$$

$$\text{Such that: } g_j(\{u\}) \geq 0 \quad j = 1,\ldots,J$$

where $\{u\}$ = the global set of nodal variables,
$f(\{u\})$ = a nonquadratic potential energy function,
and $g_j(\{u\})$ = a differentiable function describing the j'th contact surface.

Frictionless nonlinear contact problems can be represented by equation (1). A large amount of information is available on methods for treating these problems. We present a brief summary of several methods used so that they can be compared to the iterative element level elimination method.

Lagrange multipliers are used to attach the constraint equation to the function being minimized[5-13]. Although this method is not attractive from a theoretical point of view, since it introduces possible saddle points, it has

found widespread use because the Lagrange multipliers represent contact pressures. This method solves the nonlinear programming problem described by equation (1) by formulating it as follows.

$$\text{Stationary points:} \quad f(\{u\}) + \sum_{j=1}^{J} \lambda_j g_j \quad ; \qquad \{u,\lambda\} \subset R^{N+J} \tag{2}$$

where λ_j = j'th Lagrange multiplier.

There are many methods used to construct representations like equations (2) and to solve them. The paper by Simo, Wriggers and Taylor[11] describes the Lagrange multiplier method in detail and introduces a perturbed form which is a mixed penalty / Lagrange multiplier method.

The penalty method[6,14-18] also attaches the constraint equations to the function being minimized. In doing so, it maintains a minimization problem. No variables are added to the analysis set but the large penalty parameters needed can cause ill - conditioning of the modified function's tangent matrix. The problem given by equation (1) is solved using the penalty method as follows.

$$\text{Minimize:} \quad f'(\{u\}) = f(\{u\}) + \sum_{j=1}^{J} \gamma_j g_j^2 \quad ; \qquad \{u\} \subset R^N \tag{3}$$

where γ_j = j'th large constant which may depend on the tangent matrix of f({u}). (See references 17,18).

Nonlinear programming problems can sometimes be made to look like a quadratic programming problem if sucessive trial vectors {u} are sufficiently close[19,20]. If, in addition, the constraint equations can be linearized, then the revised simplex method method for quadratic programming can be used in an iterative scheme to solve equation (1)[19-21]. In this case the problem is formulated as follows.

$$\begin{aligned} &\text{Minimize:} \quad f'(\{u\}) = \frac{1}{2}\{u\}^T[K_o]\{u\} - \{P_o\}^T\{u\} \\ &\text{Such that:} \quad [A]\{x\} - \{b\} \geq 0 \\ &\text{Set:} \quad \{u_o\} = \{u\} \\ &\text{Repeat until contact set, } \{x\}\text{, does not change,} \end{aligned} \tag{4}$$

where $[K] = \left.\frac{\partial^2 f}{\partial \{u\}^2}\right|_{\{u_o\}}$,

$\{P_o\}^T = \{u_o\}^T[K_o]$,

$[A]\{x\} - \{b\}$ = the linearization of equations g_j, in (1)

$\{x\} \subset \{u\}$ is a trial set of variables in the contact set,

$\{u\}$ = a trial vector near $\{u_o\}$,

and $\{u_o\}$ = a vector which minimizes $f(\{u\})$ but does not satisfy the constraints.

This method can be useful when the functions $\{f, g_j\}$ can each be expanded in a Taylor series and when small changes in the constraint set are expected.

A less often used method of enforcing constraints during minimization is to solve the J constraint equations for the relations for J variables (in terms of the remaining N-J variables). The relations are then substituted into the function being minimized. That is, they are eliminated[6]. This yields an unconstrained minimization problem. The method is given as follows.

$$\text{Solve equations } g_j(\{u\}) = 0 \quad j = 1,\ldots J$$

$$\text{and get} \qquad x_j = F_j(\{v\}) \quad j = 1,\ldots J \tag{5}$$

where $\{v\} \subset R^{N-J}$

Substitute (5) into (1) to obtain the unconstrained minimization problem.

$$\text{Minimize: } f(\{v\}) \; ; \quad \{v\} \subset R^{N-J} \tag{6}$$

We have intentionally presented elementary descriptions of the above methods so that the relationship between these methods and the element level elimination proposed here can be easily identified.

ELEMENT LEVEL ELIMINATION METHOD

The elimination method (eqns 5,6) is typically not used when the constraints are nonlinear since it is difficult to determine the functions $F_i(\{v\})$ given by equation (5). Also, when the minimization problem involves many variables, as in a finite element problem, it is difficult to automate a nonlinear elimination method. The method we propose here avoids the difficulties associated with determining equation (5) by working with the derivatives of the constraint equations. We return to solving equation (1) with $\{u\}$ equal to the vector of element nodal variables. Expanding (1) in a Taylor series we have:

$$f(\{u\}) = f(\{u_o\} + \{\Delta u\})$$

$$= f_o + \{g_o\}^T\{\Delta u\} + \frac{1}{2}\{\Delta u\}^T[K_o]\{\Delta u\} + \ldots \tag{7}$$

where $\{u\}$ = a vector "near" $\{u_o\}$ which is closer to the minimum,
$\{u_o\}$= the current location

$$\{\Delta u\} = \{u\} - \{u_o\}$$

$$\{g_o\} = \left.\frac{\partial f}{\partial \{u\}}\right|_{\{u_o\}}$$

$$[K_o] = \left.\frac{\partial^2 f}{\partial \{u\}^2}\right|_{\{u_o\}}$$

Since f is derived from an energy principle and has been discretized by the finite element method, we can express equation (7) as

$$f(\{u\}) = \sum_e \left((f_{eo} + \{g_{eo}\}^T \{\Delta u_e\} + \frac{1}{2} \{\Delta u_e\}^T [K_{oe}] \{\Delta u_e\} + \ldots \right) \quad (8)$$

Assuming the constraint equations in (1) are differentiable we have

$$dg_j = 0 \quad (9)$$

For simplicity we assume one constraint equation (i.e. J=1)
Then, we can write (9) as

$$\sum \frac{\partial g}{\partial x_i} dx_i = 0 \quad (10)$$

where $\{x\} \subset \{u\}$

For small displacement increments

$$\sum \frac{\partial g}{\partial x_i} \Delta x = 0 \quad (11)$$

Solving (11) for a displacement increment Δx_i in the set $\{x\}$ we have

$$\Delta x_\ell = - \frac{1}{\frac{\partial g}{\partial x_\ell}} \sum_{i \neq \ell} \frac{\partial g}{\partial x_i} \Delta x_i \quad (12)$$

This suggests that we can eliminate Δx_ℓ in favor of $g(\{x\})$ equal to zero at the element level. That is,

$$\{\Delta u_e\} = [A_e]\{\Delta u_{er}\} \tag{13}$$

for an element

where $[A_e] = [A_e(\{u_e\})]$

$$\{u_{er}\} = \{\Delta u_e\} - \{\Delta x_\ell\}$$

Thus, $[A_e]$ represents a constraint matrix which depends on the displacements and $\{\Delta u_{er}\}$ the reduced set of element variables. Substituting (13) into (8) we have

$$f(\{u_c\}) = \sum_\ell \Big((f_{eo} + \{g_{eo}\}^T[A_e]\{\Delta u_{er}\} + \frac{1}{2}\{\Delta u_{er}\}^T[A_e]^T[K_{oe}][A_e]\{\Delta u_{er}\} + \ldots \Big) \tag{14}$$

In equation (14) we identify the reduced element gradient and tangent matrices as

$$\{g_{er}\} = [A_e]^T\{g_{eo}\} \tag{15}$$

and $$[K_{er}] = [A_e]^T[K_{eo}][A_e]$$

Global gradient and tangent matrices can now be assembled in the standard way for the "reduced incremental variable set". The Newton - Raphson method is then used to find $\{u\}$. The ℓ_2 norm of the reduced gradient is checked at the new location. If not zero then $\{u\}$ is set to $\{u_o\}$ and the process is repeated. It should be noted that the x_ℓ associated with the eliminated Δx_ℓ must be updated by solving a one dimensional nonlinear equation obtained from the constraint equation at each iteration.

A rule must be made for determining when a variable which is a member of the constrained set must be released. That is, when should a point in the contact set be released? We can write

$$\Delta f = \frac{\partial f}{\partial u_1}\Delta u_1 + \frac{\partial f}{\partial u_2}\Delta u_2 + \ldots + \frac{\partial f}{\partial u_n}\Delta u_n = [g]^T\{\Delta u\} \tag{16}$$

where $u_i \subset \{u\}$

which simply states that $\{g\}$ contains information on how $f(\{u\})$ changes with respect to $\{u\}$. If we consider changes in one variable at a time we obtain

$$\frac{\partial f}{\partial u_i} > 0 \longrightarrow \text{negative } \Delta u_i \text{ decreases } f$$

$$\text{and} \quad \frac{\partial f}{\partial u_i} < 0 \longrightarrow \text{positive } \Delta u_i \text{ decreases } f \tag{17}$$

The rule can then be stated as:

RELEASE RULE:
Find the direction to decrease the energy. Assume it is Δu_i^d. Then, if $u_i + \Delta u_i^d$ does not satisfy the constraint equation, keep u_i in the constrained set. Otherwise, release it.

In the case of two dimensional contact with physical variables x and y at the nodes and corresponding unit vectors $\hat{i}$, $\hat{j}$ the release rule can be simplified as follows, see Figure 1.

$$\text{Let} \qquad \vec{g}_{x_i y_i} = \frac{\partial f}{\partial x_i}\,\hat{i} + \frac{\partial f}{\partial y_i}\,\hat{j}$$

$$= \text{gradient terms for node i} \tag{18}$$

$\hat{n}$ = unit normal to contact surface

$$\text{and} \qquad \alpha = \vec{g}_{x_i y_i} \cdot \hat{n}$$

Then, the release rule for two - dimensional contact, with (x,y) nodal variables, becomes:

RELEASE RULE FOR TWO - DIMENSIONAL CONTACT

If $\alpha > 0$ Keep in constraint set.
If $\alpha < 0$ Release from constraint set.

We note that α equals the component of generalized force outward from the surface and $\vec{g}_{x,y}$ equals the contact force.

ELASTICA BENDING AROUND AN ELLIPSE

We selected the problem of an elastica bending around an ellipse, as shown in Figure 2, to demonstrate the element level elimination algorithm. The aspect ratio, a, is varied to obtain different contact problems. If "a" is large then the contact region changes rapidly with a small change in load, P. When "a" is small a large load is needed for initial contact and the contact region changes more slowly with increasing load P. Contact solutions were obtained for aspect ratios of 0.75, 1.00 and 1.50. Here, we present some details on how the solutions were obtained. First we show the ABAQUS solution (Lagrange multiplier method), second a penalty method and third the element level elimination algorithm.

For all solution methods, the elastica was of length π. One end of the elastica was fixed at the origin and a vertical load P was applied to the other end. For each aspect ratio of the ellipse, the load history used is summmarized in Table 1. Young's modulus was one and Poisson's ratio was zero. To approximate an inextensible elastica, a ratio of the cross sectional area

to the moment of inertia equal to 10^5 was selected[19,20]. The elastica was discretized into forty elements. Two noded beam elements with cubic interpolation functions were chosen to model the elastica. For the element level elimination method and the penalty method a beam element developed by Fried[8] was used. Each node has four degrees of freedom, see Figure 3. In this element the axial strain is based on the definition of engineering strain. For the Lagrangian multiplier method a beam element ('B23' element) with three degrees of freedom per node was selected in ABAQUS. A Lagrangian axial strain is used in this element formulation.

Lagrange Multiplier Method

To solve the problem of an elastica bending around an ellipse using the Lagrange multiplier method, the finite element code ABAQUS was used. A two noded planar interface element ('IRS21' element) was chosen to detect contact between the beam element and the rigid surface or ellipse. This contact element enforces a linear pressure distribution between nodes and has integration points at the nodes. The material properties cited above were input using the 'general beam section' option. Tolerances were set at 0.4% of the applied load and at one percent of the moment. Smaller tolerances did not change the ABAQUS finite element solution. The ellipse was defined by the user subroutine RSURFU. At each integration point of the planar interface element, the penetration distance into the ellipse is calculated. To do this the coordinates of the point on the ellipse closest to the integration point must be found. The direction cosines of the tangent and the rate of change of the tangent along the surface at this point on the ellipse are also determined.

To find the point on the ellipse closest to a given point along the elastica, the Newton - Raphson method was used. Given the equation for a point (x,y) on the ellipse, we need to minimize the distance between the point (x,y) and the integration point(x_o,y_o) along the elastica. This can be done using the same elimination method we use to define surface contact. We proceed as follows.

$$\text{Minimize: } \Pi = (x - x_o)^2 + (y - y_o)^2 \qquad \text{(distance)} \qquad (19)$$

$$\text{Given: } \quad f(x) = \frac{x^2}{a^2} + (y + 1)^2 = 1 \qquad \text{(constraint equation)} \qquad (20)$$

Reduce the variables (x,y) to x using the constraint equation.

$$\Delta y = \frac{x}{a^2(1 + y)} \Delta x \qquad (21)$$

$$\text{From} \quad \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} 1 \\ \dfrac{x}{a^2(y + 1)} \end{bmatrix} \{\Delta x\} \qquad (22)$$

That is, we have $\{\Delta u\} = [A]\{\Delta u_r\}$ in (22).

Using equation (19), the gradient and tangent stiffness matrices are derived.

$$\{g\} = \begin{bmatrix} 2(x - x_o) \\ 2(y - y_o) \end{bmatrix}$$

$$[K] = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \tag{23}$$

Next we solve for the reduced gradient and tangent stiffness matrices.

$$\{g_r\} = [A]^T\{g\}$$

$$= 2(x - x_o) + \frac{2x(y - y_o)}{a^2(y + 1)}$$

$$[K_r] = [A]^T[K][A] \tag{24}$$

$$= 2 + \frac{2x^2}{a^4(y + 1)^2}$$

Applying the Newton - Raphson method to solve (24) using an initial guess for (x_1, y_1) yields

$$\{x_2\} = \{x_1\} - [K_r(x_1)]^{-1}\{g_r(x_1)\} \tag{25}$$

or, here

$$x_2 = x_1 - \left[\frac{(x - x_o) + \dfrac{x_1(y_1 - y_o)}{a^2(y + 1)}}{1 + \dfrac{x_1^2}{a^4(y_1 + 1)^2}} \right]$$

To update the y coordinate, we use the constraint equation. That is,

$$y_2 = -1 + \left[1 - \frac{x_2^2}{a^2} \right]^{1/2} \tag{26}$$

This process is repeated until the reduced gradient approaches zero.

For an initial guess the principle of similar triangles was used, refer to Figure 4.

$$x_1 = \frac{x_o}{R}$$

$$y_1 = 1 - \left[1 - \frac{x_1}{a^2} \right]^{1/2} \tag{27}$$

where R = the length of the segment $\overline{OX}$

Once the point (x,y) on the ellipse closest to the integration point on the elastica is located the direction cosines of the unit tangent $\hat{t}$ at this point are:

$$\hat{t} = \frac{a^2(y + 1)\,\hat{i} - x\,\hat{j}}{\left(a^4(y + 1)^2 + x^2\right)^{1/2}} \tag{28}$$

The rate of change of the tangent along the ellipse can be expressed in terms of the curvature κ and the normal to the ellipse at the point (x,y).

$$\frac{d\hat{t}}{ds} = \kappa\hat{n} \tag{29}$$

where

$$\kappa = - \left[a^2(y + 1)^3 \left[1 + \frac{x^2}{a^4(y + 1)^2} \right]^{3/2} \right]^{-1}$$

$$\hat{n} = \frac{x\,\hat{i} + a^2(y + 1)\,\hat{j}}{\left(a^4(y + 1)^2 + x^2\right)^{1/2}}$$

Penalty Method

To obtain a penalty solution we selected a penalty term which would add the square of the minimum distance to the contact distance to the potential energy. This is one of many possible variations of equation (3). The constraint equation (20) enters indirectly when the location of the minimal distance point on the surface is determined. Thus we satisfy the constraint in a least squares sense by minimizing the distance between the node (x_n, y_n) and the ellipse. That is, minimize

$$\Pi_e' = \Pi_e + \int_n \gamma_{np} \left((x_n - x_e)^2 + (y_n - y_e)^2\right) \tag{30}$$

where $\gamma_{np} = c_n k_{nn} 10^p$ = the n'th penalty parameter

p = a variable for convergence studies

$$c_n = \begin{cases} \frac{1}{2} & \text{if one or no adjacent nodes have violated the constraint equation} \\ 1 & \text{if both adjacent nodes have violated the constraint equation} \end{cases}$$

k_{nn} = the diagonal term from the tangent stiffness matrix associated with the x_n displacement

(x_e, y_e) = the point on the ellipse closest to the node (x_n, y_n) on the elastica. This point is found using the solution method described in the previous section.

The computation of the element gradient and tangent stiffness matrices are shown below.

$$\{g_e'\} = \frac{\partial \Pi_e'}{\partial \{u_e\}}$$

$$= \{g_e\} + \left\{ \frac{\partial \Pi_e'}{\partial x_n} \right\} + \left\{ \frac{\partial \Pi_e'}{\partial y_n} \right\} \tag{31}$$

where $\frac{\partial \Pi_e'}{\partial x_n} = 2(x_n - x_o)$ (in row for x_n)

$$\frac{\partial \Pi_e'}{\partial y_n} = 2(y_n - y_o) \quad \text{(in row for } y_n\text{)}$$

Similarly,

$$[k_e'] = \frac{\partial^2 \Pi_e'}{\partial \{u_e\}^2}$$

$$[k_e'] = [k_e] + \left[\frac{\partial^2 \Pi_e'}{\partial x_n^2}\right] + \left[\frac{\partial^2 \Pi_e'}{\partial y_n^2}\right] \tag{32}$$

where

$$\frac{\partial^2 \Pi_e'}{\partial x_n^2} = \frac{\partial^2 \Pi_e'}{\partial y_n^2} = 2 \qquad \text{(in both } (x_n, x_n) \text{ and } (y_n, y_n) \text{ locations)}$$

The global gradient and tangent stiffness matrices are then assembled in the usual manner. The Newton - Raphson method is applied to minimize the global form of equation (30). To study convergence, the analysis was completed with values of 10^p equal to 10E-5, 10E-2, and 10E-1.

Element Level Elimination

For this method, we need to define the matrix $[A_e]$ from equation (13). From the constraint equation (20) we can compute Δy in terms of Δx.

$$\Delta y = -\frac{x}{a^2(y+1)}\,\Delta x = F(x)\,\Delta x \tag{33}$$

Thus, when a node on the elastica is in the contact set we can compute the element energy gradient and tangent matrices for which the relation (33) holds by using equation (15). After assembly, the number of variables in the global gradient and tangent matrices will be reduced by the number of nodes in the contact set.

When the node in contact is the first node along the element, $[A_e]$ becomes:

$$[A_e] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ F(x) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{34}$$

where $\{\Delta u_{er}\}^T = \{ \dot{x}_1 \; \dot{x}_1 \; y_1 \; \dot{x}_2 \; x_2 \; \dot{y}_2 \; y_2\}$

When the second node of the element is in contact the new $[A_e]$ can be readily seen from equation (34). When both nodes of the element violate the constraint, $[A_{er}]$ reduces to:

$$[A_e] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ F(x) & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & F(x) & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{35}$$

and $\{u_{er}\}^T = \{x_1 \; \dot{x}_1 \; \dot{y}_1 \; x_2 \; \dot{x}_2 \; \dot{y}_2\}$

The reduced element gradient and tangent matrices can then be computed using equation (15).

RESULTS

To demonstrate the use of the element level elimination method, the problem of an elastica bending around a frictionless rigid surface in the form of an ellipse was solved. For comparison, this contact problem was also solved using the Lagrange multiplier (ABAQUS) and penalty methods. Similar results were obtained for all methods. The displacements both in the contact region and at the tip were in good agreement.

A comparison of the deformed configuration of the elastica bending around a circle (a=1.00, Figure 2), at a load of 0.55, shows close agreement between the element level elimination and Lagrange multiplier methods, see Table 2. The deformed configurations at loads of 0.60, 1.00, and 1.25 are shown in Figure 5. The elastica first made contact with the circle at a load of 0.55. The contact solutions do have some differences. The ABAQUS solution and penalty solution show regional contact while the element level elimination method shows point contact. When two nodes lie along the circle, the elastica is in contact with the circle at some point between those nodes. Thus, a solution with two nodes in contact implies point contact with the rigid surface. The location of the contact surface on the circle vs load is shown in Figure 6. The final node in contact at a given load is the same, refer to Table 3. The regional solution obtained by ABAQUS and the penalty method at a given load was also an iterative solution obtained by the element level elimination method. By considering energy minimization and using the Release Rule for Two - Dimensional Contact, the additional contact points found by the other methods were released. These released points lie close to the circle; the maximum distance between a node and the circle was 10E-5. The

ABAQUS pressure forces for these additional nodes in contact were an order of magnitude lower than the largest pressure force.

For aspect ratios of 0.75 and 1.50, the contact solution was identical for all methods. (Refer to Tables 4 and 5.) The elastica makes initial contact with the ellipse at a load of 1.70 and has regional contact when the aspect ratio is 0.75. For an aspect ratio of 1.50, initial contact occurs at a load of 0.17 and point contact results. Deformed configurations of the elastica at various loads for the aspect ratios of 0.75 and 1.50 are shown in Figure 8. A summary of the results for the element level elimination contact solution is found in Figure 9.

Tables 3 through 5 also demonstrate the influence of the penalty parameter on the contact solution. When the penalty parameter equals 10E-5, the optimal solution is not obtained. The distance between the nodes in contact and the ellipse is 10E-2. When the penalty term is increased, the contact set becomes smaller and the distance bewteen the nodes in contact and the ellipse is 10E-5. At higher values, though, problems with convergence and "chattering", and oscillation between two different contact solutions was observed.

SUMMARY

An element level elimination algorithm for the analysis of frictionless geometrically nonlinear constraint problems was presented. This algorithm is easy to implement within a finite element code. The release of a nodal variable from the constraint set is based on energy minimization principles. Ill - conditioning of the tangent stiffness matrix is avoided. To demonstrate this algorithm, the problem of an inextensible elastica bending around an ellipse was solved. For comparison, solutions to this problem were also obtained using a penalty method and the Lagrange multiplier (ABAQUS) method.

REFERENCES

1. Y. Toda and G.C. Lee, 'Finite element solution to an elastica problems of beams', Int. J. Num. Meth. Engng. **2**, 229-241 (1970).

2. I. Fried, 'Finite element computation of large elastic deformations', The Mathematics of Finite Elements and Applications IV, MAFELAP 1981, ed. J.R. Whiteman, Academic Press 1982.

3. I. Fried, 'Nonlinear finite element computation of the equilibrium, stability and motion of the extensional beam and ring', Comp. Meth. Appl. Mech. Eng. **38**, 29-44 (1983).

4. B.W. Golley, 'The finite element solution of a class of elastica problems', Comp. Meth. Appl. Mech. Eng. **46**, 159-168 (1984).

5. ABAQUS Version 4.5, Hibbitt, Karlsson, and Sorenson, Providence, R.I., 1987.

6. G.V. Reklaitis, A. Ravindran, and K.M. Ragsdell, Engineering Optimization Methods and Applications, John Wiley and Sons, New York, 1983.

7. T.J.R. Hughes, R.L. Taylor, and W. Kanoknnukulchai, 'A finite element method for large displacement contact and impact problems', Formulations and Computational Algorithms in Finite Element Analysis, eds. K.J. Bathe, J.T. Oden and W. Wunderlich, MIT (1976).

8. J. Tseng and M.D. Olson, 'The mixed finite element method applied to two - dimensional elastic contact problems', Int. J. Num. Meth. Engng. 17, 991-1014 (1981).

9. O.S. Narayanaswarmy, 'Processing nonlinear multipoint constraints in the finite element method', Int. J. Num. Meth. Engng. 21, 1283-1288 (1985).

10. K.J. Bathe and A. Chaudhary, 'A solution method for planar and axisymmetric contact problems', Int. J. Num. Meth. Engng. 21, 65-88 (1985).

11. J.C. Simo, P.Wriggers and R.L. Taylor, 'A perturbed Lagrangian formulation for the finite element solution of contact problems', Comp. Meth. Appl. Mech. Eng. 50, 163-180 (1985).

12. B. Nour - Omid and P. Wriggers, 'A two level iteration method for solution of contact problems', Comp. Meth. App. Mech. Eng. 54, 131-144 (1986).

13. P.P. Lazaridis and P.D. Panagiotopoulos, 'Boundary variational principles for inequality structural analysis problems and numerical applications', Comput. Struct. 25, 35-49 (1987).

14. J.N. Reddy, 'The penalty function method in mechanics a review of recent advances', Penalty - Finite element Methods in Mechanics, AMD - Vol. 51, ASME, 1982, ed. J.N. Reddy.

15. N. Kikuchi, 'A smoothing technique for reduced integration penalty methods in contact problems', Int. J. Num. Meth. Engng. 18, 343-350 (1982).

16. T. Endo, J.T. Oden, E.B. Becker and T. Miller, 'A numerical analysis of contact and limit - point behavior in a class of problems of finite elastic deformations', Comput. Struct. 18, 899-910 (1984).

17. C.A. Felippa, 'Error analysis of penalty function techniques for constraint definitions in linear algebraic systems', Int. J. Num. Meth. Engng. 11, 709-728 (1977).

18. C.A. Felippa, 'Iterative procedures for improving penalty function solutions of algebraic systems', Int. J. Num. Meth. Engng. 12, 821-836 (1978).

19. A.R. Johnson and C.J. Quigley, 'Buckled elastica in contact - finite element solutions', Transactions of the Second Army Conference on Applied Mathematics and Computing, NTIS (AD - P004904), February 1985.

20. A.R. Johnson and C.J. Quigley, 'Frictionless geometrically nonlinear contact using quadratic programming', submitted to Int. J. Num. Meth. Engng.

21. M.H. Rusin, 'A revised simplex method for quadratic programming, SIAM J. Appl. Math. 20(2), 143-160 (1971).

22. R. Chand, E.J. Haug and R. Kim, 'Analysis of unbonded contact problems by means of quadratic programming', J. Opt. Theory Appl. 20(2), 171-189 (1976).

23. E. Haug, R. Chaud and K. Pan, 'Multibody elastic contact analysis by quadratic programming', J. Opt. Theory Appl. 21(2), 189-198 (1977).

Table 1 Load History for Each Ellipse Aspect Ratio

Load Number	Aspect Ratio of Ellipse		
	0.75	1.00	1.50
1	0.05	0.05	0.05
2	0.10	0.10	0.10
3	0.15	0.15	0.15
.	.	.	0.16
.	.	.	0.17
.	1.65	0.50	0.18
.	1.70	0.55	.
.	1.75	0.60	.
.	.	.	.
.	.	.	.
.	.	.	.
N	2.25	1.30	0.30

Table 2 Deformed Configuration of the Elastica Bending Around a Circle at a Load of 0.55 (a = 1.00, Figure 2.)

Node	Degree of Freedom	Element Level Elimination Solution	ABAQUS Solution
1	X	0.00	0.00
	Y	0.00	0.00
	Θ	0.00	0.00
2	X	7.84601E-2	7.84585E-2
(contact)	Y	-3.0827E-2	-3.0826E-2
	Θ	-7.8116E-2	-7.8113E-2
3	X	0.15645	0.15645
	Y	-1.2168E-2	-1.2168E-2
	Θ	-0.1533	-0.1533
4	X	0.23357	0.23357
	Y	-2.6955E-2	-2.6955E-2
	Θ	-0.2251	-0.2251
.	.	.	.
	.	.	.
	.	.	.
.	.	.	.
	.	.	.
	.	.	.
	.	.	.
.	.	.	.
	.	.	.
	.	.	.
	.	.	.
41	X	1.8571	1.8576
	Y	-2.287	-2.287
	Θ	-1.247	-1.247

Table 3. Nodes in Contact vs Load for Circular Contact Surface (a = 1.00, Figure 2.)

Load	Elememt Level Elimination Method**		Lagrange Multiplier Method		Penalty Method**	
	Iteration	Nodes	Iteration	Nodes	Penalty Parameter 10^{P}*	Nodes
0.55	1 2	 2	1	2	10E-5 10E-2 10E-1	2 2 2
0.60	1 2	2 2,3	1 2 3	 2,3	10E-5 10E-2 10E-1	2-5 2,3 2,3
0.65	1 2 3	2,3 2-4 3,4	1 2 3 4	 3,4	10E-5 10E-2 10E-1	2-7 2-4 2-4
0.70	1 2 3 4 5	3,4 3-5 2-5 2,4,5 4,5	1 2	 4,5	10E-5 10E-2 10E-1	2-8 3,4 4,5
0.75	1	4,5	1	2-5	10E-5 10E-2 10E-1	2-9 *** 2,4,5
0.80	1 2 3 4 5 6 7 8 9	4,5 4,5,7 3-5,7 3-7 3-6 3,5,6 2,3,5,6 2,5,6 5,6	1 2 3 4 5	 2,5,6	10E-5 10E-2 10E-1	2-10 2,4,5 2,5,6

* The entire penalty term is $c_n k_{nn} 10^{P}$, see eqn (30).
** The norm of the gradient was less than or equal to 10E-8.
*** Convergence of the Newton - Raphson soltuion was not obtained after 20 iterations.

Table 4. Nodes in Contact vs Load for Elliptical Contact Surface (a = 0.75, Figure 2.)

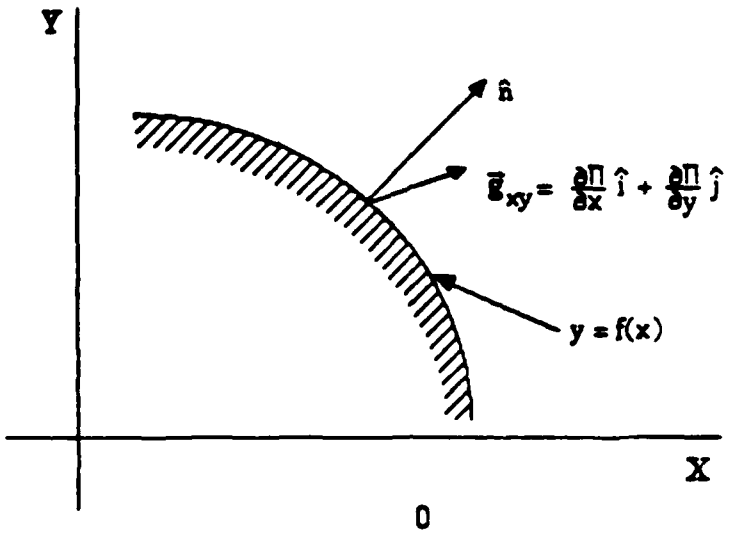
Load	Elememt Level Elimination Method**		Lagrange Multiplier Method		Penalty Method**	
	Iteration	Nodes	Iteration	Nodes	Penalty Parameter 10^{p}*	Nodes
1.70	1	2	1	2	10E-5 10E-2 10E-1	2 2 2
1.75	1	2	1	2	10E-5 10E-2 10E-1	2,3 2 2
1.80	1	2	1	2	10E-5 10E-2 10E-1	2-4 2,3 2
1.85	1 2	2 2,3	1	2,3	10E-5 10E-2 10E-1	2-5 2,3 2,3
1.90	1	2,3	1	2,3	10E-5 10E-2 10E-1	2-6 2,3 2,3
1.95	1	2,3	1	2,3	10E-5 10E-2 10E-1	2-6 2,3 2,3
2.00	1	2,3	1	2,3	10E-5 10E-2 10E-1	2-7 2-4 2-4
2.05	1 2	2,3 2-4	1 2	2-4	10E-5 10E-2 10E-1	2-8 2-4 2-4
2.10	1	2-4	1	2-4	10E-5 10E-2 10E-1	2-9 *** 2-4

* The entire penalty term is $c_n k_{nn} 10^{p}$, see eqn (30).
** The norm of the gradient was less than or equal to 10E-8.
*** Convergence of the Newton - Raphson soltuion was not obtained after 20 iterations.

Table 5. Nodes in Contact vs Load for Elliptical Contact Surface (a = 1.50, Figure 2.)

Load	Elememt Level Elimination Method**		Lagrange Multiplier Method		Penalty Method**	
	Iteration	Nodes	Iteration	Nodes	Penalty Parameter 10^{P}*	Nodes
0.17	1 2 3 4 5 6	5 2,5 2,3,5 2-5 2-4 2,3	1 2 3 4 5	 2,3	10E-5 10E-2 10E-1	2-5 2,3 2,3
0.18	1 2 3 4	2,3 2-5 2-4 3,4	1 2 3 4	 3,4	10E-5 10E-2 10E-1	2-7 3,4 3,4
0.19	1 2 3	3,4 3-5 4,5	1 2 3 4	 4,5	10E-5 10E-2 10E-1	2-8 4,5 4,5
0.20	1 2 3	4,5 4-6 5,6	1 2 3	 5,6	10E-5 10E-2 10E-1	2-9 5,6 5,6
0.21	1 2 3	5,6 6 6,7	1 2	 6,7	10E-5 10E-2 10E-1	2-10 6,7 6,7
0.22	1	6,7	1	6,7	10E-5 10E-2 10E-1	2-11 6,7 6,7
0.23	1 2 3 4 5	6,7 7 7,9 7-9 7,8	1 2	 7,8	10E-5 10E-2 10E-1	2-11 7,8 ***

* The entire penalty term is $c_n k_{an} 10^{P}$, see eqn (30).
** The norm of the gradient was less than or equal to 10E-8.
*** Convergence of the Newton - Raphson soltuion was not obtained after 20 iterations.



Let $\alpha = \hat{n} \cdot \vec{g}_{xy}$

If $\alpha > 0$ then releasing node increases energy. HOLD.

If $\alpha < 0$ then release.

If more than one point is in contact, then add or release according to largest value of α.

Figure 1 Release rule for two dimensional contact.

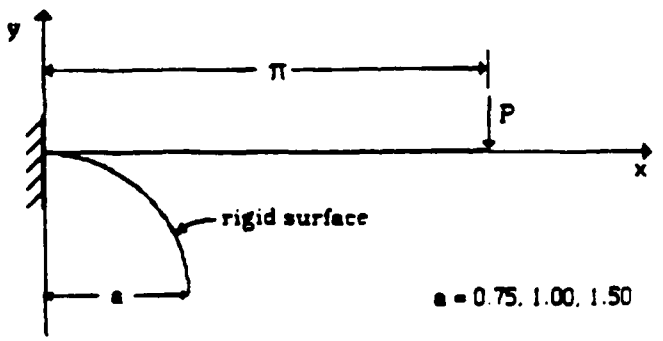
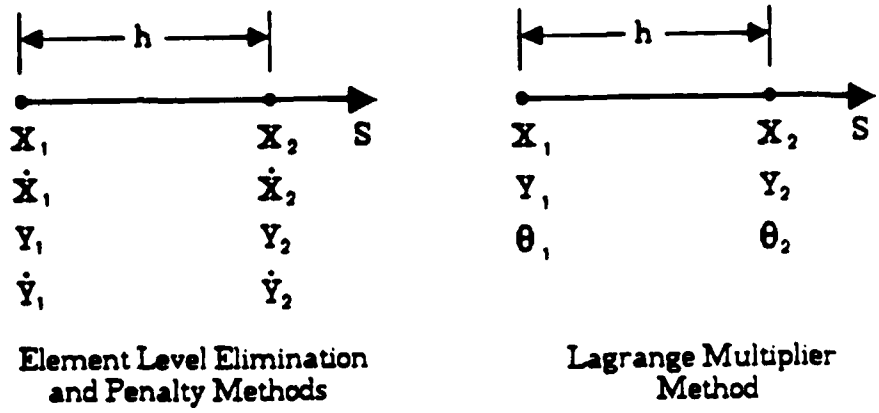


Figure 2. The elastica bending around an elliptical rigid surface.



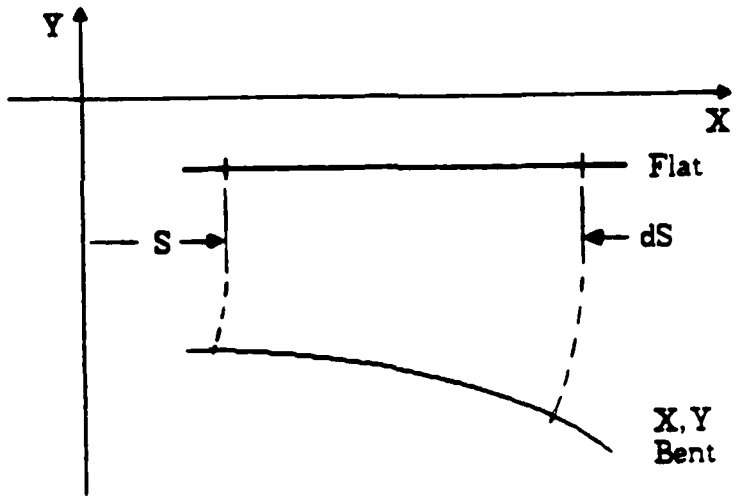


Figure 3. Element description.

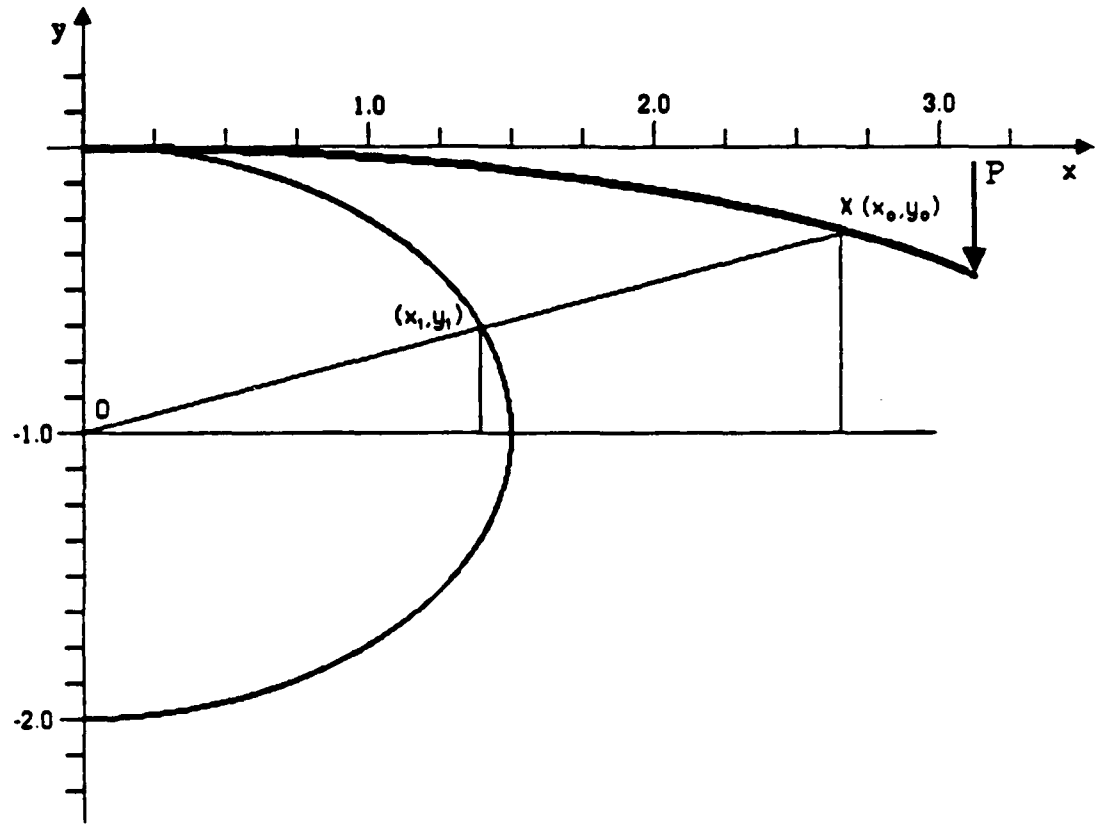


Figure 4. Determination of (x_1, y_1) using similar triangles

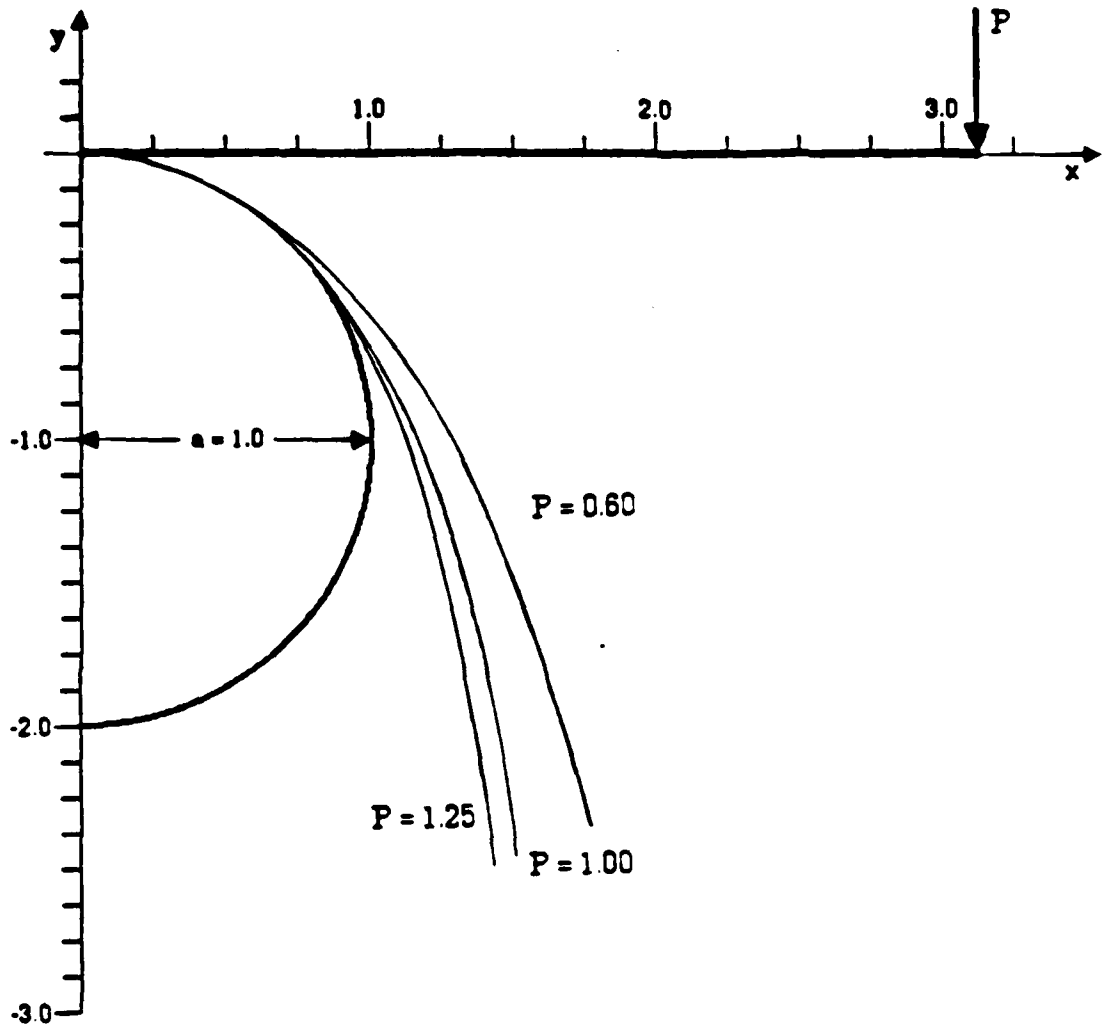


Figure 5 Deformed configuration of the elastica bending around a circle.

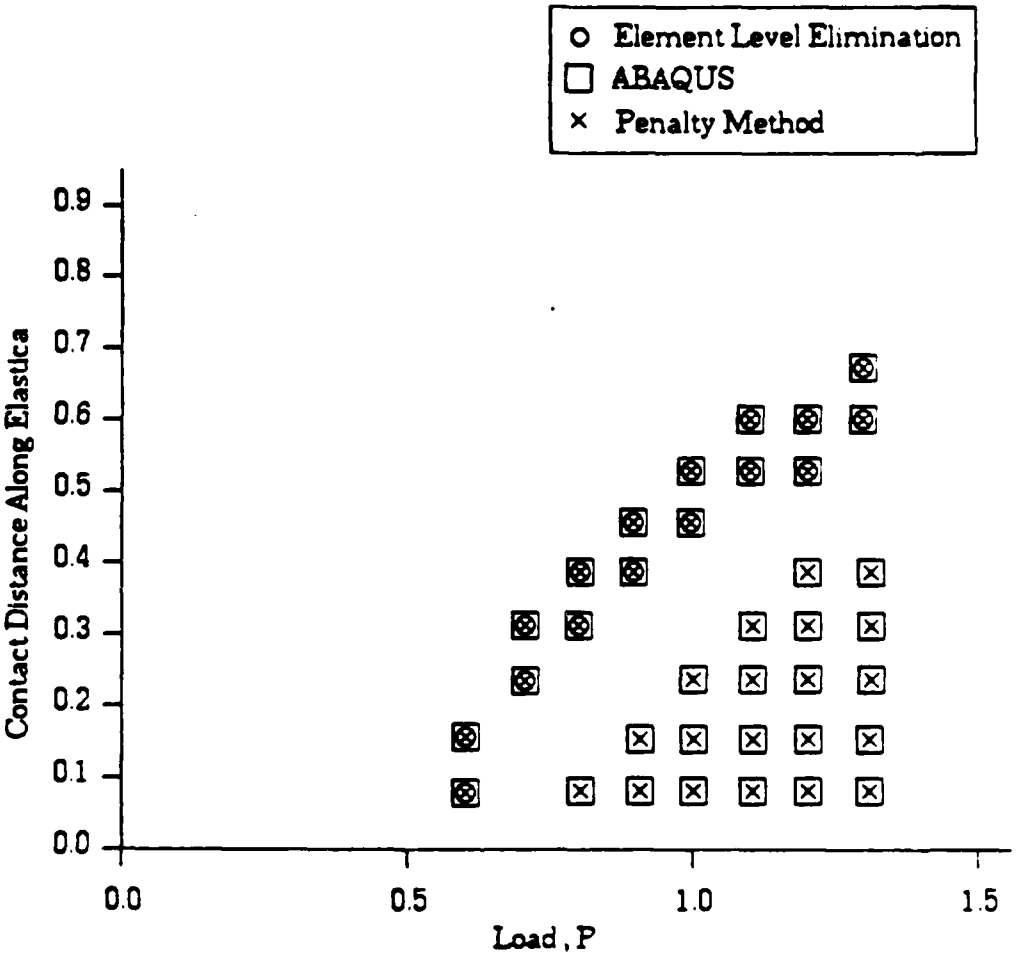


Figure 6. Location of contact surface vs. load for the elastica bending around a circle

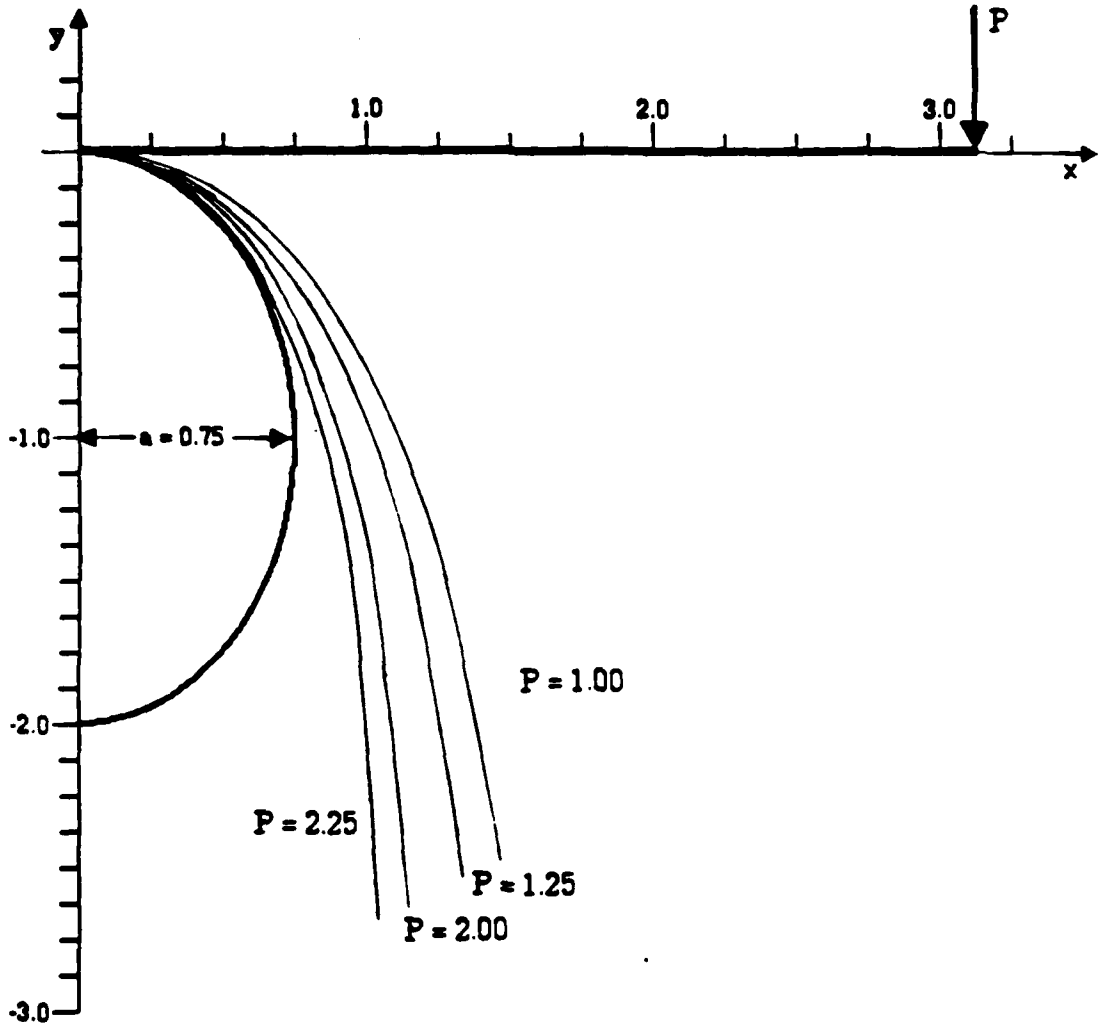


Figure 7. Deformed configuration of the elastica bending around an ellipse, aspect ratio = 0.75.

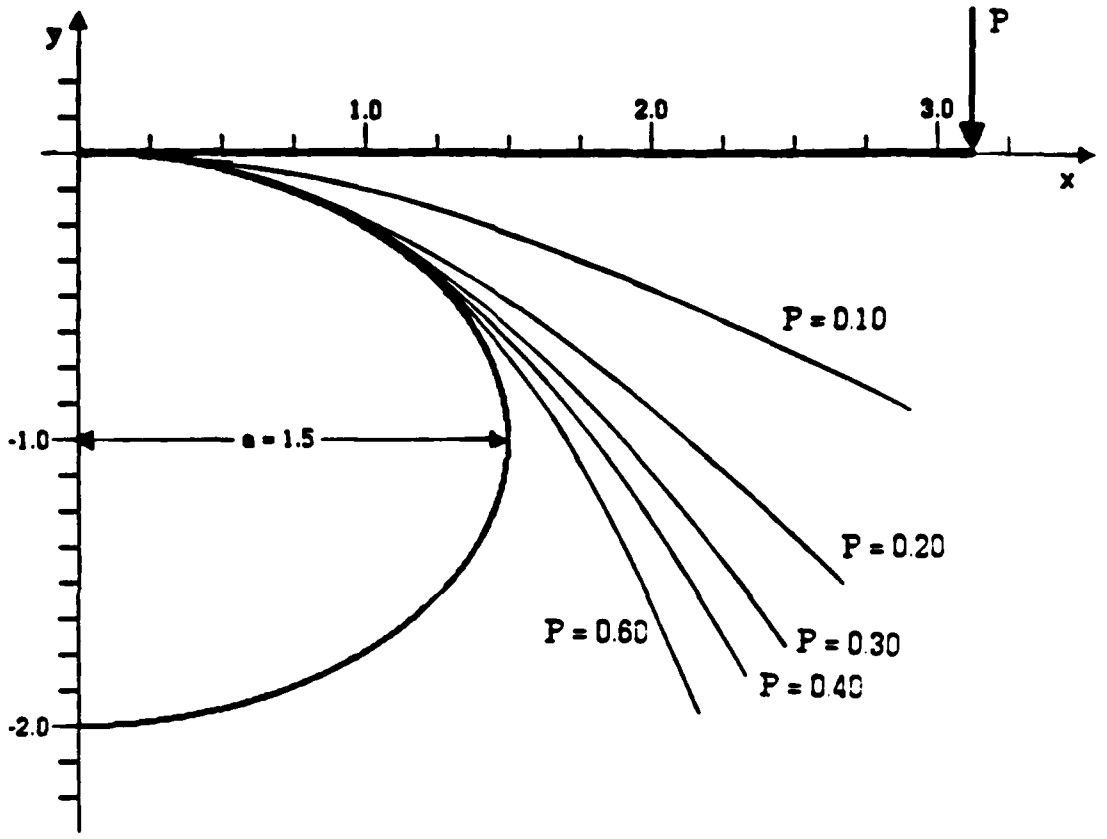


Figure 8 Deformed configuration of the elastica bending an ellipse, aspect ratio = 1.50.

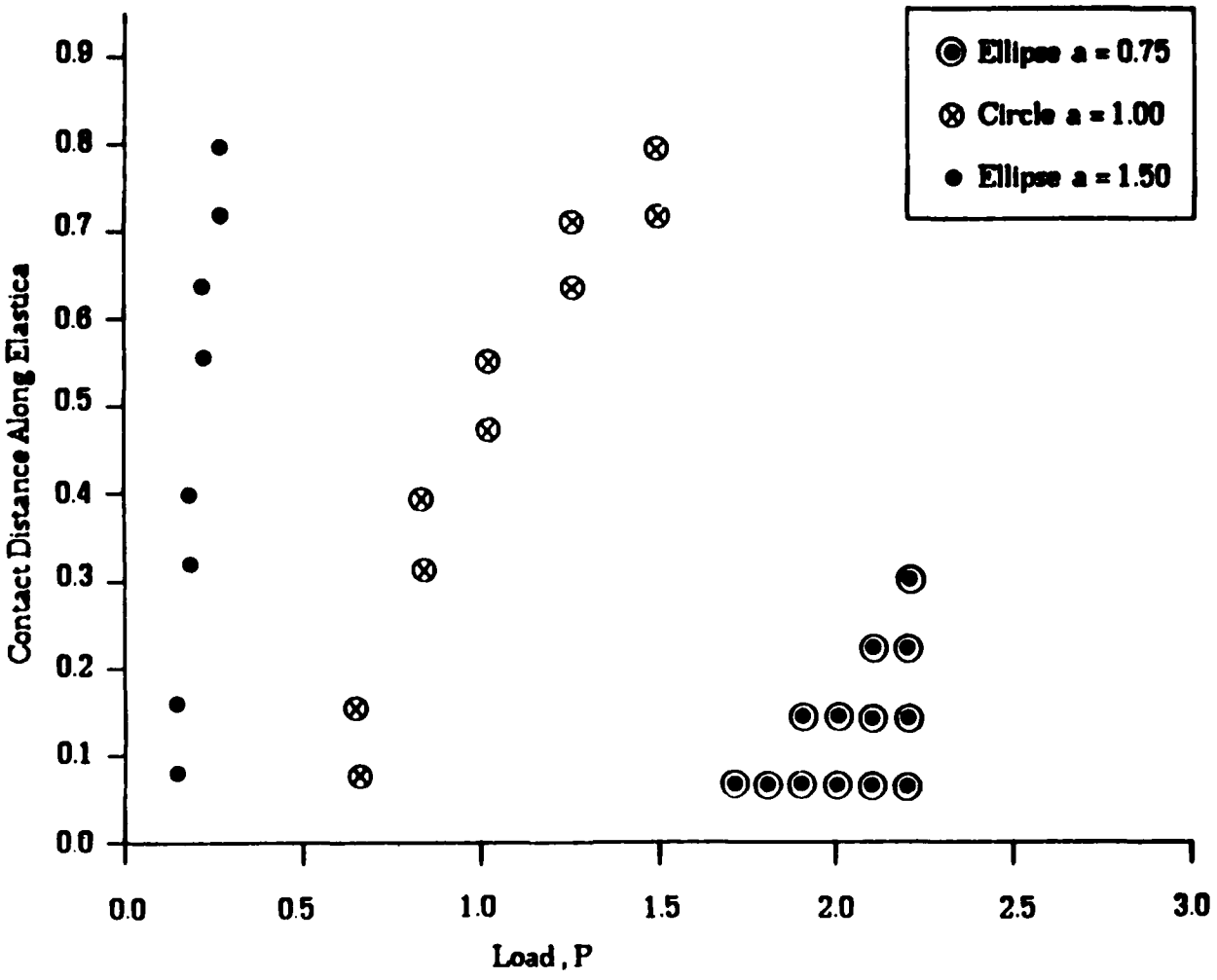


Figure 9. The location of the contact surface on the ellipse vs. load for the element level elimination method.

Condition of the Finite Element Stiffness Matrix of Highly Irregular Triangular Grids

I. Fried*

Department of Mathematics
Boston University
Boston, Mass. 02215

A.R. Johnson†

United States Laboratory Command
Army Materials Technology Laboratory
Watertown, Mass. 02172-0001

* Professor
† Mechanical Engineer

Abstract. Penetration of a sharp object causes large concentrated deformations in an elastomer solid. The nonlinear nearly incompressible elastic stress analysis of the solid is done with quadratic triangular elements and displacements referring to an immovable grid. A lower order triangular mesh for a linear thermal analysis is conveniently layed with vertices at the displaced nodes. This gives rise to highly irregular grids of slender elements near the point of maximum penetration. The condition of the global thermal (stiffness) matrix is estimated in terms of the element geometry. It is concluded that no significant decline in the condition of the matrix takes place inspite of the high deformation.

Introduction. To set up the finite element stiffness and mass matrices for plane thermal analysis we need to evaluate

$$I_1 = \int_\Delta (u_x^2 + u_y^2) dx dy \quad \text{and} \quad I_2 = \int_\Delta u^2 dx dy \tag{1}$$

over the typical triangular element Δ, for a linearly assumed temperature distribution u. Consider Δ with three sides l_1, l_2, l_3 and area A. If $u_e^T = (u_1, u_2, u_3)$ is the nodal unknowns vector for Δ, then I_1 and I_2 becomes the quadratic form $I_1 = u_e^T k_e u_e$, with

$$k_e = \frac{1}{8A}\left(l_1^2 \begin{bmatrix} 2 & -1 & -1 \\ -1 & & 1 \\ -1 & 1 & \end{bmatrix} + l_2^2 \begin{bmatrix} & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & \end{bmatrix} + l_3^2 \begin{bmatrix} & 1 & -1 \\ 1 & & -1 \\ -1 & -1 & 2 \end{bmatrix} \right) \tag{2}$$

and $I_2 = u_e^T m_e u_e$, with

$$m_e = \frac{A}{6} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \tag{3}$$

The matrices k_e and m_e are said to be the element stiffness and mass matrices, respectively.

Assembly of k_e and m_e over all Ne finite elements in the grid produces the corresponding global matrices K and M in the manner

$$u^T K u = \sum_e u_e^T k_e u_e \quad , \quad u^T M u = \sum_e u_e^T m_e u_e \tag{4}$$

where u is the global vector of nodal unknowns, and where e indicates summation over all triangles.

With minimization undertaken under the constraints of the boundary conditions our thermal problem is such that

$$\mu_1 = \mu_1(h) = \min_u \frac{u^T K u}{u^T M u} > 0 \tag{5}$$

where h is a linear measure of the element, and

$$\lim_{h \to 0} \mu_1(h) = \lambda_1 > 0 \tag{6}$$

where λ_1 is the fundamental eigenvalue of the problem describing differential operator. For a reasonably fine mesh it is safe to assume $\lambda_1 = \mu_1$.

Actually, for a sufficiently fine mesh we may use the lumped element mass matrix

$$m_e = \frac{A}{3}\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \tag{7}$$

instead of (3).

We denote by λ_1^K and λ_N^K the smallest (1st) and largest (Nth) eigenvalues of K. With proper boundary conditions K is positive definite and we want to estimate its spectral condition number

$$C_2(K) = \frac{\lambda_1^K}{\lambda_N^K} \tag{8}$$

as a function of the inherent and discretization parameters of the problem.

Global bounds. From Rayleigh's theorem we have that

$$\begin{aligned} \lambda_1^K &\le u^T K u \le \lambda_N^K \\ \lambda_1^M &\le u^T M u \le \lambda_N^M \end{aligned} \tag{9}$$

if $u^T u = 1$; while eq. (5) assures us that

$$\frac{u^T K u}{u^T M u} \ge \lambda_1 \tag{10}$$

for an arbitrary vector u that satisfies the essential boundary conditions.

If we choose u in eq. (10) so that $u^T K u = \lambda_1^K$, then we have from eq. (9) that

$$\lambda_1^K \ge \lambda_1 \lambda_1^M \tag{11}$$

On the other hand if we start with $u^T K u / u^T M u = \lambda_1$, then we obtain

$$\lambda_1^K \le \lambda_1 \lambda_N^M \tag{12}$$

or combined

$$\lambda_1 \lambda_1^M \le \lambda_1^K \le \lambda_1 \lambda_N^M \tag{13}$$

The usefulness of eq. (13) lies in the fact that M is positive definite with a spectral condition number that is independent of h.

The bounds in (13) are most critical and to make them tightest we want $\lambda_N^M / \lambda_1^M$ as close to 1 as possible. This can be achieved with a *nonuniform* density distribution, and

$$I_2 = \int_\Delta \rho u^2 dx dy, \quad \int_{\sum_e \Delta} \rho dx dy = 1 \ , \ \rho(x,y) \ge 0 \tag{14}$$

instead of (1). A variable density distribution affects λ_1, that need be assessed for it. We shall not pursue this matter here as we shall see it not very essential to the present situation.

Element bounds. On matrices that are not diagonally dominant, and the high order finite element and finite difference matrices are such, Gerschgorin's theorem fails at the lower end of the eigenvalues spectrum. The eigenvalue bounds for the *global* stiffness matrix are written here in terms of the eigenvalues of the *element* matrices. We shall show them sharp and most convenient in the finite element analysis.

If u denotes the global unknown vector and u_e the one for the typical eth element, then

$$u^T u \leq \sum_e u_e^T u_e \leq u^T u \; p_{max} \tag{15}$$

where p_{max} denotes the maximum number of elements that share a common node. Six is a typical value for p_{max} in plane problems.

To write bounds on λ_N^K we choose a normalized u, $u^T u = 1$, and such that

$$\lambda_N^K = u^T K u = \sum_e u_e^T k_e u_e \tag{16}$$

If $\lambda_n^{k_e}$ denotes the largest eigenvalue of the positive semi definite, k_e, then for any u_e

$$u_e^T k_e u_e \leq \lambda_n^{k_e} u_e^T u_e \tag{17}$$

and eq. (16) yields with it

$$\lambda_N^K \leq \max_e(\lambda_n^{k_e}) \sum_e u_e^T u_e \leq p_{max} \max_e(\lambda_n^{k_e}) \tag{18}$$

A lower bound on λ_N^K is obtained from $\lambda_N^K \geq u^T K u$, $u^T u = 1$. Choosing $u_i = 0$ at all nodes except for u_e that corresponds to the maximum eigenvalue of k_e produces the desired upper bound, and we have

$$\begin{aligned} \max_e(\lambda_n^{k_e}) &\leq \lambda_N^K \leq p_{max}(\lambda_n^{k_e}) \\ \max_e(\lambda_n^{m_e}) &\leq \lambda_N^M \leq p_{max}(\lambda_n^{m_e}) \end{aligned} \tag{19}$$

Since the element stiffness matrix k_e is usually only positive *semi* definite the bound

$$\lambda_1^K \geq \min_e(\lambda_1^{k_e}) \tag{20}$$

where $\lambda_1^{k_e}$ is the lowest eigenvalue of k_e, reduces to the trivial $\lambda_1^K \geq 0$. But the element mass matrix m_e is positive definite, $\lambda_1^{m_e} > 0$ for all e, and

$$\lambda_1^M \geq \min_e(\lambda_1^{m_e}) \tag{21}$$

is useful.

We combine eqs. (13), (19) and (21) to write

$$\lambda_1 \min_e(\lambda_1^{m_e}) \leq \lambda_1^K \leq \lambda_1 p_{max} \max_e(\lambda_n^{m_e}) \tag{22}$$

Nearly collapsed triangles. The element stiffness matrix k_e in eq. (2) is of rank two. For $u_e^T = (1,1,1)$ we have $u_e^T k_e u_e = 0$. To write the two nonzero eigenvalues of k_e we introduce the notation $r_i = l_i^2/(2A)$, $i = 1,2.3$, and have that

$$\lambda_{2,3}^{k_e} = \frac{1}{4}(r_1 + r_2 + r_3 \pm \sqrt{2}\sqrt{(r_1 - r_2)^2 + (r_2 - r_3)^2 + (r_3 - r_1)^2}) \tag{23}$$

To observe what happens to λ_N^K when elements collapse we consider the triangle in Fig. 1, and readily compute for it

$$2A = l^2 \sin\gamma,\ r_1 = r_2 = \frac{1}{\sin\gamma}\ ,\ r_3 = \frac{2(1-\cos\gamma)}{\sin\gamma} \tag{24}$$

If $\gamma \cong 0$, $\lambda_3^{k_e} = \gamma^{-1}$, and

$$\frac{1}{\gamma} \le \lambda_N^K \le \frac{6}{\gamma} \tag{25}$$

while if $\gamma \cong 180^o$, $\lambda_3^{k_e} = 3(180-\gamma)^{-1}$, and

$$\frac{3}{180-\gamma} \le \lambda_N^K \le \frac{18}{180-\gamma} \tag{26}$$

From the lumped m_e in eq. (7) we derive

$$\lambda_1^M \ge \frac{1}{3}\min_e(A_e) \tag{27}$$

and consequently

$$\lambda_1^K \ge \frac{1}{3}\lambda_1 \min_e(A_e) \tag{28}$$

which assures us that $\lambda_1^K > 0$ if $\lambda_1 > 0$. But with a careful consideration of the specific mesh we can do better than (28). Consider the mesh in Fig. 2 that includes one slender element with area A_1. As $A_1 \to 0$ the mass of this element reduces to zero but because it shares nodes with large elements λ_1^M is nearly unaffected by a small A_1. Actually, the smallest mass is at point B.

Equation (22) guarantees that under the circumstances of Fig. 2, λ_1^K is not likely to change much with A_1. We shall be more specific about that in the next section.

Penetrated elastomer. Track pads that run over sharp objects suffer very large localized deformations. Figure 3 shows the deformation of, an originally straight, cylinder ABCD made of rubberlike material, as a result of point C penetrating the body along axis AC. Elastic computation is done with *quadratic* elements and a nearly incompressible material.

First order triangular elements are judged adequate for a superposed thermal analysis of the solid and for computational convenience the new mesh is drawn with vertices fixed at the *displaced* nodes. The resulting triangular grid is shown in Fig. 3. Sharp elements are created near point C prompting us to suspect a loss of conditioning. Notice that the elastic deformation is nearly area preserving but each new individual triangle need not have the same area.

It is the theoretical and computational conclusion of this paper that the large deformations and slender elements observed in Fig. 3 have only a marginal influence on the condition of the global thermal matrix.

The dangerously collapsed triangles in Fig. 3 are with a small angle γ and we have from eq. (25) that

$$\frac{1}{\sin\gamma} \leq \lambda_N^K \leq \frac{6}{\sin\gamma} \tag{29}$$

and λ_N^K is nearly proportional to γ^{-1}.

We observe that the smallest mass is at point D, while the biggest mass is at the interior points. Hence

$$\lambda_1^M = \frac{h^2}{6} \quad \text{and} \quad \lambda_N^M = h^2 \tag{30}$$

and we have from eq. (13) that

$$\lambda_1 \frac{h^2}{6} \leq \lambda_1^K \leq \lambda_1 h^2 \tag{31}$$

Consequently

$$\frac{1}{\lambda_1 h^2 \sin\gamma} \leq C_2(K) \leq \frac{36}{\lambda_1 h^2 \sin\gamma} \tag{32}$$

The number λ_1 is not known exactly, and under large deformations it is slightly displacements dependent. To have an idea of what λ_1 can be we recall that for a unit square membrane edge fixed $\lambda_1 = 2\pi^2$. But even without a numerical value for λ_1, equation (32) clearly tells us how the spectral condition number $C_2(K)$ of the global thermal matrix K depends on h and γ.

In the mesh of Fig. 3 the temperature is prescribed along AB and at point C. Using conjugate gradients to minimize and maximize $u^T K u / u^T u$ we compute

$$\lambda_1^K = 0.0285,\ \lambda_N^K = 7.82\ ,\ C_2(K) = 274$$

for the undeformed mesh, and

$$\lambda_1^K = 0.0323\ ,\ \lambda_N^K = 16.6\ ,\ C_2(K) = 514$$

for the deformed mesh. In agreement with the theoretical prediction of eq. (31). λ_1^K is nearly independent of the deformation.

About three significant digits are lost in the thermal finite element analysis of the deformed body in Fig. 3, a wholly tolerable loss on computers that typically carry seven digits in single precision and 16 in double.

References

1. I. Fried, Bounds on the extremal eigenvalues of the finite element stiffness and mass matrices and their spectral condition number. *J. Sound and Vibration* (1972), 407-418.

2. I. Fried, Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices. *Int. J. Solids Structures* (1973), 1013-1034.

3. I. Fried, *Numerical Solution of Differential Equations* (1979). Academic Press.

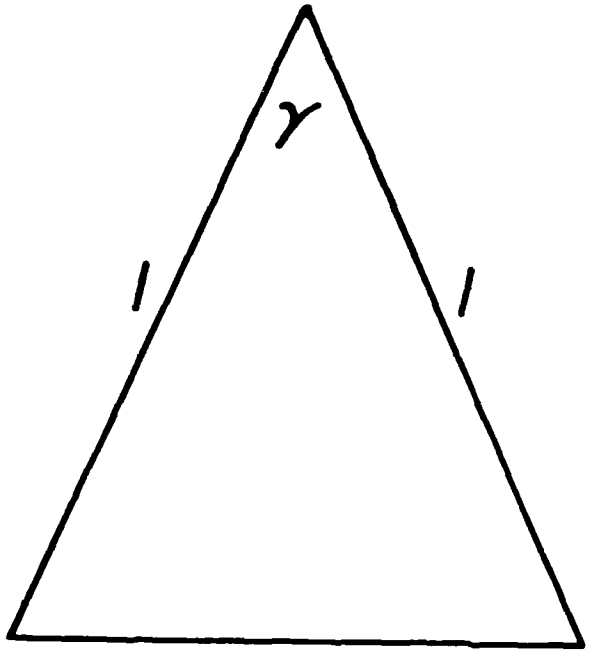


Fig.1

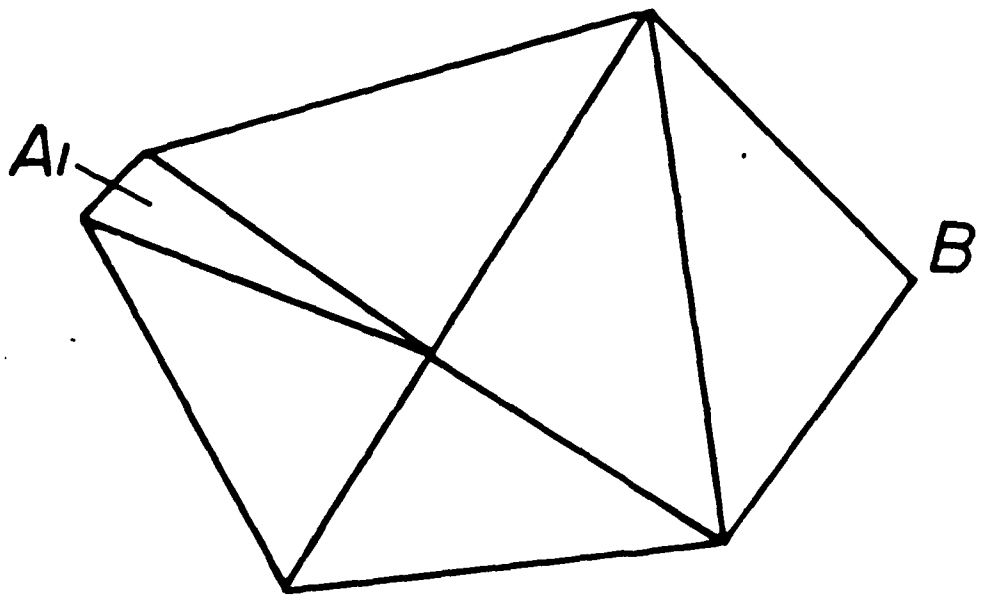


Fig.2

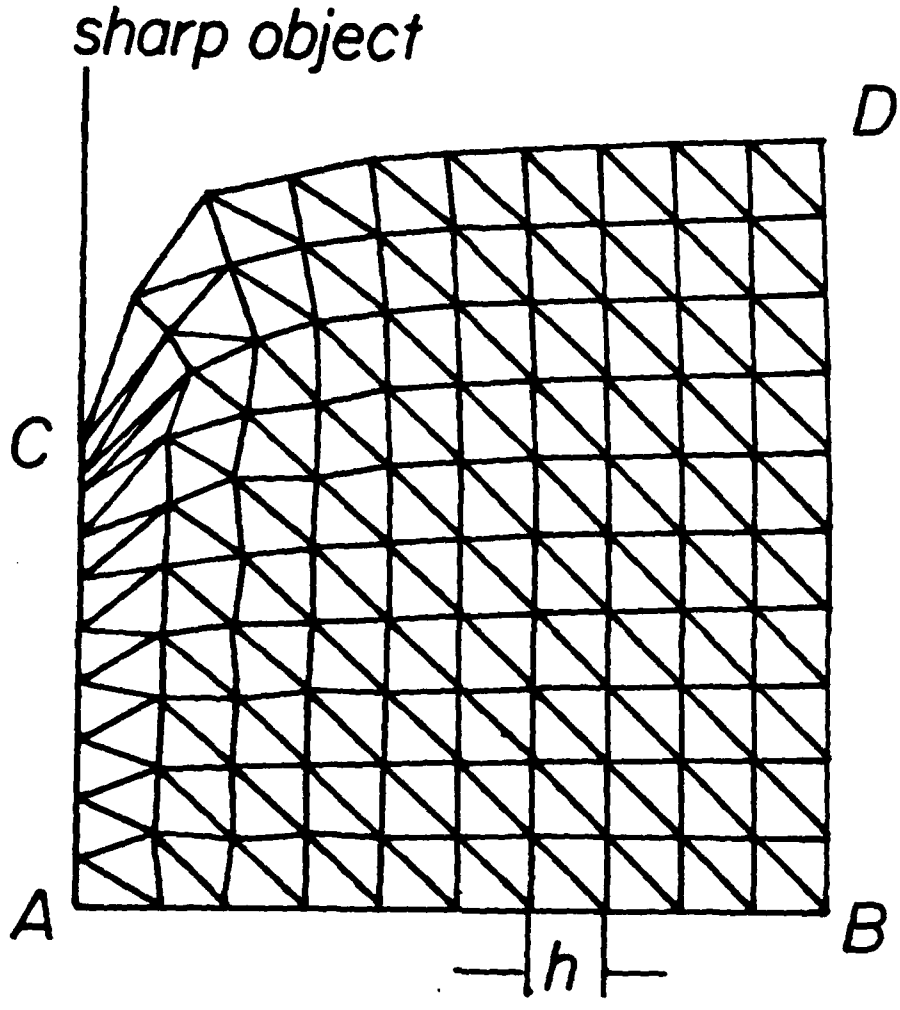


Fig. 3

A SIMPLE ANALYSIS OF SWAGE AUTOFRETTAGE PROCESS

Peter C. T. Chen
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

ABSTRACT. Many solutions have been reported for the hydraulic autofrettage process. In this paper a simple analysis of the swage autofrettage process is presented. The contact pressure at different locations is determined as a function of interference. The deformation and stress distribution during autofrettage is obtained. At the end of the autofrettage process, the permanent bore enlargement and residual stresses are calculated. Numerical results are presented in graphical forms.

I. INTRODUCTION. To increase the maximum pressure a cylinder can contain without plastic deformation and to enhance its fatigue life, residual stresses are often produced in cylinders through autofrettage [1]. Many solutions have been reported for the hydraulic autofrettage process [2-6]. The thick-walled cylinders were subjected to uniform internal pressure of sufficient magnitude to cause plastic deformation and then the pressure was removed.

A more economical way of producing residual stresses in thick-walled cylinders is the swage autofrettage process. This process is carried out by a swage, the diameter of which is greater than the inner diameter of the cylinder. This swage is driven through the cylinder from one end to the other. A rigorous analysis of this process is difficult. In this paper a simple analysis of the swage autofrettage process is reported. The swage mandrel and the cylinder are made of tungsten carbide and steel, respectively. A two-dimensional plane-strain analysis is used to determine the contact pressure at different locations of the cylinder as a function of interference. The deformation and stress distribution during autofrettage are obtained. At the end of the autofrettage process, the permanent bore enlargement and residual stresses are calculated.

II. ELASTIC SWAGING. The swage mandrel is assumed to be a short cylindrical bar driven through a long thick-walled cylinder from one end to the other. The diameter of the mandrel (2c) is a constant, but the inner and outer diameters (2a and 2b) of the tube are variables. When the difference between c and a is positive, we have interference I. For small values of interference, the stress state in the swaging assembly is elastic. The stresses and displacement in the tube are

$$\begin{matrix}\sigma_r \\ \sigma_\theta\end{matrix} = \frac{P}{1 - a^2/b^2}\left[\frac{a^2}{b^2} \mp \frac{a^2}{r^2}\right]$$

$$\frac{u}{r} = \frac{P/E}{1 - a^2/b^2}\left[(1+\nu)\frac{a^2}{r^2} + (1-\nu-2\nu^2)\frac{a^2}{b^2}\right] \qquad (1)$$

and in the mandrel

$$\sigma_r = \sigma_\theta = -P$$

$$u/r = -(1-\nu_1-2\nu_1^2)P/E_1 \qquad (2)$$

where E, ν, and E_1, ν_1 are the material constants of the tube and mandrel, respectively. At the interface, u_a (tube) - u_a (mandrel) = I by the compatibility requirement. The interference pressure (p) is a function of the interference (I) given by

$$p = \frac{EI}{a}\left(1 - \frac{a^2}{b^2}\right)/\left[(1+\nu) + (1-\nu-2\nu^2)\frac{a^2}{b^2} + (1-\nu_1-2\nu_1^2)\left(1 - \frac{a^2}{b^2}\right)E/E_1\right] \qquad (3)$$

For sufficiently large values of the interference, the stresses in the tube reach the yield limit. Assuming that Tresca's yield condition governs the behavior of the material, the tube first becomes plastic at the interference when the stresses satisfy $\sigma_\theta - \sigma_r = \sigma_o$, where σ_o is the initial tensile yield stress. The solution for the critical interference pressure to cause incipient plastic deformation is

$$p^* = \tfrac{1}{2}\,\sigma_o\,(1 - a^2/b^2) \qquad (4)$$

and it follows from Eq. (3) that the interference for the onset of plastic flow is

$$I^* = \frac{\sigma_o}{E}\,\frac{a}{2}\left[(1+\nu) + (1-\nu-2\nu^2)\frac{a^2}{b^2} + (1-\nu_1-2\nu_1^2)\left(1 - \frac{a^2}{b^2}\right)E/E_1\right] \qquad (5)$$

which reduces to $I^* = (1-\nu^2)\,a\,\sigma_o/E$ for the special case ($E_1 = E$, $\nu_1 = \nu$).

III. SWAGING BEYOND THE ELASTIC LIMIT. For values of interference larger than that given by Eq. (5), a plastic zone forms in the tube, so that for $a \leq r \leq \rho$ the tube is plastic, while for $\rho \leq r \leq b$ the tube material is still in an elastic state. The elastic-plastic interface radius ρ is a function of the interference I.

We assume that the steel tube is elastically-ideally plastic, obeying the Tresca's yield criterion and the associated flow theory, but the tungsten carbide mandrel is elastic. This assumption is justified because the strength ratio of tungsten carbide to steel is about three. For loading beyond the elastic limit, the closed-form solution has been found by Koiter [2]. The expressions for the stresses and displacement in the tube are

$$\begin{matrix}\sigma_r/\sigma_o \\ \sigma_\theta/\sigma_o\end{matrix} = \tfrac{1}{2}\left(\mp 1 + \frac{\rho^2}{b^2}\right) - \log\frac{\rho}{\rho}\ , \quad \text{in } (a \leq r \leq \rho) \qquad (6)$$

$$\begin{matrix}\sigma_r/\sigma_o \\ \sigma_\theta/\sigma_o\end{matrix} = \tfrac{1}{2}\left(\frac{\rho^2}{b^2} \mp \frac{\rho^2}{r^2}\right)\ , \quad \text{in } (\rho \leq r \leq b) \qquad (7)$$

$$\frac{E}{\sigma_o}\frac{u}{r} = (1-2\nu)(1+\nu)\frac{\sigma_r}{\sigma_o} + (1-\nu^2)\frac{\rho^2}{r^2} \tag{8}$$

where the elastic-plastic interface (ρ) is related to the internal pressure (p) by

$$P/\sigma_o = \tfrac{1}{2}(1 - \rho^2/b^2) + \log(\rho/a) \tag{9}$$

For swaging beyond the elastic limit, the compatibility requires u_a (tube) - u_a (mandrel) = I at the interface, i.e.,

$$\frac{E}{\sigma_o}\frac{I}{a} = (1-\nu^2)\frac{\rho^2}{a^2} - \frac{P}{\sigma_o}[(1-2\nu)(1+\nu) - (1-\nu_1-2\nu_1^2)\frac{E}{E_1}] \tag{10}$$

Equations (9) and (10) give us a parametric representation of relating p to I through the parameter ρ. The contact pressure at different locations can thus be determined as a function of the interference I.

IV. UNLOADING ANALYSIS. After swaging, the permanent bore enlargement and residual stresses can be calculated by an unloading analysis. Let a double prime denote a component in the residual state, i.e., $\sigma_\theta'' = \sigma_\theta + \sigma_\theta'$. Assuming elastic unloading, the solution is given by

$$\begin{matrix}\sigma_r' \\ \sigma_\theta'\end{matrix} = \frac{p}{b^2/a^2 - 1}\left[\pm\frac{b^2}{r^2} - 1\right] \tag{11}$$

$$E\,u'/r = -[(1-\nu) + (1+\nu)b^2/r^2]p/(b^2/a^2-1) \tag{12}$$

In a recent paper [6], this author presented a more rigorous elastic-plastic unloading analysis based on a new theoretical model considering the Bauschinger and hardening effects during unloading. This mode is a very good representation for the material behavior of the high strength steel used in gun barrels [7]. Taking into account the Bauschinger effect (f) and the strain-hardening during unloading (m'), we have obtained a closed-form solution. On unloading, yielding will occur for $a \leq r < \rho'$ with $\rho' < \rho$. The stresses in the reverse yielding zone ($a \leq r < \rho'$) are given by

$$\sigma_r'/\sigma_o = P/\sigma_o - \tfrac{1}{2}\beta_2'(1+f)(\rho'/a)^2(1-a^2/r^2) - (1-\beta_2')(1+f)\log(r/a) \tag{13}$$

$$\sigma_\theta'/\sigma_o = \sigma_r'/\sigma_o - (1+f)[1 + \beta_2'(\rho'^2/r^2-1)] \tag{14}$$

where

$$\beta_1' = (1-m')/[m' + \frac{\sqrt{3}}{2}\frac{(1-m')}{(1-\nu^2)}] \quad , \quad \beta_2' = m'\beta_1'/(1-m') \tag{15}$$

The stresses in the elastic zone ($\rho' \leq r \leq b$) are

$$\begin{matrix} \sigma_r'/\sigma_o \\ \sigma_\theta'/\sigma_o \end{matrix} = \tfrac{1}{2}(1+f)[\pm (\rho'/r)^2 - (\rho'/b)^2] \qquad (16)$$

The displacement for the entire tube ($a \leq r \leq b$) is

$$(E\sigma_o)u'/r = (1-2\nu)(1+\nu)(\sigma_r'/\sigma_o) - (1-\nu^2)(1+f)(\rho'/r)^2 \qquad (17)$$

The residual stresses and displacement are found by addition

$$\sigma_r'' = \sigma_r + \sigma_r' \quad , \quad \sigma_\theta'' = \sigma_\theta + \sigma_\theta' \quad \text{and} \quad u'' = u + u' \qquad (18)$$

V. NUMERICAL RESULTS AND DISCUSSION. The material constants used in the calculations are $E = 206.84$ GPa, $\nu = 0.3$, $\sigma_o = 1.29$ GPa, $m' = 0.3$ for the high strength steel and $E_1 = 610.19$ GPa, $\nu_1 = 0.258$ for the tungsten carbide mandrel. The radius of the mandrel is a constant $c = 58.42$ mm, but the thickness of the tube varies along the axial direction with the inner radius (a) increasing slightly and the external radius (b) tapering more rapidly. The values of a and b at four typical sections are $a_j = 56.96, 57.82, 57.99, 58.63$ mm and $b_j = 157.50, 106.75, 83.00, 83.00$ mm, for $j = 1,2,3,4$, respectively. The corresponding values of wall ratio are $b_j/a_j = 2.765, 1.846, 1.431, 1.42$ at four sections. The interference during swaging (I) is the positive difference between c and a. The values of I at four sections are $I_j = 1.46, 0.60, 0.43, -0.21$ mm for $j = 1,2,3,4$. The negative value of I_4 means that there is no contact between the mandrel and the tube. For the positive values of interference, the contact pressure and the stress distribution during swaging can be obtained using the methods presented in Sections II and III. The information after swaging can be obtained by the unloading analysis presented in Section IV.

The numerical results are presented in terms of the dimensionless quantities defined by

$$\bar{r} = r/a \quad , \quad \bar{p} = P/\sigma_o \quad , \quad \bar{\sigma}_\theta = \sigma_\theta/\sigma_o$$

$$\bar{I} = (E/\sigma_o)I/a \quad , \quad \bar{u} = (E/\sigma_o)u/a \quad , \quad \text{etc.} \qquad (19)$$

The contact pressure ($\bar{p}$) and hoop stress ($\bar{\sigma}_\theta$) at the interface are presented as functions of the interference (I) in Figures 1, 2, and 3 for wall ratios b/a = 2.765, 1.846, 1.431, respectively. The results for swaging within and beyond the elastic limit are included. The pressure is a monotonous increasing function of the interference, but the maximum value of hoop stress occurs at the onset of plastic flow as shown in the dotted curves. Initial yielding occurs at $\bar{I}^* = 0.774, 0.799, 0.830$, and fully plastic flow occurs at $\bar{I}^{**} = 6.638, 2.909, 1.751$ for three different wall ratios, respectively. The actual values of interference (I) at three chosen sections are $\bar{I}_1 = 4.10$, $\bar{I}_2 = 1.66$, $\bar{I}_3 = 1.19$. These values indicate that the swaging is partially plastic at these sections in zones 1, 2, and 3. The corresponding locations of elastic-plastic boundary are given by $\rho/a = 2.2001, 1.4196, 1.19205$, and the amounts of overstrain are 68, 49.6, and 44.6 percent, respectively. Also shown in Figures 1, 2, and 3 are the

values of contact pressure ($\bar{p}$ = 0.972, 0.555, 0.671) and the hoop stress at the interface $\sigma_\theta = 1 - p$. The distributions of hoop stresses during swaging are shown in Figure 4 for typical sections in three zones. The maximum value of hoop stress occurs at the elastic-plastic boundary. The information for the displacement and stresses after swaging can be obtained by an unloading analysis. The distributions of residual hoop stresses are shown in Figure 5 for the chosen sections in three zones. Elastic unloading analysis is justified in zone 3, but reverse yieldings occur in zones 1 and 2 with ρ'/a = 1.305 and 1.014, respectively. Finally the distributions of residual displacements (u'') at typical sections in three zones are presented in Figure 6. Also shown in this figure are the experimental data at the bore. The agreement between the calculated and experimental data is excellent in zone 1, but not so good in zones 2 and 3. This suggests that a more refined analysis is needed for sections with smaller wall ratios. An investigation based on the finite element method is being made and the results will be reported in the near future.

REFERENCES

1. Davidson, T.E. and Kendall, D.P., "The Design of Pressure Vessels for Very High Pressure Operation," Mechanical Behavior of Materials Under Pressure, (H.L.P. Pugh, ed.), Elsevier Co., 1970.

2. Hill, R., The Mathematical Theory of Plasticity, Oxford University Press, London, 1950.

3. Bland, D.R., "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and Temperature Gradients," Journal of Mechanics and Physics of Solids, Vol. 4, 1956, pp. 209-229.

4. Franklin, G.J. and Morrison, J.L.M., "Autofrettage of Cylinders: Prediction of Pressure/External Expansion Curves and Calculation of Residual Stresses," Proceedings of the Institute of Mechanical Engineers, Vol. 174, 1960, pp. 947-974.

5. Chen, P.C.T., "The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes," Proceedings of Army Sumposium on Solid Mechanics, The Role of Mechanics in Design-Ballistic Problems, 1972, pp. 243-253.

6. Chen, P.C.T., "The Bauschinger and Hardening Effect on Residual Stresses in an Autofrettaged Thick-Walled Cylinder," Journal of Pressure Vessel Technology, Vol. 108, February 1986, pp. 108-112.

7. Milligan, R.V., Koo, W.H., and Davidson, T.E., "The Bauschinger Effect in a High Strength Steel," Journal of Basic Engineering, Vol. 88, pp. 480-488.

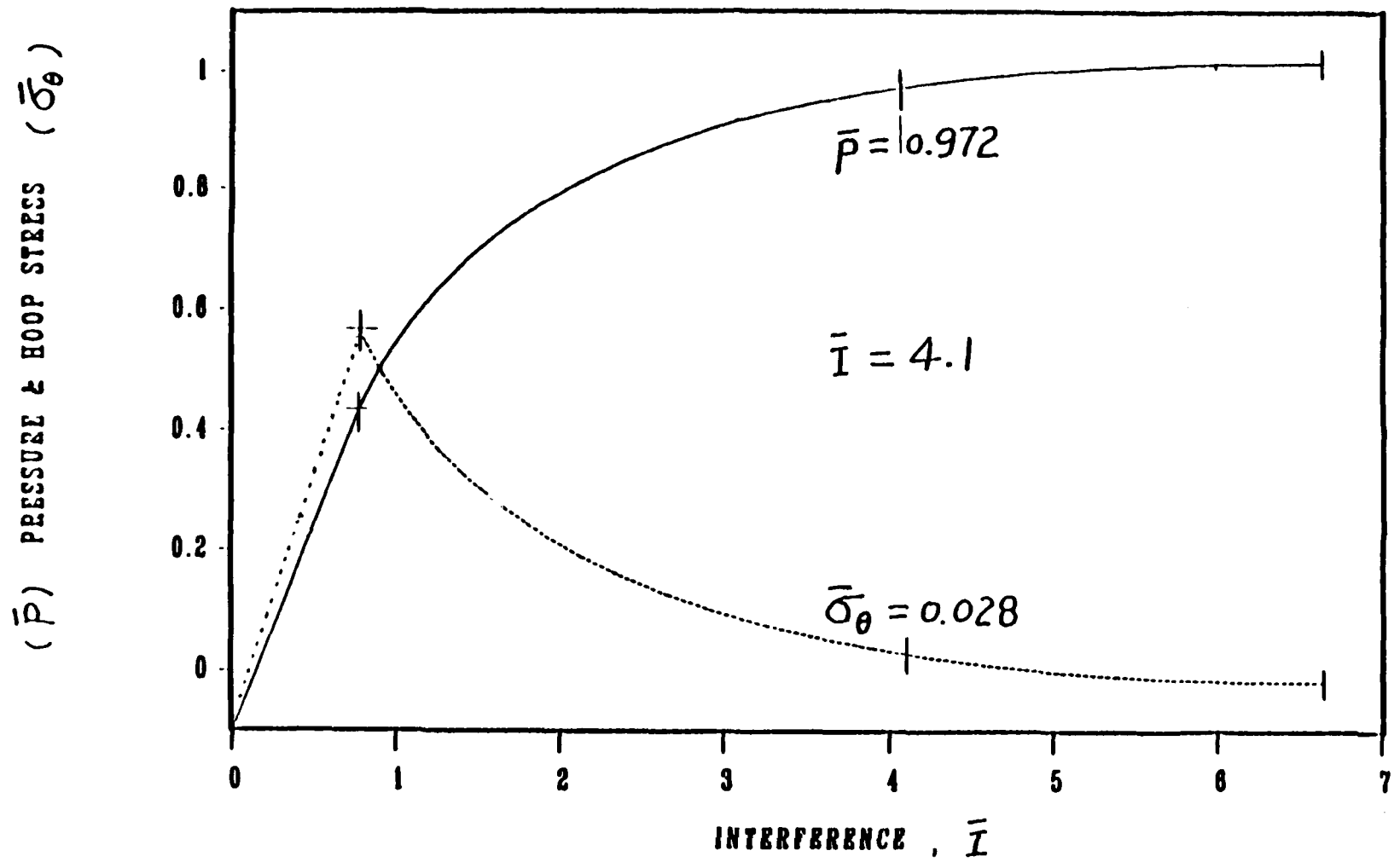


Figure 1

Contact pressure and hoop stress at the interface as functions of interference for a section in zone 1.

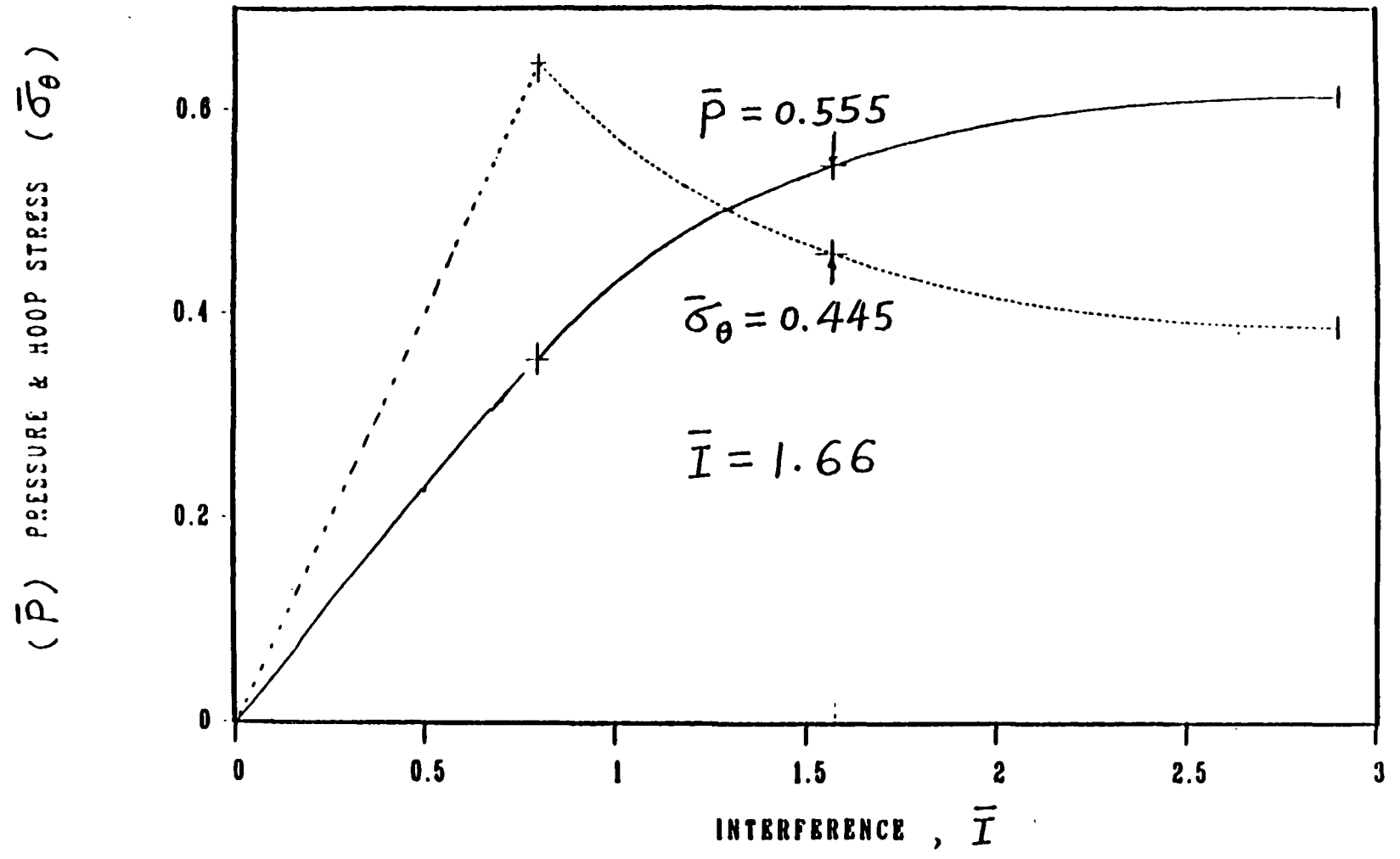


Figure 2

Contact pressure and hoop stress at the interface as functions of interference for a section in zone 2.

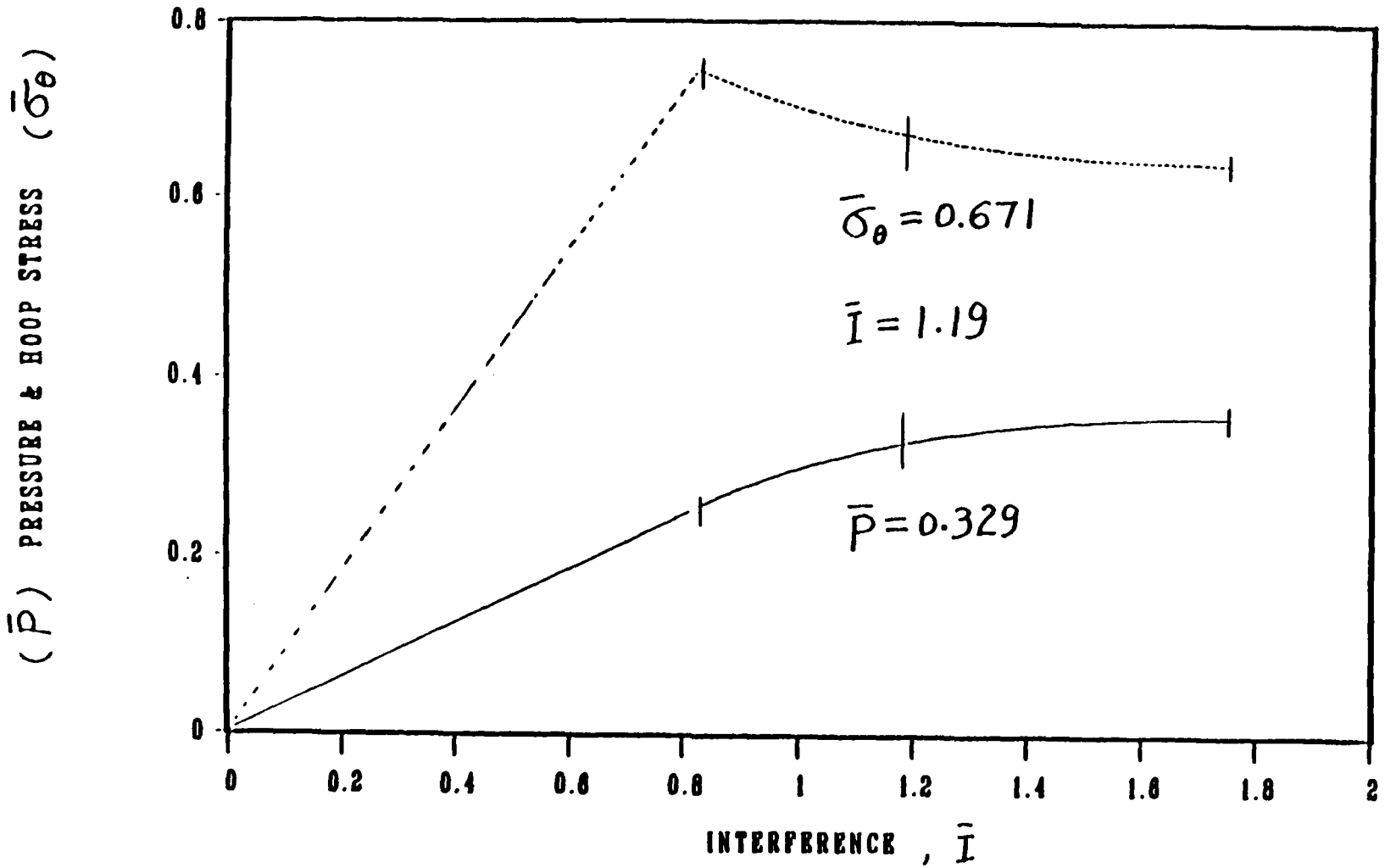


Figure 3

Contact pressure and hoop stress at the interface as functions of interference for a section in zone 3.

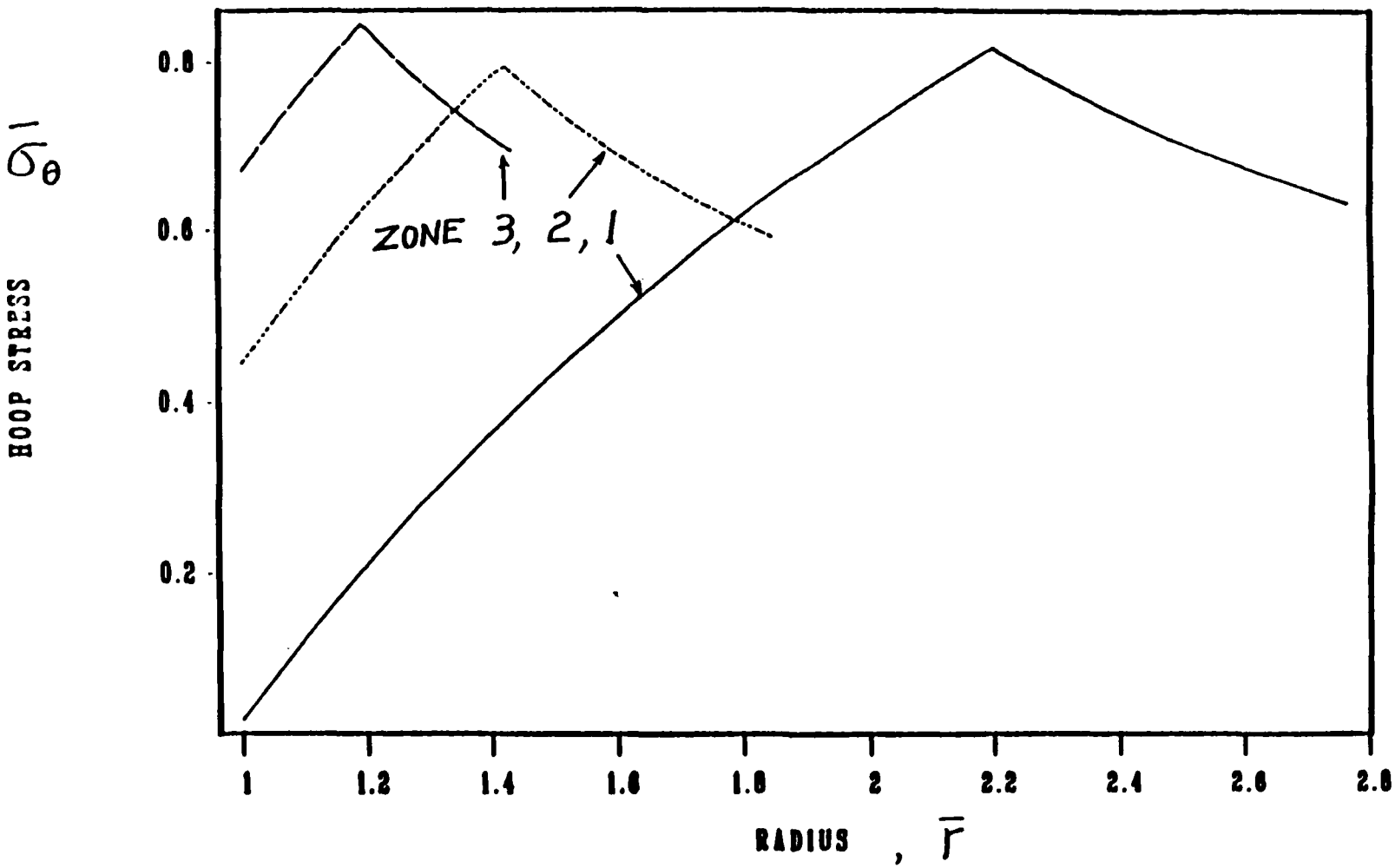


Figure 4

The hoop stress distributions at three sections.

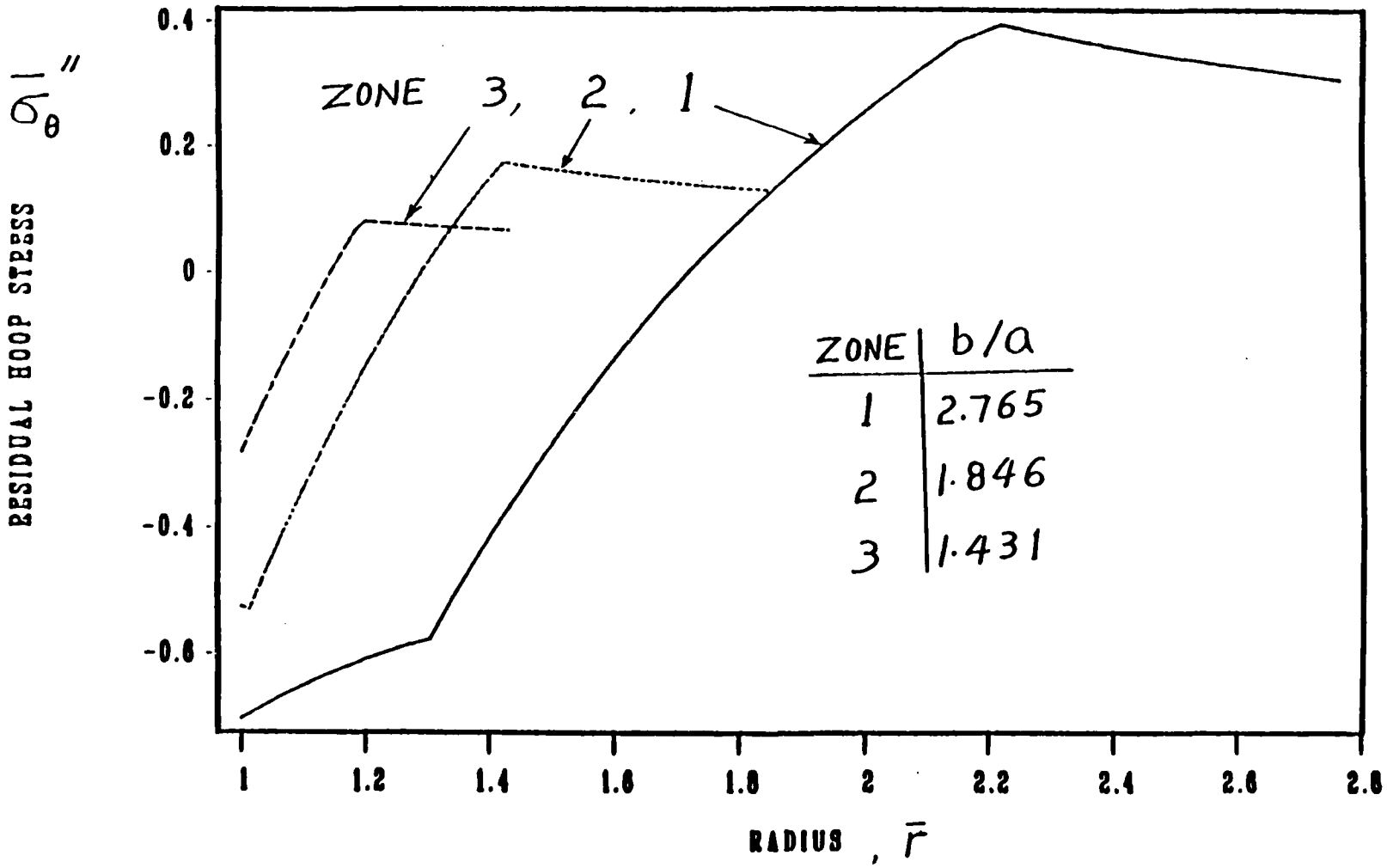


Figure 5

The distributions of residual hoop stresses at three sections.

Figure 6

The distributions of residual displacements at three sections.

OPTIMAL DESIGN OF A TWO-WAY CONDUCTOR

Gilbert Strang
Department of Mathematics
Massachusetts Institute of Technology
Cambridge MA 02139

Robert Kohn
Courant Institute
New York University
New York NY 10012

Introduction

Optimal design presents an extreme case of non-smooth mechanics. The unknown becomes the density of material, and in an ordinary design the density takes the value 0 or 1. It describes a shape which has least weight subject to the constraints. However the optimal design is frequently not at all ordinary. It is given by the "weak limit" of a sequence of designs in which the density oscillates more and more rapidly between 0 and 1. In other words the average density can take fractional values, and no ordinary shape achieves the minimum weight.

Mathematically this is an instance of the relaxation of a nonconvex problem. That is a special topic in the calculus of variations, to widen the class of admissible functions so that the problem becomes correctly posed and its minimum is achieved. To the given nonconvex problem we associate a relaxed problem with the same minimum. The solutions of the relaxed problem are the weak limits of minimizing sequences in the original problem.

Our application of this technique is to a question of "optimal bounds" for composite materials. Its solution has been a major achievement of the Tartar-Murat method of

compensated compactness. That method also searches for functionals that are weakly lower semicontinuous--which is the key property implied by convexity. If the unknown is a vector function like $(u_1(x,y),u_2(x,y))$ then convexity must be replaced by quasiconvexity; this is crucial below. Our goal is to give a variational statement of the problem of optimal bounds, and an alternate approach to its solution--in which we find the relaxed problem and solve it.

The Design of a Conductor

How should a fixed number of resistors be arranged, in order to maximize the current? For current in one direction the answer is easy: They go in parallel. The combined resistance is the harmonic mean of the individual resistances--or equivalently, the net conductance is the sum of the conductances. The problem becomes more serious if we are measuring two currents, north-south and east-west. In that case resistors in one direction contribute little or nothing to flow in the other direction. An optimal two-way design is not clear. It is a much more complicated series-parallel connection, and the rules of the competition become important.

We propose to make the problem continuous rather than discrete. Instead of current between nodes, we measure current across a unit square. In that square we place conducting material--as much as we have, in the best orientation we can find. If the area of conducting material is A, it leaves an insulated area $1 - A$ through which nothing flows. Then we impose a unit voltage difference between the left and right sides of the square, or the top and bottom, and measure the current.

For flow in one direction, the design problem is easy. By

placing the conductor in a strip across the square, or in several parallel strips, the current is maximized. Suppose the specific resistance of the conductor is also unity (completing a wanton destruction of dimensional arguments). When $A = 1$ and the square is full, the net current is 1. As the area (and strip width) A goes below 1, the horizontal current remains equal to A. The overall resistance is $1/A$ in that direction; the resistance in the vertical direction is infinite.

The real problem is to design a single conductor to carry flow in both directions, up the square as well as across. Those measurements are done separately. A voltage difference between $x = 0$ and $x = 1$ produces horizontal current, and between $y = 0$ and $y = 1$ it produces vertical current. A strip that does well for one does badly for the other. The question now involves a function of two variables:

> To achieve a horizontal current $C < 1$ and a vertical current $D < 1$, what is the minimum possible conducting area A ?

For $C = \frac{1}{3}$ and $D = 0$, the minimum area is $A = \frac{1}{3}$. The conducting material is in horizontal strips. For $C = \frac{1}{3}$ and $D = \frac{1}{3}$, the natural construction is to use both horizontal and vertical strips. If their widths are $\frac{1}{3}$, then the area they cover is

$$A = \frac{1}{3} + \frac{1}{3} - \frac{1}{9} = \frac{5}{9} \text{ (after subtracting } \frac{1}{9} \text{ for overlap).}$$

In general the area occupied by a two-strip design is $C + D - CD$. This "Red Cross design" will certainly carry the required currents (or more), with voltage differences equal to

1. The question is whether the area $A = C + D - CD$ is minimal. The answer is no.

The simplest design (Fig. 1) is not optimal. In fact it carries more current than required; we have considerably underestimated its conductance, by taking it to be C horizontally and D vertically. When the strips have width $\frac{1}{2}$, the actual currents (in each direction separately) are $1/\sqrt{3}$. Heuristically, part of the horizontal current makes use of the vertical strip. The computation uses Laplace's equation in the cross, with potentials 1 and 0 on the left and right sides. This design can achieve currents $C = D = 1/\sqrt{3}$ with area $\frac{3}{4}$ (which is less than $C + D - CD$). Nevertheless the area can still be reduced.

In this note we describe one possible optimal design (it is already known). More precisely, we describe a sequence of designs whose areas approach the minimum value A. That value is achieved in the limit, which becomes a composite material--a mixture of conductors and insulators with a properly chosen microstructure. The effective conductances of this limiting composite can be computed, and they are C and D. It is a straightforward problem in homogenization, except that the goal is not the usual one--to compute effective conductances for a given microstructure. Our problem is optimization, to find the best composite.

The design is not unique. For equal values of C and D the composite will be isotropic, and an optimal design was found by Hashin and Shtrikman [1]. They filled the square with circular disks, each consisting of a conducting ring around a smaller insulated disk. With the right ratio of radii, and a packing by infinitely many disks, the properties are optimal. The extension to anisotropic designs ($C \neq D$) was found by

Tartar and Murat [2], who replaced the circles by ellipses. But the real achievement of these authors was not the construction of optimal designs; it was the proof that no other design could be better. That is a subtle problem, to admit all microstructures. It led Tartar and Murat to develop the theory of "compensated compactness," a systematic approach to weak limits--when functions (or designs) can oscillate more and more rapidly, and only certain average values have a stable meaning. That theory has extremely valuable applications, far outside the present problem.

Our goal is to contribute one more proof that the area A, given below, is actually minimal. It is based squarely on [3], in which we computed the minimum value of a specific nonconvex functional. That nonconvexity is typical of optimal design theory, in which the original statement is a "0-1 problem"--there is a conductor or an insulator in each subregion. Just as integer programming is nonconvex and difficult in comparison with linear programming, so our continuous problem needs to be relaxed to a variational problem with reasonable solutions and the same minimum. Those reasonable solutions will be the weak limits, or averages, of the unreasonable designs which appear in the 0-1 formulation. In other words, we allow ourselves to construct composite materials out of the original materials, and this homogenization process gives to the original nonconvex problem a new and more satisfactory form. In the case of one current it becomes convex. In our present case of two currents it becomes polyconvex, and can be solved.

The construction will not be based on circles or ellipses. A different class of designs was developed by Lurie and Cherkaev [4], who stayed with strips but made them extremely thin. In the limit it is the density and direction of the

strips that determines everything--it decides the conducting area A and the macroscopic properties of the composite. With this "strip construction" the calculations become easier; the next section nearly returns to resistors in series and parallel. The construction is also a realization in physical terms of the mathematical process of convexification--to fill in the line segments between any pair of points, and then to fill in line segments to any of the new points, and so on. In the case of two currents and a conductivity matrix, these become line segments between matrices whose difference has rank one. That is algebraically more delicate, and in the present construction it produces "strips of strips"--but the underlying idea is not changed.

One contribuiton of this note is to give a fresh statement of the variational problem. We have found it useful to ask for the minimum area A as a function of C and D, rather than to describe all the conductivity tensors that can be achieved with prescribed area A. (The two forms are equivalent; it is like giving a function A(C,D) instead of its level sets.) We will not study the worst composites, which are also of interest and are closely related. Finally we reemphasize that the construction is easier to discover than a proof of its optimality, but nevertheless several proofs have been given: Kohn and Milton [5] have provided a comprehensive analysis of the problem of optimal bounds.

The principal application is to structural problems--the weight minimization of an elastic body subject to constraints from the loads. This is the shape optimization pioneered by Michell and Prager, and highly developed in the work of Rozvany [6,7]. Mathematically it rests on the relaxation of variational problems. Our joint paper [3] gives the underlying theory, which leads to a systematic procedure for computing the

theory, which leads to a systematic procedure for computing the relaxed problem--in which homogenization is successful and the minimum weight is attained. The design problem also extends to plates, where the appearance of more and more stiffeners in the numerical solution of a 0-1 problem led Olhoff and Cheng [3] to discover the approach to a composite. The plate equation is of higher order than our electrical conduction problem, but we anticipate that the strip construction still leads to an optimal design--and that the theory of homogenization (or relaxation, or polyconvexity) will yield a proof that this construction is optimal.

Strips of Strips

We go back to the Red Cross pattern, in order to improve on it. The improvement comes by making it easier for vertical current to use the horizontal strip. As it stands, the current has to make a long excursion; the vertical current away from the main vertical strip is exponentially small. We divide that strip into N thinner vertical strips, equally spaced across the square (Fig. 1b). As $N \to \infty$, that part becomes a composite--still with infinite resistance in the horizontal direction. When the density of vertical strips is E, and the height of those strips is $1-C$ (as before), the vertical resistance of the composite is $(1-C)/E$. Since the composite is in series with a conducting strip of resistance C (to vertical flow) the effective properties are:

$$\text{vertical resistance} \quad C + \frac{1-C}{E}$$

$$\text{horizontal resistance} \quad \frac{1}{C}$$

$$\text{conducting area} \quad A = C + E(1-C) \ .$$

The desired value of vertical resistance is $1/D$, to produce current D with unit voltage drop. Therefore

$$C + \frac{1-C}{E} = \frac{1}{D}, \text{ or } E = \frac{D-CD}{1-CD} .$$

The total conducting area is

$$A = C + \frac{D-CD}{1-CD}(1-C) = \frac{C+D-2CD}{1-CD} .$$

This is the optimal value established (but differently expressed) by Tartar and Murat.

For small currents the area is close to $C + D$; the economy from overlapping use is small. However the first correction term is $-2CD$, better than $C + D - CD$ from the cross pattern. For large currents the improvement increases. In the example that previously filled $5/9$ of the square, we now have:

$$C = \frac{1}{3} \text{ and } D = \frac{1}{3} \text{ require only the area } A = \frac{1}{2} .$$

In that case the density of vertical strips is $E = 1/4$.

Note that we have not filled the square with a single homogenized composite. That is easy to do, keeping the properties optimal. If we alternate rapidly between M horizontal strips of conductor and composite (Fig. 2), then the properties are not changed. As $M \to \infty$ this produces a homogeneous material formed from "strips of strips." This would be the local construction in each small square of a larger design, in which the conductances and area fraction A may vary throughout the region. That global problem has a straightforward variational statement, after the local solution

has produced the "relaxed" integrand. In other words, the present construction produces a family of optimal composites to be called on for a globally optimal design [3].

In a manufactured design, M and N are finite. However the design will not approach optimality if M = N. That choice would homogenize the simple Red Cross pattern, without improving it. The composite of vertical strips ($N \to \infty$) need not be completed before M increases--we can allow $N = M^2$--but it must proceed more quickly. Of course the vertical and horizontal directions could be reversed, to give a different design that is equally optimal (and elliptical inclusions are a third possibility). It is not known how to describe all composites that achieve the optimal bounds. Only the bounds themselves are known, and we come now to the proof that A above is minimal.

The Variational Problem

Suppose S is the open unit square, partly insulated and partly conducting. When a unit voltage is applied between $x = 0$ and $x = 1$, current flows. It is described by a vector whose divergence is zero (there are no sources inside the square). Therefore the vector has the form $(\partial u/\partial y, -\partial u/\partial x)$ for some stream function $u(x,y)$. For any u this vector has divergence $\partial^2 u/\partial x\partial y - \partial^2 u/\partial y\partial x = 0$. It gives the magnitude $|\nabla u|$ and also the direction of the current at each point. In an insulated region, the magnitude is $|\nabla u| = 0$ and the stream function is constant. At the boundary of such a region the normal derivative from both sides is $\partial u/\partial n = 0$. At the lower boundary of the square we impose $u = 0$, and at the upper boundary $u = C$ --in order that a current C shall flow from left to right. (The increase $u(Q) - u(P)$ in the stream

function gives the flow across a path from P to Q.) Since the conducting material has unit specific resistance, the heat loss--which is I^2R in a single resistor--is $\iint |\nabla u|^2 dxdy$. That equals current times voltage, or C times 1. This current is to be achieved in the smallest possible conducting area. That area is identified by the condition $\nabla u \neq 0$ --current is flowing--and the problem becomes:

Minimize the area in which $\nabla u \neq 0$, subject to

$$\iint_S |\nabla u|^2 \, dxdy = C, \quad u(x,0) = 0, \quad u(x,1) = C \; .$$

This one-dimensional problem is solved by a horizontal conducting strip of height C. The stream function can be $u = y$ for $y \leq C$, $u = C$ for $y \geq C$. Then $|\nabla u| = 1$ in the strip and $\nabla u = 0$ elsewhere. The constraints are met, and the strip area C is minimal.

Note: The constraint $\iint |\nabla u|^2 dxdy = C$ has used the fact that the actual current minimizes this integral, and therefore satisfies Laplace's equation in the conducting area. The physical argument based on heat loss can be replaced by Green's identity

$$\iint |\nabla u|^2 \, dxdy = \iint u(-u_{xx}-u_{yy}) + \int u \frac{\partial u}{\partial n} \, ds \; .$$

On the right side the only nonzero term is the integral of u $\partial u/\partial n$ along the top of the square, where $u = C$ and $\int \partial u/\partial n \, ds$ = voltage drop = 1. Therefore $\iint |\nabla u|^2 \, dxdy = C$.

It is important to see that the problem above, while not difficult, is also not convex. The minimization of area is

actually the minimization of $\iint 1_{\{\nabla u \neq 0\}} dxdy$, where the symbol 1_K represents a characteristic function--the function which equals one in the set K where $\nabla u \neq 0$, and zero outside. It is the nonconvex 0-1 function illustrated by Fig. 3a. The value zero looks isolated, but if $\nabla u = 0$ in a large set, then the integral is small. The goal is to achieve $\nabla u = 0$ as often as possible, and the constraint is introduced through a Lagrange multiplier λ: the functional becomes

$$L(u,\lambda) = \iint [1_{\{\nabla u \neq 0\}} + \lambda |\nabla u|^2] \, dxdy - \lambda C.$$

It is this integrand $F = 1 + \lambda |\nabla u|^2$, with the isolated value $F(0) = 0$, which is illustrated in Fig. 3b. It needs to be relaxed.

For a problem in which the unknown is a scalar, the relaxation of F is the same as its convexification. We may replace F by the largest convex function that satisfies $F_c \leq F$, without changing the minimum value of the integral. (The minimizing function u^* may be changed radically. For the original L it may not have existed.) In this problem F_c grows linearly with $|\nabla u|$, up to the point where $\lambda |\nabla u|^2 = 1$ and F_c is tangent to F. Prior to that point the convexified functional is

$$L_c(u,\lambda) = \iint F_c \, dxdy - \lambda C = \iint 2\lambda^{1/2} |\nabla u| \, dxdy - \lambda C \, .$$

The minimizing u^*, which must go from zero at $y = 0$ to C at $y = 1$, can be taken linear: $u^* = yC$. Then $|\nabla u^*| = C$ and the functional is

$$L_c(u^*,\lambda) = 2\lambda^{1/2}C - \lambda C \ .$$

The maximum over λ occurs at $\lambda^* = 1$, and yields the minimum area subject to the constraint: optimal area = C. We note that $\lambda^*|\nabla u^*|^2 = C^2 < 1$, so the minimum does occur in the range where F_c is strictly below F. This is the "homogenized" regime, oscillating between insulator and conductor--between $F = 0$ and $F = 1 + \lambda|\nabla u|^2$--in which the average of F is F_c.

One further comment on this easy problem, the design of a one-way conductor. It was made to look difficult by relaxation! A simpler optimizer is the one proposed at the start, with stream function $u = y$ for $y \le C$ and $u = C$ elsewhere. That choice leads to $\lambda|\nabla u|^2 = 1$ in one strip and $\nabla u = 0$ in the complementary strip, and no relaxation occurs. Each region is fully conducting or fully insulated; the 0-1 problem attains the same minimum as the homogenized problem. In fact the homogenized solution is the one suggested by Fig. 2, in which the horizontal conductor is split into M strips, with $M \to \infty$. The result is a composite conductor (horizontal only; the vertical part of Fig. 3 is not present) through which the current is uniform. That corresponds to our relaxed solution $u^* = yC$ over the whole square.

Thus the one-way problem illustrates relaxation in a case where it is not needed. The minimum area C is also attained in the unrelaxed problem. However our proof that this is the minimum used convexification: for $\lambda = 1$ and any admissible u,

$$\text{area} = \iint [1_{\{\nabla u \neq 0\}} + |\nabla u|^2]dxdy - C \ge \iint 2|\nabla u|dxdy - C \ge C \ .$$

In the two-way problem relaxation is absolutely needed--the original has no solution--but a simplex convexification is no longer correct.

The Two-Way Problem

The variational statement involves two stream functions $u(x,y)$ and $v(x,y)$. The unknown is now a vector. Its first component is constrained by $u(x,0) = 0$ and $u(x,1) = C$ and $\int\int |\nabla u|^2 \, dxdy = C$, as before. The second component v reflects the vertical current D, which is required to flow when a unit voltage drop is imposed between the bottom and top of the square. In the region where both currents are zero, $\nabla u = 0$ and $\nabla v = 0$, conducting material is not needed. The conductor occupies the set

$$K = \{\nabla u \neq 0\} \cup \{\nabla v \neq 0\} ,$$

whose area it is our goal to minimize. The problem becomes:

$$\underline{\text{Minimize area}}\ (K) = \int\int 1_K \, dxdy \text{ subject to } \int\int |\nabla u|^2 dxdy \leq C ,$$
$$\int\int |\nabla u|^2 dxdy \leq D,\ u(x,0) = 0,\ u(x,1) = C, v(0,y) = 0,\ v(1,y) = D.$$

The strip design proposed earlier has area $A = (C+D-2CD)/(1-CD)$. We now show that this is minimal.

The problem is again nonconvex because of the 0-1 characteristic function 1_K. Introducing the constraints by Lagrange multipliers λ and μ, the unrelaxed functional is

$$L(u,v,\lambda,\mu) = \int\int [1_K + \lambda |\nabla u|^2 + \mu |\nabla v|^2] dxdy - \lambda C - \mu D. \quad (1)$$

It could be convexified, but the minimum of L_c is too low; it is below that of L. The correct relaxation is the <u>quasiconvexification</u> L_r --the largest functional below L which is weakly lower semicontinuous in H^1. Its minimizing functions u^*, v^* will be the weak limits of minimizing (but highly oscillatory) sequences for L.

The difficulty is to compute the relaxed form L_r. That was the goal of our paper [3]. The property of quasiconvexity is difficult to verify, but in several important examples a stronger property holds and can be tested. This stronger property of the relaxation $L_r = \int\int F_r(\nabla u, \nabla v)\, dxdy$ is

polyconvexity: F_r is a convex function of ∇u and ∇u and the Jacobian determinant $J = |\nabla u \ \nabla v|$.

The Jacobian is itself nonconvex, so that a polyconvex function need not be convex. It will be the upper envelope of a family of <u>multilinear</u> functions--linear in J as well as ∇u and ∇v --in the same way that convex functions are envelopes of linear functions of ∇u and ∇v. In this problem the unrelaxed integrand is

$$F = \begin{cases} 0 & \text{if } \nabla\bar{u} = \nabla\bar{v} = 0 \\ 1+|\nabla\bar{u}|^2+|\nabla\bar{v}|^2 & \text{otherwise.} \end{cases}$$

The notation has incorporated λ and μ into $\bar{u} = \lambda^{1/2}u$ and $\bar{v} = \mu^{1/2}v$. We note that the Lagrange multipliers are nonnegative because the constraints are inequalities--the designer is happy if the conductor offers <u>less</u> resistance than specified to one or other of the currents. In the optimal

design we expect equality, $\int\int |\nabla u|^2 = C$ and $\int\int |\nabla v|^2 = D$.

The relaxation of F is known from [3]:

$$F_r = \begin{cases} 2\rho - 2|J| & \text{if } \rho \le 1 \\ 1 + |\nabla \bar{u}|^2 + |\nabla \bar{v}|^2 & \text{if } \rho \ge 1 \end{cases}$$

where $\rho = (|\nabla \bar{u}|^2 + |\nabla \bar{v}|^2 + 2|J|)^{1/2}$ and $J = |\nabla \bar{u}\ \nabla \bar{v}|$. We show below that F_r is polyconvex and $F_r \le F$. We need not show that F_r is the correct relaxation, although it is; no quasiconvex function is between F and F_r. That fact is not required for our specific (and self-contained) problem. We know that the constraints can be satisfied in a conducting area $A = (C + D - 2CD)/(1-CD)$, and our only task is to prove that the area cannot be smaller.

Provided F_r is polyconvex, the associated variational problem can be solved:

$$\text{Minimize } \int\int F_r \, dxdy \text{ subject to } \begin{matrix} \bar{u}(x,0) = 0, \ \bar{u}(x,1) = \lambda^{1/2} C \\ \bar{v}(0,y) = 0, \ \bar{v}(1,y) = \mu^{1/2} D. \end{matrix}$$

The constraints are satisfied by the linear functions

$$\bar{u} = \lambda^{1/2} Cy \quad \text{and} \quad \bar{v} = \mu^{1/2} Dx \ .$$

For those functions the Jacobian is constant, and the fundamental condition for quasiconvexity is that such a candidate--if it satisfies the boundary conditions, and is therefore admissible--is always minimizing. Therefore the minimum value of $\int\int F_r \, dxdy$, after integration of a constant over the unit square, is

$$2\rho - 2|J| = 2(\lambda C^2+\mu D^2+2\lambda^{1/2}\mu^{1/2}CD)^{1/2} - 2\lambda^{1/2}\mu^{1/2}CD$$
$$= 2(\lambda^{1/2}C+\mu^{1/2}D-\lambda^{1/2}\mu^{1/2}CD) \ .$$

Remembering the final terms $-\lambda C - \mu D$ in the Lagrangian (1), we are left with a maximization over λ and μ:

$$A' = \max_{\lambda,\mu} \min_{u,v} L_r$$
$$= \max_{\lambda,\mu} 2(\lambda^{1/2}C+\mu^{1/2}D - \lambda^{1/2}\mu^{1/2}CD) - \lambda C - \mu D. \quad (2)$$

Differentiating with respect to λ and μ, the Lagrange multipliers are

$$\lambda^{1/2} = \frac{1-D}{1-CD} \quad \text{and} \quad \mu^{1/2} = \frac{1-C}{1-CD} \ .$$

Then substituting into (2) yields

$$A' = \frac{C+D-2CD}{1-CD} \text{ (which coincides with } A) \ . \quad (3)$$

This calculation assumed that the minimum occurs when $\rho \leq 1$. That is easily verified. In fact ρ turns out to equal the density of conducting material--and in the end $\rho = A$, because the density has this constant value over a unit square.

To repeat the main line of the argument: The area of the design cannot go below A', because

i) $F_r \leq F$ and thus $L_r \leq L$ for each nonnegative λ and μ

ii) F_r is polyconvex, so that the associated functional attains its minimum

iii) that constrained minimum is A', which coincides

with the area A approached by the strip construction. The minimum of the relaxed problem is attained by linear stream functions u^* and v^*, corresponding to uniform flow through the square--which in the relaxed problem is covered by a homogeneous composite. In fact F_r was computed in [3] precisely by applying the strip construction. Our observation here is that we need only its polyconvexity, in order to solve the variational problem in this paper--and that this problem is a restatement of the optimal bound problem. Thus the argument depends on establishing that something is lower semicontinuous!--which Tartar and Murat did in another way.

The proof of polyconvexity could display the multilinear functions whose envelope is F_r, but the result comes more neatly as follows. Start with the convex function

$$c(t) = \begin{cases} 2t & 0 \le t \le 1 \\ 1+t^2 & t \ge 1 \end{cases} .$$

Then consider the two functions

$$F_{\pm}(\nabla u, \nabla v, J) = c([\,|\nabla u|^2 + |\nabla u|^2 \pm 2 \det [\nabla u \; \nabla v]]^{1/2}) \mp 2J .$$

For either sign, the quantity Q in brackets is a nonnegative quadratic form in $\nabla u, \nabla v$, and therefore a sum of squares. Its square root t is a convex function, and $c(t)$ is convex and increasing. Therefore the composition $c(t(\nabla u, \nabla v))$ is convex. The linear terms $\mp 2J$ leave $F_{\pm}$ convex, as functions with an extra argument. Then because F_r is the maximum of the two functions $F_{\pm}$, when J is identified with $\det[\nabla u \; \nabla v]$, F_r must be polyconvex.

For that last step, note that $c = 1 + t^2$ for large t

and the functions $F_{\pm}$ become $1 + |vu|^2 + |vv|^2$. For small t the comparison between F_+ and F_- rests on the inequality

$$2(r+s)^{1/2} - s \geq 2(r-s)^{1/2} + s \ .$$

This holds for $r \geq s \geq 0$, $r + s \leq 1$. In our case $r = |vu|^2 + |vv|^2$ and $s = 2|J|$. Thus the maximizing choice of sign is the one for which $\pm \det [vu\ vv]$ equals the absolute value $|J|$. With that choice the argument t coincides with ρ in the definition of F_r, and $\max F_{\pm}$ coincides with F_r.

Finally F_r is below F because $2t$ is below $1 + t^2$. The difference $(1-t)^2 = (1-\rho)^2$ in the range $0 < \rho < 1$ is the saving in area achieved by homogenization.

Acknowledgement. This research was supported by the National Science Foundation (84-03222: G.S.) and (83-12229: R.K.), by the Army Research Office (DAAL03-86-K0171: G.S.), and by ONR Grant N0014-83-K-0536 (R.K.). This paper appears in Nonsmooth Mechanics, edited by P. D. Panagiotopoulos, copyright Birkhäuser Basel 1987.

REFERENCES

1. Z. Hashin and S. Shtrikman, A variational approach to the theory of the effective magnetic permeability of multiphase materials, J. Appl. Phys. 33 (1962) 3125-3131.

2. L. Tartar, Estimations fines de coefficients homogénéisés, Colloque en l'honneur d'E. De Giorgi, P. Kree, ed., Pitman (1986); F. Murat and L. Tartar, Calcul des variations et homogénéisation, in Les Methodes de l'Homogénéisation: Theorie et Applications en Physique, Eyrolles, Paris, 1985.

3. R. Kohn and G. Strang, Optimal design and relaxation of variational problems, Comm. Pure Appl. Math. 39 (1986) 113-137,139-182,353-377.

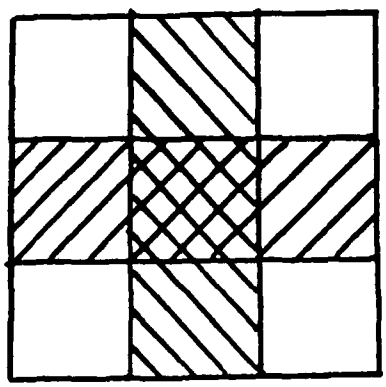
4. K. A. Lurie and A. V. Cherkaev, Exact estimates of conductivity of composites, Proc. Roy. Soc. Edinburgh 99A (1984) 71-87; G-closure of a set of anisotropically conducting media in the two dimensional case, J. Opt. Th. Appl. 42 (1984) 283-304.

5. R. Kohn and G. Milton, On bounding the effective conductivity of anisotropic composites, Homogenization and Effective Moduli of Materials and Media, J. Ericksen, D. Kinderlehrer, R. Kohn, J. L. Lions, eds., Lecture Notes, Springer, 1986.

6. G. I. N. Rozvany, Optimal Design of Flexural Systems, Pergamon, 1976.

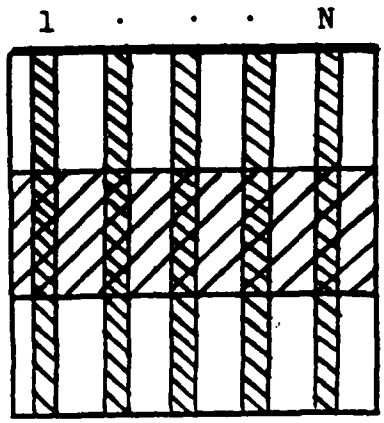
7. G. I. N. Rozvany, Structural layout theory--the present state of knowledge, Optimum Structural Design, E. Atrek, R. H. Gallagher, K. M. Ragsdell, O. C. Zienkiewicz, eds., Wiley, 1984.

8. K. T. Cheng and N. Olhoff, An investigation concerning optimal design of solid elastic plates, Int. J. Solids Structures 16 (1981) 305-323; et seq. 18 (1982) 153-170.



strip width C

strip width D



total width E

Fig. 1. (a) Red Cross design (b) Approach to a composite

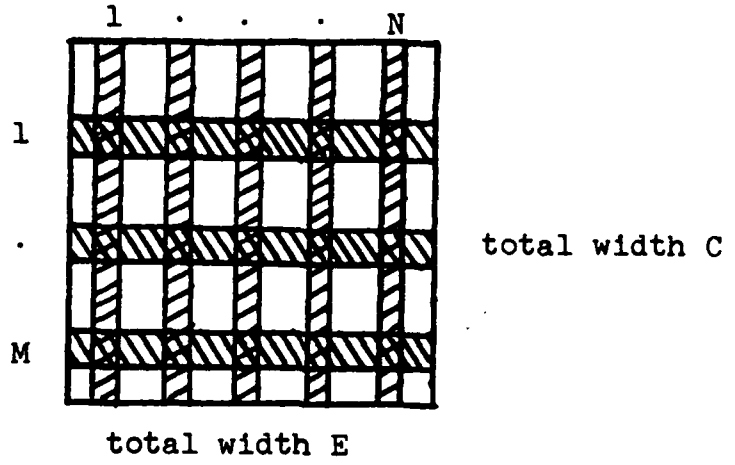


Fig. 2. Approach to a fully homogenized composite: N >> M

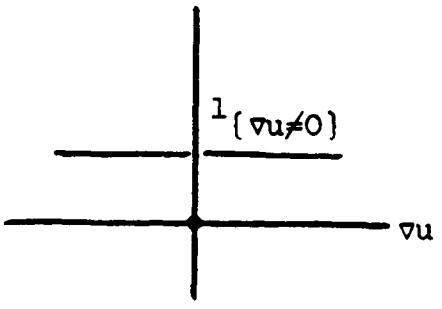
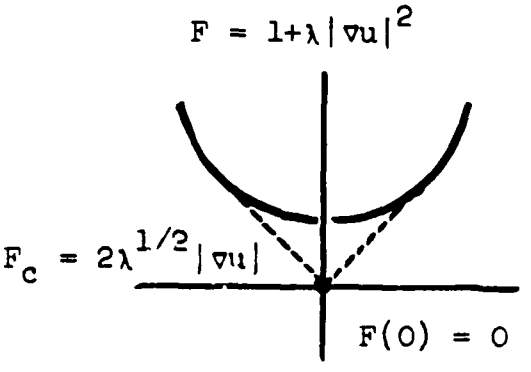


Fig. 3.(a) Characteristic function



(b) Convexified function

ON A REFINED NONLINEAR THEORY OF LAMINATED COMPOSITE PLATES

J. N. Reddy*
Department of Engineering Science and Mechanics
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

Abstract. This paper summarizes the results of research on the development of a refined shear deformation theory of plates and its analytical solutions in the linear case. The detailed results are reported in two technical papers, which will appear elsewhere. A third-order, nonlinear shear deformation plate theory that accounts for parabolic distribution of transverse shear stresses through thickness and moderate rotation terms is presented. The Levy type analytical solutions are developed for the linear case.

I. INTRODUCTION. The advent of new composite materials and their increasing use in various fields of advanced technology has generated a new interest in the development and solution of consistent refined theories of anisotropic composite plates and shells. This interest is due to the fact that the classical plate theory, in terms of its basic assumptions (i.e. the Kirchhoff hypothesis), comes in conflict with real behavior of these new materials. For example, in contrast to the basic assumption of infinite rigidity in transverse shear in the classical plate theory, the new composite materials exhibit a finite rigidity in transverse shear. This property requires the incorporation of transverse shear deformation effects.

In addition to other shortcomings, the classical plate theory involves a contradiction between the number of boundary conditions physically required to be fulfilled on a free boundary and the number available in theory, which is to be consistent with the order of the associated governing equations (see Stoker [1]). The non-fulfillment of boundary conditions on the bounding surfaces constitutes another feature of the classical theory. In recent years attempts were made to refine the classical theory by: (i) incorporating transverse shear effects, (ii) removing the contradiction which concerns the number of boundary

* Clifton C. Garvin Professor

conditions to be prescribed at each edge, and (iii) fulfilling the boundary conditions on the bounding surfaces and, in the case of laminated composite plates and shells, of the continuity conditions at the interfaces between the contiguous layers. In addition, the refined transverse shear deformation theories can be used to model such anisotropic plates and shells whose material exhibits high degree of anisotropy, and are not restricted to the thinness requirement implied by the classical laminate theory. Another feature of refined laminate theories concerns the adequate incorporation of the dynamical effects allowing the evaluation of the lowest and higher natural frequencies.

The shear deformation theories known in the literature can be grouped into two classes: (1) stress-based theories, and (2) displacement-based theories. The first stress-based transverse shear deformable plate theory is due to Reissner [2-4]. The distribution across the thickness of the transverse normal and shear stresses is determined through integration over the thickness of the equilibrium equations of the 3-D elasticity theory. The associated field equations and boundary conditions expressed in terms of 2-D quantities can be determined by using the variational principles of the 3-D elasticity theory, or by considering the moments of n^{th} order of the basic equations of 3-D elasticity theory. Both methods allow the reduction of the 3-D problems to a 2-D equivalent one.

The pioneering work of the displacement-based theories is due to Basset [5]. Based on Basset's representation of displacement field, Hildebrand, Reissner and Thomas [6] developed a variationally consistent first order theory for shells. The field equations were derived using the principle of minimum total potential energy. By using the displacement representation of Basset, Mindlin [7] extended Hencky's theory [8] of isotropic plates to the dynamic case. The shear deformation theory of Hencky-Mindlin is referred as the first-order transverse shear deformation theory. Recently, Reddy [9-11] developed a variationally consistent third-order shear deformation theory that accounts for parabolic distribution of transverse shear stresses through thickness and the von Kármán strains.

In geometrically nonlinear theories of elastic anisotropic plates one often assumes that the strains and rotations about the normal to the midplane are infinitesimal, and retains the products and squares of the derivatives of the transverse deflection in the strain-displacement equations (the von Kármán assumption). The von Kármán nonlinear theory does not account for moderate rotation terms that could be of significance in the analysis (especially in stability problems) of plates while accounting for the transverse normal and shear strains. The small strain and moderate rotation concept was used in the classical theory of plates and shells by Sanders [12], Koiter [13], Reissner [14] and Pietraszkiewicz [15], and in first-order plate and shell theories by Naghdi and Vangsarnpigoon [16], and Librescu and Schmidt [17].

In the present study, the third shear deformation with moderate rotation terms (see Reddy [18]) is reviewed, and analytical solutions of the lineary theory (see Khdeir, Reddy and Librescu [19]) are discussed.

II. A THIRD-ORDER THEORY. Consider a laminated plate composed of N orthotropic layers, symmetrically located with respect to the midplane of the laminate. The governing equations of the refined theory are based on the following displacement field [9-11]:

$$u_1 = u + z[\psi_x - \frac{4}{3}(\frac{z}{h})^2(\psi_x + \frac{\partial w}{\partial x})]$$

$$u_2 = v + z[\psi_y - \frac{4}{3}(\frac{z}{h})^2(\psi_y + \frac{\partial w}{\partial y})]$$

$$u_3 = w, \tag{1}$$

where (u_1,u_2,u_3) are the displacements along the x, y and z coordinates respectively, (u,v,w) are the corresponding displacements of a point on the midplane of the laminate, and ψ_x and ψ_y are the rotations of a transverse normal about the y- and x-axes, respectively.

The cubic variation of u_1 and u_2 through laminate thickness introduces higher-order resultants

$$P_i = \int_{-\frac{h}{2}}^{\frac{h}{2}} \sigma_i z^3 dz \qquad (i = 1,2,6)$$

$$(R_1,R_2) = \int_{-\frac{h}{2}}^{\frac{h}{2}} z^2(\sigma_5,\sigma_4)dz,$$

and laminate stiffnesses

$$(F_{ij},H_{ij}) = \int_{-\frac{h}{2}}^{\frac{h}{2}} Q_{ij}(z^4,z^6)dz \qquad (i,j = 1,2,6)$$

$$(D_{ij},F_{ij}) = \int_{-\frac{h}{2}}^{\frac{h}{2}} Q_{ij}(z^2,z^4)dz \qquad (i,j = 4,5).$$

For symmetrical cross-ply laminated plates, the following stiffness coefficients vanish [9]:

$$B_{ij} = E_{ij} = 0 \text{ for } i,j = 1,2,4,5,6$$

$$A_{16} = A_{26} = D_{16} = D_{26} = F_{16} = F_{26} = H_{16} = H_{26} = 0$$

$$A_{45} = D_{45} = F_{45} = 0.$$

This implies that the effect of coupling between stretching and bending vanishes. For such laminates the governing equations are given by (see [11]):

$$\frac{4}{3h^2}\left[F_{11}\frac{\partial^3\psi_x}{\partial x^3} + H_{11}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^3\psi_x}{\partial x^3} + \frac{\partial^4 w}{\partial x^4}\right) + F_{12}\frac{\partial^3\psi_y}{\partial x^2\partial y} + H_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^3\psi_y}{\partial x^2\partial y}\right.\right.$$

$$\left. + \frac{\partial^4 w}{\partial x^2\partial y^2}\right) + F_{12}\frac{\partial^3\psi_x}{\partial y^2\partial x} + H_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^3\psi_x}{\partial y^2\partial x} + \frac{\partial^4 w}{\partial x^2\partial y^2}\right) + F_{22}\frac{\partial^3\psi_y}{\partial y^3}$$

$$+ H_{22}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^3\psi_y}{\partial y^3} + \frac{\partial^4 w}{\partial y^4}\right) + 2F_{66}\left(\frac{\partial^3\psi_y}{\partial x^2\partial y} + \frac{\partial^3\psi_x}{\partial y^2\partial x}\right) + 2H_{66}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^3\psi_x}{\partial y^2\partial x}\right.$$

$$\left.\left. + \frac{\partial^3\psi_y}{\partial x^2\partial y} + \frac{2\partial^4 w}{\partial x^2\partial y^2}\right)\right] - \frac{4}{h^2}\left[D_{55}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial\psi_x}{\partial x}\right) + F_{55}\left(-\frac{4}{h^2}\right)\left(\frac{\partial\psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}\right)\right.$$

$$\left. + D_{44}\left(\frac{\partial^2 w}{\partial y^2} + \frac{\partial\psi_y}{\partial y}\right) + F_{44}\left(-\frac{4}{h^2}\right)\left(\frac{\partial\psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2}\right)\right] + \left[A_{55}\left(\frac{\partial\psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}\right)\right.$$

$$+ D_{55}\left(-\frac{4}{h^2}\right)\left(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}\right) + A_{44}\left(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2}\right) + D_{44}\left(-\frac{4}{h^2}\right)\left(\frac{\partial \psi_y}{\partial y}\right.$$

$$\left. + \frac{\partial^2 w}{\partial y^2}\right)] + q = 0, \tag{2a}$$

$$D_{11}\frac{\partial^2 \psi_x}{\partial x^2} + D_{12}\frac{\partial^2 \psi_y}{\partial x \partial y} + F_{11}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x^2} + \frac{\partial^3 w}{\partial x^3}\right) + F_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_y}{\partial x \partial y}\right.$$

$$\left. + \frac{\partial^3 w}{\partial x \partial y^2}\right) + D_{66}\left(\frac{\partial^2 \psi_x}{\partial y^2} + \frac{\partial^2 \psi_y}{\partial x \partial y}\right) + F_{66}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial y^2} + \frac{\partial^2 \psi_y}{\partial x \partial y} + \frac{2\partial^3 w}{\partial x \partial y^2}\right)$$

$$- [A_{55}\left(\psi_x + \frac{\partial w}{\partial x}\right) + D_{55}\left(-\frac{4}{h^2}\right)\left(\psi_x + \frac{\partial w}{\partial x}\right)] - \frac{4}{3h^2}[F_{11}\frac{\partial^2 \psi_x}{\partial x^2}$$

$$+ H_{11}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x^2} + \frac{\partial^3 w}{\partial x^3}\right) + F_{12}\frac{\partial^2 \psi_y}{\partial x \partial y} + H_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_y}{\partial x \partial y} + \frac{\partial^3 w}{\partial x \partial y^2}\right)$$

$$+ F_{66}\left(\frac{\partial^2 \psi_y}{\partial x \partial y} + \frac{\partial^2 \psi_x}{\partial y^2}\right) + H_{66}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial y^2} + \frac{\partial^2 \psi_y}{\partial x \partial y} + \frac{2\partial^3 w}{\partial x \partial y^2}\right)]$$

$$+ \frac{4}{h^2}[D_{55}\left(\frac{\partial w}{\partial x} + \psi_x\right) + F_{55}\left(-\frac{4}{h^2}\right)\left(\psi_x + \frac{\partial w}{\partial x}\right)] = 0, \tag{2b}$$

$$D_{66}\left(\frac{\partial^2 \psi_x}{\partial x \partial y} + \frac{\partial^2 \psi_y}{\partial x^2}\right) + F_{66}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x \partial y} + \frac{\partial^2 \psi_y}{\partial x^2} + 2\frac{\partial^3 w}{\partial x^2 \partial y}\right) + D_{12}\frac{\partial^2 \psi_x}{\partial x \partial y}$$

$$+ D_{22}\frac{\partial^2 \psi_y}{\partial y^2} + F_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x \partial y} + \frac{\partial^3 w}{\partial x^2 \partial y}\right) + F_{22}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_y}{\partial y^2} + \frac{\partial^3 w}{\partial y^3}\right)$$

$$- [A_{44}\left(\psi_y + \frac{\partial w}{\partial y}\right) + D_{44}\left(-\frac{4}{h^2}\right)\left(\psi_y + \frac{\partial w}{\partial y}\right)] - \frac{4}{3h^2}[F_{66}\left(\frac{\partial^2 \psi_y}{\partial x^2} + \frac{\partial^2 \psi_x}{\partial y \partial x}\right)$$

$$+ H_{66}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x \partial y} + \frac{\partial^2 \psi_y}{\partial x^2} + \frac{2\partial^3 w}{\partial x^2 \partial y}\right) + F_{12}\frac{\partial^2 \psi_x}{\partial x \partial y} + H_{12}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_x}{\partial x \partial y}\right.$$

$$\left. + \frac{\partial^3 w}{\partial x^2 \partial y}\right) + F_{22}\frac{\partial^2 \psi_y}{\partial y^2} + H_{22}\left(-\frac{4}{3h^2}\right)\left(\frac{\partial^2 \psi_y}{\partial y^2} + \frac{\partial^3 w}{\partial y^3}\right)] + \frac{4}{h^2}[D_{44}\left(\frac{\partial w}{\partial y}\right.$$

$$\left. + \psi_y\right) + F_{44}\left(-\frac{4}{h^2}\right)\left(\frac{\partial w}{\partial y} + \psi_y\right)] = 0. \tag{2c}$$

Here q denotes distributed transverse load, and A_{ij}, D_{ij}, F_{ij}, H_{ij} are the plate stiffnesses defined by

$$(D_{ij},F_{ij},H_{ij}) = \int_{-h/2}^{h/2} Q_{ij}^{(k)}(z^2,z^4,z^6)dz \quad (i,j = 1,2,6)$$

$$(A_{ij},D_{ij},F_{ij}) = \int_{-h/2}^{h/2} Q_{ij}^{(k)}(1,z^2,z^4)dz \quad (i,j = 4,5) \tag{3}$$

where $Q_{ij}^{(k)}$ denote the plane-stress reduced orthotropic moduli of the k-th lamina. The boundary conditions of the refined theory are of the form

Specify:
$$\left.\begin{array}{l} w \text{ or } Q_n \\ \frac{\partial w}{\partial n} \text{ or } P_n \\ \psi_n \text{ or } M_n \\ \psi_{ns} \text{ or } M_{ns}. \end{array}\right\} \text{on } \Gamma \tag{4}$$

Here Γ denotes the boundary of the midplane Ω of the plate, and

$$M_n = \hat{M}_1 n_x^2 + \hat{M}_2 n_y^2 + 2\hat{M}_6 n_x n_y$$

$$M_{ns} = (\hat{M}_2 - \hat{M}_1)n_x n_y + \hat{M}_6(n_x^2 - n_y^2)$$

$$P_n = P_1 n_x^2 + P_2 n_y^2 + 2P_6 n_x n_y$$

$$P_{ns} = (P_2 - P_1)n_x n_y + P_6(n_x^2 - n_y^2)$$

$$Q_n = \hat{Q}_1 n_x + \hat{Q}_2 n_y + \frac{4}{3h^2}\left(\frac{\partial P_{ns}}{\partial s} + \frac{\partial P_n}{\partial n}\right)$$

$$\hat{M}_i = M_i - \frac{4}{3h^2} P_i \quad (i = 1,2,6)$$

$$\hat{Q}_i = Q_i - \frac{4}{h^2} R_i \quad (i = 1,2)$$

$$\frac{\partial}{\partial n} = n_x \frac{\partial}{\partial x} + n_y \frac{\partial}{\partial y} \quad , \quad \frac{\partial}{\partial s} = n_x \frac{\partial}{\partial y} - n_y \frac{\partial}{\partial x}. \tag{5}$$

The stress resultants appearing in Eq. (5) can be expressed in terms of the generalized displacements (w,ψ_x,ψ_y) as:

$$M_1 = D_{11} \frac{\partial \psi_x}{\partial x} + D_{12} \frac{\partial \psi_y}{\partial y} + F_{11}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + F_{12}(-\frac{4}{3h^2})(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})$$

$$M_2 = D_{12} \frac{\partial \psi_x}{\partial x} + D_{22} \frac{\partial \psi_y}{\partial y} + F_{12}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + F_{22}(-\frac{4}{3h^2})(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})$$

$$M_6 = D_{66}(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x}) + F_{66}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} + 2\frac{\partial^2 w}{\partial x \partial y})$$

$$Q_2 = A_{44}(\psi_y + \frac{\partial w}{\partial y}) + D_{44}(-\frac{4}{h^2})(\psi_y + \frac{\partial w}{\partial y})$$

$$Q_1 = A_{55}(\psi_x + \frac{\partial w}{\partial x}) + D_{55}(-\frac{4}{h^2})(\psi_x + \frac{\partial w}{\partial x})$$

$$P_1 = F_{11} \frac{\partial \psi_x}{\partial x} + F_{12} \frac{\partial \psi_y}{\partial y} + H_{11}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + H_{12}(-\frac{4}{3h^2})(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})$$

$$P_2 = F_{12} \frac{\partial \psi_x}{\partial x} + F_{22} \frac{\partial \psi_y}{\partial y} + H_{12}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + H_{22}(-\frac{4}{3h^2})(\frac{\partial \psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})$$

$$P_6 = F_{66}(\frac{\partial \psi_y}{\partial x} + \frac{\partial \psi_x}{\partial y}) + H_{66}(-\frac{4}{3h^2})(\frac{\partial \psi_x}{\partial y} + \frac{\partial \psi_y}{\partial x} + 2\frac{\partial^2 w}{\partial x \partial y})$$

$$R_2 = D_{44}(\frac{\partial w}{\partial y} + \psi_y) + F_{44}(-\frac{4}{3h^2})(\frac{\partial w}{\partial y} + \psi_y)$$

$$R_1 = D_{55}(\frac{\partial w}{\partial x} + \psi_x) + F_{55}(-\frac{4}{3h^2})(\frac{\partial w}{\partial x} + \psi_x). \tag{6}$$

The Lévy method can be used to solve Eqs. (2) for rectangular plates for which two opposite edges are simply supported. The other two edges can each have arbitrary boundary conditions. Here we assume that the edges parallel to the y-axis are simply supported, and the origin of the coordinate system is taken as shown in Fig. 1. The simply supported

boundary conditions can be satisfied by trigonometric functions in x. The resulting ordinary differential equations in y can be solved using the state-space concept (see [20]).

Following the Lévy type procedure, we assume the following representation of the displacements and loading:

$$w(x,y) = \sum_{m=1}^{\infty} W_m(y)\sin\alpha x$$

$$\psi_x(x,y) = \sum_{m=1}^{\infty} X_m(y)\cos\alpha x$$

$$\psi_y(x,y) = \sum_{m=1}^{\infty} Y_m(y)\sin\alpha x$$

$$q(x,y) = \sum_{m=1}^{\infty} Q_m(y)\sin\alpha x, \qquad (7)$$

where $\alpha = \frac{m\pi}{a}$ and W_m, X_m, Y_m and Q_m denote amplitudes of w, ψ_x, ψ_y and q, respectively. Substituting Eqs. (7) into Eqs. (2), we obtain

$$e_1 W_m'''' + e_2 W_m'' + e_3 W_m + e_4 X_m'' + e_5 X_m + e_6 Y_m''' + e_7 Y_m' + Q_m = 0$$

$$e_8 W_m'' + e_9 W_m + e_{10} X_m'' + e_{11} X_m + e_{12} Y_m' = 0$$

$$e_{13} W_m''' + e_{14} W_m' + e_{15} X_m' + e_{16} Y_m'' + e_{17} Y_m = 0, \qquad (8)$$

where primes on the variables indicate differentiation with respect to y, and

$$e_1 = -\left(\frac{4}{3h^2}\right)^2 H_{22}$$

$$e_2 = 2\left(\frac{4}{3h^2}\right)^2 \alpha^2 (H_{12} + 2H_{66}) + A_{44} - \frac{8}{h^2} D_{44} + \left(\frac{4}{h^2}\right)^2 F_{44}$$

$$e_3 = -\alpha^2\left[\left(\frac{4}{3h^2}\right)^2 \alpha^2 H_{11} + \left(\frac{4}{h^2}\right)^2 F_{55} - \frac{8}{h^2} D_{55} + A_{55}\right]$$

$$e_4 = \alpha \frac{4}{3h^2}\left[-F_{12} + \frac{4}{3h^2} H_{12} - 2F_{66} + \frac{8}{3h^2} H_{66}\right]$$

$$e_5 = \alpha^3 \frac{4}{3h^2}\left(F_{11} - \frac{4}{3h^2} H_{11}\right) + \alpha\left[\frac{8}{h^2} D_{55} - \left(\frac{4}{h^2}\right)^2 F_{55} - A_{55}\right]$$

$$e_6 = \frac{4}{3h^2}\left(F_{22} - \frac{4}{3h^2} H_{22}\right)$$

$$e_7 = \alpha^2 \frac{4}{3h^2}\left[-F_{12} - 2F_{66} + \frac{4}{3h^2}\left(H_{12} + 2H_{66}\right)\right] - \frac{8}{h^2} D_{44} + \left(\frac{4}{h^2}\right)^2 F_{44} + A_{44}$$

$$e_8 = e_4 , \quad e_9 = e_5$$

$$e_{10} = D_{66} - \frac{8}{3h^2} F_{66} + \left(\frac{4}{3h^2}\right)^2 H_{66}$$

$$e_{11} = \alpha^2\left[-D_{11} + \frac{8}{3h^2} F_{11} - \left(\frac{4}{3h^2}\right)^2 H_{11}\right] + \frac{8}{h^2} D_{55} - \left(\frac{4}{h^2}\right)^2 F_{55} - A_{55}$$

$$e_{12} = \alpha\left[D_{12} + D_{66} - \frac{8}{3h^2}\left(F_{12} + F_{66}\right) + \left(\frac{4}{3h^2}\right)^2 \left(H_{12} + H_{66}\right)\right]$$

$$e_{13} = -e_6 , \quad e_{14} = -e_7 , \quad e_{15} = -e_{12}$$

$$e_{16} = D_{22} - \frac{8}{3h^2} F_{22} + \left(\frac{4}{3h^2}\right) H_{22}$$

$$e_{17} = \alpha^2\left[-D_{66} + \frac{8}{3h^2} F_{66} - \left(\frac{4}{3h^2}\right)^2 H_{66}\right] + \frac{8}{h^2} D_{44} - \left(\frac{4}{h^2}\right)^2 F_{44} - A_{44}.$$

Equations (8) can be written as

$$\begin{aligned} W_m'''' &= c_1 W_m'' + c_2 W_m + c_3 X_m + c_4 Y_m' + c_o Q_m \\ X_m'' &= c_5 W_m'' + c_6 W_m + c_7 X_m + c_8 Y_m' \\ Y_m'' &= c_9 W_m''' + c_{10} W_m' + c_{11} X_m' + c_{12} Y_m, \end{aligned} \tag{10}$$

where

$$c_1 = \left(\frac{e_4^2}{e_{10}} + \frac{e_4 e_6 e_{12}}{e_{10} e_{16}} - \frac{e_6 e_7}{e_{16}} - e_2\right) / \left(e_1 + \frac{e_6^2}{e_{16}}\right)$$

$$c_2 = \left(\frac{e_4e_5}{e_{10}} + \frac{e_5e_6e_{12}}{e_{10}e_{16}} - e_3\right)/\left(e_1 + \frac{e_6^2}{e_{16}}\right)$$

$$c_3 = \left(\frac{e_{11}e_4}{e_{10}} + \frac{e_{11}e_6e_{12}}{e_{10}e_{16}} - e_5\right)/\left(e_1 + \frac{e_6^2}{e_{16}}\right)$$

$$c_4 = \left(\frac{e_6e_{17}}{e_{16}} + \frac{e_4e_{12}}{e_{10}} + \frac{e_6e_{12}^2}{e_{10}e_{16}} - e_7\right)/\left(e_1 + \frac{e_6^2}{e_{16}}\right)$$

$$c_o = -\frac{e_{16}}{e_1e_{16} + e_6^2}$$

$$c_5 = -e_4/e_{10},\ c_6 = -e_5/e_{10},\ c_7 = -e_{11}/e_{10},\ c_8 = -e_{12}/e_{10}$$

$$c_9 = e_6/e_{16},\ c_{10} = e_7/e_{16},\ c_{11} = e_{12}/e_{16},\ c_{12} = -e_{17}/e_{16}. \qquad (11)$$

The linear system of ordinary differential equations (10) with constant coefficients can be reduced to a single matrix differential equation using the state-space concept (see [20])

$$\underline{x}' = A\underline{x} + \underline{b}. \qquad (12)$$

This can be done by introducing the variables

$$x_1 = W_m,\ x_2 = W_m',\ x_3 = W_m'',\ x_4 = W_m'''$$

$$x_5 = X_m,\ x_6 = X_m',\ x_7 = Y_m,\ x_8 = Y_m', \qquad (13)$$

where

$$\underline{x}' = \begin{Bmatrix} x_1' \\ x_2' \\ x_3' \\ x_4' \\ x_5' \\ x_6' \\ x_7' \\ x_8' \end{Bmatrix},\ \underline{b} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ C_oQ_m \\ 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \qquad (14)$$

and

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ c_2 & 0 & c_1 & 0 & c_3 & 0 & 0 & c_4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ c_6 & 0 & c_5 & 0 & c_7 & 0 & 0 & c_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & c_{10} & 0 & c_9 & 0 & c_{11} & c_{12} & 0. \end{bmatrix} \tag{15}$$

The solution of Eq. (12) is given by

$$\underline{x} = e^{Ay}\underline{K} + e^{Ay} \int e^{-Ay}\underline{b}dy, \tag{16}$$

where $\underline{K}$ is a constant vector to be determined from the boundary conditions, e^{Ay} denotes the product,

$$e^{Ay} = [c] \begin{bmatrix} e^{\lambda_1 y} & & & \\ & e^{\lambda_2 y} & & \underset{\sim}{0} \\ & & \ddots & \\ \underset{\sim}{0} & & & e^{\lambda_8 y} \end{bmatrix} [c]^{-1}, \tag{17}$$

[c] is the matrix of distinct eigenvectors, $\lambda_i (i = 1,2,3,\ldots 8)$ are the eigenvalues associated with matrix A, and $[c]^{-1}$ is the inverse of the eigenvectors matrix [c].

The following boundary conditions are used on the remaining two edges (i.e., the edges parallel to the x-axis) at $y = \mp \frac{b}{2}$.

<u>simply supported</u>: $w = \psi_x = P_2 = M_2 = 0$

<u>clamped</u>: $w = \frac{\partial w}{\partial y} = \psi_x = \psi_y = 0$

free:

$$P_2 = M_2 = 0$$

$$M_6 - \frac{4}{3h^2} P_6 = 0$$

$$Q_2 - \frac{4}{h^2} R_2 + \frac{4}{3h^2} \left(\frac{\partial P_6}{\partial x} + \frac{\partial P_2}{\partial y}\right) = 0. \qquad (18)$$

Numerical results are presented for symmetric cross-ply (0°/90°/0°) plates subjected to uniformly distributed load (q_o), as shown in Fig. 1. The following material properties are used in the calculations:

$$E_1 = 19.2 \times 10^6 \text{ psi} \quad , \quad E_2 = 1.56 \times 10^6 \text{ psi}$$
$$G_{12} = G_{13} = 0.82 \times 10^6 \text{ psi} \quad , \quad G_{23} = 0.523 \times 10^6 \text{ psi} \qquad (19)$$
$$\nu_{12} = 0.24$$

The following notation has been used throughout the figures:

SS - simply supported at y = -b/2 and at y = b/2.

CC - clamped at y = -b/2 and at y = b/2.

FF - free at y = -b/2 and at y = b/2.

SC - simply supported at y = -b/2 and clamped at y = b/2.

SF - simply supported at y = -b/2 and free at y = b/2.

CF - clamped at y = -b/2 and free at y = b/2. (20)

The aspect ratio, a/b, is taken to be 4.

To show the effect of transverse shear strains on the deflections plots of nondimensionalized center deflection, $\bar{w} = 10^3 w(a/2,0)h^3 E_2/(q_o a^4)$, versus side to thickness ratio of various plates are presented in Figs. 2-4. The shear deformation effect is more significant in cross-ply plates than in orthotropic plates. Also, the first order shear deformation theory (FSDT) over predicts deflections relative to the higher order theory (HSDT).

Figures 5 and 6 contain plots of the transverse shear stress σ_{13} through laminate thickness for various boundary conditions. The stresses were computed using lamina constitutive relations. The transverse shear stresses are constant and parabolic through thickness

of each lamina, respectively, for the first- and higher-order theories. The discontinuity at interface of lamina is due to the mismatch of the material properties. When the stresses $(\sigma_x,\sigma_y,\sigma_{xy})$ obtained from the constitutive equations are substituted into the equilibrium equations of elasticity and integrated through thickness to determine the transverse shear stresses, the resulting stresses will be continuous through the thickness.

III. A MODERATE-ROTATION THEORY. The theory is a generalization of the classical plate theory, the first-order shear deformation plate theory, and the third-order shear deformation theories of Reddy [9-11]. The theory is based on an assumed displacement field and orders of magnitudes of linear strains and rotations.

Points of a three dimensional continuum V are denoted by their orthogonal curvilinear coordinates $\underset{\sim}{x} = (x^1,x^2,x^3)$. Covariant and contravariant base vectors at points of the continuum are denoted by $\underset{\sim}{g}_i$ and $\underset{\sim}{g}^i$, respectively. Latin indices are assumed to have values 1, 2, 3, and the Greek indices have values 1, 2. The laminated plate continuum in the undeformed configuration is defined by the Cartesian product of points in the midplane Ω and the normal $[-h/2, h/2]$:

$$V = \Omega \times [-\frac{h}{2}, \frac{h}{2}]$$

where h denotes the constant thickness of the laminate. Let x^α denote the curvilinear inplane coordinates and x^3 be the normal to Ω. The metric tensor components of Ω are denoted by

$$g_{\alpha\beta} = \underset{\sim}{g}_\alpha \cdot \underset{\sim}{g}_\beta \quad , \quad g^{\alpha\beta} = \underset{\sim}{g}^\alpha \cdot \underset{\sim}{g}^\beta \quad , \quad g^{33} = g_{33} = 1$$

$$\underset{\sim}{g}_\alpha = \frac{\partial \underset{\sim}{r}}{\partial x^\alpha} \quad , \quad \underset{\sim}{g}_\alpha \cdot \underset{\sim}{g}^\beta = \delta_\alpha^\beta \quad , \quad \underset{\sim}{g}_3 = \underset{\sim}{n}, \tag{21}$$

where $\underset{\sim}{r}$ is the position vector of a particle (x^α,x^3) at time t, δ_α^β is the Kronecker delta, and $\underset{\sim}{n}$ is the unit normal to the boundary of Ω.

The displacement vector of a point in the plate at time t is of the form

$$\underset{\sim}{u} = u_\alpha \underset{\sim}{g}^\alpha + u_3 \underset{\sim}{n} \tag{22}$$

where the Einestein summation convention on repeated subscripts is assumed. The covariant components of the Green-Lagrange strain tensor are given by

$$\varepsilon_{ij} = \frac{1}{2}\,(u_{i|j} + u_{j|i} + u_{m|i}u^{m|j}) \tag{23}$$

where a vertical line denotes covariant differentiation. The strain components ε_{ij} can be expressed in terms of the linearized strains e_{ij} and rotations ω_{ij} as

$$\varepsilon_{ij} = e_{ij} + \frac{1}{2}\,e_{mi}e^m_j + \frac{1}{2}\,(e_{mi}\omega^m_j + e_{mj}\omega^m_i) + \frac{1}{2}\,\omega_{mi}\omega^m_j, \tag{24}$$

where

$$e_{ij} = \frac{1}{2}\,(u_{i|j} + u_{j|i}) \quad , \quad \omega_{ij} = \frac{1}{2}\,(u_{i|j} - u_{j|i}). \tag{25}$$

Following [17], we now assume that the strains ε_{ij} and rotations ω_{ij} are of the following magnitude:

$$\varepsilon_{ij} = O(\theta^2) \quad , \quad \omega_{\alpha\beta} = O(\theta^2) \quad , \quad \omega_{\alpha 3} = O(\theta) \quad , \quad \theta << 1. \tag{26}$$

Equation (26) implies that the strains and the rotations about the normal to the midplane are small, and that the rotations of a normal to the midplane are moderate. Such assumptions are justified in view of the large inplane rigidity and transverse flexibility of composite laminates.

Neglecting terms of order (θ^4) and higher in the strain displacement equations (24), we obtain

$$\varepsilon_{\alpha\beta} = e_{\alpha\beta} + \underline{\frac{1}{2}\,(e_{3\alpha}\omega^3_\beta + e_{3\beta}\omega^3_\alpha)} + \frac{1}{2}\,\omega_{3\alpha}\omega^3_\beta$$

$$\varepsilon_{\alpha 3} = e_{\alpha 3} + \underline{\frac{1}{2}\,(e_{\lambda\alpha}\omega^\lambda_3 + e_{33}\omega^3_\alpha)} + \frac{1}{2}\,\underline{\omega_{\lambda\alpha}\omega^\lambda_3}$$

$$\varepsilon_{33} = e_{33} + \underline{e_{\lambda 3}\omega^\lambda_3} + \frac{1}{2}\,\omega_{\lambda 3}\omega^\lambda_3 \tag{27}$$

where the underlined terms are of order (θ^3).

The present theory is based on the following assumed variation of the displacement components across the plate thickness:

$$u_\alpha(x^\beta,x^3,t) = u^0_\alpha(x^\beta,t) - x^3 u^0_{3|\alpha} + f(x^3)u^1_\alpha(x^\beta,t)$$

$$u_3(x^\beta,x^3,t) = u^0_3(x^\beta,t) + \hat{u}^0_3(x^\beta,t), \tag{28}$$

where f is a specified function of the thickness coordinate x^3. Note that the transverse deflection is assumed to be independent of x^3 and consists of two parts, one due to bending and the other due to transverse shear. The particular form of displacement field is assumed in order to include the displacement fields of the classical plate theory (set $\hat{u}^0_3 = 0$ and $u^1_\alpha = 0$), the first-order shear deformation theory [set $u^0_3 = 0$ and $f(x^3) = x^3$], and the third-order shear deformation theory of Reddy [9] [set $u^0_3 = 0$ and $f(x^3) = x^3[1 - \frac{4}{3}(\frac{x^3}{h})^2]$], among others.

For the displacement field in Eq. (28), the strains for the moderate rotation theory become [consistent with the assumptions in Eq. (26)],

$$\varepsilon_{\alpha\beta} = \varepsilon^0_{\alpha\beta} + x^3\varepsilon^1_{\alpha\beta} + f\kappa_{\alpha\beta}$$

$$\varepsilon_{\alpha 3} = \varepsilon^0_{\alpha 3} + g\hat{\varepsilon}^0_{\alpha 3} + x^3\kappa^0_{\alpha 3} + gx^3\kappa^1_{\alpha 3} + f\varepsilon^1_{\alpha 3} + fg\tilde{\varepsilon}^1_{\alpha 3}$$

$$\varepsilon_{33} = \varepsilon^0_{33} + g\tilde{\varepsilon}^0_{33} + g^2\hat{\varepsilon}^0_{33}, \tag{29}$$

where $g = df/dx^3$, and

$$\varepsilon^0_{\alpha\beta} = \tfrac{1}{2}(u^0_{\alpha|\beta} + u^0_{\beta|\alpha}) + \tfrac{1}{2}(u^0_{3|\alpha} + \hat{u}^0_{3|\alpha})(u^0_{3|\beta} + \hat{u}^0_{3|\beta})$$

$$\varepsilon^1_{\alpha\beta} = -u^0_{3|\alpha\beta}\ ,\ \kappa_{\alpha\beta} = \tfrac{1}{2}(u^1_{\alpha|\beta} + u^1_{\beta|\alpha})\ ,\ \varepsilon^0_{\alpha 3} = \tfrac{1}{2}(\hat{u}^0_{3|\alpha} - u^0_{\beta|\alpha}u^0_{3|\beta})$$

$$\hat{\varepsilon}^0_{\alpha 3} = \tfrac{1}{2}(u^1_\alpha + u^0_{\beta|\alpha}u^1_\beta)\ ,\ \kappa^0_{\alpha 3} = \tfrac{1}{2}u^0_{3|\lambda\alpha}u^0_{3|\lambda}\ ,\ \kappa^1_{\alpha 3} = -\tfrac{1}{2}u^0_{3|\lambda\alpha}u^1_\lambda$$

$$\varepsilon^1_{\alpha 3} = -\tfrac{1}{2}u^1_{\lambda|\alpha}u^0_{3|\lambda}\ ,\ \hat{\varepsilon}^1_{\alpha 3} = \tfrac{1}{2}u^1_{\lambda|\alpha}u^1_\lambda\ ,\ \varepsilon^0_{33} = \tfrac{1}{2}u^0_{3|\alpha}u^0_{3|\alpha}$$

$$\tilde{\varepsilon}^{o}_{33} = - u^{o}_{3|\alpha} u^{1}_{\alpha} \ , \ \hat{\varepsilon}^{o}_{33} = \frac{1}{2} u^{1}_{\alpha} u^{1}_{\alpha}. \tag{30}$$

The dynamic version of the principle of virtual displacements is used to derive variationally consistent equations of motion associated with the displacement field in Eq. (28). The principle can be stated, in the absence of body forces and prescribed tractions, as

$$0 = \int_{o}^{T} [\int_{V} (\sigma^{ij} \delta \varepsilon_{ij}) dV + \int_{\Omega} q \delta u_3 dA - \int_{V} \rho (\dot{u}_i \delta \dot{u}_i) dV] dt \tag{31}$$

where σ^{ij} denote the contravariant components of the symmetric stress tensor, $q = q(x^{\alpha})$ is the distributed transverse force per unit area, and ρ is the density of the material of the plate. The superposed dot denotes the time derivative, $\dot{u} \equiv \partial u / \partial t$. We introduce the couples and inertias,

$$(N^{\alpha\beta}, M^{\alpha\beta}, P^{\alpha\beta}) = \int_{-\frac{h}{2}}^{\frac{h}{2}} \sigma^{\alpha\beta} (1, x^3, f) dx^3$$

$$(Q^{\alpha}, \hat{Q}^{\alpha}, R^{\alpha}, \hat{R}^{\alpha}, S^{\alpha}, \hat{S}^{\alpha}) = \int_{-\frac{h}{2}}^{\frac{h}{2}} \sigma^{\alpha 3} (1, g, x^3, x^3 g, f, fg) dx^3$$

$$(N^3, \tilde{N}^3, \hat{N}^3) = \int_{-\frac{h}{2}}^{\frac{h}{2}} \sigma^{33} (1, g, g^2) dx^3 \tag{32}$$

$$I_o = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho dx^3 \ , \ I_1 = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho x^3 dx^3 \ , \ I_1^f = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho f dx^3$$

$$I_2 = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho (x^3)^2 dx^3 \ , \ \hat{I}_2 = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho x^3 f dx^3 \ , \ I_2^f = \int_{-\frac{h}{2}}^{\frac{h}{2}} \rho f^2 dx^3. \tag{33}$$

The equations of motion of the theory are obtained by substituting Eq. (30) for the strains in terms of the displacements $(u^0_\alpha, u^0_3, \hat{u}^0_3, u^1_\alpha)$ into Eq. (31), integrating by parts to transfer differentiation from the displacements to the stress resultants and couples, collecting the coefficients of the various virtual displacements, and invoking the fundamental lemma of the calculus of variations. We obtain the following six equations:

$$\delta u^0_\alpha: \quad N^{\alpha\beta}|_\beta - \underline{(Q^\beta u^0_{3|\alpha})|_\beta} + \underline{(\hat{Q}^\beta u^1_\alpha)|_\beta} = I_0 \ddot{u}^0_\alpha - I_1 \ddot{u}^0_{3|\alpha} + I^f_1 \ddot{u}^1_\alpha$$

$$\delta u^0_3: \quad M^{\alpha\beta}|_{\alpha\beta} + [N^{\alpha\beta}(u^0_{3|\beta} + \hat{u}^0_{3|\beta})]|_\alpha - \underline{(Q_\alpha u^0_{\beta|\alpha})|_\beta}$$

$$- \underline{(R^\alpha|_\alpha u^0_{3|\beta})|_\beta} + \underline{(\hat{R}^\alpha u^1_\beta)|_{\beta\alpha}} - \underline{(S^\alpha u^1_{\beta|\alpha})|_\beta} + \underline{(N^3 u^0_{3|\alpha})|_\alpha}$$

$$- \underline{(\tilde{N}^3 u^1_\alpha)|_\alpha} = q + I_0(\ddot{u}^0_3 + \ddot{\hat{u}}^0_3) + I_1 \ddot{u}^0_{\alpha|\alpha} + \hat{I}_2 \ddot{u}^1_{\alpha|\alpha} + I_2 \ddot{u}^0_{3|\alpha\alpha}$$

$$\delta \hat{u}^0_3: \quad [N^{\alpha\beta}(u^0_{3|\beta} + \hat{u}^0_{3|\beta})]|_\alpha + Q^\alpha|_\alpha = q + I_0(\ddot{u}^0_3 + \ddot{\hat{u}}^0_3)$$

$$\delta u^1_\alpha: \quad P^{\alpha\beta}|_\beta - \hat{Q}^\beta(\delta_{\alpha\beta} + \underline{u^0_{\alpha|\beta}}) + \underline{\hat{R}^\beta u^0_{3|\alpha\beta}} + \underline{\hat{S}^\beta|_\beta u^1_\alpha}$$

$$- \underline{(S^\beta u^0_{3|\alpha})|_\beta} - \underline{\hat{N}^3 u^1_\alpha} + \underline{\tilde{N}^3 u^0_{3|\alpha}} = I^f_1 \ddot{u}^0_\alpha - \hat{I}_2 \ddot{u}^0_{3|\alpha} + I^f_2 \ddot{u}^1_\alpha, \quad (34)$$

where the underlined terms are entirely due to the inclusion of moderate rotations (i.e., over and above the von Kármán nonlinear terms).

Equations (34) can be specialized to the three different theories discussed earlier. The equations are summarized below:

(i) Classical Plate Theory $(\hat{u}^0_3 = 0, \quad u^1_\alpha = 0)$

$$N^{\alpha\beta}|_\beta - \underline{(Q^\beta u^0_{3|\alpha})|_\beta} = I_0 \ddot{u}^0_\alpha$$

$$M^{\alpha\beta}|_{\alpha\beta} + (N^{\alpha\beta} u^0_{3|\beta})|_\alpha - \underline{(Q^\alpha u^0_{\beta|\alpha})|_\beta} - \underline{(R^\alpha|_\alpha u^0_{3|\beta})|_\beta} + \underline{(N^3 u^0_{3|\alpha})|_\alpha}$$

$$= q + I_0 \ddot{u}^0_3 + I_1 \ddot{u}^0_{\alpha|\alpha} - I_2 \ddot{u}^0_{3|\alpha\alpha} \quad (35)$$

(ii) First-Order Shear Deformation Plate Theory

$(u_3^0 = 0, \quad f = x^3)$

$$N^{\alpha\beta}|_\beta + \underline{(Q^\beta u_\alpha^1)}|_\beta = I_0 \ddot{u}_\alpha^0 + I_1 \ddot{u}_\alpha^1$$

$$Q^\alpha|_\alpha + N^{\alpha\beta}\hat{u}_{3|\beta}^0)|_\alpha = q + I_0 \ddot{\hat{u}}^0_3$$

$$M^{\alpha\beta}|_\beta - Q^\beta(\delta_{\alpha\beta} + \underline{u_{\alpha|\beta}^0}) - \underline{N^3 u_\alpha^1} + \underline{R^\beta|_\beta u_\alpha^1} = I_1 \ddot{u}_\alpha^0 + I_2 \ddot{u}_\alpha^1. \qquad (36)$$

(iii) Third-Order Shear Deformation Plate Theory

$(u_3^0 = 0, \quad f = x^3[1 - \frac{4}{3}(x^3/h)^2])$

$$N^{\alpha\beta}|_\beta + \underline{(\hat{Q}^\beta u_\alpha^1)}|_\beta = I_0 \ddot{u}_\alpha^0 + I_1^f \ddot{u}_\alpha^1$$

$$Q^\alpha|_\alpha + (N^{\alpha\beta}\hat{u}_{3|\beta}^0)|_\alpha = q + I_0 \ddot{\hat{u}}^0_3$$

$$P^{\alpha\beta}|_\beta - \hat{Q}^\beta(\delta_{\alpha\beta} + \underline{u_{\alpha|\beta}^0}) + \underline{\hat{S}^\beta|_\beta u_\alpha^1} - \underline{\hat{N}^3 u_\alpha^1} = I_1^f \ddot{u}_\alpha^0 + I_2^f \ddot{u}_\alpha^1 \ . \qquad (37)$$

Note that several other theories can be obtained from Eq. (35) as special cases. Analytical solutions to the linear version of the third-order theory were presented in Section II.

Acknowledgements. The research summarized herein was conducted under Grant DAAG 29-85-K-00017 from the Mathematical Sciences Division of the Army Research Office (ARO). The author gratefully acknowledges the encouragement and support of Dr. Jagdish Chandra of ARO. Also, the author is pleased to acknowledge the technical contributions to this research made by Dr. A. A. Khdeir and Professor L. Librescu. It is a great pleasure to acknowledge the prompt and skillful typing of the manuscript by Mrs. Vanessa McCoy.

References

1. Stoker, J. J., "Mathematical Problems Connected with the Bending and Buckling of Elastic Plates," Bull. Amer. Math. Soc., Vol. 48, pp. 247-261, 1942.

2. Reissner, E., "On the Theory of Bending of Elastic Plates," J. Math. Phy., Vol. 23, pp. 184-191, 1944.

3. Reissner, E., "The Effect of Transferse Shear Deformation on the Bending of Elastic Plates," J. Appl. Mech., Vol. 12, No. 1, pp. A-69 to A-77, 1945.

4. Reissner, E., "On Bending of Elastic Plates," Q. Appl. Math., Vol. 5, pp. 55-68, 1947.

5. Basset, A. B., "On the Extension and Flexure of Cylindrical and Spherical Thin Elastic Shells," Phil. Trans. Royal Soc., (London), Ser. A, Vol. 181, No. 6, pp. 433-480, 1890.

6. Hildebrand, F. B., Reissner, E. and Thomas, G. B., "Notes on the Foundations of the Theory of Small Displacements of Orthotropic Shells," NACA Technical Note No. 1833, March 1949.

7. Mindlin, R. D., "Influence of Rotatory Inertia and Shear on Flexural Motions of Isotropic, Elastic Plates," J. Appl. Mech., Vol. 18, pp. 31-38, 1951.

8. Hencky, H., "Über Die Berücksichtigung Der Schubverzerrung in Ebenen Platten," Ing. Arch., Vol. 16, pp. 72-76, 1947.

9. Reddy, J. N., "A Refined Nonlinear Theory of Plates with Transverse Shear Deformation," Int. J. Solids & Struct., Vol. 20, No. 9/10, pp. 881-896, 1984.

10. Reddy, J. N., "A Simple Higher-Order Theory for Laminated Composite Plates," J. Appl. Mech., Vol. 51, pp. 745-752, 1984.

11. Reddy, J. N., Energy and Variational Methods in Applied Mechanics, John Wiley, New York, 1984.

12. Sanders, J. L., "Nonlinear Theories for Thin Shells," Quart. Appl. Math., Vol. 21, pp. 21-36, 1963

13. Koiter, W. T., "On the Nonlinear Theory of Thin Elastic Shells," Proc. Kon. Ned. Ak. Wet., Series B., Vol. 69, pp. 1-54, 1966.

14. Reissner, E., "Rotationally Symmetric Problems in the Theory of Thin Elastic Shells," Proc. 3rd U.S. Nat. Congr. of Appl. Mech., pp. 51-69, 1958.

15. Pietraszkiewicz, W., "Lagrangian Description and Incremental Formulation in the Nonlinear Theory of Thin Shells," Int. J. Non-Linear Mechanics, Vol. 19, pp. 115-140, 1984.

16. Naghdi, P. M. and Vongsarnpigoon, L., "A Theory of Shells with Small Strains Accompanied by Moderate Rotations," Arch. for Rational Mechanics and Analysis, Vol. 83, pp. 245-283, 1983.

17. Librescu, L. and Schmidt, R., "Higher-Order Moderate Rotation Theories for Elastic Anisotropic Plates," Finite Rotations in Structural Mech., (Proc. of Euromech Symposium), No. 197, Jablonna, Poland, 1985, W. Pietraszkiewicz (ed.), Springer-Verlag, Berlin, pp. 158-174, 1986.

18. Reddy, J. N., "A Small Strain and Moderate Rotation Theory of Elastic Anisotropc Plates," J. Appl. Mech., in press.

19. Khdeir, A. A., Reddy, J. N. and Librescu, L., "Analytical Solutions of a Refined Shear Deformation Theory for Rectangular Composite Plates," Int. J. Solids & Struct., in press.

20. Franklin, J. N., Matrix Theory, Prentice-Hall; Englewood Cliffs, N.J., 1968.

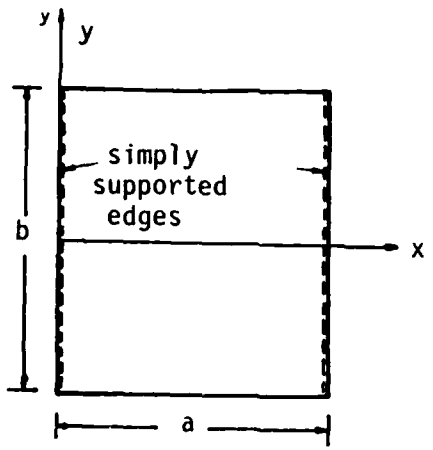


Fig. 1 Geometry and coordinate system used in the study.

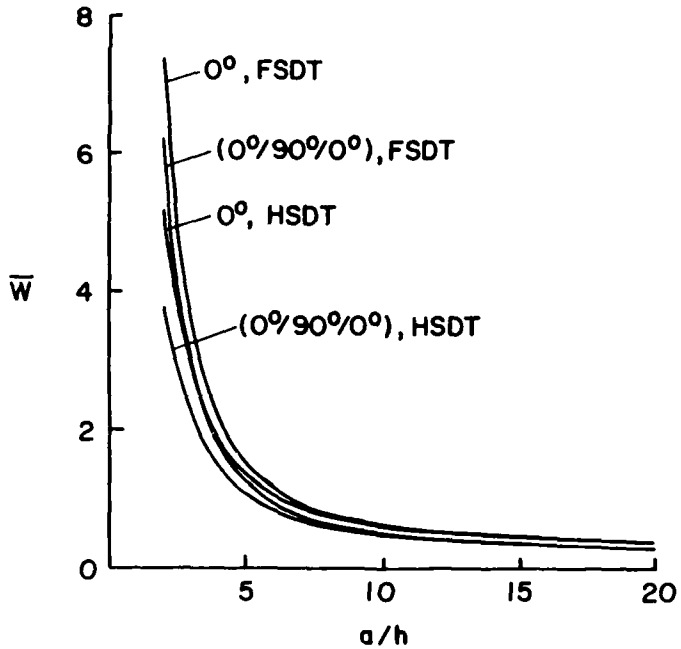


Fig. 2 Nondimensionalized center deflection versus side-to-thickness ratio of SC plates.

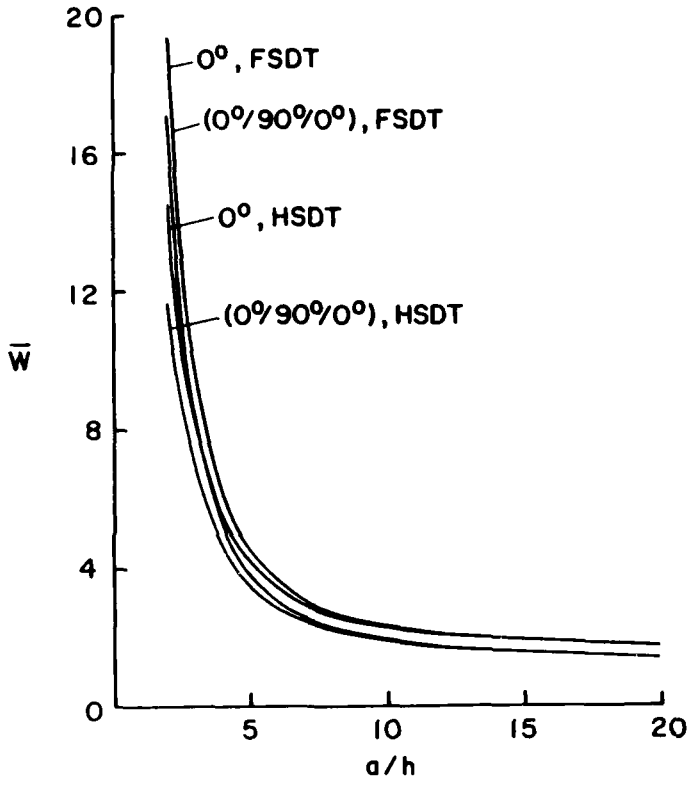


Fig. 3. Nondimensionalized center deflection versus side-to-thickness ratio of FC plates.

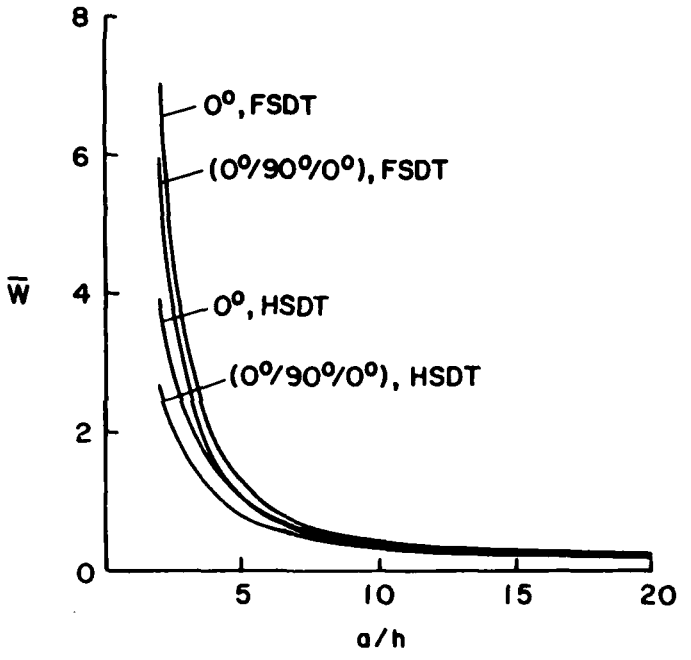


Fig. 4. Nondimensionalized center deflection versus side-to-thickness ratio of CC plates.

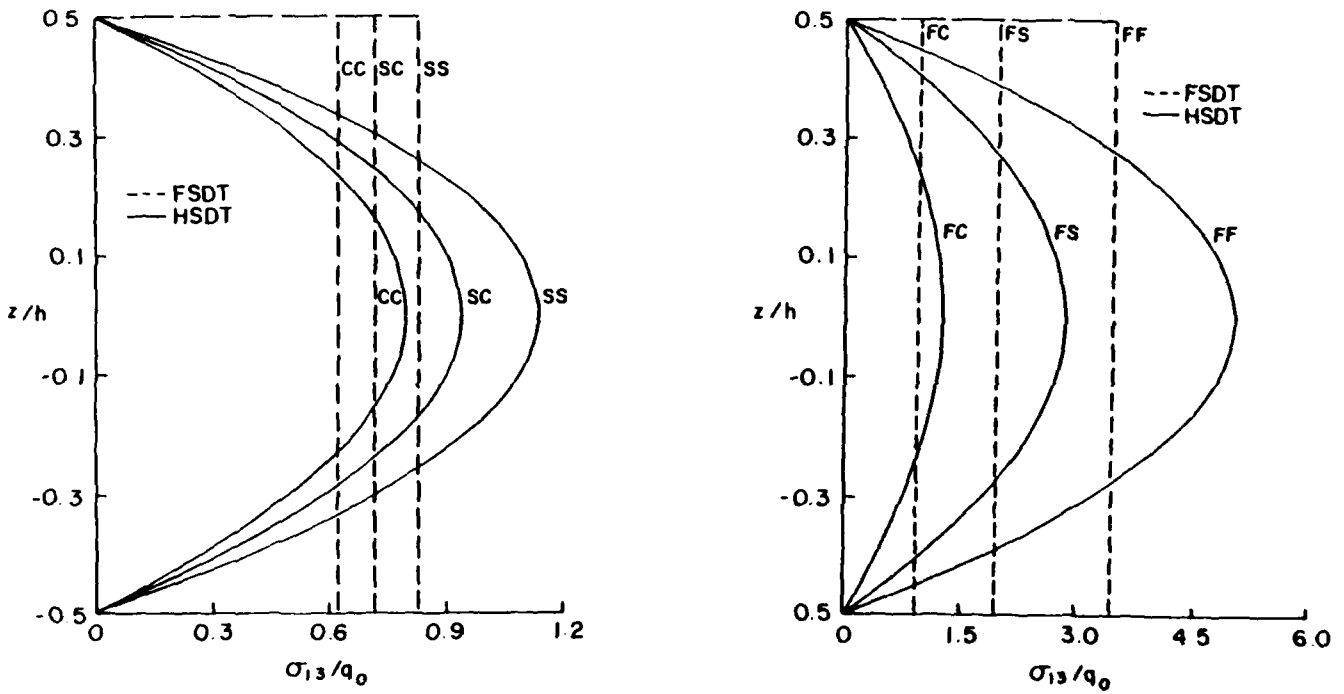


Fig. 5. Variation of the transverse shear stress through the thickness of orthotropic plates ($h/a = 0.14$).

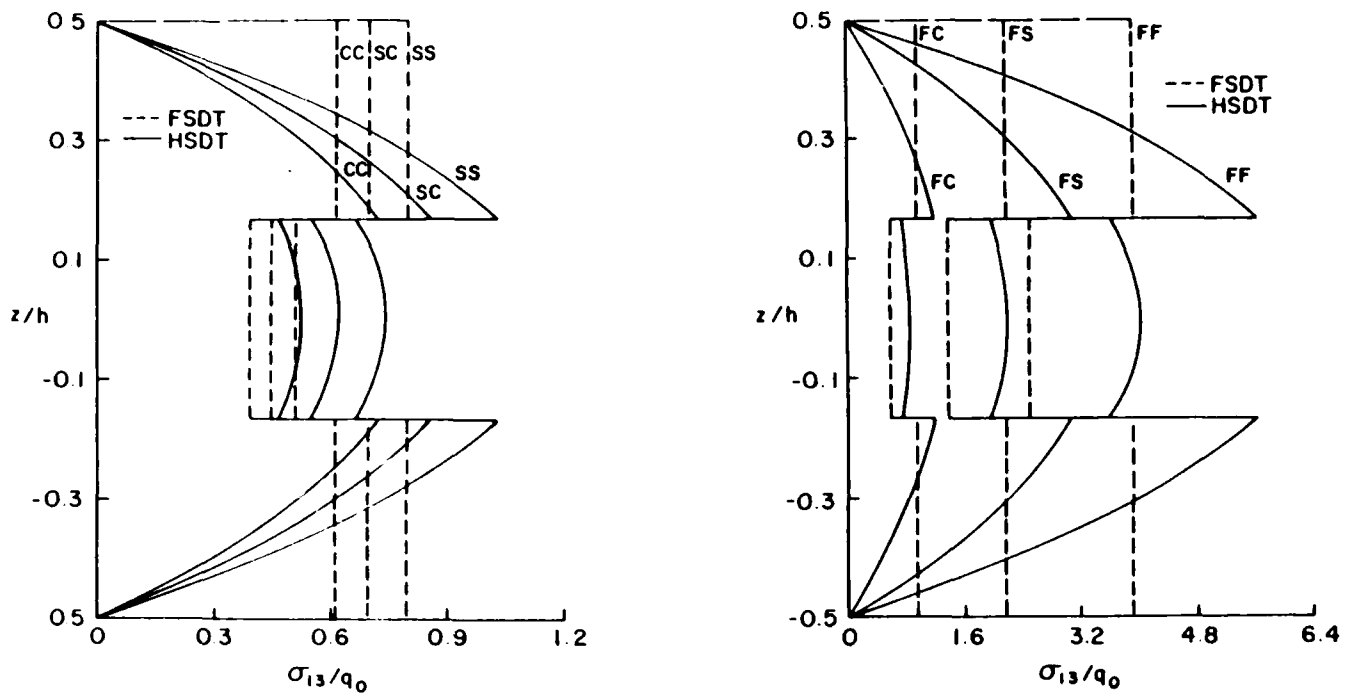


Fig. 6. Variation of the transverse shear stress through the thickness of cross-ply (0/90/0) laminates (h/a = 0.14).

NUMERICAL SOLUTION OF PARABOLIC PROBLEMS IN HIGH DIMENSIONS

Edward Dean, Roland Glowinski, Chin-Hsien Li
University of Houston
Department of Mathematics
4800 Calhoun Road
Houston, TX 77006

ABSTRACT. The main goal of this paper is to discuss the numerical solution of mathematical problems of parabolic type when the space dimension is high and/or the number of discretization points is quite large. In such cases, we can take advantage of the evolution nature of the problem under consideration to derive numerical methods quite easy to implement and well suited to vector and/or parallel computers. Operator splitting methods are one of the key ingredients of such a methodology. We shall illustrate the methods described in this paper by solving the time dependent Navier-Stokes equations for incompressible viscous fluids, a variational problem originating from the physics of liquid crystals, and finally advection-diffusion problems in very high dimension associated to the solution of the Zakai equation in stochastic optimal control.

1. GENERALITIES AND SYNOPSIS.

Linear and nonlinear Parabolic Problems for *Partial Differential Operators* occur in many branches of Natural and Engineering Sciences. One of

the main goals of this paper is to discuss the numerical solution of such problems by methods taking advantage of the *evolution nature* of the problem under consideration, and also well-suited to *vector* and/or *parallel computers*.

Operator splitting methods are definitely a key to the solution of such problems and a description of these methods will be given in Section 2. In Section 3, we shall consider the solution of the *Navier-Stokes equations* for unsteady incompressible viscous flows, then in Section 4 the solution of *nonconvex variational problems* originating from the physics of *liquid crystals*. Finally in Section 5, we shall consider *time dependent advection diffusion problems* whose solution is an important part of some solution methods for those complicated *Zakai equations* originating from *Stochastic Optimal Control*; we shall discuss there various solution methods using first and second order *upwinding* and also the *modified method of characteristics* when the diffusion coefficients are small.

The techniques described in Sections 3, 4, 5 will be illustrated by numerical experiments.

2. DESCRIPTION OF SOME BASIC OPERATOR SPLITTING METHODS FOR TIME DEPENDENT PROBLEMS.

2.1. GENERALITIES.

Let V be a *Banach space*; we consider in V the following *initial value problem*

$$\frac{du}{dt} + A(u) = 0, \tag{2.1}$$

$$u(0) = u_0 , \tag{2.2}$$

where, in (2.1), $A(\cdot)$ is a linear or nonlinear operator.

We suppose that A has the following *nontrivial* decomposition property

$$A = A_1 + A_2 \tag{2.3}$$

(by *nontrivial* we mean that A_1 and A_2 are "individually" simpler than A).

There are many techniques to achieve the numerical integration of the initial value problem (2.1), (2.2) by taking advantage of the decomposition property (2.3). We shall describe some of them just below (more techniques are described in, e.g., [1]). Before giving these descriptions let's introduce some helpful notation.

In the sequel $\Delta t\ (>0)$ will be a time discretization step and $u^{n+\alpha}$ will denote an approximation of $u((n+\alpha)\Delta t)$. The first scheme to be described is the Peaceman-Rachford scheme (cf. Sec. 2.2) and then what we call a θ - scheme (cf. Sec. 2.3).

2.2. THE PEACEMAN-RACHFORD SCHEME.

The principle of that scheme, introduced in [2], is quite simple:

Consider the time interval $[n\Delta t, (n+1)\Delta t]$ and suppose that u^n is known; introducing the mid-point $(n+1/2)\Delta t$ we integrate (2.1) over $[n\Delta t, (n+1/2)\Delta t]$ by a scheme which is of *backward Euler type* for A_1 (implicit in A_1) and of the *forward Euler type* for A_2 (explicit); on $[(n+1/2)\Delta t, (n+1)\Delta t]$ we exchange the roles of A_1 and A_2. The above program is definitely realized by the following scheme:

$$u^0 = u_0\ ; \tag{2.4}$$

then for $n \geq 0$, u^n *being known, we compute* $u^{n+1/2}$ *and* u^{n+1} *by solving successively*

$$\frac{u^{n+1/2} - u^n}{\Delta t/2} + A_1(u^{n+1/2}) + A_2(u^n) = 0 \ , \tag{2.5}$$

$$\frac{u^{n+1} - u^{n+1/2}}{\Delta t/2} + A_1(u^{n+1/2}) + A_2(u^{n+1}) = 0 \ . \tag{2.6}$$

We observe that, initialization excepted, A_1 and A_2 play a *symmetric* role in the above scheme.

To study some of the basic properties of scheme (2.4) - (2.6) , such as *accuracy* and *stability* , we consider the particular case where

(i) $V = \mathbb{R}^N$.

(ii) A is an $N \times N$ *symmetric* and *positive definite* matrix; $u_0 \in \mathbb{R}^N$.

In such a case, the exact solution of (2.1) , (2.2) is known and is given by

$$u(t) = e^{-tA} u_0 \ .$$

Concerning the decomposition of A we decompose it as follows:

$$A = A_1 + A_2 \ ; \ A_1 = \alpha A \ , \ A_2 = \beta A \ , \ \textit{with} \ \alpha + \beta = 1 \ , \ 0 < \alpha, \beta < 1. \tag{2.8}$$

Stability Properties of Scheme (2.4) - (2.6): We have from (2.5) , (2.6) , (2.8) ,

$$u^{n+1} = \left[I + \beta \frac{\Delta t}{2} A\right]^{-1} \left[I - \alpha \frac{\Delta t}{2} A\right] \left[I + \alpha \frac{\Delta t}{2} A\right]^{-1} \left[I - \beta \frac{\Delta t}{2} A\right] u^n \tag{2.9}$$

Using a vector basis consisting of eigenvectors of A, we have from (2.9)

$$u_i^{n+1} = \frac{(1-\alpha \frac{\Delta t}{2} \lambda_i)(1-\beta \frac{\Delta t}{2} \lambda_i)}{(1+\alpha \frac{\Delta t}{2}\lambda_i)(1+\beta \frac{\Delta t}{2} \lambda_i)} u_i^n \quad , \tag{2.10}$$

where λ_i $(>0, \ \forall\ i = 1, \ldots N)$ is the i^{th} eigenvalue of A; we suppose that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_N$. Consider now the *rational function* R_1 defined by

$$R_1(x) = \frac{(1-\frac{\alpha}{2} x)(1-\frac{\beta}{2} x)}{(1+\frac{\alpha}{2} x)(1+\frac{\beta}{2} x)} \ ; \tag{2.11}$$

we observe that $|R_1(x)| < 1$ for all $x>0$, implying, in that simple case, the *unconditional stability* of scheme (2.4) - (2.6). Since

$$\lim_{x \to +\infty} R_1(x) = 1 \ , \tag{2.12}$$

we observe that for *stiff* problems, i.e. problems, such that $\lambda_N/\lambda_1 >> 1$, scheme (2.4) - (2.6) is not very good to damp *simultaneously* the components of u^n associated to the large and to the small eigenvalues of A. From this observation, we can expect that scheme (2.4) - (2.6) is not well suited to "capture" the steady state solutions of stiff problems (like those obtained from the discretization of partial differential equations); this has been confirmed by numerical experiments.

Accuracy Properties of Scheme (2.4) - (2.6): Since

$$e^{-x} = 1 - x + \frac{x^2}{2} + x^2 \epsilon(x) \quad , \tag{2.13}$$

and, from (2.11),

$$R_1(x) = 1 - x + \frac{x^2}{2} + x^2\,\eta(x)\,, \tag{2.14}$$

with $\lim_{x\to 0} \epsilon(x) = \lim_{x\to 0} \eta(x) = 0$, we have that scheme (2.4) - (2.6) is

second order accurate in the simple case considered here. We observe from (2.9), that if one takes $\alpha = \beta = 1/2$, then the two linear systems, which have to be solved at each full step are, in fact, associated to the same matrix $I + \Delta t\, A/4$.

2.3 The θ - Scheme.

In order to construct operator splitting methods better suited than scheme (2.4) - (2.6) to the numerical integration of stiff initial value problems (2.1), (2.2), we introduce first $\theta \in (0, .5)$ and then associate to θ the decomposition of interval $[n\Delta t, (n+1)\Delta t]$ given by

$$[n\Delta t, (n+1)\Delta t] = [n\Delta t, (n+\theta)\Delta t] \cup [(n+\theta)\Delta t, (n+1-\theta)\Delta t] \cup [(n+1-\theta)\Delta t, (n+1)\Delta t]\,.$$

A numerical method for (2.1), (2.2) taking advantage of (2.3) and of the above splitting of $[n\Delta t, (n+1)\Delta t]$ is defined as follows:

$$u^0 = u_0\,; \tag{2.15}$$

then for $n \geq 0$, u^n *being known, we compute* $u^{n+\theta}$, $u^{n+1-\theta}$ *and* u^{n+1} *by solving successively*

$$\frac{u^{n+\theta} - u^n}{\theta\Delta t} + A_1(u^{n+\theta}) + A_2(u^n) = 0\,, \tag{2.16}$$

$$\frac{u^{n+1-\theta} - u^{n+\theta}}{(1-2\theta)\Delta t} + A_1 (u^{n+\theta}) + A_2 (u^{n+1-\theta}) = 0 , \tag{2.17}$$

$$\frac{u^{n+1} - u^{n+1-\theta}}{\theta\Delta t} + A_1 (u^{n+1}) + A_2 (u^{n+1-\theta}) = 0 . \tag{2.18}$$

Stability and Accuracy Properties of Scheme (2.15) - (2.18): taking the same model problem as in Section 2.2, we have (with $\theta' = 1 - 2\theta$)

$$u^{n+1} = (I + \alpha\theta\Delta t\, A)^{-1} (I - \beta\theta\Delta t A) (I + \beta\theta'\Delta t A)^{-1} (I - \alpha\theta'\Delta t A)$$
$$(I + \alpha\theta\Delta t A)^{-1} (I - \beta\theta\Delta t A)\, u^n. \tag{2.19}$$

which implies

$$u_i^{n+1} = \frac{(1-\beta\theta\Delta t\lambda_i)^2 (1-\alpha\theta'\Delta t\lambda_i)}{(1+\alpha\theta\Delta t\lambda_i)^2 (1+\beta\theta'\Delta t\lambda_i)}\, u_i^n . \tag{2.20}$$

Consider now the rational function R_2 defined by

$$R_2(x) = \frac{(1-\beta\theta x)^2 (1-\alpha\theta' x)}{(1+\alpha\theta x)^2 (1+\beta\theta' x)} ; \tag{2.21}$$

since

$$\lim_{x \to +\infty} |R_2 (x)| = \beta/\alpha \tag{2.22}$$

we should prescribe

$$\alpha \geq \beta \tag{2.23}$$

to have, from (2.19) , (2.20) , the stability of scheme (2.15) - (2.18) for the *large eigenvalues* of A. Concerning the accuracy of scheme (2.15) - (2.18) we can show that

$$R_2(x) = 1 - x + \frac{x^2}{2}\left\{1 + (\beta^2-\alpha^2)(2\theta^2-4\theta+1)\right\} + x^2\eta(x) , \qquad (2.24)$$

with $\lim_{x\to 0} \eta(x) = 0$. It follows from (2.24) that scheme (2.215) - (2.18) is *second order accurate* if either

$$\alpha=\beta \ (= 1/2 \ \textit{from} \ (2.8)) , \qquad (2.25)$$

or

$$\theta = 1 - 1/\sqrt{2} = .29289. . . . ; \qquad (2.26)$$

scheme (2.15) - (2.18) is only *first order accurate if neither* (2.25) *nor* (2.26) holds. If one takes $\alpha = \beta = 1/2$, it follows from (2.20), (2.21) that scheme (2.15) - (2.18) is *unconditionally stable* for all $\theta\in(0, 1/2)$; however, since (from (2.22)) we have

$$\lim_{x\to+\infty} |R_2(x)| = 1 , \qquad (2.27)$$

the remark stated for scheme (2.4) - (2.6) concerning the integration of stiff systems still holds. In general, we shall choose α and β in order to have the same matrix for all the partial steps of the integration procedure; i.e. , α , β , θ have to satisfy

$$\alpha\theta = \beta(1-2\theta) , \qquad (2.28)$$

which implies

$$\alpha = (1-2\theta)/(1-\theta) , \quad \beta = \theta/(1-\theta) . \qquad (2.29)$$

Combining (2.23) , (2.29) we obtain

$$0 < \theta \leq 1/3 \quad . \tag{2.30}$$

For $\theta = 1/3$, (2.29) implies $\alpha = \beta = 1/2$; the resulting scheme is just a variant of scheme (2.4) - (2.6) .

If $0 < \theta < 1/3$, and if α and β are given by (2.29) , we have then

$$\lim_{x \to +\infty} |R_2(x)| = \beta/\alpha = \theta/(1-2\theta) < 1 \quad . \tag{2.31}$$

Actually, we can prove that $\theta \in [\theta^*, 1/3]$ (with $\theta^* = .087385580 \ldots.$) and α , β given by (2.29) imply the *unconditional stability* of scheme (2.15) - (2.18) . Moreover, if $\theta \in (\theta^* , 1/3)$, property (2.31) makes that scheme (2.15) - (2.18) has good asymptatic properties as $n \to +\infty$ and for example is well suited to compute steady state solutions. If $\theta = 1 - 1/\sqrt{2}$ (resp. $\theta = 1/4$) , we have $\alpha = 2 - \sqrt{2}$. $\beta = \sqrt{2} - 1$, $\beta/\alpha = 1/\sqrt{2}$ (resp. $\alpha = 2/3$, $\beta = 1/3$, $\beta/\alpha = 1/2$).

2.4. Further Comments on Operator Splitting Methods.

Integration schemes related to (2.15) - (2.18) have been discussed in [3] (see also [4] - [6]) . Concerning the convergence of the above schemes, the convergence of the Peaceman-Rachford scheme (2.4) - (2.6) has been proved in [7] (see also [8]) under quite general *montonicity* assumptions on A_1 and A_2 (in fact these operators can even be *multivalued*). These are not such general results at the moment for scheme (2.15) - (2.18) (see however the discussion in [9]). In [10] , one can find splitting methods derived from the Lie-Trotter formula and applicable to situations in which $A = A_1 + A_2 + A_3$; these methods however may be inaccurate for steady state calculations; indeed splitting methods for more than two operators are also discussed in, e.g., [1] , [11] , [12] .

To conclude Section 2, we would like to describe a variation of scheme (2.4) - (2.6) (due to Douglas and Rachford; cf. [13]) ; in some occasions it seems to behave better than (2.4) - (2.6) as a tool to capture steady state solutions of systems such as (2.1) , (2.2) , however, as a method for the numerical integration of (2.1) , (2.2) it is only first order accurate. In addition to that, although more robust than scheme (2.4) - (2.6) , it also suffers from the basic drawback of not being well suited to the numerical integration of stiff differential systems.

The *Douglas-Rachford scheme* is described by

$$u^0 = u_0 \ ; \tag{2.32}$$

then for $n \geq 0$, u^n *being known, we compute* $\overset{v}{u}^{n+1}$ *and* u^{n+1} *as the solutions of*

$$\frac{\overset{v}{u}^{n+1} - u^n}{\Delta t} + A_1 (\overset{v}{u}^{n+1}) + A_2(u^n) = 0 \ , \tag{2.33}$$

$$\frac{u^{n+1} - u^n}{\Delta t} + A_1 (\overset{v}{u}^{n+1}) + A_2 (u^{n+1}) = 0 \ . \tag{2.34}$$

The convergence of scheme (2.32) - (2.34) is proved in [7] , [8] for A_1 , A_2 *monotone* (possibly multivalued) operators.

3. APPLICATION TO THE NAVIER-STOKES EQUATIONS FOR INCOMPRESSIBLE VISCOUS FLUIDS.

3.1. GENERALITIES. SYNOPSIS.

In this section, we shall discuss the application of the operator splitting methods described in Section 2 to the numerical simulation of *incompressible*

viscous flows modeled by the *Navier-Stokes equations*. We shall only give here the general principle of such numerical treatment, referring for more details to [14] - [18] .

Let us consider a Newtonian incompressible viscous fluid. If Ω and Γ denote the flow region ($\Omega \subset \mathbb{R}^N$, N = 2, 3 in practice) and its boundary, respectively, then this flow is governed by the following *Navier-Stokes equations*

$$\frac{\partial \underset{\sim}{u}}{\partial t} - \nu \Delta \underset{\sim}{u} + (\underset{\sim}{u}.\underset{\sim}{\nabla})\underset{\sim}{u} + \underset{\sim}{\nabla} p = \underset{\sim}{f} \quad in \quad \Omega \quad , \tag{3.1}$$

$$\underset{\sim}{\nabla}.\underset{\sim}{u} = 0 \quad in \quad \Omega \quad (incompressibility\ condition). \tag{3.2}$$

In (3.1) , (3.2) ,

(a) $\underset{\sim}{\nabla} = \left\{\frac{\partial}{\partial x_i}\right\}_{i=1}^{N}$, $\Delta = \underset{\sim}{\nabla}^2 = \sum_{i=1}^{N} \frac{\partial^2}{\partial x_i^2}$, $x = \{x_i\}_{i=1}^{N}$ the *generic point of* $\mathbb{R}^N$,

(b) $\underset{\sim}{u} = \{u_i\}_{i=1}^{N}$ is the *flow velocity*,

(c) p is the *pressure* ,

(d) ν is a *viscosity parameter*,

(e) $\underset{\sim}{f}$ is a *density of external forces*.

In (3.1) , $(\underset{\sim}{u}.\underset{\sim}{\nabla})\underset{\sim}{u}$ is a *symbolic notation* for the nonlinear vector term

$$\left\{\sum_{j=1}^{N} u_j \frac{\partial u_i}{\partial x_j}\right\}_{i=1}^{N} .$$

Boundary and initial conditions have to be added to (3.1) , (3.2) ; here, we shall only consider *Dirichlet boundary conditions* such as

$$\underset{\sim}{u} = \underset{\sim}{g} \quad on \quad \Gamma \ , \tag{3.3}$$

with, from the incompressibility condition (3.2) ,

$$\int_{\Gamma} \underset{\sim}{g}\cdot\underset{\sim}{n}\ d\Gamma = 0 \ , \tag{3.4}$$

with $\underset{\sim}{n}$ the *outward unit vector normal to* Γ .

Finally we shall prescribe as initial condition

$$\underset{\sim}{u}(x,0) = \underset{\sim}{u}_0(x) \quad a.e.\ on \quad \Omega \ , \ with \ \ \underset{\sim}{\nabla}\cdot\underset{\sim}{u}_0 = 0 \ . \tag{3.5}$$

Boundary conditions more complicated than (3.3) are discussed in, e.g., [14] , [18].

The Navier-Stokes equations for incompressible viscous fluids have been motivating a very large number of papers, books, reports, symposia, workshops, etc. Mentioning all of them is impossible and we therefore refer to the references in [14] - [18] .

The difficulties with the Navier-Stokes equations (even for flows at low Reynolds numbers, in bounded regions Ω) are

(i) the *nonlinear* term $(\underset{\sim}{u}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}$ in (3.1) ,

(ii) the *incompressiblity* condition (3.2) ,

(iii) the fact that their solutions are *vector-valued* functions of x, t , whose components are coupled by the nonlinear term $(\underset{\sim}{u}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}$ and by the incompressibility condition $\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0$.

Using the operator splitting methods of Section 2 for the time discretization of the Navier-Stokes equations, we shall be able to decouple those difficulties associated to the nonlinearity and the incompressibility, respectively.

3.2. Time Discretization by Operator Splitting Methods.

We shall concentrate on the θ-scheme since, from our numerical experiments, it seems to be the one giving the best results. We have then

$$\underset{\sim}{u}^0 = \underset{\sim}{u}_0 \; ; \tag{3.6}$$

then for $n \geq 0$, *starting from* $\underset{\sim}{u}^n$ *we solve*

$$\begin{cases} \dfrac{\underset{\sim}{u}^{n+\theta} - \underset{\sim}{u}^n}{\theta\Delta t} - \alpha\nu\Delta\underset{\sim}{u}^{n+\theta} + \underset{\sim}{\nabla} p^{n+\theta} = \underset{\sim}{f}^{n+\theta} + \beta\nu\Delta\underset{\sim}{u}^n - (\underset{\sim}{u}^n\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^n \quad in \quad \Omega \; , \\ \underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+\theta} = 0 \quad in \quad \Omega \; , \\ \underset{\sim}{u}^{n+\theta} = \underset{\sim}{g}^{n+\theta} \quad on \quad \Gamma \; , \end{cases} \tag{3.7}$$

$$\begin{cases} \dfrac{\underset{\sim}{u}^{n+1-\theta} - \underset{\sim}{u}^{n+\theta}}{(1-2\theta)\Delta t} - \beta\nu\Delta\underset{\sim}{u}^{n+1-\theta} + (\underset{\sim}{u}^{n+1-\theta}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^{n+1-\theta} \\ = \underset{\sim}{f}^{n+1-\theta} + \alpha\nu\Delta\underset{\sim}{u}^{n+\theta} - \underset{\sim}{\nabla} p^{n+\theta} \quad in \quad \Omega \; , \\ \underset{\sim}{u}^{n+1-\theta} = \underset{\sim}{g}^{n+1-\theta} \quad on \quad \Gamma \; , \end{cases} \tag{3.8}$$

$$\begin{cases} \dfrac{\underset{\sim}{u}^{n+1} - \underset{\sim}{u}^{n+1-\theta}}{\theta\Delta t} - \alpha\nu\Delta\underset{\sim}{u}^{n+1} + \underset{\sim}{\nabla} p^{n+1} = \underset{\sim}{f}^{n+1} + \beta\nu\Delta\underset{\sim}{u}^{n+1-\theta} \\ \qquad - (\underset{\sim}{u}^{n+1-\theta}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^{n+1-\theta} \quad in \quad \Omega \; , \\ \underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+1} = 0 \quad in \quad \Omega \; , \\ \underset{\sim}{u}^{n+1} = \underset{\sim}{g}^{n+1} \quad on \quad \Gamma \; . \end{cases} \tag{3.9}$$

3.3. Some Comments and Remarks Concerning Scheme (3.6) - (3.9)

Using the above operator splitting method, we have been able to decouple *nonlinearity* and *incompressibility* in the Navier-Stokes equations (3.1) , (3.2). We shall comment in the following sections on the specific treatment of the subproblems encountered at each step of algorithm (3.6) - (3.9). We shall only consider the case where the subproblems are still continuous in space (since the formalism of the continuous problems is much simpler); for the fully discrete case see [14] (with $\theta = 1/4$) and [18] where finite element approximations of (3.1) , (3.2) are discussed.

We observe that $\underset{\sim}{u}^{n+\theta}$ and $\underset{\sim}{u}^{n+1}$ are obtained from the solution of linear problems very close to the *steady Stokes problem.*

If one uses scheme (3.6) - (3.9) , the best choice for α and β is given by (2.29). With such a choice, many computer subprograms can be used for both the linear and nonlinear subproblems, resulting therefore in a quite substantial core memory savings.

3.4. Solution of the Nonlinear Subproblem (3.8)

This not the place to give a detailed discussion of solution methods for the nonlinear subproblem (3.8); we should observe however that it belongs to the following class of *nonlinear Dirichlet systems*

$$\begin{cases} \alpha \underset{\sim}{u} - \nu \Delta \underset{\sim}{u} + (\underset{\sim}{u} \cdot \nabla) \underset{\sim}{u} = \underset{\sim}{f} \quad in \quad \Omega \, , \\ \underset{\sim}{u} = \underset{\sim}{g} \quad on \quad \Gamma \, , \end{cases} \tag{3.10}$$

where α and ν are two positive parameters (with $\alpha \sim 1/\Delta t$, here) and where $\underset{\sim}{f}$ and $\underset{\sim}{g}$ are two given functions defined on Ω and Γ , respectively.

Several solution methods for (3.10) are discussed in [14] - [18], including *Newton's method* and *nonlinear least squares conjugate gradient* (see also [19] for further details). In the case of the nonlinear least squares conjugate gradient methods, we have been using algorithms preconditioned by *discrete variants* of the elliptic operator

$$\underset{\sim}{v} \rightarrow \alpha \underset{\sim}{v} - \nu \Delta \underset{\sim}{v} \qquad (3.11)$$

with homogeneous Dirichlet boundary conditions. In the case of flows at large Reynold numbers the viscosity parameter ν is usually small; moreover the fast dynamics of these flows require a small Δt implying that α is a large number. From these facts, the discrete forms of the elliptic operator (3.11) are matrices whose *condition number is small* implying that simple solution methods such as successive over relaxation (S.O.R.) and nonpreconditioned conjugate gradient methods will have a very fast convergence for solving the linear systems associated to those matrices approximating operator (3.11) (relaxation methods are particularly interesting since they have very good vectorization and parallelization properties); indeed acceleration methods such as multigrid or preconditioned conjugate gradient are useless for these specific problems. Similarly the iterative solution of the discrete variants of (3.10) by the nonlinear least square conjugate gradient methods described in [14] - [18] is quite fast and obtained in 3 to 4 iterations.

3.5. Solution of the Stokes Linear Subproblems (3.7) , (3.9)

At each full step of algorithm (3.6) - (3.9) we have to solve two *linear problems* of the following type

$$\begin{cases} \alpha \underset{\sim}{u} - \nu \Delta \underset{\sim}{u} + \underset{\sim}{\nabla} p = \underset{\sim}{f} \quad in \quad \Omega , \\ \underset{\sim}{\nabla} \cdot \underset{\sim}{u} = 0 \quad in \quad \Omega , \\ \underset{\sim}{u} = \underset{\sim}{g} \quad on \quad \Gamma \quad (with \int_{\Gamma} \underset{\sim}{g} \cdot \underset{\sim}{n} \, d\Gamma = 0) , \end{cases} \tag{3.12}$$

where α and ν are two *positive* parameters, and where $\underset{\sim}{f}$ and $\underset{\sim}{g}$ are two *given* functions defined on Ω and Γ, respectively. We recall that if $\underset{\sim}{f}$ and $\underset{\sim}{g}$ are sufficiently smooth, then problem (3.12) has a unique solution in $V_g \times (L^2(\Omega)/\mathbb{R})$, with

$$V_g = \{\underset{\sim}{v} \mid \underset{\sim}{v} \in (H^1(\Omega))^N , \ \underset{\sim}{v} = \underset{\sim}{g} \quad on \quad \Gamma\} ; \tag{3.13}$$

$p \in L^2(\Omega)/\mathbb{R}$ means that p is defined only to winthin an arbitrary constant.

We refer to [18] and the references therein for the discussion of iterative and direct methods for solving problem (3.12). Our favorite method is at the moment a *conjugate gradient* variant discussed in [18] of the following algorithm (introduced in [20]):

$$p^0 \in L^2(\Omega) \quad given; \tag{3.14}$$

then for $n \geq 0$, *assuming that* p^n *is known, we compute* u^n *and* p^{n+1} *by*

$$\begin{cases} \alpha \underset{\sim}{u}^n - \nu \Delta \underset{\sim}{u}^n = \underset{\sim}{f} - \underset{\sim}{\nabla} p^n \quad in \quad \Omega , \\ \underset{\sim}{u}^n = \underset{\sim}{g} \quad on \quad \Gamma , \end{cases} \tag{3.15}$$

$$\begin{cases} -\Delta \phi^n = \underset{\sim}{\nabla} \cdot \underset{\sim}{u}^n \quad in \quad \Omega , \\ \dfrac{\partial \phi^n}{\partial n} = 0 \quad on \quad \Gamma , \ \int_{\Omega} \phi^n dx = 0 , \end{cases} \tag{3.16}$$

and with $\rho > 0$,

$$p^{n+1} = p^n - \rho(\nu \underset{\sim}{\nabla} \cdot \underset{\sim}{u}^n + \alpha\phi^n) \quad . \tag{3.17}$$

Concerning the convergence of algorithm (3.14) - (3.17) we should prove by a variant of the techniques discussed in [14, Chapter 7] the following:

PROPOSITION 3.1: *Suppose that we have*

$$0 < \rho < \frac{2}{N} \frac{1}{1 + c^2 \frac{\alpha}{\nu}} \quad , \tag{3.18}$$

with

$$c = \sup_{\phi \in H - \{0\}} \left[\frac{\| \underset{\sim}{\nabla} \phi \|_{L^2(\Omega)^N}}{\| \Delta \phi \|_{L^2(\Omega)}} \right] \quad , \quad H = \{\phi \mid \phi \in H^2(\Omega) \quad , \quad \frac{\partial \phi}{\partial n} = 0 \quad on \quad \Gamma\} \quad .$$

Then, for all $p^0 \in L^2(\Omega)$, *we have*

$$\lim_{n \to +\infty} \{\underset{\sim}{u}^n \ , \ p^n\} = \{\underset{\sim}{u}, p_0\} \quad in \quad (H^1(\Omega))^N \times L^2(\Omega) \quad , \tag{3.20}$$

where $\{\underset{\sim}{u}, p_0\}$ *is the solution of the Stokes problem* (3.12) *such that*

$$\int_\Omega p_0 \, dx = \int_\Omega p^0 \, dx \quad . \tag{3.21}$$

In fact, the convergence is linear since

$$\| \underset{\sim}{u}^n - \underset{\sim}{u} \|_{H^1(\Omega)^N} \quad and \quad \| p^n - p_0 \|_{L^2(\Omega)}$$

converge to zero at least as fast as geometric sequences whose ratio is less than one.

Using the *conjugate gradient* version of algorithm (3.14) - (3.17) described in, e.g. [18, Section 4] , the analogue of ρ is adjusted *automatically* at each iteration, making useless the calculation of c ; we obtain moreover a much faster convergence.

REMARK 3.1: It follows from [20] - [22] that if we assume that Ω is a hypercube of $\mathbb{R}^N$ and that we have *periodic boundary conditions* in (3.12) , (3.15) , (3.16) , then algorithm (3.17) converges in *one* iteration, for each $p^0 \in L^2(\Omega)$.

REMARK 3.2: The remark made in Section 3.4 concerning the solution of the linear systems associated to the discrete variants of the elliptic operator (3.11) still applies to (3.15). Therefore, to solve the discrete versions of (3.15) we shall use successive over-relation, or non-preconditioned conjugate gradient methods. In fact, our preferences go the over-relaxation methods since they are much easier to parallelize and/or vectorize. Unfortunately the same remark does not apply to the Neumann problem (3.16); for the discrete variants of this problem, we have been using a Cholesky factorization taking into account the fact that the matrix is of maximal rank minus one, and also the fact that in practice the pressure is approximated on a grid twice coarser than the velocity grid (see, again, [14], [18] for more details).

3.6. NUMERICAL EXPERIMENTS

Combining the numerical methods described in the above sections with the finite element approximations discussed in [14, Chapter 7] and [18, Section 5] we have been considering the following test problem (corresponding to a double jet in a square cavity).

Here $\Omega = (0,1)^2$, $\nu = 1/8000$, and the boundary conditions are the following

$$\underset{\sim}{g}(x_1 , x_2) = \underset{\sim}{0} \quad if \quad x_1 = 0 \quad or \quad 1 \ , \tag{3.22}_1$$

$$\begin{cases} \underset{\sim}{g}(x_1,1) = \underset{\sim}{0} \quad if \quad 0 \le x_1 \le 1/3 \ , \ 19/48 \ \le x_1 \le 29/48 \ , \\ \qquad\qquad\qquad\qquad 2/3 \ \le x_1 \le 1 \ , \\ \underset{\sim}{g}(x_1,1) = -1024 \left\{ \frac{1}{\sqrt{2}} \ , \ \frac{-1}{\sqrt{2}} \right\} (x_1 - 1/3)\,(19/48 - x_1) \quad if \\ \qquad\qquad \frac{1}{3} \le x_1 \le \frac{19}{48} \left(= \frac{1}{3} + \frac{1}{16} \right) \ , \\ \underset{\sim}{g}(x_1,1) = 1024 \left\{ \frac{-1}{\sqrt{2}} \ , \ \frac{-1}{\sqrt{2}} \right\} (2/3 - x_1)\,(x_1 - 29/48) \quad if \\ \qquad\qquad \frac{2}{3} \le x_1 \le \frac{29}{48} \left(= \frac{2}{3} - \frac{1}{16} \right) \ , \end{cases} \tag{3.22}_2$$

$$\begin{cases} \underset{\sim}{g}(x_1,0) = \underset{\sim}{0} \quad if \quad \frac{1}{16} \le x_1 \le \frac{15}{16} \ , \\ \underset{\sim}{g}(x_1,0) = -1024 \left\{ 0, \ \frac{1}{\sqrt{2}} \right\} x_1 \left(\frac{1}{16} - x_1 \right) \quad if \quad 0 \le x_1 \le \frac{1}{16} \ , \\ \underset{\sim}{g}(x_1,0) = -1024 \left\{ 0, \ \frac{1}{\sqrt{2}} \right\} (1 - x_1)\,(x_1 - \frac{15}{16}) \quad if \quad \frac{15}{16} \le x_1 \le 1 \ , \end{cases} \tag{3.22}_3$$

corresponding to *injection* of fluid by the upper apertures, and *ejection* by the two lower holes.

From (3.22) , we see that both apertures are 1/16 wide, that the two jets' inclinations are 45° , the left (resp. right) one being oriented toward the left (resp.

the right) wall. We can also see that the maximum injection velocity is one, and that the fluid is ejected from the cavity by two holes, located in the lower corners, whose width is also 1/16. Parbolic profiles of velocity have been assumed at all aperatures and holes.

Finally, we assume that the flow is initially at rest, i.e.

$$\underset{\sim}{u}(x,0) = \underset{\sim}{0} \quad \text{in} \quad \Omega \ . \tag{3.23}$$

From these characteristics, we can see that we actually need two Reynolds numbers (at least) to characterize this jet problem; indeed, if one takes the dimension of the jet apertures as characteristic length, we clearly have $\mathrm{Re} = \frac{8000}{16} = 500$, but if we consider the length of the cavity as another characteristic length the corresponding Re is now 8000; actually for the two upper corners we can also define a *local* Reynolds number of $8000/3 = 2666.66\ldots$, since $1/3$ is the distance of the apertures to the closest corner (and corresponding vertical wall).

Our goal with these numerical experiments is to simulate the bouncing of the jets on the closest vertical wall and to observe the development of the vortex pattern by visualization of the *streamlines* (the streamlines have been obtained as the contour lines of the *streamfunction* ψ , the solution of the Laplace equation

$$-\Delta\psi = \omega \ ,$$

completed by adequate Dirichlet boundary conditions (see [18, Section 6]) , with the *vorticity* ω defined by

$$\omega = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \ ) \quad .$$

Following [14, Chapter 7] and [18, Section 5] the *velocity* has been approximated by *continuous* functions, *piecewise linear* on a regular triangulation consisting

of $2 \times (128)^2$ triangles; the pressure has also been approximated by *continuous* and *piecewise linear* functions, but this time on a triangulation *twice coarser* than the velocity one. The total number of unknowns is then of the order of 32000 for the velocity and 4000 for the pressure. Concerning the time step, we have taken $\Delta t = 10^{-2}$ and used the θ-scheme (3.6) - (3.9) with $\theta = 1 - 1/\sqrt{2}$ and α, β given by (2.29).

We have shown on Figures 3.1 to 3.14 the streamlines corresponding to the computed solution at $t = 0.01, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 9.5, 10.0$, respectively. From those calculations we have been also able to follow the evolution in time of the *kinetic energy* $\frac{1}{2}\int_\Omega |\underline{u}|^2\, dx$ and of the *enstrophy* $\frac{1}{2}\int_\Omega |\omega|^2\, dx$; the corresponding results, together with further comments will be reported elsewhere. Numerical simulations done with smaller Δt give back the same numerical results.

All these calculations have been done on a CRAY-XMP 201.

4. APPLICATION TO LIQUID CRYSTAL CALCULATIONS.

We follow the presentation in [9] , [23] .

4.1. Formulation of the Problem.

"Imbedding" a *steady state* problem in a *time dependent* one is a well known method to solve the former one. A perfect illustration is given in this section where to a *nonconvex* variational problem originating from the mathematical theory of *liquid crystals* we associate a *nonlinear parabolic* "*equation*" which is solved by the operator splitting methods described in the above paragraphs.

Let Ω be a *bounded* domain of $\mathbb{R}^3$; we denote by Γ the boundary of Ω and we suppose that Γ is sufficiently smooth (Lipschitz continuous, for example). We define now

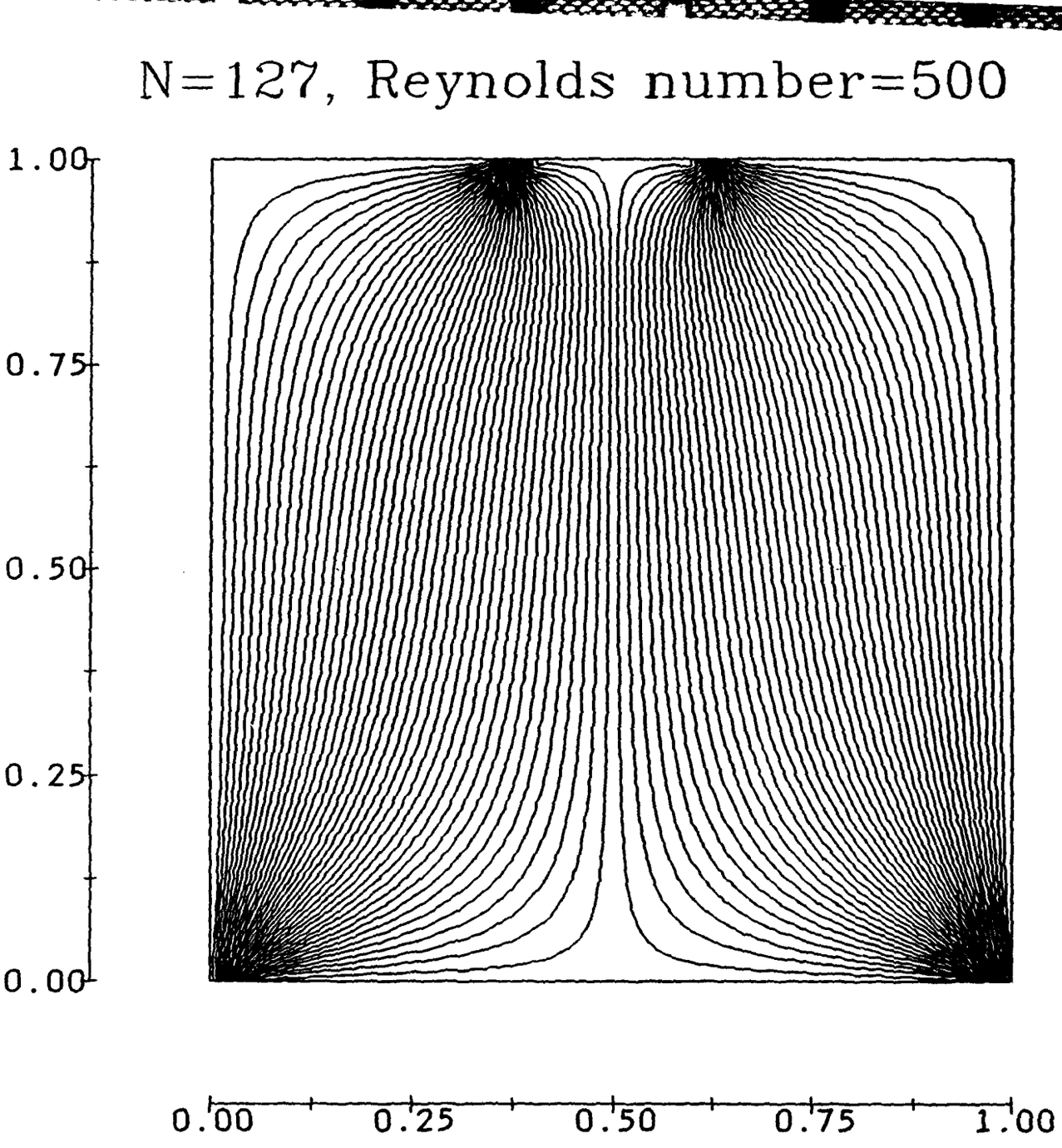


Figure 3.1 (t=0.01)

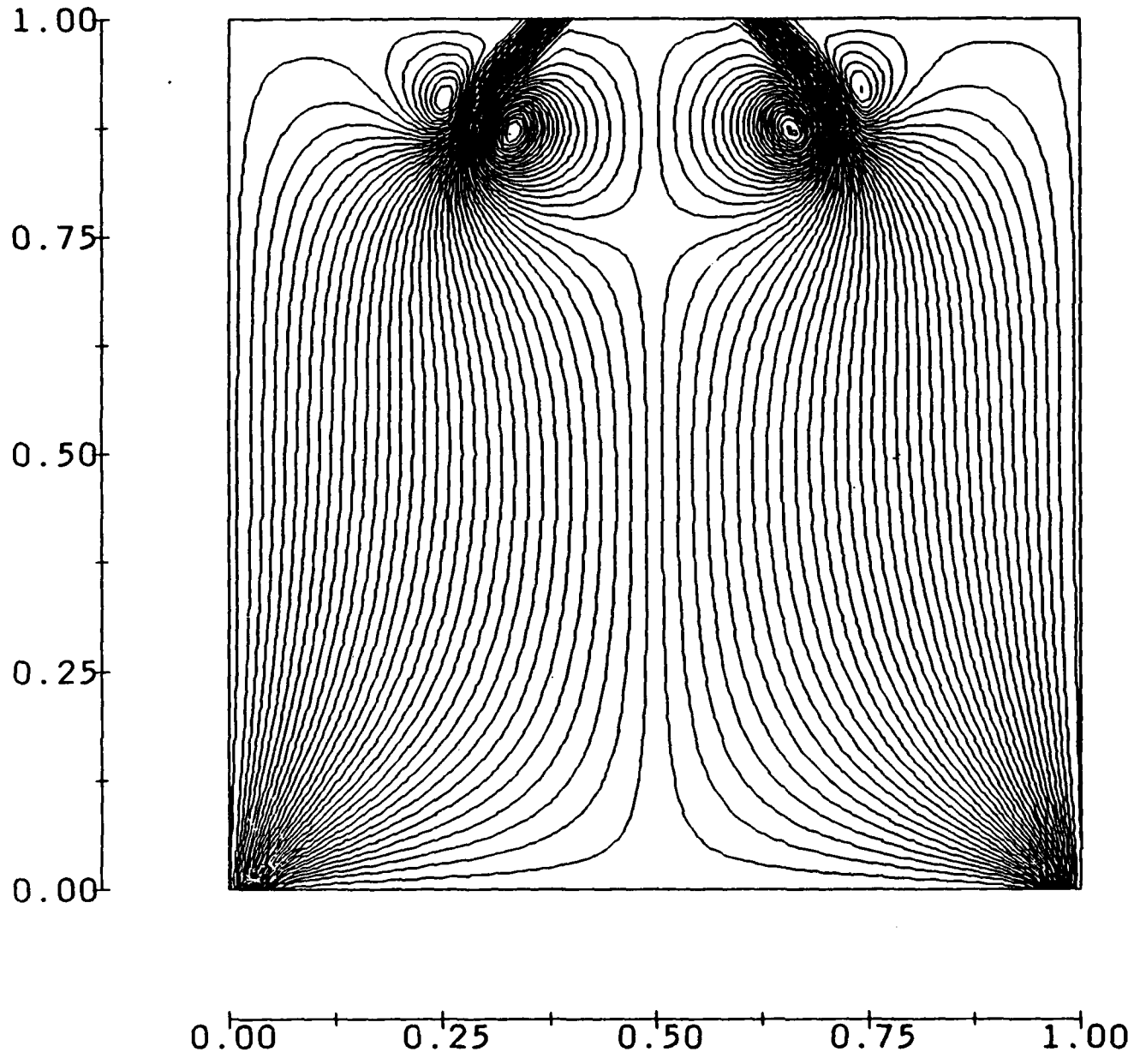

N=127, Reynolds number=500
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

Figure 3.2 (t=.5)

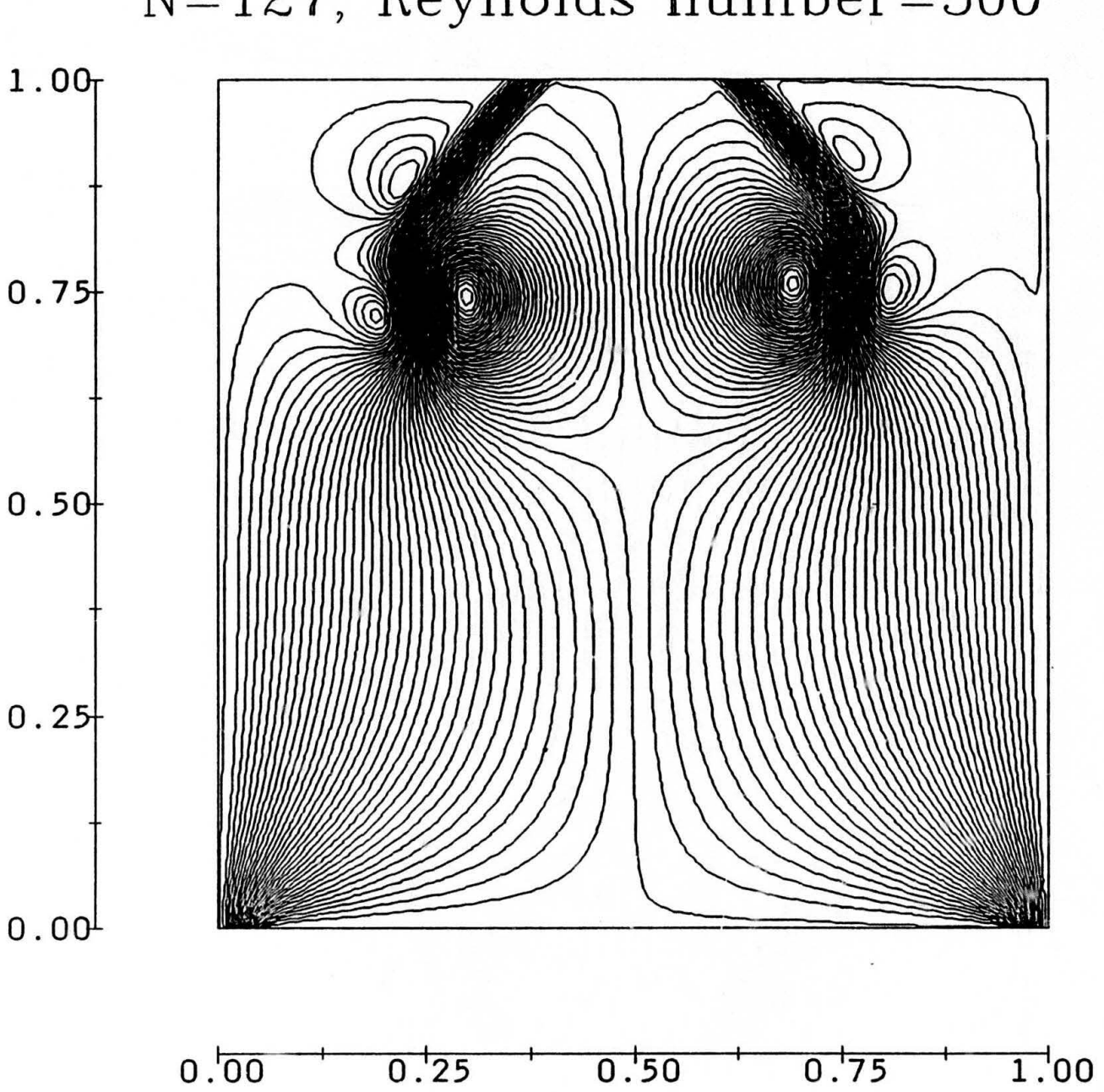

N=127, Reynolds number=500
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

Figure 3.3 (t=1)

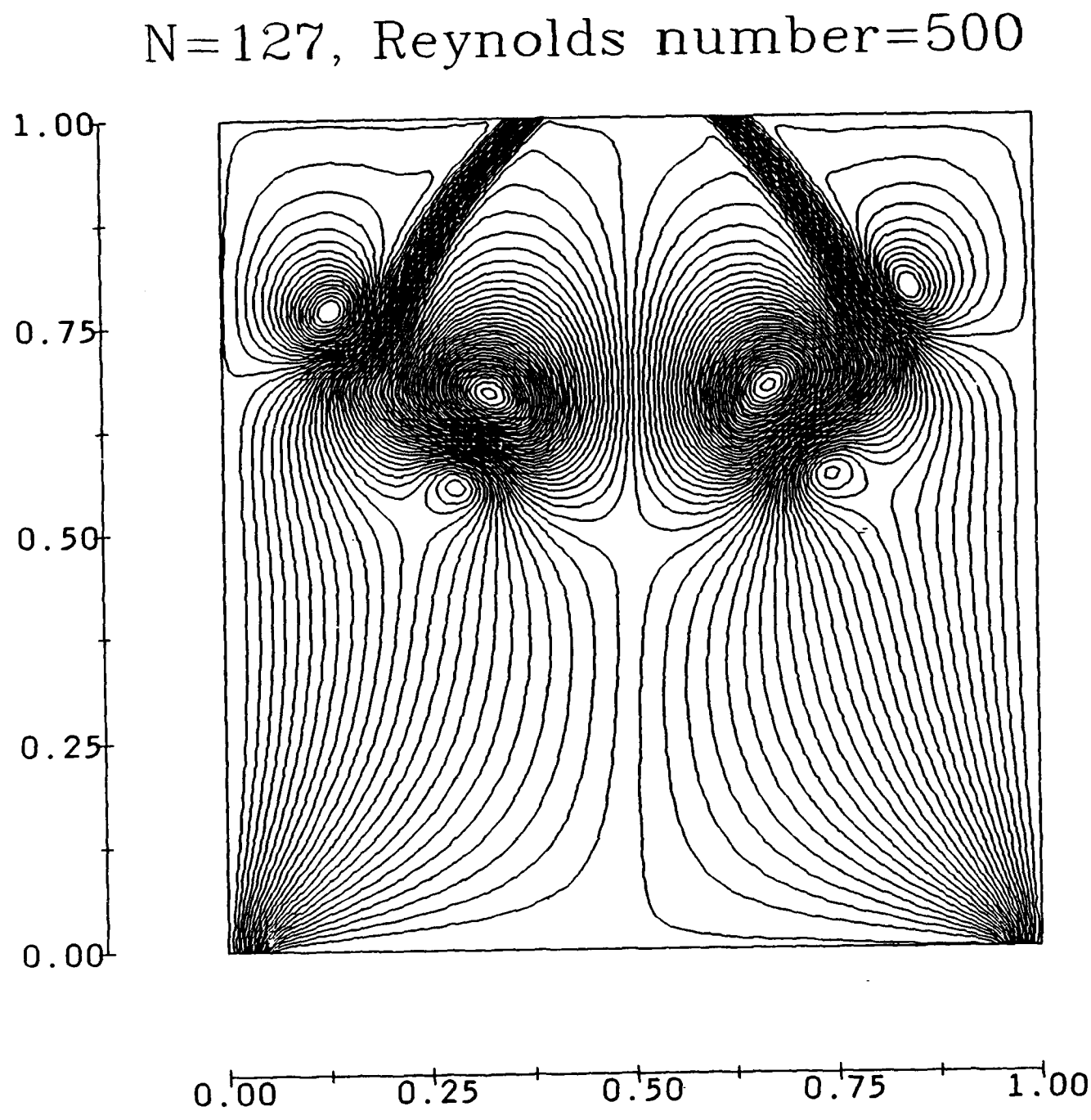


Figure 3.4 (t=1.5)

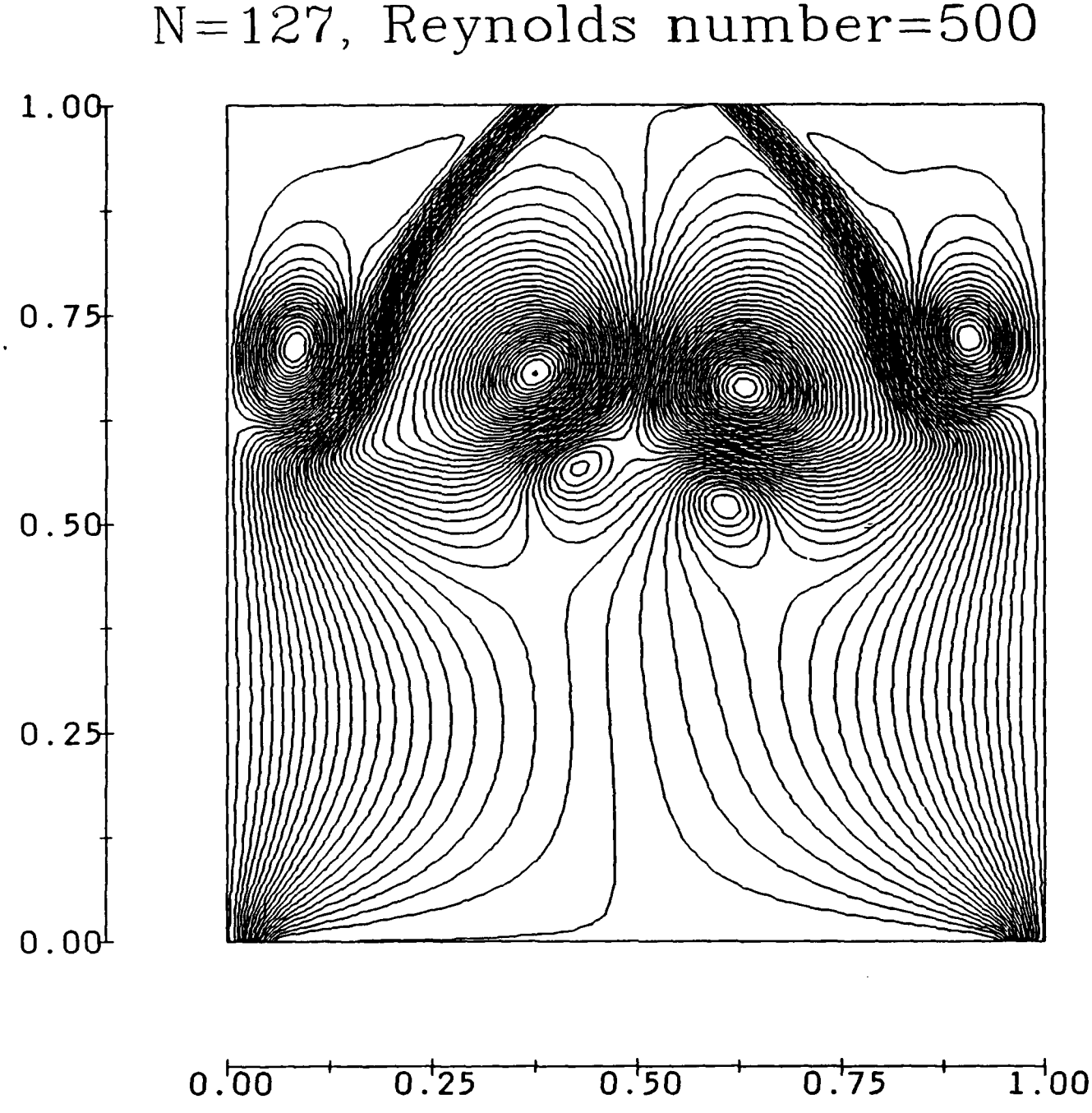

N=127, Reynolds number=500
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

Figure 3.5 (t=2)

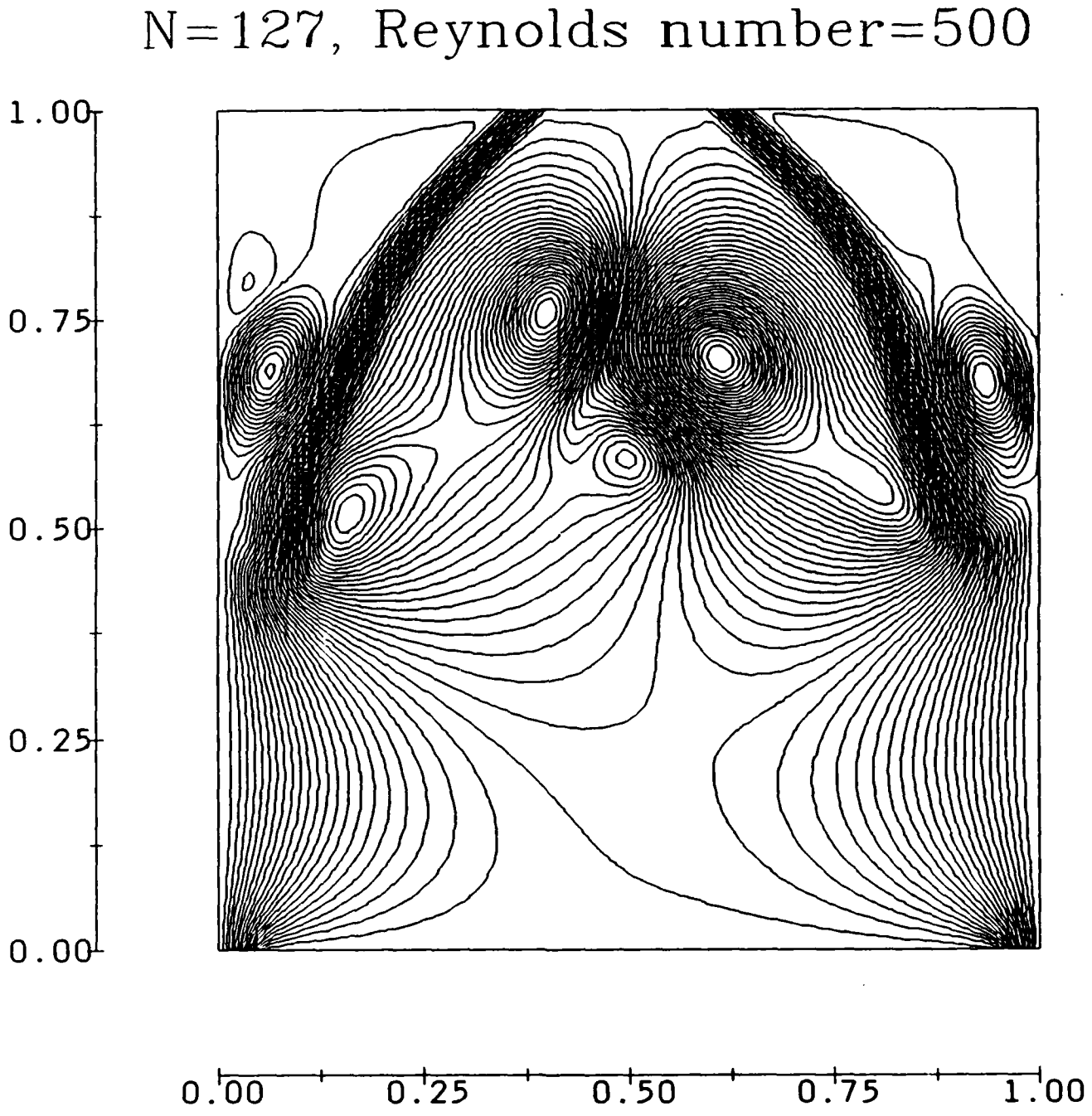


Figure 3.6 (t=2.5)

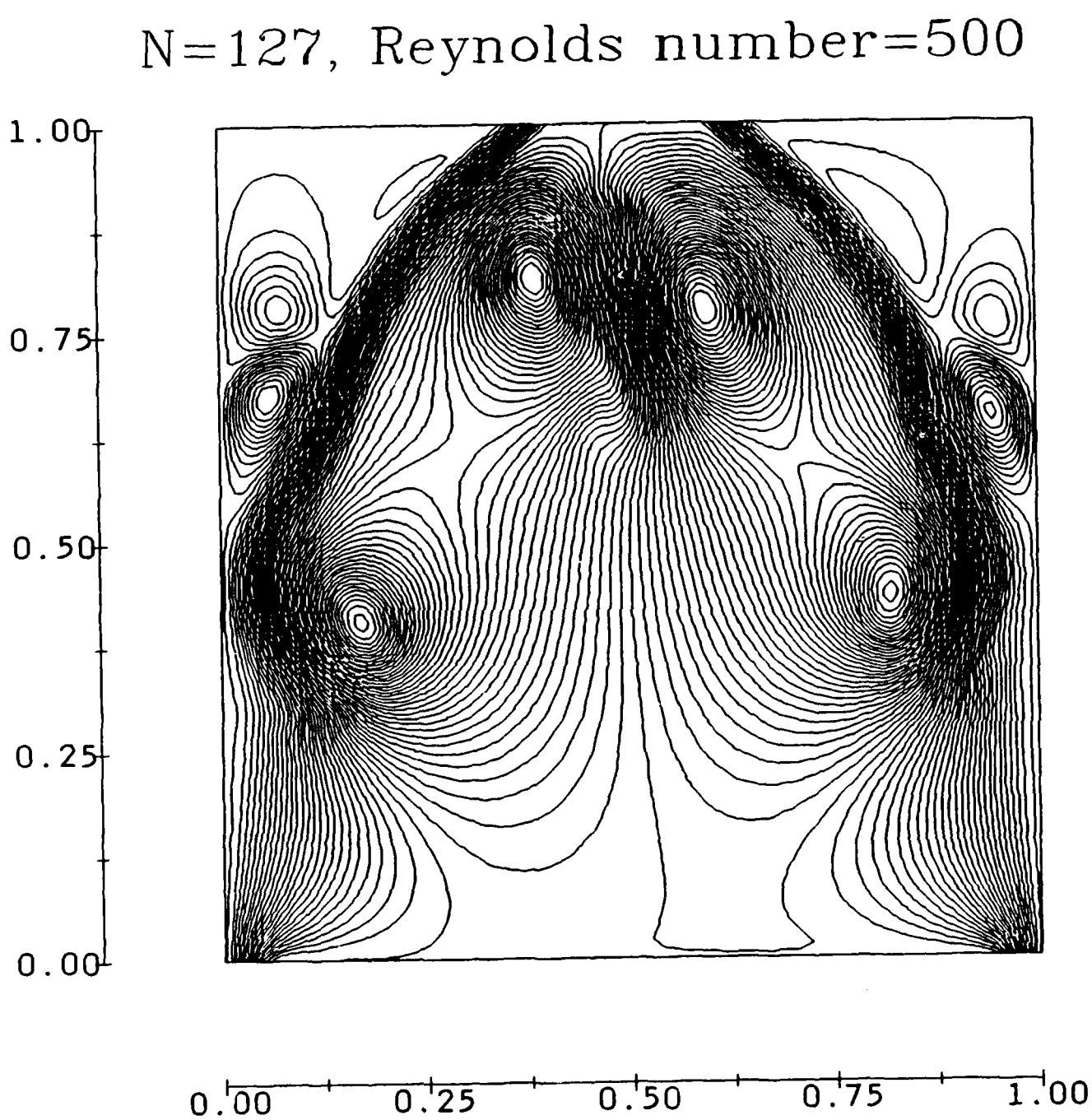

N=127, Reynolds number=500
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

Figure 3.7 (t=3)

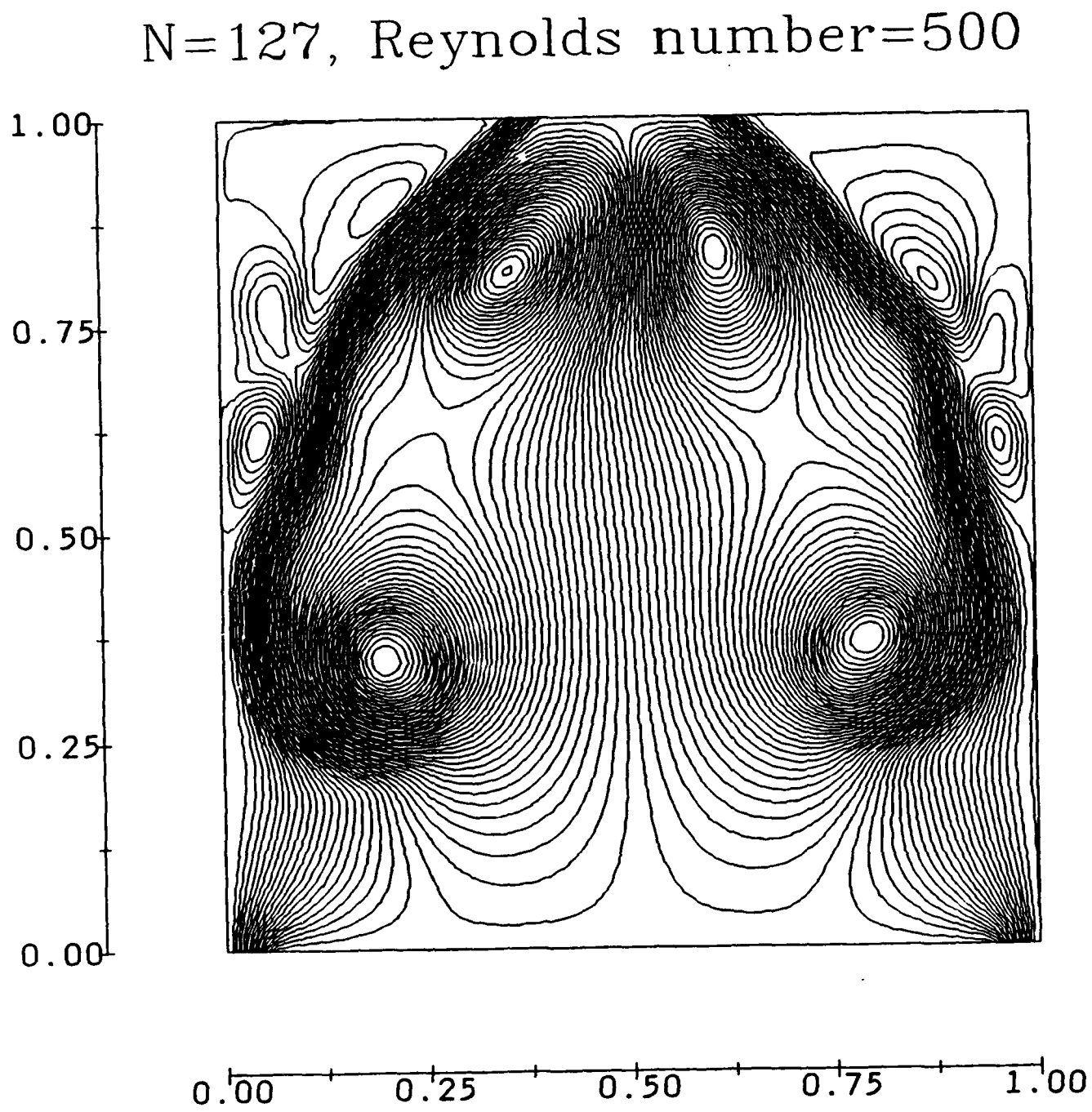


Figure 3.8 (t=3.5)

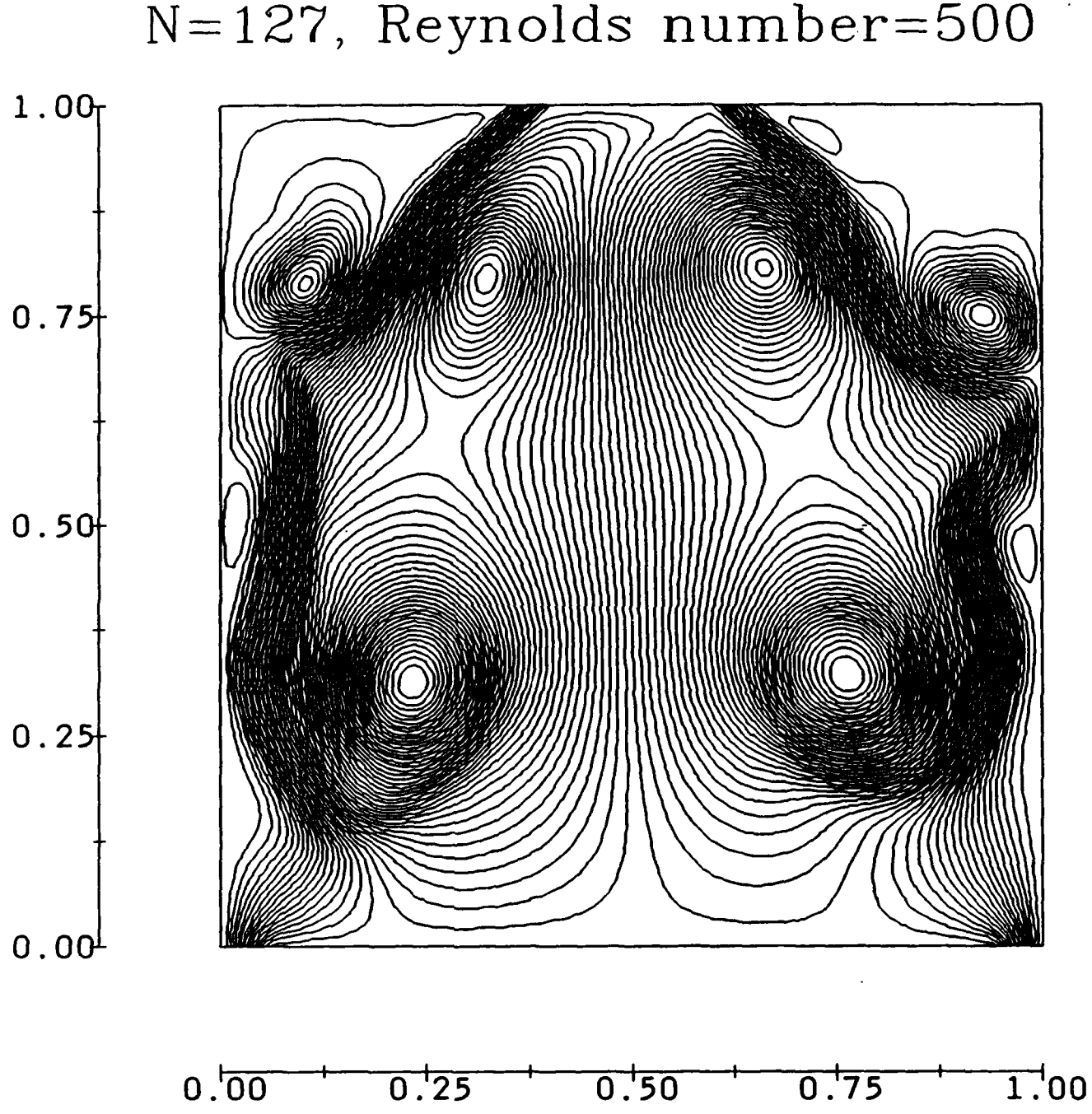

N=127, Reynolds number=500
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00

Figure 3.9 (t=4)

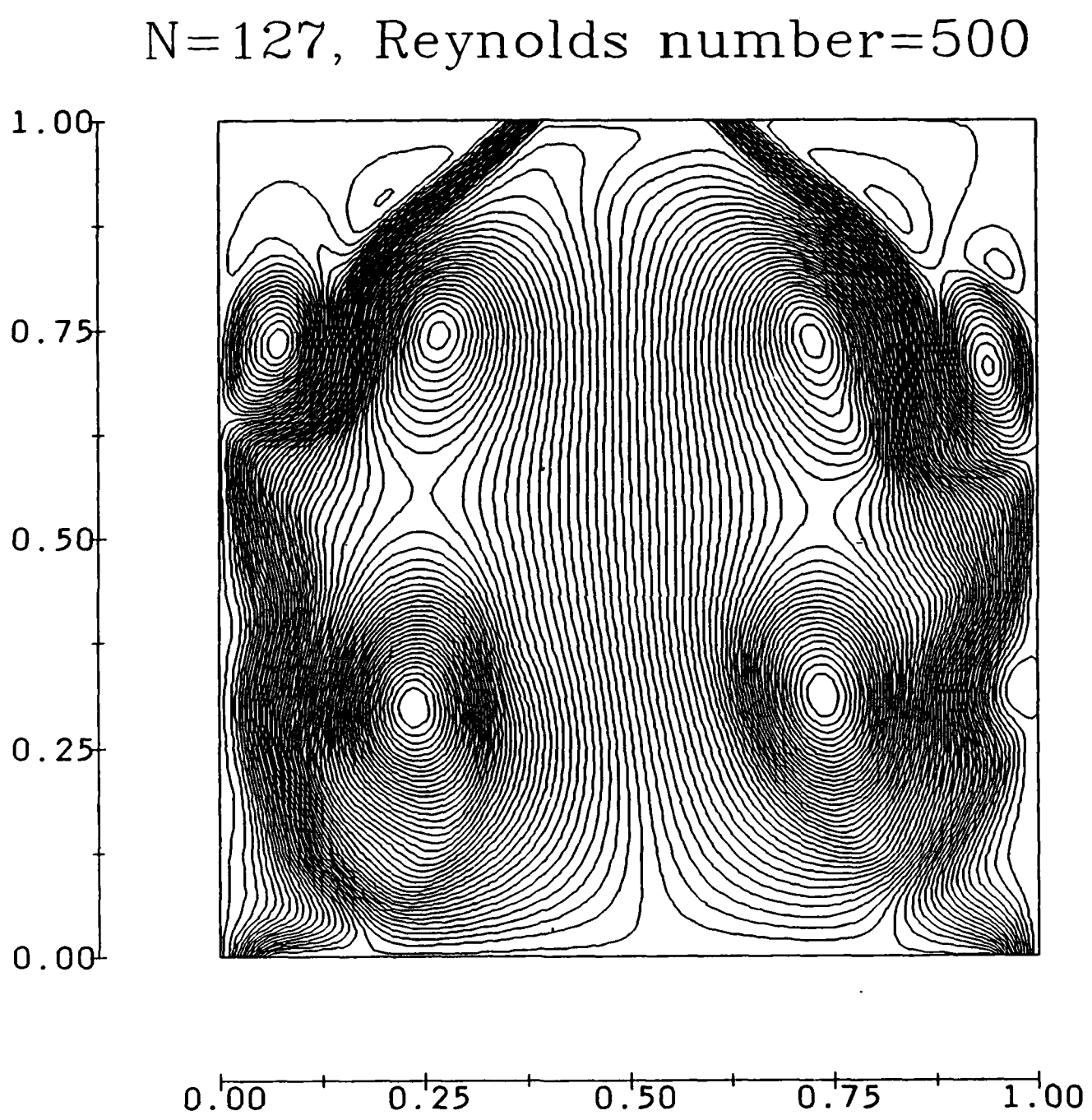


Figure 3.10 (t=4.5)

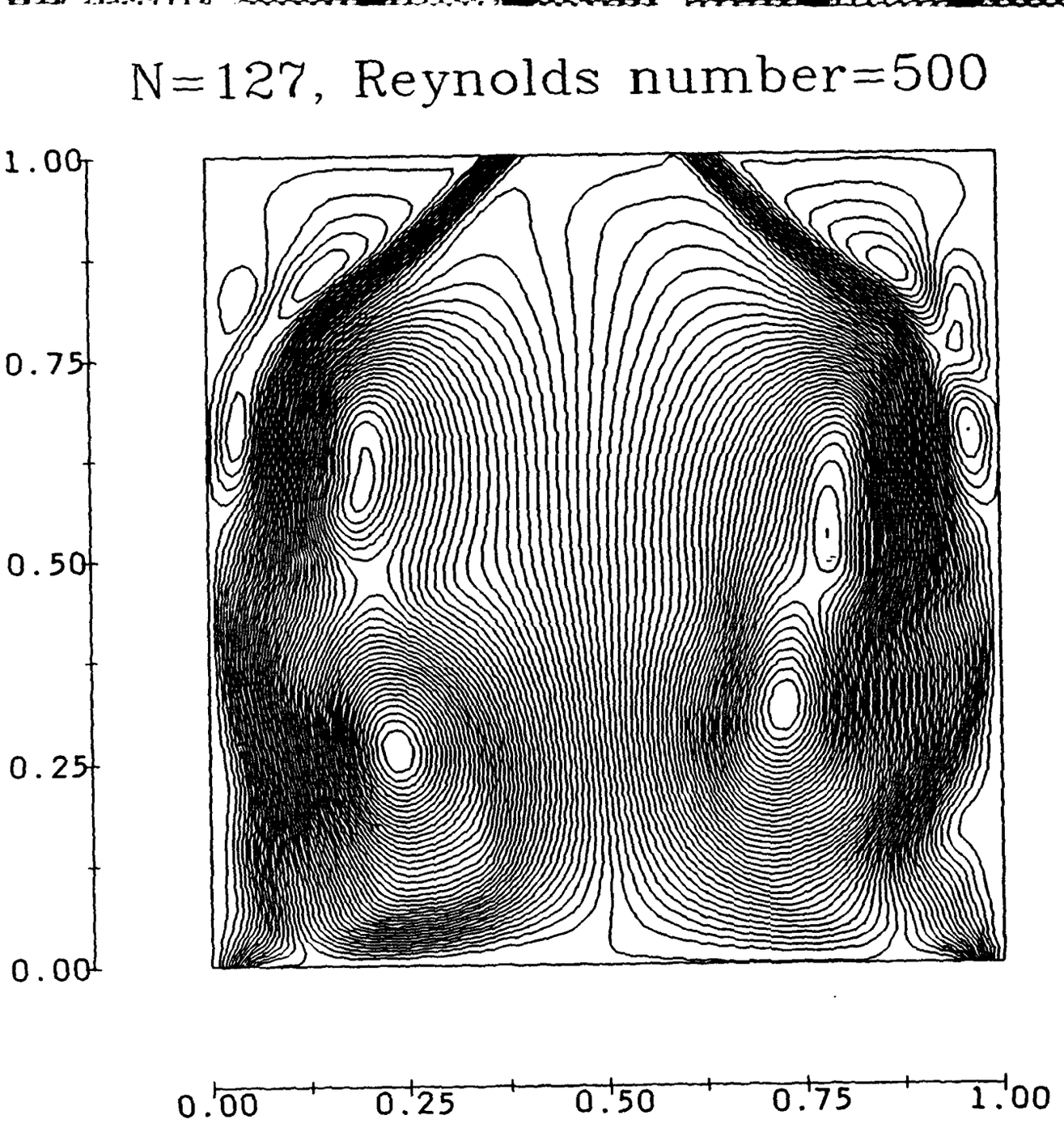


Figure 3.11 (t=5)

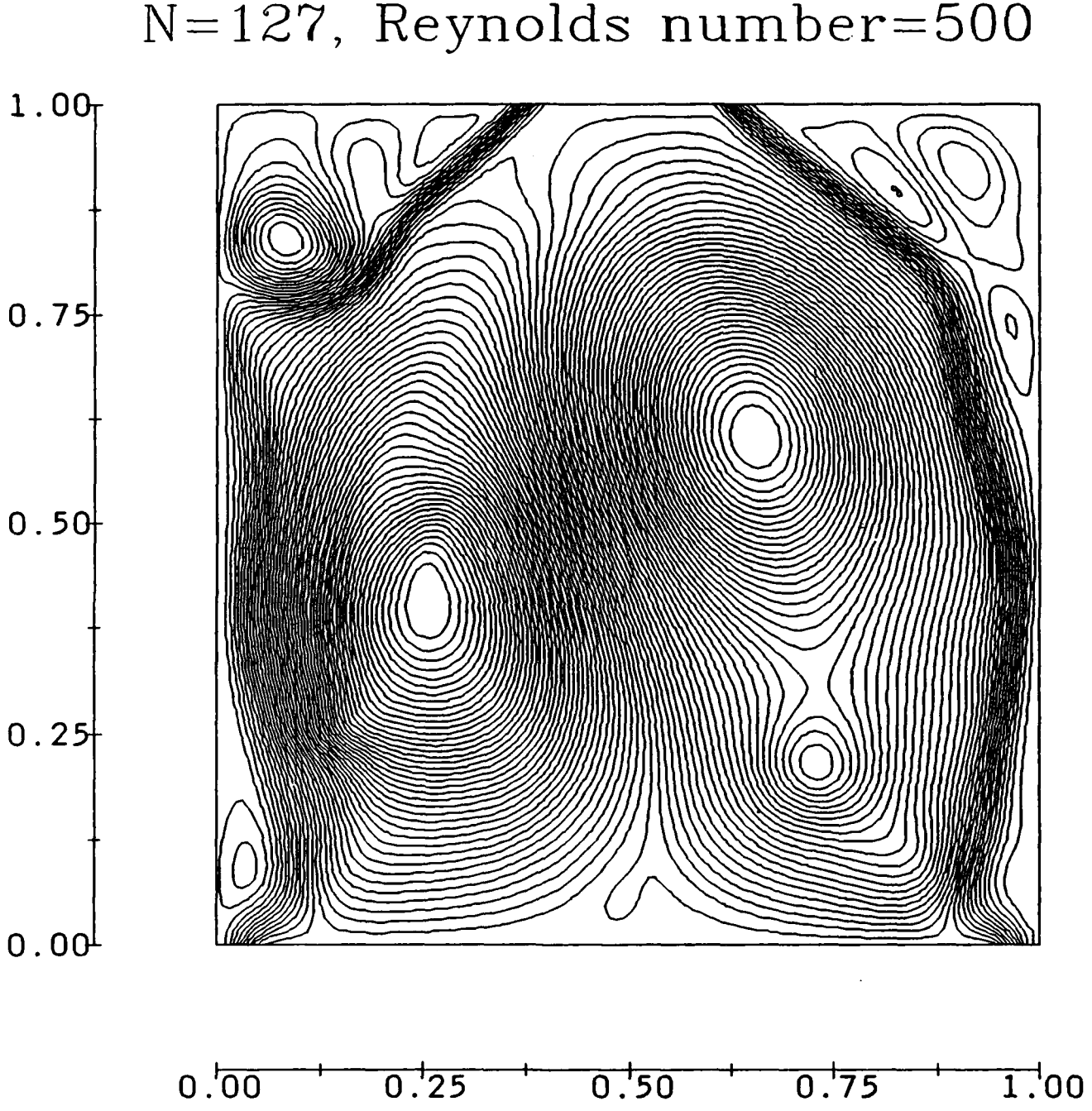


Figure 3.12 (t=9.5)

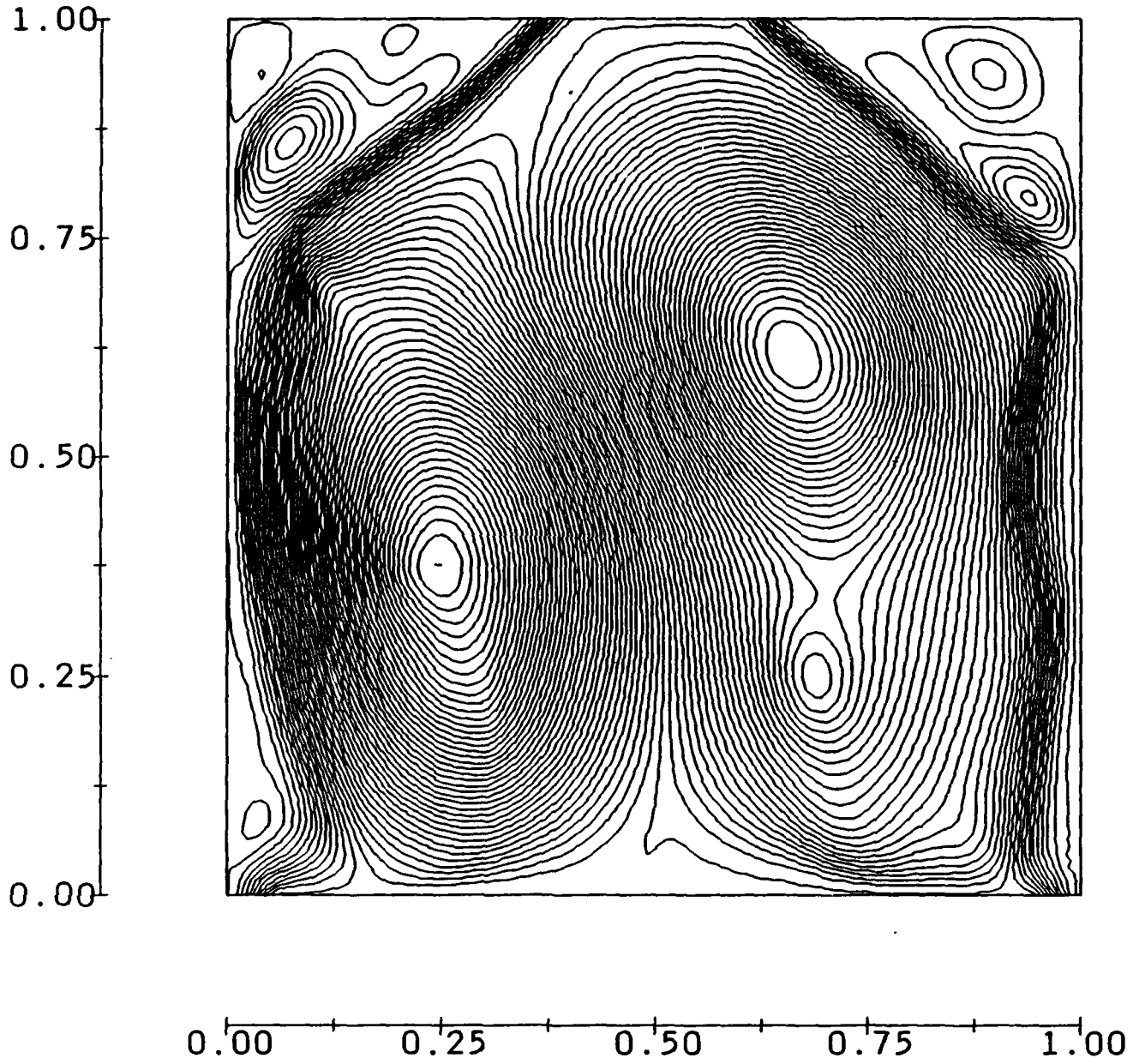


Figure 3.13 (t=10)

$$\underset{\sim}{H}^1(\Omega) = (H^1(\Omega))^3 ,$$

then, with $\underset{\sim}{v} = \{v_i\}_{i=1}^3 \in \underset{\sim}{H}^1(\Omega)$,

$$J(\underset{\sim}{v}) = \frac{1}{2}\int_\Omega |\underset{\sim}{\nabla v}|^2 dx \ \left(= \frac{1}{2}\sum_{i=1}^3 \int_\Omega |\nabla v_i|^2 dx \right) , \tag{4.1}$$

and finally

$$E = \{\underset{\sim}{v} \mid \underset{\sim}{v} \in \underset{\sim}{H}^1(\Omega) , \quad \underset{\sim}{v} = \underset{\sim}{g} \ \text{ on } \ \Gamma , \ |\underset{\sim}{v}(x)| = 1 \ \text{ a.e.}\} \tag{4.2}$$

(where $|\underset{\sim}{v}| = \left(\sum_{i=1}^3 v_i^2 \right)^{1/2}$); we suppose that $\underset{\sim}{g}$ is such that $E \neq \emptyset$.

Remark 4.1. Consider $\underset{\sim}{a} \in \mathbb{R}^3$ and define $\underset{\sim}{\phi}_a$ as the restriction to Ω of the function

$$\underset{\sim}{x} \longrightarrow \frac{\underset{\sim}{x} - \underset{\sim}{a}}{|\underset{\sim}{x} - \underset{\sim}{a}|} .$$

We clearly have $|\underset{\sim}{\phi}_a(\underset{\sim}{x})| = 1$ a.e.; furthermore, we can easily prove that $\underset{\sim}{\phi}_a \in \underset{\sim}{H}^1(\Omega)$ (even if $\underset{\sim}{a} \in \bar{\Omega}$).

We consider now the following minimization problem:

Find $\underset{\sim}{u} \in E$ *such that* $J(\underset{\sim}{u}) \leq J(\underset{\sim}{v})$ *for all* $\underset{\sim}{v} \in E$. (4.3)

Using the fact that E is *weakly closed* in $\underset{\sim}{H}^1(\Omega)$, we can easily prove that problem (4.3) has at least one solution; further mathematical properties of (4.3) are discussed in [24] , [25]. Problem (4.3) is associated to the mathematical modeling of interesting physical phenomena (as discussed in the Section 1 of [25]) , some of them occurring in the physics of *liquid crystals* (see [26] - [28] for further information on liquid crystals).

4.2. Numerical Solution of Problem (4.3).

At first glance, problem (4.3) seems to be a nontrivial problem of the Calculus of Variations. In fact, the solution of (4.3) is quite easy to achieve by the *operator splitting* methods of Section 2. This follows indeed from the fact that problem (4.3) is equivalent to

$$\begin{cases} \textit{Find } \underset{\sim}{u} \in \underset{\sim}{H}^1_g \textit{ such that} \\ J(\underset{\sim}{u}) + I\ (\underset{\sim}{u}) \leq J(\underset{\sim}{v}) + I\ (\underset{\sim}{v}) \textit{ for all } \underset{\sim}{v} \in \underset{\sim}{H}^1_g , \end{cases} \tag{4.4}$$

where (with $\underset{\sim}{L}^2(\Omega) = (L^2(\Omega))^3$)

$$\underset{\sim}{H}^1_g = \{\underset{\sim}{v} | \underset{\sim}{v} \in \underset{\sim}{H}^1(\Omega) , \quad \underset{\sim}{v} = \underset{\sim}{g} \quad \text{on} \quad \Gamma\} ,$$

$$\Sigma = \{\underset{\sim}{v} | \underset{\sim}{v} \in \underset{\sim}{L}^2(\Omega) , |\underset{\sim}{v}(x)| = 1 \quad \text{a.e.}\}$$

and where $I : \underset{\sim}{L}^2(\Omega) \rightarrow \mathbb{R} \cup \{+\infty\}$ is defined by

$$I\ (\underset{\sim}{v}) = \begin{cases} 0 \text{ if } \underset{\sim}{v} \in \Sigma , \\ +\infty \text{ if } \underset{\sim}{v} \notin \Sigma . \end{cases}$$

Using the notation of the above section, we have for (4.4) the following *Euler Lagrange "equation"*

$$\begin{cases} -\Delta \underset{\sim}{u} + \partial I\ (\underset{\sim}{u}) = \underset{\sim}{0} \textit{ in } \Omega , \\ \underset{\sim}{u} = \underset{\sim}{g} \textit{ in } \Gamma , \end{cases} \tag{4.5}$$

where $\partial I_\Sigma\ (\underset{\sim}{u})$ is the "gradient" of I_Σ at $\underset{\sim}{u}$. We associate next to the *nonlinear elliptic equation* (4.5) the *nonlinear parabolic problem*

$$\begin{cases} \dfrac{\partial \underset{\sim}{u}}{\partial t} - \Delta \underset{\sim}{u} + \partial I\ (\underset{\sim}{u}) = \underset{\sim}{0} \textit{ in } \Omega , \\ \underset{\sim}{u} = \underset{\sim}{g} \textit{ on } \Gamma , \\ \underset{\sim}{u}(0) = \underset{\sim}{u}_0 . \end{cases} \tag{4.6}$$

Concentrating on the θ-scheme (2.15) - (2.18) (since it appears as the most efficient method here) we obtain the following algorithm:

$$\underset{\sim}{u}^0 = \underset{\sim}{u}_0 \ , \quad \textit{given in} \quad \underset{\sim}{H}^1_g \ ; \tag{4.7}$$

then for $n>0$, $\underset{\sim}{u}^n$ *being known, we compute* $\underset{\sim}{u}^{n+\theta}$, $\underset{\sim}{u}^{n+1-\theta}$, $\underset{\sim}{u}^{n+1}$ *as follows:*

$$\frac{\underset{\sim}{u}^{n+\theta} - \underset{\sim}{u}^n}{\theta\Delta t} - \Delta\underset{\sim}{u}^n + \partial I_\Sigma\ (\underset{\sim}{u}^{n+\theta}) = \underset{\sim}{0} \ , \tag{4.8}$$

$$\begin{cases} \dfrac{\underset{\sim}{u}^{n+1-\theta} - \underset{\sim}{u}^{n+\theta}}{(1-2\theta)\Delta t} - \Delta\underset{\sim}{u}^{n+1-\theta} + \partial I_\Sigma\ (\underset{\sim}{u}^{n+\theta}) = \underset{\sim}{0} \ , \\ \\ \underset{\sim}{u}^{n+1-\theta} = \underset{\sim}{g} \quad \textit{on} \quad \Gamma \ , \end{cases} \tag{4.9}$$

$$\frac{\underset{\sim}{u}^{n+1} - \underset{\sim}{u}^{n+1-\theta}}{\theta\Delta t} - \Delta\underset{\sim}{u}^{n+1-\theta} + \partial I_\Sigma\ (\underset{\sim}{u}^{n+1}) = \underset{\sim}{0} \ . \tag{4.10}$$

When using algorithm (4.7) - (4.10) for practical calculations one has to give a sense to the two *multivalued* equations (4.8) and (4.10). The interpretation given to (4.8) is

$$\begin{cases} \underset{\sim}{u}^{n+\theta} \in \Sigma \quad ; \quad \underset{\sim}{u}^{n+\theta} \ \textit{minimizes over} \ \Sigma \ \text{ the functional} \\ \underset{\sim}{v} \rightarrow \frac{1}{2}\int_\Omega |\underset{\sim}{v}|^2\, dx - \int_\Omega (\underset{\sim}{u}^n + \theta\Delta t\Delta\underset{\sim}{u}^n)\cdot\underset{\sim}{v}\, dx. \end{cases} \tag{4.11}$$

The solution of problem (4.11) is clearly given by

$$\underset{\sim}{u}^{n+\theta} = \frac{\underset{\sim}{u}^n + \theta\Delta t\Delta\underset{\sim}{u}^n}{|\underset{\sim}{u}^n + \theta\Delta t\Delta\underset{\sim}{u}^n|} \ . \tag{4.12}$$

Similarly, the solution of (4.10) is given by

$$\underset{\sim}{u}^{n+1} = \frac{\underset{\sim}{u}^{n+1-\theta} + \theta\Delta t\Delta \underset{\sim}{u}^{n+1-\theta}}{|\underset{\sim}{u}^{n+1-\theta} + \theta\Delta t\Delta \underset{\sim}{u}^{n+1-\theta}|} \quad . \tag{4.13}$$

Once $\underset{\sim}{u}^{n+1}$ is known, we obtain $\partial l \quad (\underset{\sim}{u}^{n+1-\theta})$ from (4.8) and we use that information in (4.9) to compute $\underset{\sim}{u}^{n+1-\theta}$ via the solution of a *Dirichlet problem* for the *elliptic operator*

$$\underset{\sim}{v} \rightarrow \underset{\sim}{v} - (1 - 2\theta)\Delta t\Delta \underset{\sim}{v}.$$

From these observations, the only costly step of algorithm (4.7) - (4.10) is the Dirichlet problem (4.9); in fact, since in practice Δt has to be small, the discrete variants of the above elliptic operator are well conditioned matrices for which *relaxation* (and *over-relaxation*) *methods* are very efficient (see [9] , [23] for more details).

4.3. NUMERICAL EXPERIMENTS.

The numerical techniques decribed in Section 4.2 have been applied to the solution of various test problems in [9] , [23] (see also [20] for related numerical experiments). In this paper we shall only consider the test problem for which

$$\Omega = (0,1)^3 \tag{4.14}$$

and

$$\underset{\sim}{g} = \underset{\sim}{\phi}_a|_\Gamma \ , \tag{4.15}_1$$

$$\underset{\sim}{\phi}_a(\underset{\sim}{x}) = \frac{\underset{\sim}{x} - \underset{\sim}{a}}{|\underset{\sim}{x} - \underset{\sim}{a}|} \ , \quad \underset{\sim}{a} = (.5, .5, .5) \ . \tag{4.15}_2$$

It follows from [25] that if $\underset{\sim}{g}$ is defined by (4.15) , then problem (4.3) has a *unique solution* which is precisely given by

$$\underset{\sim}{u} = \underset{\sim}{\phi}_a|_\Omega \ . \tag{4.16}$$

From the simplicity of Ω, it is quite convenient to approximate problem (4.3) by a *finite difference* method such as the one described below.

Let N be a positive integer; we define a space discretization step h by h = 1 / N+1 and then the discrete set

$$\{M_{ijk}\}_{0 \le i,\, j,\, k \le N+1}, \quad \text{with} \quad M_{ijk} = (ih,\, jh,\, kh) \ .$$

With $\underset{\sim}{v}_h = (\{v^l_{ijk}\}_{l=1}^3)_{0 \le i,j,k \le N+1}$, we approximate $J(\underset{\sim}{v})$ by

$$\left\{ J_h(\underset{\sim}{v}_h) = \frac{h}{4}^3 \sum_{l=1}^{3} \sum_{1 \le i,j,k \le N} \left\{ \left| \frac{v^l_{i+1\,jk} - v^l_{ijk}}{h} \right|^2 + \left| \frac{v^l_{i-1\,jk} - v^l_{ijk}}{h} \right|^2 + \right. \right.$$

$$\left| \frac{v^l_{ij+1k} - v^l_{ijk}}{h} \right|^2 + \left| \frac{v^l_{ij-1k} - v^l_{ijk}}{h} \right|^2 + \tag{4.17}$$

$$\left. \left. \left| \frac{v^l_{ijk+1} - v^l_{ijk}}{h} \right|^2 + \left| \frac{v^l_{ijk-1} - v^l_{ijk}}{h} \right|^2 \right\} \right.,$$

and then E by

$$\left\{ E_h = \{\underset{\sim}{v}_h \mid \sum_{l=1}^{3} |v^l_{ijk}|^2 = 1 \ , \quad 1 \le i,\, j,\, k \le N; \right.$$

$$\left. \underset{\sim}{v}_{ijk} = \underset{\sim}{g}_{ijk} \quad \text{for all} \quad M_{ijk} \in \Gamma \right\} \ . \tag{4.18}$$

Finally, problem (4.3) is approximated by:

Find $\underset{\sim}{u}_h \in E_h$ *such that* $J_h(\underset{\sim}{u}_h) \leq J_h(\underset{\sim}{v}_h)$ *for all* $\underset{\sim}{v}_h \in E_h$. (4.19)

Applying the θ-scheme discussed in Section 4.2 is quite easy since the finite dimensional problem (4.19) has the same structure as (4.3).

All the calculations have been *initialized* by $\underset{\sim}{u}_h^0$, the *finite difference* approximation of the solution $\underset{\sim}{u}^0$ of the Dirichlet problem

$$\begin{cases} -\Delta \underset{\sim}{u}^0 = \underset{\sim}{0} \quad in \quad \Omega \, , \\ \underset{\sim}{u}^0 = \underset{\sim}{g} \quad on \quad \Gamma \, . \end{cases} \tag{4.20}$$

As convergence criteria, we have used (with obvious notation)

$$\frac{(h^3 \sum_{1 \leq i,j,k \leq N} |\underset{\sim}{u}_{ijk}^{n+1} - \underset{\sim}{u}_{ijk}^{n}|^2)^{1/2}}{(h^3 \sum_{1 \leq i,j,k \leq N} |\underset{\sim}{u}_{i,j,k}^{n+1}|^2)^{1/2}} \leq 10^{-4} \, . \tag{4.21}$$

Since the exact solution of problem (4.3) is known here (and is given by (4.15) , (4.26)) we can accurately estimate the $L^2(\Omega)$-norm of the approximation error; we have chosen as estimator of the $L^2(\Omega)$-error the quantity ρ_h defined by

$$\rho_h = (h^3 \sum_{1 \leq i,j,k \leq N} |\underset{\sim}{u}(M_{ijk}) - \underset{\sim}{u}_{ijk}|^2)^{1/2} \tag{4.22}$$

(we took here $\underset{\sim}{u}(1/2, 1/2, 1/2) = \{1/\sqrt{3}, \ 1/\sqrt{3} \, , \ 1/\sqrt{3}\}$) .

Using a discrete variant of algorithm (4.7) - (4.10) in which the elliptic problems (4.9) and (4.10) are solved by the *Gauss-Seidel* method we need 8 time steps to reach a steady state if one takes $h = 1/20$, $\theta = .1$ and $\Delta t = 1/200$; the corresponding CPU time is 11mn 18s on a VAX 11/780. We have then

$\rho_h = 0.39 \times 10^{-1}$. When using the Gauss-Seidel method we have initialized the calculation of $\underset{\sim}{u}_h^2$ by $\underset{\sim}{0}$ and then the calculation of $\underset{\sim}{u}_h^{n+1-\theta}$ by $\underset{\sim}{u}_h^{n+\theta}$, taking as stopping criterion one similar to (4.21) but with $\varepsilon = 10^{-6}$. In Table 4.1 we show, at each time step, the number of Gauss-Seidel iterations necessary to converge according to the above test.

Step	G. S. iterations
1	26
2	4
3	2
4 to 8	1

TABLE 4.1

Variation of the number of Gauss-Seidel iterations with the time step.

If instead of the Gauss-Seidel method, we use an *over-relaxtion* one with the optimal paramter ω the performance of our computational method is dramatically improved (particularly for the first time step), and for the above test problem (with the same values of θ, h, and Δt) we have convergence in 5 time steps (instead of 8), the CPU time being reduced to 3mn 44s (instead of 11mn 18s); actually, the L^2-error is also substantially reduced since we have now $\rho_h = 0.23 \times 10^{-1}$ (instead of 0.39×10^{-1}).

Table 4.2, just below, shows the variations of the L^2-error ρ_h as a function of h ; these results have been obtained using over-relaxation instead of the

Gauss-Seidel method. The case $h = 1/40$ has been computed on the CRAY-XMP 201 , taking approximately 12 seconds (the discrete problem (4.19) involves then 1.8×10^5 unknowns, approximately).

h	Δt	θ	ρ_h
1/10	1/100	1/10	0.74×10^{-1}
1/20	1/200	1/10	0.23×10^{-1}
1/40	1/2000	1/5	0.12×10^{-1}

TABLE 4.2

Variation of the L^2-error with h.

The results in Table 4.2 suggest that the L^2-approximation error is in $0(h)$ at best; the analysis of such an error is an interesting problem in itself.

4.4. FURTHER COMMENTS.

Relaxation methods for solving problem (4.19) are discussed in [30] ; they appear to be quite efficient, one of the reasons for such an efficiency being the fact that the *quadratic constraint*

$$|\underset{\sim}{v}(x)|^2 = 1 \quad \text{a.e.} \tag{4.23}$$

is a fairly simple one since it does not involve derivatives of $\underset{\sim}{v}$. Suppose now that instead of (4.23) we have to deal with

$$\det (\underset{\sim}{I} + \underset{\sim}{\nabla} \underset{\sim}{v}(x)) = 1 \quad \text{a.e.} \tag{4.24}$$

(nonlinear constraints such as (4.24) occur in *incompressible finite elasticity*; cf. [9] and the references therein). Relaxation methods cannot be applied any longer, at least directly. On the contrary, operator splitting techniques like those discussed in Section 2 still apply; see [9] for more details, further comments and numerical results concerning the treatment of nonlinear constraints such as (4.24).

5. NUMERICAL SOLUTION OF ADVECTION-DIFFUSION PROBLEMS IN HIGH-DIMENSION.

5.1. MOTIVATION. SYNOPSIS.

Let's consider the following *stochastic ordinary differential equation*

$$d\underset{\sim}{X} = \underset{\sim}{V}(\underset{\sim}{x})\, dt + d\underset{\sim}{W} \tag{5.1}$$

where $\underset{\sim}{X}$ is an N-dimensional vector and $\underset{\sim}{W}$ a *noise*. Assuming convenient hypotheses on the noise $\underset{\sim}{W}$, the probability of finding at time t the *state vector* $\underset{\sim}{X}$, in a neighborhood of $x \in \mathbb{R}^N$ of measure $dx = dx_1, \ldots, dx_N$, is $p(x,t)\, dx$ where the *probability density* p satisfies a *parabolic equation*. In the particular case where $\underset{\sim}{V}$ is *divergence free*, i.e.

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{V} = 0 \tag{5.2}$$

and for simple noise models, this parabolic equation reduces to

$$\frac{\partial p}{\partial t} - \varepsilon \underset{\sim}{\nabla}^2 p + \underset{\sim}{V} \cdot \underset{\sim}{\nabla} p = 0 . \tag{5.3}$$

An interesting (and difficult) case is the one where the level of noise is "weak" implying that ε is "small" ; we have then an *advection dominated advection-diffusion equation.*

The solution of such equations plays a fundamental role in the implementation of some solution methods of the *Zakai equation* occurring in *Stochastic Optimal Control.*

In the sequel we shall consider the following *initial boundary value problem*

$$\frac{\partial p}{\partial t} - \varepsilon \underset{\sim}{\nabla}^2 p + \underset{\sim}{V}\cdot\underset{\sim}{\nabla} p = f \quad in \quad \Omega \times (0,T) , \tag{5.4}_1$$

$$p = g \quad on \quad \partial\Omega \times (0,T) , \tag{5.4}_2$$

$$p(x,0) = p_0(x) \quad in \quad \Omega , \tag{5.4}_3$$

with $\Omega \subset \mathbb{R}^N$.

For solving such problems, the numerical analyst has to face two outstanding difficulties, namely

(i) When ε is *small,* the problem is *advection dominated,*

(ii) For practical problems, we usually have $N>3$.

In the following sections, we shall describe for $N = 2$ to 6 the solution of (5.4) by various *upwinding methods* and by the *modified* (i.e. backward) *method of characteristics.* Numerical results will be presented for the particular case where $\Omega = (0,1)^N$.

5.2. SOLUTION OF PROBLEM (5.4) BY FINITE DIFFERENCES AND UPWINDING METHODS.

We consider for simplicity the case where $\Omega = (0,1)^N$ with $N=2$; the extension to $N>2$ is straightforward. With I a *positive integer* , we

define h by $h=1/I+1$ and consider over $\bar{\Omega} = \Omega \cup \partial\Omega$ the discretization points

$$M_{ij} = \{ih\ ,\ jh\}\ ;\quad 0 \leq i,j \leq I+1\ . \tag{5.5}$$

At the points M_{ij} *interior to* Ω (i.e., $1 \leq i,j \leq I$) we approximate (5.4) by the following *finite difference scheme* (with $\underline{V} = (V_1\ ,\ V_2)$):

$$\left\{\begin{array}{l} \dfrac{p_{ij}^{n+1} - p_{ij}^{n}}{\Delta t} - \varepsilon\, \dfrac{p_{i+1j}^{n+1} + p_{i-1j}^{n+1} + p_{ij+1}^{n+1} + p_{ij-1}^{n+1} - 4p_{ij}^{n+1}}{h^2} \\ \\ + V_1^+(M_{ij})\ \dfrac{p_{ij}^{n+1} - p_{i-1j}^{n+1}}{h} - V_1^-(M_{ij})\ \dfrac{p_{i+1j}^{n+1} - p_{ij}^{n+1}}{h} \\ \\ + V_2^+(M_{ij})\ \dfrac{p_{ij}^{n+1} - p_{ij-1}^{n+1}}{h} - V_2^-(M_{ij})\ \dfrac{p_{ij+1}^{n+1} - p_{ij}^{n+1}}{h} \\ \\ = f(M_{ij}\ ,\ (n+1)\,\Delta t)\ , \end{array}\right. \tag{5.6}_1$$

with in $(5.6)_1$:

(i) $\Delta t\ (>0)$ a *time discretization step*;

(ii) $p_{ij}^{n} \sim p(M_{ij}\ ,\ n\Delta t)$;

(iii) $a^+ = \max(0,a)$, $a^- = \max(0,-a)$, $\forall\ a \in \mathbb{R}$;

(iv) $p_{kl}^{n+1} = g(M_{kl}\ ,\ (n+1)\Delta t)$ *if* $M_{kl} \in \partial\Omega$; $(5.6)_2$

(v) $p_{kl}^{0} = p_0(M_{kl})$. □ $(5.6)_3$

Scheme (5.6) is of the *backward Euler* type for the *time discretization* and of the *first order upwind* type for the *space discretization*. Probabilists favor the *finite difference* scheme (5.6) because it satisfied a *discrete*

maximum principle and therefore possesses a *probabilistic interpretation*. Unfortunately the above scheme is only *first order accurate*, quite dissipative and not well-suited for those situations where ε is small and $\underset{\sim}{V}$ has fast variations over the space domain Ω.

An interesting alternative to (5.6) is obtained through a space/time discretization which is second order and also of the upwind type (however, it does not satisfy the discrete maximum principle). Such a scheme is obtained as follows:

$$\begin{cases} p^0_{kl} = p_0 \ (M_{kl}) \ , \\ \\ p^1_{kl} \quad \textit{is obtained (for example) via} \ (5.6) \ ; \end{cases} \tag{5.7}_1$$

then for $n \geq 1$ *and* $2 \leq i,j \leq I-1$ *discretize* $(5.4)_1$ *by*

$$\begin{cases} \dfrac{\frac{3}{2}\, p^{n+1}_{ij} - 2\, p^{n}_{ij} + \frac{1}{2}\, p^{n-1}_{ij}}{\Delta t} - \varepsilon \, \dfrac{p^{n+1}_{i+1j} + p^{n+1}_{i-1j} + p^{n+1}_{ij+1} + p^{n+1}_{ij-1} - 4p^{n+1}_{ij}}{h^2} \\ \\ + V_1^+(M_{ij}) \, \dfrac{\frac{3}{2}p^{n+1}_{ij} - 2p^{n+1}_{i-1j} + \frac{1}{2}\, p^{n+1}_{i-2j}}{h} + V_1^-(M_{ij}) \, \dfrac{\frac{3}{2}p^{n+1}_{ij} - 2p^{n+1}_{i+1j} + \frac{1}{2}p^{n+1}_{i+2j}}{h} \\ \\ + V_2^+(M_{ij}) \, \dfrac{\frac{3}{2}p^{n+1}_{ij} - 2p^{n+1}_{ij-1} + \frac{1}{2}\, p^{n+1}_{ij-2}}{h} + V_2^-(M_{ij}) \, \dfrac{\frac{3}{2}p^{n+1}_{ij} - 2p^{n+1}_{ij+1} + \frac{1}{2}\, p^{n+1}_{ij+2}}{h} \\ \\ = f(M_{ij}, (n+1)\, \Delta t \ ) \ . \end{cases} \tag{5.7}_2$$

If $M_{ij} \in \Omega$ with either $M_{ij\pm1}$ or $M_{i\pm1j}$ on Γ it is possible that $M_{ij\pm2}$ or $M_{i\pm2j}$ does not belong to $\bar{\Omega}$; in such a case we can use to discretize $\underset{\sim}{V}\cdot\underset{\sim}{\nabla}$

at M_{ij} a first order scheme like in $(5.6)_1$ or alternatively a *centered second order* approximation like

$$\begin{cases} (\underset{\sim}{V}\cdot\underset{\sim}{\nabla}p)\,(M_{ij}) \sim V_1(M_{ij})\,\dfrac{p_{i+1j}-p_{i-1j}}{2h} \\ \qquad + V_2(M_{ij})\,\dfrac{p_{ij+1}-p_{ij-1}}{2h}\,. \end{cases} \tag{5.8}$$

The boundary conditions are treated as in $(5.6)_2$. The fact that the problems under consideration may have a fast dynamics requires the use of small h and Δt ; indeed as in Section 3 and 4 we can take advantage of the fact that Δt is small to solve the above discrete problems by successive over-relaxation since that method has good vectorization and parallelization properties (in practice few iterations will insure convergence at each time step).

Numerical experiments definitely show the superiority of the second order upwinding over the first order method (it is more accurate, less dissipative and almost as easy to implement).

5.3. SOLUTION OF PROBLEM (5.4) BY FINITE DIFFERENCES AND A BACKWARD METHOD OF CHARACTERISTICS.

As discussed in [31] , [32] (see also the references therein) the *backward or modified) method of characteristics* can be a most interesting tool for solving *advection dominated* problems.

The basic principle of the method is fairly simple and will be discussed on the continuous problem only.

Let's define the *total time derivative operator* $\dfrac{D}{Dt}$ by

$$\frac{Dp}{Dt} = \frac{\partial p}{\partial t} + \underset{\sim}{V}\cdot\underset{\sim}{\nabla} p \tag{5.9}$$

and consider the *characteristic* flow associated to $(x , t) \in \mathbb{R}^N \times \mathbb{R}_+$, i.e. the N-dimensional vector $X(\tau; x , t)$ solution of the ordinary differential system

$$\begin{cases} \dfrac{dX}{d\tau} = \underset{\sim}{V}(X) , \\ \\ X(t; x , t) = x. \end{cases} \tag{5.10}$$

With the above relations the parabolic equation $(5.4)_1$ can also be written

$$\frac{Dp}{Dt} - \varepsilon \underset{\sim}{\nabla}^2 p = f \quad in \quad \Omega \times (0, T) , \tag{5.11}$$

and discretized *along the characteristics* at time $(n+1)\Delta t$ by the elliptic equation

$$\begin{cases} \dfrac{p^{n+1}(x) - p^n\left[X(n\Delta t ; x , (n+1)\Delta t)\right]}{\Delta t} - \varepsilon \underset{\sim}{\nabla}^2 p^{n+1}(x) = f^{n+1}(x) , \\ \\ p^{n+1} = g^{n+1} \quad on \quad \Gamma ; \end{cases} \tag{5.12}$$

more sophisticated schemes can be used (cf. [33]).

In practice, to compute $\bar{p}^n(x) = p^n\left[X(n\Delta t ; x, (n+1)\Delta t)\right]$ we shall integrate (5.10) numerically , *starting from a grid point* and track back from $(n+1)\Delta t$ to $n\Delta t$.

Several situations may occur:

(i) If the characteristic curve crosses the boundary at t_x^* $(n\Delta t < t_x^* < (n+1)\Delta t)$ we shall replace Δt by $(n+1)\Delta t - t_x^*$ and take for $\bar{p}^n(x)$

the value of g at (x^*, t_n^*) where x^* is the point at the crossing of Γ and of the characteristic curve.

(ii) If $X(n\Delta t ; x , (n+1)\Delta t) \in \Omega$ it necessarily belong to a cell defined by grid points; we shall then use , for example, an interpolation technique to compute $\overline{p}^n(x)$ (see Figure 5.1)

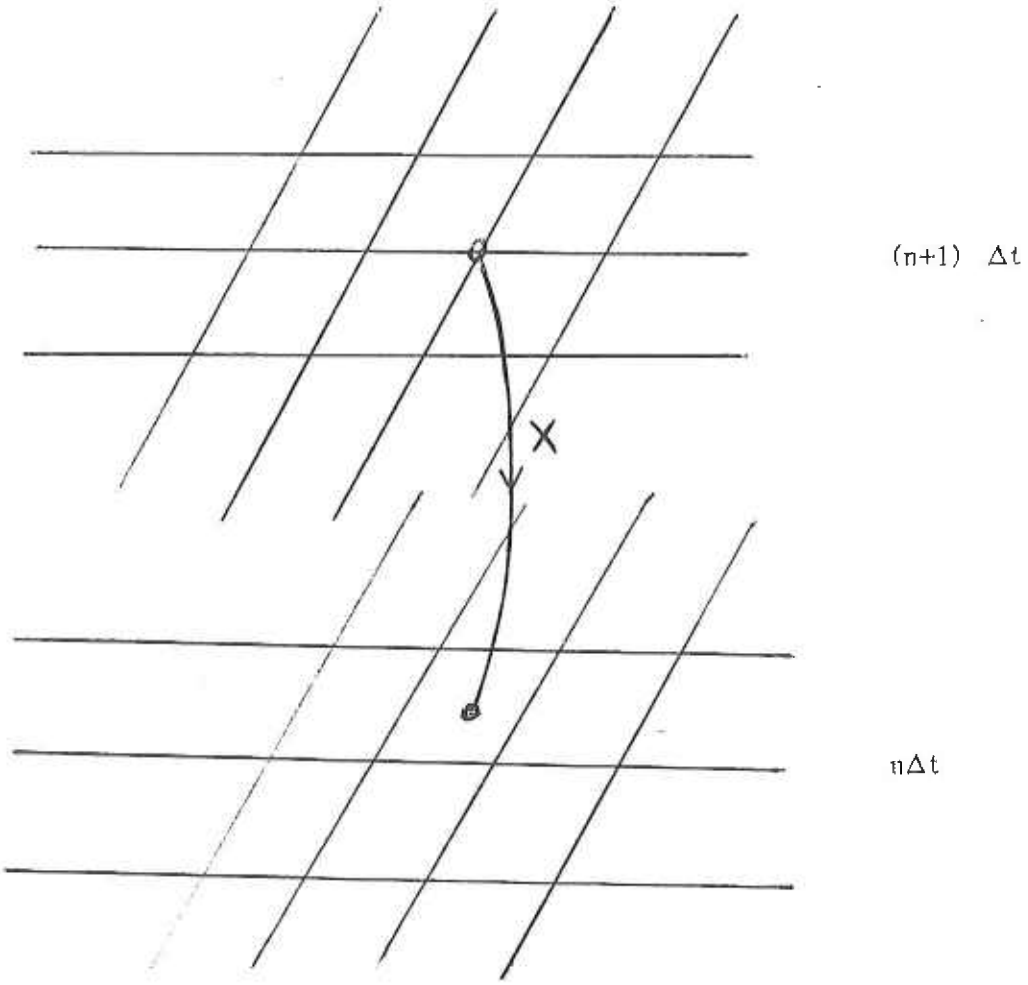


Figure 5.1: Backtracking Along the Characteristics.

Low order interpolation methods can lead to an overall method which may be quite dissipative, on the other hand high order interpolation methods are costly in high

dimension and not very easy to code leading to softwares which are not easy to vectorize or parallelize. We however think that these methods of characteristics are promising, but clearly they deserve a lot of further investigations.

The discretization of the terms associated to the elliptic operator $-\varepsilon\underset{\sim}{\nabla}^2$ in (5.12) is straightforward and is done by the same difference formula then in $(5.6)_1$ and $(5.7)_2$. The fully discrete system obtained from (5.12) is then solved by those successive over relaxation methods advocated in Section 5.2.

5.4. NUMERICAL EXPERIMENTS.

All the numerical experiments are concerned with problem (5.4) when $\Omega = (0,1)^N$, $f = 0$, $g = 0$. We have compared here the various methods discussed in Sections 5.2 and 5.3 including some variants where one only uses first order time differencing combined to second order upwinding for the space variables. We have also tested the variant of scheme (5.12) where $\frac{\partial p}{\partial t}$ has been discretized by

$$\frac{\partial p}{\partial t}(x\ ,\ (n+1)\Delta t) \sim$$

$$\frac{1}{\Delta t}\left\{\frac{3}{2}\ p^{n+1}(x) - 2p^{n}(X(n\Delta t\ ;\ x,\ (n+1)\Delta t)) + \frac{1}{2}\ p^{n-1}(X((n-1)\Delta t\ ;\ x\ ,\ (n+1)\Delta t))\right\}. \qquad (5.13)$$

The numerical experiments have been carried out for N=2,3,4,5,6.

5.4.1. TWO DIMENSIONAL EXPERIMENTS.

Data: $\Omega = (0,\ 1)^2$, $f = 0$, $g = 0$, $\varepsilon = 10^{-3}$,

$\underset{\sim}{V} = \underset{\sim}{\nabla}(ln\ \gamma)$ *with* $\gamma = \sqrt{(x-x_0)^2+(y-y_0)^2}$, $(x_0\ ,\ y_0) = (-1,\ -1)$,

$$p_0(x,y) = \begin{cases} 16^2\, x(\frac{1}{2}-x)y\,(\frac{1}{2}-y) & \text{if } (x,y) \in (0,1/2)^2\,, \\ 0 & \text{in } \Omega\setminus(0,1/2)^2\,. \end{cases}$$

The behavior of the numerical methods has been summarized in Table 5.1 , below

Method	h	Δt	CPU/time step(secs.) CRAY-XMP
2nd order upwinding with 1st order time differencing	1/10 1/20 1/40 1/80	h/2 h/2 h/2 h/2	0.0016 0.0036 0.0082 0.021
2nd order upwinding and time differencing	1/10 1/20 1/40 1/80	h/2 h/2 h/2 h/2	0.0017 0.0037 0.0081 0.020
Method of characteristics (1st order time differencing)	1/10 1/20 1/40 1/80	h h h h	0.0016 0.0029 0.0074 0.022
Method of characteristics (2nd order time differencing)	1/10 1/20 1/40 1/80	h h h h	0.0021 0.0057 0.018 0.065

Table 5.1 (*Two dimensional experiments*).

Figure 5.2 shows the trace, on the diagonal y-x=0 , of the computed solution at t=1 for various values of h (1st order time differencing, second order upwinding); Figure 5.3 shows the time evolution of this trace (computations done by the above method with h=1/40). Figure 5.4 corresponds to the same experiment than in

Figure 5.2 except that here we have been using second order time differencing and upwinding; we observe that the results for $h=\frac{1}{40}$ and $\frac{1}{80}$ are practically identical and that those obtained with $h=/20$ are indeed very close of those obtained by the previous method with $l=1/40$. Figures 5.5 and 5.6 correspond to the same experiment than in Figures 5.2 and 5.3, except that here one has used the method of characteristics of Section 5.3 (first order time differencing in Figure 5.5, 2nd order time differencing in Figure 5.6) ; we observe that the method of characteristics used here (with bilinear interpolation on the finite difference cells) is more dissipative than the 2nd order upwinding methods; we observe also that, coupled to the method of characteristics, second order time differencing seems to be slightly more dissipative than the first order one.

5.4.2. THREE-DIMENSIONAL EXPERIMENTS.

Data: $\Omega = (0,1)^3$, $f=0$, $g=0$, $\varepsilon = 10^{-3}, 10^{-4}$ and 10^{-8} , $\underline{V} = \underline{\nabla}(1/\gamma)$ with $\gamma = \sqrt{(x-x_0)^2+(y-y_0)^2+(x-z_0)^2}$, $(x_0 , y_0 , z_0) = (2,2,2,)$,

$$p_0(x, y, z) = \begin{cases} 16^3 xyz(1-x)\,(1-y)\,(1-z) & in \quad \omega = (0,1/2)^3 , \\ \\ 0 \quad in \quad \Omega/\omega \quad . \end{cases}$$

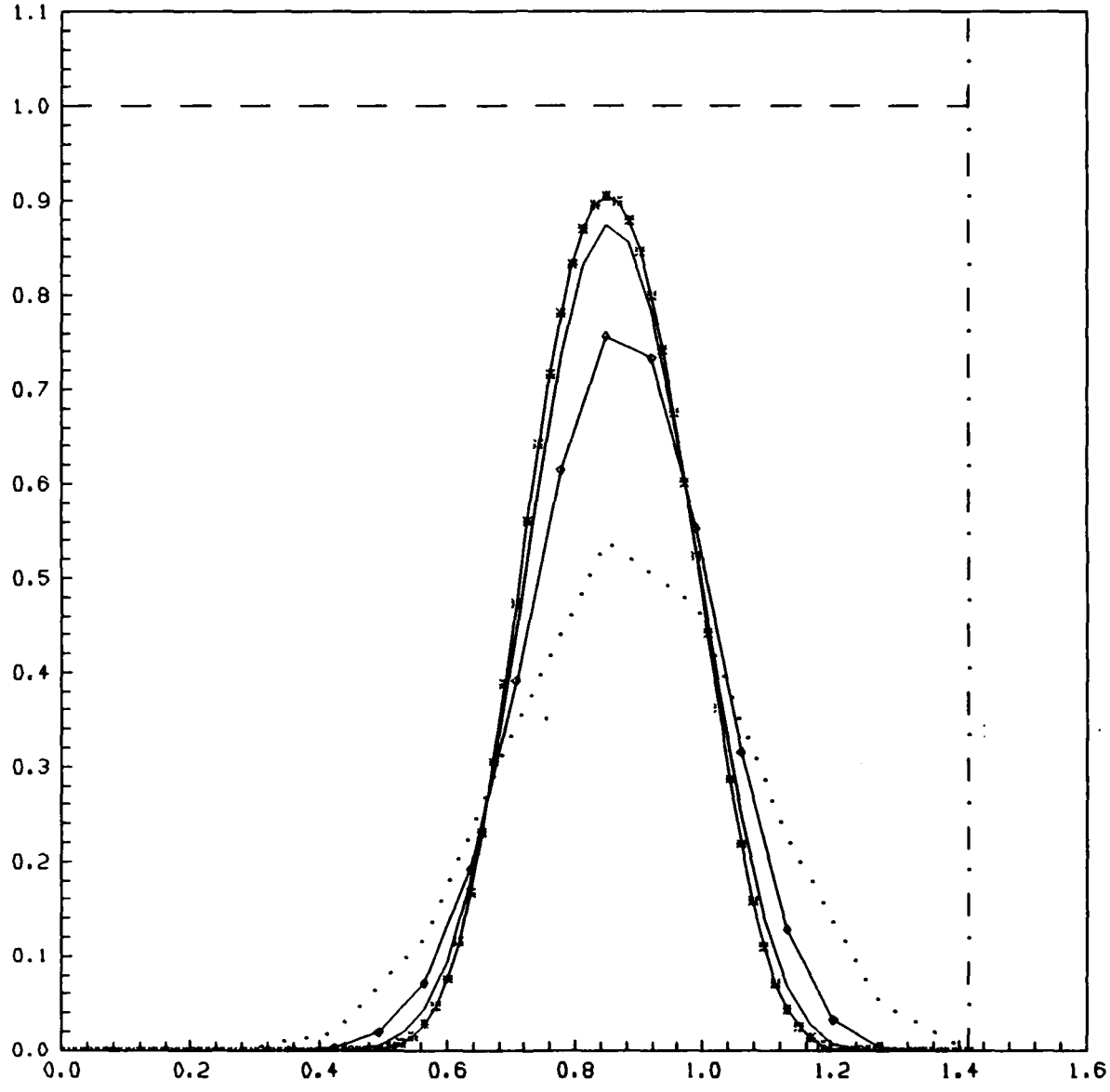


Figure 5.2

SOLUTION ON DIAGONAL AT T=1.0

"DOTTED"--H=1/10. " ◇ "--H=1/20. "SOLID"--H=1/40. "-•-"--H=1/80

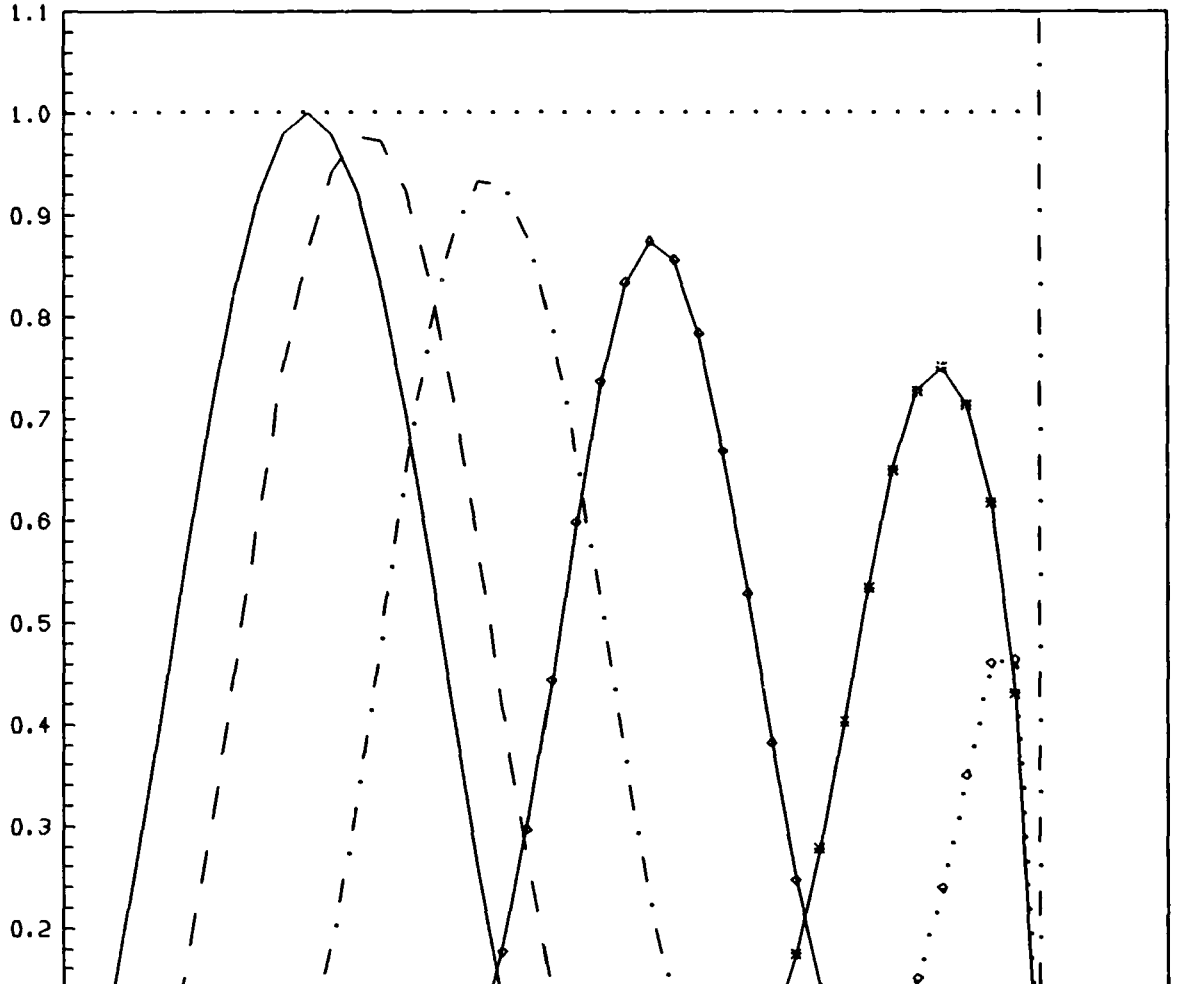
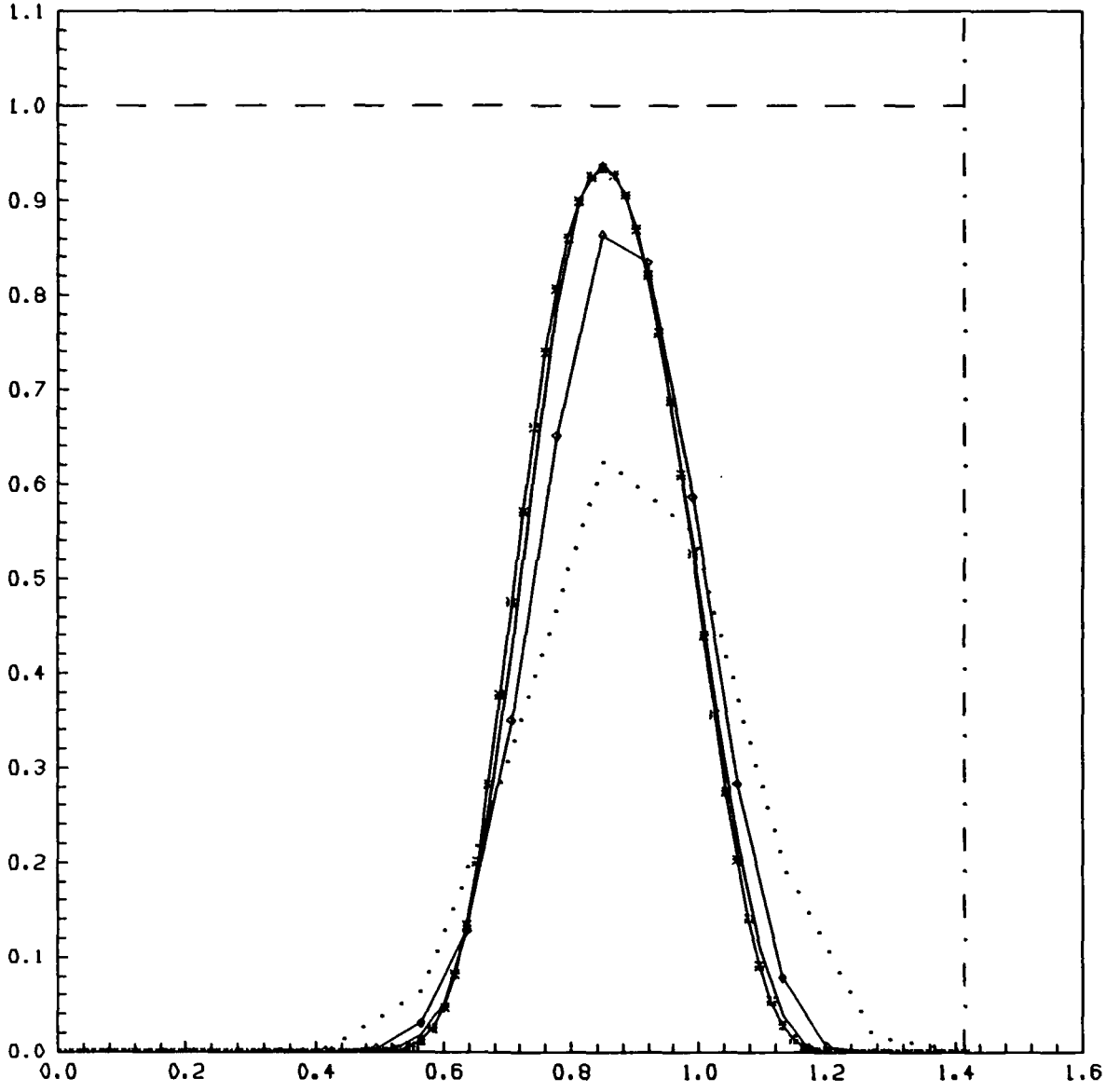


Figure 5.3

EVOLUTION OF U ON DIAGONAL

T=0.0, 0.15, 0.5, 1.0, 2.0, 2.5, 3.0; H=1/40



Figuure 5.4

SOLUTION ON DIAGONAL AT T=1.0, OBTAINED BY SECOND ORDER UPWINDING
"DOTTED"--H=1/10. "-◇-"--H=1/20. "SOLID"--H=1/40. "-•-"--H=1/80

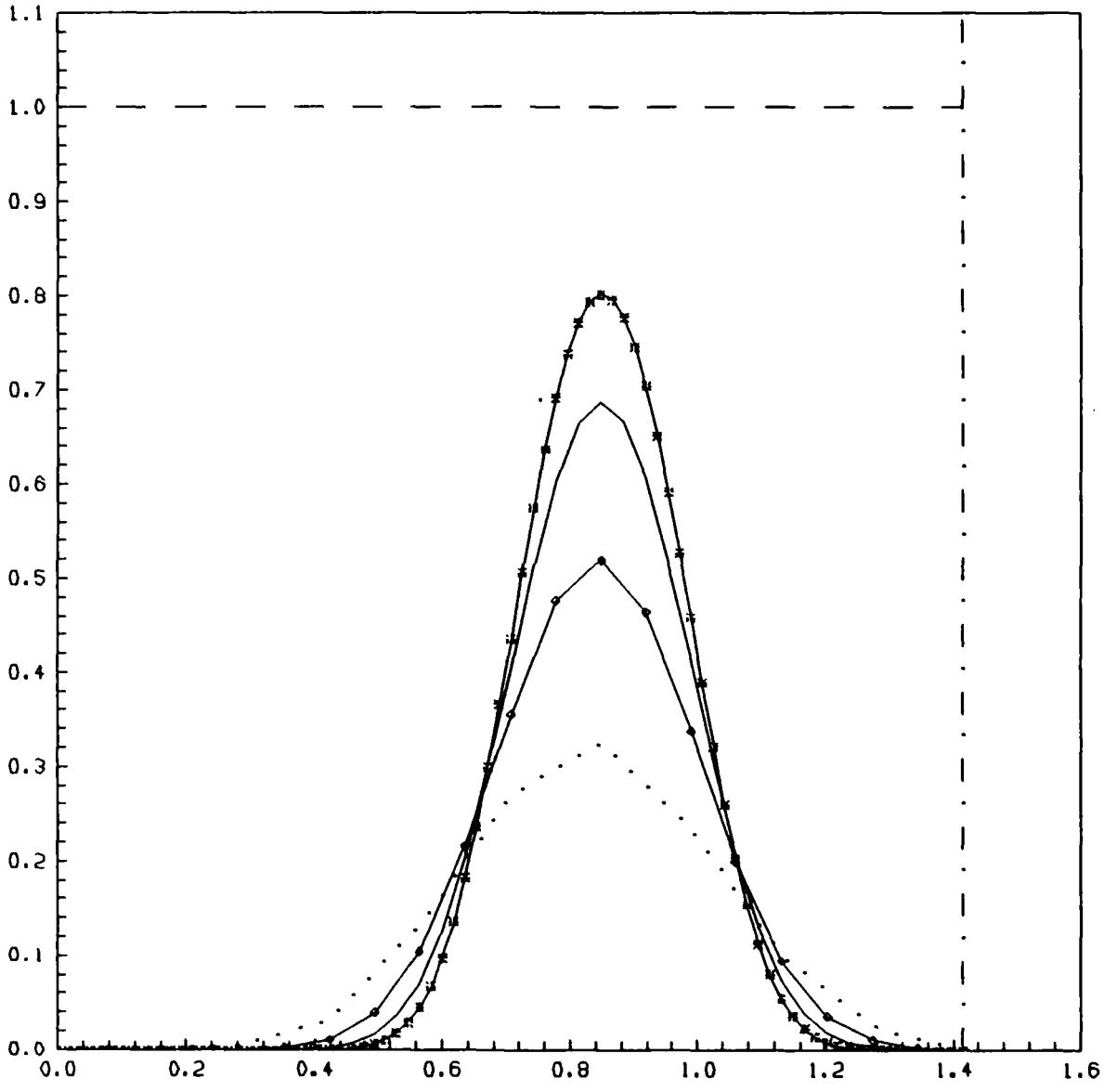


Figure 5.5

SOLUTION ON DIAGONAL AT T=1.0, OBTAINED BY METHOD OF CHARACTERISTIC
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/40, "-•-"--H=1/80

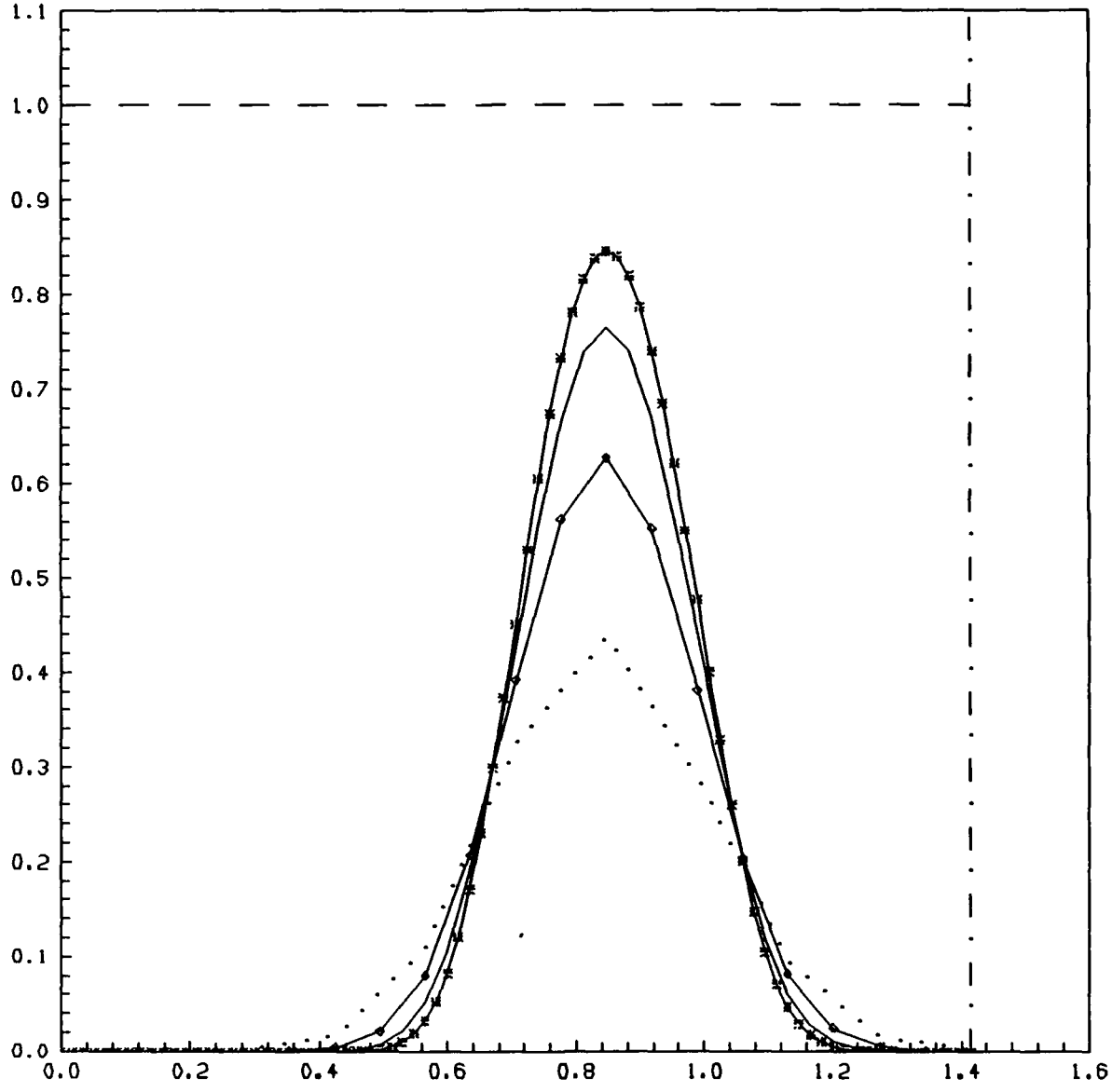


Figure 5.6

SOLUTION ON DIAGONAL AT T=1.0, OBTAINED BY METHOD OF CHARACTERISTIC
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/40, "-·-"--H=1/80

Table 5.2 summarizes some of the numerical results

Method	h	Δt	CPU/time step (secs), CRAY-XMP		
			$\epsilon = 10^{-3}$	$\epsilon = 10^{-4}$	$\epsilon = 10^{-8}$
2nd order upwinding and time differencing	1/10	h/2	0.012	0.012	0.012
	1/20	h/2	0.059	0.057	0.061
	1/40	h/2	0.28	0.28	0.29
Characteristics with 2nd order time differencing	1/10	h/2	0.022	0.019	0.017
	1/20	h/2	0.15	0.14	0.13
	1/40	h/2	1.07	1.03	1.03

Table 5.2 (*Three dimensional experiments*)

On Figure 5.7 (resp. 5.8) we have shown, at $t = 4$, the trace, on the line $x=y=z$, of the solution of (5.4) computed for various values of h by the *second order upwinding and time differencing method* (resp. the *method of characteristics with second order time differencing*). We observe again that the method of characteristics is more dissipative and less accurate than the upwinding method for which the results at $h = 1/40$ and $h = 1/80$ are practically identical. For those readers who may be surprised by the fact that the three dimensional results show more dissipation (for the same value of ϵ) than the two dimensional ones, we would like to mention that a *Fourier analysis* would show a faster time decay to zero, due the fact that, for the same rank, the eigenvalues of the Laplace operator are increasing functions of the dimension N if one considers as space domain $\Omega = (0,1)^N$.

Figure 5.9 corresponds to the same experiment than Figure 5.7, except that $\epsilon = 10^{-4}$; Figure 5.10 shows the *time evolution* of the trace, on the line $x=y=z$, of the solution of (5.4) computed by the *second order upwinding and time differencing method* for $h = 1/40$, $\Delta t = 1/80$, $\epsilon = 10^{-4}$. Figures 5.11 and 5.12 correspond to the same experiments than Figures 5.9 and 5.10, except that we have used here the method of characteristics with second order time differencing. Finally, Figures 5.13 to 5.16 correspond to the same experiments than Figures 5.9 to 5.12 except that now $\epsilon = 10^{-8}$.

5.4.3. FOUR DIMENSIONAL EXPERIMENTS.

Data: $\Omega = (0,1)^4$, $f = 0$, $g = 0$, $\epsilon = 10^{-3}$, $\underset{\sim}{V} = \underset{\sim}{\nabla}(1/\gamma^2)$ with

$$\gamma^2 = \sum_{i=1}^{4} (x_i - 2)^2 \quad \text{if} \quad x = (x_i)_{i=1}^{4} ,$$

and

$$p_0(x) = \begin{cases} 16^4 \prod_{i=1}^{4} x_i \, (1/2 - x_i) & \text{in } \omega = (0,1/2)^4 , \\ 0 & \text{in } \Omega \setminus \omega . \end{cases}$$

Table 5.3 summarizes some of the numerical results

Method	h	Δt	CPU/time step, CRAY-XMP
2nd order upwinding and time differencing	1/10	h/2	0.12 sec.
	1/20	h/2	1.1 sec.
	1/32	h/2	5.4 sec.

Table 5.3 (4^{th} dimensional experiments)

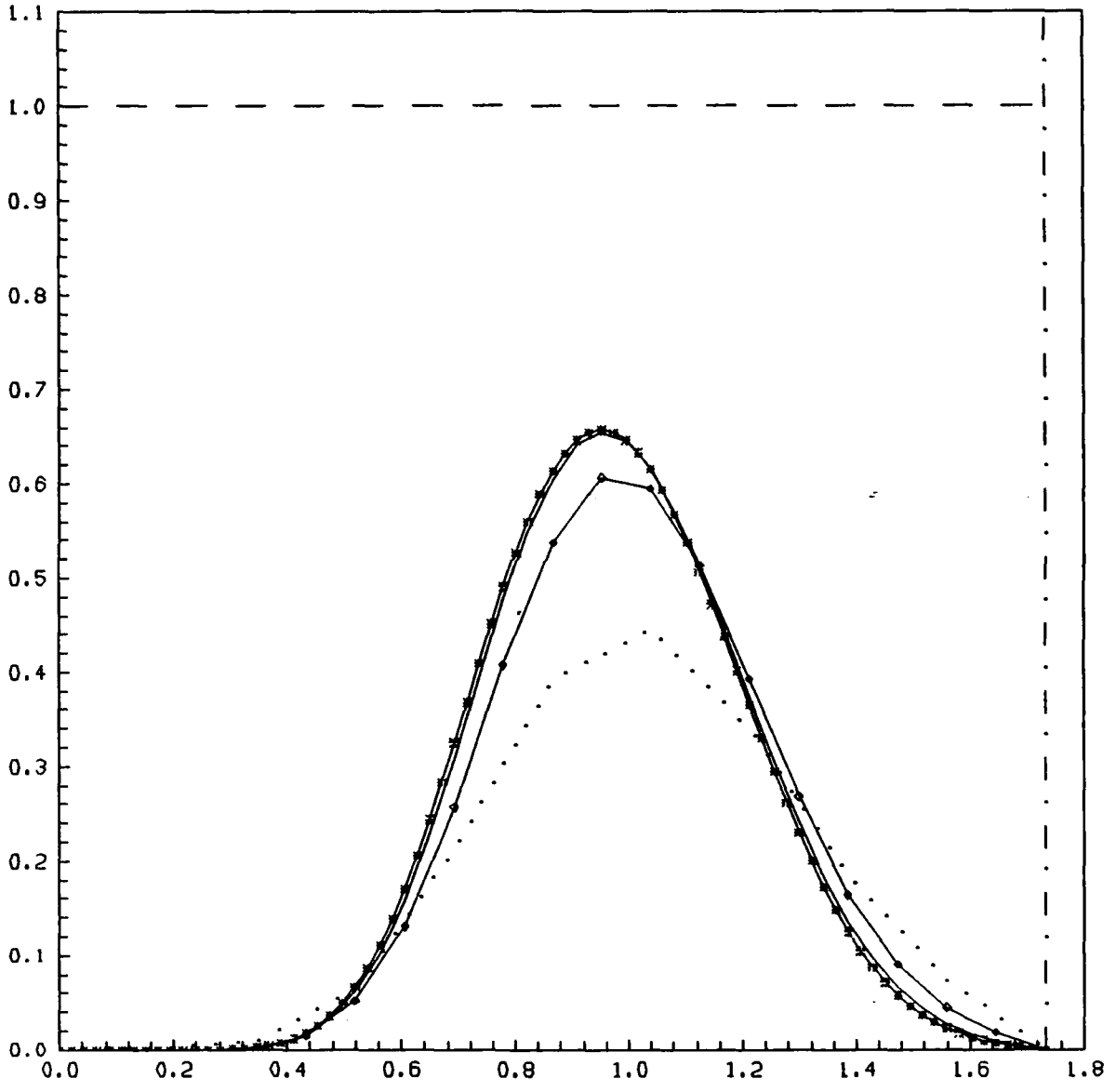


Figure 5.7

SOLUTION ON DIAGONAL AT T=4.0. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10. "-◊-"--H=1/20. "SOLID"--H=1/40. "-*-"--H=1/80

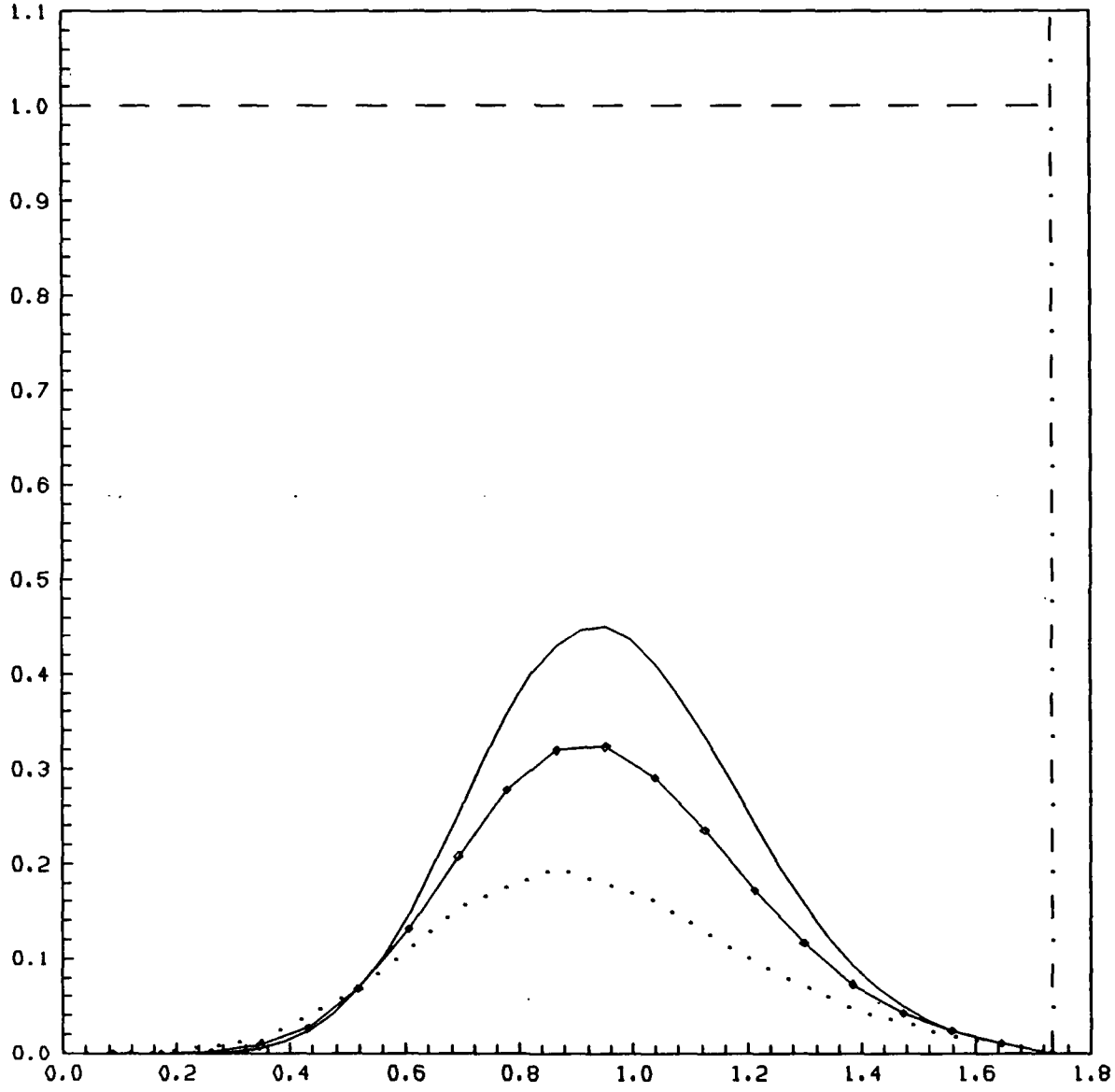


Figure 5.8

SOLUTION ON DIAGONAL AT T=4.0, METHOD OF CHARACTERISTICS WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/40, EPSI=1E-3, DT=H/2

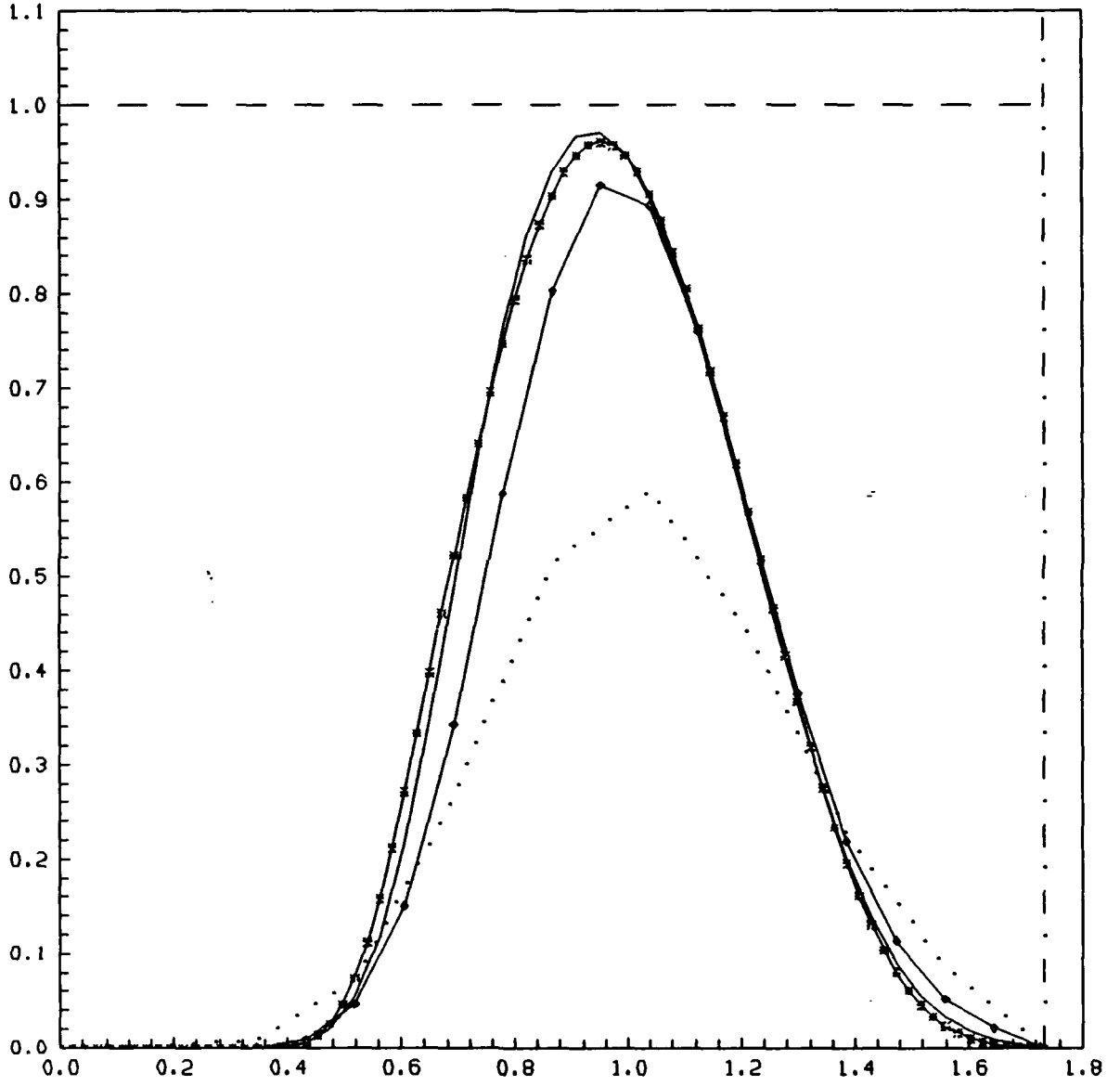


Figure 5.9

SOLUTION ON DIAGONAL AT T=4.0, SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/40, "-∗-"--H=1/80. EPSI=1E-4, DT=H/2

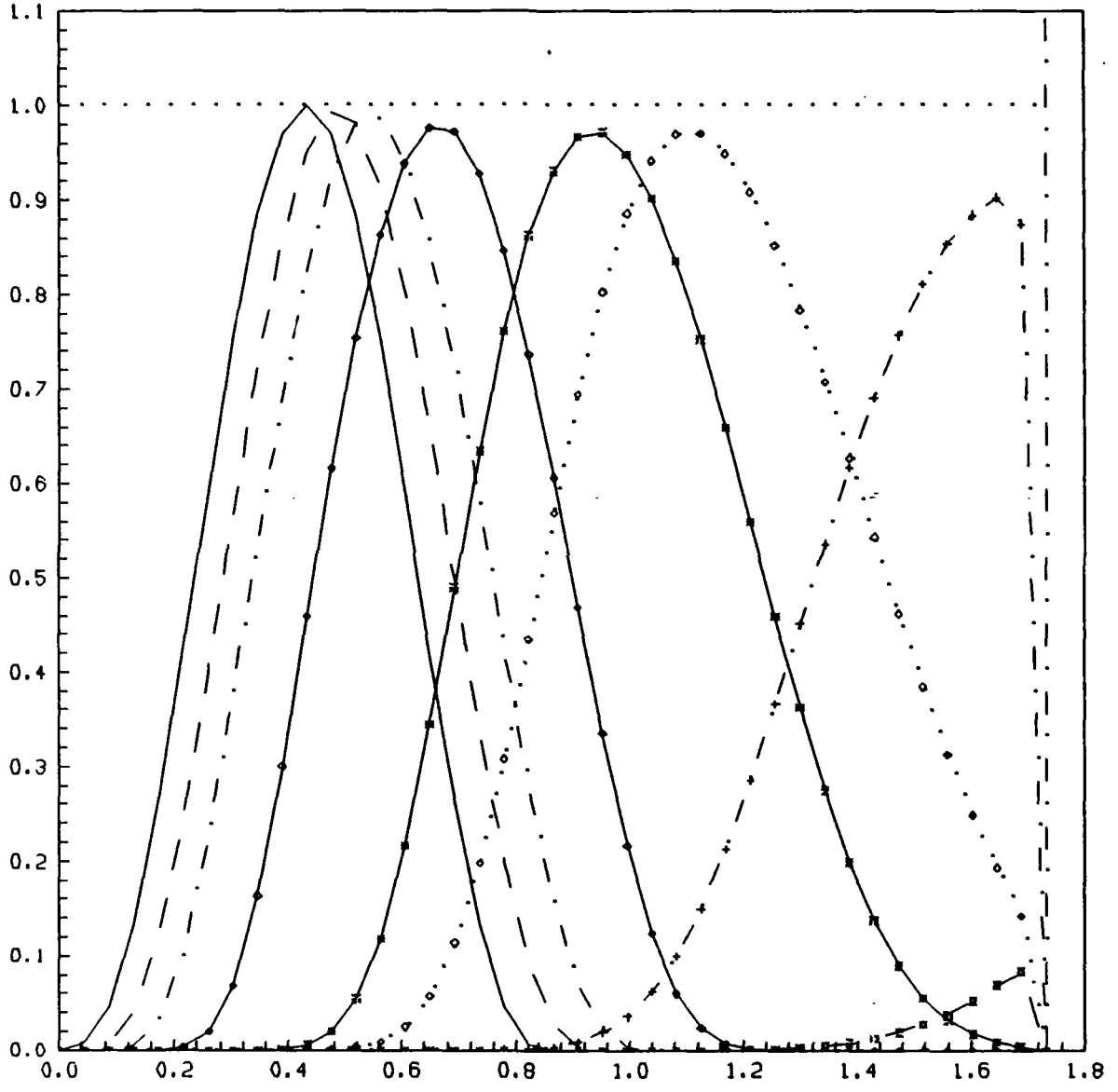


Figure 5.10

EVOLUTION OF U ON DIAGONAL. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME T=0.0, 0.5, 1.0, 2.0, 4.0, 5.0, 7.5, 10.0; H=1/40, DT=H/2, EPSI=1E-4

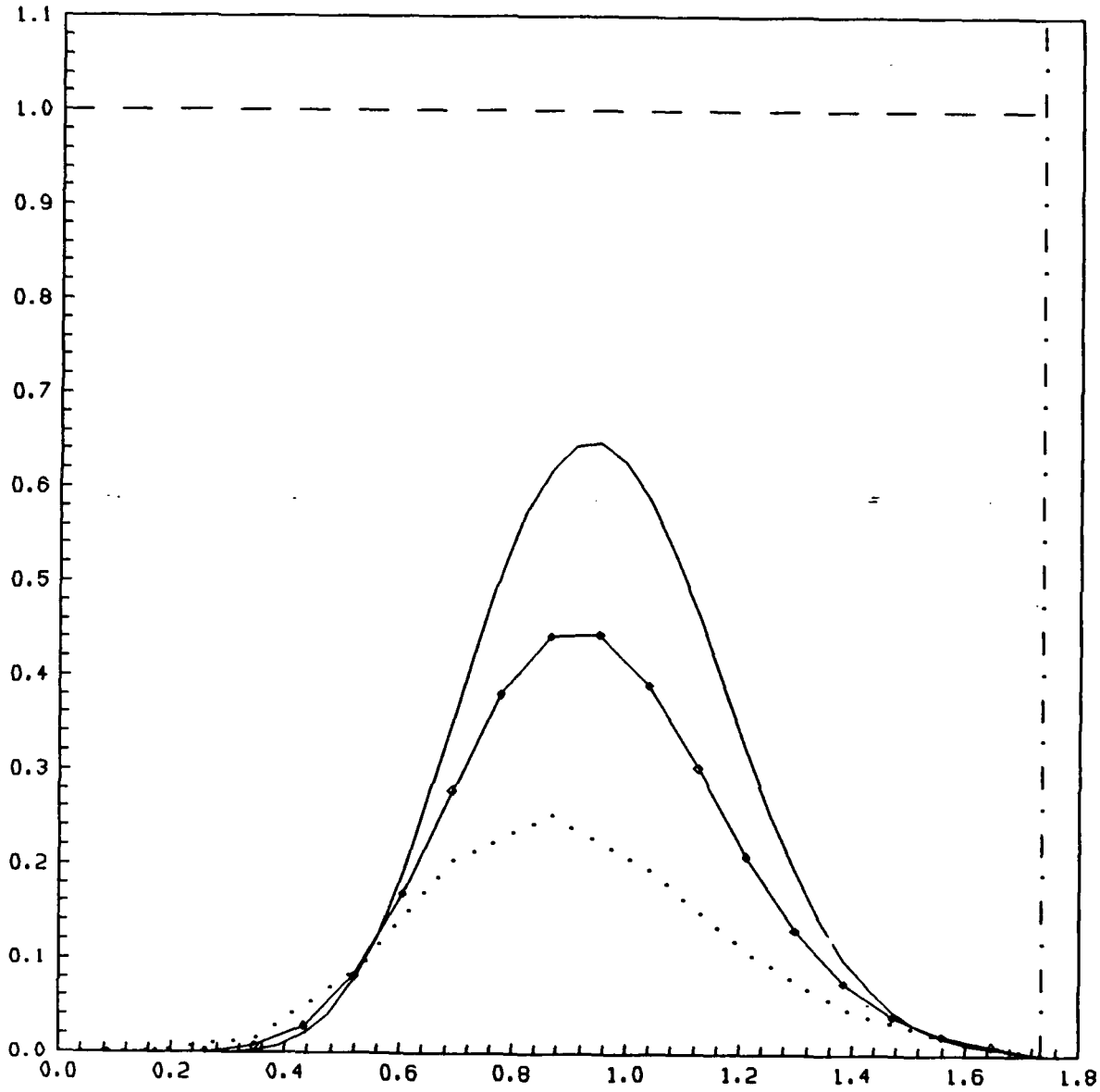


Figure 5.11

SOLUTION ON DIAGONAL AT T=4.0, METHOD OF CHARACTERISTICS WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10. "-◇-"--H=1/20. "SOLID"--H=1/40. EPSI=1E-4. DT=H/2

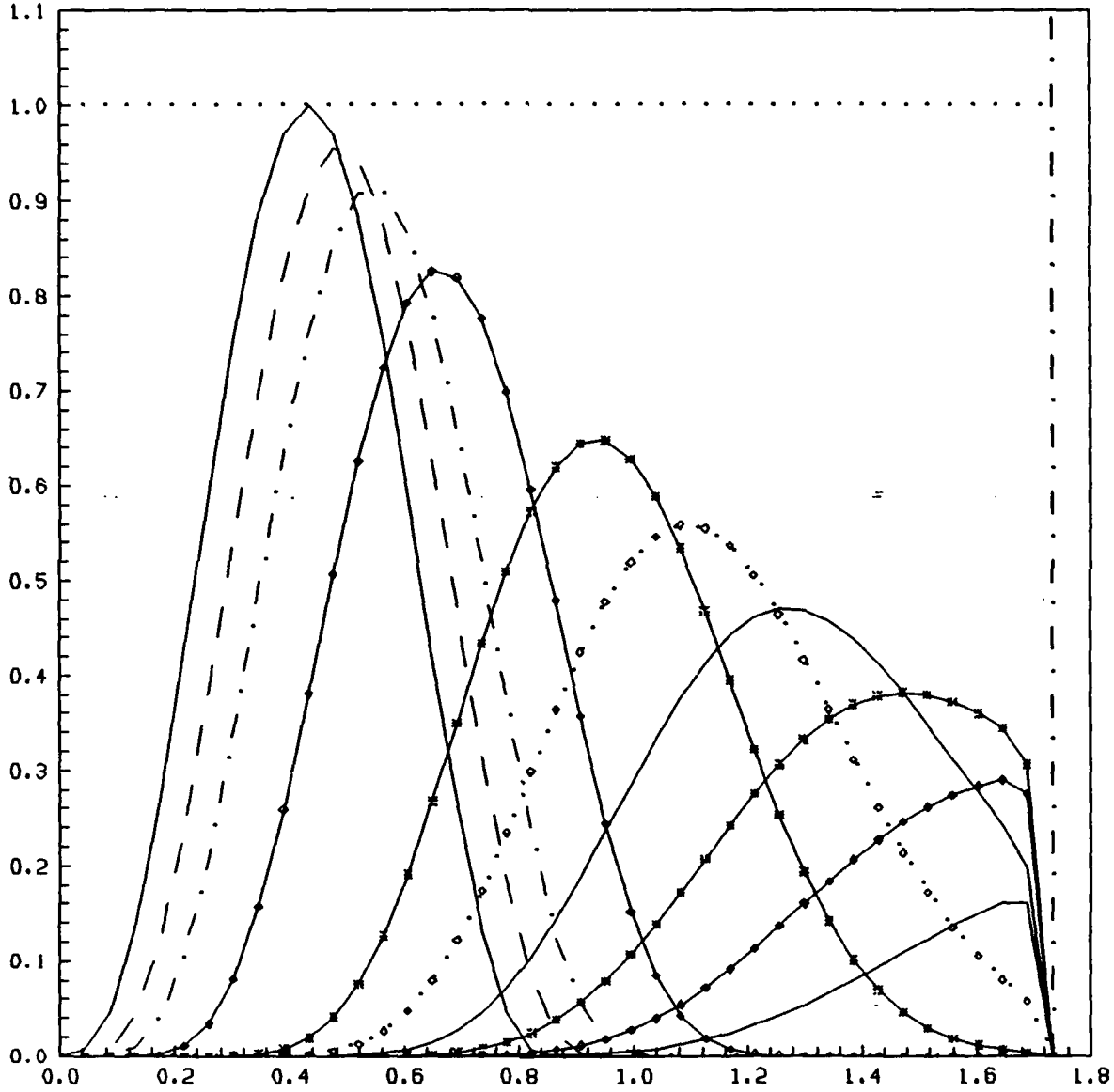


Figure 5.12

EVOLUTION OF U ON DIAGONAL. METHOD OF CHARACTERISTICS WITH SECOND ORDER DIFFER. IN TIME
T=0.0. 0.5. 1.0. 2.0. 4.0. 5.0. 6.0. 7.0. 8.0. 9.0: H=1/40. DT=H/2. EPSI=1E-4

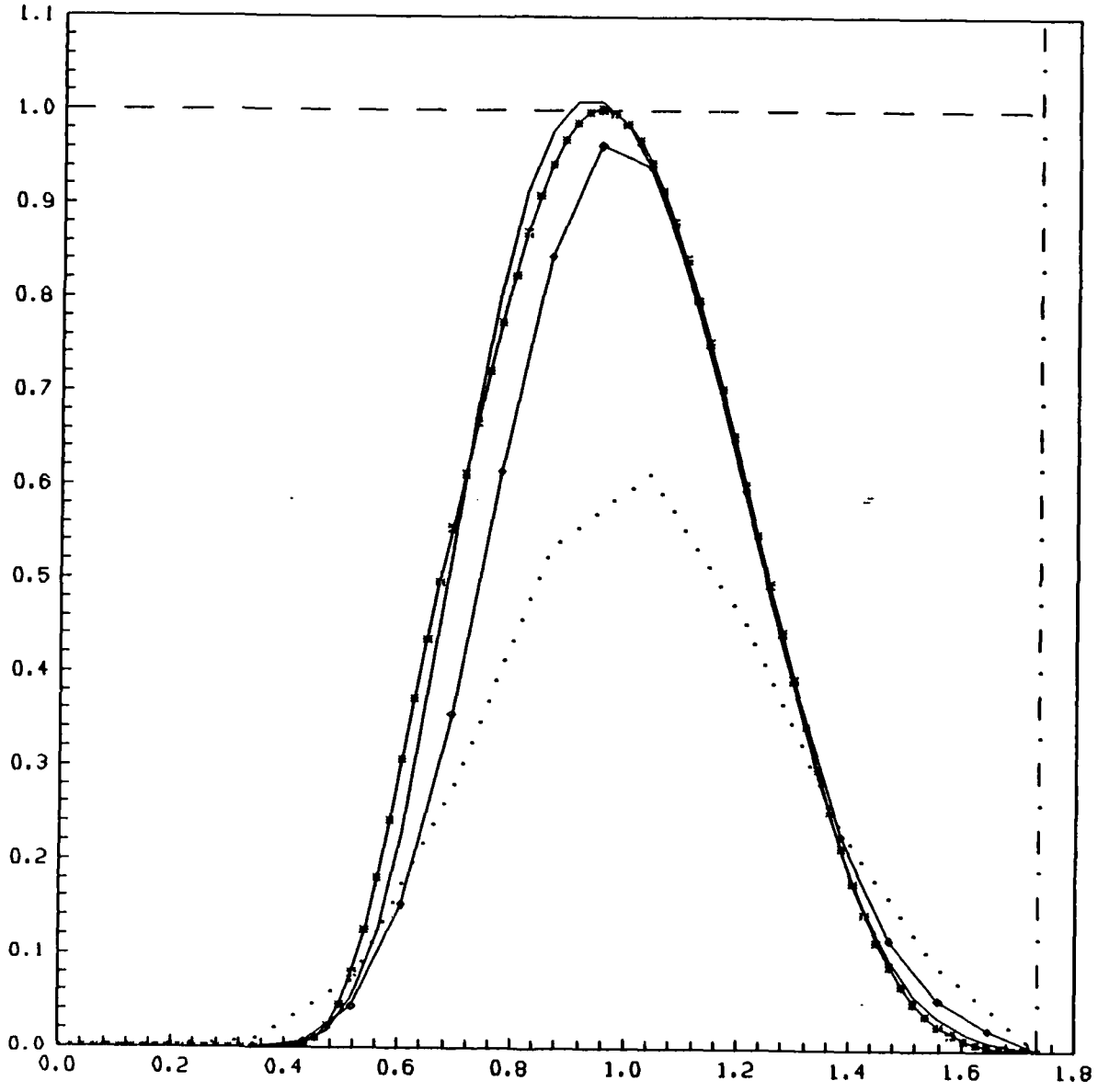


Figure 5.13

SOLUTION ON DIAGONAL AT T=4.0, SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10. "-◇-"--H=1/20. "SOLID"--H=1/40. "-·-"--H=1/80. EPSI=1E-8. DT=H/2

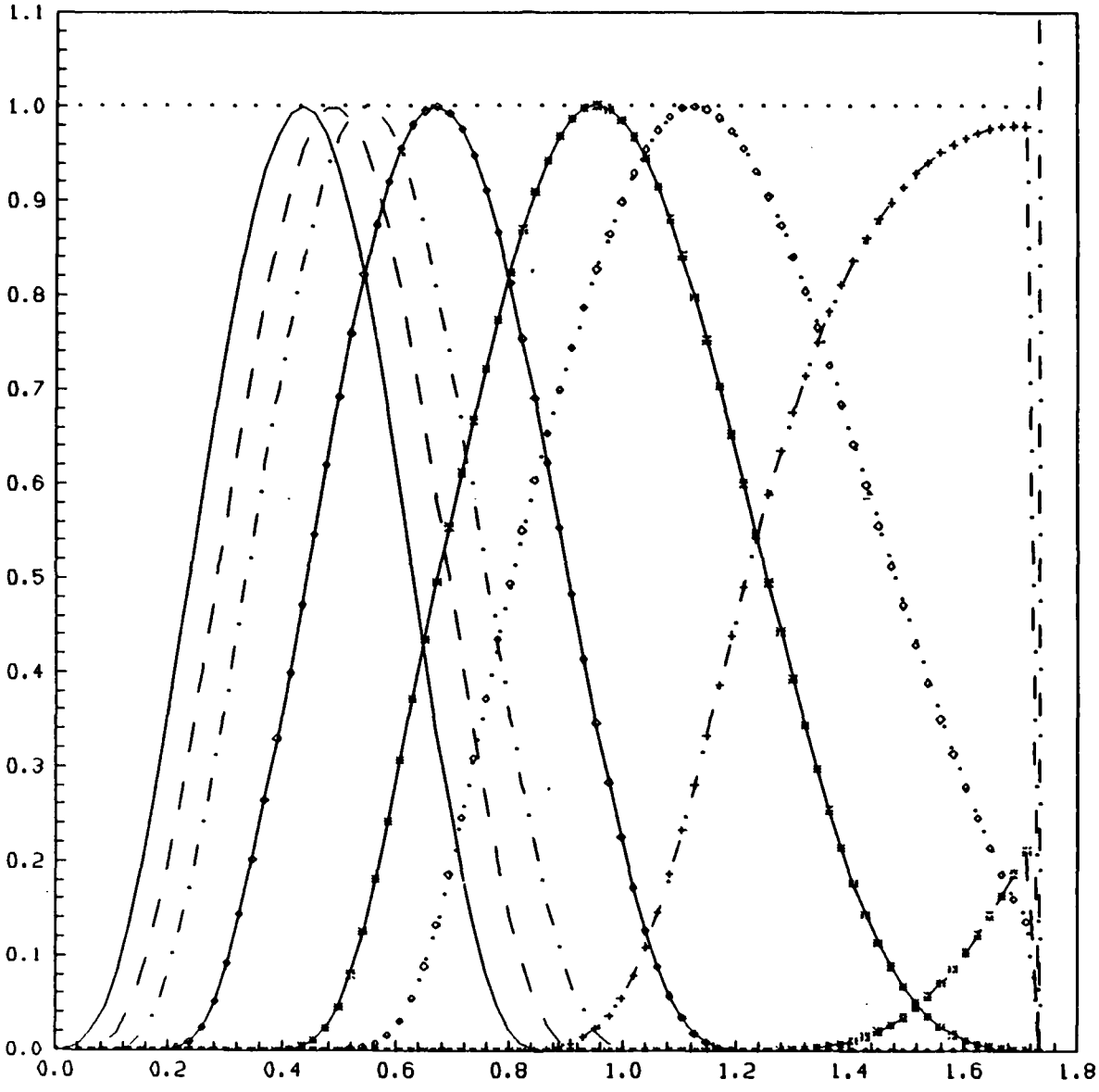


Figure 5.14

EVOLUTION OF U ON DIAGONAL. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
T=0.0, 0.5, 1.0, 2.0, 4.0, 5.0, 7.5, 10.0: H=1/80. DT=H/2. EPSI=1E-8

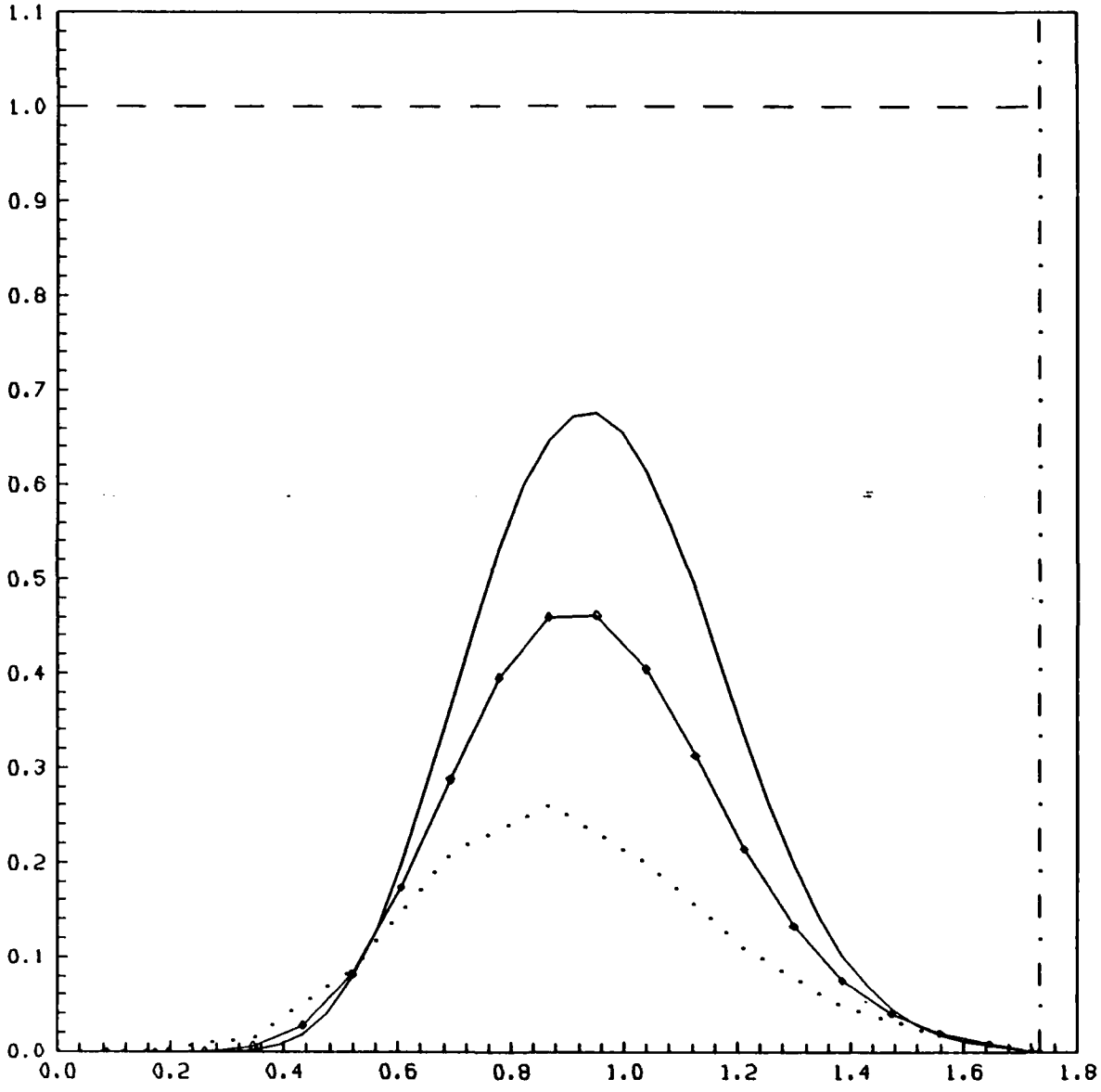


Figure 5.15

SOLUTION ON DIAGONAL AT T=4.0, METHOD OF CHARACTERISTICS WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/40. EPSI=1E-8, DT=H/2

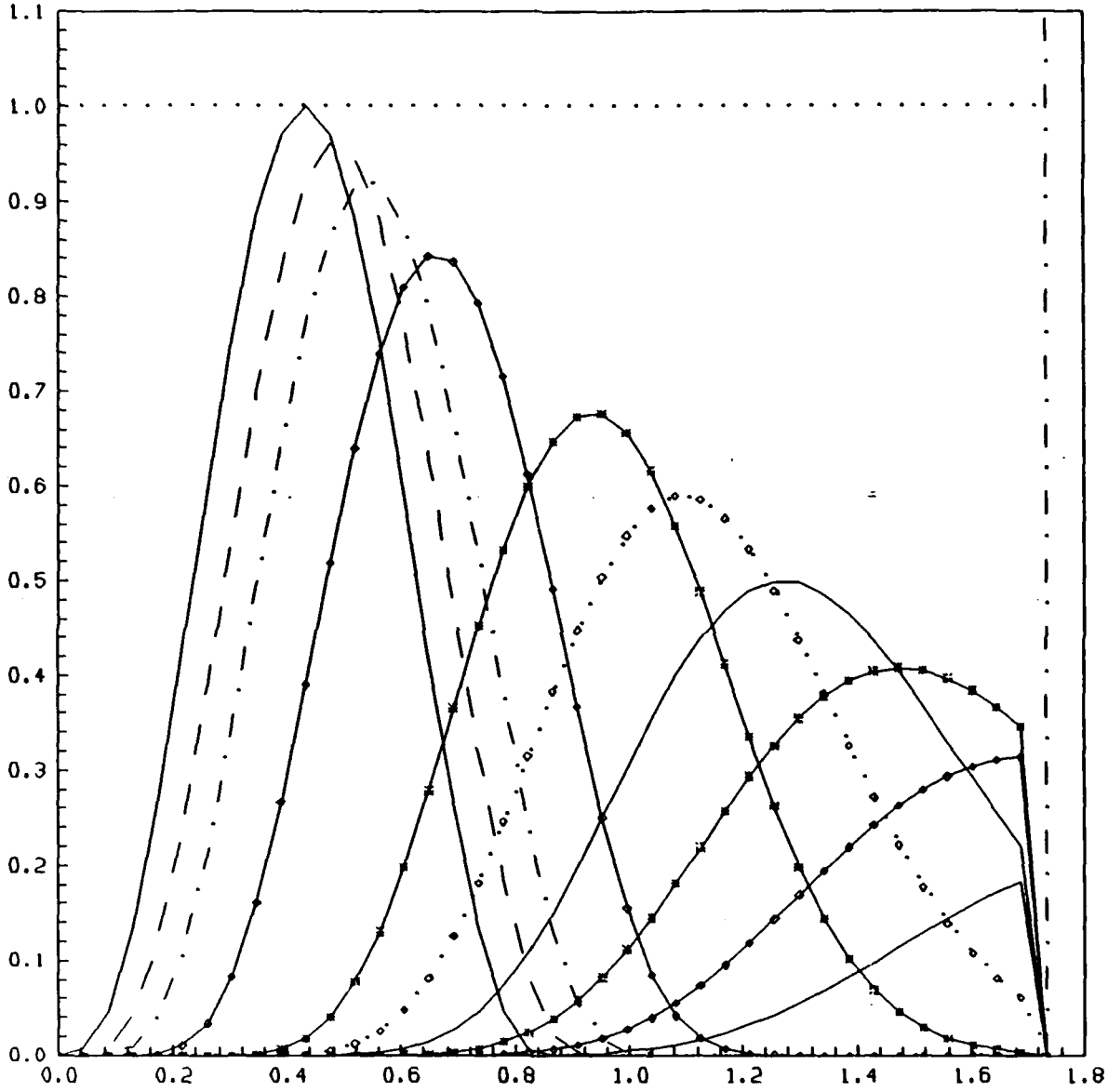


Figure 5.16

EVOLUTION OF U ON DIAGONAL. METHOD OF CHARACTERISTICS WITH SECOND ORDER DIFFER. IN TIME T=0.0, 0.5, 1.0, 2.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0: H=1/40, DT=H/2, EPSI=1E-8

Figure 5.17 shows the trace, at $t=4$ and on the line $x_1 = x_2 = x_3 = x_4$, of the solutions computed by *second order upwinding and time differencing*, for the above values of h and Δt, we observe the good agreement between the soltuions for $h=1/20$ and $h = 1/32$. Figure 5.18 shows the time evolution on the line $x_1 = x_2 = x_3 = x_4$ of the solution computed by the above method with $h=1/32$ and $\Delta t = 1/64$.

5.4.4. FIVE DIMENSIONAL EXPERIMENTS.

Data: $\omega = (0,1)^5$, $f = 0$, $g = 0$, $\epsilon = 10^{-3}$, 10^{-6},

$$\begin{cases} \underset{\sim}{V} = \underset{\sim}{\nabla}(\gamma^{-3}), \ with \\ \\ \gamma^2(x) = \sum_{i=1}^{5} (x_i-2)^2 \ \ if \ x = (x_i)_{i=1}^5, \end{cases}$$

$$p_0(x) = \begin{cases} 16^5 \prod_{i=1}^{5} x_i(1/2 - x_i) \ \ in \ \ \omega = (0,1/2)^5, \\ \\ 0 \ \ in \ \ \Omega \backslash \omega. \end{cases}$$

Since we did not have access to the full memory of the CRAY-XMP we only considered $h = 1/10$ and then $\Delta t = h/2 = 1/20$. Using the second order upwinding and time differencing method the CPU/time step ratio was 0.9 sec. if $\epsilon = 10^{-6}$.

The results displayed on Figures 5.19, 5.20 (traces on the line $x_1 = x_2 = x_3 = x_4 = x_5$) show that, as expected, the finite difference mesh is to coarse. Extrapolating from the results in lower dimensions the solution of the above test problem would require at least $h = 1/32$.

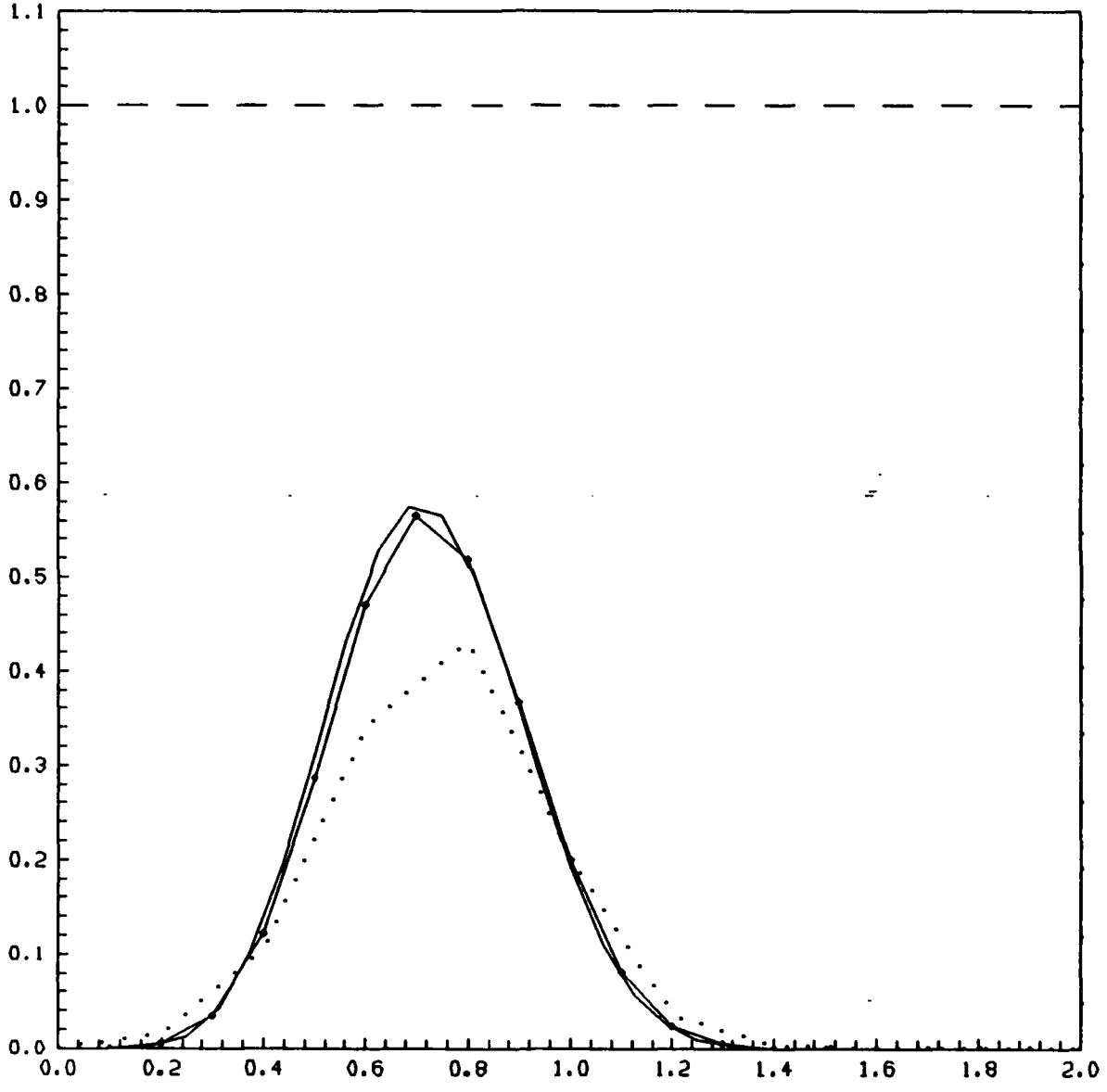


Figure 5.17

SOLUTION ON DIAGONAL AT T=4.0, SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
"DOTTED"--H=1/10, "-◇-"--H=1/20, "SOLID"--H=1/32, EPSI=1E-3, DT=H/2

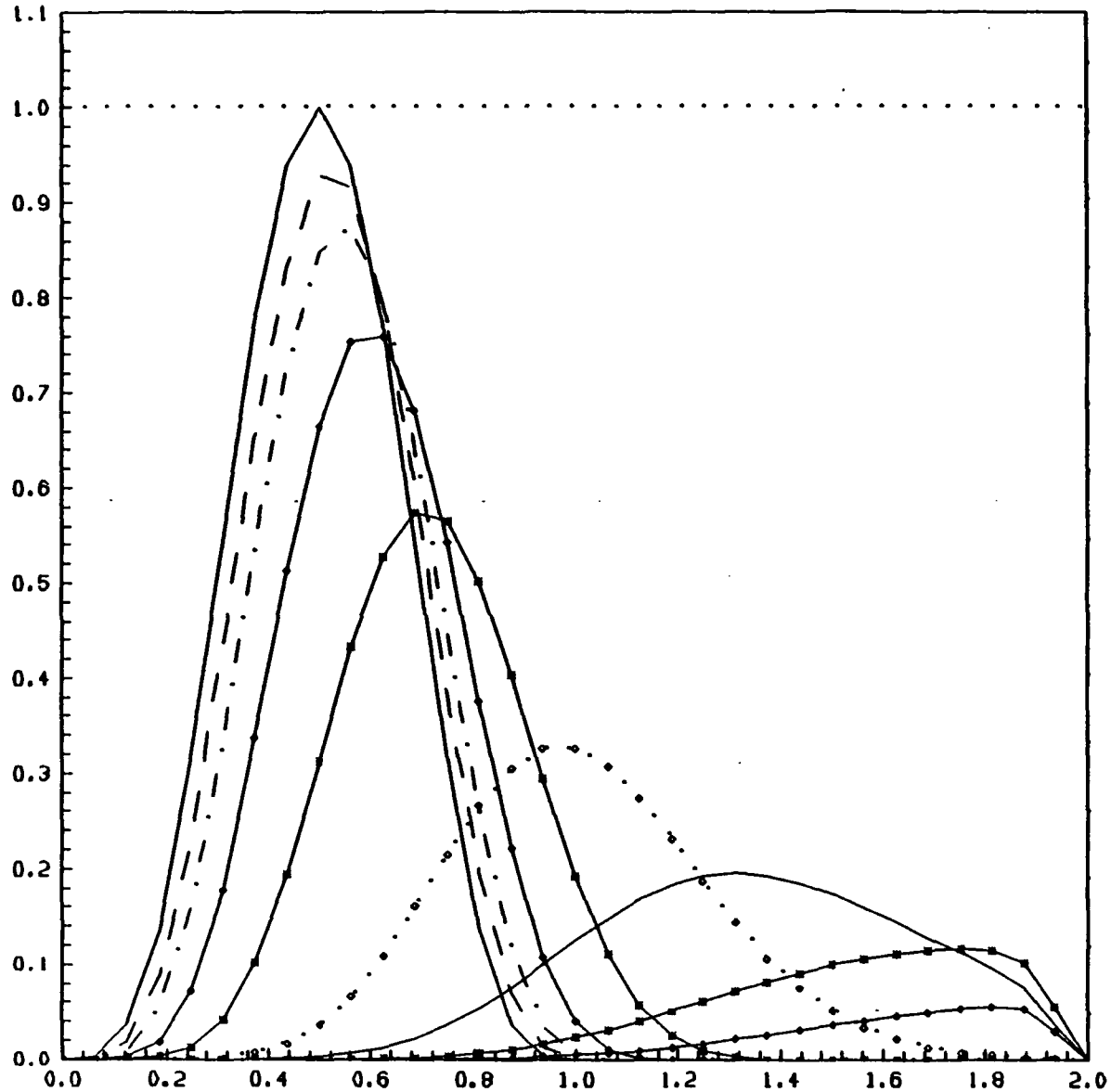


Figure 5.18

EVOLUTION OF U ON DIAGONAL. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME
T=0.0, 0.5, 1.0, 2.0, 4.0, 8.0, 12.0, 16.0, 19.0; H=1/32, DT=H/2, EPSI=1E-3

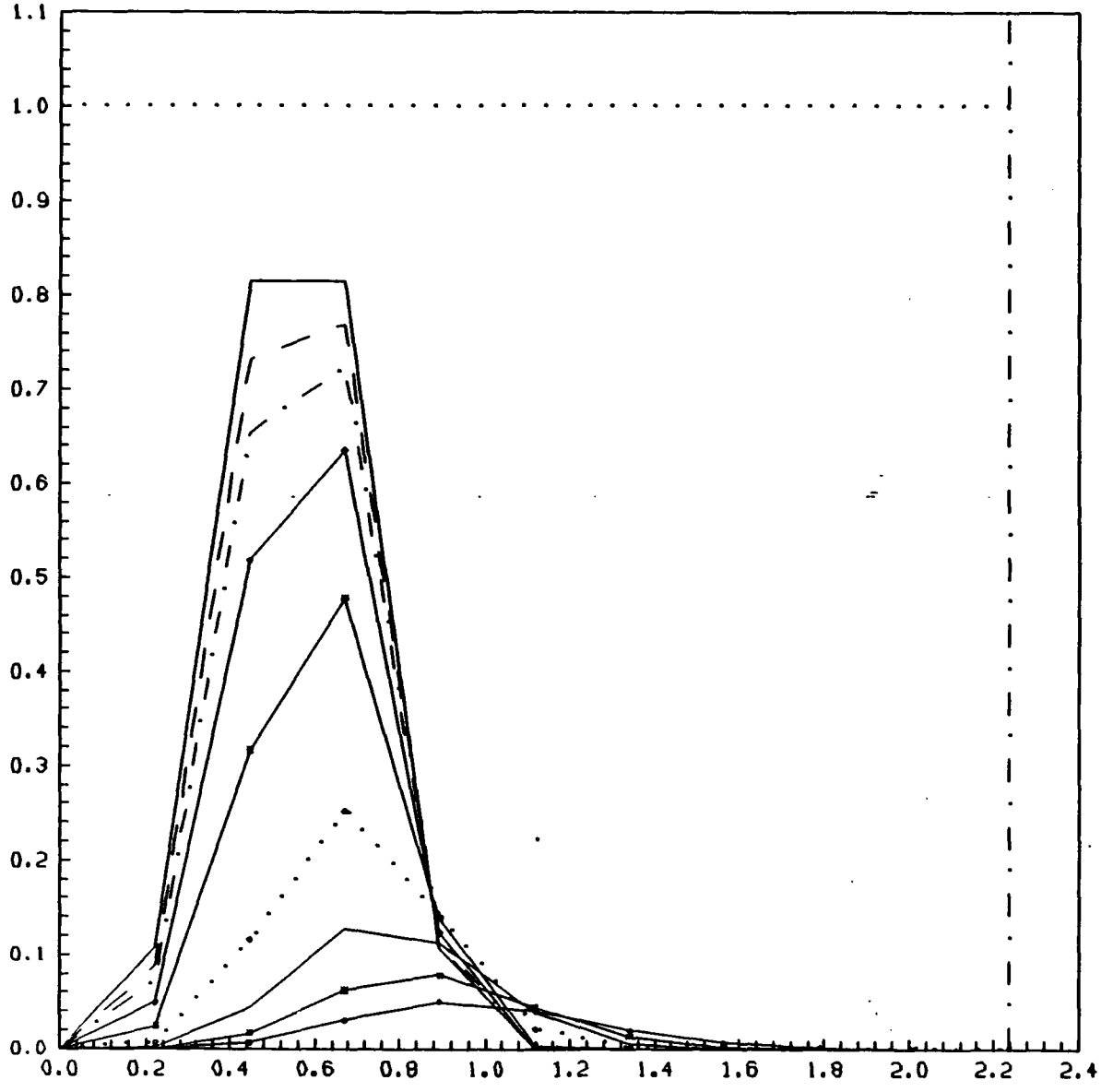


Figure 5.19

EVOLUTION OF U ON DIAGONAL. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME T=0.0, 0.5, 1.0, 2.0, 4.0, 8.0, 12.0, 16.0, 20.0: H=1/10, DT=H/2, EPSI=1E-3

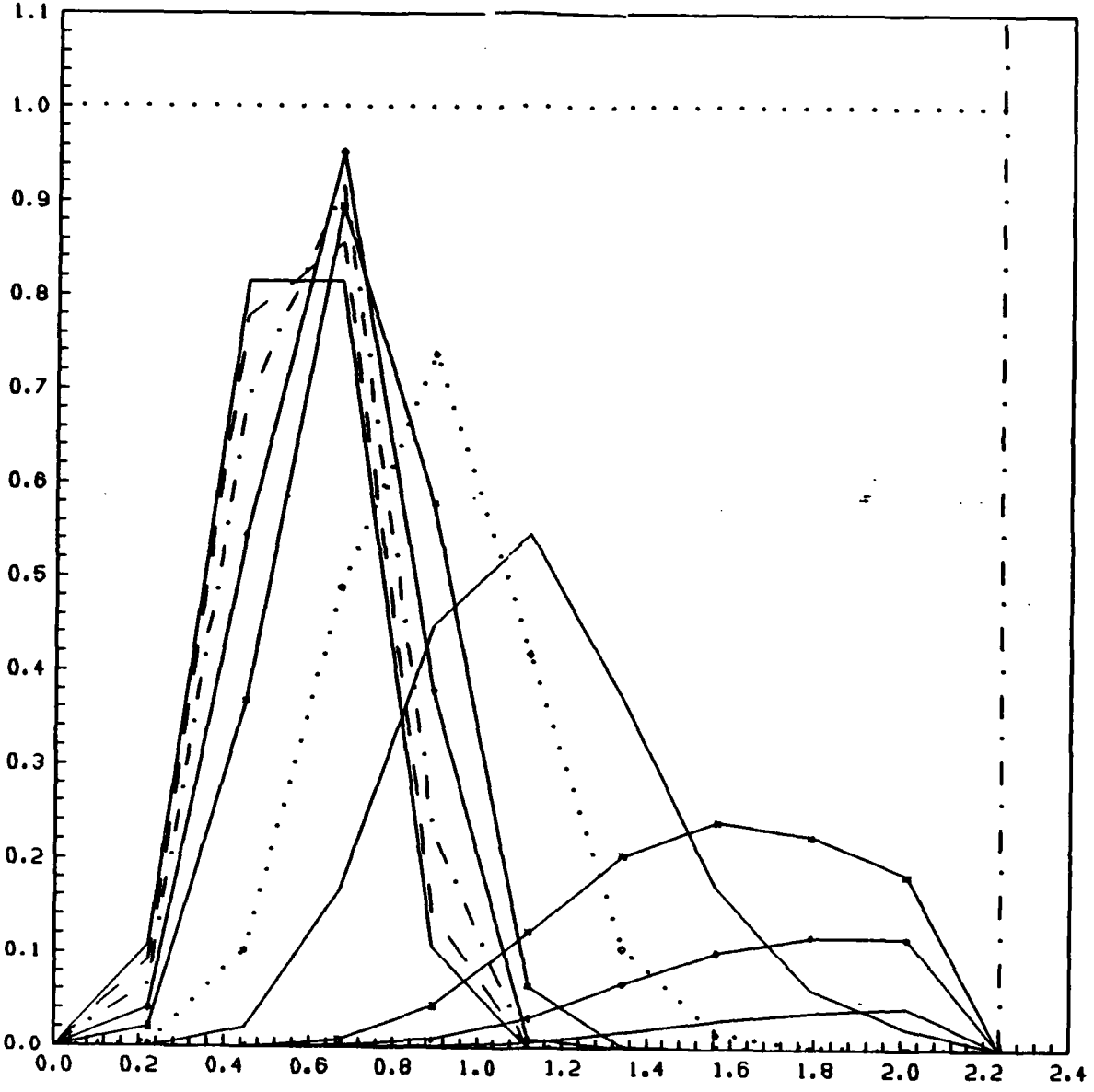


Figure 5.20

EVOLUTION OF U ON DIAGONAL. SECOND ORDER UPWINDING WITH SECOND ORDER DIFFER. IN TIME T=0.0, 1.0, 3.0, 6.0, 10.0, 20.0, 30.0, 50.0, 60.0, 70.0; H=1/10, DT=H/2, EPSI=1E-6

5.4.5. SIX DIMENSIONAL EXPERIMENTS.

Data: $f = 0$, $g = 0$, $\epsilon = 10^{-8}$, $h = 1/10$, $\Delta t = 1/20$, and

$$\begin{cases} \underset{\sim}{V} = 10\ \underset{\sim}{\nabla}(\gamma^{-4}) \quad with \\ \\ \gamma^2(x) = \sum_{i=1}^{6} (x_i - 2)^2 \quad if \quad x = (x_i)_{i=1}^{6} \ , \end{cases}$$

$$p_0(x) = \begin{cases} 16^6 \prod_{i=1}^{6} x_i(1/2 - x_i) \quad in \quad \omega = (0\ ,\ 1/2)^6 \ , \\ \\ 0 \quad in \quad \Omega \backslash \omega \ . \end{cases}$$

This test case is definitely a limit one, at least with the class of computers that the present authors can access; using *solid state device* , the CPU/time step ratio is here 173 seconds. It is clear that by a very sophisticated coding and still using the same method (*second order upwinding and time differencing*) the above performances can be improved but it is clearly the type of situations where massively parallel computing is needed.

5.5. FURTHER COMMENTS.

In the particular case where $\Omega = (0,1)^N$ *dimensional splitting methods* can be used to increase the degree of parallelism of the problem under consideration. We observe, for example, that if $N = 4$, then

$$\underset{\sim}{\nabla}^2 p = \sum_{i=1}^{2} \frac{\partial^2 p}{\partial x_i^2} + \sum_{i=3}^{4} \frac{\partial^2 p}{\partial x_i^2} \ ,$$

which suggests to apply the general methods of Section 2 with

$$A_1 = -\sum_{i=1}^{2} \frac{\partial^2}{\partial x_i^2} , \quad A_2 = -\sum_{i=3}^{4} \frac{\partial^2}{\partial x_i^2} .$$

It is our intention to start a campaign of numerical experiments to test the validity of this approach.

6. CONCLUSION.

Operator splitting methods definitely provide efficient methods for solving numerically mathematical problems modeled by parabolic equations or which can be reduced to the solution of such problems. In the case of problems in very high dimension ($N \geq 4$) the validity of this approach needs to be tested through further numerical experiments. An important conclusion which appears already is that the evolution aspect of these problems makes relaxation techniques very valuable solution methods if one uses implicit schemes for the time discretization.

ACKNOWLEDGMENTS: A large part of this research was supported by Army Grant DAAL03-86-K-0138, for which we thank Drs. J. Chandra and V. Mirelli. We would like to thank also C. Asano, J. Cahouet, J. Goussebaile, L. J. Hayes, F. Hussain, P. LeTallec, M. Luskin, J. Periaux, O. Pironneau, M. F. Wheeler. The support of CRAY Research is gratefully acknowledged, since the CRAY-XMP has been instrumental for solving the test problems presented here. Last, but not least, we would like to thank L. M. Brooks for her careful processing of this paper.

References

[1] Marchuk, G. I., *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[2] Peaceman, D. M. and Rachford, H.M., The numerical solution of parabolic and elliptic differential equations, *Journal of the SIAM* 3 , 28-41 (1955).

[3] Strang, G., On the construction and comparison of difference schemes, *SIAM J. Num. Anal.* 5 , 506-517 (1968).

[4] Beale, J. T. and Majda, A., Rates of convergence for viscous splitting of the Navier-Stokes equations, *Math. Comp.* 37 , 243-260 (1981).

[5] LeVeque, R., *Time Split Methods for Partial Differential Equations*, Ph.D. Thesis, Computer Science Department, Stanford University, Stanford, California 1982.

[6] LeVeque, R. and Oliger, J., Numerical methods based on additive splittings for hyperbolic partial differential equations, *Math. Comp.* 40 , 469-497 (1983).

[7] Lions, P. L. and Mercier, B., Splitting algorithms for the sum of two operators, *SIAM J. Num. Anal.* 16 , 964-979 (1979).

[8] Gabay, D., Application of the method of multipliers to variational inequalities, in *Augmented Lagrangian Methods*, M. Fortin, R. Glowinski eds. North-Holland, Amsterdam (1983).

[9] Glowinski, R. and Le Tallec, P., *Application of Augmented Lagrangian and Operator Splitting Methods to Nonlinear Mechanics*, to appear as a *SIAM* Monograph.

[10] Schechter, E., Sharp convergence rates for nonlinear product formulas. *Math. of Comp.* 43 , 135-155 (1984).

[11] Douglas, J. and Gunn, J. E., A general formulation of alternating direction methods, *Numer. Math* 6 , 428 (1964).

[12] Varga, R., *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, New York 1976.

[13] Douglas, J. and Rachford, H. H., On the numerical solution of the heat equation problem in 2nd and 3rd space variables, *Trans. A.M.S.* 82 , 421-439 (1956).

[14] Glowinski, R., *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York 1984.

[15] Glowinski, R., Splitting methods for the numerical solution of the incompressible Navier-Stokes equations, in *Vistas in Applied Mathematics*, A. V. Balakrishan, A. A. Doronitsyn, J. L. Lions, eds., Optimization Software, New York, 57-95 (1986).

[16] Bristeau, M. O., Glowinski, R., Mantel, B., Periaux, J. and Perrier, P., Numerical methods for incompressible and compressible Navier-Stokes problems, in *Finite Elements in Fluids* 6 , R. H. Gallagher, G. Carey, J.T. Oden, O. C. Zienkiewicz eds., J. Wiley, Chichester, 1-40 (1985).

[17] Glowinski, R., Viscous flow simulations by finite element methods and related numerical techniques, in *Progress and Supercomputing in Computational Fluid Dynamics*, E. M. Murman, S.S. Abarbanel eds., Birkhauser, 173-210 (1985).

[18] Bristeau, M. O., Glowinski, R. and Periaux, J., Numerical methods for the Navier-Stokes equations. Applications to the simulation of compressible and incompressible viscous flows, *Research Report UH/MD-4*, Department of Mathematics, University of Houston, Houston, Texas, 1987; to appear in *Computer Physics Reports*, North-Holland, 1987.

[19] Girault, V. and Raviart, P. A., *Finite Element Approximation of the Navier-Stokes Equations*, Springer-Verlag, Heidelberg 1986.

[20] Cahouet, J. and Chabard, J. P., Multi-domains and multi-solvers finite element approach for the Stokes problem, in *Innovative Numerical Methods in Engineering*, R. P. Shaw, J. Periaux, A. Chaudonet, J. Wu, C. Marino, C. A. Brebbia eds., Springer-Verlag, Berlin 317-322 (1986).

[21] Hauguel, A. and Cahouet, J., *Finite Element Methods for Incompressible Navier-Stokes Equations and for Shallow Water Equations*, Lecture Notes at the Von Karman Institute, March 1986.

[22] Glowinski, R., Goussebaile, J. and Labadie, G. (eds.), *Numerical Methods for the Stokes Problem; Application to Compressible and Incompressible Viscous Flow Simulation*, to appear.

[23] Glowinski, R. and Le Tallec, P., Augmented Lagrangian methods for the solution of variational problems, *MRC Report* 2965, Mathematics Research Center, University of Wisconsin, Madison (1987).

[24] Hart, R. and Kinderlehrer, D., Mathematical question of liquid crystal theory, in *Theory and Applications of Liquid Crystals*, J. L. Ericksen, D. Kinderhrer eds., *IMA Volumes in Mathematics and Applications* 5 , Springer-Verlag, to appear.

[25] Brezis, H., Coron, J. M. and Lieb, E., Harmonic maps with defect, *Comm. Math. Phys.*, to appear.

[26] De Gennes, P. G., *The Physics of Liquid Crystals*, Oxford 1974.

[27] Ericksen, J. L., Equilibrium theory of liquid crystals, in *Advances in Liquid Crystals*, G. H. Brown ed., 2 , Academic Press, 233-298 (1976).

[28] Chandrasekhar, S., *Liquid Crystals*, Cambridge 1977.

[29] Cohen, R., Hart, R., Kinderlehrer, D., Lin, S. Y. and Luskin, M., Minimum energy configurations for liquid crystals: Computational results, *IMA* Preprint 250 (1986), and in *Theory and Applications of Liquid Crystals*, J. L. Ericksen, D. Kinderlehrer eds., *IMA Volumes in Mathematics and Applications* 5 , Springer-Verlag, to appear.

[30] S. Y. Lin, M. Luskin, *Relaxation Methods for Liquid Crystal Problems*, *IMA* Preprint 332, Institute for Mathematics and Applications, University of Minnesota, Minneapolis, June 1987.

[31] Ewing, R. E., Russell, T. F., Wheeler, M. F., Convergence analysis of an approximation of miscble displacement in porous media by mixed finite element and a modified method of characteristics, *Computer Methods in Applied Mechanics and Engineering*, 47 , (1984), pp. 72-91.

[32] Pironneau, O., On the transport-diffusion algorithm and its application to the Navier-Stokes equations, *Numerish Math.*, 38, (1982), pp. 309-332.

[33] Russel, T. F., Time stepping along characteristics with incomplete iteration for a Galerkin approximation of miscble displacement in porous media, *SIAM J. Numer. Anal.*, 22 , (1985), 5, pp.970-1013.

Algorithms for Rational Spline Curves

Klaus Höllig[1]
Computer Sciences Department and
Mathematics Research Center
University of Wisconsin-Madison

ABSTRACT. The theory of univariate splines is well understood. However, applying the standard techniques for spline **functions** does not make use of some important features of piecewise polynomial **curves**. Since curves are invariant under reparametrization, the smoothness conditions for splines are less restrictive and standard approximation methods can be improved. This note discusses rational cubic spline curves and describes in particular two basic algorithms: the construction of smooth splines from control points and Hermite interpolation.

1. Introduction

We first review briefly two basic algorithms for "standard" cubic spline curves as a preparation for the generalizations to be discussed in the following sections. These algorithms are best described using the Bézier form for polynomials which allows a particularly simple characterization of smoothness constraints for splines. The Bézier coefficients b of a cubic polynomial p are defined by

$$p(t) = \sum_{\nu=0}^{3} b_\nu \, B_\nu(t), \quad 0 \le t \le 1,$$

where $B_\nu(t) := \binom{3}{\nu} t^\nu (1-t)^\nu$ are the Bernstein polynomials. Therefore, as is illustrated in Figure 1, a piecewise cubic spline curve p can be represented by a sequence of Bézier coefficients

$$b_\nu^j, \quad \nu = 0, \ldots, 3, \quad j = 0, \ldots, J.$$

It is assumed that $b_3^{j-1} = b_0^j$, i.e. that the curve segments join continuously. Continuity of the first and second derivatives of the parametrization is equivalent to the conditions

$$b_1^+ - b_0^+ = b_3^- - b_2^-$$
$$(b_2^+ - b_1^+) - (b_1^- - b_0^+) = (b_3^- - b_2^-) - (b_2^- - b_1^-)$$

where $b^\pm$ denote the Bézier coefficients of two adjacent curve segments. The particular form of these conditions yields a very simple algorithm for constructing the Bézier coefficients of twice continuously differentiable spline parametrizations from control points (cf. Figure 2).

[1] supported by the United States Army under Contract No. DAAG29-80-C-0041 and sponsored by the National Science Foundation under Grant No. DMS-8351187

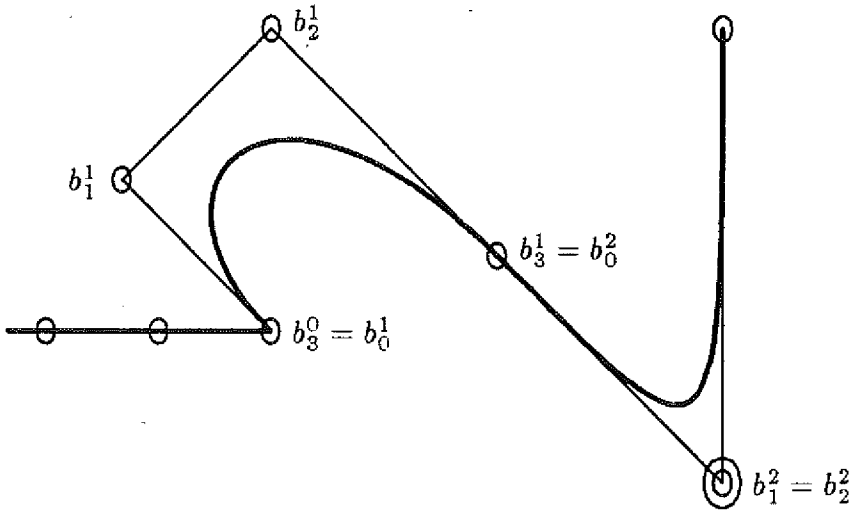


Figure 1. Bézier polygon of a cubic spline curve

Algorithm 1. $(c \Rightarrow b)$ The Bézier coefficients b corresponding to a sequence of control points c are given by

$$b_1^j := (2c^{j+1} + c^{j+2})/3, \quad b_2^{j-1} := (2c^{j+1} + c^j)/3$$
$$b_3^{j-1} = b_0^j := (b_2^{j-1} + b_1^j)/2.$$

Combining the steps in Algorithm 1, $f^j := b_3^{j-1} = b_0^j$ can be expressed in terms of the control points,

$$f^j = (c^j + 4c^{j+1} + c^{j+2})/6 \tag{1}$$

which yields

Algorithm 2. $(f \Rightarrow c)$ The control points c of the natural spline interpolant (which has zero curvature at the endpoints) corresponding to the data f^j, $j = 0, \ldots, J$, are computed by solving the linear system (1) for $j = 1, \ldots, J-1$ with the end conditions

$$c^1 = f^0, \quad c^0 = 2c^1 - c^2$$
$$c^{J+1} = f^J, \quad c^{J+2} = 2c^{J+1} - c^J.$$

Figure 3 shows an example which illustrates a slight disadvantage of the method: possible oscillations near inflection points.

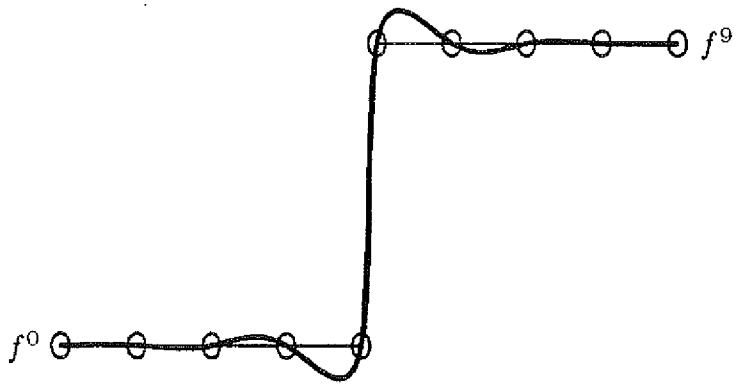


Figure 3. Natural spline interpolant

A piecewise rational curve is represented by a sequence of coefficients and weights,

$$(b_\nu^j, w_\nu^j), \quad \nu = 0, \ldots, 3, \quad j = 0, \ldots, J,$$

where, for continuity, $b_3^{j-1} = b_0^j$. To characterize higher order smoothness, we recall the definition of

Smoothness for Curves. Smoothness of a curve, $t \mapsto f(t) \in \mathbb{R}^3$, is characterized in terms of differentiability with respect to arclength $s := \int^t |f'|$. Since $dt/ds = 1/|f'|$ where $|\ |$ denotes the length of a vector, the first and second derivatives of f with respect to arclength are given by

$$\frac{d}{ds} f = f'/|f'|$$
$$\frac{d^2}{ds^2} f = (|f'|^2 f'' - (f' \cdot f'')f')/|f'|^4.$$

Taking the cross product of the second equation with $f'/|f'|$, this means that C^2-continuity is equivalent to continuity of the vectors

$$\xi_f := f'/|f'|, \quad (\kappa\eta)_f := f' \times f''/|f'|^3.$$

The vector ξ is the unit tangent vector, κ is the curvature and η is the binormal vector which is a unit vector orthogonal to the osculating plane.

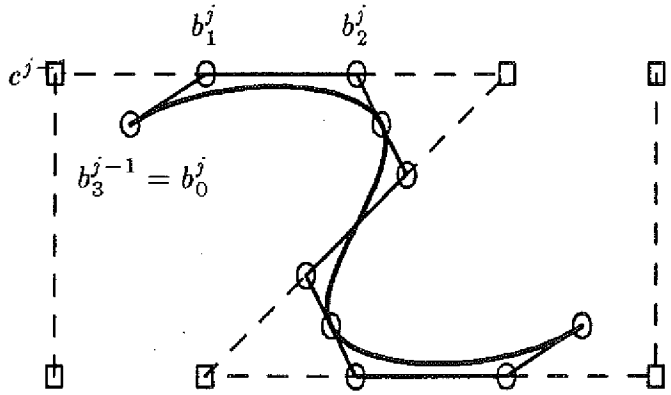


Figure 2. Control polygon, Bézier coefficients and corresponding spline curve

While straightforward, the above algorithms do not make use of the additional flexibility due to the weaker smoothness constraints for curves. This observation has led to the development of β-splines and interesting new approximation and design techniques (cf. [BBB80], [Bö85]). So far, the new geometric ideas have been primarily applied to polynomial splines. We discuss in this note the generalization of the basic algorithms to rational cubic splines. First, we describe the rational Bézier form in Section 2. Then, in Section 3 we discuss the analogues of Algorithms 1 and 2. Section 4 lists MACSYMA computations which establish the main result of this note.

2. Rational Bézier form

We review the definition and some basic facts about the Bézier form and refer to [FP79] for details. The Bézier form of a rational cubic parametrization r is defined as

$$r(t) = \frac{p}{q} = \frac{\sum_{\nu=0}^{3} w_\nu b_\nu B_\nu(t)}{\sum_{\nu=0}^{3} w_\nu B_\nu(t)}, \quad 0 \le t \le 1, \tag{2}$$

where the coefficients b_ν are vectors in $\mathbb{R}^3$ and the weights w_ν are positive numbers. The homogeneous form, i.e. multiplication with the weights in the numerator, simplifies the algebra and geometric interpretation. As in the polynomial case, the control polygon is tangent to the curve at the end points. The weights control the influence of the corresponding coefficients, i.e. increasing w_ν "pulls" the curve towards the coefficient b_ν as is illustrated in Figure 4.

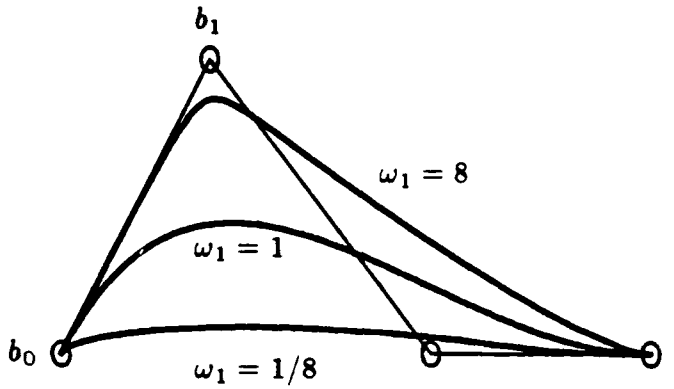


Figure 4. Rational Bézier form

Computing the vectors ξ and η for the parametrization (2) at the endpoints gives

$$\begin{aligned} \xi_r(0) &= \frac{b_1 - b_0}{|b_1 - b_0|}, \quad (\kappa\eta)_r(0) = \frac{2}{3}\,\frac{w_0 w_2}{w_1^2}\,\frac{(b_1 - b_0) \times (b_2 - b_1)}{|b_1 - b_0|^3} \\ \xi_r(1) &= \frac{b_3 - b_2}{|b_3 - b_2|}, \quad (\kappa\eta)_r(1) = \frac{2}{3}\,\frac{w_1 w_3}{w_2^2}\,\frac{(b_3 - b_2) \times (b_1 - b_2)}{|b_3 - b_2|^3}. \end{aligned} \tag{3}$$

Therefore, two adjacent curve segments with Bézier coefficients $b^{\pm}$ and weights $w^{\pm}$ join twice continuously differentiable at $b_3^- = b_0^+$ if the following two conditions are satisfied:

(C1) $b_2^-,\ b_3^- = b_0^+,\ b_1^-$ are collinear;

(C2) $b_1^-,\ b_2^-,\ b_3^- = b_0^+,\ b_1^+,\ b_2^+$ lie in a half plane and the parallelograms $R_{\pm}$ in Figure 5 satisfy

$$\frac{w_1^- w_3^-}{(w_2^-)^2}\,\frac{\text{area}(R_-)}{|b_3^- - b_2^-|^3} = \frac{w_0^+ w_2^+}{(w_1^+)^2}\,\frac{\text{area}(R_+)}{|b_1^+ - b_0^+|^3}. \tag{4}$$

3. Control points and interpolation

The geometric description of the smoothness conditions easily yields the analogue of Algorithm 1 which is a variant of the corresponding method in the polynomial case [Bõ85]. Denote by $\beta_{\pm}$ and $\delta_{\pm}$ the relative length of adjacent line segments as is indicated in Figure 5. For example,

$$|b_1^+ - b_0^-| : |b_3^- - b_2^-| = \delta_+ : \delta_-.$$

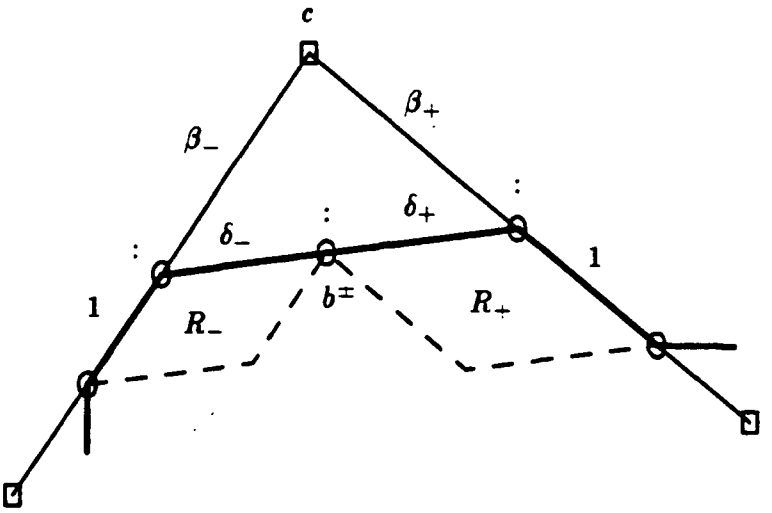


Figure 5. Geometric smoothness constraints

Then, since $\text{area}(R_+) : \text{area}(R_-) = (\delta_+\beta_-) : (\delta_-\beta_+)$, condition (4) becomes

$$\delta^2 := (\delta_+/\delta_-)^2 = (\beta_-/\beta_+)\ \frac{w_0^+ w_2^+}{(w_1^+)^2}\ \frac{(w_2^-)^2}{w_1^- w_3^-}. \tag{5}$$

This means, one can select the points b_1^-, b_2^-, b_1^+, b_2^+ and the weights $w^{\pm}$ essentially arbitrarily and it is then possible to select δ, and hence $b_3^- = b_0^+$, so that the smoothness conditions (C) are satisfied. This yields the following algorithm.

Algorithm I. $(c, w, \beta \Rightarrow b)$ The Bézier coefficients b_ν^j of a piecewise rational spline curve corresponding to the sequence of control points c^j, weights w_ν^j and parameters $\beta_\pm^j > 0$ are given by

$$b_1^j := ((1 + \beta_-^{j+2})c^{j+1} + \beta_+^{j+1}c^{j+2})/(1 + \beta_+^{j+1} + \beta_-^{j-2})$$
$$b_2^{j-1} := ((1 - \beta_-^j)c^{j+1} - \beta_-^{j+1}c^j)/(1 + \beta_-^{j+1} - \beta_-^j)$$
$$b_3^{j-1} = b_0^j := (\delta^{j-1}b_2^{j-1} + b_1^j)/(1 - \delta^{j-1})$$

where δ^{j-1} is defined in terms of w^{j-1}, w^j, $\beta_\pm^{j-1}$ according to (5).

Algorithm 1 is a special case corresponding to $w_\nu^j = \beta_\nu^j = 1$. The weights w and the parameters β permit local control of the "shape" of the curve while keeping the control points fixed. This is illustrated in Figure 6. Decreasing β increases the curvature at the knots and the curve approaches the "control polygon" which connects the points c^j. Increasing a particular weight stresses the influence of the corresponding Bézier coefficient. If this additional flexibility is not needed, the parameters can be set according to suitable optimality criteria.

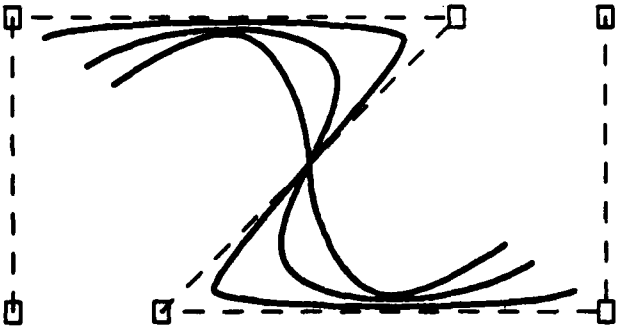


Figure 6. Control points and corresponding rational spline curve for $\beta = 1/4, 1, 4$

Algorithm 2 for interpolation requires the solution of a linear system, i.e. changes in the data have a global influence. Using the additional degrees of freedom due to the weaker smoothness constraints, it is possible to construct smooth interpolants by a local method. This method is suggested by the expressions (3) for ξ and η. Setting $\delta_i := |b_{1+2i} - b_{2i}|$, $f^i := r(i)$, $\xi^i := \xi_r(i)$, $(\kappa\eta)^i := (\kappa\eta)_r(i)$ for $i = 0, 1$ and substituting

$$b_2 - b_1 = (f^1 - f^0) - \delta_0 \xi^0 - \delta_1 \xi^1,$$

the equations for $\kappa\eta$ in (3) can be rewritten as

$$(\kappa\eta)^i = (-)^i \varrho_i \xi^i \times (f^1 - f^0) + \sigma_i \xi^1 \times \xi^0, \quad i = 0, 1, \tag{E}$$

where

$$\varrho_i := \frac{2}{3} \frac{w_i w_{2+i}}{w_{1-i}^2} \frac{1}{\delta_i^2}, \quad \sigma_i := \varrho_i \delta_{1-i}. \tag{6}$$

Since both sides of the i-th equation are orthogonal to ξ^i, the equations (E) are equivalent to a 4×4 linear system for ϱ and σ. This system has a solution with $\varrho, \sigma > 0$ if

(A) η^i lies in the interior of the cone spanned by
$(-)^i \xi^i \times (f^1 - f_0)$ and $\xi^1 \times \xi^0, \quad i = 0, 1.$

Choosing $w_1 := w_2 := 0$, the remaining weights w_0, w_3 and the parameters δ can be expressed in terms of ϱ and σ,

$$\delta_{1-i} = \sigma_i / \varrho_i, \quad w_{3i} = (3/2) \delta_i^2 \varrho_i. \tag{7}$$

The corresponding method described in Algorithm II is a generalization of Hermite interpolation.

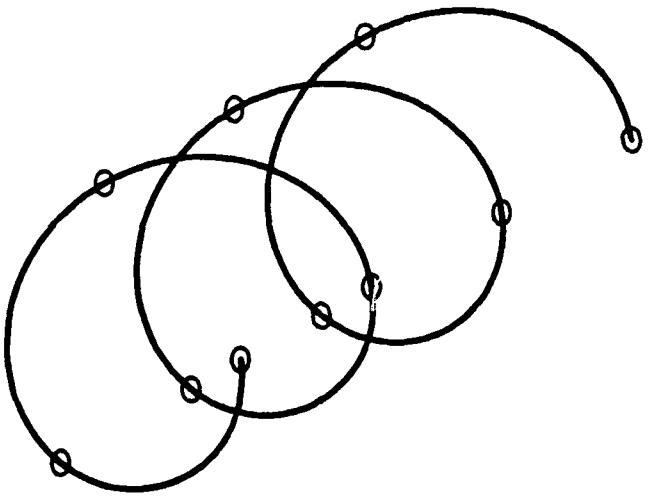


Figure 7. Rational spline interpolant of a helix

Algorithm II. $(f, \xi, \kappa\eta \Rightarrow b, w)$ The Bézier coefficients b_ν and weights w_0, w_3 of the j-th segment of a piecewise cubic rational spline r_f which matches the unit tangent vectors ξ^j and the vectors $(\kappa\eta)^j$ at the points f^j can be determined by solving the system (E) with

$$f^i := f^{j+i}, \quad \xi^i := \xi^{j+i}, \quad \eta^i := \eta^{j+i}$$

provided that condition (A) holds.

The following Theorem shows that, for data corresponding to a smooth curve, condition (A) is satisfied if the interpolation points f^j are sufficiently close. Moreover, the interpolant is of high order accuracy and has good shape preserving properties for smooth data. This is illustrated in Figure 7. If condition (A) is not valid for a particular curve segment, then the given curvatures and binormals cannot be interpolated and have to be modified or possibly chosen differently for adjacent segments. An interesting open problem is whether for given points f the vectors ξ and $\kappa\eta$ can be chosen so that (A) is valid and the resulting scheme remains accurate and shape preserving.

Theorem. Assume that the data f^j, ξ^j, η^j in Algorithm II correspond to a smooth curve f with nonvanishing curvature κ and torsion τ (cf. FP79, p. 102 for definitions). If the distance $h := \max |f^j - f^{j-1}|$ between adjacent points is sufficiently small, then for each pair of adjacent points, condition (A) is satisfied and hence the system (E) has a unique solution with $\varrho, \sigma > 0$. Moreover, the corresponding piecewise rational interpolant r_f is 6-th order accurate, i.e.

$$\operatorname{dist}(f, r_f) = O(h^6).$$

For planar curves, a similar result was obtained in [BHS87] for interpolation with piecewise cubic polynomials. For the rational case, the proof is somewhat simpler, since the weights w and the parameters δ can be expressed explicitly in terms of the data.

Proof. Consider a typical curve segment of the interpolant, e.g. corresponding to the end points f^0, f^1. Without loss we assume that f is parametrized with respect to arclength s and that

$$f^0 = [0 \quad 0 \quad 0], \quad \xi^0 = [1 \quad 0 \quad 0], \quad \eta^0 = [0 \quad 0 \quad 1], \tag{8}$$

and $f^1 := f(s)$. Denote by $r(t,s) = p(t,s)/q(t,s)$, $0 \le t \le 1$, the rational interpolant. We will show that

(i) for sufficiently small s, the system (E) has a unique solution with $\varrho, \sigma > 0$;

(ii) $q(t,0) = 1$;

(iii) $\left(\partial_t r_1(t,s)/s\right)_{|s=0} = 1$;

(iv) $\partial_t^i[p(t,s),\, q(t,s)] = O(s^i), \quad i = 1,2,3.$

Assertions (i) and (ii) guarantee that r is well defined for small s, i.e. as the distance of the points f^0 and f^1 becomes small. Assertion (iii) implies that the derivative of the first coordinate $x = r_1$ of r satisfies

$$c_0 s \le \partial_t r_1(t,s) \le c_1 s \tag{9}$$

for some constants c_ν and s sufficiently small. In particular, the function r_1 is monotone increasing in t. With $\check{r}_1$ denoting its inverse, i.e. $x = r_1(\check{r}_1(x,s),s)$, the rational interpolant has the equivalent parametrization

$$x \rightarrow R(x,s) := [x \quad r_2(\check{r}_1(x,s),s) \quad r_3(\check{r}_1(x,s),s)].$$

Similarly, f can be parametrized with respect to the first coordinate.

$$x \rightarrow F(x).$$

Since the interpolation conditions are invariant under reparametrization, the unit tangent, curvature and binormal of R and F match at $x_0 := 0$ and $x_1 := r_1(1,s) = f_1(s)$. Using that $R_1(x) = F_1(x) = x$, this implies that the derivatives of R and F match at these points up to second order. From the standard error estimate for interpolation of functions it follows that

$$|R(x,s) - F(x)| = O((x_1 - x_0)^6) = O(s^6)$$

provided that the derivatives of R with respect to x up to order 6 are bounded, uniformly in s. This follows from (iv). To see this we note that

$$R_\nu(r_1(t,s),s) = r_\nu(t,s), \quad x = r_1(t,s).$$

and compute the derivatives of R_ν inductively using the chain rule. This shows that $\partial_x^j R_\nu(x,s)$ is a sum of terms of the form

$$r_\nu^{(k)} r_1^{(\ell_1)} \cdot \ldots \cdot r_1^{(\ell_m)} / (r_1^{(1)})^{k+\ell_1+\cdots+\ell_m}$$

where superscripts denote differentiation with respect to t and all functions are evaluated at (t,s) with $t = r_1(x,s)$. Since $r^{(j)}$ is a sum of terms of the form

$$p^{(k)} q^{(\ell_1)} \cdot \ldots \cdot q^{(\ell_m)} / q^{m+1}, \quad j = k + \ell_1 + \cdots + \ell_m,$$

the boundedness of R_ν is a consequence of (ii), (iv) and (9).

It remains to verify assertions (i)-(iv). This requires elaborate Taylor expansions which are done via MACSYMA as is described in the final section.

4. MACSYMA computations

Below we list a MACSYMA program for proving (i)-(iv) of the previous section. The computation is divided into four main steps: Taylor expansion of the data; solution of equations (E): Bézier form of r; verification of (i)-(iv). To speed up the computations, we use Taylor expansion to simplify intermediate results. The order of truncation will be justified at the end of this section.

Auxiliary functions. The following auxiliary functions will be used in the program: (c1) is de Casteljau's algorithm for evaluating a polynomial at t from its Bézier coefficients b; (c2) is the vector product; (c3) generates the first $n-1$ terms of a power series with coefficients a_i; (c4) computes the Taylor expansion of the solution of the differential equation $x'(t) = f(x(t))$, $x(0) = x0$.

```
(c1) bezier_form(b,n,t) :=
        if n=0 then row(b,1)
        else t*bezier_form(submatrix(1,b),n-1,t)
                - (1-t)*bezier_form(submatrix(n-1,b),n-1,t)$

(c2) cross_product(a,b) :=
        [a[2]*b[3]-a[3]*b[2], a[3]*b[1]-a[1]*b[3], a[1]*b[2]-a[2]*b[1]]$

(c3) power_series(a,n,t) :=
        if n=0 then a[0]
        else power_series(a,n-1,t) + a[n]*t^n/factorial(n)$
```

```
(c4) solve_ode(f,x0,n,t) :=
        if n=0 then x0
        else (x1: solve_ode(f,x0,n-1,t),
                x1 + subst(0,t,diff(f(x1),t,n-1))*t^n/factorial(n))$
```

Taylor expansion of the data. The data at the left endpoint are given by (8). To obtain Taylor expansions for $f^1 = f(s)$, $\xi^1 = \xi(s)$ and $(\kappa\eta)^1 = \kappa(s)\eta(s)$, (c10) approximately solves the Frenet differential equations [FP79, p.103],

$$\begin{aligned} f' &= \xi \\ \xi' &= \kappa\varsigma \\ \varsigma' &= \tau\eta - \kappa\xi \\ \eta' &= -\tau\varsigma, \end{aligned}$$

where ς with $\varsigma(0) := [0 \quad 1 \quad 1]$ is the normal vector. This yields Taylor expansions for the data in terms of the Taylor coefficients u_i and v_i of curvature and torsion (cf. (c6), (c7)).

```
(c5) (kappa(s) := power_series(u,5,s), tau(s) := power_series(v,5,s))$

(c6) kappa(s);
(d6)        u5*s^5/120 + u4*s^4/24 + u3*s^3/6 + u2*s^2/2 + u1*s + u0

(c7) tau(s);
(d7)        v5*s^5/120 + v4*s^4/24 + v3*s^3/6 + v2*s^2/2 + v1*s + v0

(c8) (f[0]: [0,0,0], xi[0]: [1,0,0], zeta[0]: [0,1,0], eta[0]: [0,0,1])$

(c9) g(a) := [a[2], kappa(s)*a[3], tau(s)*a[4]-kappa(s)*a[2], -tau(s)*a[3]]$

(c10) ffs: solve_ode(g,[f[0], xi[0], zeta[0], eta[0]],6,s)$

(c11) (f[1]: ffs[1], xi[1]: ffs[2], zeta[1]: ffs[3], eta[1]: ffs[4])$

(c12) (kappa[0]: kappa(0), kappa[1]: kappa(s))$
```

Solution of equations (E). All vectors in the i-th equation in (E) are orthogonal to ξ^i and therefore the i-th equation is equivalent to the 2×2 system

$$eqn^i_{\nu,1} = eqn^i_{\nu,2}\,\varrho_i + eqn^i_{\nu,3}\,\sigma_i, \quad \nu = 0, 1, \tag{10}$$

obtained in (c13) by forming the dot product of the i-th equation with the vectors η^i and ξ^{1-i}. (c14) solves the system (10) by backward substitution, using that $eqn^i_{3,3} = 0$. The parameters δ and weights w are computed in (c15) and (c18) according to (7).

```
(c13) for i:0 thru 1 do
        eqn[i]: (matrix1: matrix(eta[i],xi[1-i]),
                 matrix2: matrix(kappa[i]*eta[i],
                                 (-1)^i*cross_product(xi[i],f[1]-f[0]),
                                 cross_product(xi[1],xi[0])),
                 ratsimp(taylor(matrix1.transpose(matrix2),s,0,6)))$

(c14) for i:0 thru 1 do
        (rho[i]: eqn[i][2,1]/eqn[i][2,2],
        sigma[i]: (eqn[i][1,1]-eqn[i][1,2]*rho[i])/eqn[i][1,3])$

(c15) for i:0 thru 1 do
        delta[1-i]: taylor(ratsimp(sigma[i]/rho[i]),s,0,3)$

(c16) delta[0];
(d16)     s/3 + (u0 v1 + 2 v0 u1) s^2/(36 u0 v0) + (9 u0^2 v0 v2
                 + 12 u0 v0^2 u2 - 10 u0^2 v1^2 + 2 u0 v0 u1 v1
                 - 10 v0^2 u1^2 + 6 u0^2 v0^4 + 6 u0^4 v0^2) s^3/(540 u0^2 v0^2)

(c17) delta[1]-delta[0];
(d17)     -(u0 v1 + 2 v0 u1) s^2/(18 u0 v0) - (u0^2 v0 v2 + 2 u0 v0^2 u2
                 - u0^2 v1^2 - 2 v0^2 u1^2) s^3/(36 u0^2 v0^2)

(c18) for i:0 thru 1 do
        w[i]: taylor(ratsimp((3/2)*rho[i]*delta[i]^2),s,0,2)$

(c19) w[0];
(d19)     1 + (24 u0^2 v0 v2 + 12 u0 v0^2 u2 - 35 u0^2 v1^2 - 8 u0 v0 u1 v1
                 - 20 v0^2 u1^2 + 36 u0^2 v0^4 + 36 u0^4 v0^2) s^2/(720 u0^2 v0^2)

(c20) w[1]-w[0];
(d20)     0
```

Bézier form of r. Statements (c21) and (c22) define the polynomials $p(\cdot, s)$ and $q(\cdot, s)$ using de Casteljau's Algorithms in terms of the Bézier coefficients.

```
(c21) p: (wb: matrix(w[0]*f[0], f[0]+delta[0]*xi[0],
                     f[1]-delta[1]*xi[1], w[1]*f[1]),
          ratsimp(taylor(bezier_form(wb,3,t)[1],s,0,2)))$

(c22) q: (w: matrix(w[0], 1, 1, w[1]),
          ratsimp(taylor(bezier_form(w,3,t)[1,1],s,0,2)))$
```

Verification of (i)-(iv). As is shown by (c23)-(c26), the dominant part of the system (10), as $s \to 0$ is given by

$$\begin{bmatrix} u_0 \\ u_0^2 v_0 s^2/2 \end{bmatrix} = \begin{bmatrix} u_0 s^2/2 & -u_0 s \\ u_0^2 v_0 s^4/12 & 0 \end{bmatrix} \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \tag{11}$$

which proves (i). Clearly, (c27) proves (ii). (c28) computes the numerator of $\partial_t r_1(t,s)/s$, where $r_1 = p_1/q$, and evaluates it at $s = 0$. In conjunction with (ii) this establishes (iii). Assertion (iv) is equivalent to the statement that

$$\partial_t^i \partial_s^j [p,q](t,s) = 0, \quad j < i \leq 3.$$

This is checked by (c29) which displays $(\partial_t^i [p(t,s), q(t,s)])/s^{i-1}$ evaluated at $s = 0$.

```
(c23) taylor(eqn[0][1],s,0,2);
(d23)       [u_0, u_0 s^2/2, -u_0 s - u_1 s^2/2]

(c24) taylor(eqn[1][1]-eqn[0][1],s,0,2));
(d24) [u_1 s + u_2 s^2/2, 0, 0]

(c25) taylor(eqn[0][2],s,0,4);
(d25)       [v_0 u_0^2 s^2/2 + (v_1 u_0^2 + 2u_1 v_0 u_0) s^3/6 - (v_0 u_0^4 + (v_0^3 - v_2) u_0^2
                    -(3u_2 v_0 + 3v_1 u_1) u_0) s^4/24, v_0 u_0^2 s^4/12, 0]

(c26) taylor(eqn[1][2]-eqn[0][2],s,0,4);
(d26)       [(u_0^2 v_1 + 2u_0 v_0 u_1) s^3/6 + (u_0^2 v_2 + 2u_0 v_0 u_2
                    +4u_0 u_1 v_1 + 2v_0 u_1^2) s^4/12, 0, 0]

(c27) subst(0,s,q);
(d27)       1

(c28) subst(0,s,ratsimp((diff(p[1],t)*q-p[1]*diff(q,t))/s)):
(d28)       1
(c29) for i:1 thru 3 do
          disp(subst(0,s,ratsimp(diff([p,q],t,i)/s^(i-1)))):
                    0. 0. 0, 0
                    0. 0. 0, 0
                    0. 0. 0, 0]
(d29)       done
```

It remains to justify the various orders of truncation in the intermediate Taylor expansions. Since multiplication never decreases the order of validity of truncated Taylor

expansions, we must only consider (c15) and (c18). To indicate the range of significant terms in an expansion

$$\varphi(s) = \varphi_j s^j + \varphi_{j+1} s^{j+1} + \ldots,$$

we use the notation

$$\varphi \sim [j, J]$$

if the coefficients up to index J agree with the exact expansion of φ. By (11), and since the data are computed exact up to order 6 by (c10), the coefficients in the system (10) satisfy

$$eqn^i \sim \begin{bmatrix} [0,6] & [2,6] & [1,6] \\ [2,6] & [4,6] & 0 \end{bmatrix}.$$

This shows that

$$\begin{aligned} \varrho &\sim [2,6]/[4,6] \sim [-2,0] \\ \sigma &\sim ([0,6] - [2,6] * [-2,0])/[1,6] \sim [-1,1] \\ \delta &\sim [-1,1]/[-2,0] \sim [1,3] \\ w &\sim [-2,0] * [1,3]^2 \sim [0,2] \end{aligned}$$

and hence all subsequent expansions are exact at least up to order 2 which is what is needed for the proof of (iv).

References

[BBB85] B. Barsky, R.H. Bartels and J.C. Beatty, An introduction to the use of spline functions in computer graphics. ACM Siggraph, San Francisco, 1985.

[Bö85] W. Böhm, Curvature splines, Computer-Aided Geometric Design **2** (1985), 313–324.

[BHS87] C. de Boor, K. Höllig and M. Sabin, High accuracy geometric Hermite interpolation, Proc. Conf. on CAD, G. Farin ed., Oberwolfach 1987.

[FP79] I.D. Faux and M.J. Pratt, Computational Geometry for Design and Manufacture, Ellis Horwood, 1979.

Convexity-Preserving Quasi-interpolation and Interpolation by Box Spline Surfaces

Charles K. Chui[1]
Department of Mathematics and
Center for Approximation Theory
Texas A & M University
College Station, TX 77843

Harvey Diamond[2]
Department of Mathematics
West Virginia University
Morgantown, WV 26506

Louise A. Raphael[2,3]
Department of Mathematics
Howard University
Washington, D.C. 20059

Abstract. Given a strictly convex function of two variables f(x,y), a C^1 box spline series approximant of this function, based on data {f(ih,jh)} given at points uniformly spaced by h in the x and y directions, will also be convex if h is sufficiently small. Explicit numerical bounds on h are provided in this paper which guarantee convexity of these box spline series approximants. The bounds are of the form $h \le c\varepsilon/L$ where $\varepsilon>0$ is the minimum value of $D_u^2 f$, L is a Lipschitz constant associated with the continuity of the second derivatives of f, and the constant c depends on the particular box spline approximant being used.

I. Formulation

Given data $\{f(ih,jh)\} \equiv \{f_{ij}\}$, $\forall\ (i,j) \in \mathbb{Z}^2$, representing values of a smooth function f(x,y) at the set of points {(ih,jh)} with uniform spacing of h in the x and y directions, and a C^1 locally supported box spline $\phi(x,y)$, there are various ways of determining the coefficients $\{c_{ij}\}$ of a box spline series

$$(1) \qquad s_h(x,y) \equiv \sum_{i,j} c_{ij}\phi\left(\frac{x-ih}{h}, \frac{y-jh}{h}\right)$$

such that the order of approximation of f by s_h is

$$|f(x,y)-s_h(x,y)| = O(h^3)$$

and the partial derivatives of f are simultaneously approximated with

$$|D^\alpha f(x,y)-D^\alpha s_h(x,y)| = O\left(h^{3-|\alpha|}\right)\ ,\ |\alpha|=1,2\ .$$

---

[1]Supported by NSF Grant No. DMS-8602337 and ARO Contract No. DAAG29-84-K-0154

[2]Supported by ARO Contract No. DAAG29-84-G-0004

[3]Present address: National Science Foundation, Washington, D.C. 20550

In particular, if we denote by $D_u^2 f$ the second derivative of f in the direction of the unit vector u, we have

(2) $$|D_u^2 f(x,y) - D_u^2 s_h(x,y)| = O(h).$$

If f(x,y) is a strictly convex function, say $D_u^2 f(x,y) \geq \varepsilon > 0$ for some constant ε independent of u and (x,y), then for sufficiently small h, the approximation property (2) implies that the spline approximant s_h is also convex. The purpose of this paper is to provide explicit bounds on h which guarantee convexity of s_h.

The function f is assumed to be strictly convex and to have Lipschitz continuous second derivatives, which is sufficient for (2) to hold. More specifically, we assume

(3)
$$\text{a) } f \in C^2(\mathbb{R}^2)$$
$$\text{b) } D_u^2 f(x,y) \geq \varepsilon > 0 \ \forall\ u,x,y$$
$$\text{c) } |D^\alpha f(x_2,y_2) - D^\alpha f(x_1,y_1)| \leq L\left[(x_2-x_1)^2+(y_2-y_1)^2\right]^{1/2}, \ |\alpha| = 2.$$

The remainder of the paper is as follows. In Section II we introduce box splines and the class of approximants whose convexity-preserving properties will be studied. In Section III we illustrate the estimation techniques which lead to the bounds on h which preserve convexity. In Section IV these bounds are presented. We do not present the details of the derivation of the bounds in this paper. Rather, the reader is referred to the paper [CDR2] for a detailed treatment of shape preservation in bivariate spline approximation.

II. Box Splines and Spline Approximants

Let $V = \{v_1, v_2, .., v_n\}$ be a set of (generally nondistinct) integer vectors in $\mathbb{R}^2$ which also spans $\mathbb{R}^2$. The box spline $\phi_V(x,y)$ is defined to be the probability density of the random linear combination $\Sigma t_i v_i$ where the random n-tuple $(t_1, t_2, .., t_n)$ in $\mathbb{R}^n$ is uniformly distributed in the "box" $[-1/2, 1/2]^n$. Box splines are locally supported and piecewise polynomial. Explicit formulas for the polynomial pieces can be recursively calculated fairly easily; see [CL] for details of the construction and for several examples important to the analysis described in Section III.

For reasons of simplicity, we restrict ourselves as to the possible members of V. We define the vectors $e_1=(1,0)$, $e_2=(0,1)$, $e_3=(1,1)$ and $e_4=(-1,1)$ and require that $v_i \in \{e_1, e_2, e_3, e_4\}$. With this restriction, we use the notation $\phi_{n_1 n_2 n_3 n_4}(x,y)$ instead of $\phi_V(x,y)$, the four indices n_1, n_2, n_3, n_4 indicating the number of times the vectors e_1, e_2, e_3, e_4 appear in V

respectively. We will further assume that $n_1>0$ and $n_2>0$. Finally, if the set **V** is implicit or immaterial we simply write $\phi(x,y)$ for the box spline.

Roughly speaking, the more vectors in **V**, the smoother the box spline ϕ. Specifically,

(4) $\quad \phi \in C^{\rho}$ where $\rho = n - \max\{n_i\} - 2$, $n = \sum_1^4 n_i$

The *spline space* $\mathcal{S}$ is formed by taking linear combinations of the integer translates of $\phi(x,y)$. An element $s(x,y) \in \mathcal{S}$ is given by

$$s(x,y) = \sum_{i,j} c_{ij}\phi(x-i,y-j)$$

The space $\mathcal{S}$ contains all polynomials of degree $\rho+1$. An important property of the spline space is that $\sum_{i,j} \phi(x-i,y-j) \equiv 1$.

Next we define the *scaled spline space* $\mathcal{S}_h$ with elements $s_h(x,y)$ given by

(5) $$s_h(x,y) = \sum_{i,j} c_{ij}\phi\left(\frac{x-ih}{h}, \frac{y-jh}{h}\right)$$

The function $\phi_V\left(\frac{x}{h}, \frac{y}{h}\right)$ will be denoted by $B_h(x,y|V)$, and dropping the **V** for convenience, we may write

(6) $$s_h(x,y) = \sum_{i,j} c_{ij}B_h(x-ih,y-jh)$$

for elements of $\mathcal{S}_h$. The approximation power of the space $\mathcal{S}_h$ is $O(h^{\rho+2})$:

If the partial derivatives of order $\rho+1$ of a function $f(x,y)$ are Lipschitz continuous, there exist functions $s_h \in \mathcal{S}_h$ such that

$$|f(x,y)-s_h(x,y)| = O(h^{\rho+2}).$$

Details on the above results can be found in the survey paper [H].

We introduce next the class of approximants which realize the optimal approximation order of $\mathcal{S}_h$ and whose convexity-preserving properties are the subject of this paper. The results which follow are developed in detail in [CD] and [CDR1].

We denote by F the vector of data values $\{f_{ij}\}$. Define the finite difference operator M acting on data vectors as follows:

$$(MF)_{ij} = \sum_{r,s} m_{rs} f_{i-r,j-s}$$

where

$$m_{ij} = \begin{cases} 1-\phi(0,0) & \text{if } i=j=0 \\ -\phi(i,j) & \text{otherwise} \end{cases}$$

The symmetry of $\phi(x,y)$ about the origin and the fact that $\Sigma\phi(i,j)=1$ imply that M can be expressed as a sum of central second difference operators.

The family of approximants $q_k(f)(x,y)$ is then defined as follows:

$$q_0(f)(x,y) = \sum_{i,j} f_{ij} B_h(x-ih, y-jh)$$

$$q_1(f)(x,y) = \sum_{i,j} (F+MF)_{ij} B_h(x-ih, y-jh) \tag{7}$$

$$q_k(f)(x,y) = \sum_{i,j} (F+MF+\ldots+M^kF)_{ij} B_h(x-ih, y-jh)$$

$$q_\infty(f)(x,y) = \sum_{i,j} (F+MF+\ldots)_{ij} B_h(x-ih, y-jh) = \lim_{k\to\infty} q_k$$

The limit required to calculate q_∞ will exist only if n_3 or n_4 is zero. In this case q_∞ is the unique cardinal interpolant of the data.

The approximation order of the q_k is as follows:

a) *If* $2k+2 < \rho+2$ *then* $|f(x,y)-q_k(f)(x,y)|=O(h^{2k+2})$, *provided that* $f \in C^{2k+1}$ *and the partial derivatives of* f *order* $2k+1$ *are Lipschitz continuous.*

b) *If* $2k+2 \geq \rho+2$ *then* $|f(x,y)-q_k(f)(x,y)|=O(h^{\rho+2})$, *provided that* $f \in C^{2\rho+1}$ *and the partial derivatives of* f *order* $2\rho+1$ *are Lipschitz continuous.*

If $2k+2 \geq \rho+2$ then $q_k(f)(x,y)$ provides the optimal order of approximation and is referred to as a *quasi-interpolant* of f, unless $k=\infty$ in which case q_k is referred to as the *interpolant.*

Our results in Section IV concern the convexity of the q_k.

We require one more result, due to Dahmen and Micchelli [DM]:

(8) *If* V_1 *and* V_2 *are two sets of integer vectors with* $V_1 \subset V_2$, *then if* $\sum_{i,j} c_{ij} B_h(x-ih, y-jh|V_1)$ *is convex, so is* $\sum_{i,j} c_{ij} B_h(x-ih, y-jh|V_2)$.

III. Methods of Analysis

In the analysis of convexity of the q_k we concern ourselves first with the C^1 box splines ϕ_{1111}, ϕ_{221}, and ϕ_{122}. (In the latter two the index n_4 is zero and is omitted.) The convexity of these is relatively simple to investigate as their second derivatives are piecewise constant in the first case and piecewise linear in the other two. We obtain for each the Hessian matrix $H_s(x,y)$ of an arbitrary $s_h(x,y)$ in terms of its coefficients c_{ij}. This is effected with the aid of the formula [H], $v\cdot\nabla\phi_V = \delta_v^*\phi_{V\setminus\{v\}}$ where v is an element of V, ∇ is the gradient, and δ_v^* is the centered difference operator given by $\delta_v^*\phi(\cdot) = \phi(\cdot+v/2)-\phi(\cdot-v/2)$. The Hessian matrices thus obtained characteristically have as their entries a second difference of the coefficients appropriate to the second derivative being calculated, with

perhaps an additional third order difference. For example, using ϕ_{221} we present in Figure 2 the value of $H_s(x,y)$ at the point $(x,y)=(ih+h/2, jh+h/2)$ inside the triangular piece shown in Figure 1 (on this triangular piece, ϕ_{221} is a cubic polynomial and the entries of H_s are linear).

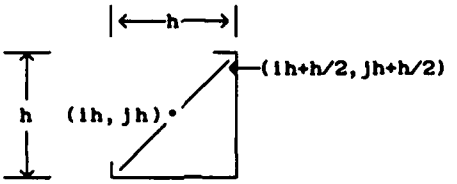
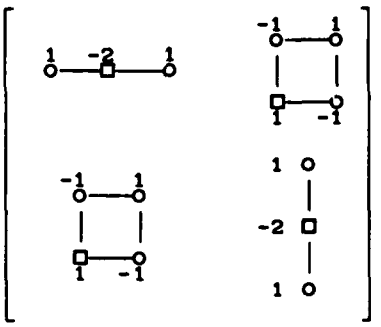


Figure 1

$$H_s(ih+h/2, jh+h/2) = \frac{1}{h^2}$$



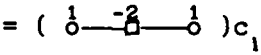
c_{ij}

Figure 2

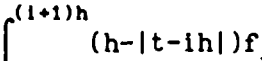
The entries in the matrix of Figure 2 are schematics for difference operators where the symbol "□" identifies the multiplier of c_{ij} and multipliers of adjacent coefficients are shown beside the symbols "o".

Next, the idea is to compare, for the approximants q_k, $D_u^2 s_h = u^T H_s u$ with $D_u^2 f = u^T H_f u$, where u is a unit vector and $H_f = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}$ is the Hessian of $f(x,y)$. If the difference is less than ε (the lower bound on $D_u^2 f$), we can conclude that s_h is convex.

We proceed to illustrate the comparison method. Again, we use as an example the case ϕ_{221} and we compare $H_s(ih+h/2, jh+h/2)$ at the upper corner of the triangle in Figure 1 with $H_f(ih, jh)$ in the case of $s_h(x,y)=q_0(f)(x,y)$. For this choice of s_h we have $c_{ij}=f_{ij}$. Consider now the (1,1) entry of H_s. We have

(9) $H_s^{(1,1)}(ih+h/2, jh+h/2) = ($  $)c_{ij} = ($ $)f_{ij}$

and the latter second difference may be written in integral form:

(10) $($  $)f_{ij} = \frac{1}{h^2} \int_{(i-1)h}^{(i+1)h} (h-|t-ih|) f_{xx}(t, jh)\,dt$

Comparing this quantity with $H_f^{(1,1)}(ih, jh) = f_{xx}(ih, jh)$, we obtain

$$\left|H_s^{(1,1)}(ih+h/2, jh+h/2) - H_f^{(1,1)}(ih, jh)\right| =$$

$$\left| \frac{1}{h^2} \int_{(i-1)h}^{(i+1)h} (h-|t-ih|)\left[f_{xx}(t, jh)-f_{xx}(ih, jh)\right]dt \right| \leq$$

$$\left|\frac{1}{h^2} \int_{(i-1)h}^{(i+1)h} (h-|t-ih|)L|t-ih|dt \right| = hL/3$$

The inequality was obtained by applying the Lipschitz condition in (3) satisfied by f_{xx}.

A similar technique is then used to estimate the analogous differences at the other entries of the Hessian matrices H_s and H_f. We then obtain an estimate of the form

$$(11) \qquad |D_u^2 f(ih, jh) - D_u^2 q_0(f)(ih+h/2, jh+h/2)| \leq u^T \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} u$$

where the entries of the matrix are proportional to hL and we showed above that $\alpha = hL/3$. If we denote by λ the largest eigenvalue of the matrix, then using the assumption $D_u^2 f(ih, jh) \geq \varepsilon$, we have

$$D_u^2 q_0(f)(ih+h/2, jh+h/2) \geq \varepsilon - \lambda \geq 0 \text{ if } \lambda \leq \varepsilon.$$

Since λ has the form hL/c we can obtain an inequality of the form $h \leq c\varepsilon/L$ as a sufficient condition for convexity of q_0 in the triangle of Figure 1, locally at the upper corner. Then similar estimates must be made for the other two corners; these estimates are adequate as $D_u^2 q_0$ is linear in the triangle so positivity at the three corners implies positivity in the interior.

The investigation of sufficient conditions for convexity of q_k, $k > 0$, requires estimates of the size of powers of the finite difference operator M (which is itself a sum of second difference operators) applied to the Hessian matrix in Figure 2 with $c_{ij} = f_{ij}$. These estimates, which involve fourth order differences, are rather complicated so we will not go into the details here, although manipulation of integral representations such as (10) and application of the Lipschitz condition is again the technique. In any case, inequalities of the form (11) can be obtained when $s_h = q_k$, where the entries of the matrix are each proportional to hL, this in turn leading to a sufficient conditon of the form $h \leq c\varepsilon/L$ for convexity of q_k.

The above technique was used to obtain sufficient conditions for convexity of the q_k for ϕ_{1111}, ϕ_{221}, and ϕ_{122} (see Table 1 in Section IV).

Consider now a smoother spline obtained by adding vectors to the vector set V of one of these three fundamental splines. By the result (8) it is sufficient to examine the Hessian matrix of the appropriate fundamental spline. Thus for instance, in examining convexity in the spline space based on ϕ_{222} (obtained by adding the vector e_3 to the vector set V of ϕ_{221}), a sufficient condition for convexity is that $u^T H_s u \geq 0$, where the Hessian H_s used is that in Figure 2. The coefficients c_{ij} however, are those appropriate to the approximants based on ϕ_{222} through the finite difference operator M.

IV. Results

We assume throughout this section that f(x,y) satisfies the hypotheses (3). Table 1 below provides for the fundamental splines ϕ_{1111}, ϕ_{221}, and ϕ_{122}, sufficient conditions on h guaranteeing convexity of $q_k(f)(x,y)$.

	Spline		
k	ϕ_{221}	ϕ_{122}	ϕ_{1111}
0	.909 ε/L	.480 ε/L	.727 ε/L
1	.546 ε/L	.290 ε/L	.425 ε/L
$1 < k < \infty$	.172 ε/L	.097 ε/L	$(1.50+.976k)^{-1}\varepsilon/L$
$k = \infty$	.172 ε/L	..097 ε/L	–

Table 1

A value of h smaller than the tabulated figure guarantees convexity. Since k=1 provides the quasi-interpolant for the splines considered above, the next approximant of interest is the interpolant, $k = \infty$, for which an estimate was computed in the cases of ϕ_{221}, and ϕ_{122}; however this estimate also applies to intermediate values of k. Of course for ϕ_{1111} the approximant for $k=\infty$ does not apply.

The next tables give sufficient conditions for convexity for splines obtained from the fundamental ones above by adding one or more of the vectors e_1, e_2, e_3, e_4 to V.

Three parameters related to a box spline ϕ appear in the table. These are:

a) $\bar{r} = \Sigma\phi(i,j)(i^2+j^2)^{1/2}$

b) #supp(ϕ) = number of points (i,j) at which $\phi(i,j)\neq 0$

c) $a = 1 - \min_{\mu,\nu}\left\{\sum_{m,n}\phi(m,n)\exp(i\mu m+i\nu n)\right\}$ where $i = \sqrt{-1}$, $(\mu,\nu) \in \mathbb{R}^2$

For box splines with $n_3=0$ or $n_4=0$, $a < 1$ always holds. For the other

cases, where $n_3 \geq 1$ and $n_4 \geq 1$, $a \geq 1$ always holds. (We assume that $n_1 \geq 1$ and $n_2 \geq 1$ in all the cases under consideration.) For ϕ_{1111}, as well as in several other cases, a=1. Although it is not proven that this is always true when all four vectors appear in V we will only present estimates based on the assumption a=1.

The Table 2 concerns box splines obtained from ϕ_{221} with $n_4=0$.

k	$n_1 \geq 2,\ n_2 \geq 2,\ n_3 \geq 1,\ n_4 = 0$
0	$(1.10)^{-1}\varepsilon/L$
1	$(1.10+2\bar{r})^{-1}\varepsilon/L$
2	$(1.10+6\bar{r})^{-1}\varepsilon/L$
$3 \leq k \leq \infty$	$\left[1.10+ 6\bar{r} + \frac{2a^2(2-a)}{(1-a)^2}\,\bar{r}\{\#\mathrm{supp}(\phi)\}^{1/2}\right]^{-1}\varepsilon/L$

Table 2

In cases where $n_4=0$ but Table 2 does not apply (i.e. the box spline cannot be derived from ϕ_{221}) the bounds of Table 3, which apply to box splines derived from ϕ_{122}, can be used.

k	$n_1=1$ or $n_2=1,\ n_3 \geq 2,\ n_4 = 0$
0	$(2.52)^{-1}\varepsilon/L$
1	$(2.52+4\bar{r})^{-1}\varepsilon/L$
2	$(2.52+12\bar{r})^{-1}\varepsilon/L$
$3 \leq k \leq \infty$	$\left[2.52+12\bar{r} + \frac{4a^2(2-a)}{(1-a)^2}\,\bar{r}\{\#\mathrm{supp}(\phi)\}^{1/2}\right]^{-1}\varepsilon/L$

Table 3

For box splines with four directions, i.e. box splines for which $n_3 \geq 1$ and $n_4 \geq 1$, as stated above, we assume that a=1. If the box spline in question can be derived from ϕ_{221} we obtain the more favorable bounds on h summarized in Table 4. The first three entries in Table 4 are the same as in Table 2 since these are derived without using the value of a but using only the fact that $\Sigma\phi(i,j)=1$.

k	$n_1 \geq 2,\ n_2 \geq 2,\ n_3 \geq 1,\ n_4 \geq 1$
0	$(1.10)^{-1}\varepsilon/L$
1	$(1.10+2\bar{r})^{-1}\varepsilon/L$
2	$(1.10+6\bar{r})^{-1}\varepsilon/L$
$3 \leq k < \infty$	$\left[1.10+ 6\bar{r} +(k+1)(k-2)\bar{r}\{\#\mathrm{supp}(\phi)\}^{1/2}\right]^{-1}\varepsilon/L$

Table 4

If the box spline cannot be derived from ϕ_{221}, i.e. $n_1=1$ or $n_2=1$, then, using the Hessian of ϕ_{1111} as a sufficient condition for convexity, we obtain the bounds of Table 5.

k	$n_1=1$ or $n_2=1,\ n_3 \geq 1,\ n_4 \geq 1$
0	$(1.64)^{-1}\varepsilon/L$
1	$(1.64+3\bar{r})^{-1}\varepsilon/L$
2	$(1.64+9\bar{r})^{-1}\varepsilon/L$
$3 \leq k < \infty$	$\left[1.64+ 9\bar{r}+\frac{3}{2}(k+1)(k-2)\bar{r}\{\#\mathrm{supp}(\phi)\}^{1/2}\right]^{-1}\varepsilon/L$

Table 5

This completes the summary of the results. Since the estimates in Tables 2-5 are rather general, in specific cases a more careful analysis using the techniques of [CDR2] may result in a sharper bound for h.

References

[CD] C.K. Chui and H. Diamond, A natural formulation of quasi-interpolation by multivariate splines, Proc. Amer. Math. Soc. 99 (1987), 643-646

[CDR1] C.K. Chui, H. Diamond, and L.A. Raphael, Interpolation by multivariate splines, Math. Comp., to appear

[CDR2] C.K. Chui, H. Diamond, and L.A. Raphael, Shape-Preserving quasi-interpolation and interpolation by box spline surfaces, CAT 146, Texas A & M University, 1987

[CL] C.K. Chui and M.J. Lai, Computation of box splines and B-splines on triangulations of nonuniform rectangular partitions, Approximation Theory and Its Applications, to appear

[DM] W. Dahmen and C.A. Micchelli, Convexity of multivariate Bernstein polynomials and box spline surfaces, Studia Scient. Math. Hung., to appear

[H] K. Hollig, Box splines, in **Approximation Theory V**, ed. by C.K. Chui, L.L. Schumaker, and J.D. Ward, Academic Press, New York, 1986, pp. 71-95

KNOT SELECTION FOR LEAST SQUARES APPROXIMATION USING THIN PLATE SPLINES[#]

John R. McMahon[*] and Richard Franke[+]

[*] Department of Mathematics
United States Military Academy
West Point, New York 10996

[+] Department of Mathematics
Naval Postgraduate School
Monterey, California 93943

ABSTRACT. Given a large set of scattered data (x_i,y_i,f_i), a method for selecting a significantly smaller set of knot points which will represent the larger set is described. The algorithm for selection of the knot point locations is based on the minimization of the sum of the squares of the difference between the average number of points per Dirichlet tile and the actual number of points in each tile, subject to the constraint that each knot is located at the centroid of its tile. Using the least squares Thin Plate Spline approximation method for constructing surfaces, various test surfaces are examined and compared to surfaces obtained using the smoothing spline and the bicubic Hermite approximation methods.

I. INTRODUCTION. The problem of fitting a surface to small sets of given data has been addressed in many different ways and several programs are currently available which enable one to deal with the problem effectively. The methods available involve either interpolation or approximation; solving the interpolation problem involves 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 involves a linear combination of functions with no greater complexity than the natural logarithm of the distance function.

[#] The work of the second author was supported in part by the Office of Naval Research under Program Element 61153N, Project NO. BR033-02-WH.

Interpolation of scattered data by the method of TPS was developed from engineering considerations by Harder and Desmairis. [1] It can be thought of as a two dimensional generalization of a 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 function.

A relationship between the basis functions which span a certain higher dimensional function space and the data exists as seen in the one dimensional analogue found in cubic spline interpolation. The term knot refers to the places at which two adjacent cubic polynomials are joined. The particular set of basis functions in a cubic spline interpolation depend on the knot points, as well as the data points. However, in approximation, the data points and the knot points may not necessarily coincide. Furthermore, the particular basis functions found in approximation may easily depend on the knot points, and the data points as well. 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: the knots. Therefore, the problem at hand is one of selecting the knot points, and hence the basis functions.

This approach differs from the use of smoothing splines, which were introduced by Wahba and Wendelberger [3] in the multidimensional case, and called Laplacian Smoothing Splines (LSS). LSS minimize a certain functional which is a linear combination of a term measuring fidelity to the data and one measuring smoothness of the function (a generalization of the usual thin plate spline functional). In this case, there is still one basis function for each data point, but the interpolation condition is relaxed.

Given a 'large' set of data points, (x_i,y_i,f_i), $i = 1,\ldots,N$, we wish to find a smaller set of knot points, (x_j,y_j), $j = 1,\ldots,K$, 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 would essentially be the same.

Approximation by least squares TPS is straightforward. We construct the TPS function

$$F(x,y) = \sum_{j=1}^{K} A_j d_j^2 \log(d_j) + ax + by + c$$

where $d_j^2 = (x-x_j)^2 + (y-y_j)^2$ and the coefficients A_j 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 ith 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 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:

$$\sum_{j=1}^{K} A_j = 0, \quad \sum_{j=1}^{K} A_j x_j = 0, \quad \sum_{j=1}^{K} A_j y_j = 0.$$

Attempts have been made to minimize the error function by considering it to be a function of the knot point locations as well as the coefficients, wherein a total of 3K parameters are involved. As reported on by Schmidt [2], the initial knot configuration was taken to be of tensor product form. The overall minimization process is a large non-linear one, and is complicated by possible coalescense of knots as well as non-unique solutions (as indicated by consideration of one-dimensional cases). Also, the objective function may have many local minima so that avoiding poor local minima or searching for better local minima may be necessary. Because of these kinds of problems, our goal is to decouple the knot selection process from the least squares process.

When data are somewhat uniformly distributed, methods involving tensor product cubic splines may be desirable. Tensor product methods place knot locations on a grid, which does not necessarily reflect the actual disposition of the data points; in fact, there could be no data nearby. Even though these problems are surmountable, they could lead to non-uniqueness of solutions and a minimum norm solution that is not aesthetically appealing.

A different point of view is considered here where the knot point locations are predetermined based on two criteria. Specifically, we shall make assumptios relating the density of data to the dependent variable and mandating the importance of each individual data point. Solution of the overdetermined system of equations follows the knot point selection. A summary of the approach and its results will be presented. Examples are given which illustrate rather well the ability of the scheme to select knot locations which reflect the underlying density of the data. Actual surface fitting and comparison with two other methods, the Laplacian smoothing splines of Wahba and Wendelberger [3], and the tensor product bicubic Hermite method due to Foley [4], are also reported on.

II. THE KNOT SELECTION PROCESS. Given 'a priori' flexibilty in knot placement, the problem becomes the selection of knot location, followed by solution of the system by least squares. Since the selection of knot location is to be decoupled from the solution of the least squares problem, some assumptions must be made in order to develop an algorithm for the knot selection process.

First, we assume that the independent variable data reflects something about the behavior of the dependent variable. For example, the density of the data points may be dependent on the curvature of the surface. Hence, where relatively many data points are found, the function is assumed to be changing behavior rapidly, whereas a low density of data indicates slowly changing behavior. Although this assumption is not universally satisfied in practice, it does not seem unreasonable one.

The second assumption is that each data point is equally important in defining the underlying surface. Therefore the number of data points represented by each knot should be the same or nearly the same. This leads to 'equal representation' of the data points by the knot points where each data point is 'close' to a knot point. A key advantage is achieved in pursuing this approach in the form of a natural heuristic for moving the knots around the plane in searching for the optimal knot configuration. This point will be elaborated on later in the paper.

Our knot selection algorithm is based on these last two assumptions. First, we wish to minimize the sum of the distances squared from each data point to the nearest knot point; that is, minimize the 'global' value,

$$GN^2 = \sum_{i=1}^{N} \underset{j}{\text{MIN}} \left[(x_i-x_j)^2 + (y_i-y_j)^2\right] .$$

This is global in the sense that it accounts for the contributions from all of the K tiles. The expression leads naturally to a 'default' Dirichlet Tesselation, a partitioning of the plane with respect to the knot points (Figure 1.1). Thus, each data point belongs to some knot point according to the Dirichlet tile in which it lies. Data points on any of the tile boundaries (ties) must be resolved by a determination of which tile they belong to or some sharing mechanism.

Differentiation of GN^2 with respect to x_j and y_j show that at the minimum, each knot point will occupy the centroid of its tile with respect to the data points inside that tile. Given some initial configuration of knot points with its default Dirichlet Tesselation (our initial guess for the initial configuration was taken to be quasi-gridded), the following algorithm for iteration to a local minimum GN^2 value is employed:

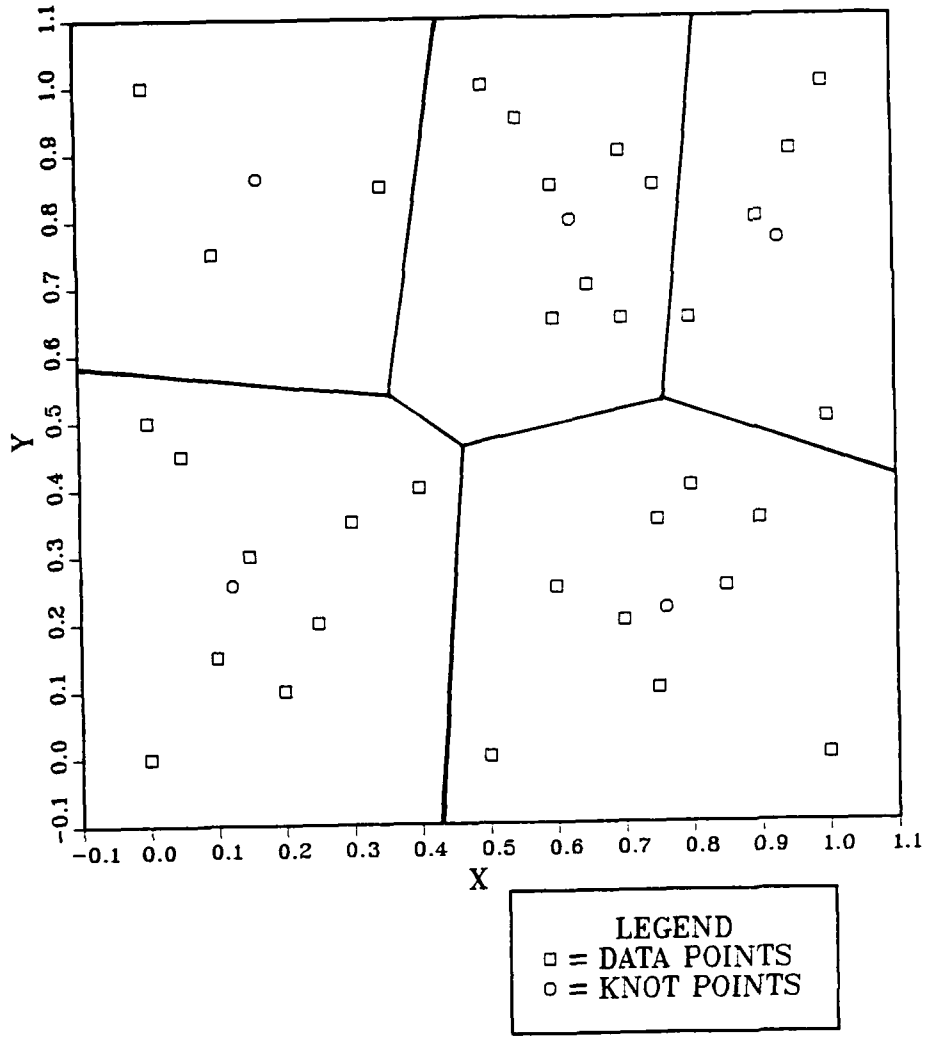


Figure 1.1 A Dirichlet Tesselation with 5 Tiles. It is constructed by connecting the perpendicular bisectors to the lines joining each of the knot points.

(a) compute the centroid of each tile with respect to the data points contained within each tile;

(b) move the knots to the corresponding centroids, which results in a new Dirichlet Tesselation and a new set of knot point - data point associations; this is the configuration for the next iteration.

(c) quit when two successive iterations yield the same knot locations, which means that a minimum global value of GN^2 has been found.

This algorithm was formulated in discussions at the Istituto per le Applicazioni della Matematica e dell'Informaica in 1983 [5], after the problem was posed by G. Nielson and R. Franke.

We note that the value of GN^2 will necessarily decrease as the iterations continue, until two successive iterations yield

the same configuration; this will be proven below. In the case where no data points lie in a tile for some knot point, the knot point is moved to the nearest data point. This mechanism avoids knots without data points. Futhermore, if a data point lies on a tile boundary, it is assigned to the knot with the smallest subscript (amongst the appropriate choices of knot points). Employment of a different criterion for the resolution of ties will yield different results. We note that knots cannot coalesce.

The following theorem is pertinent to this algorithm.
THEOREM: The function GN^2 decreases with each iteration which involves movement of a knot point.
PROOF [5]: Write GN^2 in the more convenient form

$$GN^2 = \sum_{j=1}^{K} \sum_{i \in I_j} [(x_i - x_j)^2 + (y_i - y_j)^2] \qquad (1)$$

where $I_j = \{i: (x_i, y_i)$ belongs to $(x_j, y_j)\}$. In (1), the interior sum is the sum of the distances from the data points in a tile to the knot point in that tile, and the exterior sum is over all K of the tiles. Let a prime denote the new knot points and index sets. This form leads to the expressions,

$$(x'_j, y'_j) = \left(\sum_{i \in I_j} x_i/p_j, \sum_{i \in I_j} y_i/p_j \right),$$

where p_j is the number of indices in each set I_j. The set I_j contains the indices for the data points in the tile for the j^{th} knot point. The new knot points will lead to a new tesselation, followed by the new index sets I_j'. Then the expression (1) is greater than or equal to

$$\sum_{j=1}^{K} \sum_{i \in I_j} [(x_i - x'_j)^2 + (y_i - y'_j)^2] \qquad (2)$$

because the new knot point locations minimize the contribution of the interior sums. This expression (2), in turn, is greater than or equal to

$$\sum_{j=1}^{K} \sum_{i \in I'_j} [(x_i - x'_j)^2 + (y_i - y'_j)^2] \qquad (3)$$

since an index i moves to another set only in the case wherein the corresponding data point is now closer to a different knot point, thus decreasing its contribution to the global GN^2 value.

Finding a local minimum of GN^2 is well-served by this algorithm; however, as seen next in a one dimensional example, the function GN^2 is rife with local minima, and the local minimum value found depends on the initial configuration of knots used. We can draw similar conclusions for the multi-dimensional case based on the one dimensional analogy.

KNOT FIXED AT 0.5

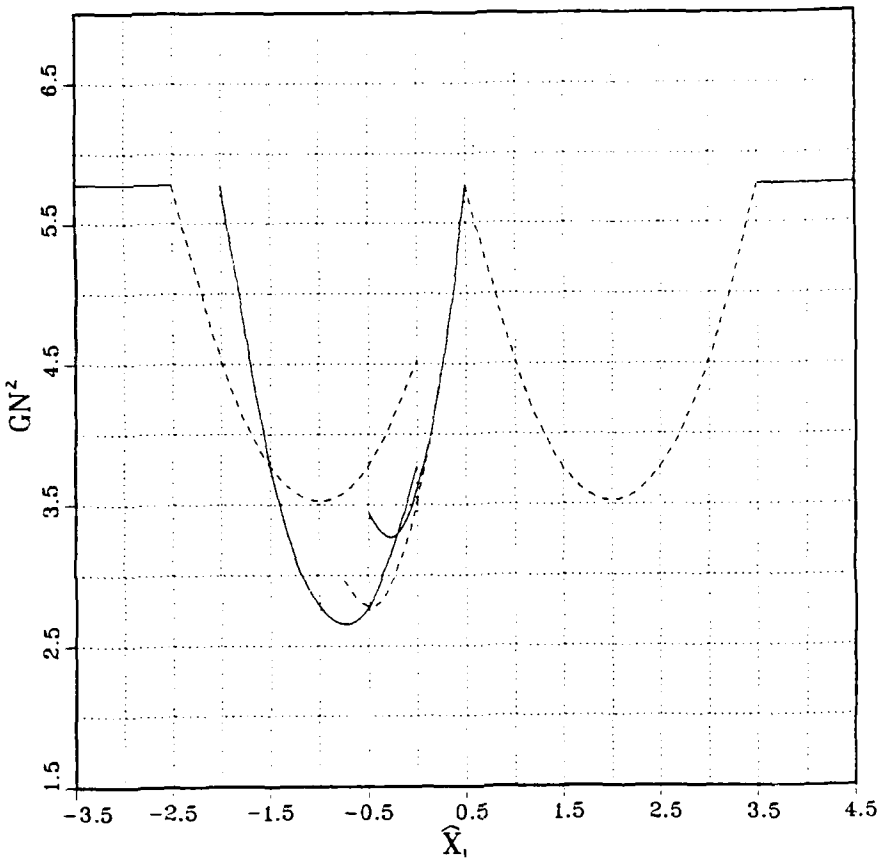


Figure 1.2 One Dimensional Cross-section of GN^2. The dependent variable in each of the graphs is the value of the GN^2 function as a function of the one variable knot point; the other knot point in each figure is fixed.

The function GN^2 is the sum of N continuous, piecewise quadratic functions, $g_i(x_j,y_j)$, i = 1,...,N, selected from a larger set of N x K functions $g_i(x_1,y_1,...,x_k,y_k)$, i = 1,...,N, such that

$$g_i(x_j,y_j) = \underset{j}{\text{MIN}}\ (d_{ij})^2 = \underset{j}{\text{MIN}}\ [(x_i-x_j)^2 + (y_i-y_j)^2].$$

Thus, GN^2 is a function of 2K independent variables, which are the knot point coordinates $(x_1,y_1,.....,x_k,y_k)$; each data point (x_i,y_i), i = 1,...,N, is fixed.

Suppose we wish to represent a set of five data points { -1,-0.5, 0, 0.33, and 2 }, with a set of two knot points, $\{x_1,x_2\}$. Then the GN^2 function is:

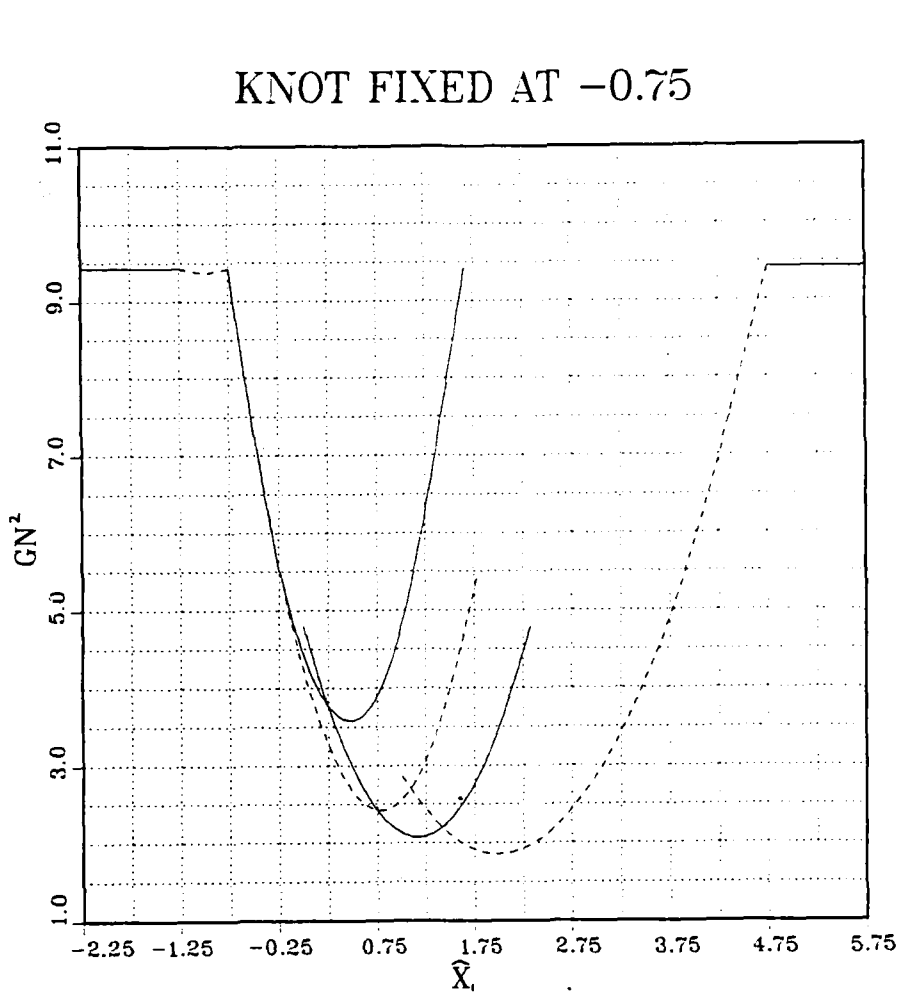


Figure 1.3 One Dimensional Cross-section of GN^2. Note how the scales of each of the Figures 1.2 through 1.5 vary so as to accomodate the ranges of the piecewise quadratic function, GN^2.

$$GN^2 = \text{MIN}\ [(x_1+1)^2,(x_2+1)^2] + \text{MIN}\ [(x_1+0.5)^2,(x_2+0.5)^2 + \text{MIN}\ [x_1^2,x_2^2] + \text{MIN}\ [(x_1-0.33)^2,(x_2-0.33)^2] + \text{MIN}\ [(x_1-2)^2,(x_2-2)^2]$$

where x_1 and x_2 are the as yet unspecified locations of the knots, the variables on which the optimization will occur. The one dimensional GN^2 function is piecewise quadratic, consisting of the continuous component quadratic functions $g_i(x_j,y_j)$.

Now we fix one of the knot points at some reasonable location. Figures 1.2 through 1.5 depict a series of cross-sectional views of the GN^2 function where one of the knot points is fixed at various reasonable locations. We can see the behavior of the GN^2 function as the varilable knot location changes. The minimization occurs in two distinct steps: first, each data point is assigned to the closest knot, and second, the centroid of each tile is computed based on the data points which have been assigned to each of them.

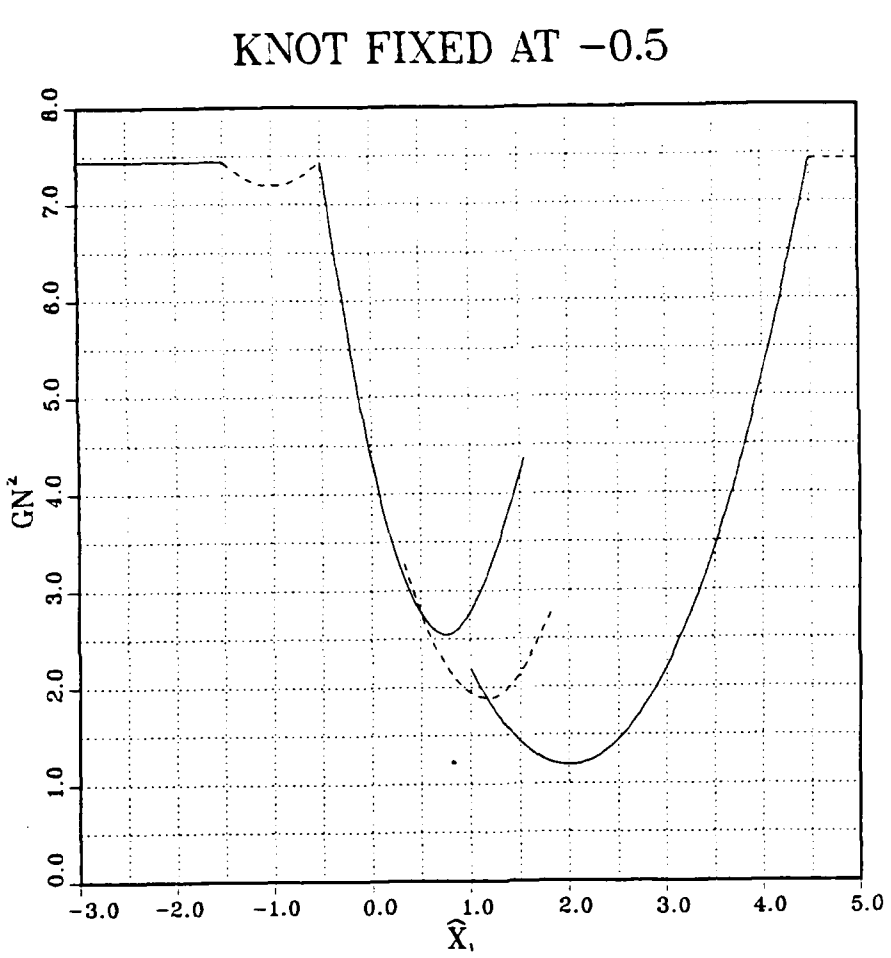


Figure 1.4 One Dimensional Cross-section of GN^2. In each of the Figures 1.2 through 1.5, the data points are fixed at the locations -1, -0.5, 0, 0.33, and 2.

The places at which the tile boundaries move across a data point are seen as corners or the intersections of two adjacent pieces of the composite quadratic function. As a direct result of the centroid requirement, the GN^2 function will stabilize at that local minimum value corresponding to the particular piece of the quadratic in which the variable knot is placed. The local minimum value will frequently occur out of the domain of the corresponding quadratic piece so that stabilization of the variable knot point will occur when the local minimum value is found within the domain of the appropriate quadratic piece.

This phenomenon is referred to as 'cascading' to a local minimum value. In spite of the cascading phenomenon, the minimum global value of the GN^2 function will not necessarily be attained, since the cascading stops as soon as the local minimum value is found to be within the domain of the quadratic. A local maximum value occurs when the two knot points coincide, leading to an apparent coalescing of knots. However, such coalescing is avoided in the algorithm by movement of the knot point to the nearest data point whenever a tile is left empty.

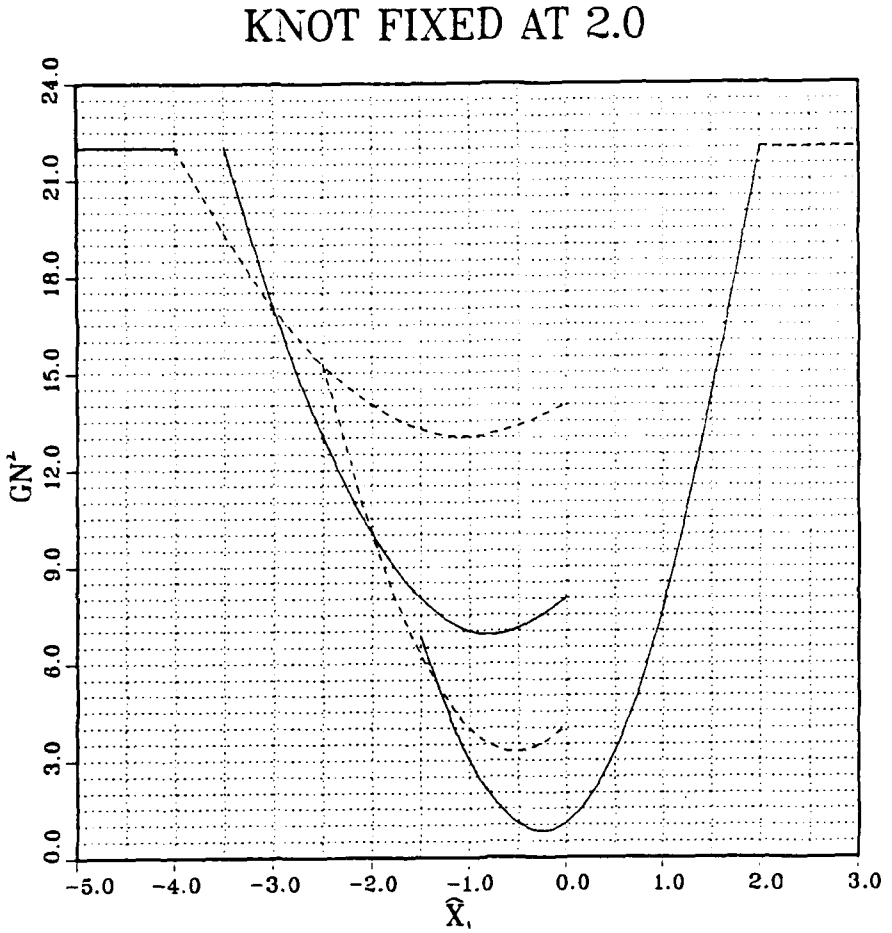


Figure 1.5 One Dimensional Cross-section of GN^2. Observe the phenomenon called cascading where the variable knot point seeks that location which minimizes the GN^2 function value, subject to it being within the domain of the quadratic.

Tables I and II summarize two possible knot movement scenarios given the same initial guess for the knot point locations. The scenarios differ in that the first one employs the specific criterion for breaking ties wherein the data point is assigned to the knot with the smallest subscript. The other scenario employs an alternative tie-breaking scheme. For a fixed set of data points, the initial guess, which can be generated by the program or provided by the user, leads directly to the assignment of the data to the closest knot point. This is followed by the determination of the new knot location via the centroid criterion. The process continues until stabilization occurs. Used in conjunction with the Figures 1.2 through 1.5, these tables yield valuable insight into how the one dimensional case works, and lend themselves to understanding the multidimensional case.

TABLE I

ITERATION	KNOT POINT X_1	DATA POINT ASSIGNMENT	KNOT POINT X_2	DATA POINT ASSIGNMENT	SEE FIG.
0 (Initial Guess)	-0.75	{-1,-0.5}	0.5	{0,0.33,2}	1.2
1 (New Knot)	-0.75	{-1,-0.5,0}	0.78	{0.33,2}	1.3
2 (New Knot)	-0.5	{-1,-0.5,0}	1.167	{0.33,2}	1.4
3 (New Knot)	-0.5	{-1,-0.5,0}	1.167	{0.33,2}	1.4
		S T A B I L I Z A T I O N			

Table I. Trial one employs the tie breaking criterion described earlier where the data point is assigned to the knot point with the smallest subscript. Both tables are read in zig-zag fashion following the flow of each iteration separately, but in tandem with the other knot point assignments.

TABLE II

ITERATION	KNOT POINT X_1	DATA POINT ASSIGNMENT	KNOT POINT X_2	DATA POINT ASSIGNMENT	SEE FIG.
0 (Initial Guess)	-0.75	{-1,-0.5}	0.5	{0,0.33,2}	1.2
1 (New Knot)	-0.75	{-1,-0.5,0}	0.78	{0.33,2}	1.3
2 (New Knot)	-0.5	{-1,-0.5,0}	1.167	{0.33,2}	1.4
3 (New Knot)	-0.5	{-1,-0.5,0,0.33}	1.167	{2}	1.4
4 (New Knot)	-0.292	{-1,-0.5,0,0.33}	2.0	{2}	1.5
5 (New Knot)	-0.292	{-1,-0.5,0,0.33}	2.0	{2}	1.5
		S T A B I L I Z A T I O N			

Table II. In trial two, the tie is resolved differently, so that an alternative data point assignment is made at iteration 3. Hence, the final outcome is significantly different. Note how each iteration is referenced to one of the Figures 1.2 through 1.5.

The algorithm for finding the minimum local value of GN^2 performs inconsistently as seen in the cascading phenomenon wherein the GN^2 function may have several local minimum values. We are lead to consideration of a somewhat different criterion for locating the best configuration of knot points. We wish to exploit the second assumption specified earlier, while still taking advantage of the minimization of the GN^2 function.

Since each data point is assumed to be equally important, the Dirichlet tile for each knot should contain about the same number of data points. Thus, we wish to minimize the sum of the squares of the differences between the number of knots in each tile and the average number that should belong to each tile; that is, minimize the quantity

$$D = \sum_{j=1}^{K} (N_j - N/K)^2 ,$$

where N_j is the number of data points in the j^{th} tile. The new algorithm for determining knot locations is based on the minimization of D, subject to the constraint that each knot be located at the centroid of its tile.

This new optimization leads to a natural heuristic for moving knots from a stable configuration to a possibly better configuration. We call the current configuration of knots a base configuration, and iterate through the algorithm as follows:

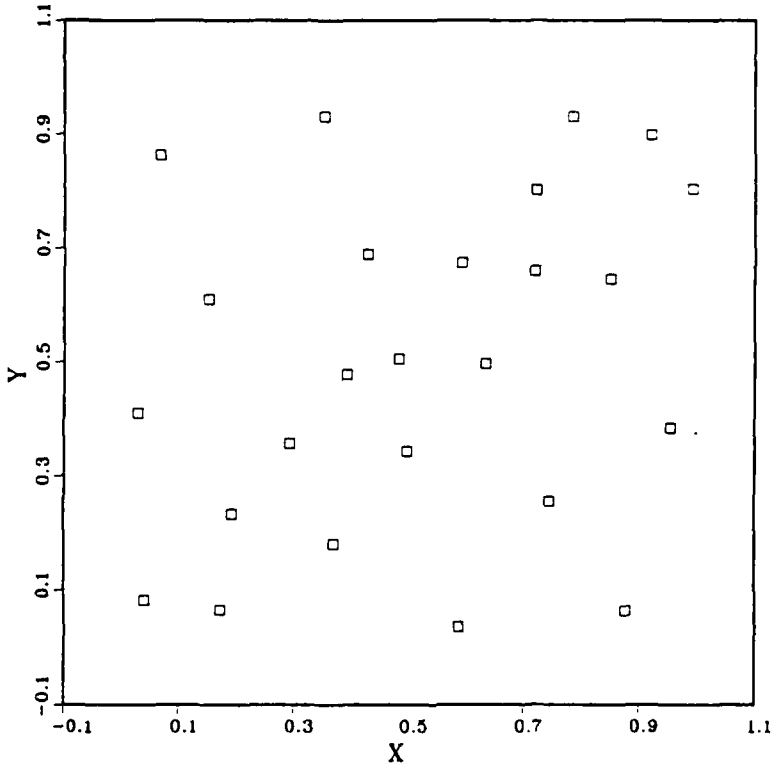
(a) generate a new guess for the knot locations by moving the knot(s) with the smallest number of data points in their tiles toward the knot(s) with the largest number of knot(s) in their tile; the distance moved is initally a large fraction of the total distance between the knots.
(b) iterate to a stable configuration using the first algorithm, compute the values of GN^2 and D, and compare D to the smallest value achieved to date, as represented by that of the base configuration;
(c) repeat the process above when a smaller value of GN is obtained with the present configuration as the base configuration;
(d) when a smaller value of GN^2 is not found, take a shorter step in the movement of the knot(s) and repeat the process above;
(e) continue with smaller and smaller steps until a smaller value of D is found (or an equal value of D with a smaller GN^2 value) until the knot locations return to the base configuration;
(f) perform the search in the symmetrical way when the base configuration is returned to; that is, move the knot(s) with the largest number of data points in their tile(s) toward the knot(s) with the smallest number of knots in their tile(s);
(g) quit when no smaller value of D is found.

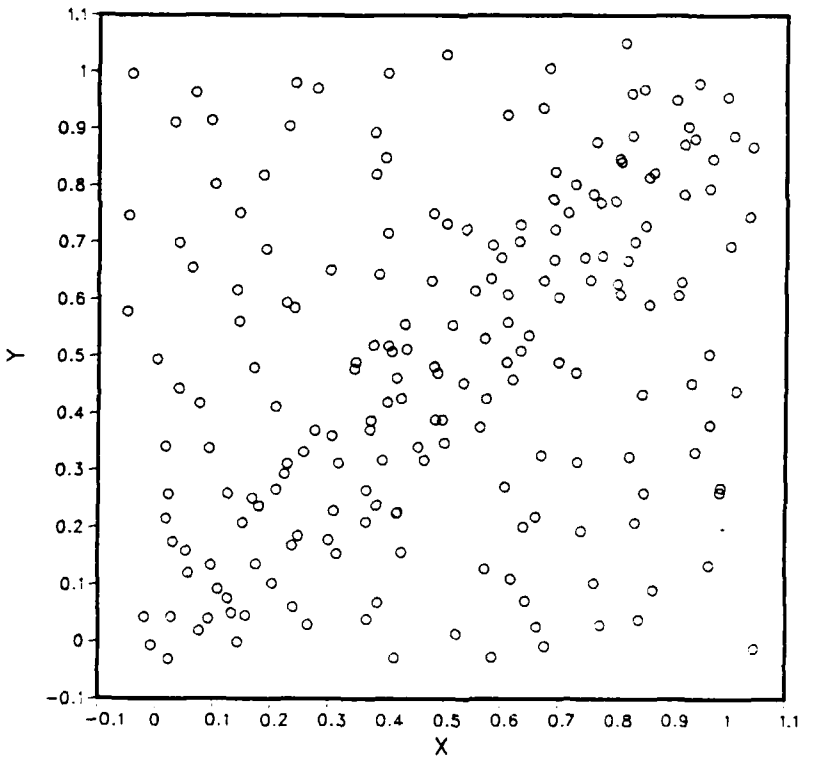
The movement of the knots is justified by the rationale that a more equitable distribution of data points can be found by moving the tile boundaries across data points. Note that the first algorithm for computing the GN^2 function value is embedded in this new algorithm.

Once the knot point locations have been determined, the least squares problem is solved using the public domain software, LINPACK [6]. We call the N+3 x K+3 matrix of equations A, the K+3 x 1 column matrix of unknown coefficients x, and the N+3 x 1 column matrix of dependent variables b, so that the least squares problem can be posed as solution of the system $Ax=b$. The LINPACK subroutines employ a QR decomposition of matrix A, which can then be written as $QRx=b$. Multiplication by Q^T yields $Rx=Q^Tb$ since $Q^TQ=I$. R is a rectangular matrix with dimensions N+3 x K+3, which is zero below its main diagonal, so that multiplication by the block matrix R_{11}^{-1} yields the result $x=R_{11}^{-1}Q^Tb$. Thus, the computation of x requires only the matrix-vector multiplication Q^Tb, followed by back substitution in the triangular system $R_{11}x=Q^Tb$. Using a Householder algorithm for the QR decomposition, numerical stability is guaranteed. Finally, with the known coefficients in hand, a grid of surface values can be computed, which can be subsequently used by a user provided plotting routine to generate a plot of the surface.

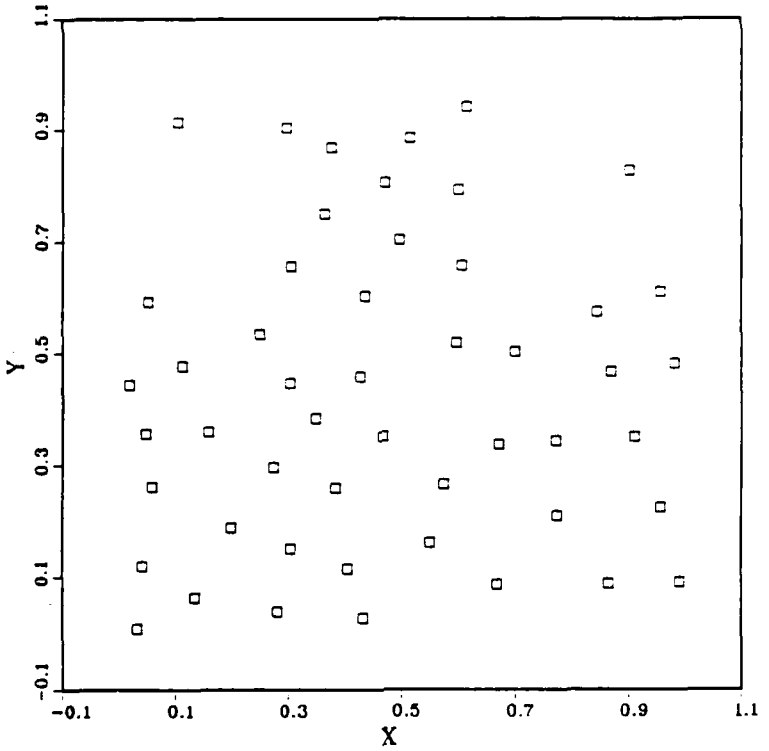
III. RESULTS AND EXAMPLES. Using the least squares algorithm for the a priori selection of the knot point locations, experiments were conducted to test the scheme using different sets of test data. This was followed by verification of the scheme on a large set of real data. Results from two sets of the test data are presented here: one consisting of 200 data points called 'Cliff', and one consisting of 500 data points called 'Humps and Dips'. Both sets of data were generated using known functions (see Franke[7]) in a way that forced the disposition of points to be proportional to the curvature of the sampled function. Figures 1.6 through 1.11 portray the test data sets graphically, and illustrate the optimized knot point configurations found using the least squares algorithm. Figure 1.11 is particularly encouraging, since it depicts actual hydrographic data collected in Monterey Bay. We note that the assumption regarding the density of the data being indicative of the behavior of the dependent variable is not actually satisfied in this case due to the source of the data. Nonetheless, these results demonstrate the ability of the algorithm to produce representative sets of knots.

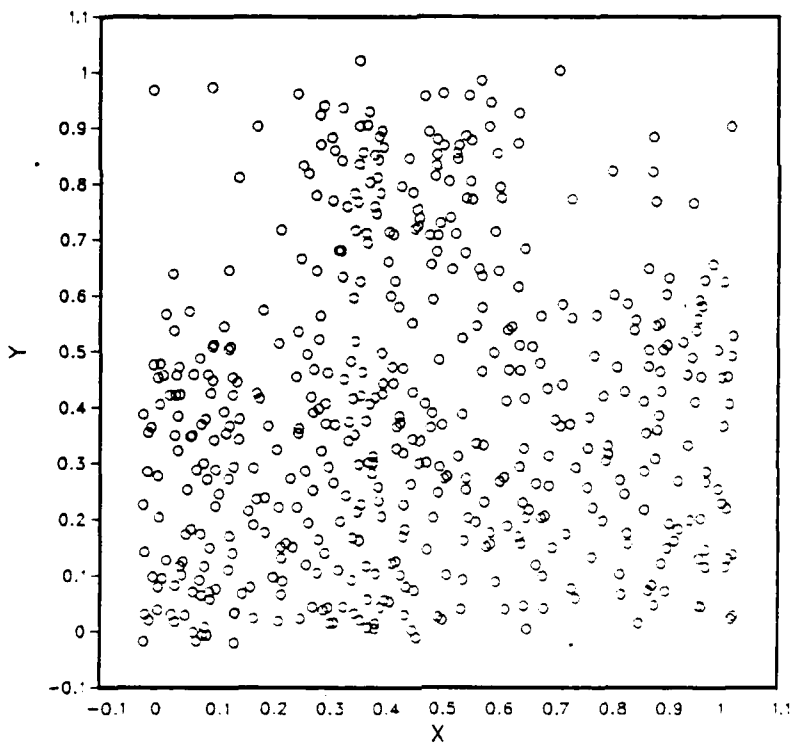
We also investigated how closely the constructed surface F and the 'true' surface resemble one another. This comparison is made in the context of the root-mean-squared error (RMS) of both the residuals (at the data points) and on a rectangular grid of locations in the plane. Tables III and IV provide a comparison of the RMS error for the test data sets using the least squares algorithm developed here, the method of Wahba and Wendelberger [3], and the method of Foley [4]. The dependent variables of the experimental data sets were generated in two ways: 1) using a known function, and 2) contaminating it by the injection of independent, normally distributed random errors with a composite standard deviation of less than 0.05. The actual standard deviation was about 0.0485.



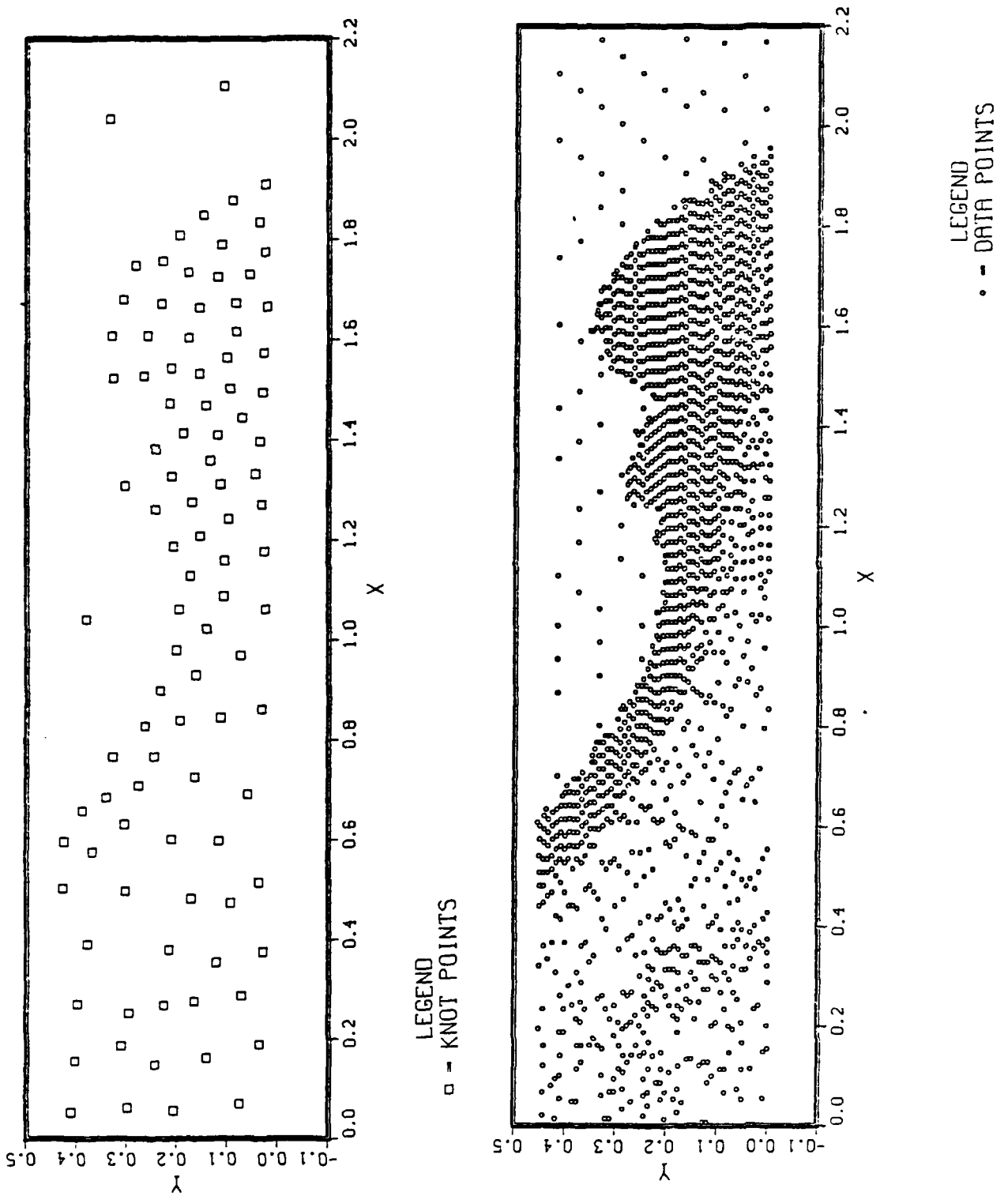


Figures 1.6 and 1.7. The 'Cliff' data set. Note the relatively dense disposition of data points across the diagonal where the underlying surface drops off. The 25 knot points used clearly reflect the behavior of the data set, as expected.





Figures 1.8 and 1.9 The 'Humps and Dips' data set. Note how clumps of data appear in three portions of the plane, indicating that the underlying surface is undergoing change. A set of 50 knot points was used to represent the data.



Figures 1.10 and 1.11 Hydrographic Data from Monterey Bay, Ca. Here, 1669 data points (soundings) are represented by 100 knot points. The results are very reasonable.

TABLE III

COMPARISON OF RMS ERRORS ON 'CLIFF' 200 POINTS

METHOD	NUMBER OF DATA POINTS/ KNOT POINTS	NO ERRORS IN DATA		CONTAMINATED DATA	
		RESIDUAL	GRID	RESIDUAL	GRID
LSTPS	200/20	.01562	.01474	.05214	.01795
LSTPS	200/25	.01179	.01154	.04805	.02040
FOLEY	200/5X5	.00777	.00613	.05996	.04819
LSTPS	200/35	.00626	.00616	.04590	.02146
FOLEY	200/6X6	.00512	.00417	.05113	.03745
SMOOTHING	200	0.0	.00096	.04272	.01806

TABLE IV

COMPARISON OF RMS ERRORS ON 'HUMPS & DIPS' 500 POINTS

METHOD	NUMBER OF DATA POINTS/	NO ERRORS IN DATA		CONTAMINATED DATA	
		RESIDUAL	GRID	RESIDUAL	GRID
LSTPS	500/20	.02402	.02517	.05256	.02738
LSTPS	500/25	.01664	.01766	.04818	.02283
FOLEY	500/5X5	.01346	.01230	.05844	.03767
LSTPS	500/50	.00645	.00845	.04544	.01961
FOLEY	500/7X7	.00645	.00552	.05696	.04864

Tables III and IV. Comparison of RMS errors in two sets of data, each with an exact and a contaminated version, using three methods for surface construction.

In the first case, we would expect to see an overall decrease in the RMS error on both the residuals and the grid as the number of knot points used to represent the data is increased. In the contaminated case, the dependent variable at each data point is the sum of the unknown underlying function value and the error function value so that the difference between the constructed surface and the 'true' surface is entirely attributable to the presence of error in the data. Thus, we expect the RMS error in the residuals to match the composite standard deviation of the random error injected into the contaminated data. At the grid points, we expect the RMS error to be smaller than the composite standard deviation, since the grid sample is larger (33x33) and the errors are distributed more evenly throughout the entire region of interest. In the case where no error is present, we expect that the difference between the constructed surface and the 'true' surface is entirely due to 'slack' in the constructed surface. We anticipate that the RMS error in the residuals would be approximately equal to the RMS error on the grid, thereby giving evidence that the error in the constructed surface is uniformly distributed over the entire region of interest.

Some observations can be made regarding Tables III and IV. The general trend of the RMS error on both the residuals and the grid is to decrease as the number of knot points is increased. As expected with the exact data, the RMS error of the residuals and the RMS error on the grid are roughly equivalent. For the contaminated data, the RMS error of the residuals roughly matches the composite standard deviation of the data, and the RMS error on the grid is smaller than the RMS error of the residuals, as expected. The discrepancy between the RMS error on the grid and the RMS error in the residuals cannot be totally attributed to the injected error; it is the result of 'undersmoothing', where the constructed surface tends to fit the error rather than the data.

In comparing the least squares to the smoothing spline method in the exact data case, we note that the smoothing spline method yields a residual RMS error of 0. This could be expected, since there is no error in the data and the spline of interpolation is chosen. On the grid, the RMS error is small since some amount of error on the grid is expected. When the data not contaminated, the RMS error of the least squares algorithm begin to approach those of the smoothing splines method only as the number of knots used becomes large. We also note that in the 500 data point set (Humps and Dips), no comparison is made since a potential limit for computing smoothing splines is 200-300 data points.

In comparing Foley's method to the least squares method for the contaminated case, the RMS error on the residuals is nearly equal to the composite standard deviation injected into the data. However, on the grid, the least squares method does better, an indication that smoothing is occurring, as expected. We also note that an increase in the number of grid points does not

significantly improve the RMS error in Foley's method, even though an increase in the number of knots in the least squares method usually yields improved results. We used the default local approximations in Foley's method, and we note that performance of the method may be improved through the use of lower degree local approximations to estimate the grid values to be used.

Finally, we note that the search for a best knot configuration can turn out to be rather expensive. For a large number of data points with a moderately large number of knot points, the computational effort could be excessive, althogh we are investigating ways of speeding up the algorithm. Furthermore, as we noted earlier, the end results are dependent on the initial guess, although they generally look quite good for any reasonable initial guess.

REFERENCES

1. Harder, R. L. and Desmarais, R. N., "Interpolation Using Surface Splines," J. Aircraft, V. 9, No. 2, February, 1972.

2. Schmidt, R., "A Contribution to Surface Approximation with Irregularly Distributed Data," International Series of Numerical Mathematics, V. 75, 1985.

3. Wahba, G., and Wendelberger, J., "Some New Mathematical Methods for Variational Objective Analysis Using Splines and Cross Validation," Monthly Weather Review, V. 108, August, 1980.

4. Foley, T. A., "Scattered Data Interpolation and Approximation," Lawrence Livermore National Laboratory RPT. No. UCID-20346, February 11, 1985.

5. Due to Nielson, Utreras, Lenarduzzi, and Franke, Istituto per le Applicazioni della Matematica e dell' Informatica, Milano, 1983.

6. Dongara, J. and others, LINPACK User's Guide, SIAM, 1979.

7. Franke, R., "Scattered Data Interpolation: Tests of Some Methods", Math. Comp. 38 (1982) 181-200.

A Rapid, Backscatter Simulation Technique for Complex B-Spline Target Models

Karl D. Reinig
Advanced Electronics Systems Laboratory
Sensors and Signal Processing Technology Division
Harry Diamond Laboratories, U.S. Army LABCOM
2800 Powder Mill Rd, Adelphi, MD 20783

ABSTRACT. This paper describes a method for rapidly evaluating the simulated radar backscatter signatures of B-spline target models moving relative to a source/receiver. A geometric optics approach is used to estimate the radar return from a complex target surface described by a bi-cubic B-spline mesh. The method exploits the second-order continuity of bi-cubic B-spline surfaces to reduce the problem of finding all the specular points associated with each new trajectory position to that of tracking the motion of existing points. In particular, it is shown that the locations of the annihilations and creations of specular paths may be predicted for an entire trajectory, eliminating the need to search the whole surface for specular points as the target moves relative to the source/receiver. The method is shown to work for the multiple-bounce case as well.

1 Introduction

Consider the scenario depicted in figure 1. A source/receiver (S/R) moving along some trajectory illuminates a target of interest. It is desired to estimate the return from the target as the S/R moves along the trajectory. Notice that whether figure 1 describes a target detection/identification problem or the terminal phase of a guided munition is mostly determined by the trajectory being considered. A backscatter simulation technique which places few or no restrictions on the paths of the trajectories to be simulated would therefore find use in all phases of seeker munitions studies. In addition, of course, the relative motion between the S/R and the target could be due strictly to the motion of the target. Thus the scenario also includes the return from passing targets.



Figure 1: Target Encounter Scenario

This paper discusses the use of a geometric optics approach to simulate the radar return from complex but generally smooth targets. The overall simulation method can be broken into two basic parts. First, the surface of a target is described using bi-variate piecewise polynomial functions such as bi-cubic tensor product B-spline surfaces. Second, for each location along any desired trajectory, the positions on the surface (the specular points) from which a ray leaving the source would be reflected back to the receiver are found along with their local principal radii of curvature. The locations of the specular points are used along with their local radii of curvature to calculate the discrete radar cross sections. The overall return from the complex target is then found by coherently summing the expected discrete return from each specular point.

Previous studies have demonstrated the usefulness of the geometric optics approach for computing the expected return from a composite of simple analytic shapes [1]. However, as the targets of interest become more complex or the desire to match their surfaces more accurately increases, the use of simple analytical shapes to describe the target surface often becomes impractical. Three-dimensional faceted models exist for most targets of

interest, including tanks, helicopters, and jet aircraft. These faceted models generally use several thousand facets to describe the target surface and contain a great deal of detail. Unfortunately, the faceted models do not directly give useful geometric optics information. For example, both principal radii of curvature of a faceted model are unbounded everywhere except at facet edges where they are undefined. This paper focuses on a technique which exploits the second-order continuity of bi-cubic B-spline surfaces to reduce the problem of finding all the specular points associated with each new trajectory position to that of tracking the motion of existing points. For a complex target, the reduction in the problem results in multiple orders of magnitude in savings. In addition, it is shown that the technique can be easily extended to the multiple-bounce case, with the potential for even greater savings.

2 Single-Bounce Return Problem

The geometry of the specular return problem from a single patch of an arbitrary B-spline surface is shown in figure 2. $\mathbf{R}_t(\Delta)$ is the current position

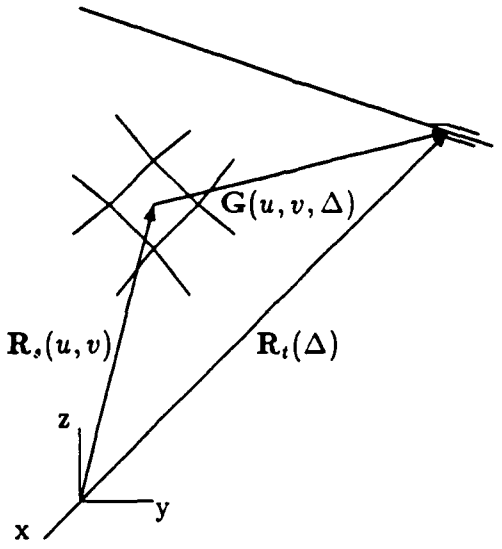


Figure 2: Specular Return Geometry

of the projectile along a linear trajectory. $\mathbf{R}_s(u,v)$ describes the target surface as a function of the two parameters u and v. And $\mathbf{G}(u,v,\Delta)$ is the difference between the two vectors $\mathbf{R}_t(\Delta)$ and $\mathbf{R}_s(u,v)$. It will be assumed throughout this paper that the projectile has an unobstructed view of the surface being considered; i.e., the problem of shadowing will not be addressed here. A necessary and sufficient condition for a point on the surface to be a specular point relative to the position $\mathbf{R}_t(\Delta)$ is that the l_2 norm of $\mathbf{G}(u,v,\Delta)$ be either a local maximum or minimum with respect to the two target surface parameters u and v. Finding all the specular points of a given surface (for a given trajectory position) is therefore the same as finding all u and v which satisfy the two nonlinear equations

$$F_1(u,v,\Delta) = \frac{\partial \|\mathbf{G}(u,v,\Delta)\|^2}{\partial u} = 0 \tag{1}$$

and

$$F_2(u,v,\Delta) = \frac{\partial \|\mathbf{G}(u,v,\Delta)\|^2}{\partial v} = 0. \tag{2}$$

If $\mathbf{G}(u,v)$ is given by a tensor product of cubic B-splines on a uniform grid, both $F_1(u,v)$ and $F_2(u,v)$ may be written explicitly in terms of u and v. In general though, solving for the u and v (call them u^* and v^*) which satisfy equations (1) and (2) requires a numerical technique. Simple application of Newton's method for nonlinear equations will find solutions to (1) and (2) provided the search is begun "close enough" to (u^*,v^*). The question of how close is close enough is a complex one, which ultimately depends on the variation of the surface being considered.

3 Twinkles

Suppose the coordinates of a specular point are known for a particular value of Δ and we wish to observe the motion of the specular point as Δ changes. The following argument is a trivial extension of that given by Longuet-Higgins for the case of a time-varying analytic surface [2]. Taking

the differential of equations (1) and (2) with respect to u, v, and Δ yields

$$\begin{pmatrix} \frac{\partial^2\|G\|^2}{\partial u^2} & \frac{\partial^2\|G\|^2}{\partial u\partial v} \\ \frac{\partial^2\|G\|^2}{\partial u\partial v} & \frac{\partial^2\|G\|^2}{\partial v^2} \end{pmatrix} \begin{bmatrix} \frac{du}{d\Delta} \\ \frac{dv}{d\Delta} \end{bmatrix} = \begin{bmatrix} -\frac{\partial^2\|G\|^2}{\partial u\partial\Delta} \\ -\frac{\partial^2\|G\|^2}{\partial v\partial\Delta} \end{bmatrix}. \tag{3}$$

Denote the two-by-two matrix of equation (3) as $\mathbf{J}(u, v, \Delta)$. Assuming the elements $\mathbf{J}(u, v, \Delta)$ are continuous functions of u, v, and Δ, if the matrix is nonsingular, $du/d\Delta$ and $dv/d\Delta$ will both be finite, which implies that the changes in u and v can be kept as small as desired by choosing the change in Δ small enough. Longuet-Higgins [2] refers to the vanishing of the determinant of $\mathbf{J}(u, v, \Delta)$ as a twinkle. The physical significance of this result is that as the S/R moves across a second-order continuous surface, specular points cannot suddenly appear or disappear unless $\mathbf{J}(u, v, \Delta)$ is locally singular. The observation that specular points move in continuous paths broken only when $\mathbf{J}(u, v, \Delta)$ is singular leads directly to the following conclusion. If for any given trajectory, it were possible to determine all the points (u, v, Δ) for which $\mathbf{J}(u, v, \Delta)$ is singular, it would no longer be necessary to search the entire surface for specular points at different positions along the trajectory. It would only be necessary to find all the specular points corresponding to one trajectory position ($\mathbf{R}_{t1}$ for example) and then track their motion, picking up or losing specular paths only at twinkles.

4 Finding Twinkles

Let

$$F_1 = \frac{\partial^2 \|\mathbf{G}\|^2}{\partial u}$$

$$F_2 = \frac{\partial^2 \|\mathbf{G}\|^2}{\partial v}$$

and

$$F_3 = \frac{\partial^2 \|\mathbf{G}\|^2}{\partial u^2}\frac{\partial^2 \|\mathbf{G}\|^2}{\partial v^2} - \left(\frac{\partial^2 \|\mathbf{G}\|^2}{\partial u\partial v}\right)^2.$$

Then the Newton step toward the parametric coordinates of a twinkle must satisfy

$$\begin{pmatrix} \frac{\partial F_1}{\partial u} & \frac{\partial F_1}{\partial v} & \frac{\partial F_1}{\partial \Delta} \\ \frac{\partial F_2}{\partial u} & \frac{\partial F_2}{\partial v} & \frac{\partial F_2}{\partial \Delta} \\ \frac{\partial F_3}{\partial u} & \frac{\partial F_3}{\partial v} & \frac{\partial F_3}{\partial \Delta} \end{pmatrix} \begin{bmatrix} S_u \\ S_v \\ S_\Delta \end{bmatrix} = \begin{bmatrix} -F_1 \\ -F_2 \\ -F_3 \end{bmatrix}.$$

Thus, local search techniques exist for finding all potential locations on the target surface, as a function of the trajectory position, for which the specular paths are discontinuous.

5 Specular Paths Near a Twinkle

By themselves, the conditions for a twinkle do not tell whether a particular twinkle represents a birth or death of a pair of specular points with respect to a chosen trajectory direction (e.g., time). Once again, a straightforward extension of the analysis by Longuet-Higgins [2] gives a method for describing the motion of specular paths near a twinkle, including whether the twinkle represents the birth or death of a pair of paths. Define

$$a_{ijk} = \left. \frac{\partial^{i+j+k} \|\mathbf{G}(u,v,\Delta)\|^2}{\partial u^i \partial v^j \partial \Delta^k} \right|_{u=v=\Delta=0}, \tag{4}$$

where the coordinate system is chosen such that $u = v = \Delta = 0$ at the twinkle and

$$a_{000} = a_{100} = a_{010} = a_{110} = a_{200} = 0. \tag{5}$$

It can be easily seen from Longuet-Higgins analysis that near a twinkle the u, v coordinates of a specular point are given by

$$\begin{aligned} u &= \pm \left[\frac{-2a_{101}\Delta}{a_{300}} \right]^{1/2} \\ v &= \frac{a_{210}a_{101} - a_{300}a_{011}}{a_{300}a_{020}}. \end{aligned} \tag{6}$$

Equations (6) show that if a_{101}/a_{300} is positive, then two solutions exist when Δ is less than zero, and no solutions exist when Δ is greater than

zero; i.e., an annihilation of a pair of specular paths occurs. Similarly, the creation of a pair of specular paths occurs when a_{101}/a_{300} is negative. It remains to determine a transformation of coordinates for which equations (6) hold. Simply choosing the origin of the new coordinate system to be the location of the twinkle ensures that a_{000} is equal to zero. It is useful to consider the surface formed by letting the function $\|\mathbf{G}(u, v, \Delta)\|^2$ be the w coordinate in an orthogonal u, v, w coordinate frame as shown in figure 3.

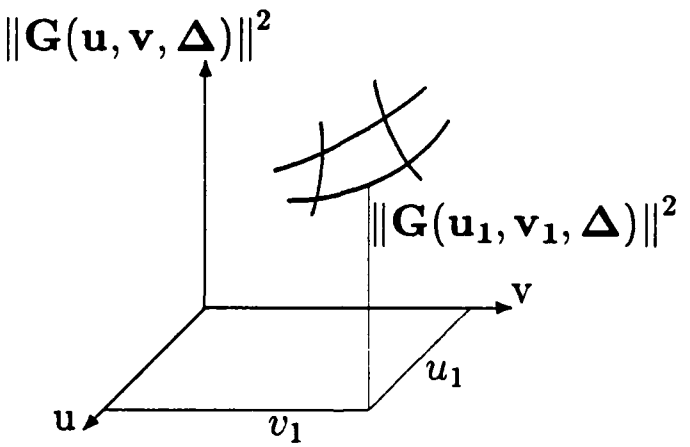


Figure 3: Distance Surface

For lack of a better term, this surface will be referred to as the "distance surface." The condition for a twinkle may then be interpreted physically as the vanishing of the Gaussian curvature of the distance surface. Or alternately stated, at a twinkle, one of the two principal radii of curvature of the distance surface is equal to zero. The curvature of any smooth surface, at a given point, in the direction δu, δv may be written as (see, among others, Faux and Pratt [3])

$$\kappa_n = (\delta u)^2 a_{200} + (\delta u)(\delta v) a_{110} + (\delta v)^2 a_{020}. \tag{7}$$

Now suppose a rotation of coordinates is made such that the u coordinate is aligned with the principal radii of curvature having zero value (that such

a direction exists is ensured at a twinkle). Then with δv = zero, equation (7) yields

$$0 = (\delta u)^2 a_{200},$$

which implies that a_{200} = zero. It can be seen that the twinkle condition implies

$$a_{200}a_{020} - a_{110}^2 = 0.$$

Therefore, in the rotated coordinate system, a_{110} must also be equal to zero and equations (5) are satisfied. Define the new coordinates u', v', and Δ' by

$$\begin{aligned} u &= u'\cos\theta - v'\sin\theta + u_{tw} \\ v &= u'\sin\theta + v'\cos\theta + v_{tw} \\ \Delta &= \Delta' + \delta_{tw}, \end{aligned}$$

where u_{tw}, v_{tw}, and Δ_{tw} are the original coordinates of the twinkle. Then if θ is the angle which rotates the original u axis into the direction of zero curvature, the signs of a_{101} and a_{300}, in the new coordinates, will tell if the twinkle represents a birth or a death. In addition, the motion of the specular points in the vicinity of the twinkle (in the new coordinates) will be given by equations (6).

6 Single-Bounce Example

Figures 4 and 5 show an example of the use of twinkles for tracking single-bounce specular paths on a B-spline surface. Figure 4 shows the B-spline surface control mesh and desired trajectory, as well as the locations along the trajectory where specular path discontinuities are expected (based on a search for twinkles). Each twinkle location along the trajectory is labeled B or D depending on whether the twinkle represents the birth or death of a pair of specular paths, and lines have been drawn connecting them with their associated locations on the target surface. In addition, the predicted paths of the specular points in the area near each twinkle are shown. Figure 5 shows the results after tracking the the specular paths for 1000

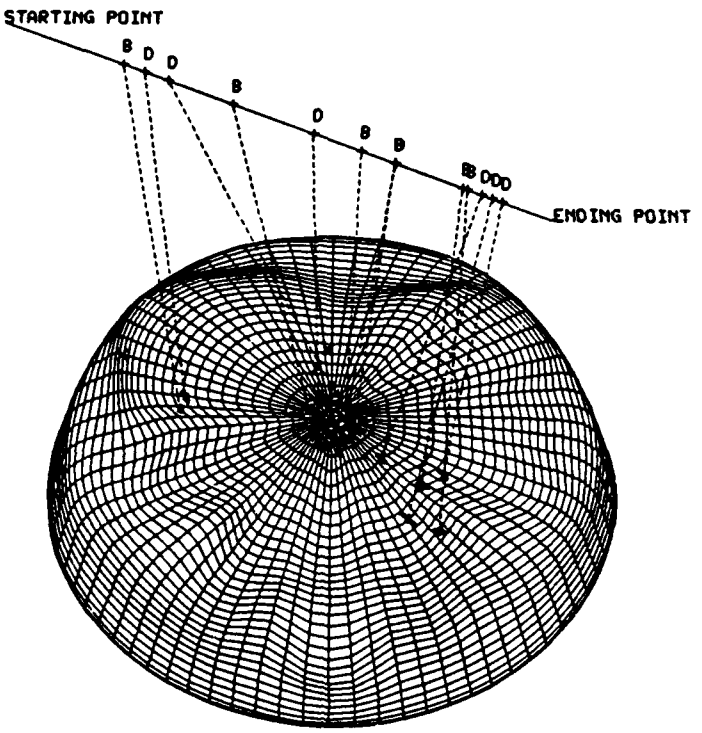


Figure 4: Twinkles and Initial Specular Points

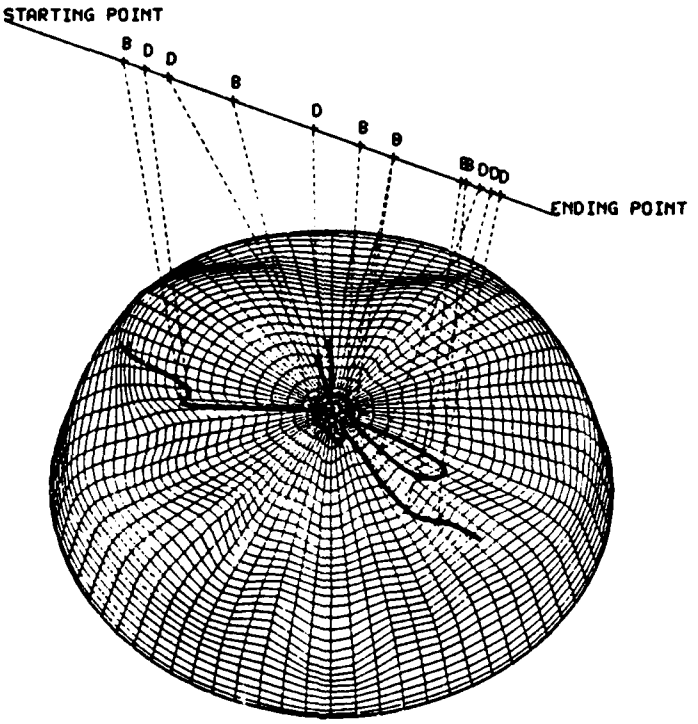


Figure 5: Specular Point Paths

locations along the trajectory. A comparison of figures 4 and 5 shows that the specular paths did in fact remain continuous everywhere except at the twinkles and moved as predicted in the regions near each twinkle. Because completely searching the entire target for specular points at each trajectory location was unnecessary, the entire simulation took only a couple of minutes. Even if the global search for specular points could be reduced to 10 seconds per trajectory location, the simulation would have taken over 2-1/2 hours without the use of twinkles.

7 Nth-Order Specular Points

Often multiple-bounce return, as depicted in figure 6, produces a significant contribution to the overall target backscatter. A weak form of Fermat's principle of optics guarantees that any multiple-bounce return path will be stationary with respect to the 2n surface parameters of the bounce. In terms of the distances between bounces, this becomes

$$\frac{\partial \sum_{i=1}^{n} d_i}{\partial u_j} = 0 = \frac{\partial \sum_{i=1}^{n} d_i}{\partial v_j}$$

for all j = 1,..., n. Noting that

$$\frac{\partial d_i}{\partial u_j} = \frac{\partial d_i}{\partial v_j} = 0$$

whenever $i - 1 \leq j$ or $j \geq i + 2$, we get the 2n conditions

$$\frac{\partial (d_i + d_{i+1})}{\partial u_i} = 0 = \frac{\partial (d_i + d_{i+1})}{\partial v_i}.$$

Denote $d_i + d_{i+1}$ by G_i. Then, taking the differential of the previous 2n equations with respect to the 2n surface parameters (u_i, v_i i = 1, n) and the trajectory parameter Δ yields

$$\begin{pmatrix} \frac{\partial^2 G_1}{\partial u_1^2} & \frac{\partial^2 G_1}{\partial u_1 \partial v_1} & \cdots & \frac{\partial^2 G_1}{\partial u_1 \partial u_n} & \frac{\partial^2 G_1}{\partial u_1 \partial v_n} \\ \frac{\partial^2 G_2}{\partial u_1 \partial v_1} & \frac{\partial^2 G_2}{\partial v_1^2} & \cdots & \frac{\partial^2 G_2}{\partial u_n \partial v_1} & \frac{\partial^2 G_2}{\partial v_1 \partial v_n} \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \frac{\partial^2 G_n}{\partial u_1 \partial v_n} & \frac{\partial^2 G_n}{\partial v_1 \partial v_n} & \cdots & \frac{\partial^2 G_n}{\partial u_n \partial v_n} & \frac{\partial^2 G_n}{\partial v_n^2} \end{pmatrix} \begin{bmatrix} \frac{du_1}{d\Delta} \\ \frac{dv_1}{d\Delta} \\ \cdot \\ \cdot \\ \cdot \\ \frac{dv_n}{d\Delta} \end{bmatrix} = \begin{bmatrix} -\frac{\partial^2 G_1}{\partial u_1 \partial \Delta} \\ -\frac{\partial^2 G_2}{\partial v_1 \partial \Delta} \\ \cdot \\ \cdot \\ \cdot \\ -\frac{\partial^2 G_n}{\partial v_n \partial \Delta} \end{bmatrix} . \quad (8)$$

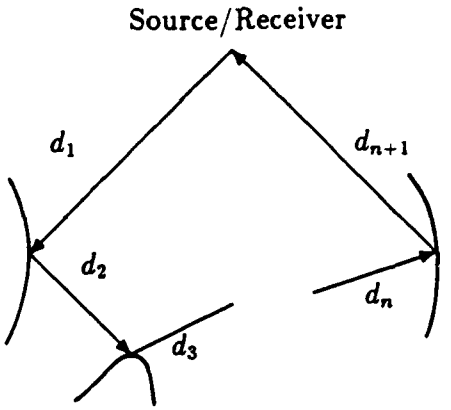


Figure 6: Multiple-Bounce Return

The same argument which led to the conclusion that single-bounce specular paths can only be created or annihilated at twinkles may be directly extended to include multiple-bounce twinkles; i.e., the vanishing of the determinant of the 2n x 2n Jacobian of equation (8) is required for a discontinuous motion of the nth-order specular paths.

8 Double-Bounce Example

Figures 7 and 8 show an example of the use of double-bounce twinkles for tracking double-bounce specular points on a simple B-spline surface. Figure 7 shows a control mesh for a simple crescent-shaped ribbon which has been tilted slightly so it may be viewed. A short trajectory is shown along with the only double-bounce twinkle associated with that trajectory and surface. In addition, a triangle has been drawn to show the double-bounce path associated with the twinkle. Figure 8 shows the results of searching the surface for double-bounce specular points at 40 locations along the trajectory. As expected, the number of double-bounce specular points associated with each trajectory location before the twinkle did not change (there were none). At the twinkle, two sets of specular paths were created (pairs of specular bounces associated with the same double-bounce path are shown connected by a line). The two sets of double-bounce paths moved in generally opposing directions from their origin.

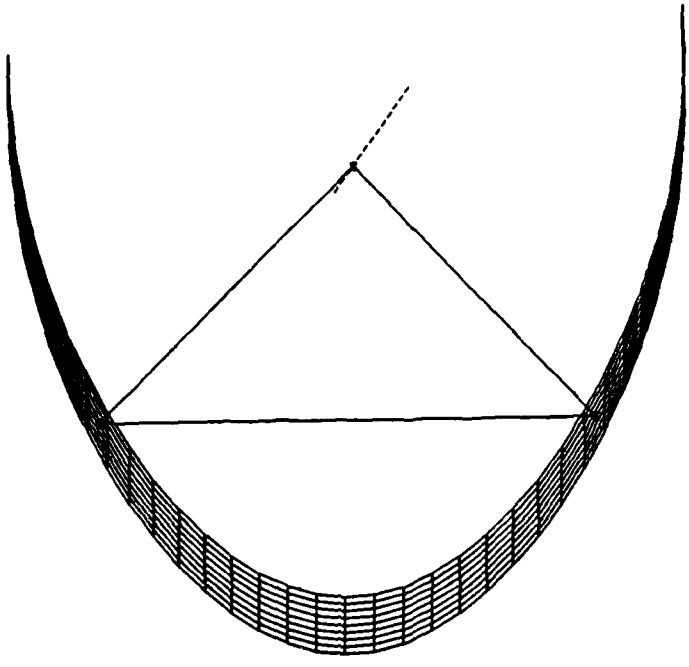


Figure 7: Double-Bounce Twinkle

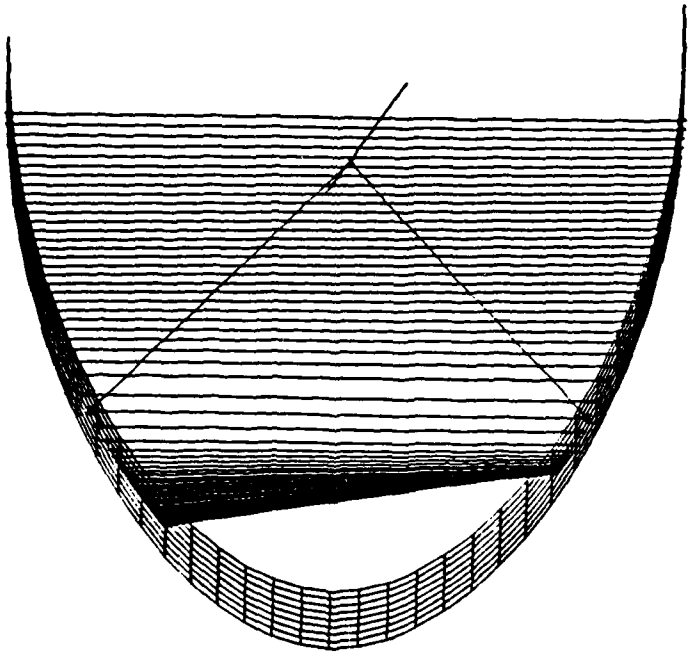


Figure 8: Double-Bounce Specular Point Paths

9 Conclusion and Near-Term Efforts

The use of twinkles to reduce the problem of finding specular points for many locations along a trajectory to that of tracking their motion can greatly reduce the computation time required to find the geometric optics return from a complex surface. Two fundamental problems are left to be solved before the method can be used widely. The first is the current lack of B-spline models which describe targets of interest. While spline modeling of complex targets is not expected to be easy, software packages such as the one developed by the Alpha_1 group at the University of Utah offer user-friendly tools which should allow for the practical development of detailed target models. It is expected that the creation of rapid and robust algorithms for the simulation of radar backscatter from complex B-spline surfaces will result in a significant demand for the development of a library of B-spline target models of interest.

The second basic problem is the local gradient search technique currently used to find twinkles and initial specular points. The algorithms used to demonstrate the method simply start searching for twinkles in the middle of each patch and the middle of the trajectory. The algorithms can find only one twinkle on a single spline patch (there may be more) and are not even assured of finding a twinkle when one exists. These problems are typical of local gradient search techniques when no additional information is used to determine the areas to be searched. Fortunately, there are properties of B-spline or B-splinelike surfaces which can be exploited to help assure global convergence of the algorithms. In particular, as B-spline patches are recursively subdivided [4], simple geometric tests based on bounds for both principal radii of curvature of the patches can be used to determine if it is possible for a twinkle (or initial specular point) to exist on that patch. With the use of a suitable stopping point, the patches could be subdivided (the majority being thrown out at each step) until only arbitrarily small patches exist which may contain twinkles (or initial specular points). Such algorithms should prove to be both rapid and robust. In addition, the method should extend readily to the multiple-bounce case, although it is not intuitively obvious what the more generalized geometric tests should be at this time.

References

[1] John F. Dammann, Jr., *Air Target Models for Fuzing Simulations,* Harry Diamond Laboratories, HDL-TR-1960 (September 1982).

[2] M. S. Longuet-Higgins, *Reflection and Refraction at a Random Moving Surface. I. Pattern and Paths of Specular Points,* J. Opt. Soc. Am. **50** (1960), 838.

[3] Ivor D. Faux and Michael J. Pratt, Computational Geometry for Design and Manufacture, Ellis Horwood, New York, NY, (1981).

[4] Elaine Cohen, Tom Lyche, and Richard Riesenfeld, *Discrete B-Splines and Subdivision Techniques in Computer- Aided Geometric Design and Computer Graphics,* Computer Graphics and Image Processing **14** (1980), 87-111.

AN ALGORITHM FOR PROCESSING SCANNING SPECTROMETER DATA

Joseph E. Zurlinden
Directed Energy Directorate/Operations Division
U.S. Army White Sands Missile Range, New Mexico 88002-5148

ABSTRACT - A method of processing data acquired from a scanning spectrometer is described. The method employs an algorithm designed to find the spectral data from a continuous stream of data from the spectrometer and to provide the experimentor with the relative peak intensities and relative powers of the known spectral lines. This algorithm does not require the use of outside reference sources, such as an electronic pulse synchronized with the output of the spectral data, to find each frame of spectral data. This can save thirty-three percent of the memory storage otherwise required for all the data from the spectrometer, and approximately twenty-five percent of the processing time on the computer.

I. INTRODUCTION. The High Energy Laser Systems Test Facility (HELSTF) is a research center for testing effects on various materials utilizing a high powered laser. The Army provides support to users in the form of secure and safe areas for performing experiments, and data acquisition and computer systems for collecting and processing the data. The author, as the data analyst insures that the test data is acquired successfully and processed satisfactorily before it is given to the user. The author presents in this paper a description of one of the data collecting instruments used at HELSTF, a scanning spectrometer, and the software developed to process the data from the scanning spectrometer. The algorithm is not original, but its use in this particular application anywhere else is unknown to this author.

Figure 1 shows a simplified diagram of the optical setup of the scanning spectrometer. This configuration is called a double-path Czerny-Turner spectrometer. Light from the source enters through the cassegrain subsystem at the lower left of the diagram. Optimum efficiency of the spectrmeter occurs when the cassegrain optics focuses the light on the slit such that the light fully illuminates the diffraction grating. The light proceeds from the slit to the lower spherical mirror, is reflected to the reflecting diffraction grating, after which the upper spherical mirror collimates and directs the dispersed beam to the flat mirror which reflects the dispersed beam to the scanning corner reflector. The scanning subassembly consists of 24 corner reflectors attached to a rotating disk with multiple rotating speeds available. The speed chosen here is such that the spectrum is scanned 800 times each second. As the corner mirror scans the spectrum, the light is reflected back through the Czerny-Turner optics where it travels back over the original paths until it is diverted to the two indium antimonide (InSb) detectors. These detectors are cooled to 77°K with liquid nitrogen; and they operate in the photoconductive mode.

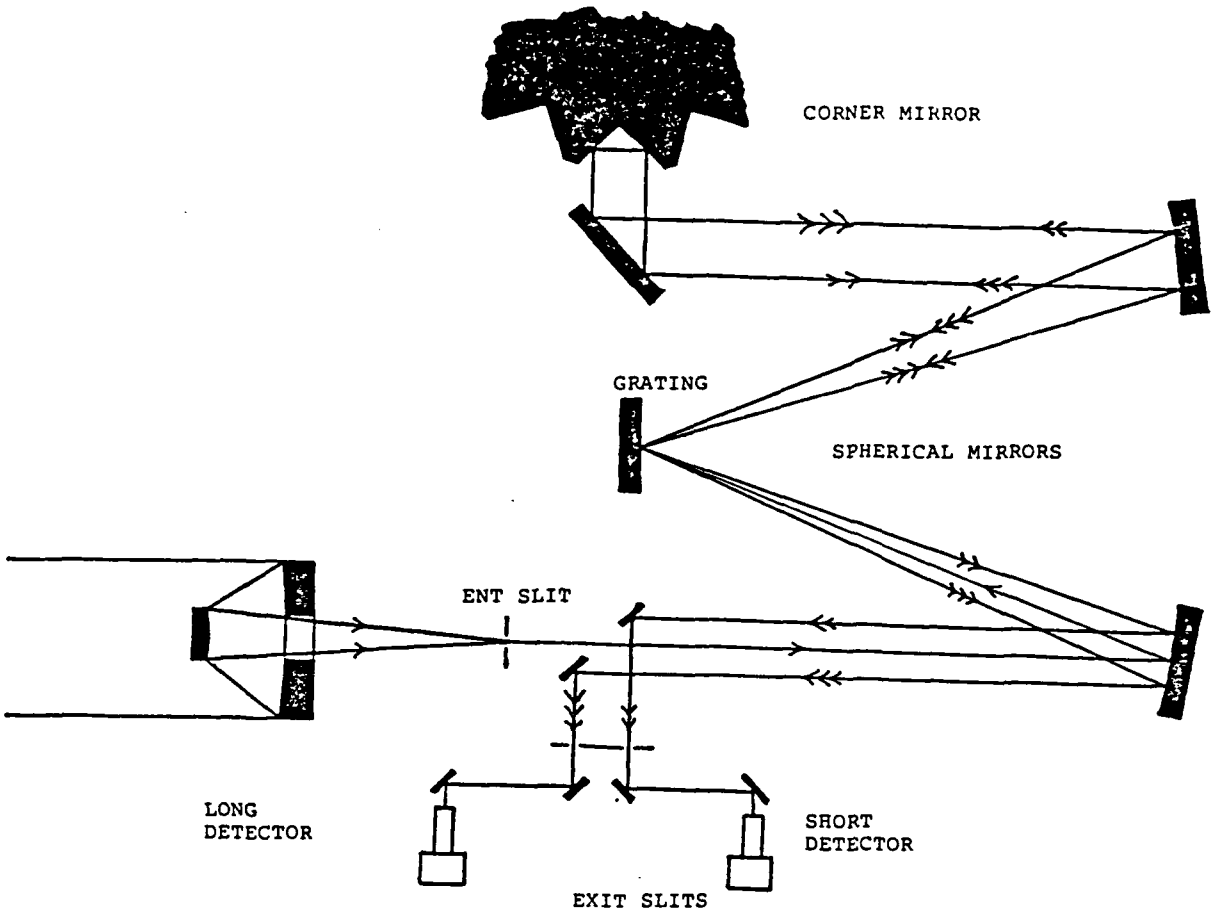


FIGURE 1. OPTICAL SCHEMATIC OF SCANNING SPECTROMETER

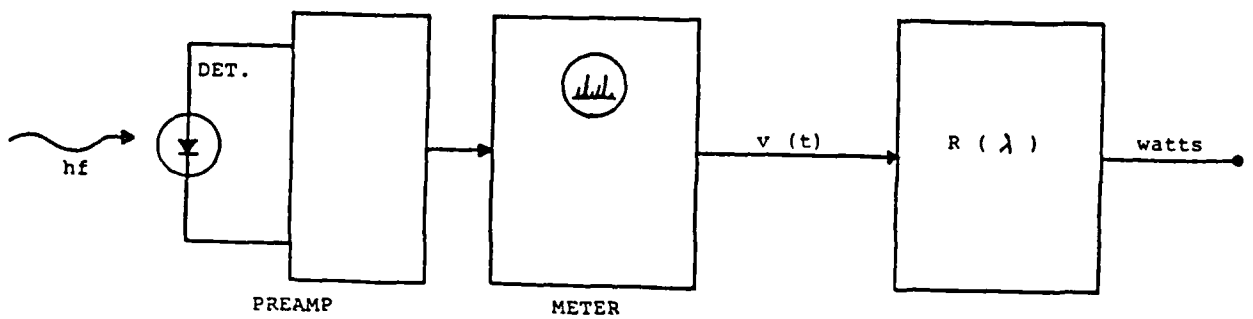


FIGURE 2. DIAGRAM OF THE PHOTOMETRIC PROCESS.

Figure 2 shows a simplified functional diagram of the photometric process. The detector sees a photon with frequency f and its conductivity is modified which produces a certain voltage level at the output of the circuit. The voltage level is measured with an electronic meter and recorded. The responsivity of the detector defines how much power corresponds to the measured voltage. There are many factors which determine the responsivity of a detector; such as the chemical composition, the environmental temperature, and so on. One can easily read more from a book on detectors. The box furthest to the right represents the process of calculating the absolute power from the responsivity values and the signal values.

The output from the detectors are fed into some amplifying electronics and output to BNC jacks. The two outputs cover the spectral region from 3.6 microns to 4.05 microns. The short wavelength region covers 3.60 to 3.85 microns and the long wavelength region covers 3.80 to 4.05 microns. Prior to each scan of the spectrum an electronic pulse, called the sync pulse, is generated with a duration of 0.250 milliseconds. This provides for a 1.0 ms duration for each frame of spectral data. This allows for a relation between the scan time and the wavelength of the spectrum. Plots of the signals are shown in Figure 3. The uppermost plot is all three signals multiplexed. The lower three plots are the three individual signals. The process of multiplexing data signals can be found in most books on digital communications. This data is from the spectrum of a chemical laser using deuterium and fluorine. The laser device is operated by TRW.

During the test, data from the spectrometer is FM recorded locally on three channels of the analogue tape. After the test, the FM tape is played back and the data digitized and multiplexed onto another tape. When one wishes to process the spectral data, it is transferred to disk and demultiplexed.

Originally the author's task was to take the demultiplexed data and develop software to determine the spectral line intensities and calculate the relative powers within the spectral lines. It was assumed that the first spectral line always occurred within a determined time interval and that the distances, in time or wavelength, between all the spectral lines remained constant. The software was designed to use as a reference point a specified level value of the leading edge of the sync pulse. Therefore, the program would read the sync pulse data and upon finding the reference point, it would know that it had to read so many points of the short and long wavelength data files before reaching the first spectral lines in each file. The program knew that the second lines were a certain number of data points from the first, the third from the second, and so on. These distances were different in the short and long data files. The program was run using test data and the output of the program seemed satisfactory; therefore, the task completed successfully, until a new requirement was generated by one of the users.

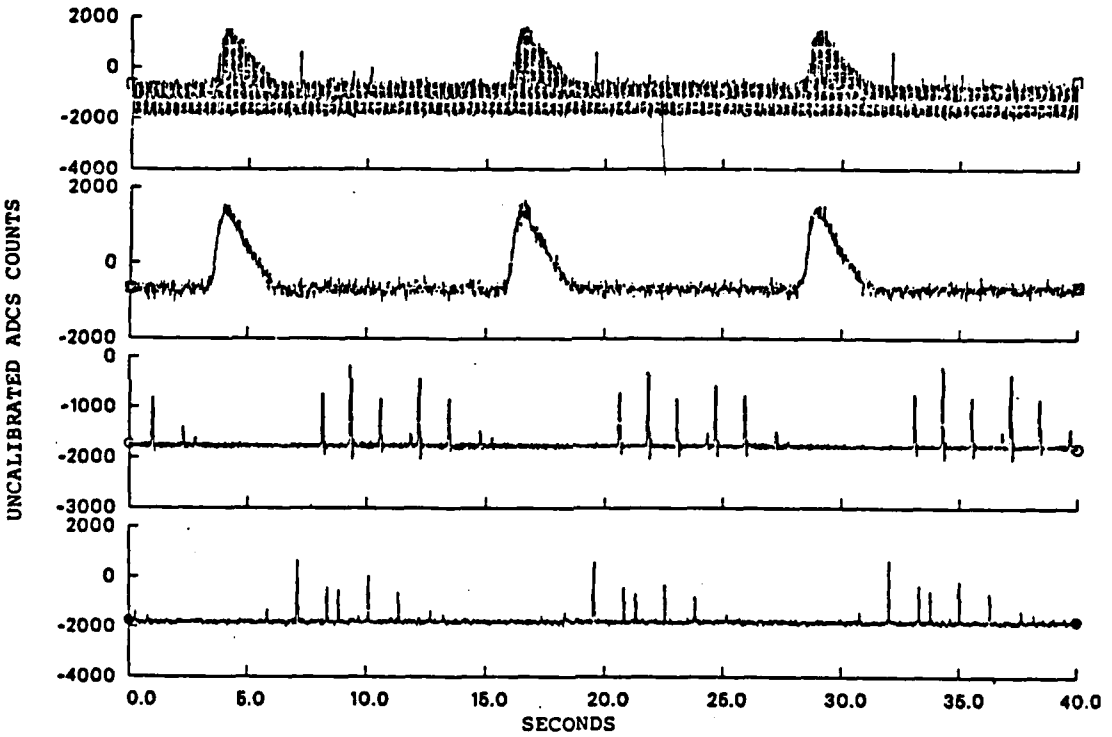


FIGURE 3. TOP: PLOT OF MULTIPLEXED DATA. SECOND, THIRD AND FOURTH PLOTS ARE OF SYNC PULSE, SHORT WAVELENGTH REGION AND LONG WAVELENGTH REGION, RESPECTIVELY.

One user was interested in seeing how each spectral line's intensity varied during the course of the test. I merely had to modify the program so that as the light intensities were determined they were written to a file: one file for each wavelength. The same was done for the energy in each line. The first time the modified program obtained outputs which resembled the plots of the spectra. A careful analysis revealed that the spectra was shifting relative to the sync pulses. However, the spacing between the spectral lines remained constant from one frame of data to the next. The cause of this shift was not immediately apparent, although it was thought that it was not in the software I had been developing. The cause for the shifting of the spectrum was found to be in the parameter values used in the demultiplexing algorithm. The error in these values caused a number of data points to be skipped over causing the spectrum data to be shifted toward the sync pulse data.

It was decided at this point not to use the sync pulse data and to find an algorithm which would identify the spectra using the characteristics of the spectrum itself. The only characteristic chosen was the spacing of the spectral lines, which one could consider as a pattern that occurred periodically many times in a long stream of time-series data. Here was a problem which was solvable using pattern recognition techniques.

II. DISCUSSION. The technique employed is based on the convolution integral. Here the integral is expressed as a summation because the data consists of discrete points. The expression is written as:

$$C(t) = \sum_{i=1}^{N} f(i)s(i-t),$$

where C(t) is the correlation at the t data point, f(i) is the ith point of the template or filter, and s(i) is the ith point of the real spectrum. As t increases to the end of the data, C(t) will vary and fluctuate from relative minima to relative maxima (see Figure 6). The number of data points in one frame of data is approximately 626 ± 4. There will be one relatve maxima for every 626 ± 4 consecutive values of C(t), and at this value of t, a spectrum begins. A complete explanation of correlation functions can be found in any text book on digital communications theory.

The software does three major functions: reads the data, scans the data determining locations of the frames of spectral data while it finds the peak intensity of each line and the relative energy in each line, and finally goes back and calculates frame averages of the spectra according to the options given to the user.

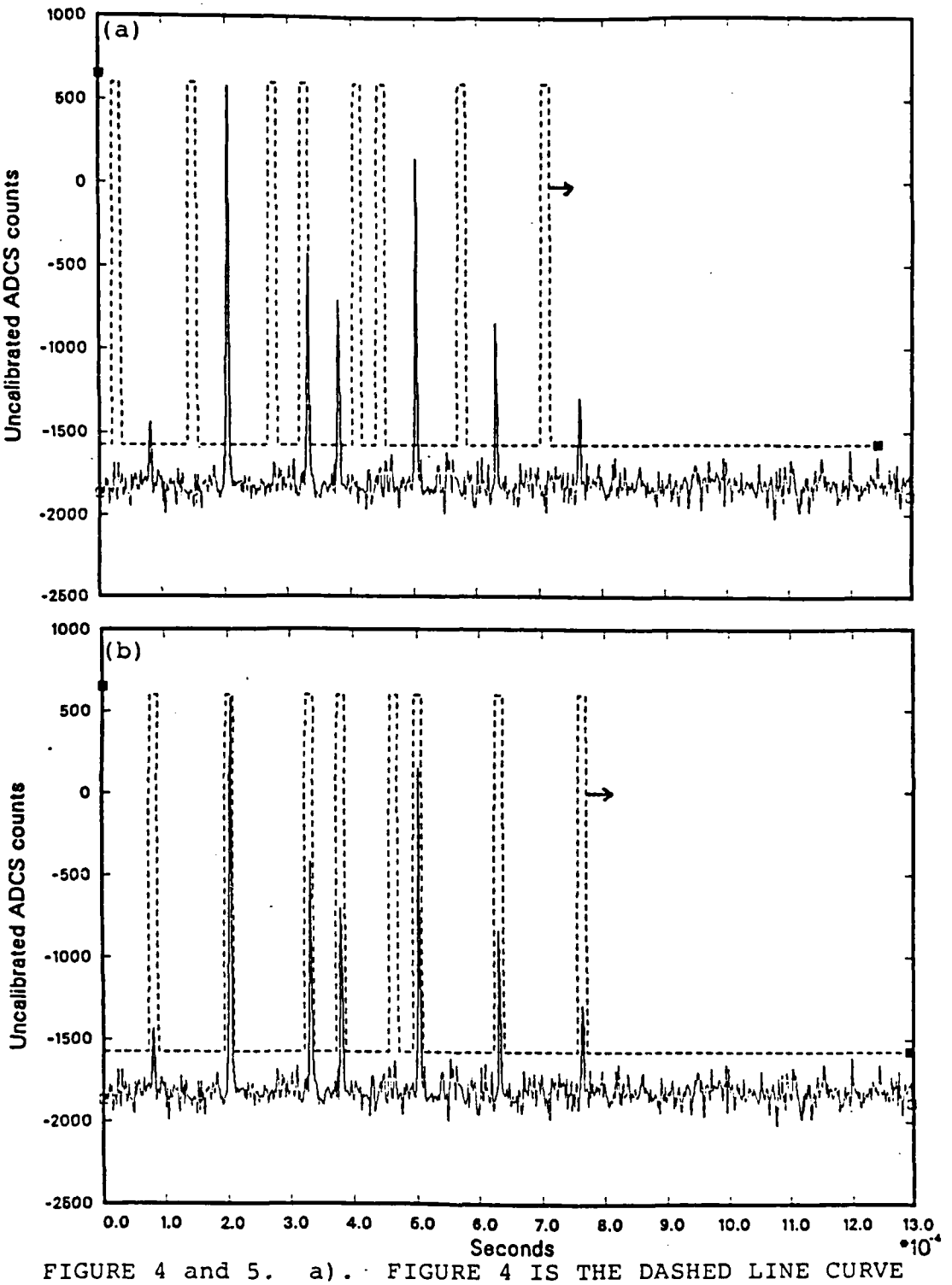


FIGURE 4 and 5. a). FIGURE 4 IS THE DASHED LINE CURVE SCANNING ACROSS FIGURE 5, THE SHORT WAVELENGTH SPECTRUM. b). HERE THE TEMPLATE IS IN THE POSITION CORRESPONDING TO A RELATIVE MAXIMA.

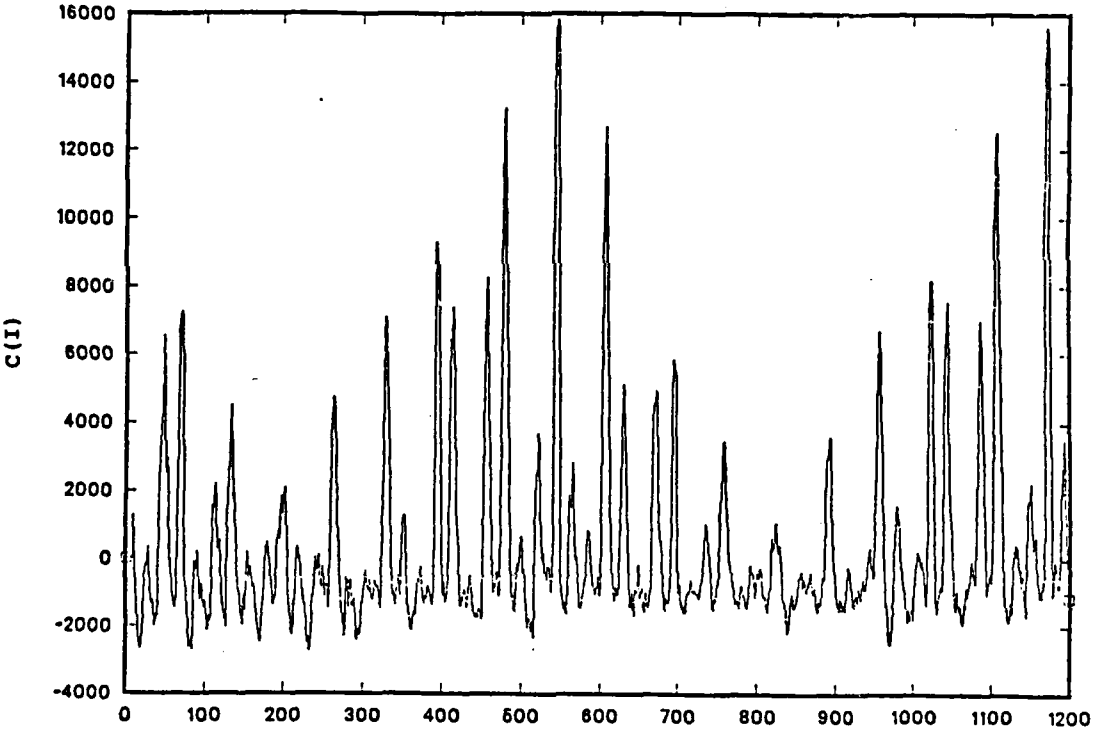


FIGURE 6. PLOT OF CORRELATION COEFFICIENTS, C(I), VERSUS THE DATA POINTS. THIS IS FROM A PARTIAL SCAN OF THE SHORT λ DATA.

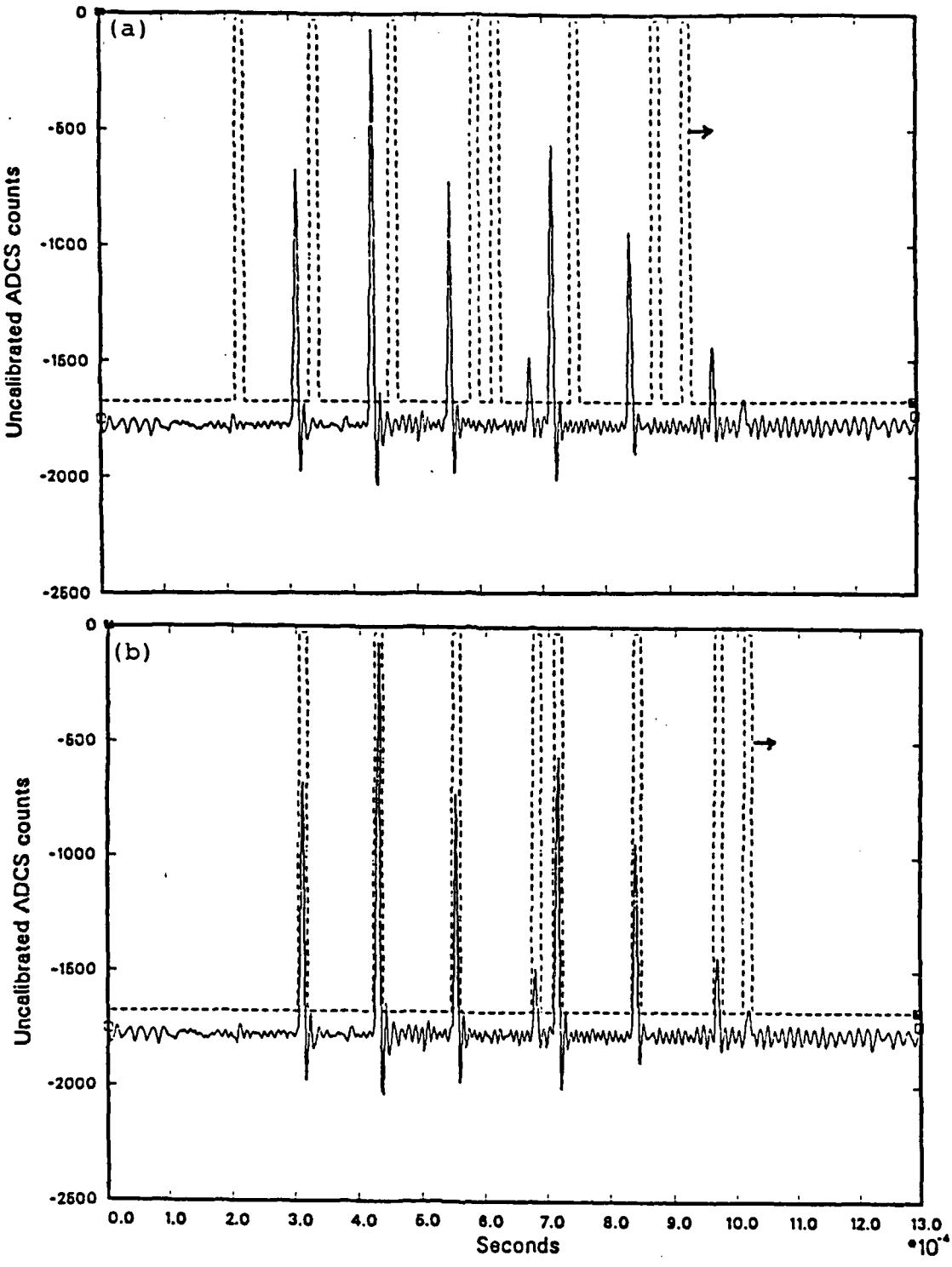


FIGURE 7 and FIGURE 8. a). FIGURE 7 IS THE DASHED LINE-CURVE SCANNING ACROSS FIGURE 8, THE LONG WAVELENGTH SPECTRUM. b). HERE THE TEMPLATE IS IN THE POSITION CORRESPONDING TO A RELATIVE MAXIMA.

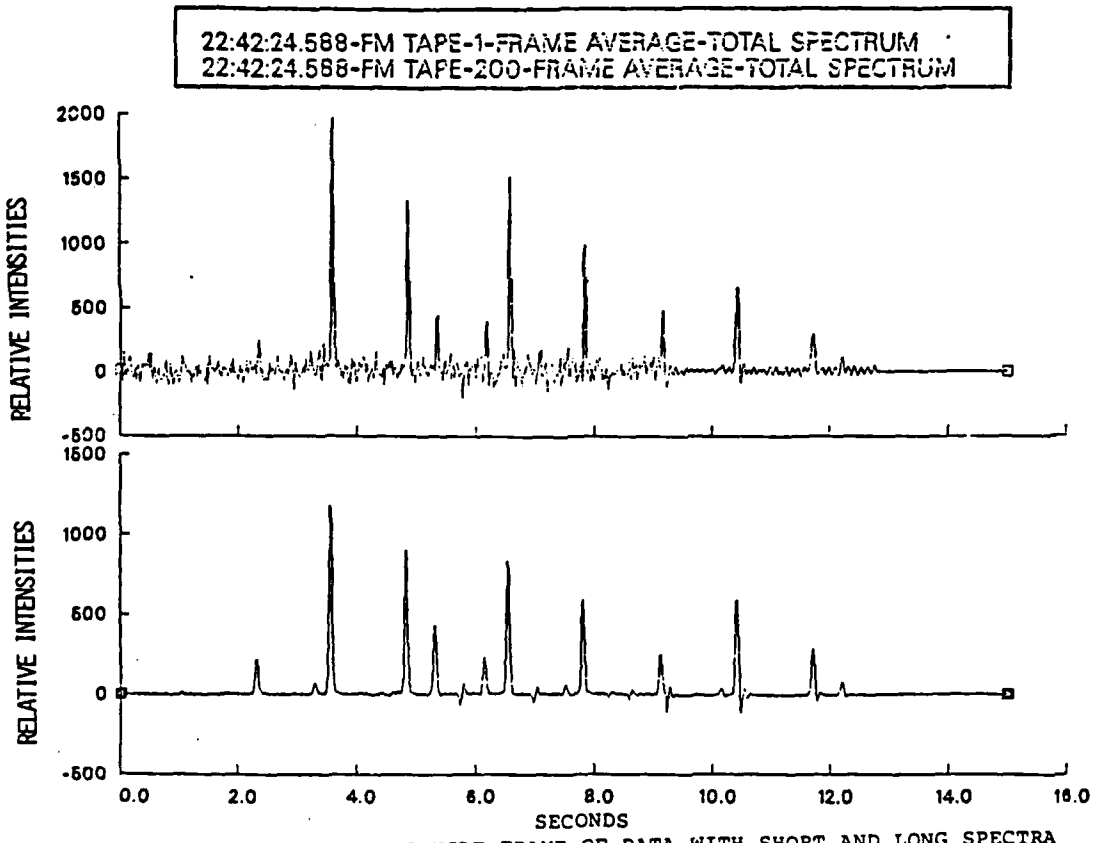


FIGURE 9. TOP PLOT: SINGLE FRAME OF DATA WITH SHORT AND LONG SPECTRA CONCATENATED. BOTTOM PLOT: RESULT FROM AVERAGING 200 FRAMES.

The user starts the program and enters values which allows the program to determine how much of the data is to be processed, whether or not averages are to be calculated, and how many are to be calculated. The data is then read and stored in memory, where now one-third less data is needed to be read and stored because the sync pulse data is no longer necessary. The spectral data is scanned and the locations of the spectra are determined and stored in memory. (See Figures 4, 5, 7, and 8). As the scanning is proceeding each value of the spectral line peak intensity is stored in a separate file, one file for each wavelength where a spectral line exists. The same is done with the energies within each line.

There are two detectors used and, therefore, values are different even for the same wavelength signal input. Therefore, the ratio of the energy contained within a line seen by both detectors is taken and the data from one detector corrected for the differences between the detectors. The short and long spectral data is concatenated at the overlaying regions to show the entire spectrum, as shown in Figure 9.

Finally the program calculates averages if they were requested. If not, the program is finished. If averages are requested, then the number of sets of frames will be averaged. If more than one set of frames is averaged, then the first set of N frames is averaged, then the next set is made up of averaging N frames starting with the second frame and including the (N+1)th frame. This continues until the program executes all the instructions based on the values the user inputs at the start of the program.

III. CONCLUSION. The degree to which the software is successful, in its ability to recognize the spectrum, is obvious as can be seen in Figure 9. The top plot shows a single frame of data. The bottom plot shows the average of 200 consecutive frames which were averaged together. The width of the lines in the averaged frame are the same as those in the nonaveraged frame. This would not be the result if the recognition of the spectrum, by the artificial spectrum (template or filter) was off by even one data point in position. This is the purpose of the algorithm because once the spectrum is located, the calculations are straight forward. The relative energies in the spectral lines were calculated using the trapezoidal rule. The software I developed was done using FORTRAN 77 on a VAX11-780 computer system.

References:

Radziemski, L. J., Intro to Spectral Analysis, 1987.

THE KP EQUATION - A COMPARISON TO LABORATORY GENERATED BIPERIODIC WAVES*

Norman W. Scheffner
U.S. Army Engineer Waterways Experiment Station
Coastal Engineering Research Center
Vicksburg, MS 39180

ABSTRACT. The propagation of waves in shallow water is a phenomenon of significant practical importance. The ability to realistically predict the complex wave characteristics occuring in shallow water regions has always been an engineering goal which would make the development of solutions to practical engineering problems a reality. The difficulty in making such predictions stems from the fact that the equations governing the complex three-dimensional flow regime can not be solved without linearizing the problem. The linear equations are solvable; however, their solutions do not reflect the nonlinear features of naturally occuring waves. A recent advance (1984) in nonlinear mathematics has resulted in an explicit solution to a nonlinear equation relevant to water waves in shallow water. The solution possesses features found in observed nonlinear three-dimensional wave fields.

The nonlinear mathematical formulation referred to above has never been compared with actual waves, so that its practical value is unknown. The purpose of the present investigation was to physically generate three-dimensional waves and compare these with exact mathematical solutions. The goals were successfully completed by first generating the necessary wave patterns with the new U.S. Army Engineer Waterways Experiment Station, Coastal Engineering Research Center's (CERC) directional spectral wave generation facility. The theoretical solutions were then formed through the determination of a unique correspondence between the free parameters of the solution and the physical characteristics of the generated wave.

I. INTRODUCTION. One of the first mathematical models of nonlinear waves in shallow water with known solutions was presented by Korteweg and deVries in their famous 1895 paper. Their formulation, known as the KdV equation, can be written in the following nondimensional form

$$f_t + 6ff_x + f_{xxx} = 0 \qquad (1)$$

in which f represents the water surface displacement, x is the direction of propagation, and t is time. This equation admits not only solitary wave solutions but also the periodic solutions commonly known as cnoidal waves.

*Presented at the 20th International Conference on Coastal Engineering, November 9-14, 1986, Taipei, Taiwan and included in the proceedings thereof entitled "Biperiodic Waves in Shallow Water"

These solutions can be written as

$$f(x,t) = 2\sigma^2 k^2 cn^2(\theta;k) - 2\sigma^2\left[\frac{E(k)}{K(k)} - 1 + k^2\right] \qquad (2)$$

where each of the terms in the solution are well documented analytic functions which can easily be computed in terms of known wave characteristics such as wave height and wavelength. Unfortunately, cnoidal wave solutions are valid only for long crested waves, e.g., waves which can be described by a single time-dependent one-dimensional surface wave pattern. Natural waves, in contrast, are composed of both long and short crested waves and can not be adequately described by this theory.

A recent advance in nonlinear mathematics has been reported by Segur and Finkel (1984). They present explicit analytical solutions to a natural three-dimensional extension of the KdV equation proposed by Kadomtsev and Petviashvili (1970), known as the KP equation shown below

$$(f_t + 6ff_x + f_{xxx})_x + 3f_{yy} = 0 \qquad (3)$$

where x now represents the primary direction of propagation; however, weak changes in the y-direction are now permitted. When no y-variations occur, the KP equation reverts to the KdV equation.

The KP equation admits an infinitely dimensional family of exact, periodic solutions (see Dubrovin, 1981 and Segur and Finkel, 1984) which can be written in the form

$$f(x,y,t) = 2\frac{\partial^2 \ln\theta}{\partial x^2} \qquad (4)$$

where θ is a Riemann theta function of genus n. Genus 1 solutions are exactly equivalent to cnoidal waves, they are permanent form, singly periodic, two-dimensional (one vertical and one horizontal) nonlinear waves. Genus 2 waves are biperiodic in that they permit the independent specification of two periodicities in both time and space. The solutions are genuinely three-dimensional, nonlinear, and propagate with permanent form at a constant velocity. Genus 3 and higher order solutions are multi-periodic and can not be characterized as permanent form with respect to any translating coordinate system as the genus 1 and 2 solutions can. This present investigation is limited to the genus 2 solutions developed by Segur and Finkel.

The construction of a genus 2 solution of the KP equation is based on the specification of the appropriate Riemann theta function. This requires the introduction of a two-component phase variable and a 2×2 real-valued Riemann matrix. The first of these, the phase variable, is shown below.

$$\phi_1 = \mu_1 x + \nu_1 y + \omega_1 t + \phi_{10}$$

and (5)

$$\phi_2 = \mu_2 x + \nu_2 y + \omega_2 t + \phi_{20}$$

Where the parameters μ_1, μ_2, ν_1, and ν_2 are wave numbers, ω_1 and ω_2 are angular frequencies, and ϕ_{10} and ϕ_{20} are constants with no dynamical significance. The second ingredient involves the specification of a real-valued, negative definite, symmetric 2 X 2 Riemann matrix as shown below.

$$B = \begin{pmatrix} b & b\lambda \\ b\lambda & b\lambda^2 + d \end{pmatrix} \tag{6}$$

The parameters b , d , and λ represent solution nonlinearity. The genus 2 theta function can now be defined in terms of the above components by the following double Fourier series:

$$\theta(\phi_1,\phi_2,B) = \sum_{m_1=-\infty}^{\infty} \sum_{m_2=-\infty}^{\infty} \exp\left(\frac{1}{2}\vec{m} \cdot B \cdot \vec{m} + \vec{m} \cdot \vec{\phi}\right) \tag{7}$$

The calculation of a general case genus 2 KP solution requires the specification of the 11 parameters shown in Equations 5 and 6. Two of these parameters (ϕ_{10} and ϕ_{20}) have no dynamical significance, their only effect is to shift the origin of the resulting solution. Dubrovin (1981) proved that a genus 2 theta function in the form of Equation 7 was a solution to the KP equation if, and only if, the solution parameters were related by four additional equations. One of these equations contains a constant of integration. Use of this additional criteria reduces the number of free parameters to 8, representing the minimum number of free parameters required to specify a general case genus 2 solution.

Genus 2 solutions of the KP equation describe a complex two-dimensional surface wave pattern. Similar features were observed by Hammack (1980) to result from the nonlinear interaction of two intersecting waves. The theoretical development by Segur and Finkel was partially prompted, in fact, by these reported waves. The development of an experimental program which would result in the generation of surface wave patterns qualitatively similar to genus 2 solutions was achieved by attempting to experimentally reproduce the conditions reported by Hammack, i.e., intersecting waves. This generation technique can best be described by presenting the analogy of interacting waves. Consider, for example, two periodic waves which intersect and pass through each other as shown in Figure 1. The angles α_1 and α_2 represent the angle of the crest of each wave front with respect to some reference line. The resulting surface wave pattern, according to linear wave theory, would simply be a superposition of the two individual waves. This would produce a diamond shaped surface pattern as indicated in Figure 1. It can be seen that certain of the basic characteristics of the individual waves, wavelength and angle of propagation for example, have been preserved.

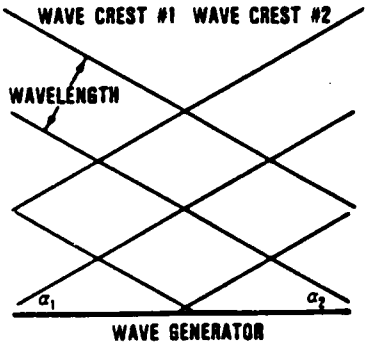


Figure 1. The Linear Intersection of Waves

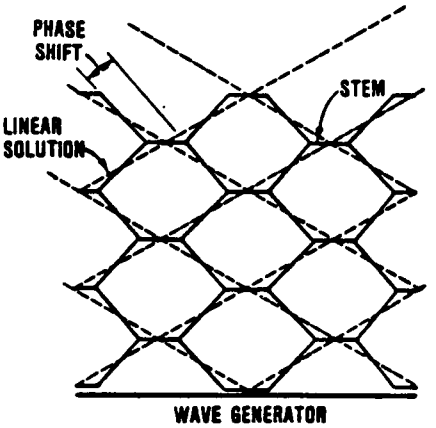


Figure 2. The Nonlinear Intersection of Waves

Now, consider the analagous case in which similarily intersecting waves interact nonlinearily with each other. This scenario is shown schematically in figure 2. The resulting wave pattern shows that a "stem of interaction" is formed at the point where the two waves cross each other. The formation of this stem region is a result of a phase shift in the crest line angles of the original waves. This phenomonon is shown in Figure 2 superimposed on the corresponding linear wave solution. The resulting surface wave pattern now assumes a hexagonal pattern in which a third wave crest, seperate of the original two, is formed. This phase shift and stem formation are indicative of the nonlinear interaction of the two waves since the exact linear solution does not predict either the phase shift or the new wave crest. Genus 2 solutions of the KP equation predict these features and was tested as a possible model for their description.

II. LABORATORY FACILITIES AND EXPERIMENTAL PROCEDURES. A project was initiated at CERC to generate three-dimensional nonlinear wave fields in the laboratory and then apply KP theory to the resulting waves in order to determine whether or not the KP equation was a model for these waves and, if so, what was the range of its applicability. This required the use of the CERC directional spectral wave generation facility. This unique wave generator, shown in Figure 3, was designed and constructed for CERC by MTS Systems Corporation of Minneapolis, Minnesota based on design specifications provided by CERC. The generator is comprised of 60 individually programmable electromechanical wave paddles. Each wave paddle is 1.5 ft wide making the generator a total of 90.0 ft wide. The generator is located in a 98.0 by 184.0 ft wave basin with 2.5 ft high side walls. Computer control of the system is provided by a Digital Equiptment Corporation (DEC) VAX 11/750 central processing unit. The above facilities were utilized to generate genus 2 candidate waves in a comprehensive experimental program.

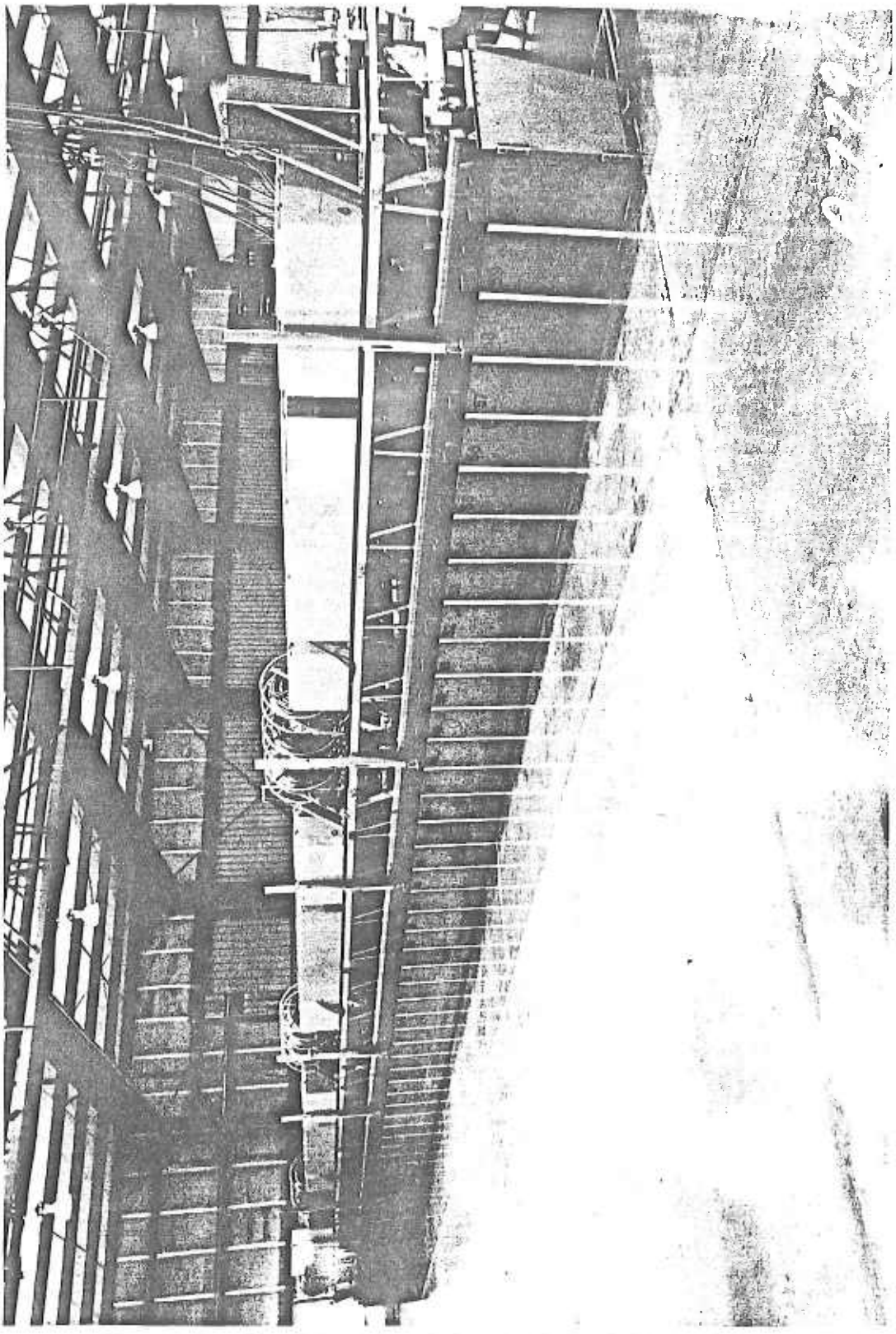


Figure 3. The Directional Spectral Wave Generator

The wave generator was programmed to simultaneously generate intersecting cnoidal wave trains. A variety of wave fields were generated by varying both the wavelength of the individual waves and their angle of intersection. Twelve wave fields, generated in this manner, were used to test the KP equation. The wave fields selected for the experimental program are presented in Table 1. Waves characterized by three wavelengths (7, 11, and 15 ft) were combined with phase shifts between adjacent wavemaker paddles. These phase shifts were approximately equivalent to the angle of the wavecrest with respect to the axis of the wave generator. The angle in the table shows the approximate correspondence between the phase lag and the angle of propagation.

Table 1
The experimental waves

Test Number	Wavelength (ft)	Phase Shift (deg)	Angle (deg)	Period (sec)
CN1007	7.0	10.0	7.45	1.378
CN1507	7.0	15.0	11.21	1.378
CN2007	7.0	20.0	15.03	1.378
CN3007	7.0	30.0	22.89	1.378
CN4007	7.0	40.0	31.23	1.378
CN1011	11.0	10.0	11.75	1.947
CN1511	11.0	15.0	17.79	1.947
CN2011	11.0	20.0	24.04	1.947
CN3011	11.0	30.0	37.67	1.947
CN1015	15.0	10.0	16.12	2.553
CN1515	15.0	15.0	24.62	2.553
CN2015	15.0	20.0	33.75	2.553

Genus 2 solutions can be visualized as a series of repeating two-dimensional permanent form surface patterns, referred to as period parallelograms. These patterns translate at a constant velocity in a constant direction. The global wave field is represented by a tiling of these basic patterns; therefore, the entire wave pattern can be exactly specified by quantifying just one period parallelogram. The location of a basic parallelogram within the hexagonal wave field of Figure 2 is shown in Figure 4. The phase variables of Equation 5 define the horizontal limits of these patterns such that each side is uniquely defined by ϕ_1 = constant and ϕ_2 = constant. The components of the Riemann matrix define the vertical and horizontal distribution within the period parallelogram.

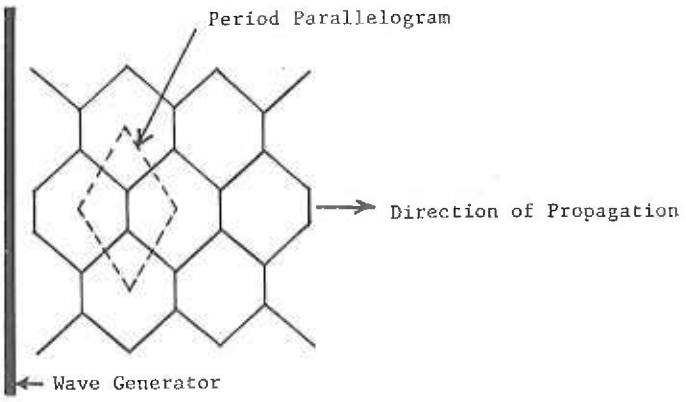


Figure 4. The Period Parallelogram

Detailed measurements of each of the generated wave fields shown in Table 1 were required in order to relate the physical characteristics of the waves to the parameters of the corresponding period parallelogram of the exact solution. This quantification was accomplished by first using overhead photography to determine the dimensions of the period parallelogram and to provide an estimate of the internal features, such as the phase shift and stem length. Knowledge of these horizontal features and their location within the wave tank were then used to locate a linear array of 9 recording wave gages in the wave basin. This approach provided a vertical wave record which could be identified with a known location within the parallelogram.

III. COMPARING THEORETICAL SOLUTIONS TO OBSERVED WAVES. The experimental program described above generates symmetric cnoidal waves ($\alpha_1 = \alpha_2$ in Figure 1) resulting in a symmetric period parallelogram. This simplification was adopted so that the generated wave patterns would all propagate perpendicularly off the axis of the wave generator, making it possible to measure all wave forms with a single stationary wave gage array. Symmetry also reduces the number of free parameters which need to be specified, for example, $\mu_1 = \mu_2$, $\nu_1 = -\nu_2$, and $\omega_1 = \omega_2$ from Equations 5. This simplification results in the requirement of only three dynamical parameters and two nondynamical parameters. The parameters choosen were b , μ , and λ along with the phase shift parameters ϕ_{10} and ϕ_{20} . The following sequence of events was used for optimizing these coefficients. Experiment CN3007 will be used to demonstrate the verification process.

Each of the waves of Table 1 were generated in the wave basin. Two overlapping photographs were taken with dual Hasselbladt model 500EL/M 70mm cameras equipped with 50mm lenses mounted 23 ft above the floor of the basin. The resulting mosaic photograph, shown in Figure 5, was used to estimate the length and width of the period parllelogram. This resulted in estimates for $\mu_1 = \mu_2$ and $\nu_1 = -\nu_2$. An estimate for the phase shift parameter λ was also determined from the photograph. The accuracy of μ , ν , and λ is a function of the distortions in the photograph. Because of this distortion, their values were considered to be initial estimates. Following the photographing of all waves, a gage spacing of 2.5 ft apart and 40.0 ft from

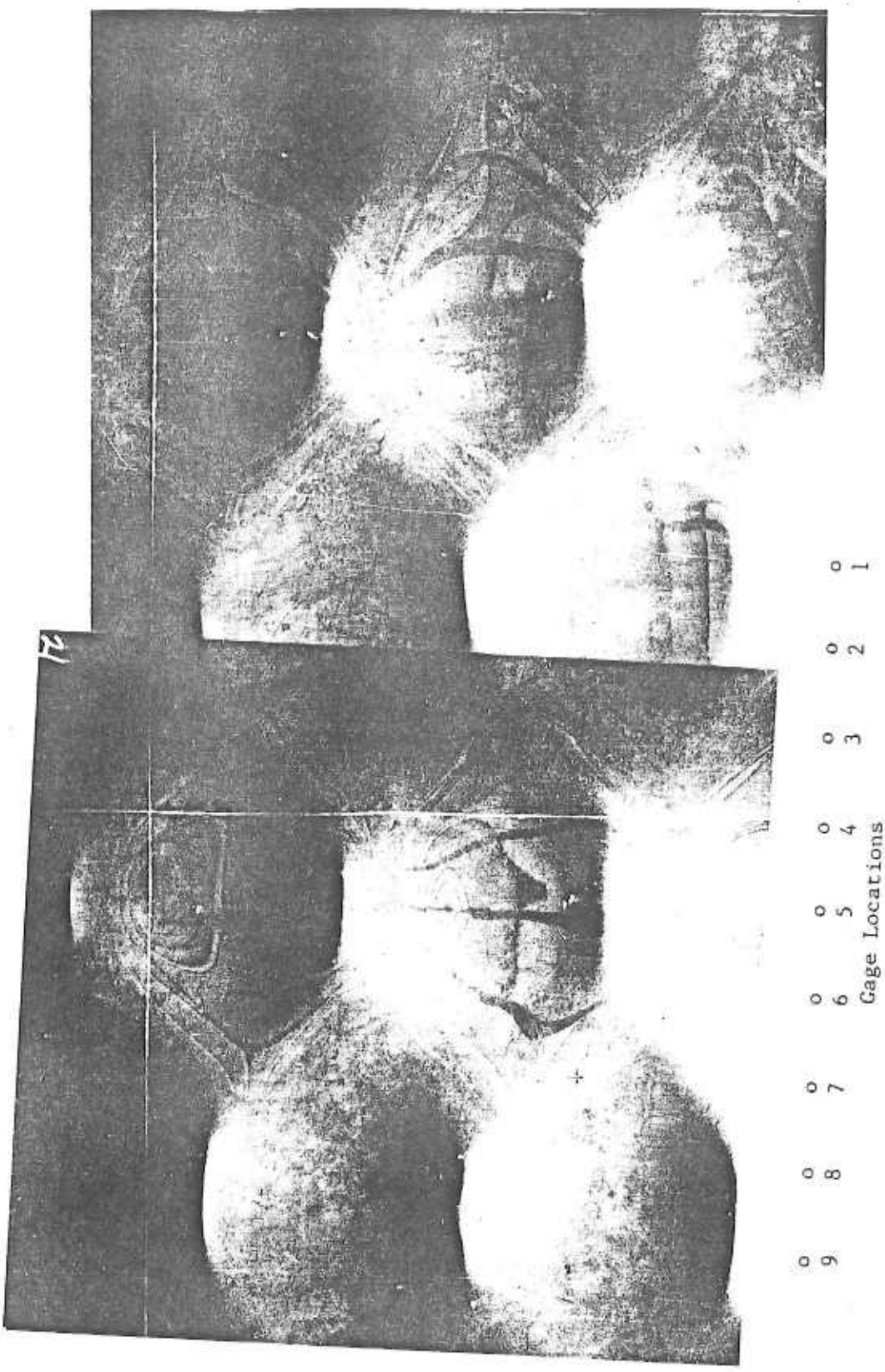


Figure 5. Overhead Mosaic Photograph of Test Wave CN3007

and parallel to the generator was selected for use in all tests. The location of each of the gages with respect to wave CN3007 is shown in Figure 5. It can be seen that each gage can be uniquely referenced according to a distance from the center of the parallelogram. Since all parallelograms are identical, wave gages located in an adjacent parallelogram can be referenced to the common center point.

Wave gages were located in the basin and each of the waves of Table 1 were regenerated. Data were sampled for each of the gages at a rate of 50 samples per second for a total of 30.0 seconds. Figure 6 shows the wave traces for CN3007. The correspondence between the wave traces and their location within the parallelogram can easily be seen. For example, gage 5 is located on a stem where only one peak per passing of the parallelogram is experienced. Gage 3 is located in the saddle region where two smaller peaks per period are seen. This comparison demonstrates the usefulness of the photographs in interpreting the data since three-dimensional effects are difficult to deduce from two-dimensional data.

The determination of the free coefficients can now be made. Known or estimated data are the period of the wave (determined from the recording wave gages), the length and width of the period parallelogram and an estimate of the phase shift parameter λ determined from the photographs, and a maximum wave height selected from the wave gage data. The following iteration procedure was used to optimize the coefficients:

a. The estimated values for $\mu_1 = \mu_2$, $\nu_1 = -\nu_2$, and λ were specified. The nondynamical parameters ϕ_{10} and ϕ_{20} were accounted for by specifying solutions to be computed at location within the period parallelogram corresponding to the location of the wave gages. A value of b was then selected such that the dimensionalized maximum KP solution was within 5.0 percent of the measured value.

b. The value of $\mu_1 = \mu_2$ was adjusted, if necessary, until the dimensionalized period was within 3.0 percent of the measured period.

c. The value of λ was adjusted, if necessary, until the dimensionalized value of $\nu_1 = -\nu_2$ was within 10.0 percent of the estimated value. A 10-percent criteria was used for this iteration since the length of the parallelogram was difficult to determine from the photographs.

d. Because of the nonlinear coupling of the solution coefficients, each adjustment affected all parameters to some extent. If corrections were found to be necessary, steps (a.) through (c.) were repeated until all of the specified tolerances were met or exceeded. Possible phasing problems regarding the gage locations within the parallelogram were rectified by adjusting the nondynamical phase parameters.

e. A KP solution corresponding to the location of each of the wave gages was calculated. A normalized plot comparing theory to measurements was made, as shown in Figure 7 for the present example. Included in each plot is the Root Mean Square (RMS) error for each comparison.

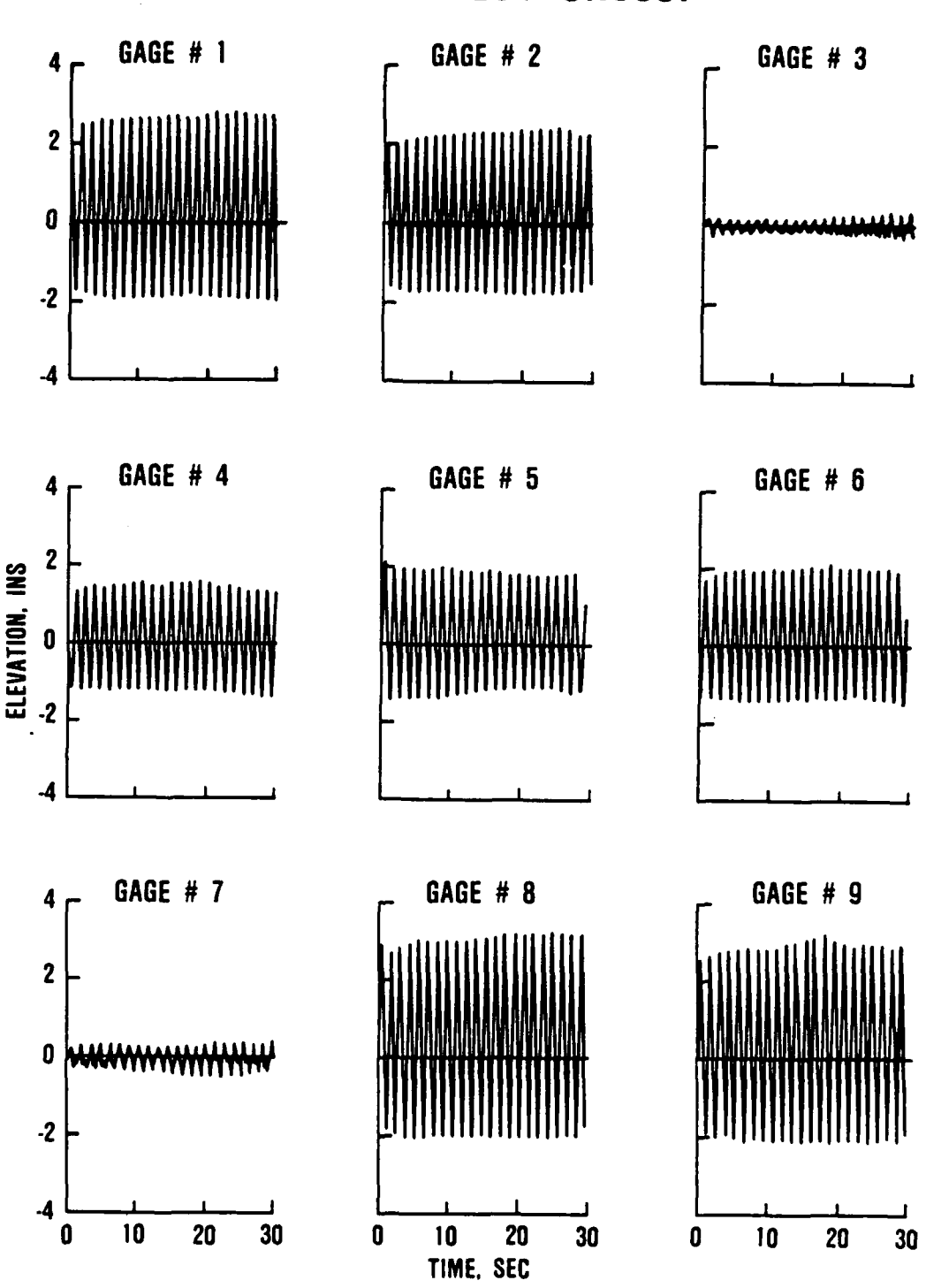


Figure 6. Wave Gage Traces for Test Wave CN3007

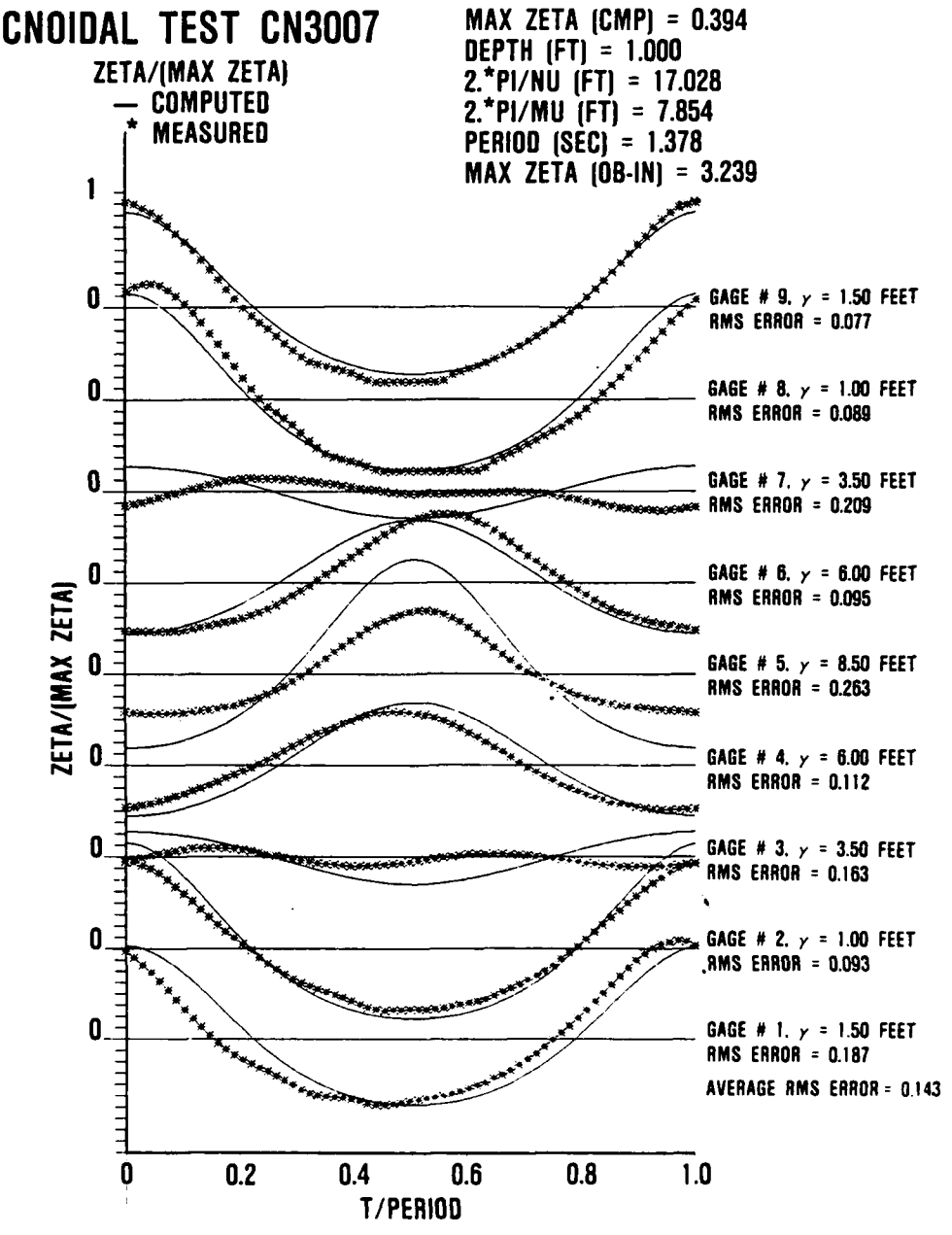


Figure 7. Theoretical and Measured Wave Profiles for the Nine Wave Gages of Test Wave CN3007

f. A normalized contour plot (Figure 8) and a three-dimensional plot (Figure 9) for each wave field was finally prepared as a visual example of the KP solution.

The above procedures were followed for each of the test wave fields of Table 1. A minimum tolerance of 5.0 percent for waveheight, 3.0 percent for period, and 10.0 percent for the Y-direction wavelength was maintained for all experiments. Table 2 presents those computed results. For each case, an average RMS error is provided which represents a simple average of the 9 RMS values computed for each gage. In no case did this error exceed 20 percent even though variations in the elevation of the basin floor of 10 percent were known to exist. Additionally, the experimental wave fields were generated almost to the point of breaking in order to span the range of solution parameters and investigate the limits of applicability of the genus 2 solutions. In view of these introduced and existing sources of potential error, the degree of fit between the generated wave fields and the exact solutions were found to be very good.

Table 2
Computed wave parameters

Test Number	Max. Height (in)	X-Wavelength (ft)	Y-Wavelength (ft)	Ave. RMS Error
CN1007	2.44	7.0	46.5	0.141
CN1507	3.59	7.2	35.1	0.188
CN2007	3.06	7.5	27.3	0.150
CN3007	3.24	7.9	17.0	0.143
CN4007	3.30	8.7	13.6	0.184
CN1011	2.23	10.7	48.0	0.174
CN1511	2.87	11.1	40.3	0.122
CN2011	3.10	11.6	27.6	0.126
CN3011	2.48	12.6	20.7	0.172
CN1015	2.65	15.0	59.3	0.120
CN1515	2.84	16.1	32.6	0.094
CN2015	2.86	17.1	29.0	0.098

IV. CONCLUSIONS. Twelve seperate nonlinear wave fields were generated for the purpose of verifying the KP equation to be an accurate model for three-dimensional nonlinear waves. Criteria were developed which provided a unique correspondence between the solution parameters of the KP equation and the physical characteristics of the laboratory generated waves. Results of these experiments showed that both the generated waves and the genus 2 solutions are remarkably robust in that both were stable over a wide range of parameters, including the near breaking of waves. The excellent degree of fit between the observed and computed solutions shows that the genus 2 solutions of the KP equation represent a viable model for three-dimensional, nonlinear, shallow water waves.

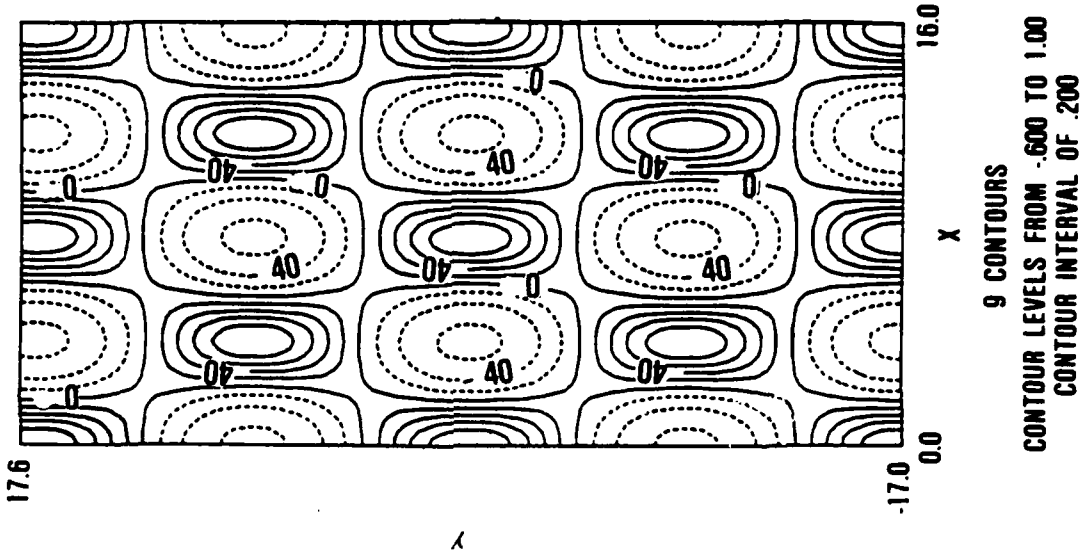


Figure 8. Normalized Contour Plot of Test Wave Field CN3007

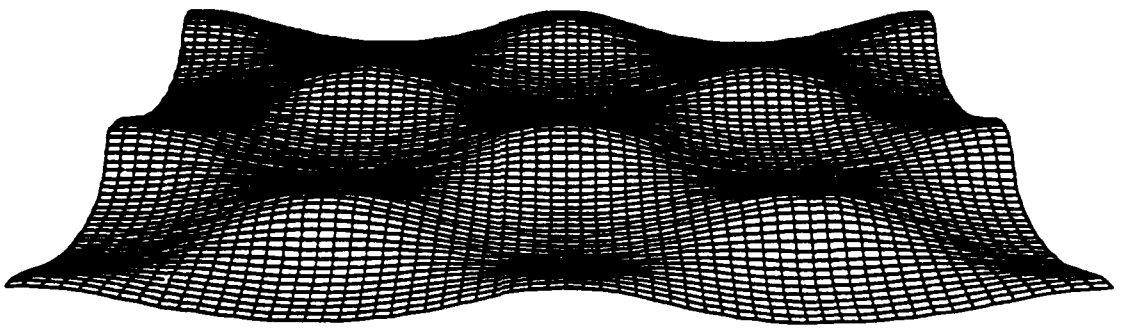


Figure 9. Three-dimensional Plot of Test Wave CN3007

V. ACKNOWLEDGEMENTS. The author is greatful to Drs. Harvey Segur and Joseph Hammack for their help and advice, both of which were necessary for the successful completion of this project. The research results contained in this paper were funded through a Department of the Army In-House Laboratory Independent Research (ILIR) program. The author wishes to acknowledge the Office, Chief of Engineers, U. S. Army Corps of Engineers, for authorizing publication of this paper.

REFERENCES.

Dubrovin,B.A. (1981), "Theta Functions and Non-Linear Equations", Russian Math. Surveys, Vol. 36:2, pp.11-92.

Hammack,J.L. (1980), Unpublished Experiments.

Kadomtsev,B.B. and Petviashvili,V.I. (1970), "On the Stability of Solitary Waves in Weakly Dispersive Media", Soviet Physics - Doklady, Vol. 15, No. 6, pp. 539-541.

Korteweg, D.J. and deVries,G. (1895), "On the Change of Form of Long Waves Advancing in a Rectangular Channel, and on a New Type of Long Stationary Waves',Phil. Mag., Ser. 5, Vol. 39, pp. 422-443.

Segur,H. and Finkel,A. (1984), "An Analytical Model of Periodic Waves in Shallow Water", Aeronautical Research Associates of Princeton, Inc., Tech. Memo. 84-12.

ASYMPTOTICS BEYOND ALL ORDERS

Harvey Segur

ARAP/Titan Co.
Princeton, NJ 08543-2229
and
Department of Mathematics
SUNY at Buffalo
Buffalo, NY 14214-3093

ABSTRACT: Conventional asymptotic methods often fail to capture effects that are transcendentally small because these effects lie beyond all orders in the asymptotic expansion. New methods have been developed recently to find transcendentally small terms in problems where they control the entire solution. This paper surveys some of these recent developments.

TEXT: When a differential equation that cannot be solved exactly contains a small parameter, conventional asymptotic methods often can be used effectively to find approximate solutions[1]. In a small but significant class of problems these conventional methods uttterly fail, because they provide no nontrivial information at any order of the expansion. In these problems it is necessary to go beyond all orders in the asymptotic expansion to answer questions of interest. "Asymptotics beyond all orders" describes the more delicate methods required to obtain information in these pathological problems.

The essential problem can be seen in a simple function like

$$f(\epsilon) = \exp(-1/\epsilon), \qquad 0< \epsilon <<1. \qquad (1)$$

The function is well-defined, and it is positive for any positive ϵ, no matter how small. The function is not analytic at $\epsilon = 0$, so it has no Taylor series there. If one ignores this fact and tries to evaluate $f(\epsilon)$ for small ϵ with a formal "Taylor series", one obtains zero at every order of the expansion:

$$f(\epsilon) \sim 0 + \epsilon\cdot 0 + \epsilon^2\cdot 0 + \epsilon^3\cdot 0 + \ldots . \qquad (2)$$

(Here $\sim$ means "is asymptotically approximated by" [1].) Clearly (2) does not imply that the function is zero, but simply that the expansion is too crude to evaluate it.

This example is too simple to be realistic, but a variation of it is common: the function is not given explicitly, as in (1), but rather it is defined implicitly by a differential equation. For example, let

$$D(y, x; \epsilon) = 0, \qquad 0 < \epsilon \ll 1, \qquad (3)$$

represent some differential equation which, along with boundary conditions, uniquely defines its solution $y(x; \epsilon)$. We imagine that the equation comes from some application, and that the question of interest is to determine the sign of $y(0; \epsilon)$, the solution of (3) at $x=0$. If the differential equation cannot be solved exactly, as most cannot, conventional asymptotic methods [1] can be used to generate an approximate solution. In the simplest cases, the series contains only increasing powers of ϵ:

$$y(x; \epsilon) \sim y_0(x) + \epsilon\, y_1(x) + \epsilon^2 y_2(x) + \ldots \quad .$$

How many terms are needed in this series depends on the problem, but in any case the sequential approximating functions are obtained explicitly. With these functions explicitly in hand, one evaluates them sequentially at $x=0$ to obtain an increasingly accurate description of the desired function, $y(0; \epsilon)$. Ordinarily this approach is successful, but occasionally one finds that at $x=0$:

$$y_0(0) = 0, \quad y_1(0) = 0, \quad y_2(0) = 0, \ldots ,$$

and one can show recursively that at every order of the expansion, $y_n(0) = 0$. In this problem, therefore, one has shown that $y(0; \epsilon)$ vanishes to all orders in the asymptotic expansion. It follows **either** that $y(0; \epsilon) = 0$, **or** that $y(0; \epsilon)$ is transcendentally small, as in (1). Thus the calculation has failed completely to answer the question of interest: whether or not $y(0; \epsilon)$ is positive. This failure persists even if the expansion is carried to all orders; in this problem it is simply too crude to answer the question.

At this point the reader may concede that conventional asymptotic methods cannot detect transcendentally small terms, but may wonder why anyone would care about such small effects. I now describe briefly some problems in which this issue has arisen, and in which the most fundamental questions about the problem hinge on

whether certain transcendentally small terms do or do not vanish. In these problems, questions like "Does a solution exist?" cannot be answered without going beyond all orders in the asymptotic expansion. This brief survey does not discuss how to solve these problems , but references to the recent literature are given below.

It should be mentioned that transcendentally small effects have been evaluated in particular linear problems over the last 30 years [2, 3, 4, 5, 6]. What distinguishes the recent flurry of activity is the realization that no linear structure is required, and the recent work has treated fully nonlinear problems. On the other hand, almost all of the recent work that has appeared in print to date is formal, with no assertion of rigor.

My first example is known as "viscous fingering of fluids", or as the "Saffman-Taylor paradox", after the famous paper by these authors [7]. Motivated by a problem of interest in petroleum engineering, Saffman and Taylor studied the slow motion of the interface between two fluids of different viscosities (such as oil and water). They found experimentally that when the fluids were confined to a narrow gap between two parallel walls (a Hele-Shaw cell), the less viscous fluid could be made to push steadily into the more viscous fluid in a single symmetric, uniformly growing "finger". They also found experimentally that the width of this finger far from the tip was extremely predictable ($\lambda = 1/2$ in their dimensionless notation) in the appropriate range of their experimental parameters.

In the same paper, the authors analyzed the (Navier-Stokes) equations of motion, seeking a steady-state solution to describe this steadily growing finger. They found that if they made certain plausible approximations, including neglecting surface tension, then they could find a continuous family of exact solutions of the resulting equations. This family was parameterized by λ, the finger width. The solution corresponding to $\lambda=1/2$ agreed well with their experimental data. However, the question of identifying the selection mechanism that picked out $\lambda=1/2$ in their experiments remained open.

The hypothesis that surface tension provided the selection mechanism was tested by McLean and Saffman [8], who developed an asymptotic expansion for the shape of the finger in powers of the (small) surface tension, starting at zeroth order with a Saffman-Taylor solution. These exact solutions were left-right symmetric, and McLean-Saffman intended to show that this symmetry was broken in the presence of any small, positive surface tension. To their surprise they found that the symmetry, and therefore the continuous family of solutions found in [7], persisted to all orders in their asymptotic

expansion; i.e., they found no analytical evidence that surface tension provided the selection mechanism observed experimentally. Adding to the confusion were numerical experiments by them [8] and by Vanden Broeck [9], using finite values for the surface tension, which seemed to indicate that surface tension **did** provide a selection mechanism.

The paradox was resolved in three papers published simultaneously [10, 11, 12]. Each of these papers showed that small surface tension does indeed break the symmetry and destroy the continuous family of solutions. However, the symmetry is broken by an exponentially small amount, so this breaking lies beyond all orders in the asymptotic expansion in [8], and it is not captured by that analysis. This is an example of a problem of physical interest in which the most basic question one can ask about the model, whether it even has a solution for arbitrary values of λ, cannot be answered without going beyond all orders in the asymptotic expansion.

A second example arises in the study of growing crystals in a supercooled melt of a pure substance. The verbal description of the problem is quite similar to that of the viscous fingers. Under appropriate conditions, a solid crystal is observed to grow into the liquid melt. The overall shape of the crystal is complicated and time-dependent, but the tip apparently grows with a nearly constant shape and at a nearly constant speed. From the speed and radius at the tip, one can form a dimensionless Peclet number, and this number is observed experimentally to depend only on the substance in question and on its temperature.

Important theoretical work was done by Ivanstov [13], who found that by neglecting surface tension, he could produce a continuous, one-parameter family of exact, steadily growing, two-dimensional crystal shapes, called "needle crystals". These were later generalized to three-dimensional needle crystals with ellipsoidal symmetry [14]. In both cases the free parameter was the Peclet number. Thus we have a second paradox: the theory predicts a needle crystal for every Peclet number, while the experiments show that one Peclet number is always selected. Again the question arises: What is the selection mechanism? In particular, does a small amount of surface tension break up the continuous family of exact solutions?

With surface tension included, the exact governing equations for this problem are quite complicated [15], and two simplified models were constructed to help to guide the analysis[16, 17]. In a paper which (embarrassingly) is still unpublished, Kruskal and

Segur[18] obtained the following results for one of these models (the "geometric model").

(a) Without surface tension, the model admits a spatially symmetric needle-crystal solution for every Peclet number.

(b) With or without surface tension, every needle-crystal solution in this model must be spatially symmetric.

(c) For small surface tension and for every Peclet number, the model admits an asymptotic expansion for a needle crystal that is symmetric to all orders in the expansion.

(d) For sufficiently small surface tension, **every** solution of the model is asymmetric. The amount of asymmetry is exponentially small, so it is missed at every order of the asymptotic expansion. Even so, it follows that the geometric model has **no** needle-crystal solutions for small surface tension, even though they exist to all orders in the asymptotic expansion.

(e) If one adds to the model a second parameter ("crystalline anisotropy"), then for each value of that parameter the model admits a needle crystal only for a discrete set of values of the Peclet number.

Some of these results were also obtained by others, using different means of analysis [19, 20]. From our standpoint, the main conclusion of all of these analyses is that the question of whether the geometric model even has a needle-crystal solution cannot be decided without going beyond all orders in the asymptotic expansion.

In more recent work [21] it has been claimed that a similar situation occurs in the full equations for needle crystals.

Now let us consider a third example in which asymptotics beyond all orders plays a decisive role. In one spatial dimension, a Klein-Gordon equation is a partial differential equation of the form:

$$u_{tt} - u_{xx} = g(u), \qquad g(0) = 0, \quad g'(0) > 0.$$

In the usual linear equation, $g(u) = mu$, where m represents "mass". Out of all possible nonlinear equations, two that have been studied extensively are the sine-Gordon equation with $g(u) = \sin u$, and the ϕ^4-model with $g(u) = 2u - 3u^2 + u^3$. The latter name comes from setting $u = \phi + 1$, after which the Lagrangean density for this model differs from that for the linear model by a term (ϕ^4).

A "breather" is defined to be a real-valued solution of a nonlinear Klein-Gordon equation that is localized in space and periodic in time, with a nontrivial period. If one thinks of the Klein-Gordon equation as a classical model of a field theory in one dimension, then any localized solution might represent an

elementary particle and a breather might represent a particle with an internal degree of freedom. Breathers have physical importance if they exist.

Breathers are known to exist for the sine-Gordon equation [22]:

$$u(x,t) = 4 \arctan \{ [\omega / \sqrt{1-\omega^2}] \operatorname{sech} \sqrt{1-\omega^2}\, x \cdot \sin \omega t \}$$

The question is whether they exist for any other Klein-Gordon equations. For small amplitudes (i.e., $u<<1$), the sine-Gordon and the ϕ^4-models approximate each other, so one expects the ϕ^4-model to admit at least an approximate breather solution for small amplitudes. It turns out that the ϕ^4-model admits an asymptotic expansion for a breather in a small amplitude limit, and that this expansion can be carried to all orders without developing any secular terms. This approximate breather was used by Dashen, Hasslacher and Neveu [23] in their quantization of ϕ^4. The question of whether the expansion represents a true breather solution was not addressed by them.

Segur and Kruskal [24] showed the ϕ^4-model admits **no** true breathers in this limit. The asymptotic expansion does represent true solutions to the equation, but none of them are both localized in space and periodic in time. Typically, these solutions radiate energy away, but at a rate that is exponentially small, and that is missed by the asymptotic expansion even when carried to all orders. Nevertheless, this exponentially small radiation rate is enough to carry away all of the energy eventually, so eventually the breather disappears.

A final example involves an ideal pendulum under the influence of small, periodic forcing [25]. It is known that a small, periodic forcing at moderate frequency of a pendulum typically destroys the integrability of the problem and introduces chaotic trajectories of the pendulum. The concrete evidence of nonintegrability (the Melnikov integral) vanishes to all orders in the high-frequency limit, but Holmes, Marsden and Scheurle [25] showed by evaluating exponentially small terms that the forcing destroys integrability in this limit as well.

Perhaps it is appropriate to conclude this survey with two general remarks. The first is that all of the problems mentioned here are pathological, in the sense that it is rare for an asymptotic expansion to yield **no** information at any order. The existence of these pathological examples does not mean that no asymptotic expansion should be trusted, but rather that they must be interpreted correctly. The second remark is that even though these

examples are pathological, they are not unphysical. Each came from a real-world problem of physical interest. Pathological problems do arise in physical contexts, but only occasionally.

ACKNOWLEDGEMENTS: The author gratefully acknowledges many valuable conversations with Martin Kruskal. The work was supported by the Army Research Office and by the Department of Energy.

REFERENCES

1. C. M. Bender & S. A. Orszag, Advanced Mathematical Methods for Scientists and Engineers, McGraw-Hill, New York, 1978.
2. V. L. Pokrovskii & I. M. Khalatnikov, Sov. Phys. JETP, **13**, 1207, 1961.
3. W. Wasow, SIAM J. Math. Anal., **5**, 673, 1974.
4. R. E. Meyer, SIAM J. App. Math, **29**, 481, 1975.
5. P. B. Chapman & J. J. Mahony, SIAM J. App. Math., **34**, 303, 1978.
6. W. Kath, "Asymptotic Evaluation of Short Wavelength Reflection Coefficients", preprint, 1985.
7. P. G. Saffman & G. I. Taylor, Proc. Roy. Soc. London A, **245**, 312, 1958.
8. J. W. McLean & P. G. Saffman, J. Fluid Mech., **102**, 455, 1981.
9. J. M. Vanden Broeck, Phys. Fluids, **26**, 2033, 1983.
10. B. I. Shraiman, Phys. Rev. Lett., **56**, 2028, 1986.
11. D. C. Hong & J. S. Langer, Phys. Rev. Lett., **56**, 2032, 1986.
12. R. Combescot, T. Dombre, V. Hakim, Y. Pomeau & A. Pumir, Phys. Rev. Lett., **56**, 2036, 1986; "Analytical Theory of the Saffman-Taylor Fingers", preprint, 1987.
13. G. P. Ivantsov, Dok. Akad. Nauk USSR, **58**, 567, 1947.
14. G. Horway & J. W. Kahn, Acta Metall., **9**, 695, 1965.
15. G. E. Nash & M.E. Glicksman, Acta Metall., **22**, 1283, 1974.
16. R. C. Brower, D. Kessler, J. Koplik & H. Levine, Phys. Rev. A, **29**, 1335, 1984.
17. E. Ben-Jacob, N. Goldenfeld, J. S. Langer & G. Schon, Phys. Rev. A, **29**, 330, 1984.
18. M. D. Kruskal & H. Segur, "Asymptotics Beyond All Orders in a Model of Dendrites", preprint, 1985.
19. R. F. Dashen, D. A. Kessler, H. Levine & R. Savit, Physica D, **21**, 376, 1986.
20. N. D. Kazarinoff & C. Lu, "A Model of the Early Stages of Growth of Dendritic Crystals", preprint, 1987.
21. M. Ben Amar & Y. Pomeau, "Theory of the Needle Crystal",

preprint, 1986.
22. M. J. Ablowitz & H. Segur, Solitons and the Inverse Scattering Transform, SIAM, Philadelphia, PA, 1981.
23. R. F. Dashen, B. Hasslacher & A. Neveu, Phys. Rev. D, **10**, 4114, 1974.
24. H. Segur & M. D. Kruskal, Phys. Rev. Lett., **58**, 747, 1987.
25. P. Holmes, J. Marsden & J Scheurle, "Exponentially Small Splitting of Separatrices", preprint, 1987

COMPUTATIONAL ISSUES IN GOAL PROGRAMMING RESOURCE ALLOCATION

Leon Medler

U.S. Army Troop Support Command
Belvoir RD&E Center
Fort Belvoir, VA 22060-5606

1. EXECUTIVE SUMMARY

The US Army Belvoir RD&E Center has been engaged in a program of research aimed at developing an efficient and fair methodology for ranking proposed RD&E programs. A linear goal programming model (LGPM) has been the foundation of this methodology. Problems with this approach have surfaced in two areas: (1) excessive computer time and (2) anomalous results. This paper reports on research conducted in order to redress these problems. Analysis of the existing model indicated that time expenditure in sensitivity excursions was the major culprit. Sensitivity analysis was being conducted by decrementing resource constraints and rerunning the LGPM "from scratch." Since virtually as much computation was involved in the reruns as in the initial run, significant computation expense was being incurred. Therefore, we modified the LGPM to start from the last solution (i.e., the last LGPM simplex tableau). This required the addition of a dual simplex algorithm to supplement the regular simplex algorithm, since the last solution becomes infeasible under certain resource constraint changes. For a 110 project prioritization problem involving ten resource levels the improved LGPM requires only 10-20% as many simplex iterations as did the old LGPM.

2. THEORY OF IMPLEMENTATION

The following discussion assumes some familiarity with the fundamentals of linear programming and linear goal programming, as might be found in Ignizio [1982]. We use the notation from this source, largely, and it tends to be standard in the linear programming literature. We will use [] to denote matrices and arrays, and $[]^T$ to denote the transpose of a matrix or vector. The following abbreviations will be used:

IBFS - initial basic feasible solution

LGPM - linear goal programming model

LP - linear program

RHS - right hand side.

In order to explain the general method for forcing a LGPM to begin at a prescribed solution, it will be easiest to explain the method for the simplest sort of LP. The generalization to the more complex preemptive LGPM is straightforward. Therefore, consider the following simple LP problem:

(1) Minimze $z = [c][x]^T$

s.t. $[A][x]^T = [b]^T$

$[x]^T >= [0]^T$,

Assume that this problem is such that an IBFS can be found by using simple slack variables. This corresponds to rewriting the problem as:

(2) Minimize $z = [c,0][x,s]^T$

s.t. $[A,I][x,s]^T = [b]^T$

$[x,s]^T >= [0]^T$.

In these last two formulations A is an m x n matrix, c and x are 1 x n matrices (vectors), s and b are both 1 x m matrices

(vectors), I is an m x m matrix, and 0 is either 1 x n or 1 x (n+m), as appropriate in context.

The LP problem is then solved using the simplex algorithm, which begins by operating on the following initial extended tableau, corresponding to using the slack variables as the initial basis:

c_j's->	c_1	c_2	...	c_n	0	0		0	
c_B	BV	x_1	x_2	...	x_n	s_1	s_2	s_m	x_B
0	s_1	$a_{1,1}$	$a_{1,2}$	...	$a_{1,n}$	1	0 ...	0	b_1
0	s_2	$a_{2,1}$	$a_{2,2}$	...	$a_{2,n}$	0	1 ...	0	b_2
.	.	.	.		.	.	.	.	.
.	.	.	.		.	.	.	.	.
.	.	.	.		.	.	.	.	.
0	s_m	$a_{m,1}$	$a_{m,2}$	...	$a_{m,n}$	0	0 ...	1	b_m
Indicators->	z_1-c_1	z_2-c_2	...	z_m-c_m	z_{n+1}	z_{n+2}	....	z_{n+m}	z

The simplex algorithm then proceeds, with the status at any particular iteration represented by an extended tableau of the general form:

c_j's->	c_1	c_2	...	c_n	0	0		0	
c_B	BV(xB) label	x_1	x_2	...	x_n	s_1	s_2	s_m	xB value
cB_1	xB_1	$y_{1,1}$	$y_{1,2}$	...	$y_{1,n}$	$y_{1,n+1}$	$y_{1,n+2}$ ...	$y_{1,n+m}$	xB_1
cB_2	xB_2	$y_{2,1}$	$y_{2,2}$	...	$y_{2,n}$	$y_{2,n+1}$	$y_{2,n+2}$ ...	$y_{2,n+m}$	xB_2
.	.	.	.		.	.	.	.	.
.	.	.	.		.	.	.	.	.
.	.	.	.		.	.	.	.	.
cB_m	xB_m	$y_{m,1}$	$y_{m,2}$	...	$y_{m,n}$	$y_{m,n+1}$	$y_{m,n+2}$ ...	$y_{m,n+m}$	xB_m
Indicators->	z_1-c_1	z_2-c_2	...	z_m-c_m	z_{n+1}	z_{n+2}	....	z_{n+m}	z

The reader will recall that at any point in time the $y_{i,j}$'s and xB_i's may be expressed in terms of the m x m matrix [B] defined as

$$[B] = [\, a(xB_1) \quad a(xB_2) \quad \ldots\ldots\ldots \quad a(xB_m) \,]$$

where a(xBj) denotes the column in the initial tableau corresponding to the basic variable now labeled xBj. The important relationships actually involve the inverse of [B]:

$$[xB]^T = [B]^{-1}[b]^T$$

$$[y_j]^T = [B]^{-1}[a_j]^T$$

where $[b]^T$ is the initial right hand side (RHS).

Once these are computed, the indicator row elements may be computed as

$$z = [cB][xB]^T$$

and

$$z_j = [cB][y_j]^T.$$

Thus, given the basis (xB_1, xB_2, ... xB_m), together with [A], [b], and [c] from the original problem, we can identify [B], find its inverse, and then calculate all the elements of the tableau needed to start the simplex procedure from that point.

3. CONCLUSIONS

Several conclusions can now be drawn.

(1) Note that to construct B we only need the identification of the basis variables, not their values.

(2) If any series of Gauss-Jordan operations leads from the initial tableau to the tableau with basis x_B, then in the final tableau the important matrix $[B]^{-1}$ actually appears, viz., as the array occupying the columns under the slack variables:

$$[B]^{-1} = \begin{vmatrix} y_{1,n+1} & \dots & y_{1,n+m} \\ . & \dots & . \\ y_{m,n+1} & \dots & y_{m,n+m} \end{vmatrix}$$

This observation is the key to efficient sensitivity analysis, and is called "A Fundamental Insight" by Hillier and Lieberman [1980].

(3) Recall from elementary linear algebra that the inverse of [B] may be calculated by forming the rectangular matrix [B|I] and conducting appropriate Gauss-Jordan operations until the identity matrix appears on the left: [I|D]. At that point the square matrix on the right is $D = [B]^{-1}$.

(4) The simplex procedure is simply a sequence of Gauss-Jordan operations, guided by rules for selecting the pivot point.

(5) The $y_{i,j}$'s and xB_i's computed using $[B]^{-1}$ in the formulas above are identical to those that would be obtained through a series of simplex operations yielding the same basis vector.

Taken together, the above suggests a framework for a simple and efficient technique to construct the simplex tableau appropriate to the desired initial basis:

Employ the basic simplex algorithm, but override the normal selection of pivot point (entering and exiting variables), instead forcing the desired variables into the basis and prohibiting their exit.

4. REFERENCES

(1) [Ignizio, 1982] Ignizio, James P., Linear Programming in Single and Multiple-Objective Systems, Prentice-Hall, Englewood Cliffs, N.J. 1982.

(2) [Hillier and Lieberman, 1980] Hillier, Frederick S., and Gerald J. Lieberman, Introduction to Operations Research, Holden-Day, Inc., San Francisco, 1980 (Third Edition).

DESIGN OF A FEELING-THINKING MACHINE

Ray Scanlon and Mark Johnson
Benet Laboratories
US Army Armament Research, Development, and Engineering Center
Watervliet, NY 12189-4050

ABSTRACT. A feeling-thinking machine has been designed using the mammalian brain as a model and current psychobiology concepts as a guide. The machine has been successfully run as a computer simulation. It mimics a primitive organism with eight functional brain centers. They are the recticular ascending substance (RAS), the amygdala, the cingulate gyrus, and medial forebrain bundle, the hippocampus, thalamus, hypothalamus, and the neocortex.

I. INTRODUCTION. Machine intelligence for autonomous systems must be capable of learning and especially thinking, if we are to go beyond the 'islands of autonomy' presently envisioned for teleoperated and remotely piloted vehicles. One approach is to use the mammalian brain as a model and investigate the possibility of duplicating its functions in electronic circuitry. This extension of neural network design is called non-living intelligence (NLI).

The brain consists of approximately 10**12 neurons intricately interconnected. Only a small part of this circuitry has been unravelled. The NLI effort at Benet Labs does not describe how the brain works, but involves electronic and computational experiments that provide insight into how the brain might work. We pursue NLI through the design and construction of feeling-thinking machines. Feeling is essential because without motivation, there is nothing. The machine must want to do things. In doing things, it will learn; and having learned, it will think. This report does not describe machines that "exhibit intelligent behavior"; but rather machines that feel, want, and think. The distant goal is to create a machine that thinks and acts like a man. This report discusses the first of a series of feeling-thinking machine designs.

II. THE MODEL. Our approach to designing this machine is to simulate a primitive organism which must survive within a contrived universe. We have given it the name "Pacrat". Pacrat's brain has eight brain centers. The electrical activity of these neural centers is not modeled, only the functional relationships. From these interactions arises a sophisticated structure which rests upon the anatomy of Pacrat's brain. The neural centers modeled are: the reticular ascending substance (RAS), the thalamus, the hypothalamus, the amygdala, the cingulate gyrus, the medial forebrain bundle, the hippocampus, and the isocortex. (Gregory, 1975, pp. 688-689)

Individual neural response is not simulated, only the activity of assemblages of neurons called codons. A codon is the result or record of an experience. It exists as the altered synapses between the neurons which constitute the assemblage. (Palm, 1982, 1986)

Pacrat, in diagrammatic form, is shown in figure 1. It shows that he has been provided with the ability to get about in his universe through four motor neurons. These are driven by the motor area of the isocortex as a final result of sensory input, channeled to the isocortex under the control of the thalamus, and filtered through the isocortex under the influence of the prevailing emotion.

Hunger is the level of neural activity in an area of the hypothalamus which we will call the hunger center. (Kissin, 1986, p. 15) The model assumes there are sensory neurons lining the stomach wall that respond to expansions and contractions of the stomach. These determine the level of activity of the hunger center. As the stomach empties, the hunger center becomes more active: as the stomach fills, it becomes less active.

Anger is also a level of neural activity in an area of the hypothalamus, but in this case the cause is the activation of certain codons in the isocortex as mediated through the amygdala. When this area is active, Pacrat experiences some level of anger or frustration. The activity in the amygdala is quickly inhibited by eating. (Flynn, et al., 1970) (LeDoux, 1986, p. 342)

Fear is the level of activity of the cingulate gyrus. This activity is, subjectively, unease escalating to terror. In Pacrat, it is assumed that sensory neurons excite the cingulate gyrus whenever his back is uncovered. This is agoraphobia, the fear of open places.

Curiosity is the level of activity of the hippocampus. It is set off by the activation of a codon in the isocortex which has not previously been excited. The continual excitation of "old" codons will allow this activity to fade away. Pacrat's hippocampus has efferents on his motor area with the result that "newness" leads to exploratory rather than hunger or fear driven activity.

All sensory input (other than olfactory) is gated through the thalamus to the isocortex. Thus the thalamus can relay or block this input. It can also inhibit the motor output that would normally result from activity in the isocortex. The thalamus does this in a rhythmic manner when the reticular ascending substance (RAS) is stimulated. The RAS is excited whenever the hypothalamus or the cingulate gyrus is active.

The thalamus extends this period of choking off sensory input when it receives impulses from a codon through synapses which have been facilitated in the past by the reward-punishment mechanism. This blocking of sensory input and an associated inhibition of motor output is the function of the thalamic reticular complex. On the other hand, if a codon is activated which has a facilitated synapse on the "goal" area of the thalamus (cf. akinetic mutism, Girvin, 1975), sensory input is gated to the isocortex and the motor output is enabled.

The normal activity of the isocortex is association. During each "moment" there is an active codon which has efferents on the motor output system. If this system is not inhibited, motor output will follow. This codon fades out as its store of strategic molecules becomes temporarily depleted. As it fades out another codon starts up and the next "moment" begins. The new codon is determined by the sensory input (if not blocked), the previously excited codon, and the current dominant emotion.

A reward-punishment mechanism is started up by the medial forebrain bundle whenever activity in the hypothalamus or cingulate gyrus is reduced. The role of this mechanism is to facilitate all recently fired synapses. (LeDoux, 1986)

III. THE IMPLEMENTATION. Pacrat exists in a contrived universe: a very simple universe which is seen as partitioned by a rectangular grid (figure 2). At each location in the grid one of Pacrat's sensory neurons, unique to that location, becomes active. This gives him a location sense. At genesis he does not know where one location is relative to another, but he does know that he is where he is. He has also been given the ability to sense his own trail, and has a general adversion to going where he has recently been. Again, the individual activity of the sensory neurons is not simulated, only the relationship with other active neurons. His burrow (or starting point) is always at row 11, column 1. This is indicated by shading that cell. Pacrat's current location is given by highlighting the cell he is in. (Walter, 1950, 1951)

One codon is active at any time and this represents a 'moment' in Pacrat's life. This codon is excited by current sensory input to the isocortex plus the previously excited codon and the prevailing emotion. The inputs are the location sense, which is gated through the thalamus (figure 1), smell, and the axonal bundles from the hypothalamus, and cingulate gyrus. A codon in the simulation is simply a vector of scalars representing the current sensory input (if any), normalized synaptic weights to the four motor neurons, associative connections to other codons, dominant emotion, and synaptic weights to the amygdala, thalamus, and hippocampus. (Mishkin, et al., 1987)

Pacrat's motivation is hunger and fear. When awake, Pacrat is forced to move by one or the other, or else he just goes to sleep as the RAS quiets down. (Kissin, 1986, Chap 2). Initially this drive is hunger. In the simulation, the distension of the stomach is represented by a scalar. This number continually decreases unless Pacrat is at a food spot and is eating. When this number is low enough, the hypothalamus responds (again a scalar) and the RAS is excited. Pacrat wakes up. He is forced to explore his universe for food to satisfy hunger. Food is placed randomly in one of three locations. The three potential food spots are highlighted on the right side of the grid, with food located in one of the cells. When he reaches a food spot, he eats, his stomach fills up, and the activity in the hypothalamus is significantly reduced. This is simulated by simply increasing the number corresponding to distension of the stomach which is sensed by the neurons lining the stomach wall. These neurons have efferents to the hypothalamus.

Pacrat can move north, south, east, and west within the boundaries of his universe. Motor neurons drive Pacrat in one of these directions one cell at a time. Each active codon in the isocortex has efferents on each of these motor neurons and the relative effectiveness of these efferents determines the direction of travel. At the outset, i.e. trial 1, there is no preferred direction of movement. The synaptic weights from any given codon to the motor neurons are identical. Pacrat moves about his universe randomly until food is found. When it is, a reward mechanism is activated through the medial forebrain bundle which facilates all recently fired synapses. This is learning and will generate a preferred direction of movement when similar codons are active in the future. The vectors representing codons are changed so that elements corresponding to synaptic connections between simultaneously active neurons are increased. Facilitation is proportionally lower for codons active earlier in time.

After hunger is satiated and the level of activity of the hypothalamus reduced, fear is no longer masked by hunger. Fear keeps the activity of the RAS high. An active cingulate gyrus drives Pacrat back to his burrow. Again, if this is the first trial, there is no preferred direction, but the codons which are activated are those associated with fear rather than hunger. The neurons in the cingulate gyrus, not the hypothalamus, excite neurons in the isocortex. When his burrow is reached, Pacrat's back is covered. The activity of the cingulate gyrus is abruptly decreased. The reward mechanism is again activated through the medial forebrain bundle and recently fired synapses are facilitated. This will generate biased movement in the future if these codons are active. Henceforth, at any cell in the grid, he will tend to go in a direction depending on which neurons are active in the brain centers. An active hypothalamus may move him east, an active cingulate gyrus with the same sensory input may drive him north.

During epigenisis, Pacrat learns to survive. Randomness forces Pacrat out of obsessive behavior patterns. Although a reward mechanism may increase synaptic strength between a codon and a given motor neuron, there is always a chance Pacrat will move in a different direction. A built in random element raises the level of activity of motor neurons with a lower synaptic weight to the currently active codon. An active hippocampus increases the effect of this random element. If this were not present, Pacrat would not survive. Once food were found, he would follow the same path again and again. However, as the synaptic strength between a codon and motor neuron is increased, it becomes more and more difficult for Pacrat to alter his behavior. He will continue searching for food in locations where it does not exist. To resolve this we have given Pacrat an amygdala. An active amygdala mediates anger. Excited neurons in the amygdala generate unique active codons in the isocortex. Figure 1 shows the role of the amygdala in Pacrat. When the reward mechanism is active, all recently fired synapses are facilitated through the medial forebrain bundle. These include synapses from the active codon in the isortex to the amygdala. Therefore, if this same codon is excited in future trials, the amygdala also becomes highly active. This high level of activity excites a region in the hypothalamus associated with anger or frustration. Unless there is a concurrent good experience, such as eating, which will inhibit the amygdala; Pacrat will become angry. In other words, he gets mad when food is not where it is supposed to be. This anger quickly drives Pacrat out of the vicinity of a food spot by exciting different codons in the isocortex. These codons do not have synaptic weights to the motor area that favor any given direction. He is effectively 'bounced' randomly to neighboring locations in the grid. Without the amygdala, Pacrat would keep looking for food in the same spot almost indefinitely. The level of activity in the hypothalamus is far greater than that of the hippocampus. When he is starving, he doesn't get bored.

The effect of the rhythmic action of the thalamus is that a moment (active codon n) that generates motor output is followed by several moments (active codons n+1, n+2,...) with motor output inhibited. This is the first of three forms that thinking takes. The sensory input is temporarily blocked, motor output inhibited, and associated codons in the isocortex are turned on. This form of thinking is implemented in Pacrat as he effectively evaluates the consequences of his last move.

A second form of thinking comes about when an excited thalamus results in an extended period of blocking of sensory input. Again, the normal state in the isocortex is association so codons continue to fade in and out. If the chain of associating codons reaches a codon with inhibitory efferents on the thalamus, activity of neurons in the reticular complex is reduced. The blocking cycle of sensory input is reduced to a minimum and Pacrat proceeds to move with intent. This form of thinking is recognition. It is initiated when Pacrat moves to an area of particular interest to him on the grid. This a location where synapses from the isocortex to the thalamus have been facilitated from previous rewards.

A third form of thinking comes about when the extended period of blocked sensory input and inhibited motor output results in slightly different associated codon chains. This can occur because of the inherent randomness of neural actions. If one of these chains results in activating a codon quicker than recent paths have done, the neurons of this codon are in a different state of molecular depletion. It has had less time to recover from the last activation. It comes on with a burble which is transmitted to the reward system, and recently fired synapses are facilitated. This is insight and is the basic mechanism of rational thought. Pacrat has demonstrated this by "thinking" of more efficient paths to food.

IV. TRIAL RUN. Figure 3 shows four static displays of a typical trial run of the simulation. Figures 3a - 3d are snapshots in trial 702. The activity of the brain centers is given by a bar chart on the left side of the display. The larger the bar, the more active that area of the brain. Even number trials (i.e., 702) display the effect of a particular neural center (i.e., hunger, anger) while odd number trials (i.e., 703) give the name of the center. The number of steps indicate the sequence of each snapshot in the trial.

In figure 3a Pacrat has just left his burrow, the starting point. The activity of the hypothalamus was high enough to activate the RAS and wake him up. Figures 3a-c show Pacrat driven by hunger. Through epigenisis, which is his previous 701 trials, he has learned. Figure 3b shows Pacrat with his motor output inhibited by the thalamus. This is shown by freezing him at his current location and dynamically displaying his codon association chain. The reward cell for this trial is in the middle of the three possible food locations (row 11, column 18). From past experience, food has been known to be located in the last reward cell (row 19, column 18). Pacrats codon chain eventually associates to this location and facilitated synapses from the isocortex stop the thalamus from blocking sensory input and inhibiting the motor area. His motor output is no longer inhibited. Figure 3c shows the active amygdala when food is not found where it was expected.

His frustration forces him out of the vicinity of the empty food cell and eventually he locates the food. Figure 3d shows Pacrat moving with intent back to his burrow, driven by fear. The activity in the thalamus (thinking) indicates it has been inhibited from blocking sensory input and from inhibiting motor output. This is a result of recognition resulting in strong inhibitory input from the isocortex.

V. IMPLEMENTING PACRAT AS A NEURAL NET. A simplified version of Pacrat has been implemented using only simulated neurons, formal definitions of neural activity, and synaptic facilitation. Neural activity is modeled using PID (proportional-integral-differential) control. The governing equations for cell activity and synaptic facilitation are given in figure 4. Synaptic facilitation has both a Hebbian (associative) and a non-Hebbian (reward-punishment) component. A selectable resting frequency and codon saturation frequency were included to help with governance of the network.

This simulation is called Mouse. Figure 5 gives a static display at one point in the simulation. The shaded circles on the left represent neurons. Mouse, like Pacrat, lives in a bounded universe. This universe is the ten by ten grid on the right. At each cell in the grid, a single sensory neuron becomes active. The color of the circles reflect the activity. The color changes gradually from blue to red to white as the activity increases. Since color is not reproduced in this report, the active neurons are circled. Each sensory neuron has excitatory efferents on each of four motor neurons. These motor neurons are labelled N (north), S (south), E (east), and W (west). When the activity of one of these motor neurons exceeds a preset threshhold, Mouse moves one cell in that direction (within the boundaries) and a different sensory neuron becomes excited. As in Pacrat, there are three potential reward cells. At the beginning of each trial, food is placed randomly in one of these cells. These cells are the three shaded cells in column 9 as shown in figure 5. The dark cell gives the location of the reward cell for that trial. When Mouse reaches a cell where food is located, a reward mechanism is activated and recently fired synapses are facilitated. The normalized synaptic weights from the sensory neurons to the motor neurons are shown by the arrows in the grid. There is always an element of randomness associated with each move, but the larger the arrow the more likely Mouse will move in that direction. Initially, (i.e., trial one) Mouse has no preferred direction of movement and the arrows have zero length and direction. Figure 5 gives the normalized weights after 1000 trials. Mouse always starts at row 6, column 1 and his current location in the grid is highlighted. In order to avoid obsessive, compulsive behavior Mouse has been given a sense of smell. He is designed to avoid his own trail. This is accomplished via four sensory neurons with inhibitory efferents on the motor neurons. These are labeled 1/N, 1/E, 1/S, and 1/W indicating their effect on that direction of travel. The necessity for these is evident if one imagines four arrows in the grid forming a loop.

Figure 5 shows Mouse after a single move. He has just moved north so there is a high level of neural activity in the neuron inhibiting motor neuron S (south). Since this is the 1000th trial, Mouse has a preferred direction of movement. At this location, as shown by the arrow, it is east. The large synaptic weighting from the currently active sensory neuron to motor neuron E (east) is raising the activity of this motor neuron more than the others. It is therefore likely that Mouse will move east.

VI. CONCLUSIONS. A brassboarded feeling-thinking machine is possible. We believe it is not practical at the moment to consider casting everything in silicon, therefore the neural network section of the machine will be emulated in a highly parallel computer ensemble.

VII. ONGOING EFFORTS. The Pacrat simulation is being completely rewritten so that the neurons are explicitly modeled. This is preparatory to moving the simulation to a transputer network running under an Occam harness.

REFERENCES

Flynn JP, Vanegas H, Foote W, and Edwards S (1970): Neural mechanisms involved in a cat's attack on a rat, in Whalen RW, Thompson RF, Verzeano M, and Weinberger NM (eds): The Neural Control of Behavior. New York, Academic Press, pp. 135-173.

Girvin JP (1975): Clinical correlates of hypothalamic and limbic system function, in Mogenson GJ and Calaresu FR (eds): Neural Integration of Physiological Mechanisms and Behavior. Toronto, University of Toronto Press, pp. 412-434.

Gregory RL (1975): Do we need cognitive concepts?, in Gazzaniga MS and Blakemore C (eds): Handbook of Psychobiology. New York, Academic Press pp. 607-628.

Kissin B (1986): Conscious and Unconscious Programs in the Brain. New York, Plenum Publishing Corp.

LeDoux JE (1986): The neurobiology of emotion, in LeDoux JE and Hirst W (eds): Mind and Brain. Cambridge, Cambridge University Press, pp. 301-358.

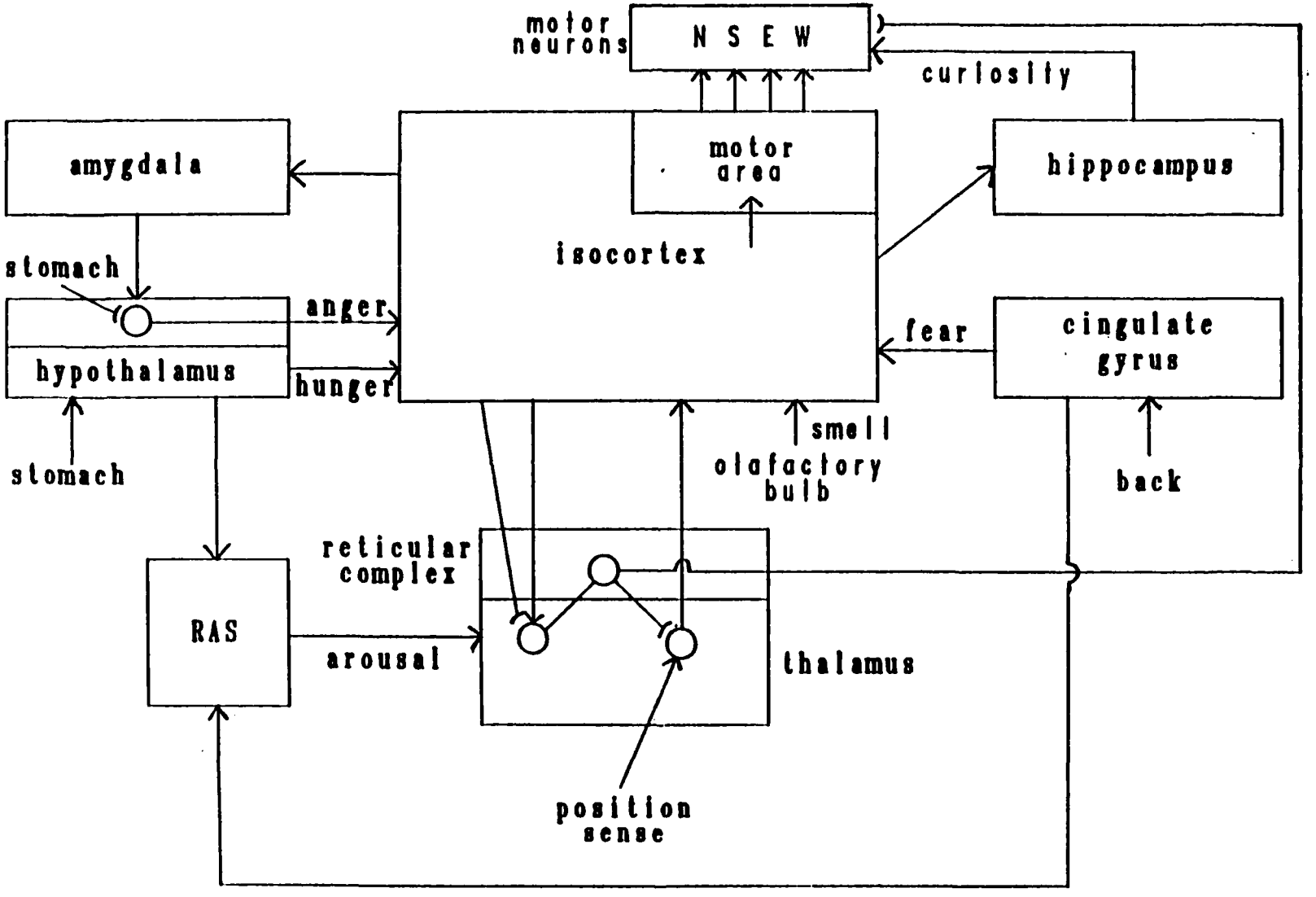
Mishkin M and Appenzeller T (1987): The anatomy of memory. Scientific American 256:N6.

Palm G (1982): Neural Assemblies. Berlin, Springer-Verlag.

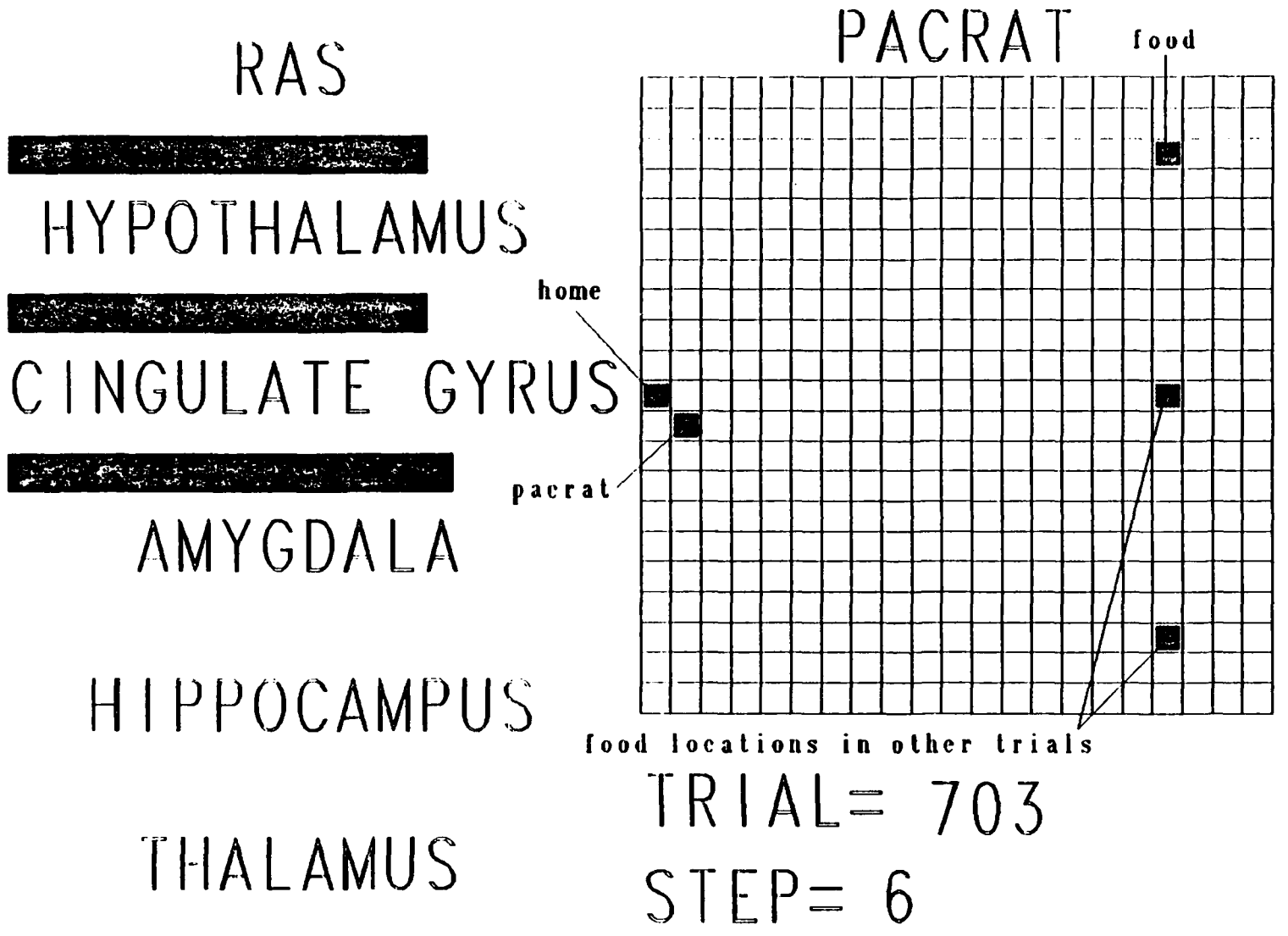
Palm G (1986): Associative networks and cell assemblies, in Palm G and Aertsen A (eds): Brain Theory. Berlin, Springer-Verlag.

Walter WG (1950): An imitation of life. Scientific American 182:N5.

Walter WG (1951): A machine that learns. Scientific American 185:N2.

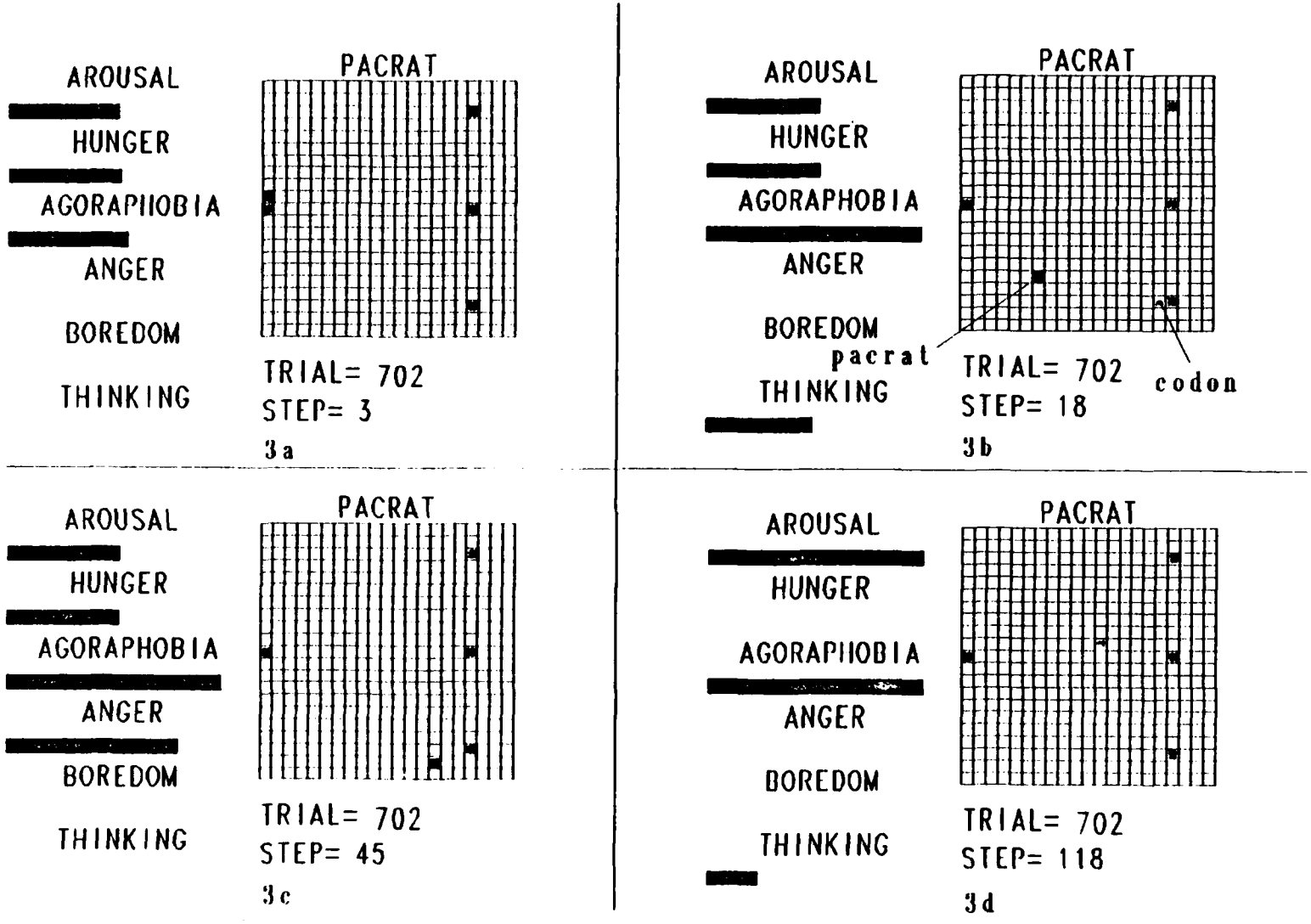


Anatomy of Pacrat
Figure 1



Pacrat

Figure 2



Snapshots of Dynamic Graphics Display
Figure 3

CELL ACTIVITY (PID)

i = postsynaptic , j = presynaptic

$$X_i = [AX_i + B \int_{-t}^{0} (K_i - X_i)dt + \dot{C}X + D*XI + E*R]^+$$

$$XI = XI + F*(\sum X_j W_{ij} - X_i)$$

X = cell activity

K_i = resting frequency

W_{ij} = synaptic weight

A,B,C,D,E,F = empirical constants

R = rectangular distribution on (0.0,1.0)

SYNAPTIC FACILITATION

$$W_{ij} = -AW_{ij} + B*H(X_i,X_j) + \Gamma(R,P)*\alpha* \int_{-t}^{0} H(X_i,X_j)dt$$

$$\alpha = C\alpha + D*H(X_i,X_j)$$

$$\Gamma(R,P) = E*R + F*P$$

$$H(X_i,X_j) = [X_i-K_i]^+[X_j-K_j]$$

A,B,C,D,E,F = empirical constants

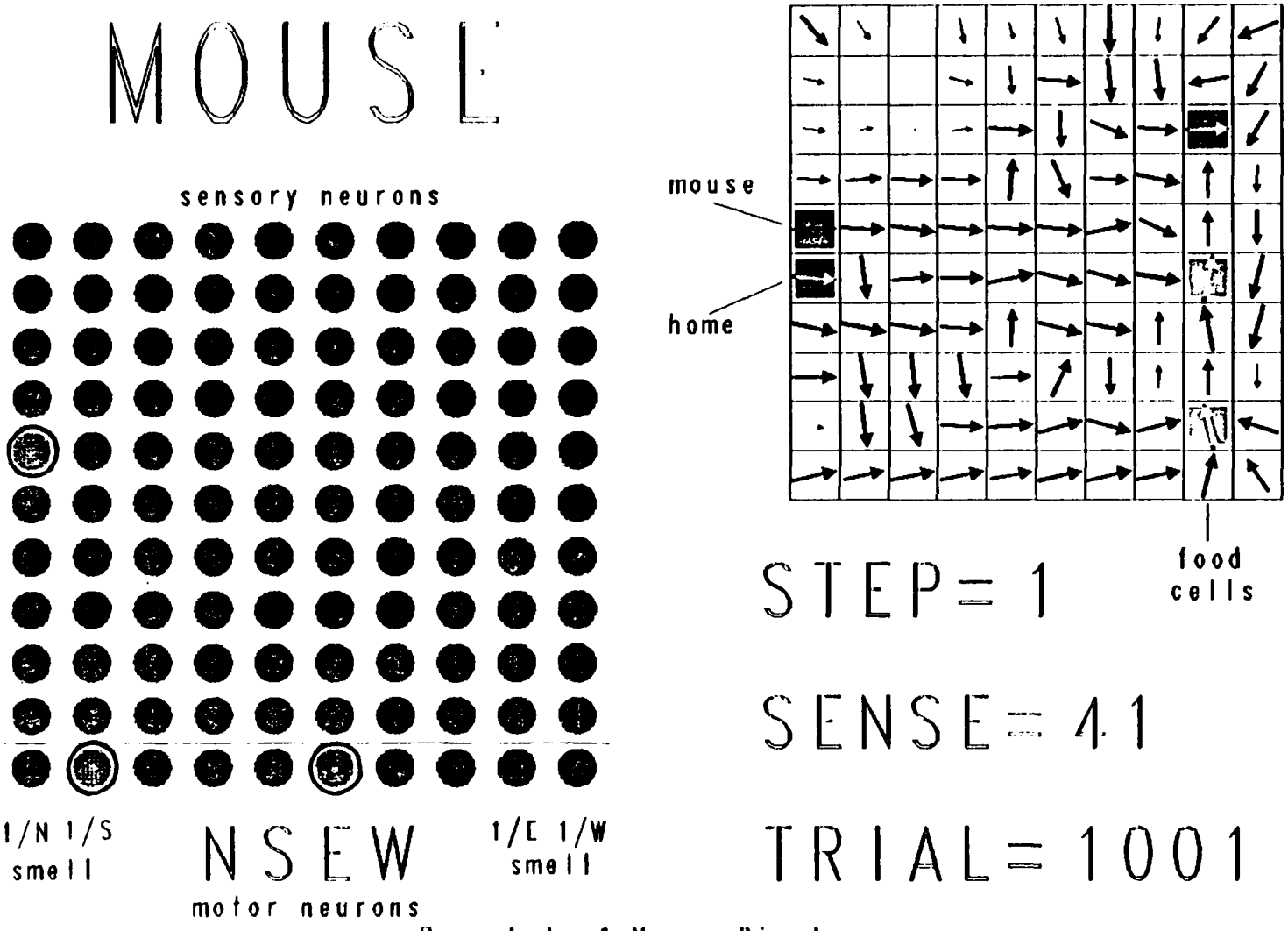
K_i,K_j = resting frequencies

R = instantaneous reward level

P = instantaneous punishment level

$R = P = 0$ or $E = F = 0$ => Hebbian

Figure 4. Cell Activity and Synaptic Facilitation.



Snapshot of Mouse Display
Figure 5

POLYNOMINAL DEFINITION OF DISCRETE FIELD POINT OF MAP OF DIFFUSION EQUATION

William F. Donovan
Mechanics and Structures Branch
Interior Ballistics Division
Ballistics Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT

The one dimensional diffusion equation, $\frac{\partial T}{\partial t} = a \frac{\partial^2 T}{\partial \chi^2}$, is given finite difference expression, transformed to geometric and then algebraic context, and then by differencing, recomposed into general proposition. Discrete terms of the algebraic transpositon take the terminating polynominal form

$$T\ (N,P) = \frac{\phi}{2^h \quad 2^j}\ (A - Bm + Cm^2 - \cdots \mp m^k)$$

where the coefficients A,B,C etc., which turn out to be rational expressions, are analyzed by differencing methods. The systematic reduction to a base-line source reveals a general behavior pattern re-expressed in compressed tables, from which the algebraic form of any (N,P) term can be recomposed.

INTRODUCTION

The diffusion equation of Physics has been used to analyze unsteady heat transfer, boundary layer velocity distribution, long line electrical voltage fluctuation and salt-solute penetration. The general mathematical expression is $\partial T/\partial t = a\, \partial^2 T/\partial \chi^2$ where particular physical constraints determine the context. There are two approaches to the problem statement and solution; the one most widely used being transformation, and the other, following finite differences, employes variations of summing averages of term values established by unique methods. This report considers an averaging type solution in algebraic format. The final result consists of a series of discrete polynomials with rational coefficients which describe the dependent variable state at each time-distance coordinate in the manner of the non-reflecting Schmidt plot.

PROCEDURE

Essentially, the differential equation is given finite difference expression which is transposed first to geometric and then to polynomial algebraic form. The polynomials, representing discrete solutions to the differential equation, are analyzed by differencing techniques whereby the numerical coefficients of common diagonal terms are found to be expressible in a generalized matrix.

From the one dimensional partial differential equation

$$\frac{\partial T}{\partial t} = a \frac{\partial^2 T}{\partial \chi^2} \tag{1}$$

where T is the dependent variable

t is the independent variable

χ is the independent variable

and a is a constant.

Heat transfer language makes

T = temperature,

t = time,

χ = distance,

and a = diffusivity.

Application of the finite difference procedure gives

$$\frac{1}{a} \frac{\Delta_t T}{\Delta t} = \frac{\Delta_\chi^2 T}{\Delta \chi^2} \tag{2}$$

with Δt the finite difference in time,

$\Delta\chi$ the finite difference in distance,

$\Delta_t T$ the time variable effecting a change in T,

and $\Delta_\chi T$ the distance variable effecting a change in T.

By expansion the equation becomes

$$\frac{1}{a}\left(\frac{T_n^{(t+1)} - T_n^t}{\Delta t}\right) = \frac{T_{(n-1)}^t - 2T_n^t + T_{(n+1)}^t}{\Delta\chi^2}$$

where subscript n refers to the χ increments and superscript t refers to the t increments. Schmidt[1] developed the graphical form shown on Figure 1 with the stepwise linear temperature gradients across adjacent layers of material. Since the change in internal energy within a layer of material over a finite time is the difference between the heat flow in and heat flow out, the corresponding temperature increment becomes a function of the ratio $\frac{\Delta\chi^2}{2a\Delta t}$ and it is convenient to select this ratio as unity.

Whereby:

$$\Delta t = \frac{\Delta\chi^2}{2a} \quad . \tag{4}$$

Also, a geometric simplification results from defining the graphical proportions as

$$m = \frac{\Delta\chi}{\Delta\chi + \Delta\chi_o} \quad . \tag{5}$$

Table 1 shows the discrete algebraic expressions for the time-temperature - distance intersections of Figure 1. Along the diagonals of Table 1 a matched power polynomial appears and the coordinate expression for T in time and distance takes the general form

$$T(N,P) = \frac{\phi}{2^h\ 2^j}\left(A - Bm + Cm^2 - \cdots \mp m^k\right) \tag{6}$$

where N is a distance index

P is a time index

---

[1] See George P. Sutton, Rocket Propulsion Elements, John Wiley & Sons, New York 1956.

$h = \frac{(P + N) - 2 - |\sin (P + N)\, \pi/2|}{2}$, the external denominator exponent,

k is the terminal exponent of m

j = (k - term exponent of m), the individual term denominator exponent,

and $\phi = m\,(To - T_1 o)$ with A,B, C .... numerical coefficients of the interior terms of the equation.

To establish T (N, P) for any φ and m (which include the physical constraints) it is necessary to establish the precise values of the coefficients A, B, C ....; and this is the object of the current investigation.

Any full expression, T (N,P), can be developed from a Gregory-Newton formulation[2] of the separate terms A, B, C, for the bounded diffusion equation as shown in Figure 2. Some interesting progressions do result but an alternate and more geometric presentation is available from the direct finite difference tables.

Table 2 is also extracted, by difference equation procedure, from Table 1 and generates the coefficients of the constant term, A, for the coordinate expression of distance and time. The starting point is within the heavy box of column 7. These numbers, 17548, 25147, 35401, 49024 and 66868, were found by direct calculation using Figure 1 and Table 1 and are the constant numerator terms only. Table 3 lists the complete polynomial expressions for these coordinates. By the usual differencing, the 7th through the 1st columns are established. It is then possible to work vertically using the regression in column 2, back to zero; and then to complete the elements of columns 1 through 6. Noting the resulting bias progression at the tops of these columns, the next step is continue diagonally (I, II, III) to column 8 and, using column 7 as a summation, verify the vertical sequence of column 8. Columns 9, 10, 11, etc., are generated similarly.

Within the individual frames containing the coefficients is a paranthecized pair of numbers which indicate the distance and time coordinate. These indices run diagonally upwards at constant distance and bi-sequentially as time. Tables 4 through 9 are formed by the same procedure and extend arbitrarily to the 6th power of m. However, a different sequence appears along diagonals I, II and III according to the power of m. Table 10 summarizes this behavior and reveals yet another correlation, shown mainly by column 4, from which the adjacent columns can be constructed ad-infinitum. The final coincidence occurs from a re-inspection of Tables 2, 4-9 where the digital vertical counting column (1, 2, 3, 4, 5 ....) conjoins the power of m and the time sequence of the first distance diagonal (1,3), (1,5), (1,7), etc., by an interval of 3 in the counter according to Table 11.

---

(2) M.R. Spiegel, Theory and Problems of Finite Differences and Finite Difference Equations, Schaum's Outline Series in Mathematics, McGraw-Hill Book Company, New York, etc., 1971, p.p. 36-44.

RESULTS

To write any term defining the dependent variable in time and distance, Tables 2 through 11 are used to form the numerical coefficients in Equation (6)

$$T(N,P) = \frac{\phi}{2^h 2^j}\left[A - Bm + Cm^2 - + \quad \pm m^k\right]$$

For T (1,15) for instance

$N = 1$

$P = 15$

$$k = \frac{P - N - \left|\sin (P - N)\,\pi/2\right|}{2} = \frac{(15 - 1) - \left|\sin 7\,\pi\right|}{2} = 7$$

$$h = \frac{P + N - 2\left|\sin (P + N)\,\pi/2\right|}{2} = \frac{(15 + 1) - 2 - \left|\sin 8\,\pi\right|}{2} = 7$$

$j =$ (k - term exponent of m) = (7 - term exponent of m) .

The finite difference tables for the (1,15), j variables are then reconstructed (Tables 2, 4 -9) using Tables 10 and 11 and the respective numerators determined.

Whereby

Term	Numerator Value	Exponent of "m"	"j"
A	51480	0	7
B	39796	1	6
C	20264	2	5
D	7050	3	4
E	1672	4	3
F	260	5	2
G	24	6	1
H	1	7	0

and

$$T(1,15) = \frac{\phi}{128}\left[\frac{51480}{128} - \frac{39796m}{64} + \frac{20264m^2}{32} - \frac{7050m^3}{16} + \frac{1672m^4}{8} - \frac{260m^5}{4} + \frac{24m^6}{2} - m^7\right]$$

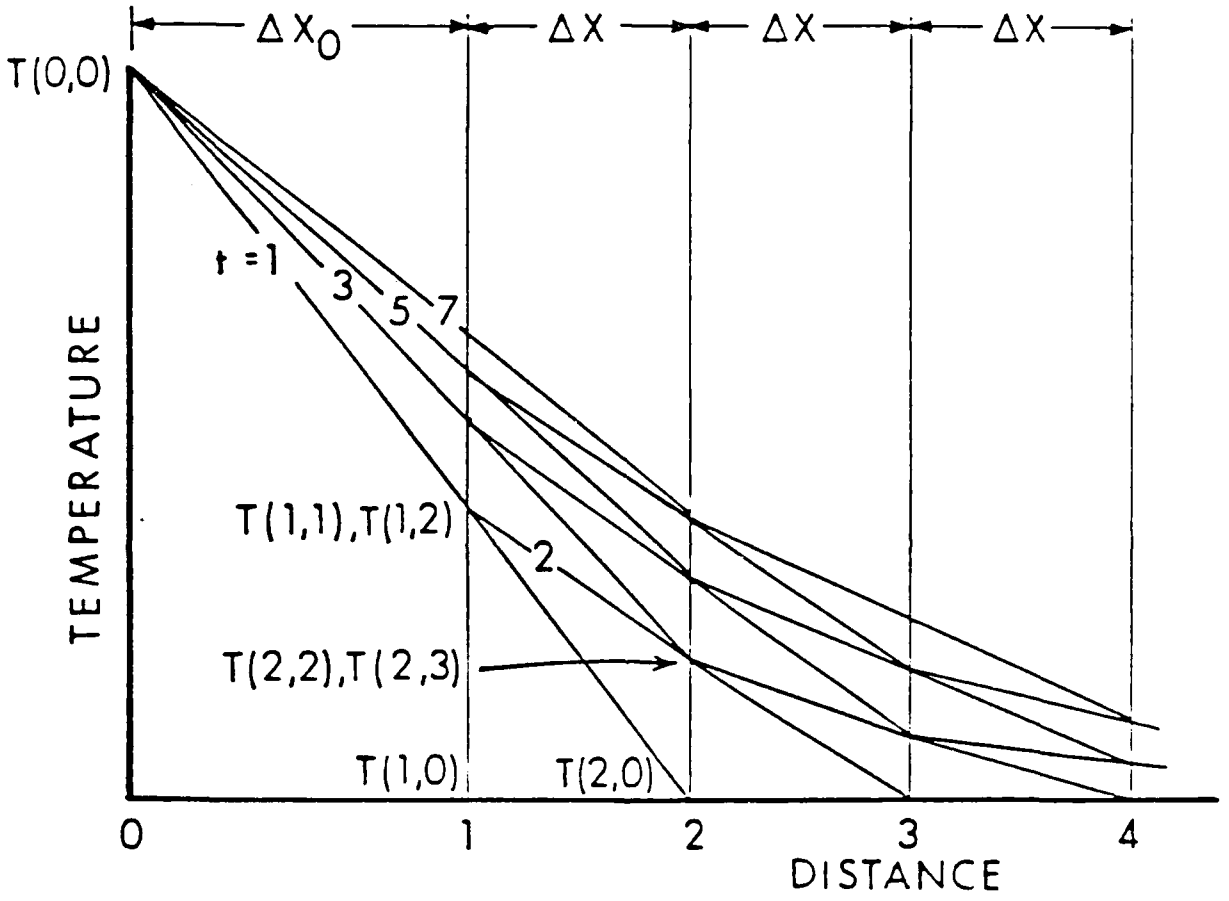


Figure 1. Schmidt Diagram

TABLE 1. Time Location

N ↓ / P →	1	2	3	4	5	6
1	Ø	Ø	(Ø/2)(3 - m)	(Ø/2)(3 - m)	(Ø/4)(30/4 - 9/2 m + m^2)	(Ø/4)(30/4 - 9/2 m + m^2)
2		Ø/2	Ø/2	(Ø/4)(7/2 - m)	(Ø/4)(7/2 - m)	(Ø/8)(38/4 - 10/2 m + m^2)
3			Ø/4	Ø/4	(Ø/8)(4 - m)	(Ø/8)(4 - m)
4				Ø/8	Ø/8	(Ø/16)(9/2 - m)
5					Ø/16	Ø/16
6						Ø/32
7						
8						

TABLE 1. (continued)

7	8
$(\emptyset/8)(35/2 - 29/2\ m + 12/2\ m^2 - m^3)$	$(\emptyset/8)(35/2 - 29/2\ m + m^2 - m^3)$
$(\emptyset/8)(35/4 - 10/2\ m + m^2)$	$(\emptyset/16)(187/8 - 69/4\ m + 13/2\ m^2 - m^3)$
$(\emptyset/16)(47/4 - 11/2\ m + m^2)$	$(\emptyset/16)(47/4 - 11/2\ m + m^2)$
$(\emptyset/16)(9/2 - m)$	$(\emptyset/32)(57/4 - 12/2\ m + m^2)$
$(\emptyset/32)(5 - m)$	$(\emptyset/32)(5 - m)$
$\emptyset/32$	$(\emptyset/64)(11/2 - m)$
$\emptyset/64$	$\emptyset/64$
	$\emptyset/128$

TABLE 1. (continued)

9
$(\emptyset/16)(630/16 - 325/8\ m + 95/4\ m^2 - 15/2\ m^3 + m^4)$
$(\emptyset/16)(187/8 - 69/4\ m + 13/2\ m^2 - m^3)$
$(\emptyset/32)(244/8 - 81/4\ m + 14/2\ m^2 - m^3)$
$(\emptyset/32)(57/4 - 12/2\ m + m^2)$
$(\emptyset/64)(68/4 - 13/2\ m + m^2)$
$(\emptyset/64)(11/2 - m)$
$(\emptyset/128)(6 - m)$
$\emptyset/128$
$\emptyset/256$

TABLE 1. (continued)

10
$(\emptyset/16)(630/16 - 325/8\ m + 95/4\ m^2 - 15/2\ m^3 + m^4)$
$(\emptyset/32)(874/16 - 406/8\ m + 109/4\ m^2 - 10/2\ m^3 + m^4)$
$(\emptyset/32)(244/8 - 81/4\ m + 14/2\ m^2 - m^3)$
$(\emptyset/64)(312/8 - 94/4\ m + 15/2\ m^2 - m^3)$
$(\emptyset/64)(68/4 - 13/2\ m + m^2)$
$(\emptyset/128)(80/4 - 14/2\ m + m^2)$
$(\emptyset/128)(6 - m)$
$(\emptyset/256)(13/2 - m)$
$\emptyset/256$
$\emptyset/512$

TABLE 1. (continued)

11
$(\emptyset/32)(1386/16 - 843/8\ m + 312/4\ m^2 - 141/4\ m^3 + 18/2\ m^4 - m^5)$
$(\emptyset/32)(874/16 - 406/8\ m + 109/4\ m^2 - 16/2\ m^3 + m^4)$
$(\emptyset/64)(1186/16 - 500/8\ m + 124/4\ m^2 - 17/2\ m^3 + m^4)$
$(\emptyset/64)(312/8 - 94/4\ m + 15/2\ m^2 - m^3)$
$(\emptyset/128)(392/8 - 298/4\ m + 16/2\ m^2 - m^3)$
$(\emptyset/128)(60/4 - 14/2\ m + m^2)$
$(\emptyset/256)(93/4 - 15/2\ m + m^2)$
$(\emptyset/256)(13/2 - m)$
$(\emptyset/512)(14/2 - m)$
$\emptyset/512$
$\emptyset/1024$

TABLE 1. (continued)

12
$(\emptyset/32)(1386/16 - 843/8\ m + 312/4\ m^2 - 141/4\ m^3 + 18/2\ m^4 - m^5)$
$(\emptyset/64)(1979/16 - 1093/8\ m + 374/4\ m^2 - 79/2\ m^3 + 19/2\ m^4 - m^5)$
$(\emptyset/64)(1186/16 - 500/8\ m + 124/4\ m^2 - 17/2\ m^3 + m^4)$
$(\emptyset/128)(1578/16 - 608/8\ m + 140/4\ m^2 - 18/2\ m^3 + m^4)$
$(\emptyset/128)(392/8 - 298/4\ m + 16/2\ m^2 - m^3)$
$(\emptyset/256)(485/8 - 123/4\ m + 17/2\ m^2 - m^3)$
$(\emptyset/256)(93/4 - 15/2\ m + m^2)$
$(\emptyset/512)(214/8 - 16/2\ m + m^2)$
$(\emptyset/512)(14/2 - m)$
$(\emptyset/1024)(15/2 - m)$
$\emptyset/1024$
$\emptyset/2048$

TABLE 1. (continued)

13
$(\emptyset/64)(3003/16 - 4165/16\ m + 1841/8\ m^2 - 532/4\ m^3 - 196/4\ m^4 + 21/2\ m^5 - m^6)$
$(\emptyset/64)(1979/16 - 1093/8\ m + 374/4\ m^2 - 79/2\ m^3 + 19/2\ m^4 - m^5)$
$(\emptyset/128)(2762/16 - 1397/8\ m + 444/4\ m^2 - 88/2\ m^3 + 20/2\ m^4 - m^5)$
$(\emptyset/128)(1578/16 - 608/8\ m + 140/4\ m^2 - 18/2\ m^3 + m^4)$
$(\emptyset/256)(2063/16 - 731/8\ m + 157/4\ m^2 - 19/2\ m^3 + m^4)$
$(\emptyset/256)(485/8 - 123/4\ m + 17/2\ m^2 - m^3)$
$(\emptyset/512)(592/8 - 139/4\ m + 17/2\ m^2 - m^3)$
$(\emptyset/512)(214/8 - 16/2\ m + m^2)$
$(\emptyset/1024)(244/8 - 17/2\ m + m^2)$
$(\emptyset/1024)(15/2 - m)$
$(\emptyset/2048)(16/2 - m)$
$\emptyset/2048$
$\emptyset/4096$

TABLE 1. (continued)

14
$(\emptyset/64)(3003/16 - 4165/16\ m + 1841/8\ m^2 - 532/4\ m^3 + 196/4\ m^4 - 21/2\ m^5 + m^6)$
$(\emptyset/128)(4387/16 - 5562/16\ m + 2285/8\ m^2 - 310/2\ m^3 + 216/4\ m^4 - 22/2\ m^5 + m^6)$
$(\emptyset/128)(2786/16 - 2794/16\ m + 444/4\ m^2 - 88/2\ m^3 + 20/2\ m^4 - m^5)$
$(\emptyset/256)(7599/32 - 3525/16\ m + 1045/8\ m^2 - 195/4\ m^3 + 21/2\ m^4 - m^5)$
$(\emptyset/256)(2063/16 - 731/8\ m + 157/4\ m^2 - 19/2\ m^3 + m^4)$
$(\emptyset/512)(2655/16 - 870/8\ m + 175/4\ m^2 - 20/2\ m^3 + m^4)$
$(\emptyset/512)(592/8 - 139/4\ m + 18/2\ m^2 - m^3)$
$(\emptyset/1024)(714/8 - 156/4\ m + 19/2\ m^2 - m^3)$
$(\emptyset/1024)(244/8 - 17/2\ m + m^2)$
$(\emptyset/2048)(276/8 - 18/2\ m + m^2)$
$(\emptyset/2048)(16/2 - m)$
$(\emptyset/4096)(17/2 - m)$
$\emptyset/4096$
$\emptyset/8192$

TABLE 1. (continued)

15
$(\emptyset/128)(6435/16 - 9949/16\ m + 5066/8\ m^2 - 3525/8\ m^3 + 836/4\ m^4 - 130/2\ m^5 + 24/2\ m^6 - m^7)$
$(\emptyset/128)(4387/16 - 5562/16\ m + 2285/8\ m^2 - 310/2\ m^3 + 216/4\ m^4 - 22/2\ m^5 + m^6)$
$(\emptyset/256)(25147/64 - 14649/32\ m + 5615/16\ m^2 - 1485/8\ m^3 + 237/4\ m^4 - 23/2\ m^5 + m^6)$
$(\emptyset/256)(7599/32 - 3525/16\ m + 1045/8\ m^2 - 195/4\ m^3 + 21/2\ m^4 - m^5)$
$(\emptyset/512)(10254/32 - 4395/16\ m + 1220/8\ m^2 - 215/4\ m^3 + 22/2\ m^4 - m^5)$
$(\emptyset/512)(2655/16 - 873/8\ m + 175/4\ m^2 - 20/2\ m^3 + m^4)$
$(\emptyset/1024)(3369/16 - 513/4\ m + 97/2\ m^2 - 21/2\ m^3 + m^4)$
$(\emptyset/1024)(714/8 - 156/4\ m + 19/2\ m^2 - m^3)$
$(\emptyset/2048)(852/8 - 174/4\ m + 20/2\ m^2 - m^3)$
$(\emptyset/2048)(138/4 - 18/2\ m + m^2)$
$(\emptyset/4096)(155/4 - 19/2\ m + m^2)$
$(\emptyset/4096)(17/2 - m)$
$(\emptyset/8192)(18/2 - m)$
$\emptyset/8192$
$\emptyset/16384$

TABLE 1. (continued)

16
$(\emptyset/128)(6435/16 - 9949/16\ m + 5066/8\ m^2 - 3525/8\ m^3 + 836/4\ m^4 - 130/2\ m^5 + 24/2\ m^6 - m^7)$
$(\emptyset/256)(76627/128 - 43947/64\ m + 25879/32\ m^2 - 8485/16\ m^3 + 1090/8\ m^4 - 283/4\ m^5 + 23/2m^6 - m^7$
$(\emptyset/512)(35401/64 - 9502/16\ m + 6835/16\ m^2 - 825/4\ m^3 + 259/4\ m^4 - 24/2\ m^5 + m^6)$
$(\emptyset/512)(10254/32 - 4395/16\ m + 1220/8\ m^2 - 215/4\ m^3 + 22/2\ m^4 - m^5)$
$(\emptyset/1024)(13623/32 - 5421/16\ m + 1414/8\ m^2 - 236/4\ m^3 + 23/8\ m^4 - m^5)$
$(\emptyset/1024)(3369/16 - 513/4\ m + 97/2\ m^2 - 21/3\ m^3 + m^4)$
$(\emptyset/2048)(4221/16 - 600/4\ m + 107/2\ m^2 - 22/2\ m^3 + m^4)$
$(\emptyset/2048)(852/8 - 174/4\ m + 20/2\ m^2 - m^3)$
$(\emptyset/4096)(1007/8 - 193/4\ m + 21/2\ m^2 - m^3)$
$(\emptyset/4096)(155/4 - 19/2\ m + m^2)$
$(\emptyset/8192)(173/4 - 20/2\ m + m^2)$
$(\emptyset/8192)(18/2 - m)$
$(\emptyset/16384)(19/2 - m)$
$\emptyset/16384$
$\emptyset/32768$

$$T(N,P) = T(N,N) = \emptyset/2^{N-1}$$

$$T(N,P) = T(N,N+2) = \emptyset/2^{N} \left[(N + 5)/2 - m\right]$$

$$T(N,P) = T(N,N+4) = \emptyset/2^{N+1} \left[(N^2 + 13N + 46)/8 - (N + 8)m/2 + m^2)\right]$$

$$T(N,P) = T(N,N+6) = \emptyset/2^{N+2} \left[(N^3 + 24N^2 + 203N + 612)/48 - (N^2 + 19N + 96)m/8 + (N + 11)m^2/2 - m^3)\right]$$

$$T(N,P) = T(N,N+8) = \emptyset/2^{N+3} \left[(N^4 + 38N^3 + 568N^2 + 3886N + 10632)/384 - (N^3 + 33N^2 + 380N + 1536)m/48 + (N^2 + 25N + 164)m^2/8 - (N + 14)m^3/2 + m^4)\right]$$

$$T(N,P) = T(N,N+10) = \emptyset/2^{N+4} \left[(N^5 + 55N^4 + 1245N^3 + 14585N^2 + 88994N + 227760)/3840 - (N^4 + 50N^3 + 971N^2 + 8722N + 30720)m/384 + (N^3 + 42N^2 + 611N + 3090)m^2/48 - (N^2 + 31N + 250)m^3/8 + (N + 17)m^4/2 - m^5)\right]$$

$$T(N,P) = T(N,N+12) = \emptyset/2^{N+5} \left[(N^6 + 75N^5 + 2395N^4 + 41865N^3 + 424564N^2 + 238060N + 5798880)/46080 - (N^5 + 70N^4 + 2015N^3 + 29930N^2 + 23030N + 737280)m/3840 + (N^4 + 62N^3 + 1487N^2 + 16402N + 70416)m^2/384 - (N^3 + 51N^2 + 896N + 5436)m^3/48 + (N^2 + 37N + 354)m^4/8 - (N + 20)m^5/2 + m^6)\right]$$

FIGURE 2. Gregory-Newton Transposition

TABLE 2. Constant Term Numerators

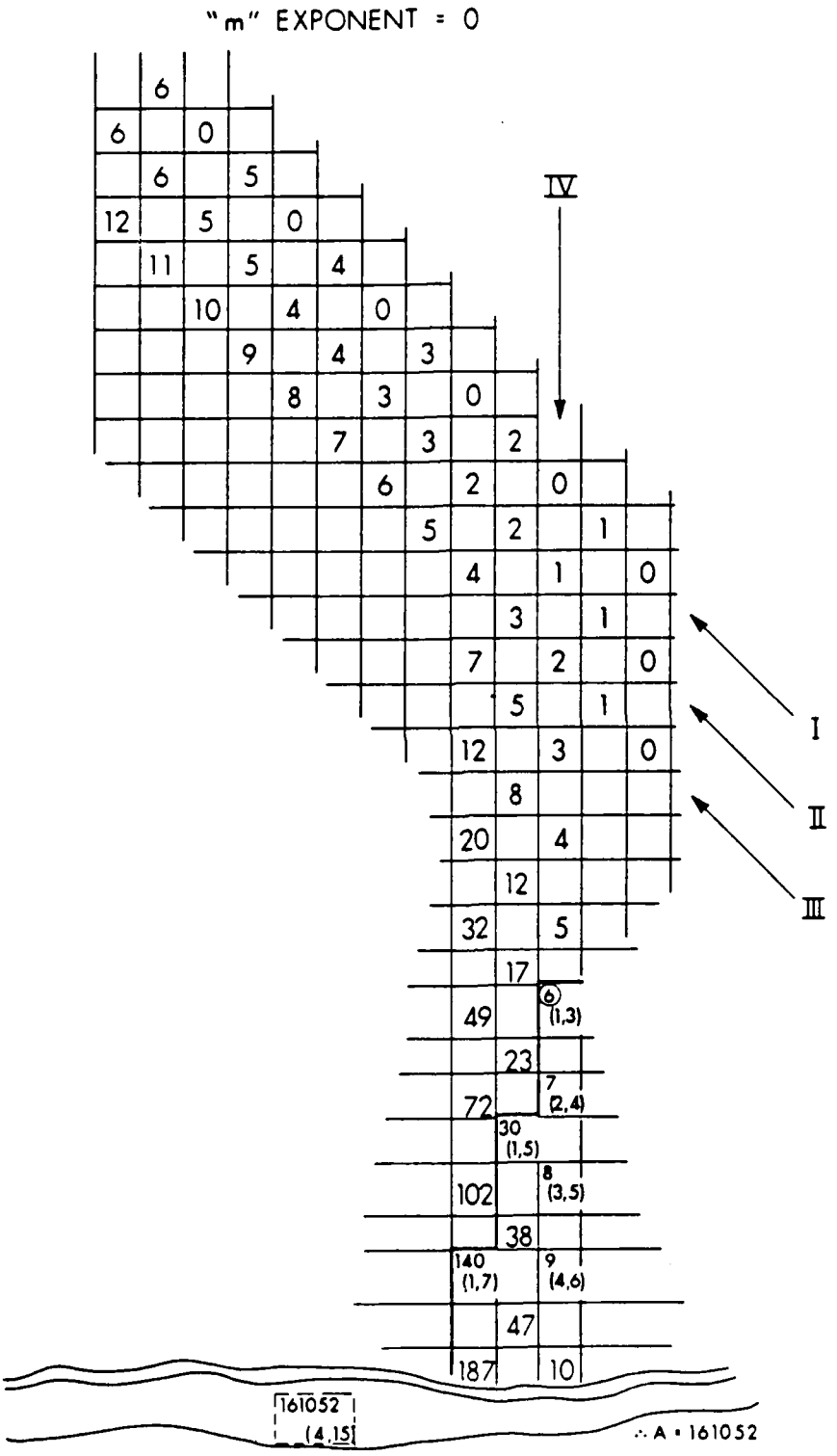


TABLE 3. Polynomial Expansions for Selected Intersections

$T(2,14) = \emptyset/128 \quad 17547/64 - 11124m/32 + 4570m^2/16 - 1240m^3/8 + 216m^4/4 - 22m^5/2 + m^6$

$T(3,15) = \emptyset/256 \quad 25147/64 - 14640m/32 + 5615m^2/16 - 1435m^3/8 + 237m^4/4 - 23m^5/2 + m^6$

$T(4,16) = \emptyset/512 \quad 35401/64 - 19044m/32 + 6835m^2/16 - 1650m^3/8 + 259m^4/4 - 24m^5/2 + m^6$

$T(5,16) = \emptyset/1024 \quad 49024/64 - 24465m/32 + 8249m^2/16 - 1886m^3/8 + 289m^4/4 - 25m^5/2 + m^6$

$T(6,16) = \emptyset/2048 \quad 66560/64 - 31096m/32 + 9877m^2/16 - 3344m^3/8 + 306m^4/4 - 26m^5/2 + m^6$

TABLE 4. 1st Power Numerators

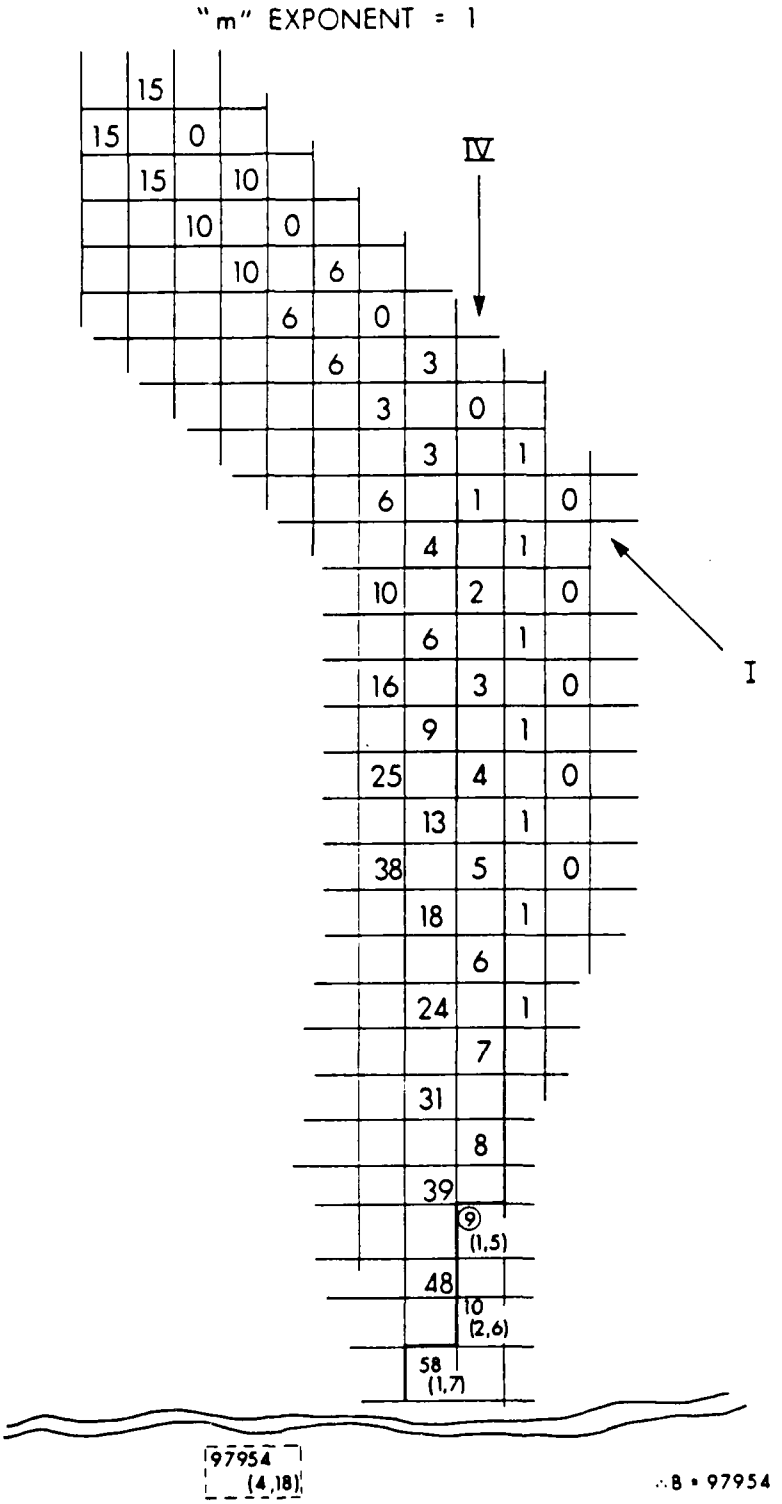


TABLE 5. 2nd Power Numerators

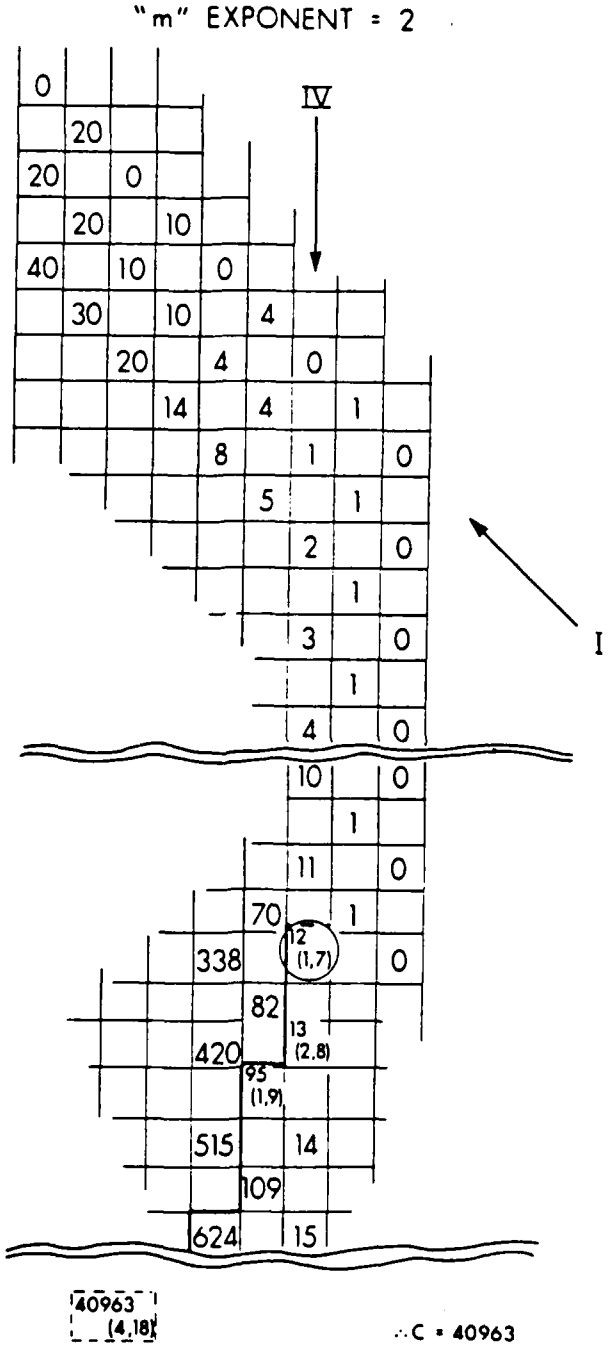


TABLE 6. 3rd Power Numerators

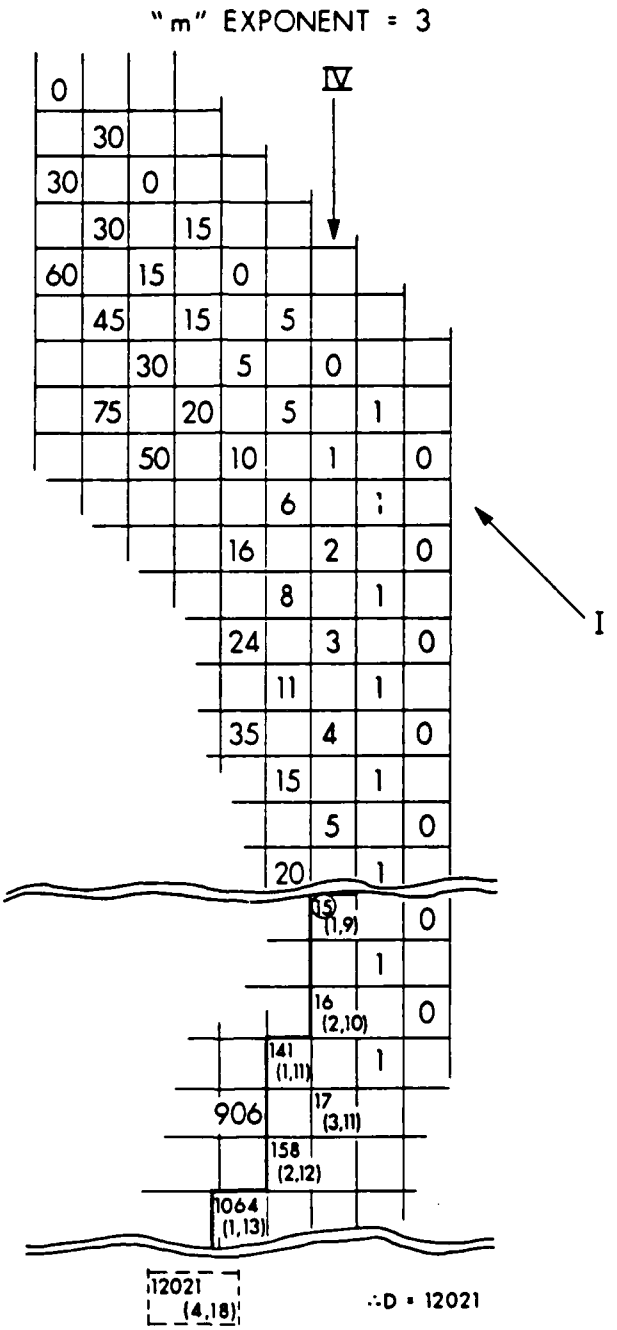


TABLE 7. 4th Power Numerators

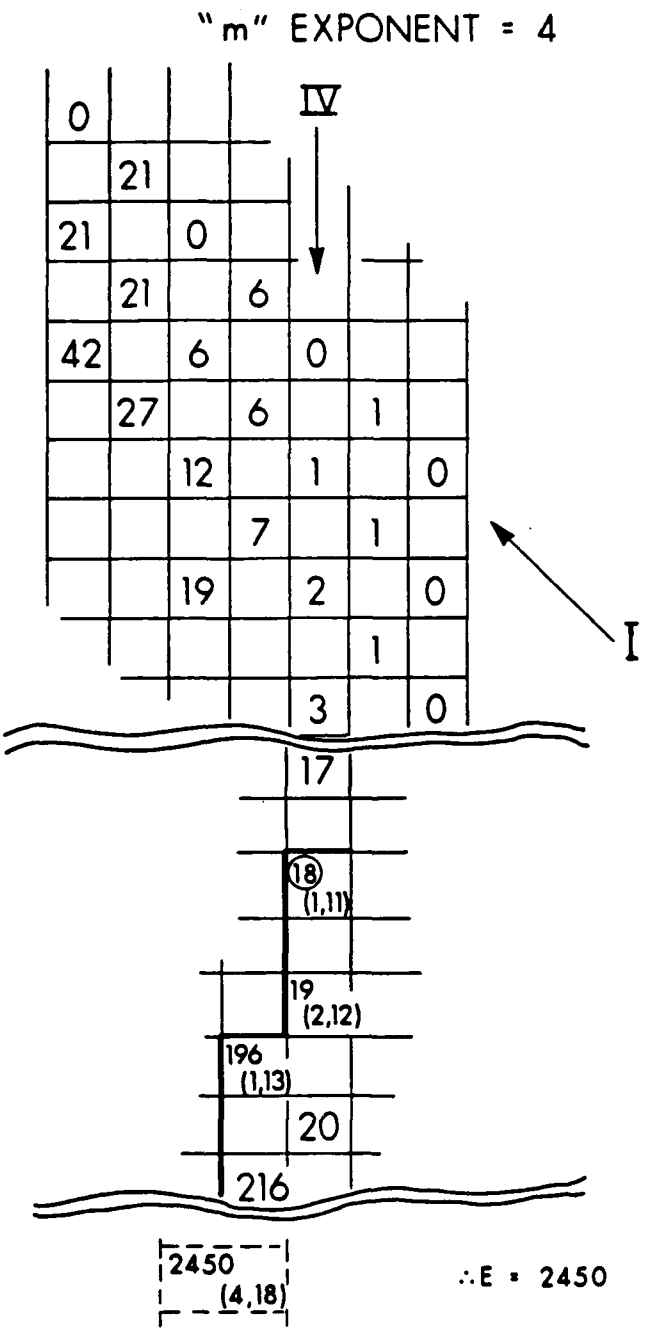


TABLE 8. 5th Power Numerators

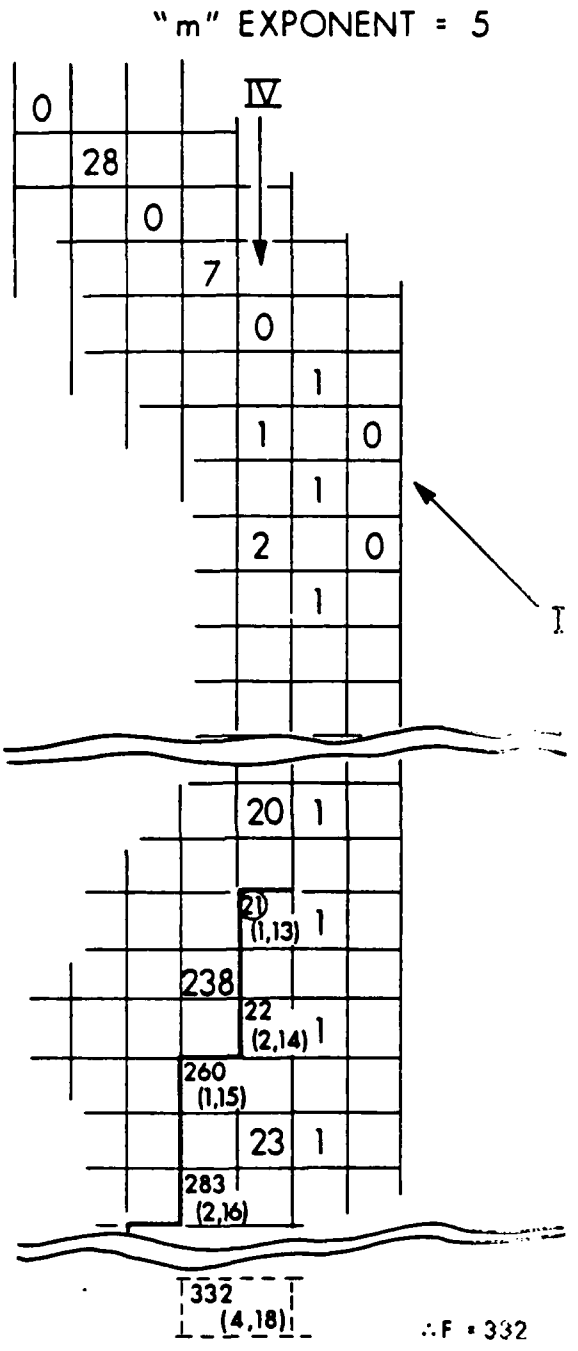


TABLE 9. 6th Power Numerators

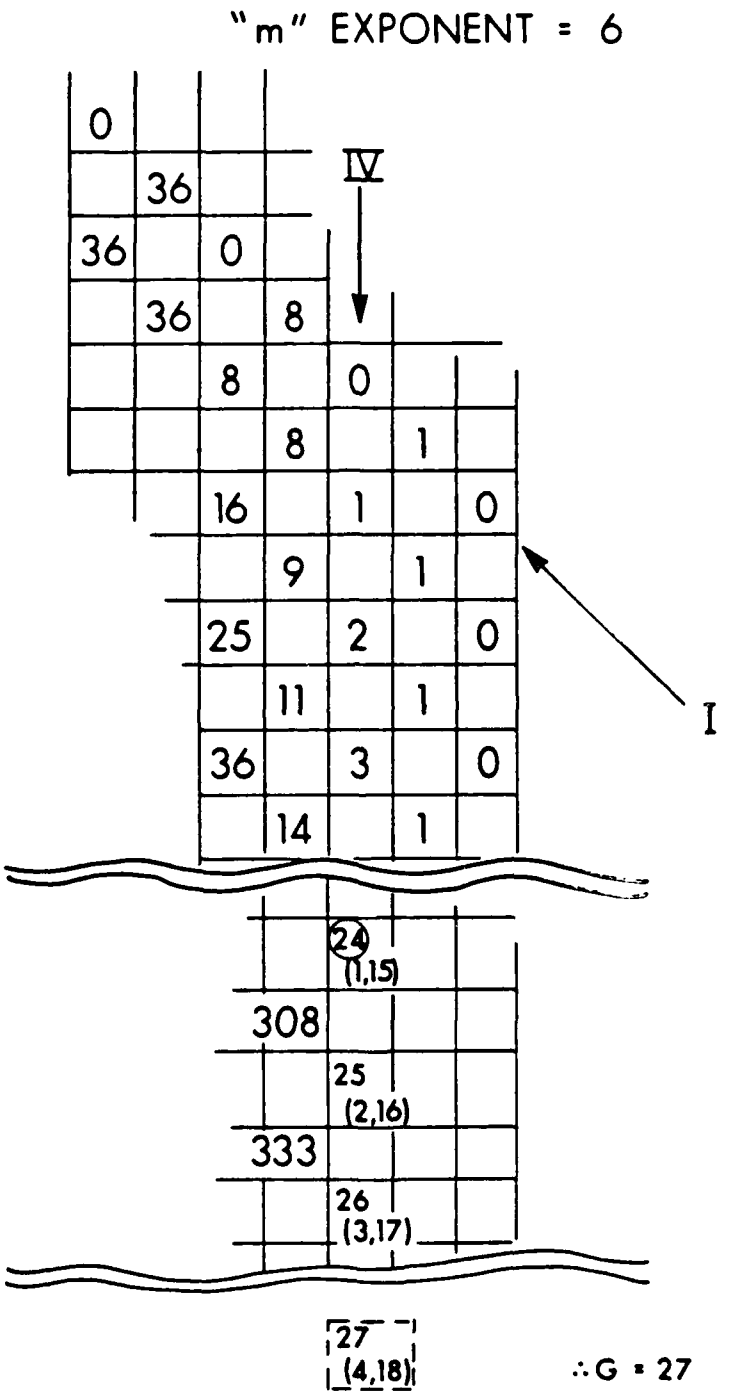


TABLE 10. Line Progression Correlations

Diagonal	Term Exponent	Column Value									
I	0	0	1	0	2	0	3	0	4	0	5
	1	0	1	0	3	0	6	0	10	0	15
	2	0	1	0	4	0	10	0	20		
	3	0	1	0	5	0	15				
	4	0	1	0	6	0	21				
	5	0	1	0	7						
II	0	0	1	1	2	2	3	3	4	4	5
	1	0	1	1	3	3	6	6	10	10	
	2	0	1	1	4	4	10	10	20		
	3	0	1	1	5	5	15	15			
	4	0	1	1	6	6	21	21			
	5	0	1	1	7						
	6	0	1	1	8						
III	0	0	1	2	3	4	5	6	7	8	9
	1	0	1	2	4	6	9	12	16	20	
	2	0	1	2	5	8	14	20	30		
	3	0	1	2	6	10	20	30			
	4	0	1	2	7	12	27				
	5	0	1	2	8						
	6	0	1	2	9						

TABLE 11. Vertical Counter Correlation

Term Exponent	"IV" Column Value	Distance-Time Coordinate
0	6	(1,3)
1	9	(1,5)
2	12	(1,7)
3	15	(1,9)
4	18	(1,11)
5	21	(1,13)
6	24	(1,15)
7	27	(1,17)
8	30	(1,19)
9	33	(1,21)

REFERENCES

1. Sutton, G.P., Rocket Propulsion Elements, John Wiley & Sons, New York 1956.

2. Spiegel, M.R., Theory and Problems of Finite Differences and Finite Difference Equations, Schaum's Outline Series in Mathematics, McGraw-Hill Book Company, New York, etc., 1971, p.p. 36-44.

3. Donovan, W.F., "Transposition of Schmidt Plot Graphics to Generalized Algebraic Expression for a Particular Heat Transfer Domain," Ballistic Research Laboratories Report (submitted).

LIST OF SYMBOLS

a	constant (diffusivity)
h	exponent of 2 in the external denominator
j	exponent of 2 in each term denominator
k	exponent of "m" in final term
m	constant $\left(\frac{\Delta_\lambda}{\Delta_\lambda + \Delta_{\lambda_o}}\right)$
n	number of λ increments
t	independent variable (time)
λ	independent variable (distance)
A,B,C,...	numerical coefficients
N	distance index
P	time index
T	Dependent variable (temperature)
∂	indicating partial differentrative
Δ	indicating difference
$\Delta_t T$	time variable effecting a change in T
$\Delta_\lambda T$	distance variable effecting a change in T
φ	external numerator

The Numerical Simulation of Richtmyer-Meshkov Unstable Interfaces: A Conference Report

John Grove [1]

Courant Institute of Mathematical Sciences
New York University
New York, New York 10012

ABSTRACT

This paper investigates the interaction between a planar shock wave and a perturbed contact discontinuity. The interaction is simulated by a front tracking method that uses a local steady state analysis to model the diffraction patterns produced by the collision This front tracking method automatically adjusts the topology of the tracked interface to account for the wave interactions. The acceleration of the contact discontinuity by the shock wave excites unstable modes in the gas interface that generate Richtmyer-Meshkov instabilities.

The following is an shortened version of a paper submitted to the SIAM Journal on Applied Mathematics.

1. Introduction

The numerical simulation of a collision between a planar shock wave and a contact discontinuity surface is discussed in this paper. An important feature of this method of simulation is the use of a front tracking algorithm that handles bifurcations of tracked waves. Front tracking sharply resolves the diffracted wave patterns that are produced as the two waves collide. It also gives a detailed picture of the growth of surface instabilities in the gas interface.

The initial small amplitude linear analysis of the shock-contact interaction is due to Richtmyer [1], and experimental confirmation was provided by Meshkov, et al. [2]. Thus this interaction is usually referred as the Richtmyer-Meshkov instability.

Recent calculations by D. L. Youngs [3] give a detailed view of this instability, including the large amplitude, late time regime. Eulerian methods are used because the extreme degree of interface complexity would lead to excessive mesh distortion in Lagrangian codes. However Eulerian codes tend to suffer from numerical diffusion, that degrades the interface. Youngs uses a volume in cell Eulerian method [4-7] with the monotonic advection method of Van Leer [8,9] to enhance the interface resolution and to minimize the numerical diffusion.

There are two principle methodological differences between the computations of Youngs and the ones presented in this paper. The first is the use of the front tracking algorithm for an exact resolution of the interface, and the resulting absence of numerical diffusion across the interface. In [4], Youngs states, "*A possible way of tracking interfaces would be to define each interface by a set of Lagrangian marker particles. However, this method becomes logically complicated if the interfaces become highly distorted or the the geometry is complex.*" It is believed that the front tracking method used here shows that this problem has

1. Supported in part by the Army Research Office, grant DAAG29-84-K-0130.

been solved for interfaces with a considerable degree of complexity. The second principal methodological difference is the use by Youngs of the monotonic advection method of Van Leer. The front tracking algorithm at present uses a second-order Lax-Wendroff method for the solution away from the tracked interface. However there is no inherent incompatibility between the method of front tracking and such second-order Godunov methods as the Van Leer scheme or the PPM method of Colella and Woodward [10], indeed upgrades of the front tracking algorithm are planned that will include such methods.

The shock-contact interaction modeled here is the original problem of Richtmyer [1], in which a shock wave collides an interface between two gases. Youngs on the other hand models the shock tube experiment of Meshkov et al. [2] in which a shock originally incident in the heavier gas collides with the contact discontinuity interface, is reflected by a rigid wall, and the reflected shock again interacts with the interface. These problems are closely related and close qualitative similarities between the results obtained here and those of Young's are observed, with the expected difference of an absence of numerical diffusion in the front tracking method. A detailed comparison has not been attempted at this point. A second difference is that in the published results of this paper the parameter ranges include strong incident shocks with pressure ratios across the shocks up to $\frac{P_1}{P_0} = 1000$ and shock Mach numbers up to 28. The most nearly comparable of the figures in the two papers is perhaps figure 4.4(e) of this paper and figure 9 frame 3 of [3]. The mesh used in the front tracking run is about 4.4 times as coarse, per mode, in the direction parallel to the interface as is that of Youngs. In spite of the coarser grid, the results show a considerably finer level of detail at the interface. This is not surprising, since the front tracking algorithm concentrates numerical power at the interface. The resolution of the solution in the untracked portion of the computational region will of course reflect the relative coarseness of the grid.

The number and types of the unstable modes that are observed in a shock wave and contact discontinuity interaction depend on the incident shock strength, the initial geometry of the two waves and the physical properties of the gases. A single mode can be isolated when the incident shock wave is planar and the contact discontinuity surface has the shape of a sine curve of a single period. More complicated initial geometries for the initial gas interface can be used to study the interaction between different unstable modes.

2. The Front Tracking Algorithm

The front tracking algorithm [11-14], is an adaptive grid method for the sharp resolution of selected waves in numerical solutions to systems of partial differential equations in two space dimensions:

$$\mathbf{w}_t + \nabla \cdot \vec{F}(\mathbf{w}) = 0. \tag{2.1}$$

Usually these waves represent discontinuities in the solution function, for example, shock waves or contact discontinuities. The selected waves are tracked by superimposing a set of one-dimensional curves onto an underlying rectangular grid. These curves correspond to the location of the tracked waves at a given time and are dynamically modified as the solution evolves in time.

Some terminology will be helpful. The basic data structures are points, bonds, curves, nodes and interfaces, see [11]. A point describes a location in space and a bond contains the information needed to describe an oriented linear segment connecting two points. A curve is an ordered set of bonds, and thus corresponds to a piecewise linear ordered curve in space. All curves are assumed to be continuous. The start and end points of a curve are called nodes. Several curves may meet at the same node. An interface is a collection of curves and nodes.

Values for the state variables that describe a solution to system (2.1) are associated with geometric points on a rectangular grid. In addition, since the tracked curves represent discontinuities in the solution function, two sets of state values are associated with each point

on a curve. These correspond to the value of the solution on either side of the curve at the particular point. States are also associated with the start and end of each curve. These states correspond to the tangential limits of the solution as the end point of the curve is approached along the curve. The solution in a neighborhood of a node is described by the start or end states of the curves going into that node.

The propagation of the solution from time t to time $t + \Delta t$ is divided into two main parts, the propagation of the tracked wave structures (the front propagation), and the updating of the values of the states at locations away from the tracked interface (the interior propagation). The simulations described in this paper used an operator split Lax-Wendroff method for the interior propagation. This finite difference method has been modified to use the states on the tracked interface as boundary data.

At each non-node point P on the tracked interface a one dimensional Riemann problem is solved for the component of system (2.1) normal the curve through P. The solution to this Riemann problem gives the wave speed and a set of updated states at P. Later a second sweep over the points is performed in which the contribution of the tangential component of the equations is included. See [13] for a description of the details of these steps. The nodes are treated separately from the non-node points on the interface, since the solution is fully two dimensional at such points and operator splitting does not apply. The states near a node and the tangents of the curves joining the node define a two dimensional Riemann problem [15], and the solution of this Riemann problem is taken as the first order solution near the node. Sometimes it is also possible to compute higher order corrections to the states near a node that include such effects as the curvature of the incoming waves and the variability of the solution near the node.

Often an explicit expression for the solution to a given two-dimensional Riemann problem is unavailable. In such cases the solution to the two-dimensional Riemann problem is approximated by finding a projection of the exact solution onto a subclass of functions that will capture the main features of the interaction. The next section will discuss such a projection for a node that corresponds to a shock-contact collision.

A more detailed description of the propagation of the tracked interface for one time step can be found in [16].

3. The Tracking of Shock-Contact Interactions

The direct simulation of the Richtmyer-Meshkov instability is based on a numerical solution to the Euler equations for a non-viscous, non-heat conducting gas.

(Conservation of mass)

$$\rho_t + (\rho u)_x + (\rho v)_y = 0, \tag{3.1a}$$

(Conservation of momentum)

$$(\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = 0, \tag{3.1b}$$

$$(\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = 0, \tag{3.1c}$$

(Conservation of energy)

$$\left\{\rho[\tfrac{1}{2}q^2 + e]\right\}_t + \left\{\rho u[\tfrac{1}{2}q^2 + i]\right\}_x + \left\{\rho v[\tfrac{1}{2}q^2 + i]\right\}_y = 0. \tag{3.1d}$$

The variables u and v are the x and y components of the gas velocity at the point (x, y), $q^2 = u^2 + v^2$. The thermodynamic variables ρ, e, p and $i = e + \frac{p}{\rho}$ are respectively the density, specific internal energy, pressure and specific enthalpy of the gas. The thermodynamic variables for each gas are related by a caloric equation of state

$$e = e(\tau, S), \tag{3.2}$$

where $e(\tau, S)$ is a convex function of the specific volume $\tau = \frac{1}{\rho}$ and specific entropy S. In

general this equation of state will be different for the two gases on opposite sides of the gas interface.

The pressure p is given by $p(\tau, S) = -\frac{\partial e}{\partial \tau}$. Often this relation can be inverted to give the entropy and hence the energy as a function of p and ρ. This expression, called an incomplete equation of state, is usually sufficient to solve the Euler equations. The numerical examples described below used a polytropic equation of state,

$$e = \frac{p\tau}{(\gamma - 1)}, \tag{3.3}$$

where the ratio between the specific heats γ is a constant satisfying $1 < \gamma \leq \frac{5}{3}$. The author and his colleagues are actively pursuing hydrodynamic simulations with more general equations of state. Thus, the equation of state dependencies in our code have been modularized to allow the optional use of other equation of state models. This modularity requires that all hydrodynamic quantities be interpreted in terms that can be expressed for a general equation of state. In particular, this involves such requirements as an equation of state independent formulation for the solution of one dimensional Riemann problems, and equation of state independent expressions for the shock polars described below, see [16].

The Richtmyer-Meshkov instability simulation is initialized at a time shortly before the incident shock wave reaches the gas interface. The incident shock is taken to be planar, and the contact discontinuity interface is given an initial geometry specified by input. If a single mode is to be isolated this initial geometry is that of a sine wave of a single period across the computational domain. The gas interface is assumed to be at rest with respect to the gas ahead of the incident shock, so the initial data is piecewise constant.

Since the two interacting waves are tracked, it is necessary to resolve the diffracted wave patterns that are produced at the point of collision between the two waves. In the simulation used here, these diffraction patterns are resolved using shock polar analysis. Briefly, the interacting waves are approximated by their tangents near the point of collision between the incident shock wave and the gas interface, and the nearby states are approximated by states that are constant between the interacting waves. It is further assumed that there exists a Galilean transformation that translates this local approximation into a stationary flow. This assumption will be in general valid provided the angle between the two incident waves is small, i.e. if the initial amplitude of the perturbation of the contact discontinuity is sufficiently small.

The analysis of the interaction between a planar shock wave and a planar contact discontinuity for a polytropic equation of state is well known, [14, 17, 18]. The type of diffraction pattern that is observed is a function of the strengths of the interacting waves, the angle at which the two waves meet, and the equations of state for the two gases. The simplest of these diffraction patterns consists of the incident shock wave and contact discontinuity, a single reflected wave that is either a shock or a Prandtl-Meyer rarefaction wave, a transmitted shock wave, and a deflected gas interface behind the point of interaction. This so called regular shock diffraction is observed provided the angle between the interacting waves is sufficiently small. Many other configurations besides the regular diffraction node are possible, these include Mach type reflections and precursor shock type configurations, see [19, 20].

The interaction between a steady state shock wave and contact discontinuity can be regarded as a Riemann problem for the steady flow Euler equations. The line perpendicular to the upstream contact becomes the space-like axis, and the line parallel to the upstream contact becomes time-like in the downstream direction. The data for the Riemann problem consists of the state behind the incident shock wave, and the upstream state on the side of the contact opposite to the incident shock. It is assumed that both states are supersonic. Again this will be the case for sufficiently small incident angles. A solution is sought for this Riemann problem in the class of self-similar functions that consist of constant states separated by downstream oriented shocks, Prandtl-Meyer rarefactions, and contact

discontinuities. The abstract structure of this solution is completely analogous to that of the time dependent Riemann problem in one space dimension. The difference comes from the increased nonlinearity of the wave curves.

The characteristics for supersonic flow come in three families: streamlines, and two sonic wave families that cross the streamlines at the Mach angle $\sin(A) = \frac{c}{q}$. Where c is the local sound speed and q is the flow speed. The characteristic family associated with streamlines is linearly degenerate and the associated waves are contact discontinuities or slip lines across which the pressure and flow angle are continuous. The sonic characteristic families are genuinely non-linear provided the fundamental derivative of gas dynamics

$$\mathcal{G} = -\frac{1}{2}\tau \frac{\partial^3 e(\tau, S)/\partial\tau^3}{\partial^2 e(\tau, S)/\partial\tau^2} \neq 0. \tag{3.4}$$

For most materials $\mathcal{G} > 0$, although $\mathcal{G}$ may be negative near phase transitions. For a polytropic gas, $\mathcal{G} = \frac{(\gamma + 1)}{2} > 1$. If it is assumed that $\mathcal{G} > 0$, then the sonic characteristic wave families support waves that are either shocks or Prandtl-Meyer rarefaction waves. Since the streamline characteristic field is linearly degenerate, it follows that the solution to the Riemann problem for steady planar flow can be found by calculating the intersection of the wave curves for the two sonic families through the data points in the pressure - flow angle phase space, see [16]. The shock portion of the these wave curves correspond to the well know shock polars as described in [21].

Figure 3.1, shows a representative shock diffraction pattern along with a pair of generic streamlines and the corresponding shock polars for the case of a reflected shock.

The application of this analysis to the direct simulation of the shock contact collision consists of calculating at each time step a new steady diffraction pattern based on the changing angles between the incident waves. The transformation from the locally steady configuration to the global reference frame for the entire simulation is found by calculating an intersection between the two propagated sections of the incident shock wave and contact discontinuity. This intersection defines the position of the point of shock diffraction at time $t + \Delta t$. The difference between the positions of this point at the beginning and end of the time step provides the transformation between the two frames of reference.

There are several important issues connected with the changes in topology for the tracked waves as they collide and interact. These include the numerical detection and identification of the tracked wave interactions, and the changes to the tracked wave structures needed to simulate the underlying physics of the interactions. See [16] for a more detailed discussion of these issues.

4. Numerical Results

Figure 4.1 shows a series of frames documenting the growth of an unstable finger in an air to sulphur-hexafluoride (SF_6) interface. Both gases are modeled as polytropic gases with $\gamma = 1.4$, and $\gamma = 1.094$ respectively. The shock wave is incident in the air and the ratio of the pressure behind the shock to the pressure in front is 10. At room temperature the SF_6 is about 5.03 times as dense as air. A net vertical velocity is given to the initial contact discontinuity. This is done since the boundaries at the top and bottom of the computational rectangle are open, and it was found that the contact discontinuity exits the computational rectangle early in the simulation if a reference frame in which the original gas interface is at rest is used.

The gas interface is flattened by the incident shock wave as the two waves collide. The diffraction of the shock wave through the interface causes the reflected and transmitted shocks to assume the geometry of the original gas interface. However, as the waves continue to propagate away from each other, the unstable mode in the contact discontinuity interface begins to grow, while the two shock waves restablize to planar curves. The two shock waves eventually exit the open boundaries, leaving the contact discontinuity as the only tracked

wave.

This simulation is interesting since during the shock diffraction portion of the run, the transmitted shock wave is nearly contiguous with the deflected contact discontinuity behind the point of diffraction. The angle between the two tracked waves is less than 1°. It is believed that one strength of the front tracking method, is the ability to resolve such closely proximate waves.

Figure 4.3 shows a similar interaction, except here the shock is incident in the heavier gas. Both gases are taken as polytropic with $\gamma = 1.4$, while the pressure ratio across the incident shock is 100 and the heavier gas is ten times as dense as the lighter gas. One notes that the phase of the contact is reversed by the shock wave collision, and the interaction produces a reflected rarefaction wave rather than a reflected shock wave. The two tracked waves on the upper side of the contact are the forward and backward edges of the reflected rarefaction wave. The long time behavior of the unstable interface is show in Figure 4.3d. There is some question about the dimple that is produced in the lower edge of the contact. This may arise as a result of numerical instability. However there is some evidence that this dimple may be physical. Further studies using finer grids and comparison with other simulations are needed to resolve this question.

In addition to calculations of the growth of a single finger, simulations that invoive several unstable modes have been performed. Figure 4.4 shows a three mode interaction with the interface separating warm air from cooler air, figure 4.5 shows the interaction of a shock wave incident in helium ($\gamma = 1.63$) with a helium to air interface.

5. Conclusions

It has been shown that front tracking offers a useful method for the simulation of shock wave and contact discontinuity interactions. It allows for a sharp resolution of the diffracted wave patterns produced by the interaction of the two waves, and a clear picture of the growth of unstable modes in the gas interface.

The framework for the resolution of tracked wave interactions has been shown to be capable of handling complicated situations. Furthermore, it is possible to include new bifurcations as they are needed, or to remove tracking when the result of a wave interaction is either too complicated or unknown.

References

1. R. D. Richtmyer, "Taylor Instability in Shock Acceleration of Compressible Fluids," *Comm. Pure and Appl. Math*, vol. 13, pp. 297-319, 1960.
2. V. Andronov, S. M. Bakhrakh, E. E. Meshkov, V. N. Mokhov, V. V. Nikiforov, A. V. Pevnitskii, and A. I. Tolshmyakov, *Sov. Phys. JETP*, vol. 44, pp. 424-427, 1967.
3. David L. Youngs, "Numerical Simulation of Turbulent Mixing by Rayleigh-Taylor Instability," *Physica* , vol. 12D, pp. 32-34, 1984.
4. David L. Youngs, "Time-Dependent Multi-Material Flow with Large Fluid Distortion," in *Numerical Methods for Fluid Dynamics*, ed. M. J. Baines, Academic Press, New York, 1982.
5. J. D. Ramshaw and J. A. Trapp, "A Numerical Technique for Low-Speed Homogeneous Two-Phase Flows with Sharp Interfaces," *J. Comput. Phys.*, vol. 21, pp. 438-453, 1976.
6. C. W. Hirt and B. D. Nichols, "Volume of Fluid (VOF) Method for the Dynamics of Free Boundaries," *J. Comput. Phys.*, vol. 39, pp. 201-225, 1981.
7. W. F. Noh and P. Woodward, "SLIC (Simple Line Interface Calculation)," in *Lecture Notes in Physics*, vol. 59, Springer Verlag, New York, 1976.

8. B. Van Leer, "Toward the Ultimate Conservative Difference Scheme. IV. A New Approach to Numerical Convection," *J. Comp. Phys.*, vol. 23, pp. 276-299, 1977.

9. B. Van Leer, "Towards the Ultimate Conservative Difference Sheme. V. A Second-Order Sequel to Godunov's Method," *J. Comp. Phys.*, vol. 32, pp. 101-136, 1977.

10. Phillip Colella and Paul R. Woodward, "The Piecewise Parabolic Method (PPM) for Gas-Dynamical Simulations," *J. Comp Phys.*, vol. 54, pp. 174-201, 1984.

11. J. Glimm and O. McBryan, "A Computational Model for Interfaces," *Adv. Appl. Math.*, vol. 6, pp. 422-435, 1985.

12. J. Glimm, E. Isaacson, D. Marchesin, and O. McBryan, "Front Tracking for Hyperbolic Systems," *Adv. in Appl. Math.*, vol. 2, pp. 91-119, 1981.

13. I-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, "Front Tracking for Gas Dynamics," *J. Comp. Phys.*, vol. 62, pp. 83-110, 1986.

14. J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, "Front Tracking and Two Dimensional Riemann Problems," *Adv. in Appl. Math.*, vol. 6, pp. 259-290, 1985.

15. James Glimm and D. H. Sharp, "An S Matrix Theory for Classical Nonliner Physics," *Found. of Physics*, vol. B16, pp. 125-141, 1986.

16. J. Grove, "Front Tracking and Shock-Contact Interactions," *Submitted to SIAM Jour. Appl. Math.*.

17. L. F. Henderson , "The Refraction of a Plane Shock Wave at a Gas Interface," *J. Fluid Mech.* , vol. 26, p. 607, 1966 .

18. A. M. Abd-El-Fattah, L. F. Henderson, and A. Lozzi, "Precursor Shock Waves at a Slow-Fast Gas Interface," *J. Fluid Mech.*, vol. 76, p. 157, 1976.

19. A. M. Abd-El-Fattah and L. F. Henderson , "Shock Waves at a Fast-Slow Gas Interface ," *J. Fluid Mech.* , vol. 86, p. 15, 1978 .

20. A. M. Abd-El-Fattah and L. F. Henderson , "Shock Waves at a Slow-Fast Gas Interface ," *J. Fluid Mech.* , vol. 89 , p. 79, 1978 .

21. R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Springer Verlag, New York, 1948.

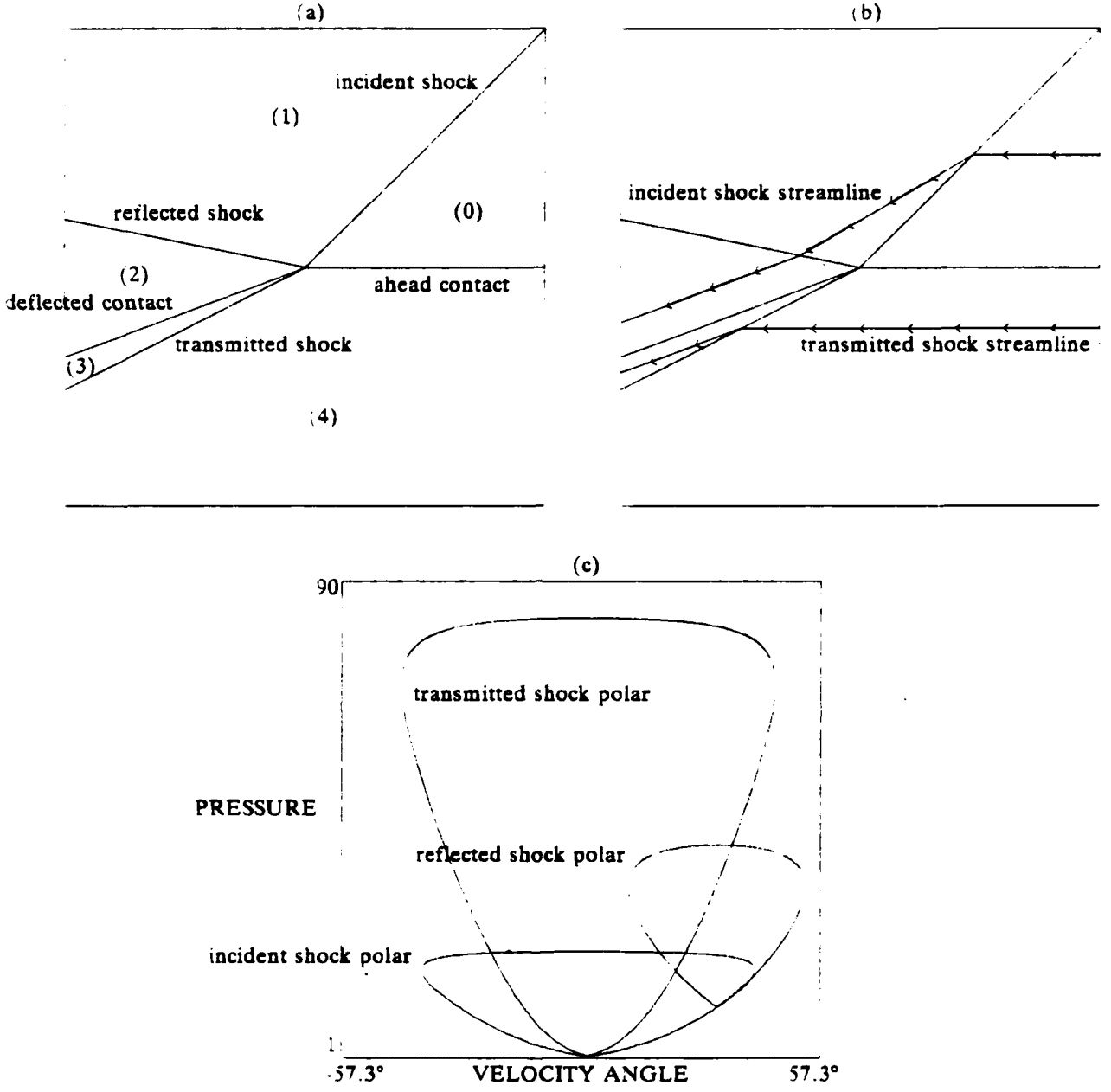


Fig. 3.1. A shock wave-contact discontinuity collision that produces a single reflected shock. The reaction occurs in air modeled as a polytropic gas with $\gamma = 1.402$. The gas on the transmitted shock side of the ahead contact discontinuity is four times as dense as the gas on the incident shock side. The angle between the incident shock and the ahead contact discontinuity is 45°, and the ratio of the pressures across the incident shock is 10. The flow is turned by about 30.3° through the incident shock, -9.8° through the reflected shock, and 20.5° through the transmitted shock.

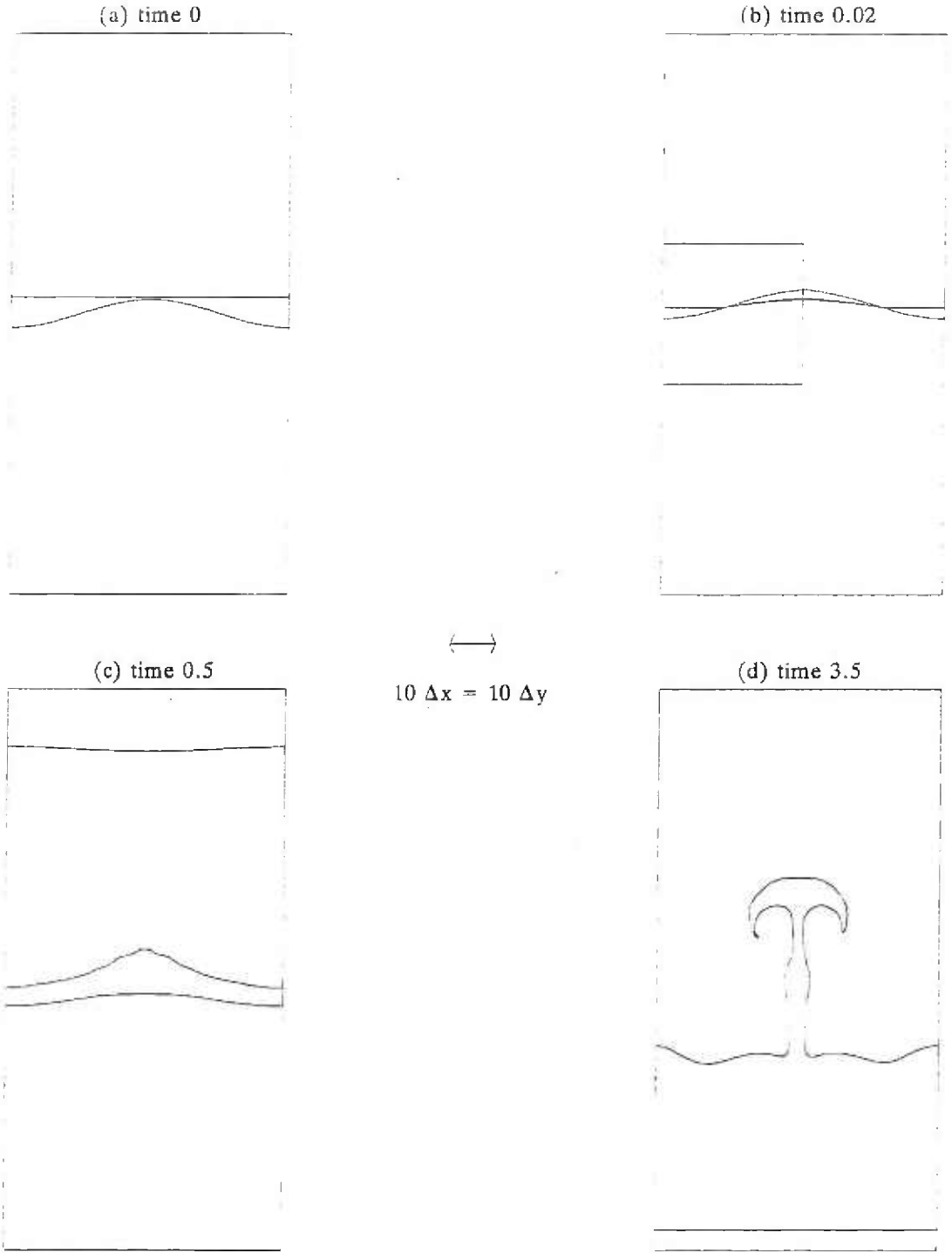


Fig. 4.1. A shock hitting a contact discontinuity separating air from the gas SF_6. The contact discontinuity curve is given an initial shape of a sine curve. The shock is incident from the air and has a pressure ratio of 10. The boxed region in Fig. 4.1b is blown up in the next figure.

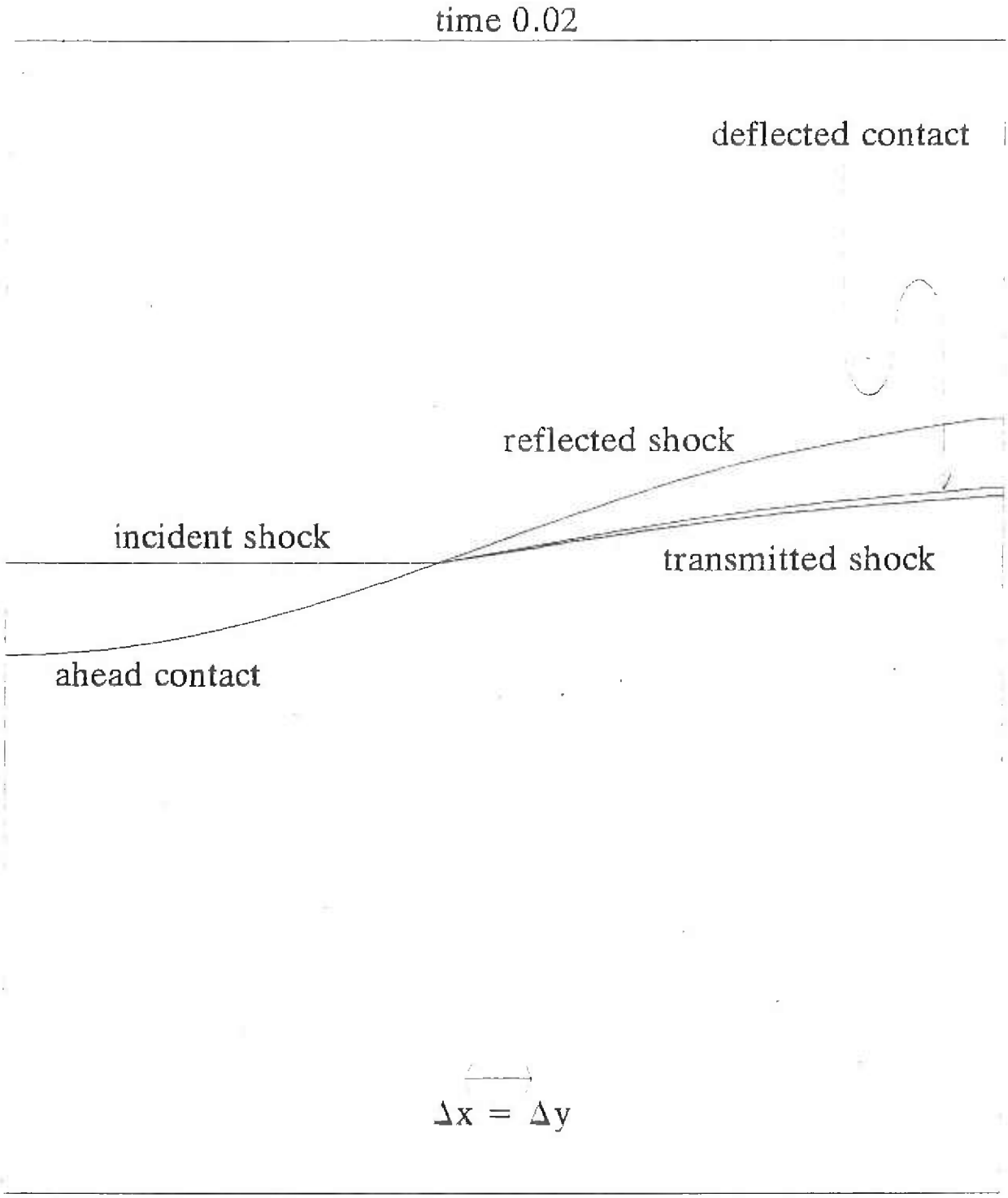


Fig. 4.2. A blowup of a subregion of Fig 4.1b showing the incident shock colliding with the ahead contact discontinuity, producing reflected and transmitted shocks.

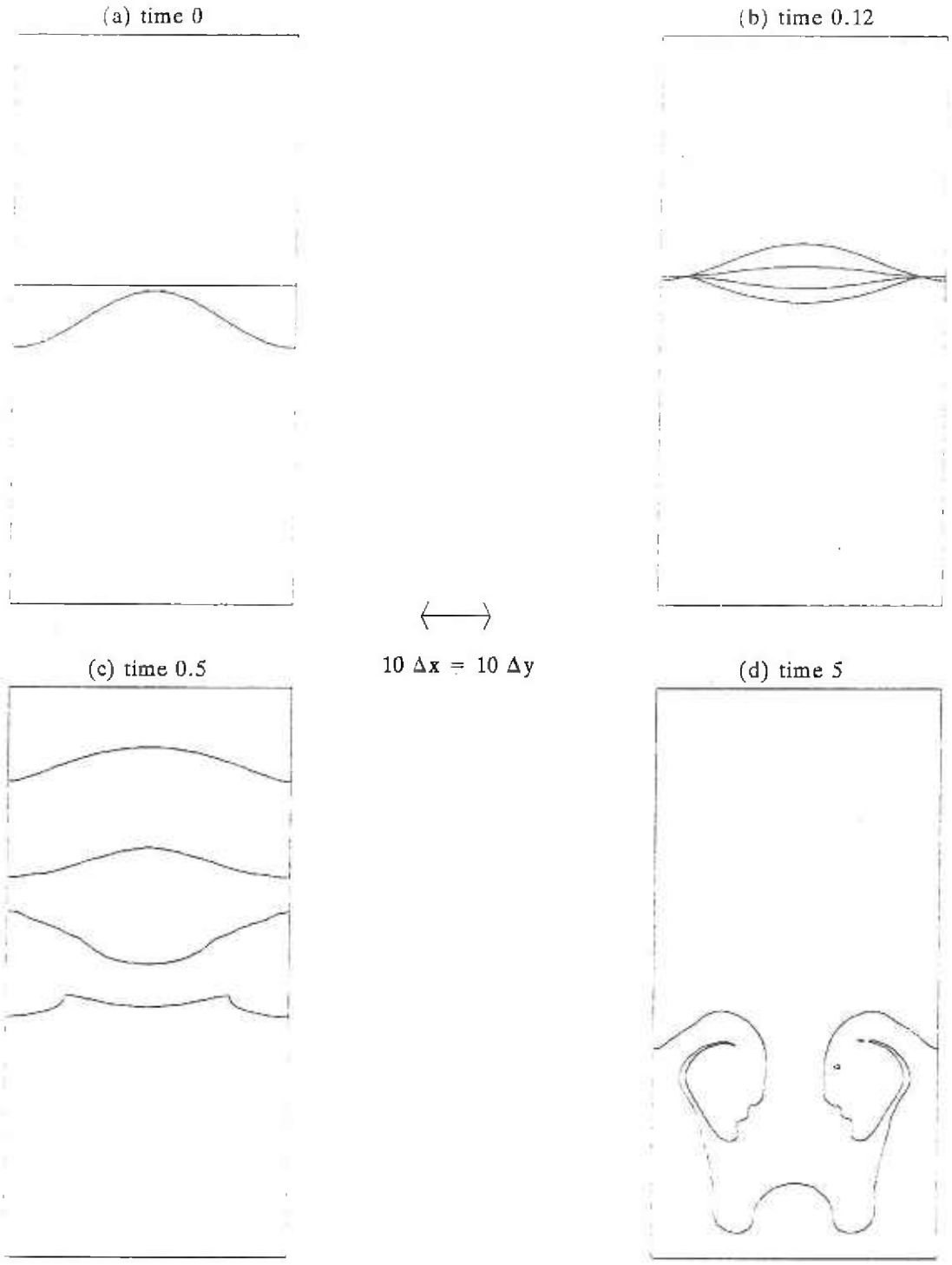


Fig. 4.3. A shock-contact interaction that produces a reflected rarefaction wave. The pressure ratio across the shock is 100 and the density ratio across the contact discontinuity is 10. Both gases are polytropic with $\gamma = 1.4$. The shock wave in incident in the heavier gas.

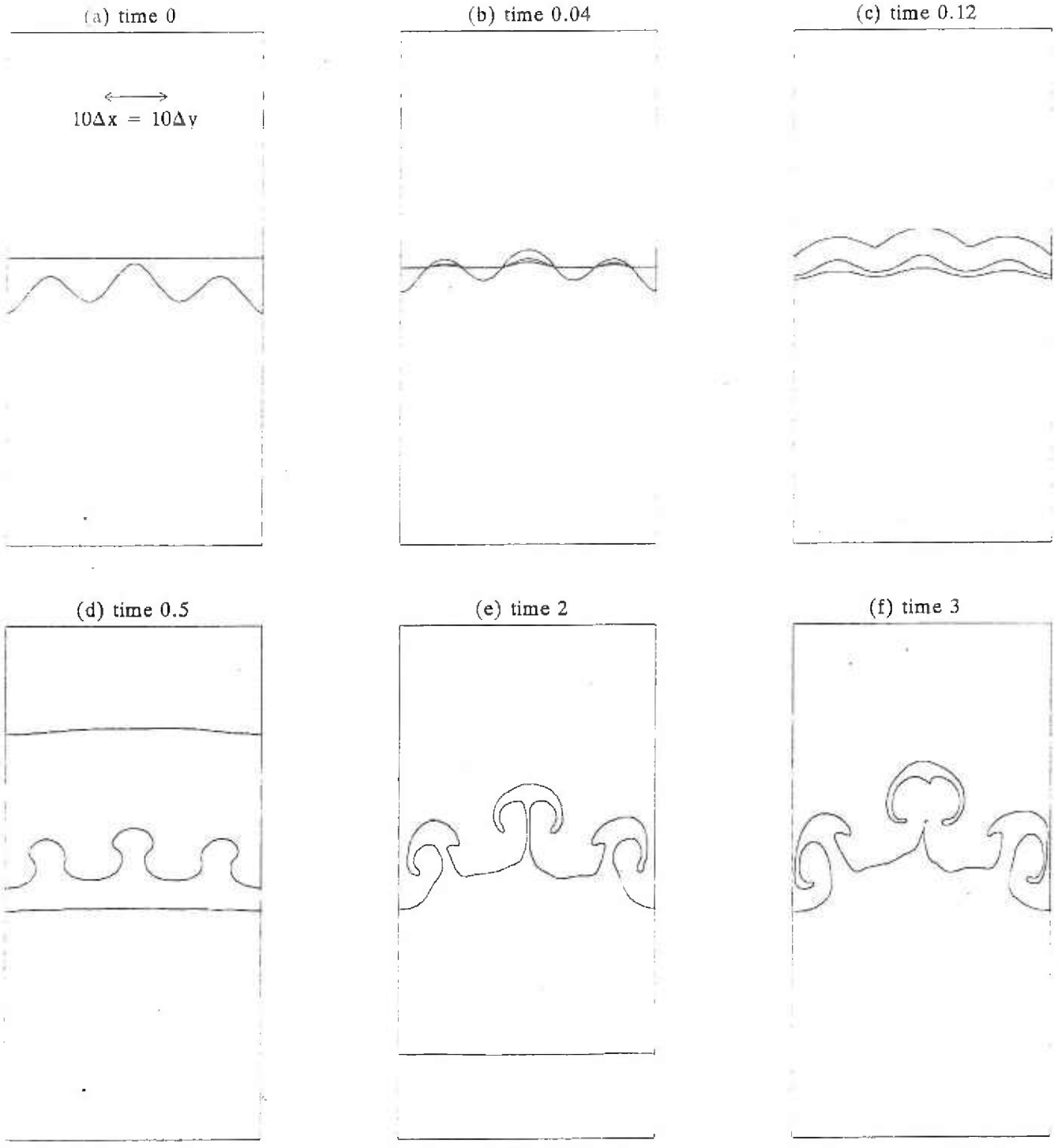


Fig. 4.4. A series of frames showing a shock contact collision interaction. Both gases are polytropic with $\gamma = 1.4$. The pressure ratio across the incident shock is 100, and the density ratio (above to below) across the original contact is 2.86. The grid is 40x80.

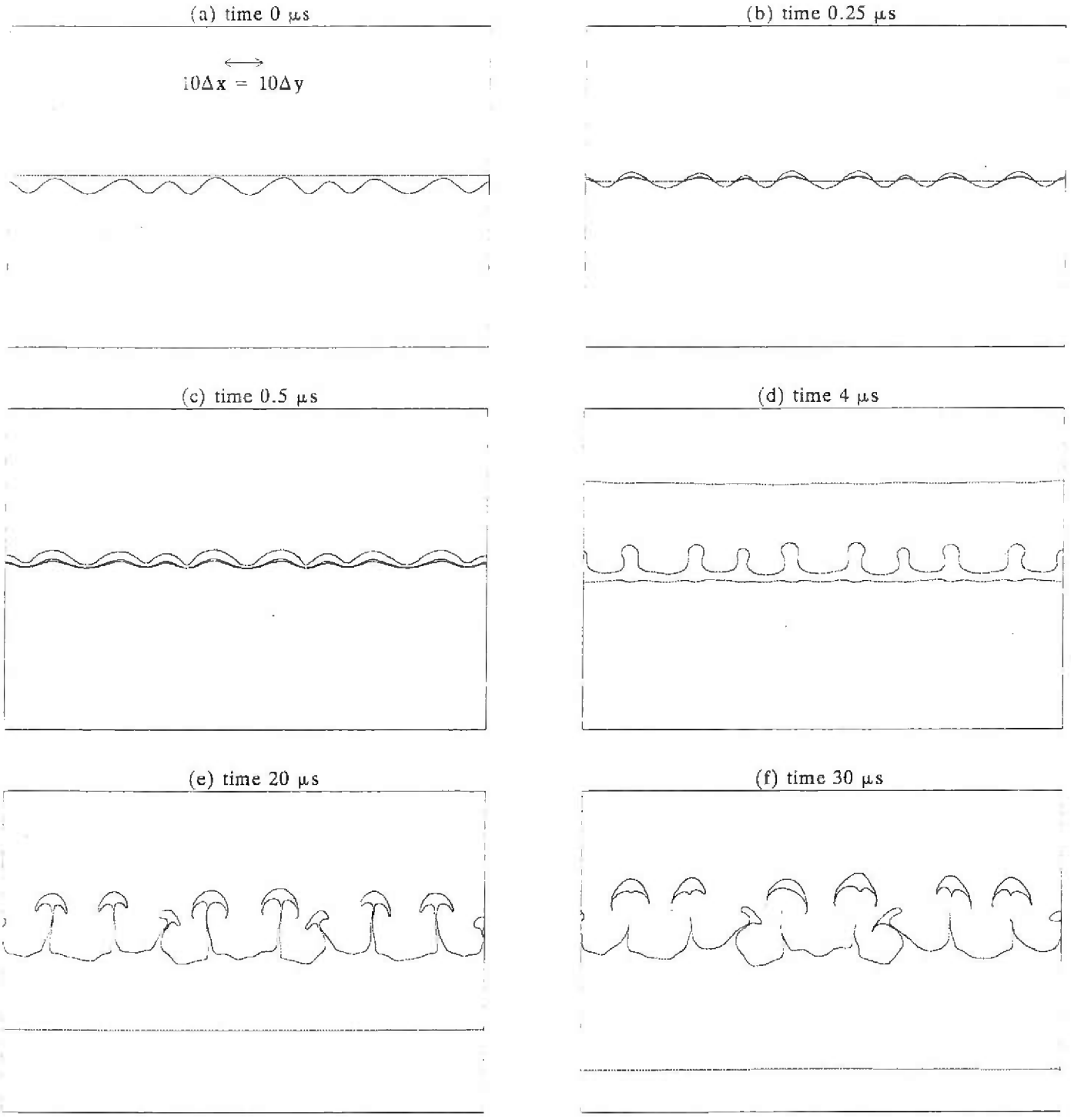


Fig. 4.5. A series of frames showing a shock in helium (γ = 1.63) colliding with an air (γ = 1.4) - helium interface. The pressure in front of the shock is 1 atm. and the pressure behind is 1000 atm.. The density of dry air at 25°C is 0.00118497 g/cc and the density of helium at the same temperature is 0.000101325 g/cc. The grid is 120x80.

A POSTERIORI ERROR ESTIMATION OF ADAPTIVE FINITE DIFFERENCE SCHEMES FOR HYPERBOLIC SYSTEMS

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, Munitions, and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Laboratory
Watervliet, NY 12189-4050

ABSTRACT. We describe several techniques that are based on Richardson's extrapolation for estimating discretization errors of finite difference solutions of one- and two- dimensional hyperbolic systems. These a posteriori error estimates are intended for use with adaptive mesh moving and local refinement procedures. Mesh moving algorithms produce nonuniform grids which necessitate special treatment of solution and error estimation techniques. The required adjustments are discussed using a two step MacCormack method as a model finite difference scheme. We also discuss automatic time step selection procedures and the effects of artificial viscosity. Extrapolation schemes that produce separate estimates of the temporal and spatial discretization errors are presented and we show how these may be used to control local mesh refinement. Several examples illustrating these techniques are presented.

1. INTRODUCTION. With the use of adaptive methods to solve time-dependent partial differential equations there exists a requirement to compute solutions on moving nonuniform grids. There is also a requirement to estimate the local discretization error as feedback to modify or refine the mesh. In this paper, we discuss the MacCormack finite difference scheme and a Richardson extrapolation-based error estimation procedure that was used in the adaptive algorithm of Arney [3] and Arney and Flaherty [4,5] to solve time-dependent hyperbolic systems in one and two space dimensions. Examples of other adaptive methods with these requirements are Rai and Anderson [30], Adjerid and Flaherty [2], Bell and Shubin [10], and Davis and Flaherty [14].

Finite difference methods use a mapping to transform the time and space variables from a moving nonuniform mesh to a stationary uniform mesh. The method used to compute the metrics of this transformation must be carefully chosen in order to preserve the stability, conservation, and accuracy of the scheme (cf., Thomas and Lombard [33,34] and Hindman [19,20]).

The MacCormack finite difference scheme has had wide use in solving Eulerian conservation laws for fluid dynamics. The recent use of artificial viscosity to make this scheme total variation diminishing (TVD) makes it more attractive as a general solver for problems with discontinuities (cf. Davis [13] and Roe [31]). The MacCormack scheme, our implementation of the differencing of the metric terms, adaptive selection of the time step, and the TVD artificial viscosity of Davis [13] are discussed in Section 2. The Richardson's extrapolation-based error estimation method produces a point wise approximation of the local discretization error which can be used to construct several global measures of the discretization error. Our error estimate and its implementation on a moving mesh are discussed in Section 3. In Section 4, we present computational results of solutions of hyperbolic problems. Computations were performed in one and two dimensions on stationary uniform and moving nonuniform grids. In Section 5 we discuss the utility of our methods, the computational results, and future work.

2. SOLUTION SCHEME. Consider the hyperbolic vector systems of conservation laws in two space dimensions

$$\vec{u}_t + \vec{f}_x(x,y,\vec{u},t) + \vec{g}_y(x,y,\vec{u},t) = 0, \ (x,y) \in D, \ t > 0 \tag{2.1}$$

$$\vec{u}(x,y,0) = \vec{u}_0(x,y), \ (x,y) \in D \cup \partial D, \tag{2.2}$$

with appropriate well-posed conditions on the boundary ∂D of a rectangular domain D.

We chose to implement the MacCormack finite difference scheme for hyperbolic problems because of its general applicability. The MacCormack scheme, like most higher-order methods, will suffer a reduction in order on a moving nonuniform grid. Despite this fact, proper mesh moving and node placement by an effective adaptive procedure provide enough efficiency and accuracy to compensate for this order reduction.

A. MacCormack Scheme

In order to discretize (2.1) we introduce a transformation

$$\xi = \xi(x,y,t), \ \eta = \eta(x,y,t), \ \tau = t, \tag{2.3}$$

from the physical (x,y,t) domain to a computational (ξ,η,τ) domain where a uniform rectangular grid will be used. Under this transformation (2.1) becomes

$$\vec{u}_\tau + \vec{u}_\xi \xi_t + \vec{u}_\eta \eta_t + \vec{f}_\xi \xi_x + \vec{f}_\eta \eta_x + \vec{g}_\xi \xi_y + \vec{g}_\eta \eta_y = 0. \tag{2.4}$$

The transformation metrics $(\xi_x, \xi_y, \xi_t, \eta_x, \eta_y, \eta_t)$ are related to the metrics $(x_\xi, x_\eta, x_\tau, y_\xi, y_\eta, y_\tau)$ by the identities

$$\xi_x = \frac{y_\eta}{J}, \quad \xi_y = \frac{-x_\eta}{J}, \quad \xi_t = \frac{(y_\tau x_\eta - s_\tau y_\eta)}{J}, \quad \eta_x = \frac{-y_\xi}{J},$$
$$\eta_y = \frac{x_\xi}{J}, \quad \eta_t = \frac{(y_\xi x_\tau - x_\xi y_\tau)}{J}, \quad J = y_\eta x_\xi - x_\eta y_\xi . \tag{2.5}$$

Using (2.5) in (2.4) gives

$$\vec{u}_\tau + \vec{u}_\xi \frac{(y_\tau x_\eta - x_\tau y_\eta)}{J} + \vec{u}_\eta \frac{(y_\xi x_\tau - x_\xi y_\tau)}{J} + \vec{f}_\xi \frac{y_\eta}{J} + \vec{f}_\eta \left(\frac{-y_\xi}{J}\right) + \vec{g}_\xi \left(\frac{-x_\eta}{J}\right) + \vec{g} \frac{x_\xi}{J} = 0. \tag{2.6}$$

This equation can be rewritten in another form in the original transformation metrics by further substitutions of (2.5) into (2.2) as

$$\vec{u}_\tau - \vec{u}_\xi(x_\tau \xi_x + y_\tau \xi_y) - \vec{u}_\eta(x_\tau \eta_x + y_\tau \eta_y) + \vec{f}_\xi \xi_x + \vec{f}_\eta \eta_x + \vec{g}_\xi \xi_y + g_\eta \eta_y = 0. \tag{2.7}$$

Some authors (cf., Hyman [22] and Thompson [35]) prefer to write this equation in still another form as

$$\vec{u}_\tau + \vec{u}_\xi \xi_t + \vec{u}_\eta \eta_t + \vec{f}_\xi \xi_x + \vec{g}_\xi \xi_y + \vec{g}_\eta \eta_y = 0. \tag{2.8}$$

A uniform space-time grid having mesh spacing $\Delta\xi \times \Delta\eta \times \Delta\tau$ is introduced onto the computational domain. The finite difference solution at $(\ell\Delta\xi, m\Delta\eta, n\Delta\tau)$ is referred to as $\vec{U}^n_{\ell,m}$. A similar notation is used for the fluxes $\vec{f}$ and $\vec{g}$ and the metrics (cf. Eq. (2.5)). The two-step MacCormack scheme [24] uses first-order forward temporal and spatial difference approximations in the predictor step, and first-order backward differences in the corrector step. The predicted solution $\vec{\bar{U}}^{n+1}_{\ell,m}$ satisfies

$$\vec{\bar{U}}^{n+1}_{\ell,m} = \vec{u}^n_{\ell,m} - \frac{\Delta\tau}{\Delta\xi}\left[(\vec{U}^n_{\ell+1,m} - \vec{U}^n_{\ell,m})(\xi_t)^n_{\ell,m} + (\vec{f}^n_{\ell+1,m} - \vec{f}^n_{\ell,m})(\xi_x)^n_{\ell,m}\right.$$
$$\left. + (\vec{g}^n_{\ell+1,m} - \vec{g}^n_{\ell,m})(\xi_y)^n_{\ell,m}\right] - \frac{\Delta\tau}{\Delta\eta}\left[\vec{U}^n_{\ell,m+1} - \vec{U}^n_{\ell,m})(\eta_t)^n_{\ell,m}\right. \tag{2.9}$$
$$\left. + (\vec{f}^n_{\ell,m+1} - \vec{f}^n_{\ell,m})(\eta_x)^n_{\ell,m} + (\vec{g}^n_{\ell,m+1} - \vec{g}^n_{\ell,m})(\eta_y)^n_{\ell,m}\right] .$$

The metrics $(\xi_x)^n_{\ell,m}$, etc. are computed by forward differences. The corrected solution $\overset{+}{U}{}^{n+1}_{\ell,m}$ satisfies

$$\overset{+}{U}{}^{n+1}_{\ell,m} = \frac{1}{2}\{\overset{+}{U}{}^{n}_{\ell,m} + \overset{+}{\underline{U}}{}^{n+1}_{\ell,m} - \frac{\Delta\tau}{\Delta\xi}[(\overset{+}{\underline{U}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{U}}{}^{n+1}_{\ell-1,m})(\xi_t)^{n+1}_{\ell,m} + (\overset{+}{\underline{f}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{f}}{}^{n+1}_{\ell-1,m})(\xi_x)^{n+1}_{\ell,m}$$

$$+ (\overset{+}{\underline{g}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{g}}{}^{n+1}_{\ell-1,m})(\xi_y)^{n+1}_{\ell,m}] - \frac{\Delta\tau}{\Delta\eta}[(\overset{+}{\underline{U}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{U}}{}^{n+1}_{\ell,m-1})(\eta_t)^{n+1}_{\ell,m} \tag{2.10}$$

$$+ (\overset{+}{\underline{f}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{f}}{}^{n+1}_{\ell,m-1})(\eta_x)^{n+1}_{\ell,m} + (\overset{+}{\underline{g}}{}^{n+1}_{\ell,m} - \overset{+}{\underline{g}}{}^{n+1}_{\ell,m-1})(\eta_y)^{n+1}_{\ell,m}]\},$$

with metrics computed by backward differences. The notation $\overset{+}{\underline{f}}{}^{n+1}_{\ell,m}$ denotes $f(\overset{+}{\underline{U}}{}^{n+1}_{\ell,m})$, etc. The use of first forward and backward difference approximations for the metrics implies that the transformation from the computational to the physical domain is piecewise trilinear in space and time for the predictor and corrector steps. Such low order difference approximations are responsible for reducing the orders of the MacCormack scheme. A smoother transformation and the use of higher-order difference approximations of the metrics could be used to maintain second-order accuracy.

It was shown by Hindman [19,20] that this differencing of Equation (2.6) produces consistent approximations. Therefore, a uniform flow solution is maintained. Other conservative forms for the transformed equations were investigated by Hindman [19] and found to be less efficient or needing special differencing of the metrics for computing consistent approximations.

Equation (2.4) is conservative on a moving mesh. We show this for a one-dimensional scalar conversation law by investigating the Rankine-Hugoniot jump conditions across a shock discontinuity. Consider a conservation law in the form

$$\frac{\partial}{\partial t}\left(\int_{-\infty}^{\infty} u\, dx\right) + f(u)\Big|_{-\infty}^{\infty} = 0 . \tag{2.11}$$

The jump conditions for a discontinuity at $x = s(t)$ satisfy

$$\dot{s} = \frac{[f]}{[u]} , \tag{2.12}$$

where $[q]$ indicates the jump in q and $\dot{s} = \frac{ds}{dt}$ denotes the shock velocity [37].

A conservation law on a moving mesh produced by a transformation of variables to a uniform stationary mesh satisfies

$$\left(\frac{\partial}{\partial\tau} + \xi_t \frac{\partial}{\partial\xi}\right)\int_{-\infty}^{\infty} ux_\xi d\xi + f(u)\Big|_{-\infty}^{\infty} = 0 . \tag{2.13}$$

Assuming the existance of a shock discontinuity $\xi = r(\tau)$ gives

$$\dot{r}x_\xi[u] + \int_{-\infty}^{r} (ux_\xi)_\tau d\xi + \int_{r}^{\infty} (ux_\xi)_\tau d\xi + f(u)\Big|_{-\infty}^{\infty} = 0 \ . \tag{2.14}$$

Using the chain rule provides an integrable form

$$\dot{r}x_\xi[u] + \int_{-\infty}^{r} (ux_\tau - f)_\xi d_\xi + \int_{r}^{\infty} (ux_\tau - f)_\xi d\xi + f(u)\Big|_{-\infty}^{\infty} = 0 \ . \tag{2.15}$$

Integration of this equation gives jump conditions in the computational domain as

$$\dot{r}x_\xi[u] - [f] + x_\tau[u] = 0 \ . \tag{2.16}$$

Since $s(t)$ and $r(\tau)$ are related by

$$\dot{s} = \dot{r}\, x_\xi + x_\tau \ , \tag{2.17}$$

the appropriate jump condition (2.12) is recovered.

B. Variable Time step.

The explicit MacCormack scheme has a stability restriction that limits the time step allowed for a given spatial mesh. For efficient computation, the time step should be adaptively set close to the maximum allowed by the Courant, Friedrichs, Lewy theorem [27]. Thus, we choose

$$\Delta\tau = \frac{0.8}{2\sqrt{2}\ \max(\psi,\omega)} \ . \tag{2.18}$$

The computational mesh has been selected to have spacing $\Delta\xi = \Delta\eta = 1$ and the constant 0.8 provides a twenty-percent margin of safety. The quantities ψ and ω are the spectral radii of one-dimensional conservation laws on moving meshes, i.e.,

$$\psi = \max\left[(\lambda_i - x_\tau)\xi_x + (\rho_i - y_\tau)\xi_y\right] \ , \tag{2.19a}$$

$$\omega = \max\left[(\lambda_i - x_\tau)\eta_x + (\rho_i - y_\tau)\eta_y\right] \ , \tag{2.19b}$$

where λ_i and ρ_i are eigenvalues of $\vec{f}_{\vec{u}}(\vec{u})$ and $\vec{g}_{\vec{u}}(\vec{u})$. These eigenvalues and the metrics in (2.19) are evaluated at the beginning of each time step.

C. Artificial Viscosity

The MacCormack scheme, being a second-order accurate centered scheme, produces spurious oscillations near discontinuities. In order to eliminate or reduce these oscillations, artificial viscosity or dissipation is added to the solution to diffuse the discontinuity. The viscosity is often problem dependent, and considerable "fine tuning" is usually needed to balance the effects of the spurious oscillations and diffusion [23].

We use an artificial viscosity model due to Davis [13] which is not problem dependent and only requires knowledge of ψ and ω. This artificial viscosity model is designed to convert the MacCormack scheme into a total variation diminishing (TVD) scheme in one-dimension. A scheme is TVD if the total variation of the solution to an initial value problem is non-increasing in time. Recent research efforts have resulted in the development of other second-order accurate TVD schemes (cf., Osher and Chakravarthy [28] and Warming and Beam [36]).

The artificial viscosity of Davis [13] is based on a flux limiter that does not depend on explicitly determining the upwind direction and, with a recent modification by Roe [31], does not affect the region of stability of the MacCormack scheme. Because the MacCormack scheme also does not determine the upwing direction, the combined use of the MacCormack scheme and Davis's artificial viscosity is computationally simpler to perform than many other TVD schemes. The artificial viscosity terms are calculated from the solution data at the beginning of the time step. For two dimensional problems separate dissipative terms are calculated in the ξ and η directions respectively.

3. ERROR ESTIMATION. Accurate a posteriori error estimation is an integral part of an adaptive software system. Error estimation can be the most expensive part of an adaptive procedure and an important goal is to find accurate and inexpensive ways of estimating the discretization error (cf., Babuska, et al. [8,9]). The error estimation technique is dependent on many factors, including the type of solver used in the algorithm, the type of error to be determined, and the norm in which the error estimate is to be measured. It is most desirable to have a procedure that provides pointwise estimates of the error which can then be used to find estimates in several local and global norms.

Mesh nonuniformity affects the accuracy and convergence of numerical schemes and error estimation. The effects of the mesh on the solution scheme have been studied by Ciment [12], Fritts [17], Hoffman [21], Osher and Sanders [29], Sanders [32], and Mastin [25]. Error analysis seems to be more natural and further developed for finite element schemes, especially for elliptic and parabolic problems (cf., Adjerid and Flaherty [1,2], Zienkiewicz et. al. [38,39], and Babuska and Rheinboldt [6,7]), where relatively inexpensive local calculations are used to provide accurate global spatial error estimates. More study needs to be done to find less expensive and more accurate error estimates for finite difference schemes for hyperbolic problems.

We calculate the local temporal and spatial portions of the discretization error, using an algorithm based on Richardson extrapolation. Flaherty and Moore [15,16] and Berger and Oliger [11] also use Richardson extrapolation to estimate error on uniform meshes for their local mesh refinement algorithms.

A. Richardson Extrapolation Error Estimation

We develop the error estimate for the second-order MacCormack scheme for a linear scalar problem in two dimensions. Separate pointwise estimates at a general spatial node i, at time t, for the local temporal error $E_i^t(t)$ and local spatial error $E_i^s(t)$ are obtained with two different extrapolation procedures.

Consider a uniform mesh with spacing $\Delta x \times \Delta y$ and time step Δt. Let the exact solution at node i and time t be denoted as $u_i(t)$, the numerical solution by the MacCormack scheme at the same point and time as $U_i(t;\Delta x,\Delta y,\Delta t)$ and the MacCormack finite difference operator as $L(\Delta x,\Delta y,\Delta t)$, i.e.,

$$U_i(t + \Delta t;\ \Delta x,\Delta y,\Delta t) = L(\Delta x,\Delta y,\Delta t)U_i(t;\Delta x,\Delta y,\Delta t). \tag{3.1}$$

Assume that the local error has a Taylor's series expansion of the form

$$u_i(t) - U_i(t;\Delta x,\Delta y,\Delta t) = \Delta t\left[c_1\Delta t^2 + c_2\Delta x^2 + c_3\Delta y^2 + \ldots\right], \tag{3.2}$$

where the constants c_1, c_2, c_3,... are independent of the mesh spacing.

To estimate the spatial component of the error, we calculate a solution on a mesh of double spatial size ($2\Delta x \times 2\Delta y$) with the same time step (Δt). The local error on this mesh satisfies

$$u_i(t + \Delta t) - U_i(t + \Delta t;2\Delta x,2\Delta y,\Delta t) = \Delta t\left[c_1\Delta t^2 + 4c_2\Delta x^2 + 4c_3\Delta y^2 + ..\right]. \tag{3.3}$$

Subtracting (3.3) from (3.2), and neglecting higher-order terms, we obtain an expression for the leading team in the spatial portion of the local error for the MacCormack scheme on the $\Delta x \times \Delta y \times \Delta t$ mesh as

$$\begin{aligned} E_i^s(t + \Delta t) &:= \Delta t\left[c_2\Delta x^2 + c_3\Delta y^2\right] \\ &= \tfrac{1}{3}\left[U_i(t + \Delta t;2\Delta x,2\Delta y,\Delta t) - U_i(t + \Delta t;\Delta x,\Delta y,\Delta t)\right]. \end{aligned} \tag{3.4}$$

Similarly, an estimate of the temporal portion of the local error, $E_i^t(t + \Delta t)$, can be calculated by computing another solution on the $\Delta x \times \Delta y$ spatial mesh using two time steps of $\Delta t/2$, subtracting this result from (3.2), and retaining leading order terms as

$$\begin{aligned} E_i^t(t + \Delta t) &:= \Delta t\left[c_1\Delta t^2\right] \\ &= \tfrac{4}{3}\left[U_i\left(t + \Delta t;\Delta x,\Delta y,2\left(\tfrac{\Delta t}{2}\right)\right) - U_i(t + \Delta t;\Delta x,\Delta y,\Delta t)\right]. \end{aligned} \tag{3.5}$$

The leading terms of the local error at node i, and time $t + \Delta t$, is

$$E_i(t + \Delta t) = E_i^t(t + \Delta t) + E_i^s(t + \Delta t) . \tag{3.6}$$

There are several disadvantages to this technique that should be noted: (i) the error cannot be calculated for nodes on or adjacent to the boundary; (ii) the solution must be smooth enough for the c_1, c_2, and c_3 to exist; (iii) the error estimation costs approximately three times more to compute than the solution; and (iv) the mesh must be uniform. Equation (3.6) may still be useful as a mesh refinement or motion indicator even in situations where jumps in the solution render it invalid as an estimate of the error.

Richardson's extrapolation can be done in a more classic manner provided that we are willing to forego separate spatial and temporal error estimates. We illustrate the method for a one-dimensional problem. In this case, the error at node i in a solution on a mesh having spacing $\Delta x \times \Delta t$ is estimated by calculating a second solution on a mesh with spacing $\Delta x/2$ using two time steps of $\Delta t/2$. According to (3.3) restricted to one-dimension, the local error on this mesh satisfies

$$u_i(t + \Delta t) - U_i\big(t + \Delta t;\Delta x/2,2(\Delta t/2)\big) = \Delta t\big[c_1\Delta t^2/4 + c_2\Delta x^2/4 + \ldots\big]. \tag{3.7}$$

Subtracting (3.7) from (3.3) and neglecting higher order terms we can obtain error estimates for either $U_i(t + \Delta t;\Delta x,\Delta t)$ or $U_i\big(t + \Delta t;\Delta x/2,2(\Delta t/2)\big)$ provided that node i is common to both meshes. Our adaptive method carries the fine grid solution forward in time; thus, we estimate its error as

$$\begin{aligned} E_i(t + \Delta t) &= \tfrac{1}{4}\,\Delta t(c_1\Delta t^2 + c_2\Delta x^2) \\ &= \tfrac{1}{3}\big[U_i\big(t + \Delta t;\tfrac{\Delta x}{2}\,2\big(\tfrac{\Delta t}{2}\big)\big) - U_i(t + \Delta t;\Delta x,\Delta t)\big]. \end{aligned} \tag{3.8}$$

Using this procedure the error can now be calculated at nodes adjacent to boundaries. Even though this error estimate costs four times more to compute that the solution, we only incur this overhead in the first level of refinement. No additional cost is incurred if portions of the mesh have to be refined, because the solution on the refined mesh has already been computed and stored while estimating the error for the coarser parent mesh.

B. Error Estimation for a Moving Nonuniform Mesh

Nonuniformity of the mesh changes the discretization error of the MacCormack scheme. For simplicity, we will determine this error and analyze its effects on the Richardson extrapolation error estimation using a linear scalar problem in one space dimension.

$$u_t + bu_x = 0 . \tag{3.9}$$

The local error for the MacCormack method on a one-dimensional moving nonuniform mesh is

$$u_i(t + \Delta t) - U_i(t + \Delta t;\Delta x, \Delta t) = \Delta t\Big[- \frac{b}{4}\left(\Delta x_r^{n+1} - \Delta x_\ell^n\right)u_{xx} - \Delta t\ b^2\Big(1 - \Big(\frac{\Delta x_r^{n+1}}{\Delta x_\ell^n}\Big)u_{xx}\Big) + c_1\Delta t^2 + c_2\Delta x^2\Big] , \tag{3.10}$$

where, Δx_ℓ^n and Δx_r^n are the mesh sizes on the left and right of node i at time step n, respectively, and $\Delta x = \max\left(\Delta x_r^{n+1}, \Delta x_\ell^n\right)$. On the moving nonuniform mesh, both the temporal and spatial error components contain second order terms whereas the error on a uniform mesh is third order. The previous analysis can be used to show that the leading component of the temporal error is

$$E_i^t(t + \Delta t) := \Delta t\Big[- \Delta t\ b^2\Big(1 - \frac{\Delta x_r^{n+1}}{\Delta x_\ell^n}\Big)u_{xx}\Big] = 2\Big[U_i\Big(t + \Delta t;\Delta x, 2\Big(\frac{\Delta t}{2}\Big)\Big) - U_i(t + \Delta t;\Delta x;\Delta t)\Big]. \tag{3.11}$$

Calculation of the spatial portion of the error is more difficult since the temporal portion of the error does not cancel upon subtraction of solutions calculated on two spatially different meshes. We overcome this difficulty and also greatly simplify the procedure in two dimensions by constraining the mesh to maintain double size increments for special nodes of the moving coarse mesh. This constrained grid structure consists of a coarse mesh, shown with darker lines in Figure 1, containing properly nested fine cells created by binary division of the sides of the coarse cells, shown by lighter lines in Figure 1. The vertices of the coarse cells are denoted as "independent moving nodes". Error estimates are calculated for these nodes. The remaining nodes in the mesh of Figure 1 are "dependent moving nodes" which must be moved to maintain the constrained grid structure. A solution is computed for these "dependent moving nodes," but no error estimate is obtained.

For the "independent moving nodes", the spatial error calculation can proceed as for a uniform mesh; therefore, the local spatial error estimate is

$$E_i^s(t + \Delta t) = \Delta t\Big[- \frac{b}{4}\left(\Delta x_r^{n+1} - \Delta x_\ell^n\right)u_{xx}\Big] = U_i(t + \Delta t;\Delta x,\Delta t) - U_i(t + \Delta t;2\Delta x,\Delta t). \tag{3.12}$$

The above analysis extends directly to two dimensions; hence, we have a Richardson extrapolation-based procedure of estimating error on a moving nonuniform grid. In practice, we test the need for local uniformity and, if found, use formulas (3.4-6) to compute error estimates.

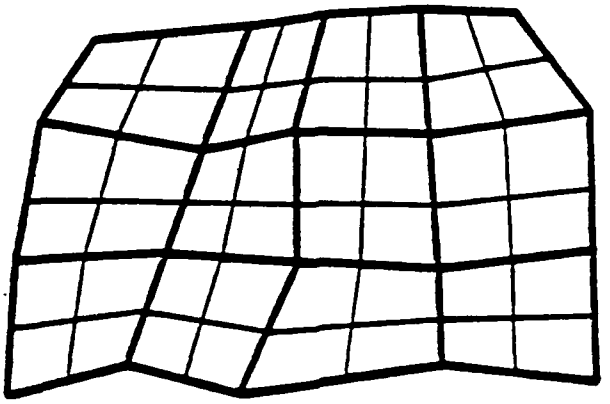


Figure 1. Spatial structure of the moving coarse mesh (bold lines) with embedded fine mesh (fine lines) used for the error estimation.

Error estimation for systems of equations involves the use of a vector norm at node i and time t. The examples of Section 4 use the maximum norm, i.e.,

$$E_i(t) := \max_{1 \le j \le N} |E_{ij}(t)|, \tag{3.13}$$

where N is the number of equations in the system and $E_{ij}(t)$ is the local error estimate for the jth component of the solution vector at node i.

4. COMPUTATIONAL EXAMPLES. The solution and local error estimation procedures are applied to four examples. In Example 4.1, we demonstrate the capability of the MacCormack scheme with Davis' (TVD) artificial viscosity on a moving nonuniform mesh. In Example 4.2, we investigate a one-dimensional problem using a modified form of the error estimate (3.8,9). Examples 4.3 and 4.4 illustrate the performance of the error estimation procedure on a problem having a smooth solution and one with a jump in the first derivative, respectively. We investigate the accuracy and convergence of the local error estimator by determining an effectivity index

$$\theta = \frac{\|E\|_1}{\|e\|_1}. \tag{4.1a}$$

at a fixed time t for several different meshes and different adaptive strategies. Here e and E are the exact and estimated errors, respectively.

The L_1 norm,

$$\|E\|_1 := \iint E dxdy \tag{4.1b}$$

is obtained by assuming E to be a piecewise constant function.

Example 4.1. Consider the initial-boundary value problem

$$u_t - yu_x + xu_y = 0, \quad t > 0, \quad -1.2 \le x \le 1.2, \; -1.2 \le y \le 1.2, \tag{4.2}$$

$$u(x,y,0) = \begin{cases} 0 & , \quad \text{if } \left(x - \frac{1}{2}\right)^2 + 1.5y^2 \ge \frac{1}{16} \\ 1 - 16\left(\left(x - \frac{1}{2}\right)^2 + 1.5y^2\right) & , \quad \text{otherwise,} \end{cases} \tag{4.3}$$

and

$$u(1.2,y,t) = u(-1.2,y,t) = u(x,-1.2,t) = u(x,1.2,t) = 0 . \tag{4.4}$$

$$u(x,y,t) = \begin{cases} 0 & , \quad \text{if } C < 0 \\ C & , \quad \text{if } C \ge 0, \end{cases} \tag{4.5a}$$

where

$$C = 1 - 16\left(\left(x\cos t + y\sin t - \frac{1}{2}\right)^2 + 1.5(y\cos t - x\sin t)^2\right). \tag{4.5b}$$

Equations (4.5) represent a moving elliptical cone rotating counter-clockwise around the origin with period 2π. This problem was proposed as a test problem by Gottlieb and Orszag [18] and was used as a test problem in a survey by McRae et al. [26].

We show the sequence of meshes that were generated at t = 0, 1.6, and 3.2 using the adaptive mesh moving method of Arney and Flaherty [4] in Figures 2, 3, and 4, respectively. Arney and Flaherty's [4] mesh moving method utilizes the error estimates of Section 3 to concentrate the mesh in the high-error region beneath the cone and to follow it as it rotates. It also increases the accuracy of the solution and reduces oscillations in the wake following the cone. However, small oscillations are still present. Next we solve this problem with the same moving mesh technique, by using Davis' [13] artificial viscosity with the MacCormack scheme. Surface and contour plots of solutions with and without artificial viscosity are shown in Figures 5 and 6. There is no artificial wake behind the cone when artificial viscosity is used. However, the artificial viscosity slightly diffuses the cone, widening its base and reducing its peak from 1.0 to 0.88.

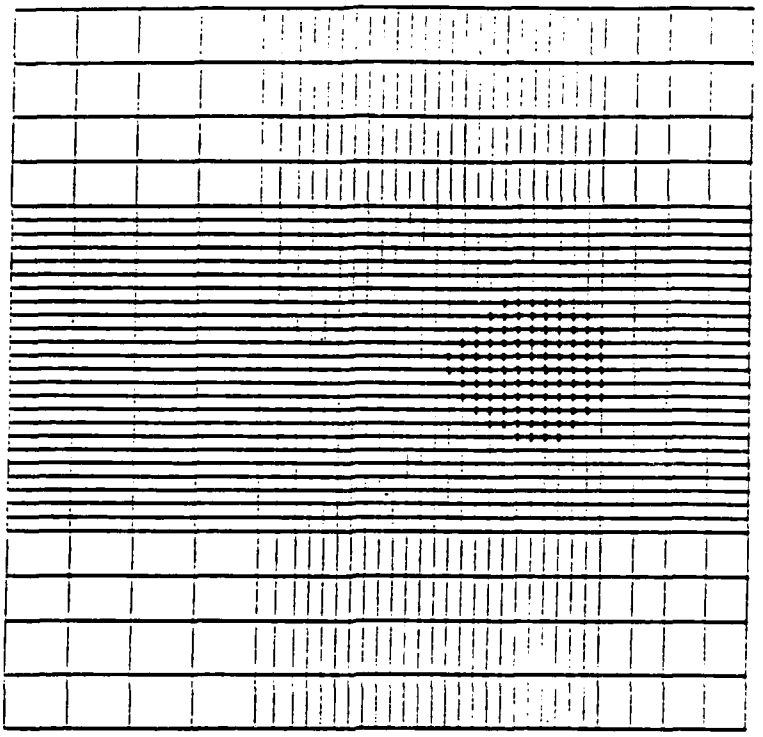


Figure 2. Initial mesh for Example 4.1.

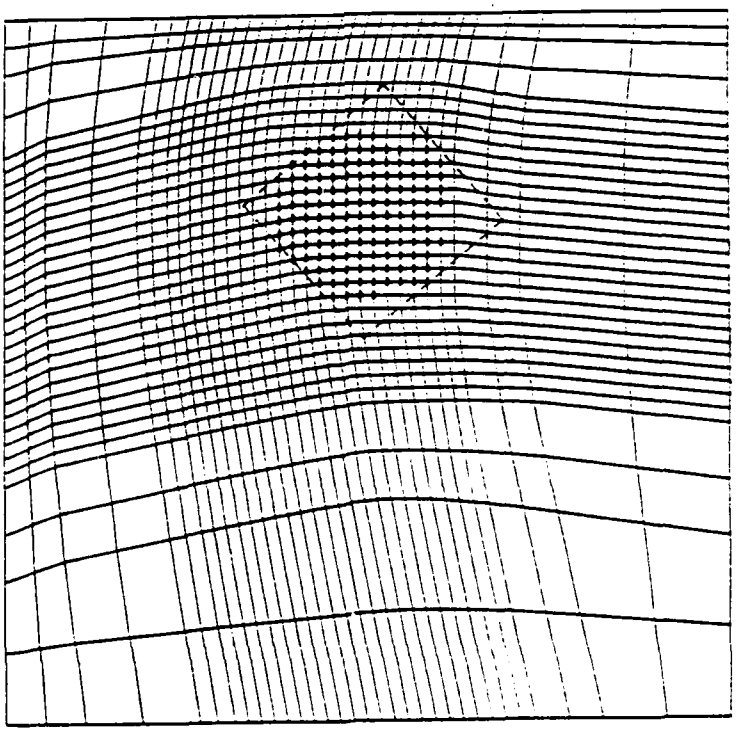


Figure 3. Mesh of Example 4.1 at t = 1.6. Nodes are moving with the rotating cone.

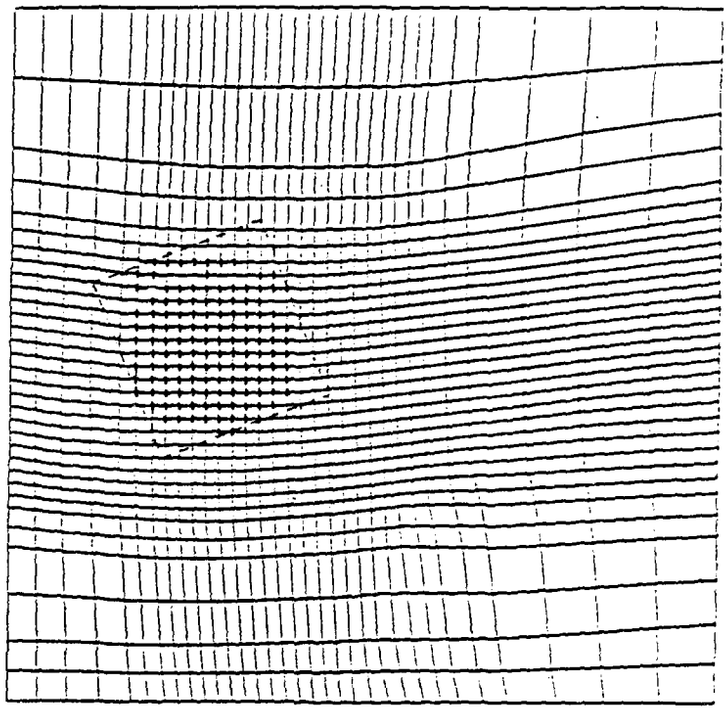


Figure 4. Mesh of Example 4.1 at t = 3.2. Nodes have moved with the cone for one-half rotation.

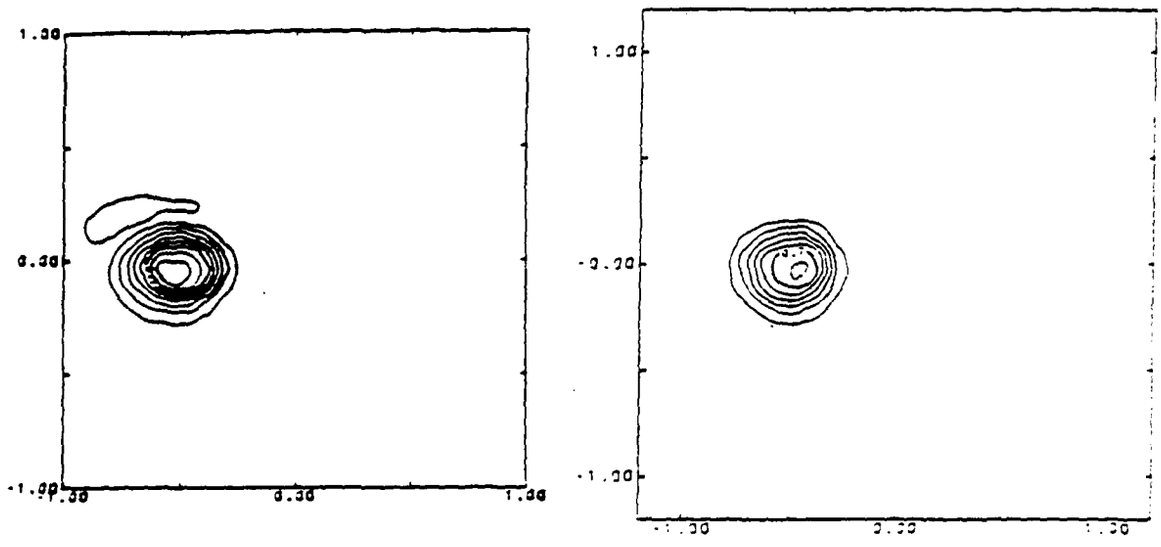


Figure 5. Contour plots of the solutions of Example 4.1 on a moving mesh without artificial viscosity (left) and with artificial viscosity (right) at t = 3.2.

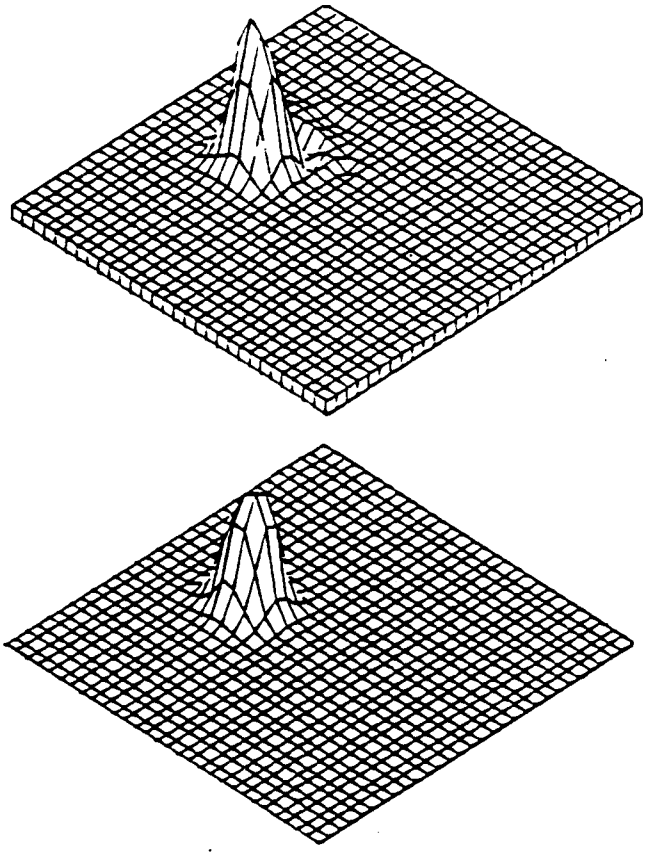


Figure 6. Surface plots of the solutions of Example 4.1 on a moving mesh without artificial viscosity (top) and with artificial viscosity (bottom) at t = 3.2.

Example 4.2. We consider an application of the direct Richardson's extrapolation error estimation procedure (3.9) to the one-dimensional linear scalar equation

$$u_t + u_x = 0, \quad t > 0 \ , \ 0 < x < 0.8, \tag{4.6}$$

with initial and Dirichlet boundary conditions specified so that the exact solution is

$$u(x,t) = \frac{1}{2}\left[1 - \tanh 100\ (x - t - 0.2)\right]. \tag{4.7}$$

This solution is a relatively steep wave that moves at unit speed across the domain.

We solved this problem for one time step on seven different uniform meshes having N computational cells per time step in order to investigate accuracy and convergence of the error estimate. Table 1 shows the results obtained from these calculations. The effectivity ratio appers to be converging to unity.

We also solved this problem using Arney and Flaherty's [5] adaptive local refinement procedure on a base mesh having $\Delta x = \Delta t = 0.1$ with a local error tolerance of 1/128. The mesh created by the local refinement algorithm is shown in Figure 7 and the solutions computed at each base time step are shown in Figure 8.

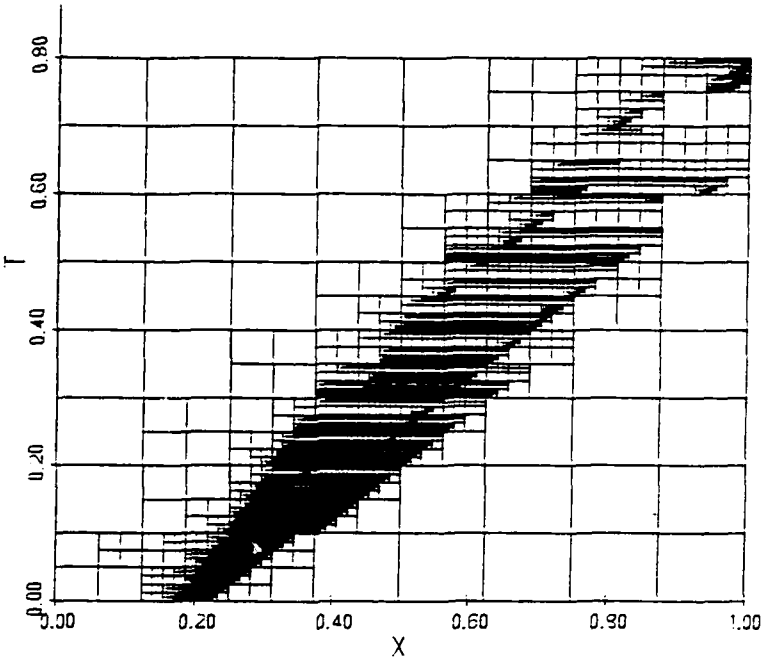


Figure 7. Adaptive Mesh of Example 4.2.

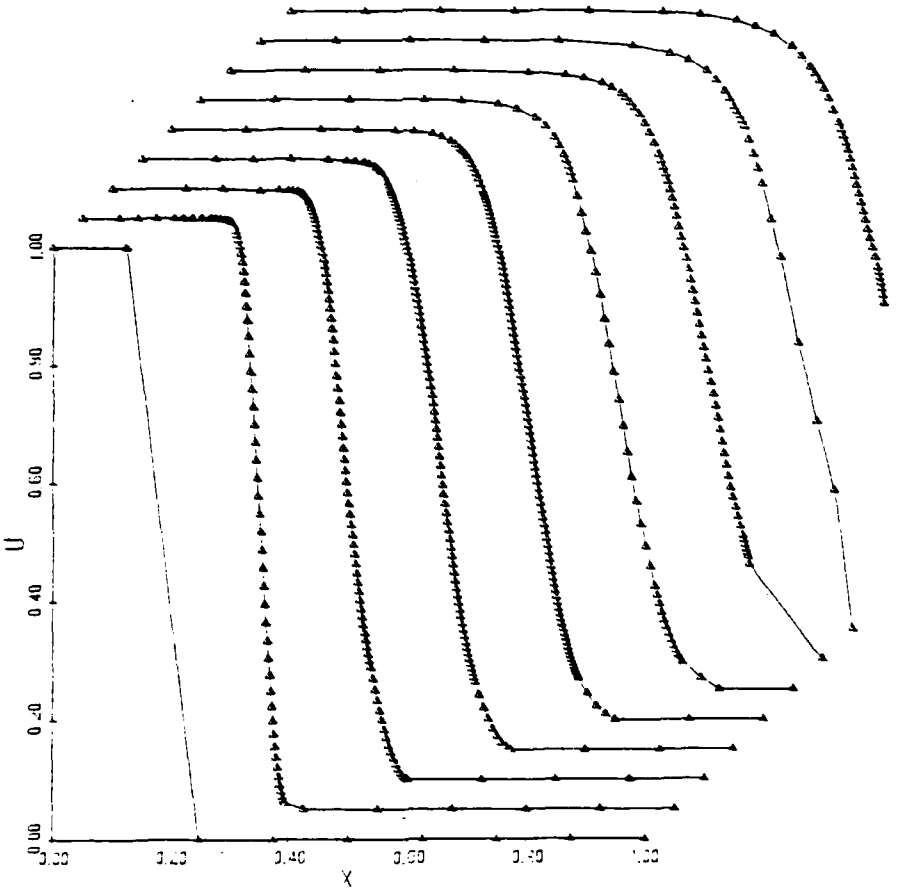


Figure 8. Solutions at base time steps for the adaptive local refinement procedure of Example 4.2.

The adaptive composite mesh of Figure 7 shows a distinct pattern associated with using the MacCormack scheme with Arney and Flaherty's [5] local refinement strategy. Spurious oscillations of the solution on the base mesh cause several levels of refinement which drastically reduce the base mesh spacing at the beginning of the each base time step. However, once these oscillations have been controlled the need for refinement is reduced at the later stages of the adaptive procedure. This situation could be alleviated by including an artificial viscosity model with the MacCormack scheme, as in Example 4.1.

Δt	N	Exact Error $\|e\|_1$	Estimated Error $\|E\|_1$	Effectivity Ratio θ
0.1	8	$.352 \times 10^{-1}$	$.467 \times 10^{-2}$	0.133
0.05	16	$.132 \times 10^{-1}$	$.234 \times 10^{-2}$	0.177
0.025	32	$.236 \times 10^{-2}$	$.106 \times 10^{-2}$	0.449
0.0125	64	$.256 \times 10^{-3}$	$.138 \times 10^{-3}$	0.539
0.00625	128	$.380 \times 10^{-4}$	$.294 \times 10^{-4}$	0.773
0.00312	256	$.403 \times 10^{-5}$	$.303 \times 10^{-5}$	0.752
0.00156	512	$.661 \times 10^{-6}$	$.538 \times 10^{-6}$	0.814

Table 1. Exact and estimated errors for different mesh sizes for Example 4.2.

Example 4.3. Consider the linear scalar hyperbolic differential equation

$$u_t + 2u_x + 2u_y = 0, \quad t > 0 , \quad 0.2 \leq x \leq 1.2 , \quad 0 > y \leq 1 , \tag{4.8}$$

with initial conditions

$$u(x,y,0) = \frac{\left(1 - \tanh 3(x - .1y + .1)\right)}{2} , \tag{4.9}$$

and with Dirichlet boundary conditions specified so that the exact solution of this problem is

$$u(x,y,t) = \frac{\left(1 - \tanh 3(x - .1y - 1.8t + .1)\right)}{2} . \tag{4.10}$$

This solution is a smooth wave that moves at an angle of 45 degrees across the domain. The problem was selected to show the convergence and accuracy of the Richardson extrapolation error estimation procedures (3.4,5) and (3.11,12.) We solve (4.8,9) for one time step, $\Delta t = 0.012$, on eight different meshes. The mesh strategy of each calculation is described as follows:

1) a stationary uniform (10 × 10) rectangular mesh,
2) a stationary uniform (20 × 20) rectangular mesh,
3) a stationary uniform (40 × 40) rectangular mesh,
4) a stationary uniform (60 × 60) rectangular mesh,
5) a stationary (40 × 40) mesh of nonuniform quadrilateral cells,
6) a moving (20 × 20) mesh with uniform rectangles,
7) a moving (20 × 20) mesh of nonuniform quadrilateral cells,
8) a moving (40 × 40) mesh of nonuniform quadrilateral cells.

Table 2 shows the results obtained from these calculations by comparing the exact errors and the effectivity indices for the eight strategies.

Strategies 1-4 show the convergence of the error estimates on uniform meshes as the number of nodes increase. These errors show a rate of convergence of $O(\Delta x^2, \Delta y^2)$, which is predicted in Eq. (3.2). Comparison of the errors of Strategies 3 and 5 show the error is cut in half by computing with a better nonuniform stationary mesh. Further comparison of Strategies 5 and 7 shows another reduction of error by half when the mesh is properly moved. The nonuniformity of the mesh in Strategies 5, 7, and 8, produces little change in the effectiveness of the error estimation. These nonuniform mesh computations indicate a convergence rate $O(\Delta x^{1.32}, \Delta y^{1.32})$.

Mesh Strategy (from above)	Exact error $\|e\|_1$	Estimated error $\|E\|_1$	Effectivity ratio θ
1	0.0111	0.0071	0.64
2	0.00370	0.00318	0.86
3	0.000942	0.000908	0.96
4	0.000367	0.000368	1.00
5	0.000399	0.000418	1.04
6	0.00136	0.00124	0.91
7	0.000411	0.000370	0.90
8	0.000167	0.000156	0.94

Table 2. Exact and estimated errors for different mesh strategies for Example 4.3.

Example 4.4. Consider the linear scalar hyperbolic differential equation

$$u_t + u_x + 0.25u_y = 0, \quad t > 0\ ,\ 0.2 \le x \le 1.2\ .\ 0 \le y \le 1\ , \tag{4.11}$$

with initial conditions

$$u(x,y,0) = \begin{cases} 0 & , \quad \text{if } y < -4x + 1.2 \\ 0.8 & , \quad \text{if } y > -4x + 1.6 \\ -8x - 2y + 3.2\ , & \quad \text{otherwise,} \end{cases} \tag{4.12}$$

and with Dirichlet boundary conditions

$$u(x,y,0) = \begin{cases} 0 & , \quad \text{if } y - 0.25t < -4(x - t) + 1.2 \\ 0.8 & , \quad \text{if } y - 0.25t > -4(x - t) + 1.6 \\ -8(x - t) - 2(y - 0.25t) + 3.2\ , & \quad \text{otherwise.} \end{cases} \tag{4.13}$$

The solution of this problem is an oblique ramp-like wave front that moves at an angle of 14 degrees across the domain. The solution has a jump in its first partial derivatives at the top and bottom edges of the wave front. We expect some difficulty in estimating the error near locations where the derivatives jump. In the region of the front itself the gradient of the solution is constant and there is no error in the solution or in the error estimate.

We solved this problem for one time step, $\Delta t = 0.015$, for the following six mesh strategies:

1) a stationary uniform (12 × 12) rectangular mesh,
2) a stationary uniform (24 × 24) rectangular mesh,
3) a stationary uniform (48 × 48) rectangular mesh,
4) a stationary uniform (64 × 64) rectangular mesh,
5) a stationary (24 × 24) mesh of nonuniform quadrilateral cells,
6) a moving (24 × 24) mesh of nonuniform quadrilateral cells.

Table 3 shows the results of these calculations.

Mesh strategy	Exact error $\|e\|_1$	Estimated error $\|E\|_1$	Effectivity ratio θ
1	0.0058	0.0016	0.28
2	0.00275	0.00110	0.40
3	0.000866	0.000479	0.55
4	0.000400	0.000222	0.56
5	0.00144	0.00078	0.54
6	0.000720	0.000349	0.49

Table 3. Exact and estimated errors for different mesh strategies of Example 4.4. The error estimate is inaccurate but the solution appears to be converging.

The results are once again as expected. The error estimate of this problem with a jump in the derivative is not as accurate as the smooth solution of Example 4.3. However, the error estimate still shows signs of converging to the exact error in L_1 for the uniform meshes of Strategies 1-4. Once again the better nodal placement of the initial mesh by the mesh generator of Arney [3] reduces the error by half from a uniform mesh. Also, moving the mesh by the method of Arney and Flaherty [4] reduces the error by half again.

5. CONCLUSION. We have shown that MacCormack's finite difference scheme and error estimation based on Richardson's extrapolation can be used on moving grids with local refinement. With proper computation of the transformation metrics and the use of TVD artificial viscosity, the MacCormack scheme is stable and is able to solve problems with sharp discontinuities.

The examples we have presented demonstrate the utility of these methods and also point out their shortcomings. Of particular concern is the lack of any error estimation near the boundaries, the poor error estimation near discontinuities, and the need to constrain the mesh to obtain any accurate error estimation. These problems must be solved in order to effectively utilize this solution scheme and error estimation procedure with an adaptive technique.

LITERATURE CITED

1. Adjerid, S., 'Adaptive Finite Element Methods for Time Dependent Partial Differential Equations,' Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY, 1985

2. Adjerid, S. and Flaherty, J.E., 'A Moving Finite Element Method for Time Dependent Partial Differential Equations with Error Estimation and Refinement,' SIAM J. Numer. Anal., 23 (1986), pp. 778-795.

3. Arney, D.C., 'An Adaptive Mesh Algorithm for Solving Systems of Time Dependent Partial Differential Equations,' Ph.D Thesis, Rensselaer Polytechnic Institute, Troy, NY, 1985.

4. 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.

5. Arney, D.C. and Flaherty, J.E., 'An Adaptive Method With Mesh Moving and Local Mesh Refinement for Time Dependent Partial Differential Equations,' Trans. Fourth Army Conf on Appl. Math. and Comput., (1987), pp. 1115-1141.

6. Babuska, I. and Rheinboldt, W.C., 'Computational Error Estimates and Adaptive Processes for Some Nonlinear Structural Problems', Comp. Meths. Appl. Mech. Engrg., 34 (1982), pp. 895-937.

7. Babuska, I. and Rheinboldt, W.C., 'A Survey of a Posteriori Error Estimates and Adaptive Approaches in the Finite Element Method', Institue for Physical Science and Technology Tech. Note BN-981, 1982.

8. Babuska, I., Zienkiewicz, O.C., Gago, J., and de A. Olivera, E.R., (Eds.), Accuracy Estimates and Adaptive Refinements in Finite Element Computations, Wiley and Sons, London, 1986.

9. Babuska, I., Chandra, J., and Flaherty, J.E. (Eds.), Adaptive Computational Methods for Partial Differtial Equations, SIAM, Philadelphia, 1983.

10. Bell, J.B. and Shubin, G.R., 'An Adaptive Grid Finite Difference Method for Conservation Laws,' J. Comput. Phys., 52 (1983), pp. 569-591.

11. Berger, M. and Oliger, J., 'Adaptive Mesh Refinement for Hyperbolic Partial Differential Equations,' J. Comput. Phys., 53 (1984), pp. 484-512.

12. Ciment, M., 'Stable Difference Schemes with Uneven Mesh Spacings', Math. Comp., 25 (1971), pp. 219-227.

13. Davis, S., 'TVD Finite Difference Schemes and Artificial Viscosity', ICASE Report No. 84-20, NASACR No. 172373, 1984.

14. Davis, S. and Flaherty, J.E., 'An Adaptive Finite Element Method for Initial-Boundary Value Problems for Partial Differential Equations,' SIAM J. Sci. Stat. Comput., 3 (1982), pp. 6-27.

15. Flaherty, J. E. and Moore, P.K., 'An Adaptive Local Refinement Finite Element Method for Parabolic Partial Differential Equations,' Proc. Conf. Accuracy Estimates and Adaptive Refinements in Finite Element Computations, Lisbon, (1984) pp. 139-152.

16. Flaherty, J. E. and Moore, P.K., 'An Local Refinement Finite Element Method for Time Dependent Partial Differential Equations,' Proc. Second Army Conf. Appl. Math and Comput., US Army Research Office Report No. 85-1, (1985), pp. 585-595.

17. Fritts, M. J., 'Numerical Approximation on Distorted Lagrangian Grids,' Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky and R. Stepleman (Eds.), IMACS, New Brunswick, (1979), pp. 137-142.

18. Gottlieb, D. and Orszag, S., Numerical Analysis of Spectral Methods: Theory and Applications, SIAM, Philadelphia, 1977.

19. Hindman, Richard, 'Generalized Coordinate Forms of Governing Fluid Equations and Associated Geometically Induced Errors,' AIAA J., 20 (1982), pp. 1359-1367.

20. Hindman, Richard, 'A Two-dimensional Unsteady Euler Equation Solver for Flows in Arbitrarily Shaped Regions Using a Modular Concept,' Ph.D. Thesis, Iowa State, Ames, Iowa, 1980.

21. Hoffman, J.D., 'Relationship between the Truncation Errors of Centered Finite-difference Approximation on Uniform and Nonuniform Meshes,' J. Comput. Phys., 46 (1982), pp. 469-474.

22. Hyman, J. M., 'Adaptive Moving Mesh Methods for Partial Differential Equations,' Los Alamos National Laboratory Report LA-UR-82-3690.

23. Lapidus, A., 'Detached Shock Calculation by Second Order Finite Differences,' J. Comput. Phys., 2 (1967), pp. 154-177.

24. MacCormack,R.W., 'The Effect of Viscosity in Hypervelocity Impact Cratering,' AIAA Paper 69-354, 1969.

25. Mastin, C., 'Error Induced by Coordinate Systems', Numerical Grid Generation, J. Thompson (Ed.), North-Holland, New York, (1982), pp. 31-37.

26. McRae, G., Goodin, W., and Seinfeld, J., 'Numerical Solution of the Atmospheric Diffusion Equation for Chemically Reacting Flows,' J. Comput. Phys., 45 (1982), pp. 1-42.

27. Mitchell, A.R. and Griffiths, D.F., The Finite Difference Method in Partial Differential Equations, Wiley, New York, 1980.

28. Osher, S. and Chakravarthy, S., 'Upwind Schemes and Boundary Conditions with Applications to Euler Equations in General Geometries,' J. Comput. Phys, 50 (1983), pp. 447-481.

29. Osher, S. and Sanders, R., 'Numerical Approximation to Nonlinear Conservation Laws with Locally Varying Time and Space Grids', Math. Comp., 41 (1983), pp. 321-336.

30. Rai, M. and Anderson, D., 'Grid Evolution in Time Asymptotic Problems,' J. Comput. Phys., 43 (1981), pp. 327-344.

31. Roe, P.L., 'Generalized Formulation of TVD Lax Wendroff Schemes,' ICASE Report No. 84-53, NASACR No. 172478, 1984.

32. Sanders, R., 'On Convergence of Monotonic Finite Difference Scheme with Variable Spatial Differencing', Math. Comp., 40 (1983), pp. 91-106.

33. Thomas, P. D. and Lombard, C. K., 'Geometric Conservation Law and its Application to Flow Computations on Moving Grids', AIAA J., 17 (1979), pp. 1030-1037.

34. Thomas, P. D. and Lombard, C. K., 'The Geometric Conservation Law - A Link between Finite Difference and Finite Volume Methods of Flow Computation on Moving Grids,' AIAA Paper No. 78-1208, 1978.

35. Thompson, J. F., 'Grid Generation Techniques in Computational Fluid Mechanics', AIAA J., 22 (1984), pp. 1505-1523.

36. Warming, R. F. and Beam, R. M., 'Upwind Second Order Difference Schemes and Applications in Aerodynamics', AIAA J., 14 (1976), pp. 1241-1249.

37. Whitham, G. B., Linear and Nonlinear Waves, Wiley-Interscience, New York, 1974. (1979), pp. 1-17.

38. Zienkiewicz, O. C. and Craig, A. W., 'Adaptive Mesh Refinement and A Posteriori Error Estimates for the P-version of the Finite Element Method,' Adaptive Computational Methods for Partial Differential Equations, I. Babuska, J. Chandra, and J. E. Flaherty (Eds.), SIAM, Philadelphia, (1983), pp. 33-56.

39. Zienkiewicz, O. C., Kelly, D. W., Gago, J., and Babuska, I., 'Hierarchical Finite Element Approaches, Error Estimates and Adaptive Refinement', Proc. MAFELAP 1981, April 1981.

AN ADAPTIVE OVERLAPPING LOCAL GRID REFINEMENT METHOD FOR TWO-DIMENSIONAL PARABOLIC SYSTEMS

Peter K. Moore

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

and

Joseph E. Flaherty

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590
and
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Laboratory
Watervliet, NY 12189-4050

ABSTRACT. We present an adaptive local refinement finite element method for solving vector systems of parabolic partial differential equations in two space dimensions and time. The algorithm uses the finite element-Galerkin method in space and backward Euler temporal integration. At each time step we obtain an estimate of the error on each element, group the elements whose error violates a user prescribed tolerance, form new local grids and solve the problem again on each of the new grids. We discuss several aspects of the algorithm, including the necessary data structures, the error estimation technique, and the determination of initial and boundary conditions at coarse-fine mesh interfaces. Finally we present several examples which demonstrate the viability of our approach.

I. INTRODUCTION. Over the past several years extensive efforts have been made in using adaptive strategies to solve partial differential equations [2, 3]. In this paper, we consider a local mesh refinement procedure for two-dimensional parabolic partial differential systems where fine meshes are introduced in regions where greater resolution is deemed necessary. Our approach permits finer meshes to overlap elements of coarser ones and is related to an earlier effort on h-refinement methods for

This research was partially supported by the U. S. Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant Number AFOSR 85-0156 and by the SDIO/IST under management of the U. S. Army Research Office under Contract Number DAAL 03-86-K-0112.

one-dimensional parabolic problems [5, 7, 10].

We consider an initial-boundary value problem for an m-dimensional vector system having the form

$$\mathbf{u}_t + \mathbf{f}(x,y,t,\mathbf{u},\mathbf{u}_x,\mathbf{u}_y) = [\mathbf{D}_1(x,y,t,\mathbf{u})\mathbf{u}_x]_x + [\mathbf{D}_2(x,y,t,\mathbf{u})\mathbf{u}_y]_y, \quad (x,y)\in\Omega, \quad t>0, \tag{1a}$$

$$\mathbf{u}(x,y,0) = \mathbf{u}_0(x,y), \quad (x,y)\in\Omega\bigcup\partial\Omega, \tag{1b}$$

$$\mathbf{u}(x,y,t) = \mathbf{g}_D(x,y,t), \quad (x,y)\in\partial\Omega_D, \quad t>0, \tag{1c}$$

$$\mathbf{D}_1\mathbf{u}_x\eta^1 + \mathbf{D}_2\mathbf{u}_y\eta^2 = \mathbf{g}_N(x,y,t), \quad (x,y)\in\partial\Omega_N, \quad t>0. \tag{1d}$$

The domain Ω is the rectangle $\{(x,y) \mid a<x<b, c<y<d\}$ with boundary $\partial\Omega = \partial\Omega_D \bigcup \partial\Omega_N$ and unit outer normal $\eta := [\eta^1,\eta^2]^T$. The system (1) is assumed to be well posed and parabolic, i.e., $\mathbf{D}_1$ and $\mathbf{D}_2$ are positive definite. We do not expect that our methods will be able to solve all problems having this generality, but our one-dimensional procedure [10] has worked well on a wide range of linear and nonlinear problems.

Our approach begins with the solution of (1) on a uniform space-time grid using finite elements in space and the backward Euler method in time. At the end of each time step, an indication of the local discretization error is generated on each finite element. In our initial investigation of one-dimensional problems [5, 7], we used an h-refinement (Richardson's extrapolation) procedure to compute a local error indicator. This has subsequently been abandoned in favor of a p-refinement approach [10], which increases the order of the trial space instead of reducing the mesh spacing. The p-refinement strategy employs nodal superconvergence to improve computational efficiency and it can be used to generate an asymptotically correct estimate of the discretization error [1, 10]. Elements having high error are grouped into rectangular regions called *megagrids* using a nearest neighbor clustering algorithm (cf. Berger and Oliger [4]). Overlapping fine uniform grids are generated within the megagrids and (1) is solved again on these grids. This process is repeated until a prescribed local error tolerance is satisfied. An illustration of a coarse spatial mesh with two megagrids and three fine grids is shown in Figure 1.

A tree is a natural data structure to manage the information associated with all of the grids. Nodes of the tree represent data at the megagrid level, with finer megagrids regarded as offspring of coarser ones. Information associated with overlapping fine grids within each megagrid are stored as records at the nodes of the tree.

A finite element problem is formulated and solved on each grid within a megagrid. This necessitates the prescription of appropriate initial and boundary conditions on each space-time grid. Since our temporal integration is implicit, prescribing boundary conditions is particularly complex in regions where meshes overlap (cf. Figure 1). An iterative procedure, analogous to Schwarz alternation (cf. Dihn et al. [6]), is used to successively calculate solutions on fine grids within each megagrid. We observe that this procedure converges for a variety of problems, but have no analysis

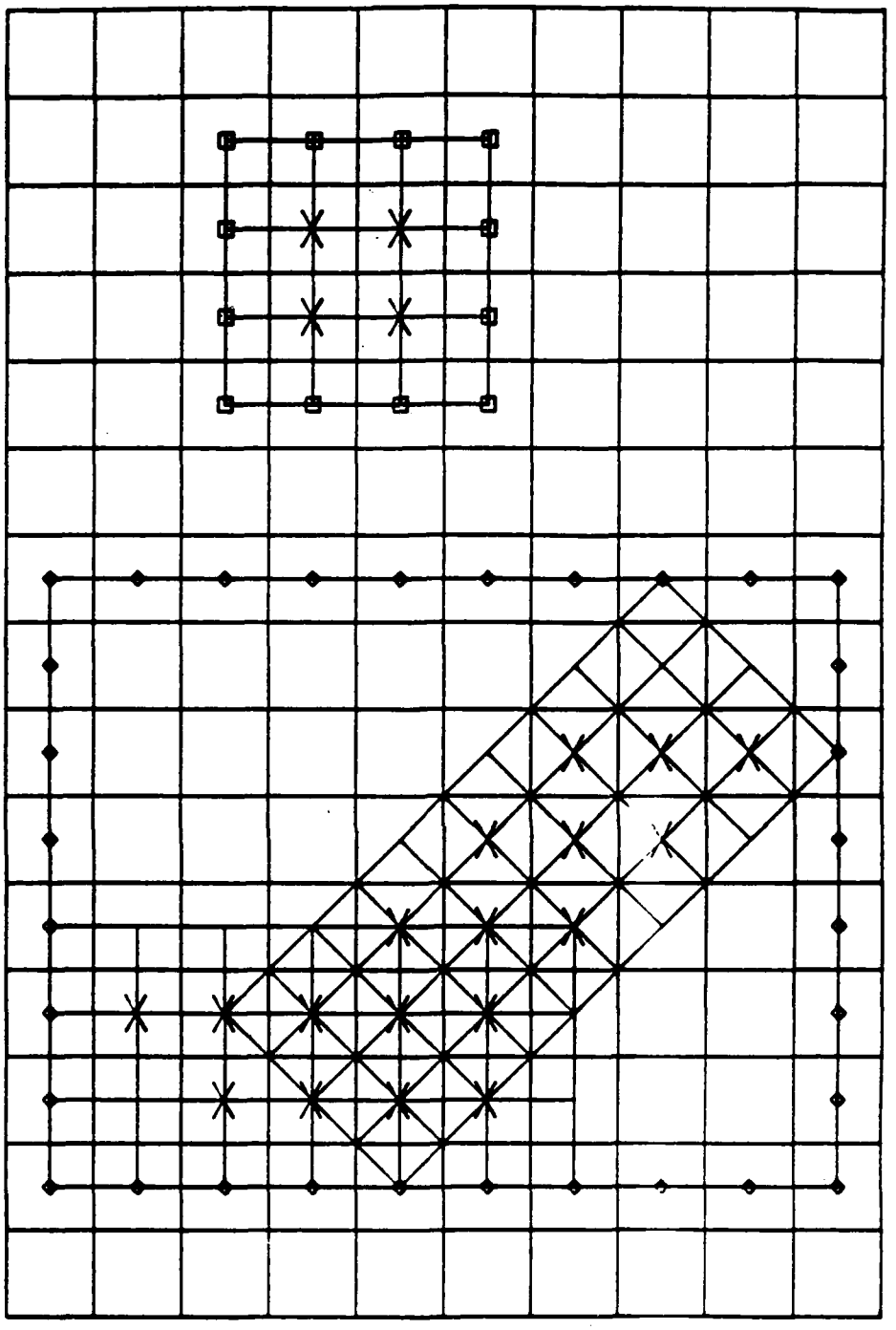


Figure 1. Coarse spatial background mesh with two offspring megagrids (marked with diamonds and squares) and their local grids. High-error elements of the coarse mesh are indicated by ×'s.

demonstrating either convergence or stability. Starius [11] obtained some stability results on a similar method for hyperbolic equations.

A description of the data structures and the local refinement procedure is given in Section II. In Section III we present the finite element method and the local error estimation technique. Section IV contains some preliminary computation results on three linear parabolic problems. Our conclusions and plans for further improvements are described in Section V. The examples indicate that the error estimation procedure converges to the true discretization error as the mesh is refined and the solution procedure based on the Schwarz alternating technique converges.

II. LOCAL REFINEMENT AND DATA STRUCTURES. We outline our procedure for solving (1) on an arbitrary hexahedral megagrid $R(\omega,p,q,F,S,L)$. The domain $\omega := \{(x,y) \mid \alpha < x < \beta, \gamma < y < \delta\}$; p and q are the times at the beginning and end of the time step, respectively; F and S point to the parent and offspring megagrids, respectively; and L is the record of information for the σ local rectangular grids within R.

A top level description of our local refinement algorithm is presented in Figure 2. A solution and error indicators are generated on R using **procedure** solve (cf. Figure 3). Elements where the error indicator exceeds a prescribed tolerance *tol* are partitioned into rectangular regions using the nearest neighbor clustering algorithm. As noted, we call these regions megagrids. Berger and Oliger's [4] bisection and merging procedure is used to generate local uniform fine grids for each megagrid. Local grids within a megagrid can overlap, but the megagrids are independent of each other; hence, each offspring megagrid may have different spatial and temporal refinement factors. This also reduces communication between megagrids and, thus, simplifies the computation of initial conditions on offspring megagrids. This representation may additionally be suitable for execution on parallel computers. Temporal refinement factors are calculated and solutions are recursively generated for each megagrid.

In order to solve problem (1), the procedure locref is invoked on the coarse grids $R(\Omega,t_k,t_{k+1},0,S,L)$, $k = 0, 1, \cdots$. Solutions satisfying the prescribed accuracy requirements are generated at each time t_k, $k = 1, 2, \cdots$.

The solution on a megagrid $R(\omega,p,q,F,S,L)$ is described by the **procedure** solve of Figure 3. Initial conditions are generated for each local computation grid contained in R. Following this, we compute an initial guess for the boundary conditions of the local grids using either the prescribed boundary data at physical boundaries or linear interpolation in time from the parent megagrid of R. A finite element solution is generated for one of the local grids and its solution is used to update boundary conditions on all other intersecting local grids. This solution process is repeated on each local grid in turn until satisfactory convergence is attained. Our procedure is, thus, similar to the Schwarz alternating principle for elliptic problems, which has been used recently to develop domain decomposition methods for parallel computation [6, 8].

A local grid is denoted as $T(x_m,y_m,d_x,d_y,s)$. Each local rectangular grid is characterized by the coordinates of its center (x_m,y_m), the lengths of its sides d_x and d_y, and the slope s of a side of the rectangle. In order to avoid ambiguity, we choose $s \geq 0$ and let d_x correspond to this side (cf. Figure 1). The number of elements $m_i \times n_i$ on local grid $T_i = T((x_m)_i,(y_m)_i,(d_x)_i,(d_y)_i,s_i)$ is determined by a single mesh

```
procedure locref (R(ω,p,q,F,S,L),tol);
  begin
    solve  (R(ω,p,q,F,S,L));
    if any error indicator > tol then
      begin
        Form offspring megagrids;
        for j := 1 to number of offspring do
          Create local rectangular grids;
        for j := 1 to number of offspring do
          Calculate the temporal refinement factor tref[j];
        for j := 1 to number of offspring do
          for i := 1 to tref[j] do
            begin
              p[i] := p+(i-1)*(q-p)/tref[j];
              q[i] := p[i] +(q-p)/tref[j];
              locref (R(ω[j],p[i],q[i],
                R(ω,p,q,F,S,L),S[j],L[j]),tol)
            end
      end
  end { locref };
```

Figure 2. Recursive local refinement algorithm for the solution of (1.1) on $R(\omega,p,q,F,S,L)$ with an error tolerance tol.

spacing parameter h_R as $m_i = round(d_x/h_R)$ and $n_i = round(d_x/h_R)$. Thus, each local grid in R has approximately the same spatial resolution. Many details of this algorithm have been omitted and additional information is presented in Moore [9]. For example, a strategy has been developed for storing the finite element solution at p and q without unnecessary duplication or copying of information.

Initial conditions for each local grid are either determined from (1b) when $p = 0$ or by bilinear interpolation using the finest grids available in the tree structure at time $p > 0$. Isolating local grids within megagrids greatly simplifies the search for data needed for this bilinear interpolation. Thus, the search for a solution value at an arbitrary point is performed at the megagrid level until the finest megagrid containing the point has been identified. The local grids of this finest megagrid provide the necessary interpolation data. Scanning the points of a grid in a predetermined order can be used to further reduce the complexity of the search procedure.

Similar considerations are required to determine boundary conditions on grid edges that are not subsets of $\partial\Omega$. Our one-dimensional techniques [10] and the explicit finite difference procedures of Berger and Oliger [4] used the notion of a "buffer" to apply boundary conditions. The idea is to enlarge a local rectangular grid by increasing d_x and d_y by two or four elements so that "artificial boundary conditions" may be obtained from data in low-error regions. However, in regions where local

```
procedure solve (R(ω,p,q,F,S,L));
  begin
    for i := 1 to number of local grids do
      begin
        Compute initial conditions for local grid
          T((x_m)_i,(y_m)_i,(d_x)_i,(d_y)_i,s_i);
        Compute boundary conditions for
          T((x_m)_i,(y_m)_i,(d_x)_i,(d_y)_i,s_i);
      end
    for j := 1 to number of iterations do
      for i := 1 to number of local grids do
        begin
          Solve the finite element problem for (1) on
            T((x_m)_i,(y_m)_i,(d_x)_i,(d_y)_i,s_i);
          if j = number of iterations then
            Compute error on T((x_m)_i,(y_m)_i,(d_x)_i,(d_y)_i,s_i)
          Update appropriate boundary conditions
        end
  end { solve } ;
```

Figure 3. Solution algorithm on megagrid $R(\omega,p,q,F,S,L)$.

grids overlap, accurate boundary conditions cannot be obtained from parent grid data even with a buffer. Buffers do provide accurate boundary conditions in regions where grids do not overlap and, for this reason, we continue to use them.

Dirichlet boundary conditions are obtained on the edges of buffered local grids by piecewise bilinear interpolation in time using solution values from the parent megagrid. In non-overlapping buffered regions, the interpolated boundary conditions satisfy the prescribed error tolerance and are, thus, expected to produce acceptable accuracy. As noted, accurate boundary conditions are obtained in regions where local grids overlap by means of the Schwarz alternating principle. Hence, we initially solve a finite element problem on local grid T_1 of R, realizing that the interpolated boundary data may be inaccurate in regions where T_1 intersects other local grids. In solving the problem on T_2 we use boundary data from T_1 with bilinear interpolation in regions where T_1 and T_2 intersect. This sequential updating procedure can be continued iteratively until satisfactory convergence is obtained. In practice, we halt the iteration after a few cycles and compute an error estimate for each local grid in R. The grids of R are refined if the error tolerance is not satisfied. Thus, we do not distinguish between failure of the Schwarz iteration to converge and failure to satisfy prescribed accuracy conditions.

Treatment of situations where local grids overlap $\partial\Omega$ are considerably more complex. A second complication arises when a local grid crosses the boundary of $\bigcup_{i=1}^{\sigma_F} (T_F)_i$, where the subscript F denotes the parent megagrid of R. These issues are

handled by regridding as described in Moore [9].

III. SPATIAL AND TEMPORAL DISCRETIZATION. As noted, the partial differential system (1) is discretized on a local grid T of R using a finite element Galerkin procedure in space and the backward Euler method in time. For each time $t \in [p,q]$, we assume that $\mathbf{u} \in H_E^1$ and select a test function $\mathbf{v} \in H_0^1$, where H^1 denotes the usual Sobolev space. Functions that further satisfy Dirichlet conditions on ∂T are said to belong to H_E^1, while functions satisfying trivial Dirichlet conditions belong to H_0^1.

The Galerkin form of (1) on T is

$$(\mathbf{v},\mathbf{u}_t) + (\mathbf{v},\mathbf{f}(\cdot,\cdot,t,\mathbf{u},\mathbf{u}_x,\mathbf{u}_y)) + A(\mathbf{v},\mathbf{u}) = \int_{\partial T \cap \partial\Omega_N} \mathbf{v}^T \mathbf{g}_N \, ds, \quad \text{for all } \mathbf{v} \in H_0^1, \tag{2a}$$

where

$$(\mathbf{v},\mathbf{u}) = \int_T \mathbf{v}^T \mathbf{u} \, dx\,dy, \tag{2b}$$

$$A(\mathbf{v},\mathbf{u}) = \int_T [\mathbf{v}_x^T \mathbf{D}_1(x,y,t,\mathbf{u})\mathbf{u}_x + \mathbf{v}_y^T \mathbf{D}_2(x,y,t,\mathbf{u})\mathbf{u}_y] \, dx\,dy. \tag{2c}$$

Initial conditions are required at $p = 0$ and these can be obtained, e.g., by L^2 or H^1 projection. Initial conditions for $p > 0$ trivially follow from the solution at the end of the previous time step.

A finite element solution of (2) is obtained by approximating H^1 by a finite dimensional subspace K of piecewise bilinear polynomials on T. The finite element solution $\mathbf{U}$ satisfies

$$(\mathbf{V},\mathbf{U}_t) + (\mathbf{V},\mathbf{f}(\cdot,\cdot,t,\mathbf{U},\mathbf{U}_x,\mathbf{U}_y)) + A(\mathbf{V},\mathbf{U}) = \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N \, ds.$$
$$\text{for all } \mathbf{V} \in K_0. \tag{3a}$$

$$\mathbf{U}(x,y,p) = \begin{cases} \mathbf{P}(\mathbf{u}_0), & p = 0 \\ \mathbf{P}(\mathbf{U}(\cdot,\cdot,p^-)), & p > 0. \end{cases} \tag{3b}$$

The projection $\mathbf{P}$ at $p = 0$ is obtained by constructing a piecewise bilinear approximation of $\mathbf{u}_0$. For $p > 0$, we proceed in a similar manner except that we construct interpolants using the finest grid solution available at $t = p^-$.

Temporal discretization of (3) is performed by the backward Euler method; thus, we determine $\mathbf{U}^q(x,y)$ as discrete approximation of $\mathbf{U}(x,y,q)$ by solving

$$(\mathbf{V},\mathbf{U}^q) + \Delta t\,[(\mathbf{V},\mathbf{f}(\cdot,\cdot,t,\mathbf{U}^q,\mathbf{U}_x^q,\mathbf{U}_y^q)) + A(\mathbf{V},\mathbf{U}^q)] = (\mathbf{V},\mathbf{U}^p) +$$

$$\Delta t \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N(x,y,q)ds, \quad \text{for all } \mathbf{V} \in K_0, \tag{4}$$

Initial conditions for the discrete system (4) follow the lines of (3b) for the semi-discrete system.

A posteriori estimates of the discretization error of the solution of (4) are obtained by means of a p-refinement technique. To begin, we calculate a second solution $\mathbf{U}_Q^q(x,y)$ of (2) using piecewise quadratic polynomials in space and trapezoidal rule integration in time. This solution is higher order in both space and time than the solution of (4); thus, the difference $\|\mathbf{U}^q - \mathbf{U}_Q^q\|_1$ furnishes an estimate of the discretization error of $\mathbf{U}^q$. The computational efficiency of this procedure can be substantially improved by using the nodal superconvergence property of finite element methods for parabolic problems [1, 10]. Nodal superconvergence implies that bilinear finite element solutions converge at a faster rate in space at nodes than elsewhere. These considerations imply that $\mathbf{U}_Q^q$ can be calculated as

$$\mathbf{U}_Q^q(x,y) \simeq \hat{\mathbf{U}}_Q^q(x,y) = \hat{\mathbf{U}}^q(x,y) + \mathbf{E}^q(x,y), \tag{5}$$

where $\hat{\mathbf{U}}^q(x,y)$ is a piecewise bilinear function and $\mathbf{E}^q(x,y)$ is a piecewise serendipity function (a biquadratic polynomial less a quartic term) that vanishes at the nodes of T. Specifically, we find that $\hat{\mathbf{U}}^q(x,y)$ satisfies

$$\left(\mathbf{V}, \frac{\hat{\mathbf{U}}^q - \mathbf{U}^p}{\Delta t}\right) + \tfrac{1}{2}[(\mathbf{V}, \mathbf{f}(\cdot,\cdot,q,\hat{\mathbf{U}}^q,\hat{\mathbf{U}}_x^q,\hat{\mathbf{U}}_y^q)) + (\mathbf{V}, \mathbf{f}(\cdot,\cdot,p,\mathbf{U}^p,\mathbf{U}_x^p,\mathbf{U}_y^p))] +$$

$$\tfrac{1}{2}[A(\mathbf{V},\hat{\mathbf{U}}^q) + A(\mathbf{V},\mathbf{U}^p)] = \tfrac{1}{2} \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N(x,y,q)ds + \tfrac{1}{2} \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N(x,y,p)ds$$

$$\text{for all } \mathbf{V} \in K_0. \tag{6}$$

Thus, a trapezoidal rule integration step is performed using the backward Euler solution $\mathbf{U}^p(x,y)$ as an initial condition. Both (4) and (6) are a nonlinear algebraic system which we solve by Newton's method. In order to reduce the computational effort associated with assembling and solving (6), the Jacobian of (4) is used for both Newton iterations. The solution of (4) is obtained first and the result $\mathbf{U}^q(x,y)$ is used as an initial guess for $\hat{\mathbf{U}}^q(x,y)$.

The piecewise quadratic correction $\mathbf{E}^q(x,y)$ satisfies

$$(\mathbf{V}, [(\hat{\mathbf{U}}^q + \mathbf{E}^q) - (\mathbf{U}^p + \mathbf{E}^p)]/\Delta t) + \tfrac{1}{2}[(\mathbf{V}, \mathbf{f}(\cdot,\cdot,q,\hat{\mathbf{U}}^q + \mathbf{E}^q, \hat{\mathbf{U}}_x^q + \mathbf{E}_x^q, \hat{\mathbf{U}}_y^q + \mathbf{E}_y^q))$$

$$+ (\mathbf{V}, \mathbf{f}(\cdot,\cdot,p,\mathbf{U}^p + \mathbf{E}^p, \mathbf{U}_x^p + \mathbf{E}_x^p, \mathbf{U}_y^p + \mathbf{E}_y^p))] + \tfrac{1}{2}[A(\mathbf{V},\hat{\mathbf{U}}^q + \mathbf{E}^q) + A(\mathbf{V},\mathbf{U}^p + \mathbf{E}^p)]$$

$$= \tfrac{1}{2} \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N(x,y,q)ds + \tfrac{1}{2} \int_{\partial T \cap \partial\Omega_N} \mathbf{V}^T \mathbf{g}_N(x,y,p)ds \quad \text{for all } \mathbf{V} \in K_0^Q \tag{7}$$

As noted, the space K_Q^q consists of piecewise serendipity functions that vanish at the vertices of the elements. Trivial initial conditions are used in the solution of (7) for $p > 0$. Interpolated values of the initial error $\mathbf{u}^0(x,y) - \mathbf{U}(x,y,0)$ onto K_Q^q are used at $p = 0$.

Linear systems associated with the application of Newton's iteration to (4), (6), and (7) are solved by the Lanczos acceleration of the Jacobi iterative method as implemented in the iterative solution package ITPACK of Young and Mai [12].

IV. EXAMPLES. We consider a sequence of three linear problems that are designed to illustrate the performance of our error estimation and local refinement procedures and convergence of the Schwarz iteration. Our results are very preliminary and additional computational work and analysis will be necessary before firm conclusions can be drawn.

Performance of our error estimation technique is measured by the effectivity ratio

$$\theta = \frac{\|\mathbf{U}^q - \hat{\mathbf{U}}_Q^q\|_1}{\|\mathbf{u}(x,y,q) - \mathbf{U}^q\|_1}, \tag{8}$$

which is a ratio of the estimated to the actual error in the H_1 norm. Ideally, the effectivity ratio should approach unity as the mesh is refined and should not differ substantially from unity over a large range of mesh spacings. The convergence of our error estimate to the true discretization error has been established for one-dimensional linear problems [10].

Example 1. Consider the linear constant coefficient heat conduction problem on $\Omega := \{ (x,y) \mid 0 < x,y < \pi \}$

$$u_t = \tfrac{1}{2}(u_{xx} + u_{yy}), \quad (x,y) \in \Omega, \quad t > 0, \tag{9a}$$

$$u(x,y,0) = \sin x \sin y, \quad (x,y) \in \Omega \cup \partial\Omega, \tag{9b}$$

$$u(x,y,t) = 0, \quad (x,y) \in \partial\Omega, \quad t > 0. \tag{9c}$$

The exact solution of this problem is

$$u(x,y,t) = e^{-t} u(x,y,0). \tag{10}$$

We solved (9) for a single time step on uniform grids having equal temporal and spatial mesh spacings of π/J, $J = 10, 20, 40$. The exact error and effectivity ratio are presented in Table 1. The results indicate that the finite element solution is converging at a linear rate and that the effectivity ratio is converging to unity.

Example 2. Consider the forced heat conduction equation on $\Omega := \{ (x,y) \mid 0 < x,y < 1 \}$

J	$\|u(x,y,\Delta t) - U^{\Delta t}\|_1$	θ
10	0.1578	1.050
20	0.0882	1.012
40	0.0469	1.003

Table 1. Error and effectivity ratio for one time step and uniform spatial meshes of spacing π/J for Example 1.

$$u_t + f(x,y,t) = u_{xx} + u_{yy}, \quad (x,y) \in \Omega, \quad t > t_0. \tag{11}$$

with $f(x,y,t)$ and the initial and Dirichlet boundary conditions specified so that the exact solution is

$$u(x,y,t) = \sin\pi t e^{t-20[(x-\frac{1}{2})^2+(y-\frac{1}{2})^2]}. \tag{12}$$

With $t_0 = 0.5$, we solve (11) for one time step on uniform grids having equal temporal and spatial meshes of $1/J$, $J = 10, 20, 40$. Results similar to those of Example 1 are displayed in Table 2. Thus, once again, the error is converging to zero at a linear rate and the effectivity ratio is tending to unity and is close to unity for all meshes. In this example, as opposed to Example 1, the effectivity ratio appears to be converging to unity from below. In practice, an upper bound is more suited to an adaptive local refinement procedure.

J	$\|u(x,y,\Delta t) - U^{\Delta t}\|_1$	θ
10	0.6796	0.996
20	0.3383	0.998
40	0.1668	0.999

Table 2. Error and effectivity ratio for one time step and uniform spatial meshes of spacing π/J for Example 2.

We also solve (11) for $0 = t_0 \le t \le 1$ using the adaptive local refinement strategy of Section II with a tolerance of 0.05 and an initial 10×10 mesh having a time step of 0.1. Surface renditions and contour plots of the solution at $t = 0.3$, 0.5, and 0.8 are

shown in Figures 4 and 5, respectively.

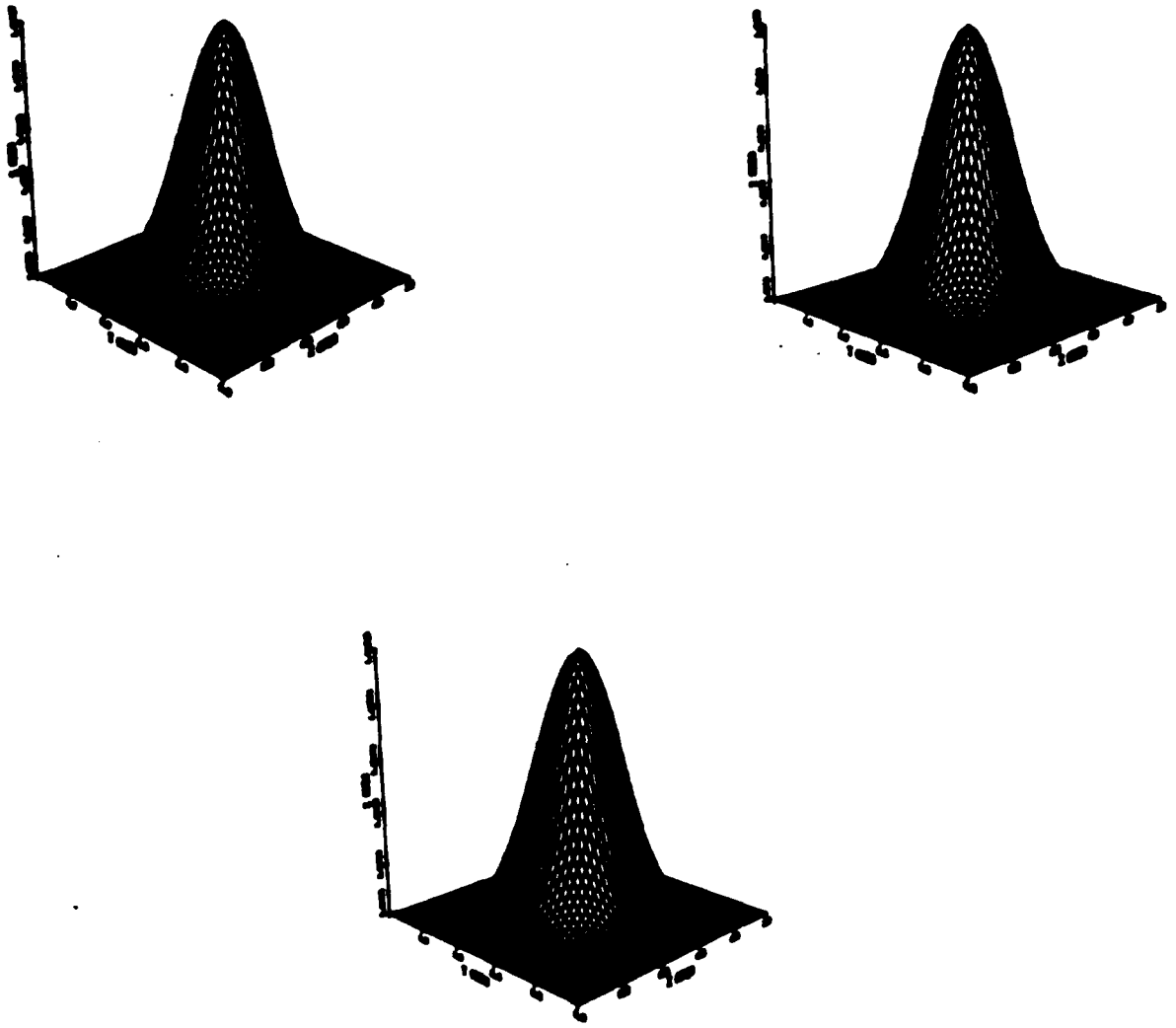


Figure 4. Surface renditions of the solution of Example 2 at $t = 0.3$ (upper left), 0.5 (upper right), and 0.8 (lower center).

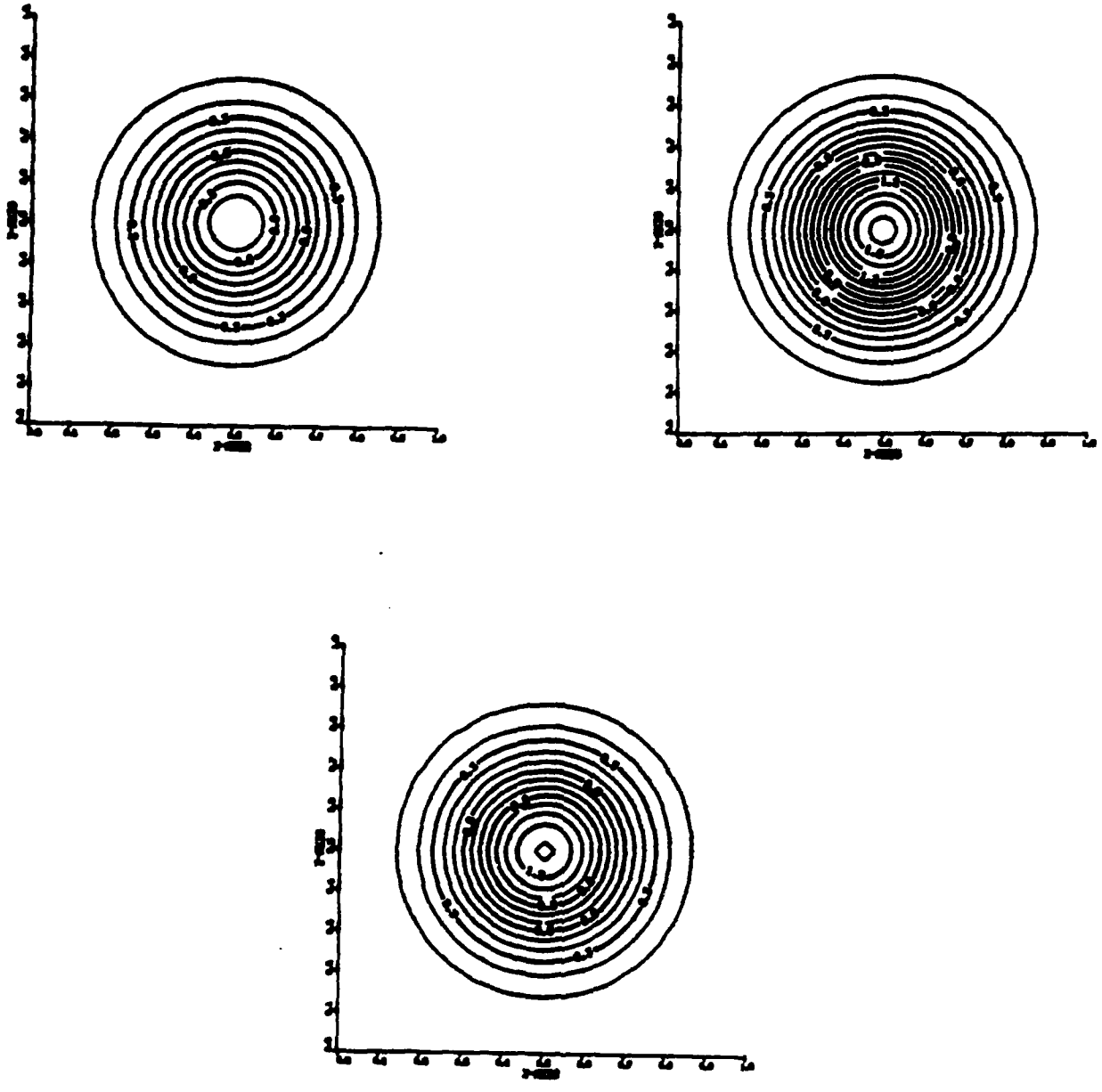


Figure 5. Contour plots of the solution of Example 2 at $t = 0.3$ (upper left), 0.5 (upper right), and 0.8 (lower center).

Example 3. Consider the forced heat conduction equation (11) on $\Omega := \{ (x,y) \mid 0 < x,y < 1 \}$ with $f(x,y,t)$ and the initial and Dirichlet boundary conditions specified so that the exact solution is

$$u(x,y,t) = 1.0 - \tanh[10(x+y-t-0.45)]. \tag{13}$$

This example is used to verify convergence of the Schwarz alternating principle. The problem is solved for a single time step with $t_0 = 0.5$ on an initial uniform coarse 10×10 mesh having a time step of 0.1 and a tolerance of 0.05. Refinement was needed at the initial time and 10 local grids, as shown in Figure 6, were introduced. The initial coarse mesh is also shown as a reference. Schwarz iterations were performed on these grids and we measure the difference in successive solutions on alternating grids on the portions of the boundaries of each local grid in regions where they overlap. The maximum such difference after each Schwarz iteration is shown in Table 3. It appears that the iteration is converging at nearly a quadratic rate.

Iteration	Maximum Difference
1	0.1506
2	0.0114
3	0.0016
4	0.0004

Table 3. Maximum difference between solutions on the boundaries of overlapping grids after each Schwarz iteration.

V. CONCLUSIONS. We developed an adaptive local mesh refinement procedure for nonlinear parabolic systems on rectangular regions. A complex tree data structure is used to manage a nest of local overlapping grids. An implicit finite element solution strategy using piecewise linear approximations and the backward Euler method is formulated. We obtain an estimate of the local discretization error of these finite element solutions using a p-hierarchical approach with piecewise serendipity approximations and trapezoidal rule integration. The Schwarz alternating principle is used to calculate boundary conditions on portions of local grids that overlap.

Our results indicate that the error estimation procedure converges to the exact local error as the mesh is refined. As noted, a proof of this convergence has been established for certain linear one-dimensional problems (cf. Moore and Flaherty [10]). It should be possible to construct a proof of convergence of the two-dimensional error estimate using the ideas developed in the one-dimensional case. The use of the Schwarz alternating principle also appears to be a very efficient method of calculating boundary conditions in overlapping-grid regions.

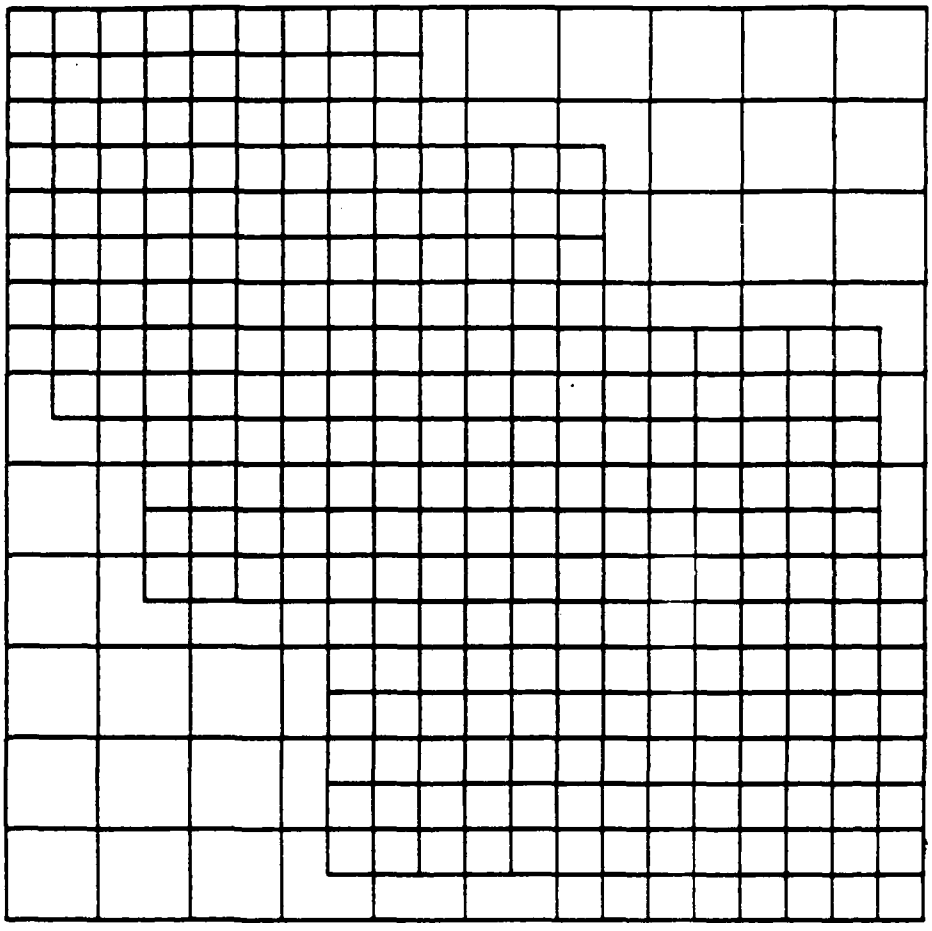


Figure 6. Local grids introduced after the initial time step for Example 3. The original coarse grid is also shown.

We are encouraged by the performance of our methods on these preliminary problems; however, several aspects of our approach need improvement. The Lanczos iteration used to solve the linear system appeared to be far less than optimal. The stopping criteria used in the ITPACK [12] implementation was too conservative for our applications. Creation of local solution grids is difficult and complex near domain boundaries. At present we know of no way of improving this defect. We have plans of extending our methods to non-rectangular domains using an overlapping-grid mesh generation procedure.

REFERENCES

1. S. Adjerid and J.E. Flaherty, Local refinement finite element methods on stationary and moving meshes for one-dimensional parabolic systems, 1987, in preparation.

2. I. Babuska, J. Chandra, and J.E. Flaherty, Eds., *Adaptive Computational Methods for Partial Differential Equations*, SIAM, Philadelphia, 1983.

3. I. Babuska, O.C. Zienkiewicz, J.R. Gago, and E.R. de A. Olivera, Eds., *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, John Wiley and Sons, Chichester, 1986.

4. M. Berger and J. Oliger, Adaptive mesh refinement for hyperbolic partial differential equations, *J. Comput. Phys.*, 53 (1984), 484-512.

5. M. Bieterman, J.E. Flaherty, and P.K. Moore, Adaptive refinement methods for non-linear parabolic partial differential equations, Chap. 19 in *Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, I. Babuska, O.C. Zienkiewicz, J.R. Gago, and E.R. de A. Olivera, Eds., John Wiley and Sons, Chichester, 1986.

6. Q.V. Dihn, R. Glowinski, and J. Périaux, Solving elliptic problems by domain decomposition methods with applications, in *Elliptic Problem Solvers II*, G. Birkhoff and A. Schoenstadt, Eds., Academic Press, Orlando, 1984, pp. 305-426.

7. J.E. Flaherty and P.K. Moore, A local refinement finite element method for time dependent partial differential equations, Trans. Second Army Conf. Appl. Math and Comput., ARO Rep. 85-1, U.S. Army Research Office, 1985, pp. 585-595.

8. R. Glowinski and M.F. Wheeler, Domain decomposition and mixed finite element methods for elliptic problems, Research Report UH/MD/-8, Department of Mathematics, University of Houston, 1987.

9. P.K. Moore, *A Local Adaptive Refinement Method for Parabolic Partial Differential Equations in One and Two Space Dimensions*, Ph.D. Dissertation, Rensselaer Polytechnic Institute, 1987, in preparation.

10. P.K. Moore and J.E. Flaherty, A local refinement finite element method for one-dimensional parabolic systems, 1987, in preparation.

11. G. Starius, On composite mesh difference methods for hyperbolic differential equations, *Numerische Math.*, 35 (1980), pp. 241-255.

12. D.M. Young and T.-Z. Mai, ITPACK 3A user's guide (preliminary version), Report CNA-197, Center for Numerical Analysis, The University of Texas, 1984.

Propagator Matrices in the Solution of EMP Problems

K. C. Heaton

Defence Research Establishment Valcartier

Abstract

A general solution to the problem of the electric and magnetic fields produced by extremely energetic explosions is difficult to obtain since boundary conditions near the explosion site, at infinity and at any intermediate conducting surface must all be satisfied. In particular, when these explosions occur near the surface of the Earth, the conductivity of the Earth is usually great enough that the tangential component of the quasi-static electric field and the normal component of the quasi-static magnetic field along the surface of the Earth must vanish.

In this work, the complete set of boundary conditions for the electric and magnetic fields at infinity, between the air and the Earth's surface, and between the air and the perfectly conducting plasma close to the explosion site are derived. The field equations, source functions and boundary conditions are written in terms of spheroidal and torsional vector fields. It is shown that, in this form, a propagator matrix formalism which automatically guarantees that all boundary conditions are satisfied can be developed to solve the equations for the electric and magnetic fields. The propagator matrix formalism developed in this work is applied to the numerical solution of Maxwell's equations for the electric and magnetic fields for the case of a typical explosion. It is found that the boundary conditions along the surface of the Earth impose consistency conditions which must be satisfied by the individual multipoles of the fields, as well as by the source current densities produced by the original explosion. Values are obtained for the electric and magnetic fields and compared with experimental results.

1 Introduction

Electric and magnetic fields of appreciable magnitudes, capable of being detected for considerable distances, accompany energetic explosions (Glasstone and Dolan 1977). When the explosions are caused by chemical explosives, the fields are generated by the compression of magnetic flux within the ionised gases at accelerating shock fronts (Wilhelm 1984, 1983) and by the dust cloud formed by the explosion (Bacon and Cherin 1984). For the case of nuclear explosions, the primary mechanism for

the production of these fields is electric currents caused by Compton scattering of electrons by X- and γ-rays, with the other two mechanisms having very minor roles or none at all. The fields caused by nuclear explosions are generally known as electromagnetic pulses (EMP) (e.g. Longmire and Gilbert 1980, Longmire 1978). As a comparison of the energies involved in each case would suggest, the fields produced by nuclear explosions are several orders of magnitude stronger than those caused by chemical explosions.

For the case of chemical explosions, the dust induced electromagnetic noise (DIEMN) is capable, at the least, of interfering significantly with radio and television broadcasts. The fields produced by the compression of ionised gases can initiate radio-controlled detonators or chemical explosives. In the case of nuclear explosions, the fields generated can produce field strengths of several kV/m over kilometre distance scales and time scales of milliseconds. In the Johnston Island test of 1962, the fields created by a nuclear explosion seem to have caused current surges in electrical equipment of sufficient magnitude to have triggered fuses in the street lighting system in Honolulu some 800 miles distant (Glasstone and Dolan 1977).

Lightning flashes have been observed to occur at times up to 1 second after a nuclear explosion at distances of .9 $\sim$ 1.4 km from the explosion site (Wyatt 1980, Uman et al 1972). These flashes are presumed to have been produced by the dielectric breakdown of the air by the electric fields generated by EMP. The most commonly used models for EMP are unable to predict electric fields of sufficient magnitude (usually believed to be $\sim$ 100 kV/m) to cause this breakdown (Wyatt 1980, Uman et al 1972).

Extensive work has been done in the past few years on the theoretical calculation of EMP effects at various stages of the explosion. Particular interest has been paid to the EMP generated by an explosion close to the surface of the Earth, especially during the so-called quasi-static phase during which the rate of change with respect to time of the electric and magnetic fields is sufficiently slow that it may be neglected in Maxwell's equations. It is well known that an electric field must vanish within a perfect conductor. In the region over which the Earth can be considered to be a perfect conductor, the quasi-static EMP field at the surface of the Earth should be zero. This boundary condition is automatically satisfied by odd multipoles of the electric field. From this condition, it has generally been assumed that the quasi-static electric field produced by a near surface blast can consist only of odd multipoles of the field throughout all space. (e.g. Downey 1983, Grover 1980)

In this paper, the complete set of boundary conditions for the quasi-static electric field and magnetic fields at infinity, between the air and the Earth's surface, and between the air and the perfectly conducting plasma close to an explosion site are derived. The field equations, boundary conditions, and source functions are expressed in terms of spheroidal and torsional vector fields. A general algorithm which uses propagator matrices and which automatically guarantees that all boundary conditions are satisfied is presented, and used to obtain numerical solutions to

Maxwell's equations for the electric and magnetic fields produced by a typical explosion. These results are then compared with experimental results. In particular, it is found that the boundary conditions along the surface of the Earth impose consistency conditions on all of the multipoles of the electric and magnetic fields, but that these conditions do not preclude the existence of even multipole fields.

2 Maxwell's Equations for the Quasi-static Phase of EMP

The time-dependent Maxwell's equations are

$$\vec{\nabla} \cdot \vec{D} = \rho_e, \tag{1}$$

$$\vec{\nabla} \cdot \vec{B} = 0, \tag{2}$$

$$\frac{\partial \vec{B}}{\partial t} = -\vec{\nabla} \times \vec{E}, \tag{3}$$

$$\frac{\partial \vec{D}}{\partial t} = -\vec{J} + \vec{\nabla} \times \vec{H}, \tag{4}$$

where $\vec{B}$ is the magnetic induction in webers/m^2, $\vec{E}$ the electric field in volts/m, $\vec{J}$ the current density in amps/m^2, $\vec{D} = \epsilon \vec{E}$ the electric displacement in coulombs/m^2, $\vec{H} = \frac{\vec{B}}{\mu}$ the magnetic field strength in amp-m, ρ_e the space charge density in coulombs/m^3, ϵ the dielectric permittivity in faradays/m, and μ the magnetic permeability in henrys/m. Throughout the course of this paper, we shall be concerned only with the calculation of the fields in air, and hence ϵ and μ will be assumed to take their free space values, ϵ_0 and μ_0.

Assuming that the fields are evaluated at times late enough that the fields are nearly constant in time, eqs. (3) - (4) become

$$\vec{\nabla} \times \vec{E} = 0, \tag{5}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J}, \tag{6}$$

in air.

Now, the current density $\vec{J}$ can be divided into two parts, the source current $\vec{J}_s$, and the conduction current $\vec{J}_c$. The source current arises from the ionisation created by the explosion; its exact form depends on whatever the dominant ionisation mechanism is at the time the fields are evaluated. For chemical explosions, this can be the ionisation created by the shock or collisions with dust particles. In nuclear explosions, $\vec{J}_s$ is created primarily by Compton scattering of the electrons in the air by γ- and X-rays. Since, to a good approximation Ohm's law is obeyed in air, one can write

$$\vec{J} = \vec{J}_s + \sigma \vec{E} \tag{7}$$

where $\vec{J}_c = \sigma \vec{E}$ and the conductivity σ is measured in 1/(ohms-m). In air, σ depends upon the value of $\vec{E}$ (e.g. Lee 1980, Longmire and Gilbert 1980). However, up to fields of strength $\sim$ 100 kV/m, this dependence is small and can be neglected.

According to the Helmholtz-Lamb decomposition theorem any vector field can be represented as the sum of a spheroidal vector field, $\vec{S}$, and a torsional vector field, $\vec{T}$, where

$$\vec{S} = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \vec{SP}_n^m, \tag{8}$$

$$\vec{T} = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \vec{T}_n^m, \tag{9}$$

and

$$\vec{SP}_n^m = U_n^m(r)\, S_n^m(\theta,\phi)\hat{r} + V_n^m(r)\frac{\partial S_n^m(\theta,\phi)}{\partial \theta}\hat{\theta} + \frac{V_n^m(r)}{\sin\theta}\frac{\partial S_n^m(\theta,\phi)}{\partial \phi}\hat{\phi}, \tag{10}$$

$$\vec{T}_n^m = -\frac{W_n^m(r)}{\sin\theta}\frac{\partial S_n^m(\theta,\phi)}{\partial \phi}\hat{\theta} + W_n^m(r)\frac{\partial S_n^m(\theta,\phi)}{\partial \theta}\hat{\phi}. \tag{11}$$

In eqs. (8)- (11) , U_n^m, V_n^m, W_n^m , are the functions containing the radial dependence of the vector associated with each surface spherical harmonic, S_n^m . The surface spherical harmonics of angular order m and rank n are defined by

$$S_n^m(\theta,\phi) = P_n^m(\cos\theta)e^{im\phi} \tag{12}$$

where the associated Legendre functions, P_n^m, are given by

$$P_n^m(\cos\theta) = (-1)^m \sin^m\theta \frac{d^m P_n(\cos\theta)}{d(\cos\theta)^m} \tag{13}$$

and the Legendre polynomials, P_n , by

$$P_n(\cos\theta) = \frac{(-1)^n}{2^n n!}\frac{d^n(\sin^{2n}\theta)}{d(\cos\theta)^n}. \tag{14}$$

r, θ, and ϕ the standard spherical polar co-ordinates with the origin located at the site of the original explosion, as shown in Fig. 1.

From the orthogonality conditions on spheroidal and torsional fields (Bullard and Gellman 1954, Smylie 1965) it is known that

$$\int_0^{2\pi}\int_0^{\pi} \vec{SP}_n^m \cdot \vec{SP}'^{m'}_{n'} \sin\theta\, d\theta\, d\phi \tag{15}$$
$$= (-1)^m \frac{4\pi}{(2n+1)}\left[U_n^m U'^{m'}_{n'} + n(n+1)V_n^m V'^{m'}_{n'}\right]\delta^n_{n'}\delta^m_{-m'},$$

$$\int_0^{2\pi}\int_0^{\pi} \vec{T}_n^m \cdot \vec{T}'^{m'}_{n'} \sin\theta \, d\theta \, d\phi = (-1)^m \frac{4\pi n(n+1)}{2n+1} W_n^m W'^{m'}_{n'} \delta^n_{n'} \delta^m_{-m'}, \tag{16}$$

$$\int_0^{2\pi}\int_0^{\pi} \vec{T}_n^m \cdot \vec{SP}^{m'}_{n'} \sin\theta \, d\theta \, d\phi = 0. \tag{17}$$

In order to satisfy eq. (5), the electric field must be entirely spheroidal, thusly:

$$\vec{E} = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \vec{\mathcal{E}}_n^m, \tag{18}$$

where

$$\vec{\mathcal{E}}_n^m = -\frac{\partial E_n^m(r)}{\partial r} S_n^m(\theta,\phi)\hat{r} - \frac{E_n^m(r)}{r} \frac{\partial S_n^m(\theta,\phi)}{\partial \theta}\hat{\theta} - \frac{E_n^m(r)}{r\sin\theta} \frac{\partial S_n^m(\theta,\phi)}{\partial \phi}\hat{\phi}. \tag{19}$$

Eqs. (18) - (19) are, of course, equivalent to stating that the electric field $\vec{E}$ must be derivable from a scalar potential.

After substituting eq. (7) into eq. (6) and taking the divergence, one obtains

$$-\vec{\nabla} \cdot (\sigma \vec{E}) = \vec{\nabla} \cdot \vec{J}_s. \tag{20}$$

The substitution of eqs. (18) - (19) into eq. (20), along with the application of eq. (15), yields

$$\sigma \frac{d^2 E_n^m}{dr^2} + \left(\frac{2\sigma}{r} + \frac{d\sigma}{dr}\right) \frac{dE_n^m}{dr} - n(n+1) \frac{\sigma}{r^2} E_n^m = \frac{dU_{J_s n}^{\ m}}{dr} + \frac{2}{r} U_{J_s n}^{\ m} - \frac{n(n+1)}{r} V_{J_s n}^{\ m}, \tag{21}$$

where it is assumed that the source current density, $\vec{J}_s$, is a spheroidal vector of the form

$$\vec{J}_s = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \vec{J}_n^m \tag{22}$$

where

$$\vec{J}_n^m = U_{J_s n}^{\ m}(r) \, S_n^m(\theta,\phi)\hat{r} + V_{J_s n}^{\ m}(r) \frac{\partial S_n^m(\theta,\phi)}{\partial \theta}\hat{\theta} + \frac{V_{J_s n}^{\ m}(r)}{\sin\theta} \frac{\partial S_n^m(\theta,\phi)}{\partial \phi}\hat{\phi}. \tag{23}$$

and the conductivity, σ , is assumed to be a function of r only. In fact, the conductivity exhibits a weak dependence on things like local field strength, angle, and water vapour content of the air. The assumption that the conductivity σ is a function only of r seems to be adequate at late times, at least as a first approximation (Grover 1980).

Using eqs. (2),(6) - (17), and the assumption that $\vec{J}_s$ is entirely spheroidal, one finds that

$$\vec{B} = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \vec{\mathcal{A}}_n^m + \vec{\mathcal{B}}_n^m, \tag{24}$$

where

$$\vec{\mathcal{A}}_n^m = \frac{\partial A_n^m(r)}{\partial r} S_n^m(\theta,\phi)\hat{r} + \frac{A_n^m(r)}{r} \frac{\partial S_n^m(\theta,\phi)}{\partial \theta}\hat{\theta} + \frac{A_n^m(r)}{r\sin\theta} \frac{\partial S_n^m(\theta,\phi)}{\partial \phi}\hat{\phi}, \tag{25}$$

$$\vec{B}_n^m = -\frac{B_n^m(r)}{\sin\theta}\frac{\partial S_n^m(\theta,\phi)}{\partial\phi}\hat{\theta} + B_n^m(r)\frac{\partial S_n^m(\theta,\phi)}{\partial\theta}\hat{\phi}. \tag{26}$$

The equations for the radial dependence of the field, A_n^m and B_n^m , are given by

$$\frac{d^2A_n^m}{dr^2} + \frac{2}{r}\frac{dA_n^m}{dr} - \frac{n(n+1)}{r^2}A_n^m = 0, \tag{27}$$

$$\frac{dB_n^m}{dr} + \frac{2}{r}B_n^m = \frac{\sigma\mu_0}{n(n+1)}\frac{dE_n^m}{dr} + \frac{\sigma\mu_0}{r}E_n^m - \frac{\mu_0}{n(n+1)}\left(U_{J_s n}^m + n(n+1)V_{J_s n}^m\right). \tag{28}$$

Grover (1980) and others (e.g. Hodgdon 1984) have derived similar equations, with the important difference that n in eq. (21) was only allowed to assume odd values. This was done in order to satisfy the boundary condition that the radial component, E_r , of the electric field must vanish identically over the surface of the Earth. However, as can be seen, if U_n^m and V_m^n are not identically zero for all even n, this ignores those multipoles excited by those current densities with even values of n. Since, in fact, the even multipoles of $\vec{J}_s$ are not all zero, another way of satisfying the boundary conditions must exist.

The boundary conditions on the fields across the boundary separating two regions, 1 and 2, are given by:

$$\hat{n}\times(\vec{E}_2 - \vec{E}_1) = 0, \tag{29}$$

$$\hat{n}\cdot(\vec{D}_2 - \vec{D}_1) = \varpi, \tag{30}$$

$$\hat{n}\cdot(\vec{B}_2 - \vec{B}_1) = 0, \tag{31}$$

$$\hat{n}\times(\vec{H}_2 - \vec{H}_1) = \vec{K}, \tag{32}$$

(Jackson 1962, Stratton 1941). In eqs. (29) - (32), the variables with subscript 1 refer to the region 1, and those with with subscript 2 refer to the region 2. ϖ is the surface charge density on the boundary between the regions, $\vec{K}$ the surface current density, and $\hat{n}$ the unit normal going from region 1 to region 2.

If the Earth is assumed to be a perfect conductor, the electric fields must vanish within it. Hodgdon (1984) has pointed out that sufficiently close to an explosion, the conductivity of the air first approaches and then surpasses that of the Earth. This implies that there is second region, distinct from the Earth, over which the boundary conditions, eqs. (29) - (32), must be applied: the region around the blast site in which the air is so highly ionised that it can be considered a perfect conductor. For simplicity, it will be assumed that this region is a hemisphere centred at $r = 0$ and with a radius r_0. Within that hemisphere, the electric and magnetic fields must vanish as well as within the Earth. At this stage, it will be assumed that the Earth can be treated as an infinite plane, located at $\theta = 90°$, and which is perfectly conducting for all $r > r_0$. It will also be assumed that all physical processes involved in the explosion and the field are symmetrical in the $x-y$ plane and hence that the resulting fields are independent of the ϕ co-ordinate. This implies that the angular

order m of the surface spherical harmonics in eqs. (19) - (21) is always 0, and hence that one is left only with a summation over the rank n.

Accordingly, the boundary conditions at the edge of the perfectly conducting hemisphere around the blast site become:

$$\hat{n} \times \vec{E}_2\Big|_{r=r_0} = 0, \tag{33}$$

$$\hat{n} \cdot \vec{D}_2\Big|_{r=r_0} = \varpi_2|_{r=r_0}, \tag{34}$$

$$\hat{n} \cdot \vec{B}_2\Big|_{r=r_0} = 0, \tag{35}$$

$$\hat{n} \times \vec{H}_2\Big|_{r=r_0} = \vec{K}_2\Big|_{r=r_0}. \tag{36}$$

In eqs. (33) - (36), ϖ_2 is the surface charge density in the air, $\vec{K}_2$ the surface current density in the air, and all other variables with a subscript 2 are to be understood to take those values which they would have in the air. The outward normal $\hat{n}$ to the hemisphere is the unit vector $\hat{r}$.

The application of eqs. (15) - (17) to eqs. (33) - (36) yields:

$$E_n^0\Big|_{r=r_0} = 0, \tag{37}$$

$$\frac{dE_n^0}{dr}\Big|_{r=r_0} = -\frac{(2n+1)}{2}\int_0^\pi \varpi_2(r_0,\theta) P_n^0(\cos\theta)\sin\theta\, d\theta, \tag{38}$$

$$\frac{dA_n^0}{dr}\Big|_{r=r_0} = 0, \tag{39}$$

$$B_n^0\Big|_{r=r_0} = -\frac{2n+1}{2n(n+1)}\int_0^\pi \vec{K}_2(r_0,\theta)\cdot\vec{S}P_n^0\sin\theta\, d\theta, \tag{40}$$

for $n \neq 0$. Equation (27) for A_n^m is simply the radial part of Laplace's equation in spherical co-ordinates, whose solution is given by

$$A_n^0 = a_n r^n + b_n r^{-(n+1)}, \tag{41}$$

where a_n and b_n are constants to be determined by the boundary conditions. Equation (41), taken in conjunction with eq. (39) and the requirement that the magnetic field be 0 as $r \to \infty$ implies that $a_n = b_n = 0$. This in turn implies that there is no spheroidal magnetic field during the quasi-static phase of EMP, only a torsional one.

The boundary conditions eqs. (33) - (36) degenerate even further for the case of the spherically symmetric or monopole part of the field (i.e. for $n = 0$). There can be no magnetic field associated with this part of the field, and so eq. (36) must be satisfied identically by insisting that the spherically symmetric part of $\vec{K}_2$ be 0. Equation (33) is automatically satisfied, leaving eq. (34) as the sole remaining condition.

The corresponding boundary conditions along the surface of a perfectly conducting Earth are:

$$\hat{n} \times \vec{E}_2\Big|_{\theta=90^\circ} = 0, \tag{42}$$

$$\hat{n} \cdot \vec{D}_2\Big|_{\theta=90^\circ} = \varpi_1, \tag{43}$$

$$\hat{n} \cdot \vec{B}_2\Big|_{\theta=90^\circ} = 0, \tag{44}$$

$$\hat{n} \times \vec{H}_2\Big|_{\theta=90^\circ} = \vec{K}_1. \tag{45}$$

In eqs. (42) - (45), ϖ_1 is the surface charge density density along the surface of the Earth, $\vec{K}_1$ the surface current density along the Earth. Again, the variables with a subscript 2 are to be evaluated in the air. Incidentally, for explosions over sea water, the surface of the Earth can be considered to be a perfect conductor much closer to the explosion site than would be the case for an explosion over soil.

The outward normal to the surface of the Earth is the unit vector along the z axis. Using this, and substituting eqs. (18), (19), (24), (25) and (26) into eqs. (42) - (45), one obtains

$$\sum_{n=0}^{\infty} \left(\sin\theta \frac{dE_n^0}{dr} P_n + \cos\theta \frac{E_n^0}{r} \frac{dP_n}{d\theta} \right)\Bigg|_{\theta=90^\circ} = 0, \tag{46}$$

$$\sum_{n=0}^{\infty} \left(\cos\theta \frac{dE_n^0}{dr} P_n - \sin\theta \frac{E_n^0}{r} \frac{dP_n}{d\theta} \right)\Bigg|_{\theta=90^\circ} = -\frac{\varpi_1}{\epsilon_0}, \tag{47}$$

$$\sum_{n=0}^{\infty} \sin\theta B_n^0(r) \frac{dP_n}{d\theta} \hat{r} + \cos\theta B_n^0(r) \frac{dP_n}{d\theta} \hat{\theta}\Bigg|_{\theta=90^\circ} = \mu_0 \left((\vec{K}_1 \cdot \hat{r})\hat{r} + (\vec{K}_1 \cdot \hat{\theta})\hat{\theta} \right). \tag{48}$$

Equation (44) is automatically satisfied for $m = 0$. The summations over the odd Legendre polynomials P_n vanish, leaving only the summations over the even polynomials to be satisfied, thusly:

$$\sum_{n=0}^{\infty} (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} \frac{dE_{2n}^0}{dr} = 0, \tag{49}$$

$$\sum_{n=0}^{\infty} (-1)^{n+1} \frac{(2n+2)(2n+1)!}{2^{(2n+2)}\left((n+1)!\right)^2} \frac{E_{2n+1}^0}{r} = -\frac{\varpi_1}{\epsilon_0}, \tag{50}$$

$$\sum_{n=0}^{\infty} (-1)^{n+1} \frac{(2n+2)(2n+1)!}{2^{(2n+2)}\left((n+1)!\right)^2} B_{2n+1}^0(r) = \mu_0 \vec{K}_1 \cdot \hat{r}. \tag{51}$$

The usual practice (Hodgdon 1984, Grover 1980) has been to satisfy the boundary condition by insisting $\frac{dE_n^0}{dr} = E_n^0 = 0$ throughout all space for the even spherical harmonics. As was indicated above, this seems unlikely if the current density depends to some degree upon the even spherical harmonics. What seems more likely is that the non-zero field at the surface of the Earth draws charges there which

arrange themselves in such a fashion so as to cancel the inducing field at the surface, but not necessarily throughout all space. Essentially, Eqs. (49) - (51) impose consistency conditions on the source current $\vec{J}_s$.

To sum up: in this section we have derived the equations governing the electric fields induced by electric currents in the atmosphere from explosions of various types. We have shown how the fields may be decomposed into multipole fields and that where the source and conduction currents are dependent upon particular multipoles, electric fields which are dependent on those multipoles are created. From this it follows that in general, both even and odd multipole fields exist as a result of an explosion.

Where the conductivity of the Earth is sufficiently high that it may be considered a perfect conductor with respect to the air, the boundary condition on the field requires that the component of the field along the ground must vanish. For the odd multipoles of the field, this condition is satisfied automatically. For the even multipoles, it is satisfied by the appearance of a surface charge density which produces a field which counteracts the original field at the surface of the Earth. However, the field which results from the sum of these two fields need not be zero everywhere else, and hence the even multipole fields can contribute to the total field.

3 Numerical Methods

Before one attempts numerical solutions of the field equations, eqs. (21) and (28), it is necessary to know the conductivity σ , and the source currents $\vec{J}_s$. Both of these depend upon the precise nature of the ionisation process. Since the most interesting cases from a theoretical standpoint occur when the fields are produced by a nuclear explosion, it was decided to choose expressions for σ and $\vec{J}_s$ appropriate to a thermonuclear explosion. Hence, at this point the further development of the field equations will be confined to the specific case of the fields generated by a thermonuclear explosion.

The total atmospheric conductivity is composed of two parts: an ionic and an electronic conductivity. Each Compton recoil electron produces about about thirty thousand ion-electron pairs. At early times, the electronic conductivity dominates; at late times, the ionic dominates. The expression for the total conductivity is hence

$$\sigma = \sigma_e + \sigma_I \tag{52}$$

where σ_e , the electronic conductivity, is given by

$$\sigma_e = e\,\mu_e\,\frac{S}{\alpha_e} \tag{53}$$

and σ_I , the ionic conductivity, is given by

$$\sigma_I = 2e\,\mu_I\left(\frac{S}{\gamma_I}\right)^{\frac{1}{2}} \tag{54}$$

(Downey 1983, Wyatt 1980, Grover 1980). In eqs. (53) and (54) e is the charge on the electron, μ_e the electron mobility, S the local ionisation rate, α_e the electron attachment rate, μ_I the ionic mobility, and γ_I the ion-ion recombination rate. The ionisation rate S is assumed to have the form

$$S = S_0 \frac{\exp(-r/\lambda)}{r^2} \tag{55}$$

where λ is the effective mean free path of the gamma rays, S_0 is a constant for a given time and yield, and r is, as above, the radial co-ordinate of a spherical co-ordinate system centred at the blast site.

For convenience, we shall define

$$\begin{aligned}
F_0(t) &= -3.9 \times 10^{-22}\, Y_0\, N_a \exp(-8.33 \times 10^2 t), \\
G_0(t) &= 8.2 \times 10^{-22}\, Y_0\, N_a \exp(-8.33 \times 10^2 t), \\
H_0(t) &= -2.8 \times 10^{-23}\, Y_0\, N_a \exp(-16.7t), \\
F(r,t) &= \frac{F_0(t)}{r^2}\left[\exp(-2.65 \times 10^{-5}\rho_0 r) - \exp(-1.04 \times 10^{-4}\rho_0 r)\right], \\
G(r,t) &= \frac{G_0(t)}{r^2}\exp(-4.61 \times 10^{-5}\rho_0 r), \\
H(r,t) &= \frac{H_0(t)}{r^2}\left[\exp(-2.20 \times 10^{-5}\rho_0 r) - \exp(-4.78 \times 10^{-5}\rho_0 r)\right], \\
X(r,t) &= F(r,t) + H(r,t), \\
Y(r,t) &= 16F(r,t) + 1.3H(r,t), \\
Q(r,t) &= G(r,t), \\
Z(r,t) &= -G(r,t).
\end{aligned} \tag{56}$$

In terms of the functions defined in eq. (56), the source current densities are given by

$$\begin{aligned}
J_r &= F(r,t)(1 + 16\cos\theta), \\
J_\theta &= G(r,t)(1 - \cos\theta),
\end{aligned} \tag{57}$$

for ground capture sources, and

$$\begin{aligned}
J_r &= H(r,t)(1 + 1.3\cos\theta), \\
J_\theta &= 0,
\end{aligned} \tag{58}$$

for air capture sources (Downey 1983). In eqs. (56) - (58), Y_0 is the total yield in kilotons, N_a is the number of neutrons/ kiloton, ρ_0 is the air density in mg/cm^3, r is the radial distance from the blast in centimetres, θ is the polar angle, t the retarded time in seconds, J_r the radial current density in abamps/cm^2 , and J_θ the polar current density in abamps/cm^2. The total current density at any retarded time t must be the vector sum of eqs. (57) - (58) . Hence, the components of the source current density are

$$J_r = X(r,t) + Y(r,t)\cos\theta, \tag{59}$$

$$J_\theta = Q(r,t) + Z(r,t)\cos\theta. \tag{60}$$

Using eqs. (56) - (59) and eqs. (15) - (17) one finds that:

$$\begin{aligned}
&U_{J_{\bullet 0}}^{0} = X(r),\\
&V_{J_{\bullet 0}}^{0} = 0,\\
&U_{J_{\bullet 1}}^{0} = Y(r),\\
&V_{J_{\bullet 1}}^{0} = -\frac{3\pi}{8}Q(r),\\
&U_{J_{\bullet n}}^{0} = 0, n > 1,\\
&V_{J_{\bullet n}}^{0} = -\frac{(2n+1)\pi}{2n(n+1)}Z(r)\sum_{k=0}^{n/2}(-1)^k\frac{(2n-2k)!(n-2k)}{2^{2n-2k}k!(n-k)!(n-2k+2)\left((n/2-k)!\right)^2}\\
&\qquad \text{for } n \text{ even }, n \geq 2\ ,\\
&V_{J_{\bullet n}}^{0} = -\frac{(2n+1)\pi}{2n(n+1)}Q(r)\sum_{k=0}^{(n-1)/2}(-1)^k\frac{(2n-2k)!}{2^{2n-2k-1}k!(n-k)!\left(((n-1)/2-k)!\right)^2(n-2k+1)}\\
&\qquad \text{for } n \text{ odd }, n \geq 3\ .
\end{aligned} \tag{61}$$

By substituting eq. (61) back into eqs. (21) and (28), it is now possible to solve numerically for the radial part of the electric potential and the electric and magnetic fields. It should, however, be noted that eq. (61) must be converted into amps/m^2 in order to be consistent with the expression for the conductivity. It is generally most convenient in numerical solutions of differential equations to use scaling factors to form dimensionless equations. By defining

$$\begin{aligned}
&\frac{dE_n^0}{dr} = y_1\frac{ML}{QT^2},\\
&E_n^0 = y_2\frac{ML^2}{QT^2},\\
&\frac{B_n^0}{\mu_0} = y_3\frac{Q}{TL},\\
&r = r^*L,\\
&t = t^*T,\\
&\sigma = \sigma^*\frac{TQ^2}{ML^3},\\
&U_{J_{\bullet n}}^{0} = \left(U_{J_{\bullet n}}^{*\,0}\right)\frac{Q}{TL^2},\\
&V_{J_{\bullet n}}^{0} = \left(V_{J_{\bullet n}}^{*\,0}\right)\frac{Q}{TL^2},\\
&F_n^0 = \left(F^{*0}{}_n\right)\frac{Q}{TL^3},
\end{aligned} \tag{62}$$

where

$$Q = L^3T\left(F_n^0(r_0)\right),$$

$$M = L^3 T \left[\left(F_n^0 \right)^2 / \left(\frac{d\sigma}{dr} \right) \right]_{r=r_0}, \tag{63}$$
$$F_n^0 = \frac{dU_{J,n}^{\ 0}}{dr} + \frac{2}{r} U_{J,n}^{\ 0} - \frac{n(n+1)}{r} V_{J,n}^{\ 0},$$

and L , T , M, and Q are the scaling factors in MKS units for length, time, mass and electric charge, respectively, with r_0 being the smallest value of r that appears in the integration, one is allowed to specify any two of L , T , M and Q as free parameters. It was found to be most convenient to specify $T = 16.7$ secs, and L as twice the maximum value of r used in the integration. Using the dimensionless quantities defined above, the field equations, eqs. (21) and (28) become

$$\begin{aligned}
\frac{dy_1}{dr^*} &= \frac{-1}{\sigma^*} \left(\frac{2\sigma^*}{r^*} + \frac{d\sigma^*}{dr^*} \right) y_1 + n(n+1) \frac{y_2}{r^{*2}} + \frac{\left(F^{*0}_n \right)}{\sigma^*}, \\
\frac{dy_2}{dr^*} &= y_1, \\
\frac{dy_3}{dr^*} &= \frac{\sigma^*}{n(n+1)} y_1 + \frac{\sigma^*}{r^*} y_2 - \frac{2y_3}{r^*} - \frac{1}{n(n+1)} \left(U^{*\ 0}_{J,n} + n(n+1) V^{*\ 0}_{J,n} \right).
\end{aligned} \tag{64}$$

Equations (64) form a system of linear differential equations which can most easily be solved by means of the propagator matrix formalism (Gilbert and Backus 1966, Smylie and Mansinha 1971). In a system of n linear homogeneous differential equations

$$\frac{d\vec{f}(r)}{dr} = \tilde{A}(r) \cdot \vec{f}(r), \tag{65}$$

where $\tilde{A}(r)$ is the matrix of coefficients, an $n \times n$ matrix $\tilde{F}(r)$ is called a fundamental matrix of the system eq. (65) if it satisfies the condition

$$\frac{d\tilde{F}(r)}{dr} = \tilde{A}(r) \cdot \tilde{F}(r), \tag{66}$$

and has an inverse for every r in its domain. Now, $\tilde{F}(r) = \tilde{P}(r, r_i)$ is called the propagator matrix of eq. (65) if

$$\tilde{F}(r_i) = \tilde{P}(r_i, r_i) = \tilde{I} \tag{67}$$

where $\tilde{I}$ is the identity matrix.

It follows (Gilbert and Backus 1966), among other things, that

$$\begin{aligned}
\tilde{P}(r_i, r_{i-2}) &= \tilde{P}(r_i, r_{i-1}) \cdot \tilde{P}(r_{i-1}, r_{i-2}), \\
\tilde{P}(r_i, r_{i-1}) &= \tilde{P}^{-1}(r_{i-1}, r_i), \\
\vec{f}(r) &= \tilde{P}(r, r_i) \cdot \vec{f}(r_i),
\end{aligned} \tag{68}$$

where $\vec{f}(r)$ is the solution to eq. (65) and $\vec{f}(r_i)$ is the solution at $r = r_i$.

Now, the system of non-homogeneous equations,

$$\frac{d\vec{f}(r)}{dr} = \tilde{A}(r) \cdot \vec{f}(r) + \vec{g}(r), \tag{69}$$

can be shown to have the solution

$$\vec{f}(r) = \int_{r_i}^{r} \tilde{P}(r,\varsigma) \cdot \vec{g}(\varsigma)\, d\varsigma + \tilde{P}(r,r_i) \cdot \vec{f}(r_i) \tag{70}$$

where $\vec{f}(r_i)$ is a solution to the non-homogeneous system , eq. (69). The boundary conditions eqs. (37) - (40) and (49) - (51) would overdetermine the system of equations (64) if ϖ_2, ϖ_1, $\vec{K}_1$, and $\vec{K}_2$ were known. Since they are not, the remaining boundary conditions are sufficient, and the unknown functions can be calculated from the general solution to eq. (64). In terms of the variables defined in eq. (62), the relevant boundary conditions are:

$$y_2|_{r=r_0} = 0, \tag{71}$$

$$\sum_{n=0}^{\infty} (-1)^n \frac{(2n)!}{2^{2n}(n!)^2} y_1^{(2n)} = 0 \tag{72}$$

where $y_1^{(2n)}$is the solution to the first of eqs.(64) associated with the $2n$th harmonic.

The boundary condition eq. (71) is satisfied if the initial solution, $\vec{y}(r_0)$ is chosen thusly:

$$\vec{y}(r_0) = \begin{bmatrix} \kappa_1 \\ 0 \\ \kappa_2 \end{bmatrix} \tag{73}$$

It was decided to set $r_0 = .4$ km. At that point, the conductivity σ of the air is approximately equal to that which is typical for the Earth's surface $\sim 10^{-3}$ (ohms-meter)$^{-1}$. In order to be consistent with the boundary conditions applied at the surface of the Earth, the air conductivity σ should be regarded as infinite there, and hence the boundary conditions, eqs. (37) - (40), apply. The constants κ_1 and κ_2 are determined by the condition that $\frac{dE_n^0}{dr} \to 0$ and $B_n^0 \to 0$ as $r \to \infty$. Since the fields were found to be small for $r > 5.0$ km, (Hodgdon 1984, Downey 1983), it was decided to apply the boundary condition $y_1 = y_3 = 0$ at values of r greater than that. For example, for $n = 1$, the boundary conditions were applied at $r = 8.1$ km, for $n = 2$ at $r = 7.0$ km, and so on. As a check, the integration was reperformed *ab initio* and extended out roughly twice as far, at which point new values for κ_1 and κ_2 were calculated. Typically, difference between the two values for the fields was found to be $\leq 1\%$ for $r \leq 4.0$ km, rising to about 3% at $r = 6.0$ km. It should be noted that the use of the boundary condition at infinity at a finite value of r implies that y_2 and hence the tangential field need not be zero there, as in fact it is not in general. However, its magnitude is sufficiently small at the point at which the boundary condition is applied that it may be neglected.

It should be noted that the equations developed above must be modified slightly when $n = 0$. In that case, $B_0^0 \equiv 0$, and eq. (28) becomes

$$\frac{dE_0^0}{dr} = \frac{U_{J,0}^0}{\sigma}. \tag{74}$$

As one would have expected, eq. (74) can also be obtained from the integration of eq. (21) with $n = 0$.

Equations (64) were solved using the propagator matrix formalism of eqs. (65) - (70) and a four-point Runge-Kutta algorithm with automatic error controls to obtain the propagator matrix. The expression $\int_{r_i}^{r} \tilde{P}(r,\varsigma) \cdot g(\vec{\varsigma})\, d\varsigma$ was evaluated using Simpson's second rule with various base point spacings and an algorithm in which global errors were controlled by halving the base point spacing until the largest difference between successive iterations was less than .2 kV/m in the electric fields for $n = 1$ and less than .02 kV/m for all other multipoles. The Runge-Kutta routines were checked for convergence by decreasing the upper error bound from 10^{-9} to 10^{-10}.

4 Numerical Results and Analysis

Figures 2-5 show the radial and tangential electric fields, electric potential and torsional magnetic field associated with the dipole (i.e. for $n = 1$ in eqs. (21), (27) and (28)) as a function of the radial co-ordinate r at various angles for a nuclear explosion of 10 megatons evaluated at a retarded time of 1 msec after the blast. Unless otherwise stated, it will henceforth be assumed that all of the fields discussed in this section are evaluated at the same retarded time of 1 msec, and that the source currents are those generated by a 10 megaton thermonuclear explosion (i.e. $Y_0 = 10^4$ in eqs. (56)). One also needs to have values for S_0, ρ_0, N_a, α_e, μ_e, γ_I, and μ_I. Following Grover (1980), S_0 in eq. (55) was set to 1.1×10^{30} ion-pairs/m-sec, a value appropriate to a 10 megaton burst. The values assumed for the other quantities were also those chosen by Grover (1980):

$$N_a = 2.0 \times 10^{23}\ \text{neutron/kT},$$
$$\rho_0 = 1.225\ \text{mg/cm}^3,$$
$$\alpha_e = 1.5 \times 10^8\ \text{sec}^{-1},$$
$$\mu_e = .25\ (\text{m}^2/\text{V-sec}),$$
$$\gamma_I = 2.0 \times 10^{-12}\ \text{m}^3/\text{sec},$$
$$\mu_I = 2.5 \times 10^{-4}\ (\text{m}^2/\text{V-sec}).$$

In reality, these values depend upon things like the field strength, air density and fraction of water vapour present. However, the average values will suffice as a first approximation. The gamma dose attenuation length λ was set to 320 metres for all calculations. Unless otherwise stated, r_0 will be assumed to be .4 km throughout.

Figure 6 shows the monopole electric field (i.e. for $n = 0$ in eq. (21)). Figures 7-10 show the quadrapole fields and potential (i.e. for $n = 2$ in eqs. (21), (27) and (28)). Figures 11-14 show the fields and potential for the sextopole fields (i.e. for $n = 3$ in eqs. (21), (27) and (28)). The corresponding graphs for the higher order multipoles show essentially the same behaviour as those for $n = 3$ and so are not reproduced here.

These graphs demonstrate several features of interest about the quasi-static electric and magnetic fields. Firstly, it is evident that the magnetic fields produced by the source currents are relatively weak. The strongest magnetic field is the one associated with the dipole ($n = 1$ multipole), which has a maximum of only $\sim 20\%$ of the typical strength of the geomagnetic field. This justifies *a posteriori* the neglect of self- consistent effects in the calculation of the fields, since such effects could only be a small perturbation to the main fields.

As one would have expected from the expressions for the source current densities, eqs. (61), the $n = 1$ fields dominate the others. However, both the $n = 0$ field and the higher order multipoles have non-negligible field strengths, especially close to the origin. For example, the $n = 0$ electric field has a peak value of $\sim 10\%$ of the $n = 1$ field, and the $n = 2$ field a peak value of $\sim 4\%$ of the $n = 1$ field.

Figures 15-16 show the sums of the radial and tangential electric fields for the multipoles from $n = 0$ to $n = 5$. As expected, the dipole ($n = 1$) field is the dominant influence, for both the radial and tangential fields, except for the total radial field at $\theta = 90°$. There, since all of the multipoles with odd values of n are identically zero, the multipoles for even values of n are the only non-zero fields.

That, of course, presents a problem since, from eq. (49), the sum of the radial fields must be zero along the surface of the Earth i.e. for $\theta = 90°$. Figure 17 shows the degree to which this consistency condition is violated. The solid line is the sum of the radial fields at $\theta = 90°$ for $n = 0, 2, 4$, with the $n = 0$ field being calculated using eq. (74). The dotted line shows the value which the $n = 0$ field would have to have in order to cancel the $n = 2$ and $n = 4$ fields at $\theta = 90°$. Since multipoles of order higher than 4 add relatively little to the total fields, their contributions have been neglected.

From the preceding discussion, it is clear that the expressions for the source current densities, eqs. (56) - (60), or for the conductivity, eqs. (52) - (55), or both, are inconsistent as they stand and must be altered. The approach taken in Heaton (1987) was to assume that the expressions for the source current densities and conductivities were correct, but that the boundary conditions were satisfied at the surface of the Earth by an induced electric potential. While possible in principle, this method leads to considerable numerical difficulties. An alternative approach would be to consider the physics of the situation more closely. Equations (52) - (60) have been derived from fits to experimental data (Downey 1983); accordingly, it is reasonable to seek the solution which changes them the least. An inspection of eqs. (59) - (60) reveals that $X(r,t)$ and $Z(r,t)$ have the least influence on the final field strengths, and of those two, $X(r)$ affects only the value of the $n = 0$

Table 1: Comparisons of Calculated EMP Electric Fields

Radius (m)	Total Field (Wyatt 1980) (kV/m)	Total Field $\theta = 0$ (Grover 1980) (kV/m)	Total Field $\theta = 0$ (Downey 1983) (kV/m)	Total Field $\theta = 0$ (Heaton 1987) (kV/m)	Total Field $\theta = 0$ (Present Work) (kV/m)
500	390	45	23	304	100
900	164	21	19	128	58
1300	114	15	13	63	25

field. Accordingly, it seems most reasonable to estimate the true field strength by replacing the value of $X(r,t)$ given in eq. (56) by $\sigma\Delta(r)$ where $\Delta(r)$ is the function graphed by the dotted line in Fig. 17.

Figures 18-19 show the total fields obtained by replacing the value of the $n = 0$ field as given by eq. (74) by $\Delta(r)$ and summing from $n = 0$ to $n = 5$, as before. A comparison of Figs. 15 and 18 shows that the effect of ensuring that the boundary conditions along the surface of the Earth are satisfied in this fashion is to decrease the peak value of the field by $\sim 9\%$.

One test of any model of EMP is whether it is capable of producing fields of sufficient intensity, usually regarded as being in excess of 100 kV/m, to cause the lightning observed during several tests. As can be seen from Figs. 2-4, the field for the $n = 1$ multipole reaches a maximum of $\sim$ 118 kV/m at .4 km and falls to less than 1 kV/m at 5.1 km. It is well known (e.g. Hodgdon 1984, Longmire and Gilbert 1980) that the dominant field is dipolar because of the $\cos\theta$ dependence of the current density. Hence, the fields displayed in Figs. 2-4 should constitute the greater part of the total electric field. It is encouraging that the magnitudes calculated are around those needed to produce nuclear lightning over some of the range in which they were observed (900 - 1400 m from the blast) at time scales of 1 msec (Wyatt 1980). Wyatt's values for the fields are listed in Table 1, and compared with the ones obtained here, as well as with Heaton's (1987), Downey's (1983) and Grover's (1980) values for the total fields. These values are necessarily adequate only for order of magnitude comparisons, because of the angular dependence of some of the field values. As can be seen, the results from the current work are considerably larger than either Grover's or Downey's results, and considerably smaller than Wyatt's or Heaton's (1987) values, reaching the 100 kV/m level only at 500 m. In Downey's and Grover's cases, the results are likely attributable to the different conductivity models used. Downey used detailed fits to the expected form of the conductivity, taking into account the air chemistry, as opposed to Grover's more approximate model. Even so, Downey only found a variation of 10% - 30% between his values and

Grover's. The major reason for the variation in the calculations of the magnitudes of the electric fields seems to be the boundary conditions at r_0. Grover and Downey felt that the condition that the field should vanish at $r = 0$ required that both the radial and tangential electric fields be zero at r_0, the point at which the integration is begun. While this satisfies the boundary conditions there, the condition is rather more restrictive than required. Wyatt neglected the boundary conditions at r_0 as did Heaton (1987). These last two sets of results, then, assume that the region of very high conductivity was sufficiently far from the region of interest that its effects on the field would be insignificant. The results in this paper are intermediate between the two groups of results, predicting fields lower than those of Wyatt and Heaton (1987) but higher than those of Grover and Downey. The boundary conditions near the explosion site chosen in this paper are more restrictive than Wyatt's and Heaton's (1987) but less restrictive than Grover's and Downey's. Wyatt and Heaton (1987) essentially applied no boundary conditions to the fields near the explosion while Grover and Downey required that both the radial and tangential components of the electric field vanish near the explosion site. The boundary conditions developed here for that region, eqs. (37) - (40), require that only the tangential electric field vanish near the explosion site. All this implies that the inner boundary conditions, eqs. (37) - (40), have a significant effect on the magnitudes of the fields, even far from the perfectly conducting region, contrary to the assumptions of Wyatt and Heaton (1987). This is borne out by an examination of the effects that the conductivities of the Earth and the perfectly conducting hemisphere around the explosion site have on the magnitude and location of the maximum electric field strength. When, as a test, r_0 was successively set to several different values, the maximum value of the $n = 1$ electric field was reduced in all cases. For $r_0 = .5$ km the maximum field strength was 116 kV/m and for $r_0 = .6$ km, it was 100 kV/m. For both cases, the maximum occurred at r_0. When the value of r_0 was decreased, the maximum electric field strength decreased more noticeably, and was not always located at r_0. For example, for $r_0 = .01$ km, the maximum field strength was 78 kV/m at .41 km, with the field strength at r_0 being only 28 kV/m. With $r_0 = .1$ km, the maximum field strength was 80 kV/m at .4 km, and at r_0, the field strength was 57 kV/m. It should be noted that the expressions for the conductivity, eqs. (52) - (55), are not really valid for $r < .4$ km, and so the results within that region should be regarded merely as being suggestive rather than definitive.

These results can best be understood by a consideration of the conductivity and source current models employed. Increasing the value of r_0 is essentially equivalent to decreasing the ground conductivity. The maximum value of the fields decreased slightly when this was done because that part of the source current density at values of $r < r_0$ was not included in the calculation, thus reducing the total field strength. Decreasing the value of r_0 is equivalent to increasing the conductivity of the ground. The vanishing of the tangential electric field at r_0 forces a proportionately smaller discontinuity in the potential as r_0 decreases, hence producing a weaker radial field there. One would expect the air around the explosion site to be divided into roughly

three regions: a completely ionised one, a partially ionised one, and an almost completely unionised one. The boundaries between these regions will obviously change at different rates with respect to time. The unionised and partially ionised regions will move inwards, and the completely ionised region will eventually vanish altogether. The results obtained above suggest that the widths of these three zones and the structure of the transitions from one to the other are crucial for the determination of the magnitudes of the fields. Since the maximum values of the fields are near to those required to initiate dielectric breakdown in the air, it may be that relatively small changes in atmospheric or ground properties might produce conditions favourable for the occurrence of nuclear lightning in one instance, and unfavourable in the next.

5 Conclusions

In this work, it has been demonstrated that the quasi-static electric fields produced by an explosion contain components that depend on both the odd and even surface spherical harmonics, and that this remains true even if the explosion occurs near a good conductor.

Expressions for the excitation function for the EMP in terms of the surface spherical harmonics were obtained, and used, along with a simple model of ionic and electronic conductivity, to obtain values for the electric and magnetic fields generated by a typical explosion. A propagator matrix algorithm for solving the EMP equations was developed. Using this algorithm, it was demonstrated that the dominant electric and magnetic fields are dipoles, but that the contribution of the other multipole fields to the total field is significant. In particular, the calculated values of the field were near those which are expected to produce the lightning which has been observed to accompany nuclear explosions. The results obtained here suggest that the detailed structure of the ionised regions around the explosion site is crucial to the existence of nuclear lightning, and that relatively small changes in a few parameters might be sufficient to permit its formation.

It is shown that the values which were used for the source currents lead to values of the electric fields which do not satisfy the boundary conditions over the surface of the Earth, and hence should be modified to take into account the boundary conditions on the fields.

Efforts are currently being made to extend this work by incorporating more accurate, self-consistent models for the conductivity and source currents, perhaps by the introduction of Monte Carlo techniques into the algorithms. As well, it is planned to modify the boundary conditions to take account of the large but finite conductivities in the Earth and around the blast site, and extend the propagator matrix formalism to the time-dependent case.

References

[1] Bacon, D.P. and Cherin, D.P. 1984, *Dust Induced Electro-Magnetic Noise (DIEMN)*, Science Applications Intl. Corp, Maclean, Virginia

[2] Bullard, E.C., and Gellman, H. 1954, *Phil. Trans. R. Soc. A*, 247, 213

[3] Downey, J.R. 1983, M. Sc. Thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

[4] Glasstone, S. and Dolan P. J. 1977, *The Effects of Nuclear Weapons*, United States Department of Defence, Washington, D.C.

[5] Gilbert, F. and Backus, G.E. 1966, *Geophysics*, XXI, no. 2, 326

[6] Grover, M.K. 1980, *Some Analytic Models for Quasi-Static Source Region EMP: Application to Nuclear Lightning*, R and D Associates, Marina Del Ray, California

[7] Heaton, K.C. 1987, *Proceedings of the 4th Army Conference on Applied Mathematics and Computing*, 387, ARO Report 87-1, U.S. Army Research Office

[8] Hodgdon, K.M. 1984, M. Sc. Thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

[9] Jackson, J. D. 1962, *Classical Electrodynamics*, John Wiley and Sons Inc., New York

[10] Lee, K.S.H. 1980, *EMP Interaction: Principles, Techniques, and Reference Data*, Dikewood Industries, Albuquerque, New Mexico

[11] Longmire, C.L. 1978, *IEEE Transactions on Antennas and Propagation*, AP-26, no. 1, 3

[12] Longmire, C.L. and Hobbs, W.E. 1979, *Fireball Effects in Late Time EMP from Surface Bursts*, Mission Research Corporation, Santa Barbara, California

[13] Longmire, C.L. and Gilbert, J. L. 1980, *Theory of EMP Coupling in the Source Region*, Mission Research Corporation, Santa Barbara, California

[14] Smylie, D.E. 1965, *Geophys. J. R. astr. Soc.*, 9, 169

[15] Smylie, D.E. and Mansinha, L. 1972, *Geophys. J. R. astr. Soc.*, 23, 329

[16] Stratton, J. A. 1941, *Electromagnetic Theory*, McGraw-Hill Book Co. Inc., New York

[17] Uman, M.A., Seacord, D.F., Price, G.H., and Pierce, E.T. 1972, *Journal of Geophysical Research*, 77, 1591

[18] Wilhelm, H.E. 1983, *Appl. Phys. B*, 31, 107

[19] _____ 1984, *J. Appl. Phys.*, 56, no. 5, 1285

[20] Wyatt, W.T. 1980, *An Improved Model for EMP Induced Lightning*, U.S. Army Materiel Development and Readiness Command, Alexandria, Virginia

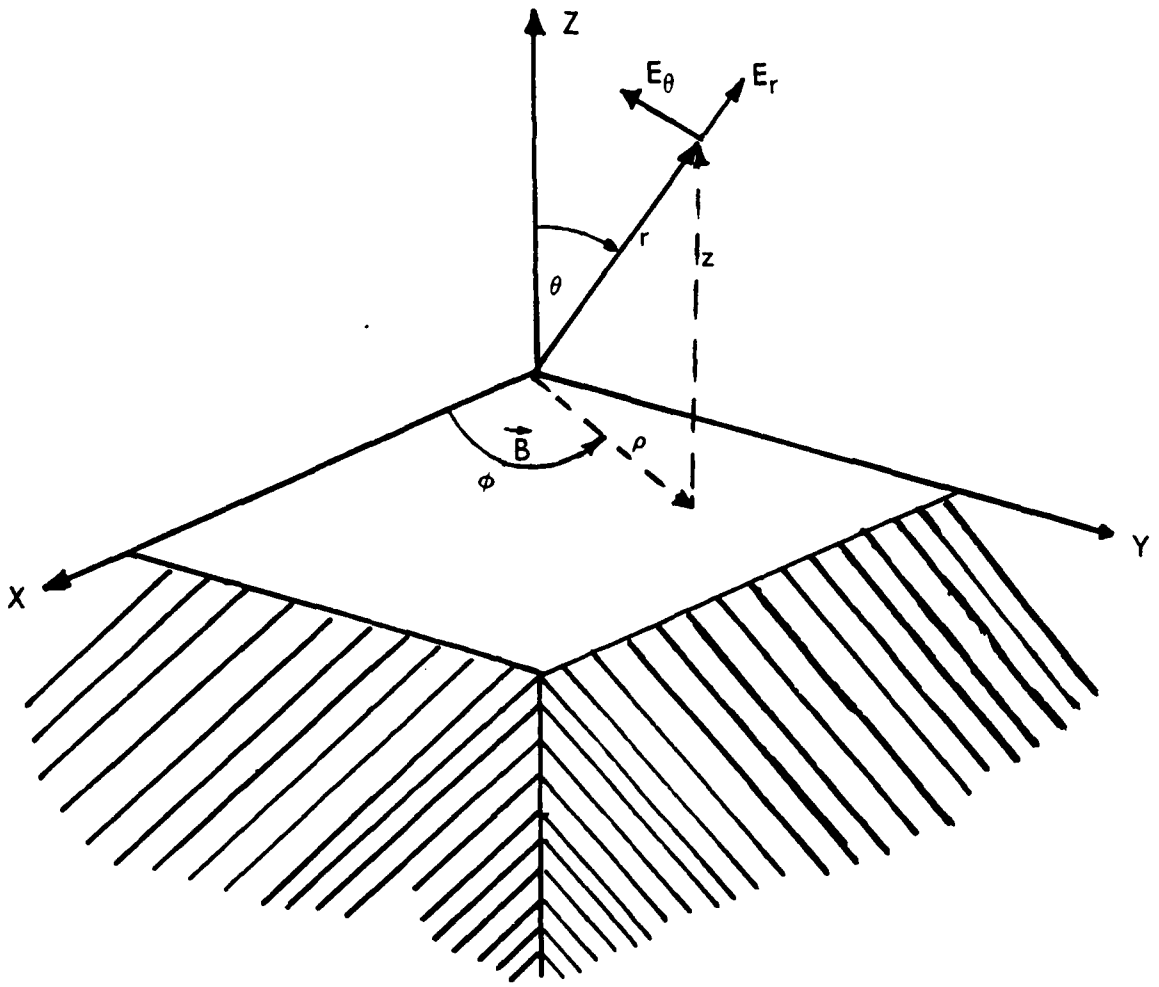
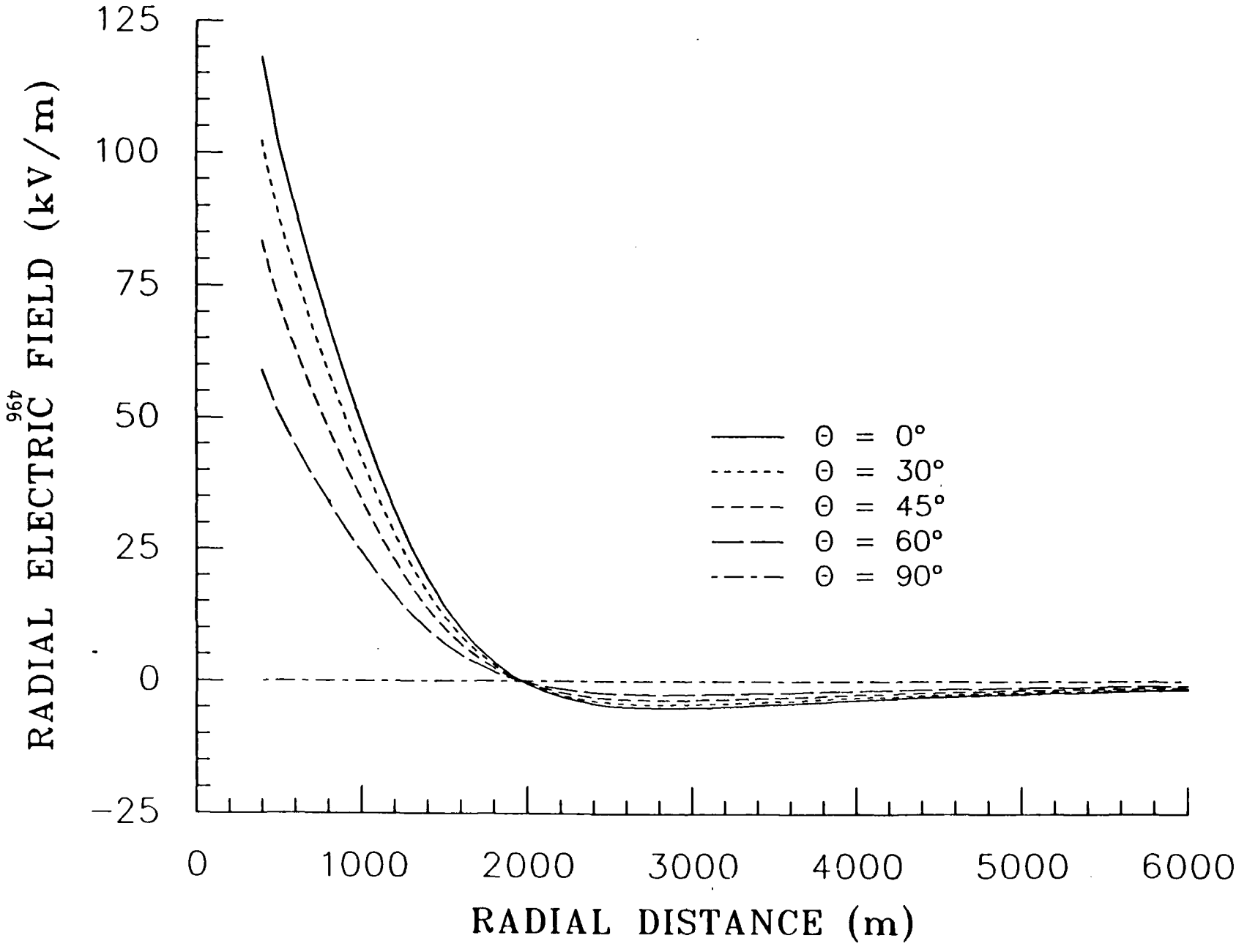


Figure 1: Relation of EMP Fields to the Earth

Fig. 2: Radial Electric Field for $n = 1$



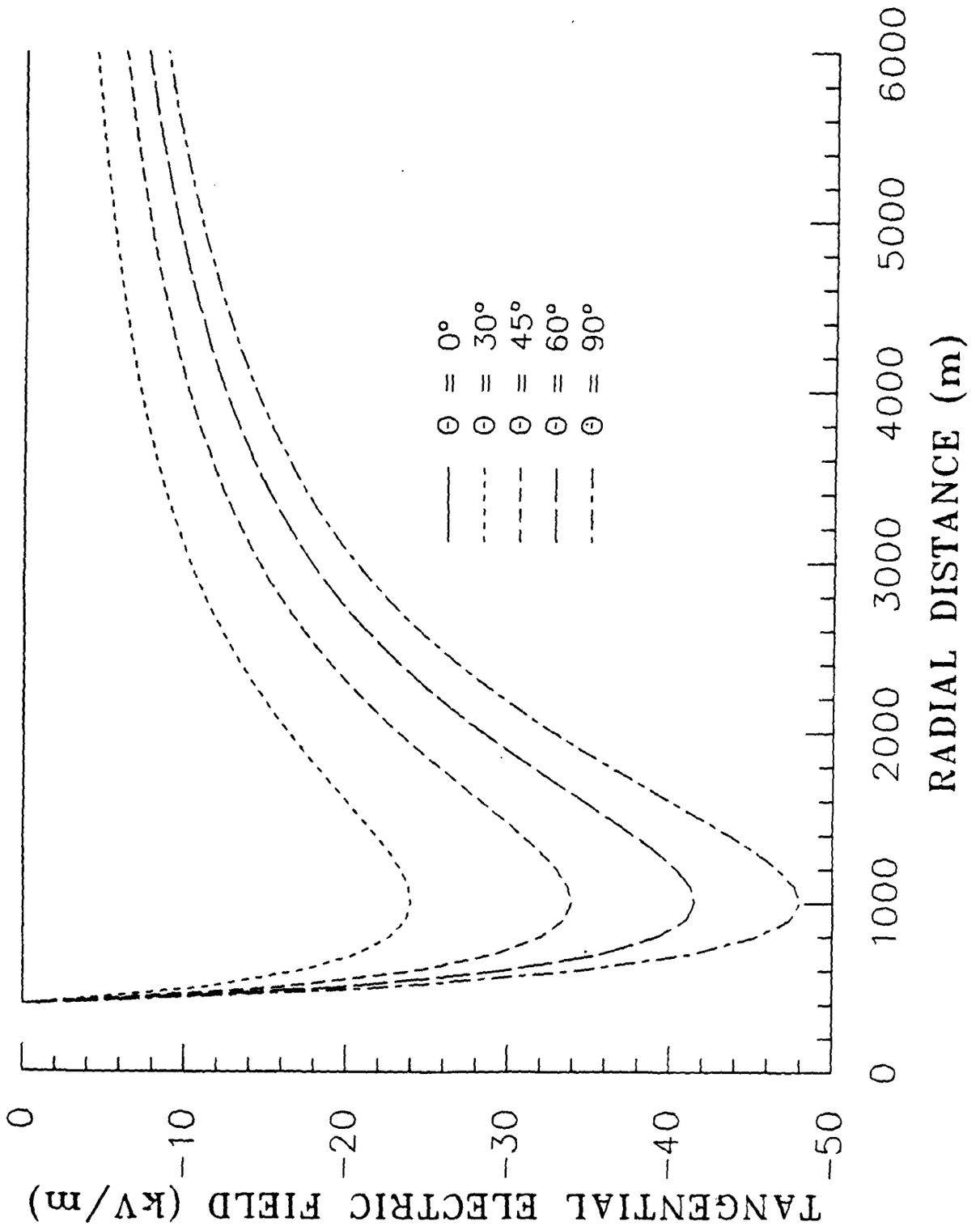


Fig. 3: Tangential Electric Field for $n = 1$

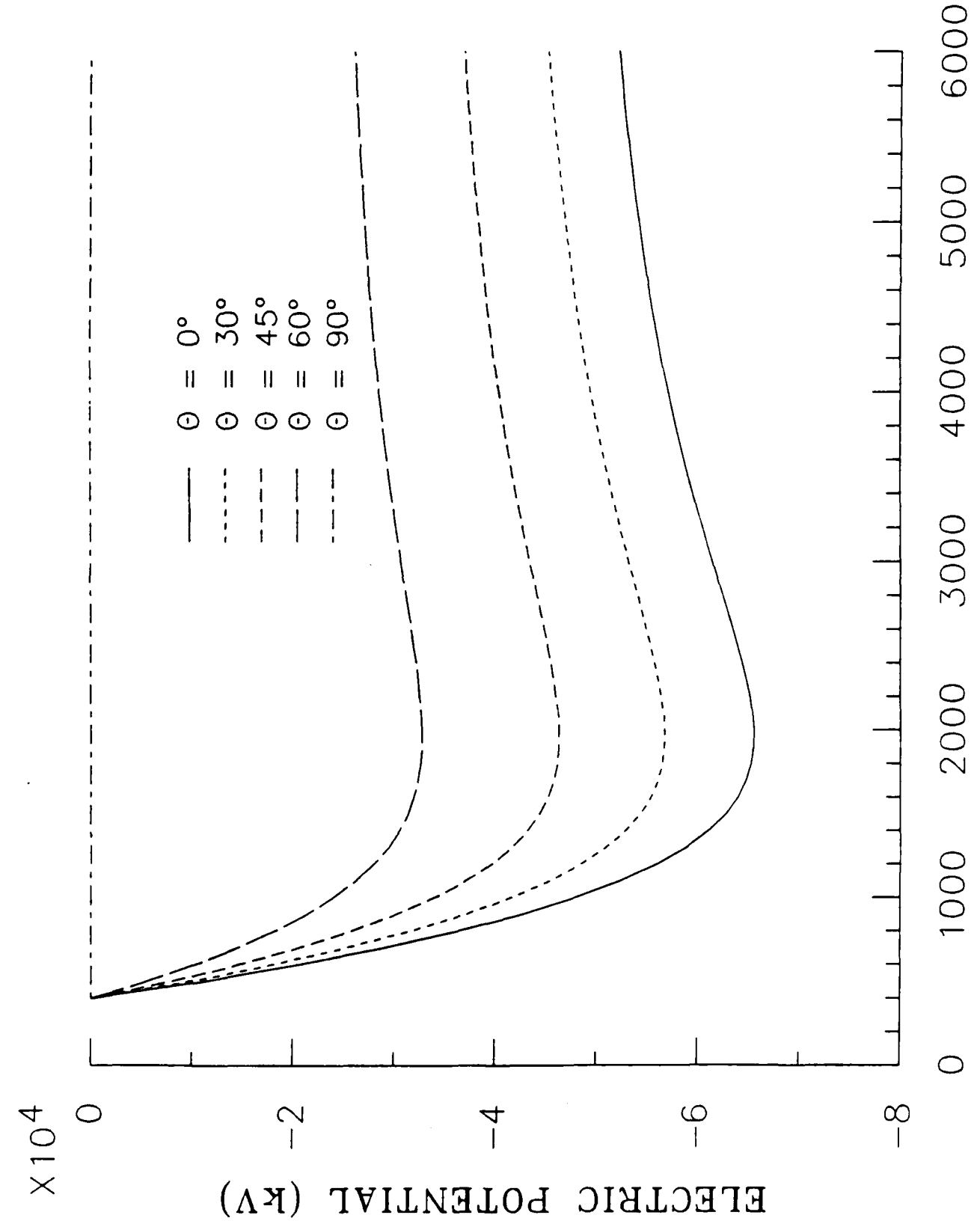


Fig. 4: Electric Potential for $n = 1$

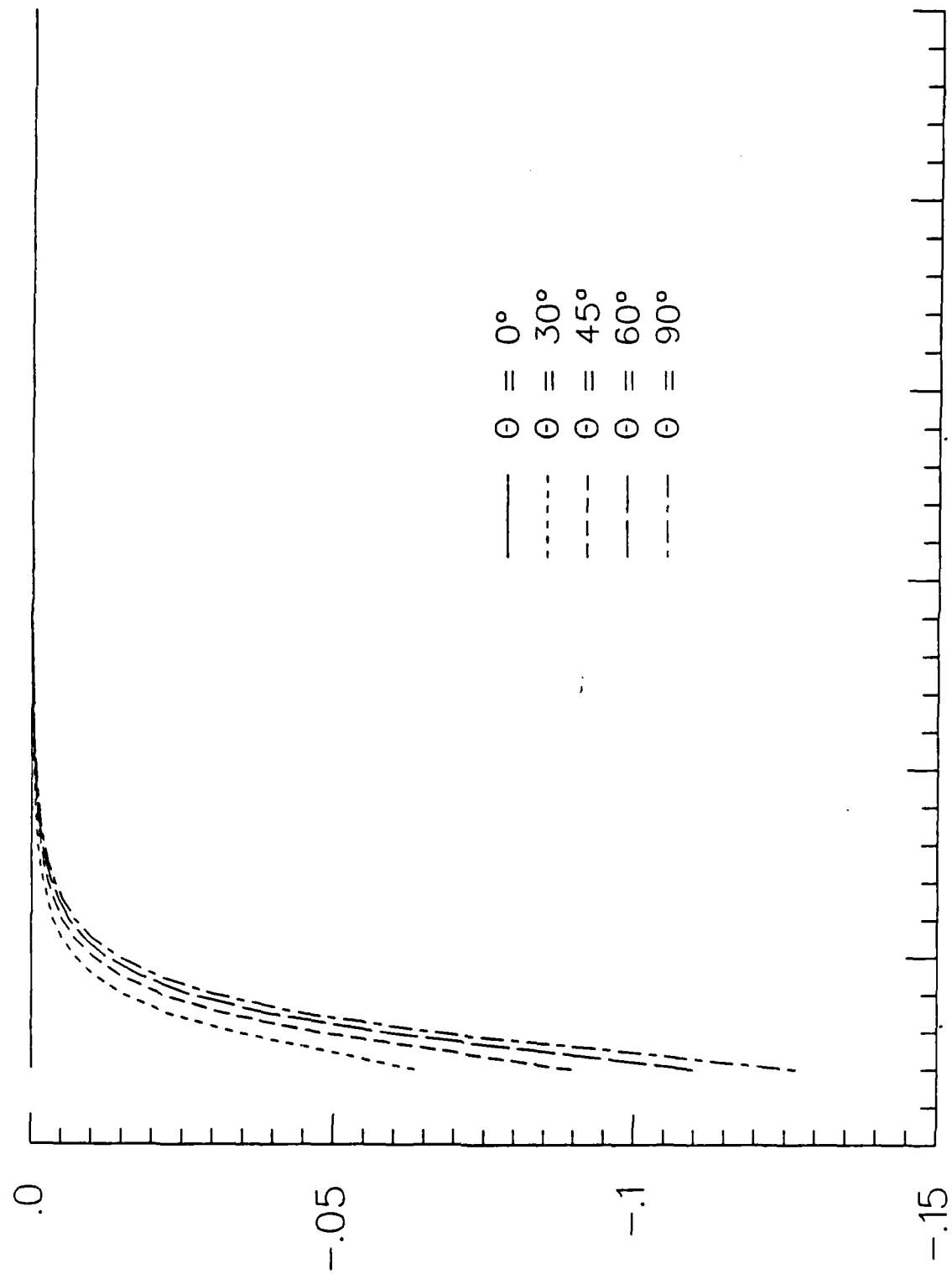
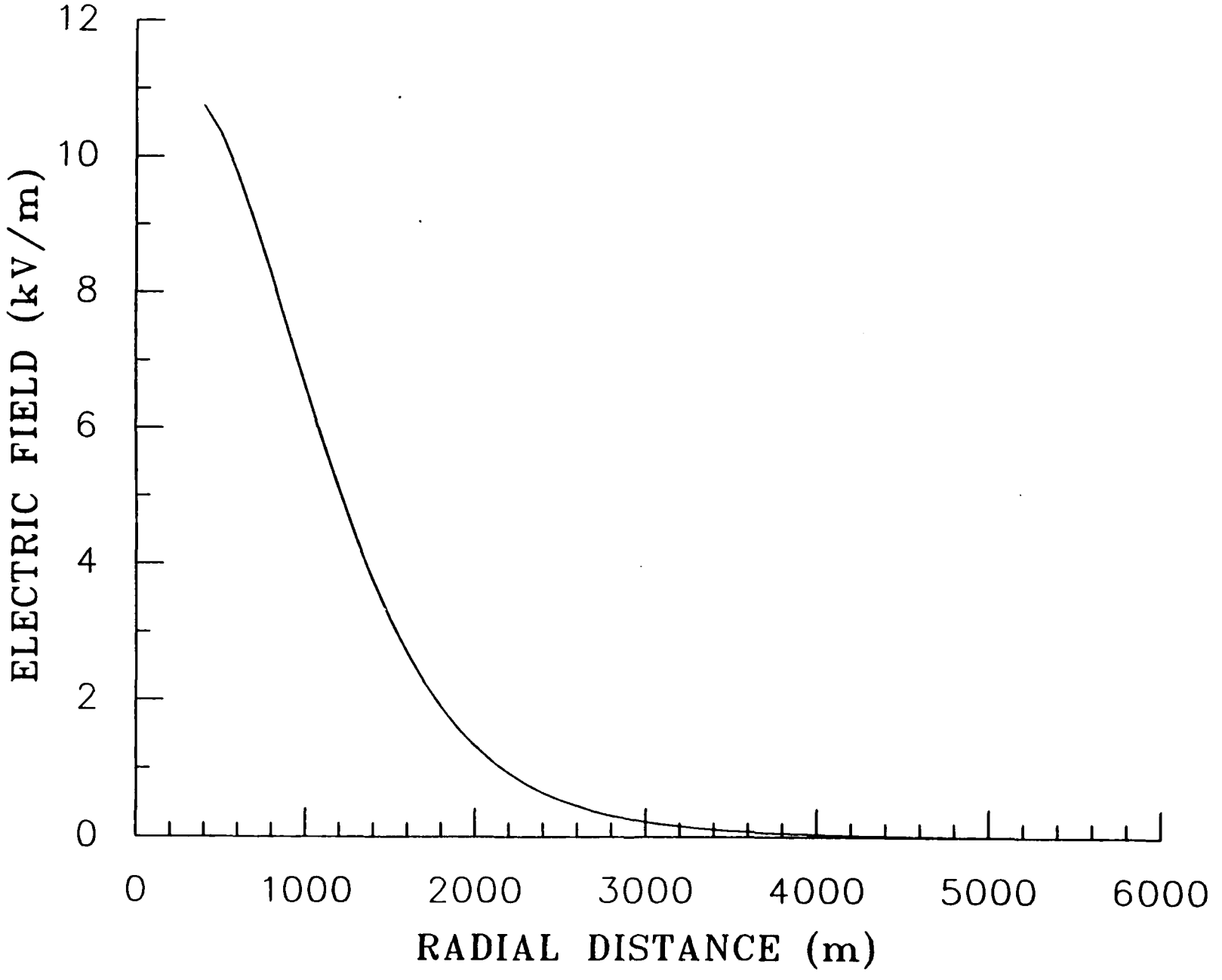


Fig. 5: Magnetic Field for $n = 1$

Fig. 6: Electric Field for $n = 0$



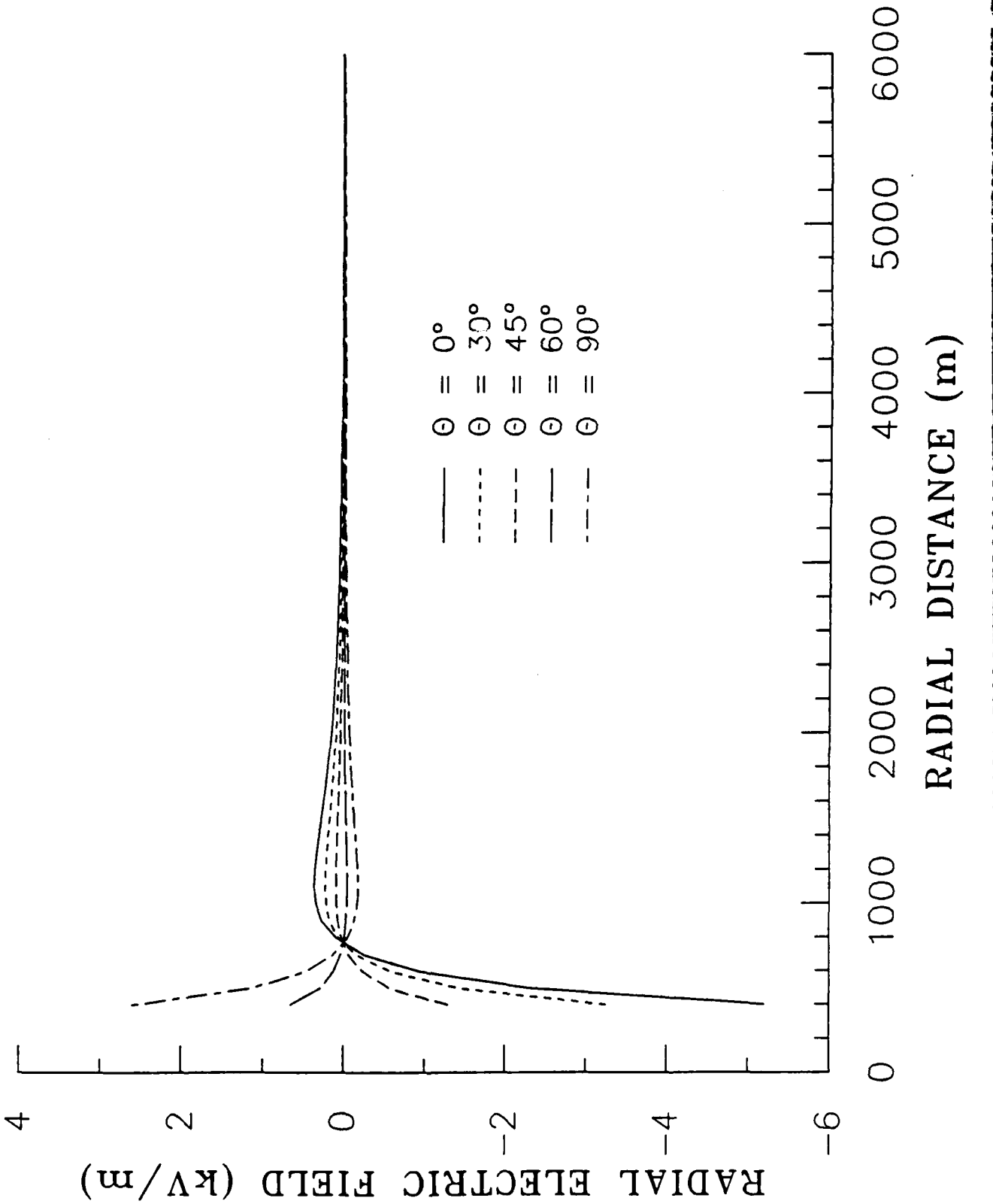


Fig. 7: Radial Electric Field for $n = 2$

Fig. 8: Tangential Electric Field for $n = 2$

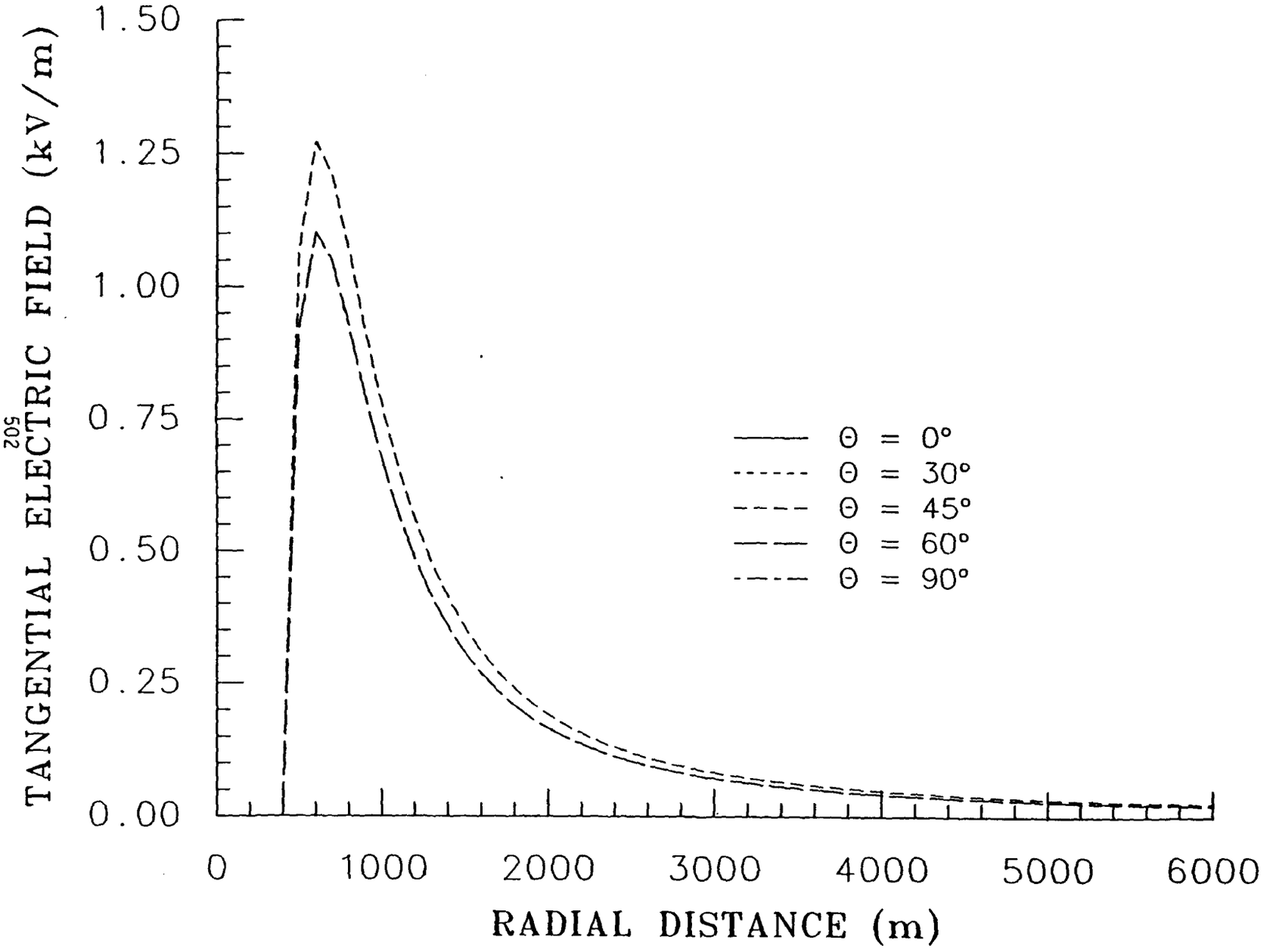


Fig. 9: Electric Potential for $n = 2$

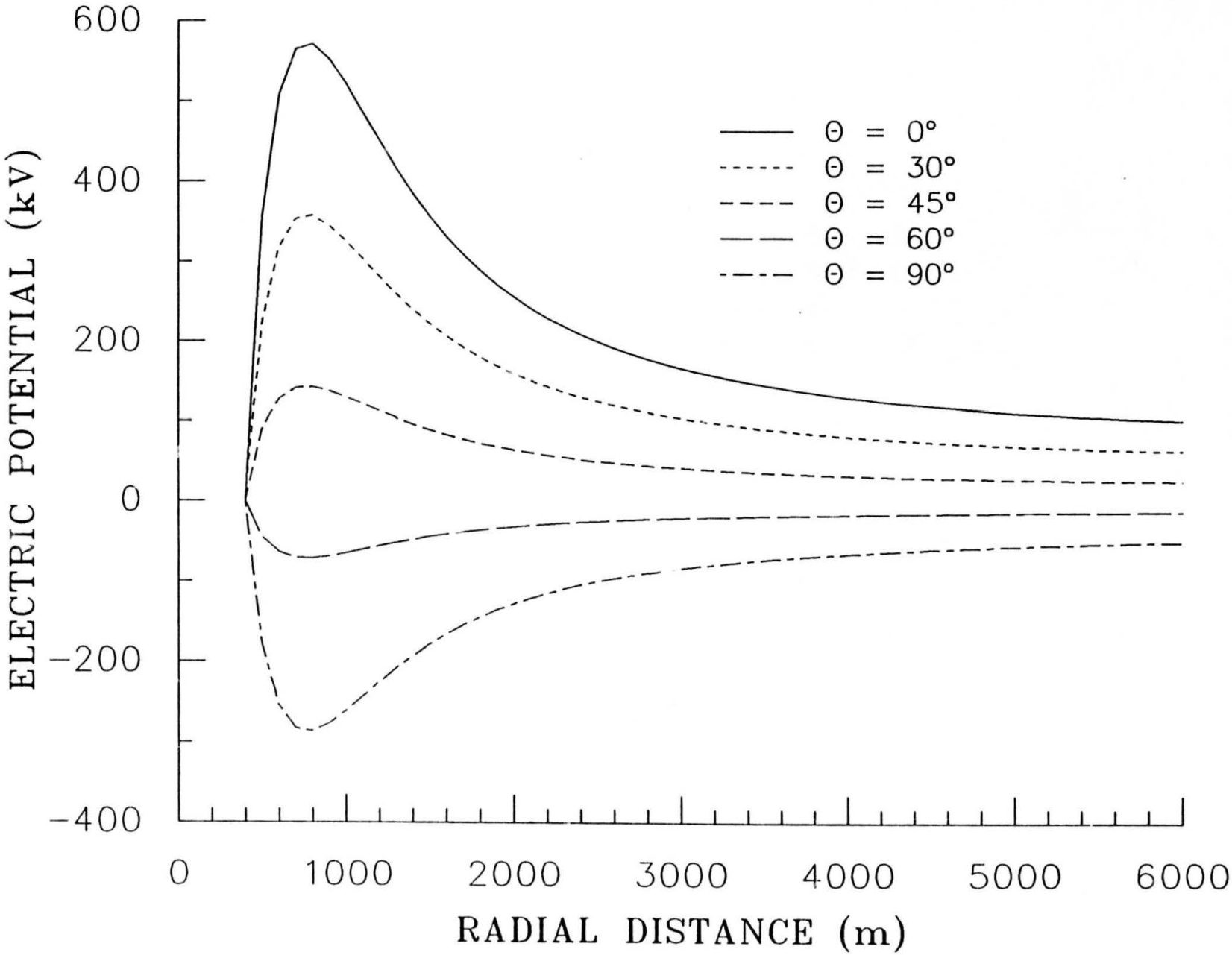
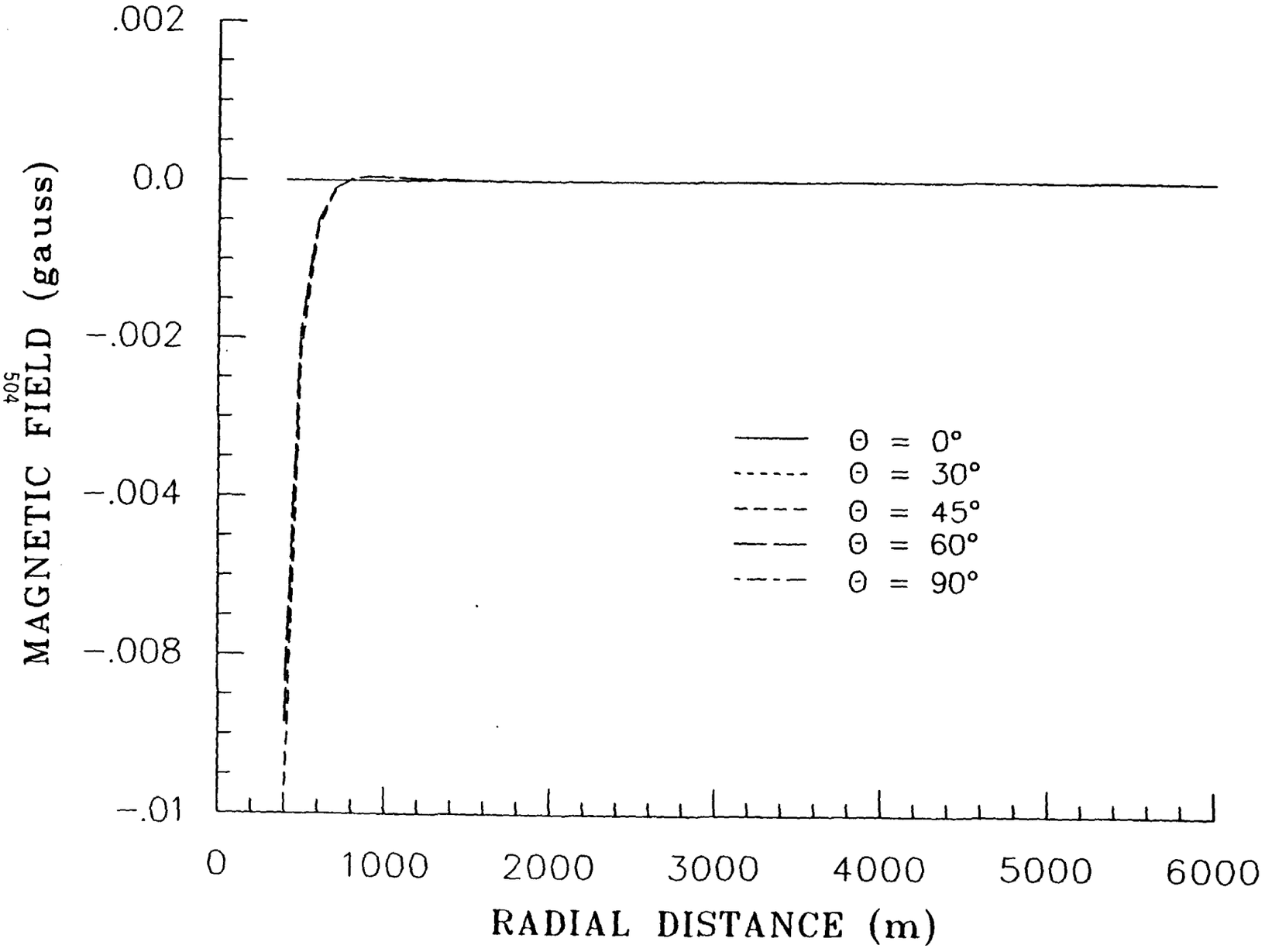


Fig. 10: Magnetic Field for $n = 2$



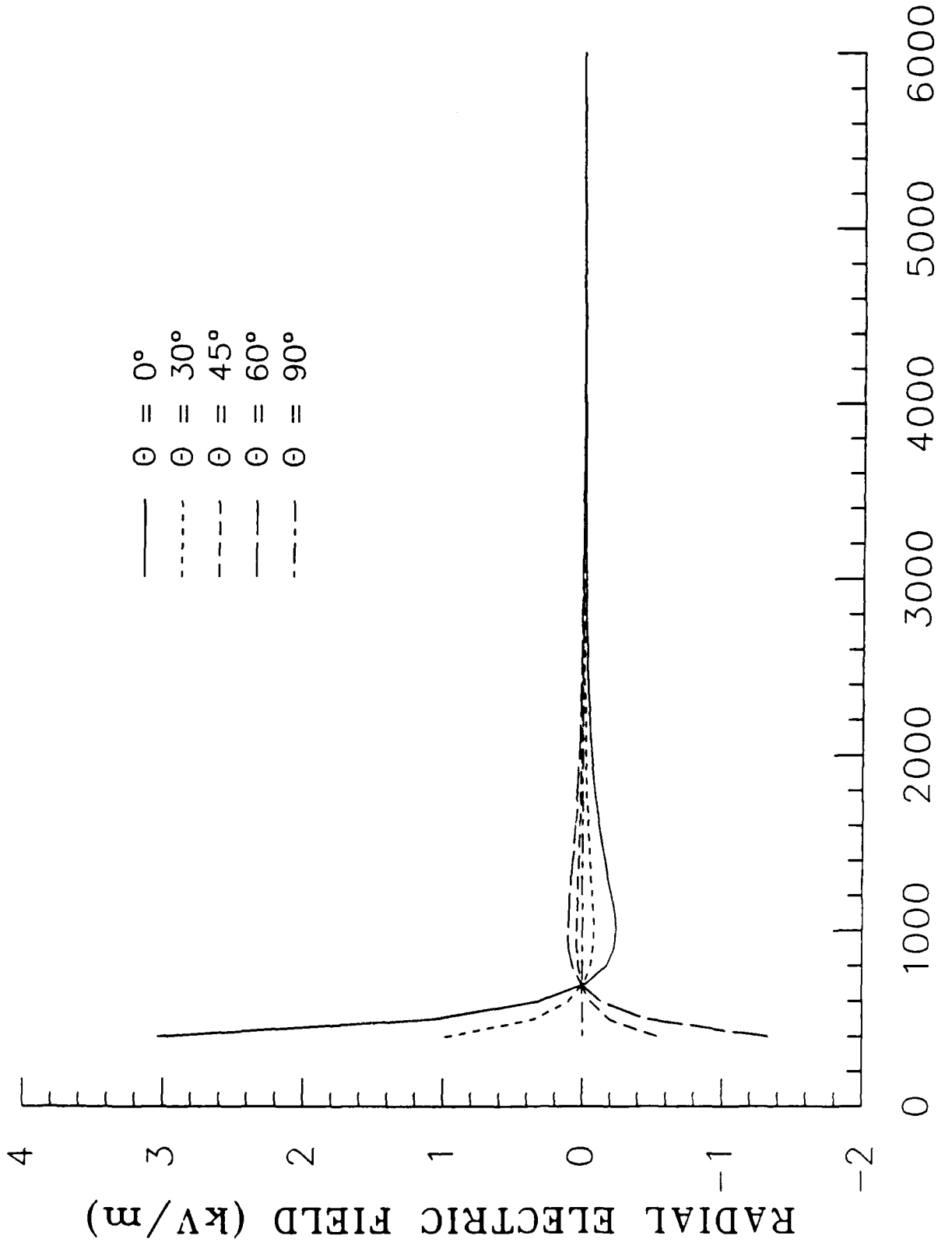


Fig. 11: Radial Electric Field for $n = 3$

Fig. 12: Tangential Electric Field for $n = 3$

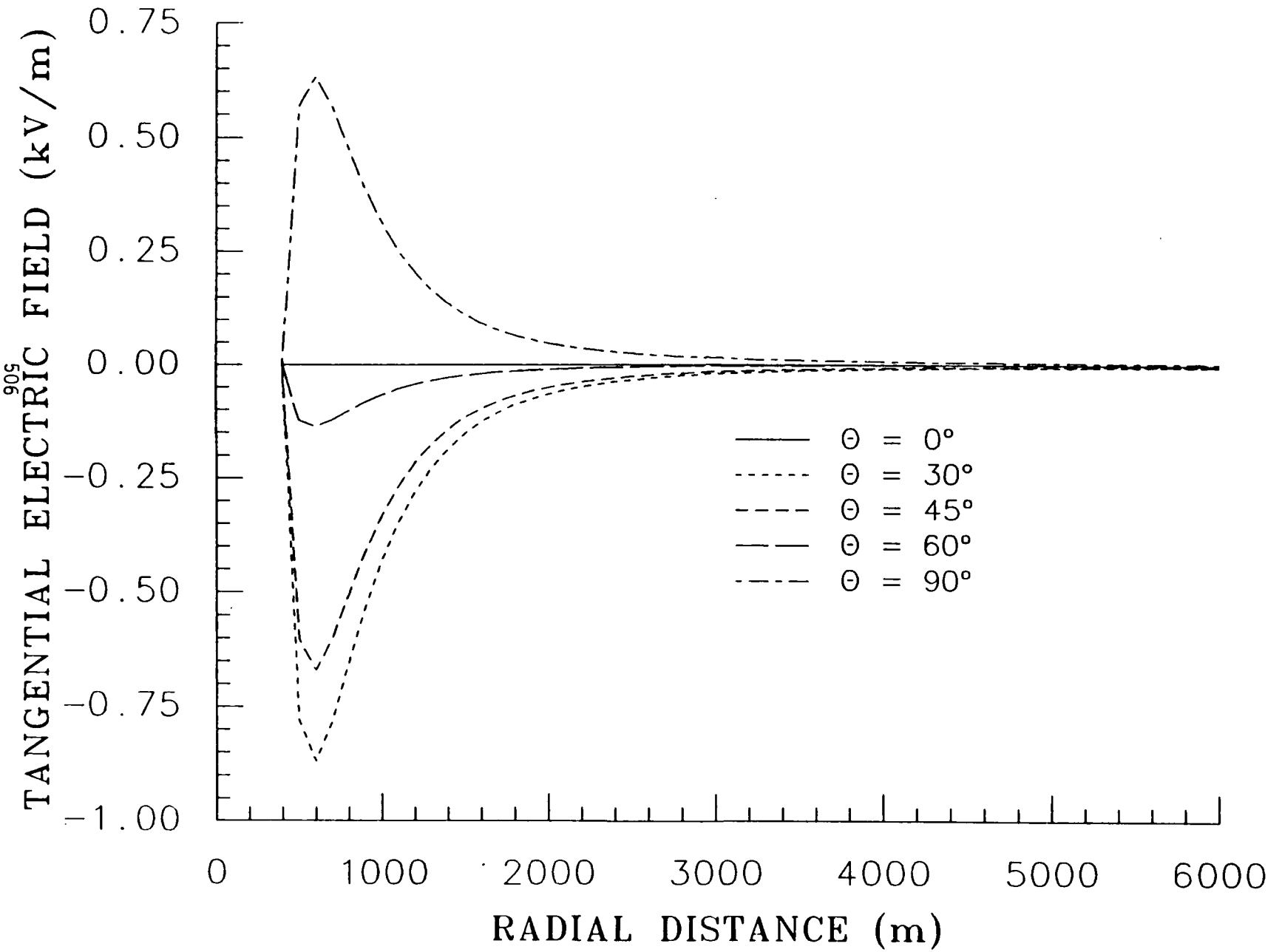
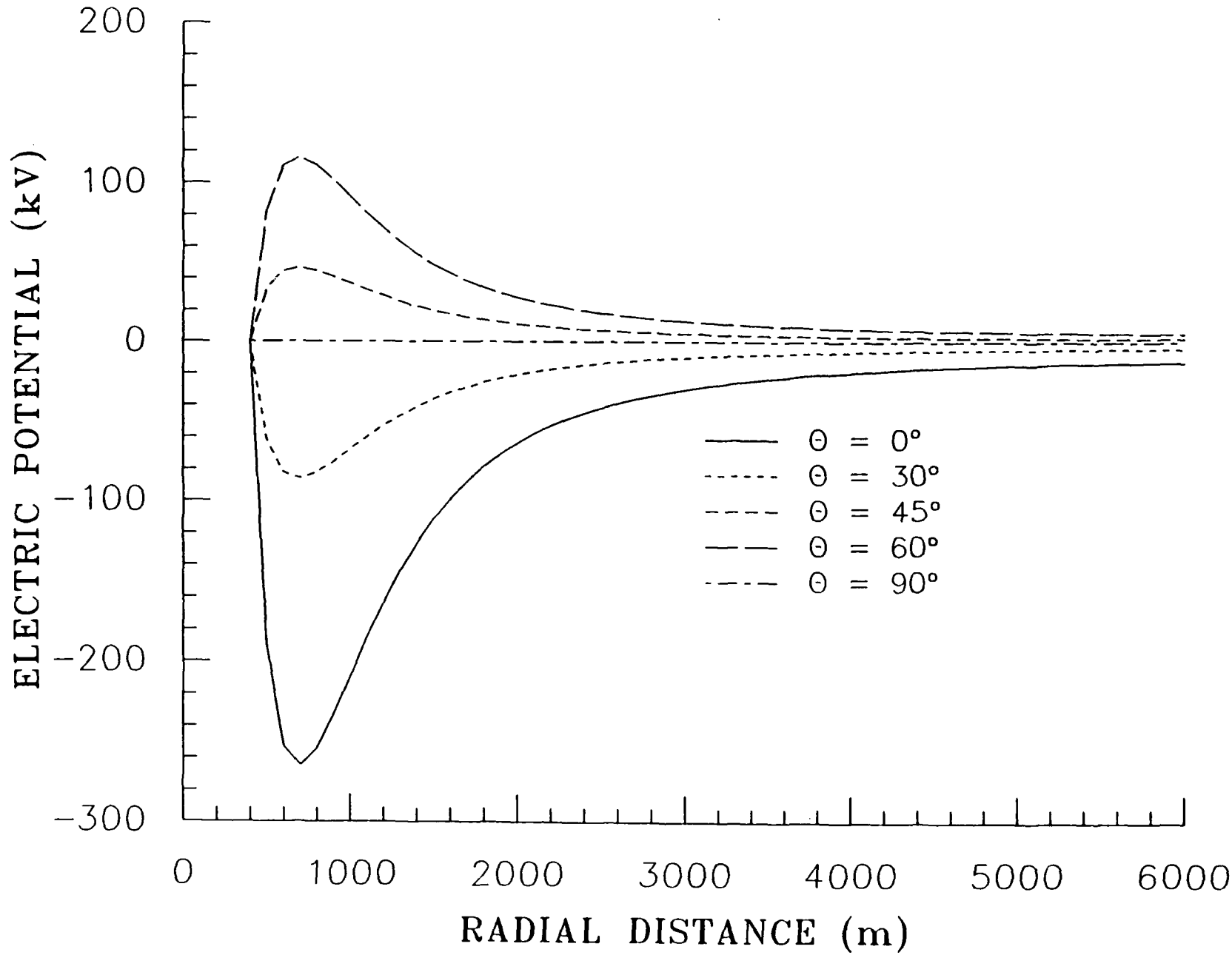


Fig. 13: Electric Potential for $n = 3$



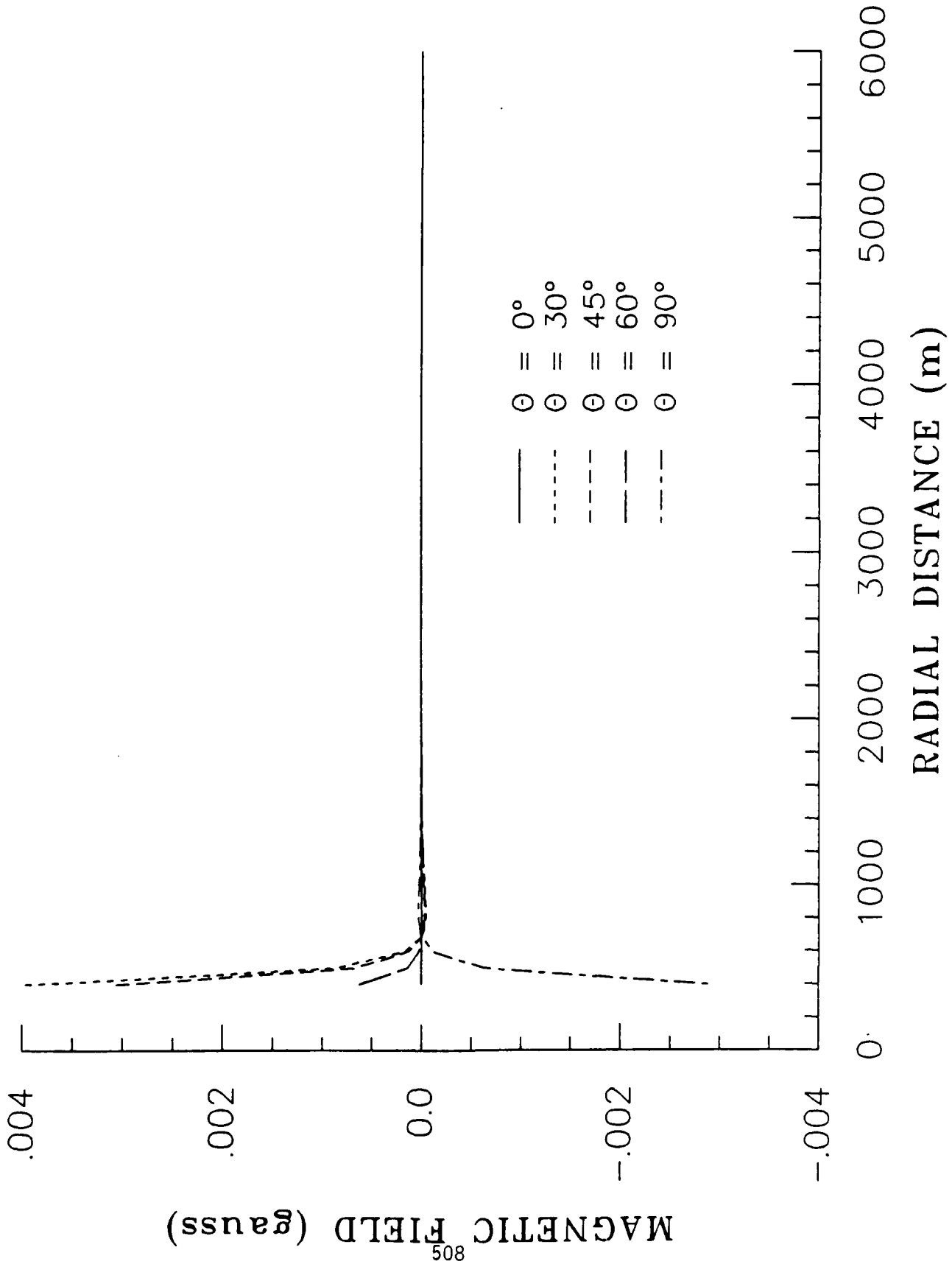


Fig. 14: Magnetic Field for $n = 3$

Fig. 15: Total Radial Electric Field

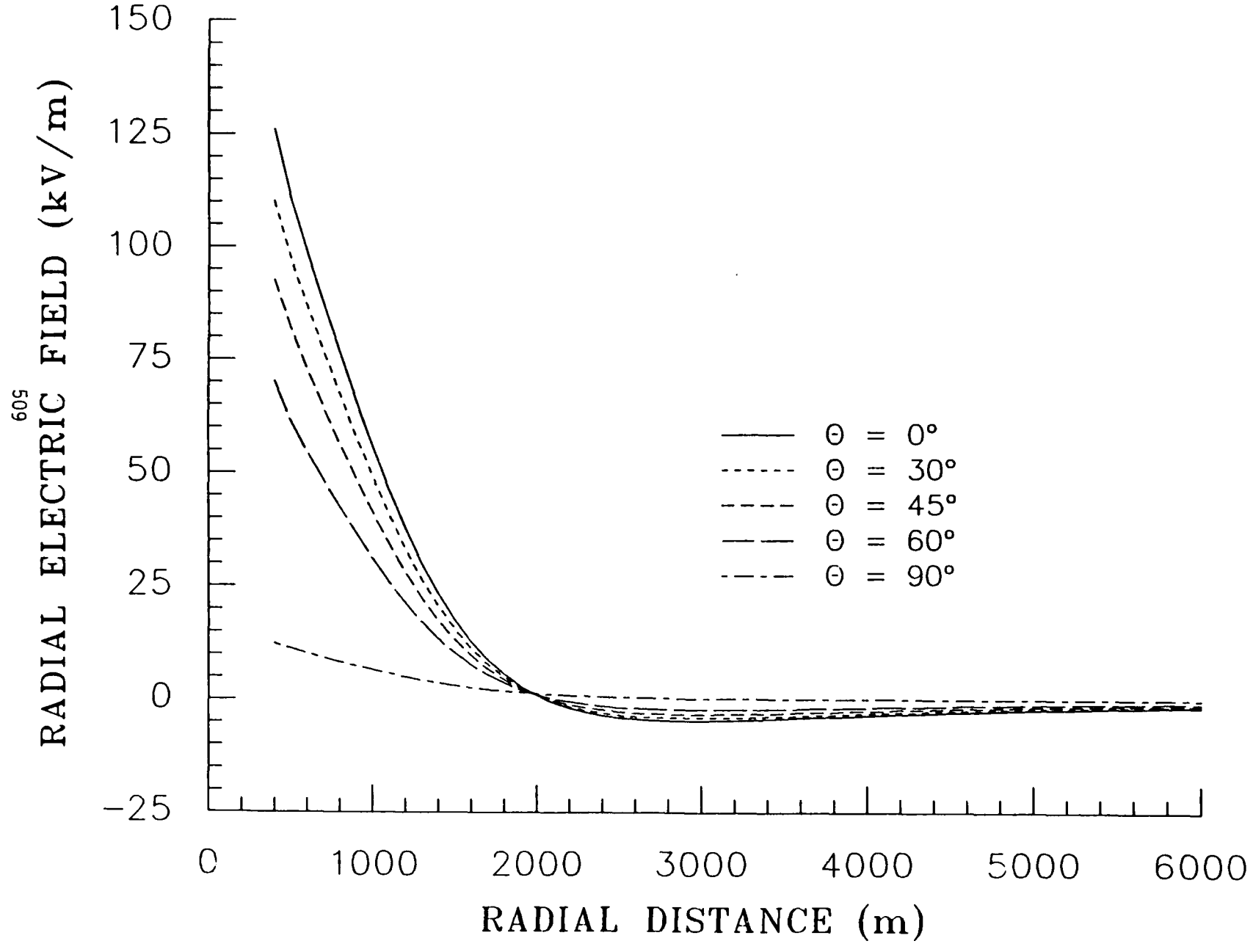


Fig. 16: Total Tangential Electric Field

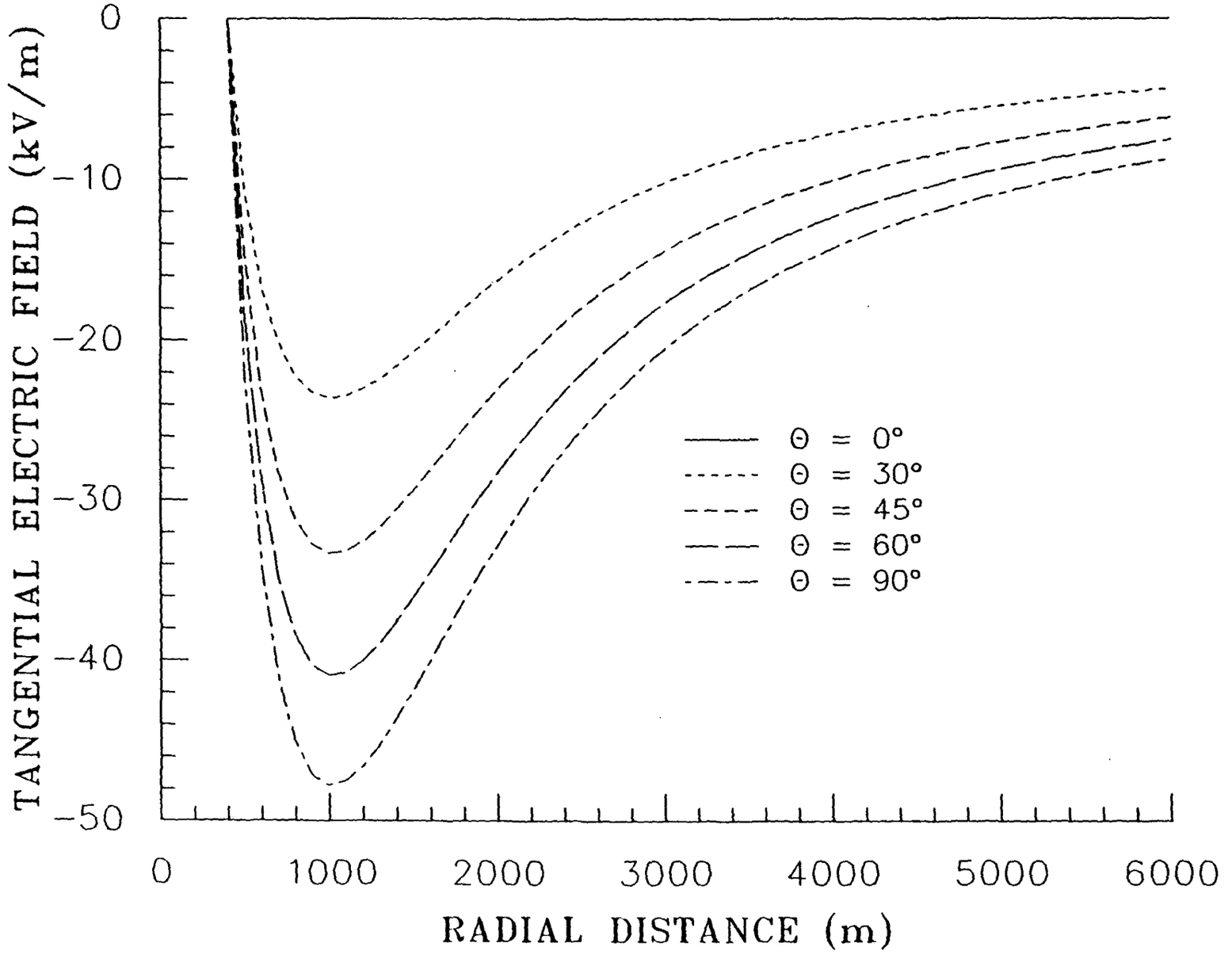
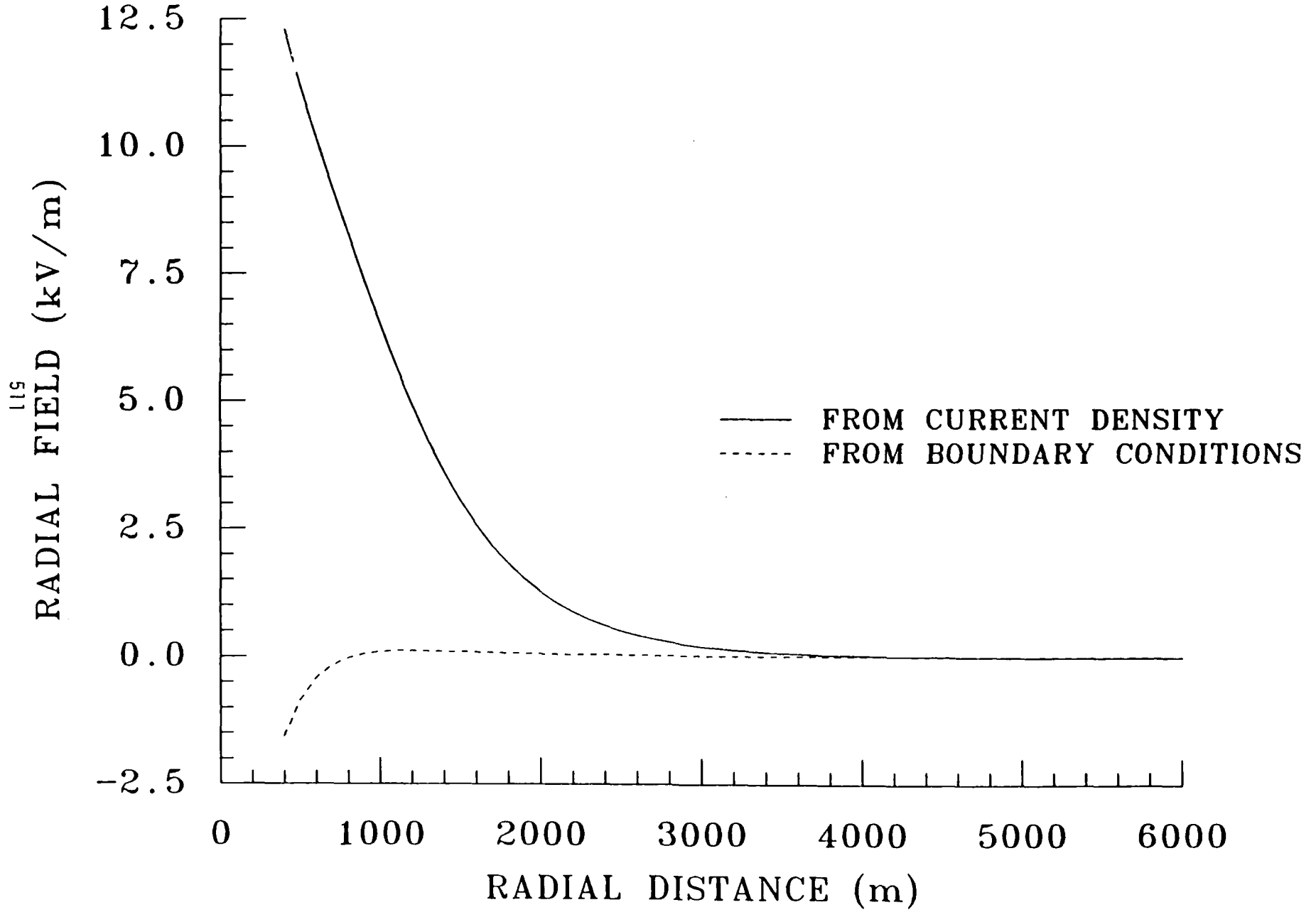


Fig. 17: Calculated Electric Fields for $n = 0$



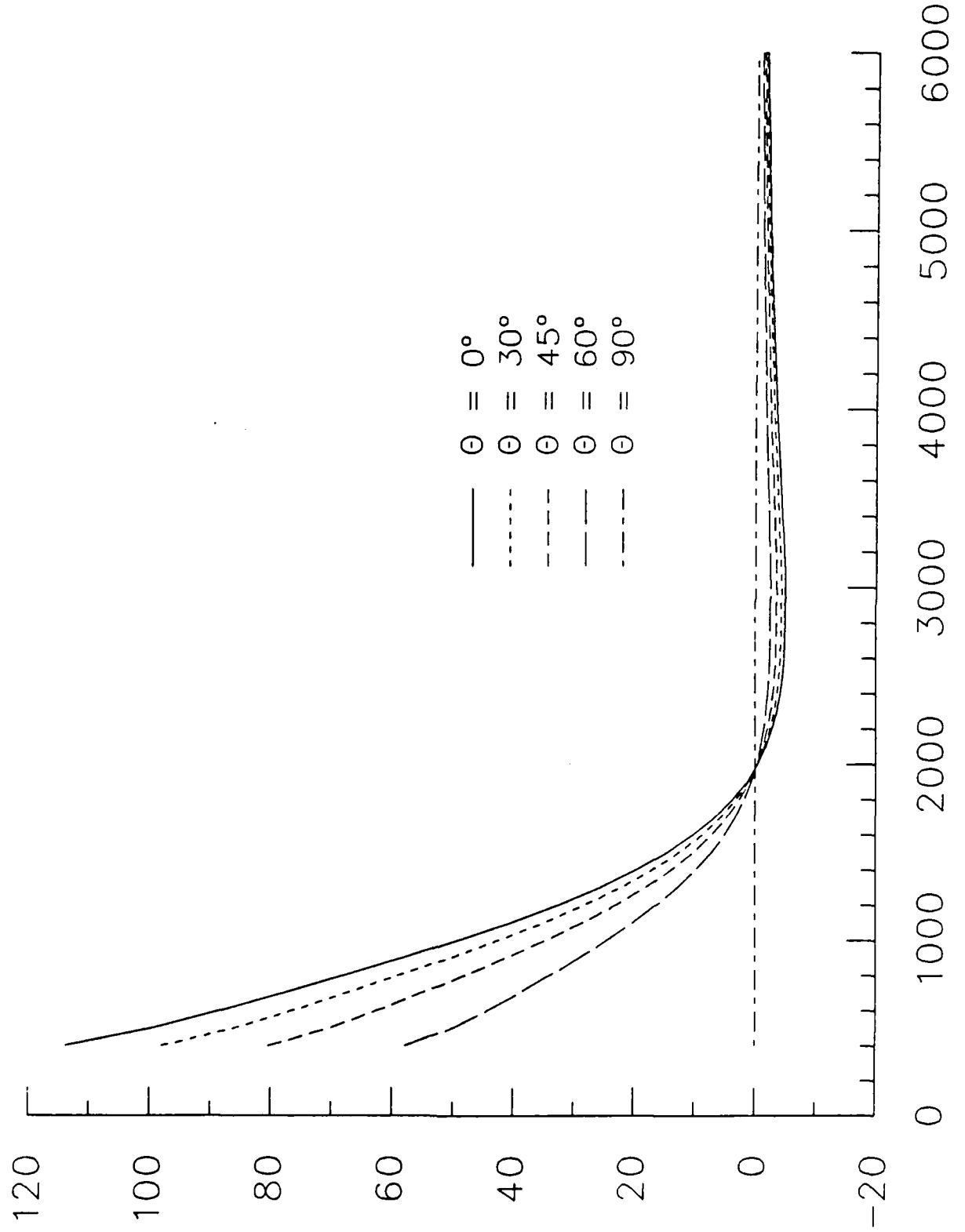


Fig. 18: Corrected Radial Electric Field

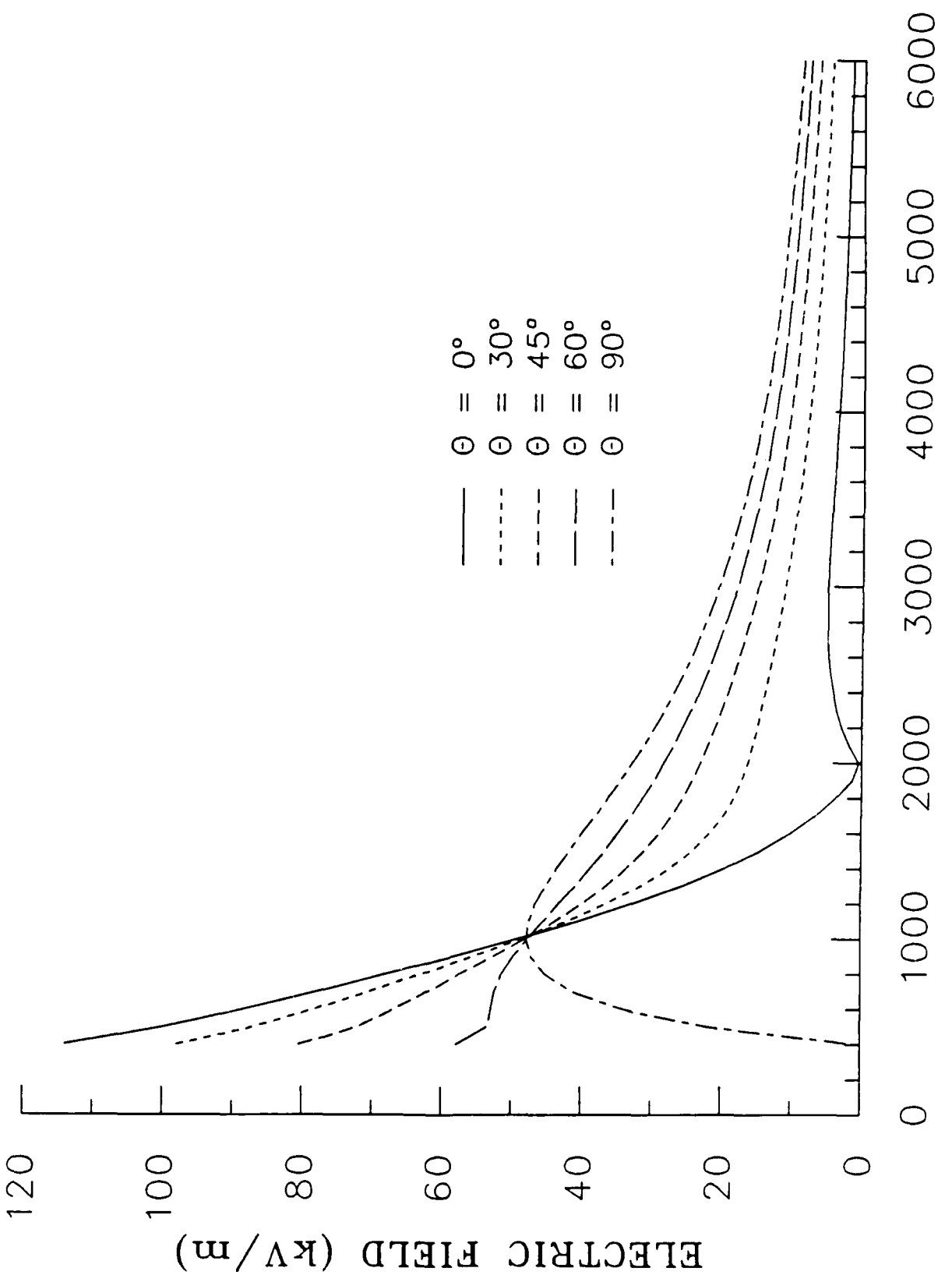


Fig. 19: Corrected Total Electric Field

SPLINE-BASED FINITE-ELEMENT METHOD FOR SOLVING A STEFAN'S PROBLEM IN A FINITE DOMAIN - FORMULATION

Shunsuke Takagi
U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, NH 03755

ABSTRACT. The finite-element method presented in this paper has two features. First, a cubic spline is included in the basis functions; the advantage of the inclusion has not yet been examined. Second, space coordinates only are used to determine the temperature distribution. The time-coordinate increment of the phase front corresponding to the space-coordinate increment can be determined consequently. In our problem, where the final position of the phase front may be determined at the start of the solution, the solution method using a space-coordinate sequence is preferred to the one using a time-coordinate sequence. This numerical method can work smoothly even for an extremely large time.

Analytical formulation only is presented in this paper.

I. INTRODUCTION. Instead of the two end conditions usually employed for determining the two extraneous unknowns, Sections 1 introduces two internal conditions and develops a cubic spline without end conditions. Section IIthen demonstrates that a cubic spline interpolating the unknown temperatures produces a set of roof-shaped basis functions. Equidistant mesh points are used for the derivation.

We have applied this finite element method for solving the freezing of water in a finite domain. The interface is always chosen as a mesh point. The ice and water domains are divided into equilength subregions, with n and m internal points, respectively, where n and m may be any integers larger than or equal to 2. Therefore, the mesh points are never fixed in this method.

Section III states the problem to be solved. The temperature distributions are determined in Section IV by using the interfacial coordinate ξ in place of time t. On the assumption that the temperature distributions at the time substitute ξ are known, Section V finds the simultaneous equations for the unknown temperatures at the time substitute $\xi+d\xi$, showing that they are quadratic. Application of Newton's approximation reduces solving a set of simultaneous quadratic equations to sequentially solving sets of simultaneous linear equations representing tangent planes of the quadratics. The time-coordinate increment dt corresponding to the space-coordinate increment $d\xi$ can be found by use of the expression of $d\xi/dt$.

In our problem, where the terminal temperatures are both given, the final interfacial position can be found at the start of the solution. We can choose, therefore, an appropriate magnitude of increment $d\xi$ at any stage of the solution. The solution method using a space-coordinate sequence is more convenient in our problem than the customary method using a time-coordinate sequence. It was experienced in the numerical computation of the analytical

solution (Ref. 1) that the former works smoothly, even for an extremely large time.

Analysis only is presented in this paper. The advantage of using a spline function in a finite-element method has yet to be clarified.

II. CUBIC SPLINE WITHOUT END CONDITIONS. Instead of the two end conditions that are usually stipulated, we adopt two internal conditions,

$$y'''_{x_1} - 0 \quad = \quad y'''_{x_1} + 0$$

$$y'''_{x_{N-1}} - 0 \quad = \quad y'''_{x_{N-1}} + 0$$

for determining a cubic spline that passes through equally spaced points $P_0(x_0, y_0), \ldots, P_N(x_N, y_N)$. Selecting y''_i, denoted by z_i, at point P_i as unknowns ($i = 0, \ldots, N$), the two internal conditions become

$$z_0 - 2z_1 + z_2 = 0$$

$$z_{N-2} - 2z_{N-1} + z_N = 0 \quad .$$

In other words, we adopt a single cubic passing through points P_0, P_1, P_2 and another through points P_{N-2}, P_{N-1}, P_N. The minimum of N in this stipulation, therefore, is 4, if the two cubics should not be the same. If the two cubics can coincide, N may be 3.

Thus determined, the equations for z_i are:

$$z_0 - 2z_1 + z_2 = 0 \qquad [0]$$

$$z_0 + 4z_1 + z_2 = A_1 \qquad [1]$$

$$z_1 + 4z_2 + z_3 = A_2 \qquad [2]$$

$$\vdots \qquad \vdots$$

$$z_{N-3} + 4z_{N-2} + z_{N-1} = A_{N-2} \qquad [N-2]$$

$$z_{N-2} + 4z_{N-1} + z_N = A_{N-1} \qquad [N-1]$$

$$z_{N-2} - 2z_{N-1} + z_N = 0 \ , \qquad [N]$$

where

$$A_i = 6(y_{i-1} - 2y_i + y_{i+1})/(\Delta x)^2 \qquad (1)$$

for $i = 1,\ldots, N-1$. The equal distance between two adjacent mesh points is denoted by Δx.

The general solution of the above equations can be found for $N \geq 6$. Define

$$f_1 = 1/6 \ , \quad f_2 = 1 \ , \quad f_3 = 4 \ ,$$

and

$$f_i = 4f_{i-1} - f_{i-2} \qquad (2)$$

recursively for $i \geq 4$. The difference equation (2) is solved to:

$$f_i = (1/(2\sqrt{3}))\{(2 + \sqrt{3})^{i-1} - (2 - \sqrt{3})^{i+1}\}$$

$$= \sum_{p=0}^{[(i-2)/2]} \binom{i-1}{2p+1} 2^{i-2-2p}\, 3^p \ , \qquad (3)$$

which happens to be valid for $i \geq 2$.

When $N = 2r$ with $r \geq 3$, we may, as proved below, transform the simultaneous equations to:

$$z_0 = 2z_1 - z_2 \qquad [0]'$$

$$z_1 = f_1 A_1 \qquad [1]'$$

$$f_3 z_2 + f_2 z_3 = \sum_{1}^{2} (-1)^{2-i} f_i A_i \qquad [2]'$$

$$\vdots \qquad\qquad \vdots$$

$$f_r z_{r-1} + f_{r-1} z_r = \sum_{1}^{r-1} (-1)^{r-1-i} f_i A_i \qquad [r-1]'$$

$$z_{r-1} + 4z_r + z_{r+1} = A_r \qquad [r]'$$

$$f_{r-1}\, z_r + f_r\, z_{r+1} = \sum_{1}^{r-1} (-1)^{r-1-i}\, f_i\, A_{N-i} \qquad [r+1]'$$

$$\vdots \qquad\qquad \vdots$$

$$f_2\, z_{N-3} + f_3\, z_{N-2} = \sum_{1}^{2} (-1)^{2-i}\, f_i\, A_{N-i} \qquad [N-2]'$$

$$z_{N-1} = f_1\, A_{N-1} \qquad [N-1]'$$

$$z_N = 2z_{N-1} - z_{N-2}\ . \qquad [N]'$$

The relation [0]' is found from [0]. The relation [1]' is found by subtracting [0] from [1]. Substituting the relation [1]' into [2] , we find [2]' . Assume that the relation

$$f_{k+1}\, z_k + f_k\, z_{k+1} = \sum_{1}^{k} (-1)^{k-i}\, f_i\, A_i$$

is valid for $k \geq 2$. Substituting z_k by

$$z_k = -4z_{k+1} - z_{k+2} + A_{k+1}\ ,$$

we find

$$(4f_{k+1} - f_k)z_{k+1} + f_{k+1}\, z_{k+2} = \sum_{1}^{k+1} (-1)^{k+1-i}\, f_i\, A_i\ .$$

Therefore if (2) is true, [r-1]' must also be true. We apply the same procedure from below starting at [N-1]' , and find [r+1]' .

We solve the three equations in the middle for z_{r-1} , z_r , and z_{r+1} . Then the rest of the unknowns can be successively determined. We express the solution in terms of a (N+1) by (N+1) matrix a_i^k defined by

$$(\Delta x)^2\, z_i = \sum_{0}^{N} a_i^k\, y_k\ , \quad i \in \{0,\ldots,N\}\ . \qquad (4)$$

Here a set-theoretic notation (Ref. 2,3) is used to show that i is a member of the set of numbers enclosed in a pair of braces. Similar notations will be used henceforth. The solution is shown in Appendix A.

When $N = 2r + 1$ with $r \geq 3$, we transform the equations [0]...[N] to

$$z_0 = 2z_1 - z_0 \qquad [0]''$$

$$z_1 = f_1 A_1 \qquad [1]''$$

$$f_3 z_2 + f_2 z_3 = \sum_1^2 (-1)^{2-i} f_i A_i \qquad [2]''$$

$$\vdots \qquad \vdots$$

$$f_{r+1} z_r + f_r z_{r+1} = \sum_1^r (-1)^{r-i} f_i A_i \qquad [r]''$$

$$f_r z_r + f_{r+1} z_{r+1} = \sum_1^r (-1)^{r-i} f_i A_{N-i} \qquad [r+1]''$$

$$\vdots \qquad \vdots$$

$$f_2 z_{N-3} + f_3 z_{N-2} = \sum_1^2 (-1)^{2-i} f_i A_{N-i} \qquad [N-2]''$$

$$z_{N-1} = f_1 A_{N-1} \qquad [N-1]''$$

$$z_N = 2z_{N-1} - z_{N-2} \; . \qquad [N]''$$

Following the same procedure as the preceding, we have in this case the two equations in the middle, which we solve for z_r and z_{r+1}. Then the rest of the unknowns can be successively determined.

The result is shown in Appendix A in terms of the matrix a_i^k. The cases N = 3, 4 and 5 are also listed in Appendix A. The reason for the restriction $r \geq 3$ is exhibited in the Appendix.

III. BASIS FUNCTIONS. We denote the cubic spline in an interval $[x_i, x_{i+1}]$ by $y(x \in [x_i, x_{i+1}])$. Given y_i, y_{i+1}, z_i, z_{i+1}, it is formulated (Ref. 4),

$$y(x\in[x_i, x_{i+1}]) = y_i \frac{x_{i+1} - x}{\Delta x} - y_{i+1} \frac{x - x_i}{\Delta x} -$$

$$- \frac{(\Delta x)^2}{6} z_i \left\{\frac{x_{i+1} - x}{\Delta x} - \left(\frac{x_{i+1} - x}{\Delta x}\right)^3\right\} - \qquad (5)$$

$$- \frac{(\Delta x)^2}{6} z_{i+1} \left\{\frac{x - x_i}{\Delta x} - \left(\frac{x - x_i}{\Delta x}\right)^3\right\} \qquad \text{for } x\in[x_i , x_{i+1}]$$

$$= 0 . \qquad \text{for } x\notin[x_i , x_{i+1}]$$

A pair of brackets enclosing two points, like those at right above, mean a closed interval spanned by the points.

The cubic spline in the whole domain $[x_0 , x_N]$ is given by

$$y(x\in[x_0, x_N]) = \sum_{i=0}^{N-1} y(x\in[x_i, x_{i+1}]) . \qquad (6)$$

Define

$$p^j(x\in[x_i, x_{i+1}]) = -\frac{1}{6}\left[\frac{x_{i+1} - x}{\Delta x} - \left(\frac{x_{i+1} - x}{\Delta x}\right)^3\right] a_i^j -$$

$$- \frac{1}{6}\left[\frac{x - x_{i+1}}{\Delta x} - \left(\frac{x - x_i}{\Delta x}\right)^3\right] a_{i+1}^j , \qquad \text{for } x\in[x_i,x_{i+1}]$$

$$= 0 , \qquad \text{for } x\notin[x_i,x_{i+1}]$$

where

$$j\in\{0,\ldots, N\} , \qquad i\in\{0,\ldots, N-1\} .$$

Then (6) may become

$$y(x\in[x_0,\ x_N]) = \sum_{i=0}^{N-1} \{y_i \frac{x_{i+1} - x}{\Delta x} + y_{i+1} \frac{x - x_i}{\Delta x} + \sum_{j=0}^{N} y_j\ p^j(x)\}_{x\in[x_i,\ x_{i+1}]}$$

The first derivatives of $p^j(x)$ are in general discontinuous at the mesh points. We rewrite the above to

$$y(x\in[x_0,\ x_N]) = \sum_{i=0}^{N} y_i\ B^i(x) \tag{8}$$

by defining the basis functions

$$B^i(x) = (\frac{x - x_{i-1}}{\Delta x})_{x\in(x_{i-1},\ x_i)} + (\frac{x_{i+1} - x}{\Delta x})_{x\in(x_i,\ x_{i+1})} +$$

$$+ \sum_{h=0}^{N-1} p^i(x\in[x_h,\ x_{h+1}]) \quad , \tag{9}$$

with a convention that

$$(\frac{x - x_{-1}}{\Delta x})_{x\in(x_{-1},\ x_0)} = 0$$

and $\tag{10}$

$$(\frac{x_{N+1} - x}{\Delta x})_{x\in(x_N,\ x_{N+1})} = 0$$

and another that the value at $x = x_i$ should be found as the limit $x \to x_i - 0$ or $x \to x_i + 0$, where a pair of parentheses enclosing two points mean an open interval spanned by them. The basis function satisfies

$$B^i(x_j) = \delta^i_j \quad , \tag{11}$$

where δ^i_j is Kronecker's δ .

IV. PROBLEM. We consider freezing of water in a finite domain $0 \le X \le \ell$. The boundary temperatures T_A at $X = 0$ and T_B at $X = \ell$ are constant, the latter being also the initial temperature. They satisfy the condition, $T_A < T_F \le T_B$, where T_F is the phase change temperature.

At $t = 0$, ice emerges at $X = 0$, whose temperature we express by $T^I(X, \kappa_I t)$, where κ_I is the thermal diffusivity of the ice. The phase front is denoted by $s(t)$. The temperature of the water is expressed by $T^W(X, \kappa_W t)$.

We introduce three nondimensional coordinates and one nondimensional constant,

$$x = X/\ell, \quad \tau = \kappa_I t/\ell^2, \quad \xi = s(t)/\ell, \quad \beta = \kappa_W/\kappa_I \ . \tag{12}$$

Then the heat equations may become:

$$\frac{\partial T^I}{\partial \xi} \frac{d\xi}{d\tau} - \frac{\partial^2 T^I}{\partial x^2} = 0 \quad ,$$

and (13)

$$\frac{\partial T^W}{\partial \xi} \frac{d\xi}{d\tau} - \beta \frac{\partial^2 T^W}{\partial x^2} = 0 \quad .$$

They are subject to the conditions:

$$T^I(0) = T_A \tag{14.1}$$

$$T^W(1) = T_B \tag{14.2}$$

$$T^I(\xi) = T^W(\xi) = T_F \tag{14.3}$$

$$\xi(0) = 0 \tag{14.4}$$

$$T^W(x,0) = T_B \tag{14.5}$$

$$\frac{d\xi}{d\tau} = (C_I/L) \left(\frac{\partial T^I}{\partial x}\right)_F - (\beta C_W/L) \left(\frac{\partial T^W}{\partial x}\right)_F \ , \tag{14.6}$$

where C_I and C_W are heat contents of ice and water, respectively, and L is the latent heat.

V. TEMPERATURES. Choosing the phase front as a mesh point, we insert equally spaced n and m internal mesh points in the ice and water domains, creating $N = n + 1$ and $M = m + 1$ subintervals, respectively. Substituting ξ for the time coordinate τ, we express the unknown temperatures at the internal mesh points by $T_i(\xi)$, where indexes $i = 1,\ldots, n$ are for the ice and indexes $i = n+1,\ldots, n+m$ for the water. Then we have $\Delta x = \xi/N$ and $(1-\xi)/M$ in the ice and water domains, respectively.

Using the basis functions the temperatures of ice and water are given by

$$T^I(x, \xi) = B_I^0(x)\ T_A + \sum_1^n B_I^i(x)\ T_i(\xi) + B_I^N(x)\ T_F\ , \qquad (15.1)$$

and

$$T^W(x, \xi) = B_W^0(x)\ T_F + \sum_1^m B_W^i(x)\ T_{n+i}(\xi) + B_W^M(x)\ T_B\ , \qquad (15.2)$$

where $B_I^k(x)$ and $B_W^k(x)$ are basis functions in the ice and water domains, respectively. Eq (4) is expressed by

$$\left(\frac{\xi}{N}\right)^2 z_k^I = a_k^0\ T_A + \sum_1^n a_k^i\ T_i(\xi) + a_k^N\ T_F \qquad k \in \{0, \ldots, N\} \qquad (16.1)$$

and

$$\left(\frac{1-\xi}{M}\right)^2 z_k^W = b_k^0\ T_F + \sum_1^m b_k^i\ T_{n+i}(\xi) + b_k^M\ T_B \qquad k \in \{0, \ldots, M\} \qquad (16.2)$$

in the ice and water domains respectively. In the latter, notation b_k^i is used instead of a_k^i . The interfacial condition (14.6) becomes

$$\frac{d\xi}{d\tau} = \frac{C_I}{L}\frac{N}{\xi}\left\{T_F - T_n(\xi) + \frac{1}{6}\left(a_n^0 + 2a_N^0\right) T_A\ +\right.$$

$$\left. +\ \frac{1}{6}\sum_1^n \left(a_n^i + 2a_N^i\right)T_i(\xi) + \frac{1}{6}\left(a_n^N + 2a_N^N\right)T_F\right\}\ +$$

$$+\ \frac{\beta C_W}{L}\frac{M}{1-\xi}\left\{T_F - T_{n+1}(\xi) + \frac{1}{6}\left(2b_0^0 + b_1^0\right)T_F\ +\right.$$

$$\left. +\ \frac{1}{6}\sum_1^m \left(2b_0^i + b_1^i\right)T_i(\xi) + \frac{1}{6}\left(2b_0^M + b_1^M\right)T_B\right\}\ , \qquad (17)$$

which is found by substituting $\left(\partial T^I/\partial x\right)_F$ and $\left(\partial T^W/\partial x\right)_F$ in (14.6) with

$$\left(\frac{\partial T^I}{\partial x}\right)_F = \frac{T_F - T_n(\xi)}{\xi/N} + \frac{\xi/N}{6}\left(z_{n-1}^I + 2z_n^I\right)$$

and

$$\left(\frac{\partial T^W}{\partial x}\right)_F = \frac{T_{n+1}(\xi) - T_F}{(1-\xi)/M} - \frac{(1-\xi)/M}{6}(2z_0^W + z_1^W) \quad ,$$

respectively, and using the equations in (16).

VI. FINITE-ELEMENT EQUATIONS. We compute the integrals,

$$\int_0^{\xi} \left(\frac{\partial T^I}{\partial \xi}\frac{d\xi}{d\tau} - \frac{\partial^2 T^I}{\partial x^2}\right) B_I^j(x)dx = 0 \quad , \quad j \in \{1, \cdots, n\} \quad , \tag{18.1}$$

and

$$\int_{\xi}^{1} \left(\frac{\partial T^W}{\partial x}\frac{d\xi}{d\tau} - \beta\frac{\partial^2 T^W}{\partial x^2}\right) B_W^j(x)dx = 0 \quad , \quad j \in \{n+1, \cdots, n+m\} \quad . \tag{18.2}$$

On the condition that the temperature distributions at the time substitute ξ are known, we rewrite the result of the integrations to the difference equations at the time substitute $\xi + \frac{1}{2}\Delta\xi$. Letting

$$y_k = T_k(\xi + \Delta\xi) \quad \text{for } k = 1, \ldots, n+m \quad ,$$

we find that the difference equations are quadratic,

$$\sum_{k=1}^{m+n} \sum_{h=1}^{m+n} \alpha_j^{kh} y_k y_h + \sum_{k=1}^{m+n} \beta_j^k y_k + \gamma_j = 0 \quad , \tag{19}$$

where $j \in \{1, \ldots, n+m\}$. It is noted that α_j^{kh} is not symmetric with regard to k and h. The process of obtaining these coefficients is shown in the next section.

To find the solution for y_k, let $y_k^{(\nu)}$ be the νth approximation. Then, applying Newton's method, the $(\nu+1)$th approximation is found by solving a set of simultaneous linear equations,

$$\sum_{h=1}^{m+n} c_j^k y_k^{(\nu+1)} = d_j \quad , \tag{20}$$

where

$$c_j^k = \sum_{h=1}^{m+n} (\alpha_j^{kh} + \alpha_j^{hk}) z_h^{(\nu)} + \beta_j^k$$

and (21)

$$d_j = \frac{1}{2} \sum_{k=1}^{m+n} \sum_{h=1}^{m+n} (\alpha_j^{kh} + \alpha_j^{hk}) z_k^{(\nu)} z_h^{(\nu)} - \gamma_j \ .$$

The time interval $\Delta\tau$ corresponding to the time-substitute interval $\Delta\xi$ can be found by use of (17).

VII. COEFFICIENTS. Integrations of the equations in (18) become simpler if the delta and epsilon notations introduced in the following by (22) and (31), respectively, are employed.

By use of the delta notations, defined by

$$\delta(k;\ i-1,\ i) = 1 \ , \quad \text{if} \quad k = i-1 \ , \tag{22.1}$$

$$= -1 \ , \quad \text{if} \quad k = i \ , \tag{22.2}$$

$$= 0 \ , \quad \text{if} \quad k \notin \{i-1,\ i\} \ , \tag{22.3}$$

where $k \in \{0, \ldots, N-1\}$, we give a unified expression to the derivatives of the basis functions,

$$\frac{dB^i}{dx} \left(x \in (x_k,\ x_{k+1})\right) = \frac{1}{\Delta x} \delta(k;\ i-1,\ i) + \frac{dp^i}{dx} \left(x \in (x_k,\ x_{k+1})\right) \ . \tag{23}$$

Because $k \in \{0, \ldots, N-1\}$, (22.1) and (22.2) are not applicable if $i = 0$ and N, respectively.

The values at the mesh points $x_k + 0$ and $x_{k+1} - 0$ are:

$$\frac{dB^i}{dx} \left(x \in (x_k,\ x_{k+1})\right)\Big|_{x_k + 0}$$

$$= \frac{1}{\Delta x} \left\{\delta(k;\ i-1,\ i) - \frac{1}{6} (2a_k^i + 2_{k+1}^i)\right\} \tag{24}$$

and

$$\frac{dB^i}{dx} \left(x \in (x_k,\ x_{k+1})\right)\Big|_{x_{k+1} - 0}$$

$$= \frac{1}{\Delta x} \left\{\delta(k+1;\ i,\ i+1) + \frac{1}{6} (a_k^i + 2a_{k+1}^i)\right\} \ . \tag{25}$$

Including the terminal temperatures, we rewrite (15.1) to

$$T^I(x,\xi) = \sum_{i=0}^{N} B^i(x)\, T_i^I(\xi) \ , \tag{26}$$

where we do not attach subscript I to $B^i(x)$ for simplicity.

Because

$$\sum_{k=0}^{N-1} \left[\frac{dB^i(x)}{dx}\, B^j(x)\right]_{x_k}^{x_{k+1}} = 0 \ , \tag{27}$$

where $i \in \{0,.., N\}$ and $j \in \{1,.., N-1\}$, the partial integration of (18.1) yields

$$\frac{d\xi}{dt} \sum_{i=0}^{N} \frac{dT_i^I}{d\xi} \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} B^i(x)\, B^j(x)\, dx \ +$$

$$+ \sum_{i=0}^{N} T_i^I(\xi) \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} \frac{dB^i(x)}{dx}\, \frac{dB^j(x)}{dx}\, dx = 0 \ . \tag{28}$$

To prove (27), use (11), (A.2), (A.3), and (A.4), the latter three of which are in the Appendix.

Use of (23) yields

$$\int_{x_k}^{x_{k+1}} \frac{dB^i(x)}{dx}\, \frac{dB^j(x)}{dx}\, dx$$

$$= \frac{1}{\Delta x}\, \delta(k;\, i-1,\, i)\, \delta(k;\, j-1,\, j) +$$

$$+ \int_{x_k}^{x_{k+1}} \frac{dp^i}{dx}\left(x \in (x_k,\, x_{k+1})\right) \cdot \frac{dp^j}{dx}\left(x \in (x_k,\, x_{k+1})\right) dx \ , \tag{29}$$

where it is considered that

$$p^j\left(x \in (x_k,\, x_{k+1})\right)\Big|_{x_k} = 0$$

and

$$p^j(x\in(x_k,\ x_{k+1}))\Big|_{k_{k+1}} = 0 \quad .$$

Letting $\lambda = (x-x_k)/\Delta x$, the second summand in (29) is integrated to

$$\frac{1}{36\Delta x}\int_0^1 \left[(2-6\lambda+3\lambda^2)a_k^i + (1-3\lambda^2)a_{k+1}^i\right]\left[(2-6\lambda+2\lambda^2)a_k^j + (1-3\lambda^2)a_{k+1}^j\right]d\lambda$$

$$= \frac{1}{36\Delta x}\left\{\frac{4}{5}\,(a_k^i\ a_k^j + a_{k+1}^i a_{k+1}^j) + \frac{7}{10}\,(a_k^i\ a_{k+1}^j + a_{k+1}^i\ a_k^j)\right\} \quad .$$

When (29) is substituted in (28), the first summand in (29), i.e. the product of the delta notations, produces

$$F = \frac{1}{\Delta x}\sum_{i=0}^{N} T_i^I(\xi)\sum_{k=0}^{N-1}\delta(k;\ i-1,\ i)\ \delta(k;\ j-1,\ j) \ ,$$

where $j\in\{1,\ldots,\ N-1\}$. Considering that the second multiplicand delta notation in the above is nonzero only for $k = j-1$ and $= j$, we get

$$F = \frac{1}{\Delta x}\sum_{i=0}^{N} T_i^I(\xi)\left[\delta(j-1;\ i-1,\ 1) - \delta(j;\ i-1,\ i)\right] \ ,$$

which becomes

$$F = \frac{1}{\Delta x}\left[-\ T_{j-1}^I(\xi) + 2T_j^I(\xi) - T_{j+1}^I(\xi)\right\} \quad .$$

Thus the second summand in (28) is evaluated,

$$\Delta x\sum_{i=0}^{N} T_i^I(\xi)\sum_{k=0}^{N-1}\int_{x_k}^{x_{k+1}}\frac{dB^i(x)}{dx}\ \frac{dB^j(x)}{dx}\,dx$$

$$= -\ T_{j-1}^I(\xi) + 2T_j^I(\xi) - T_{j+1}^I(\xi) + \frac{1}{36}\sum_{i=0}^{N} T_i^I(\xi)\ \Lambda_I^{ij} \ , \qquad (30)$$

where

$$A_I^{ij} = \sum_{k=0}^{N-1} \left\{ \frac{4}{5} \left(a_k^i \, a_k^j + a_{k+1}^i \, a_{k+1}^j \right) + \frac{7}{10} \left(a_k^i \, a_{k+1}^j + a_{k+1}^i \, a_k^j \right) \right\} ,$$

in which

$$i \in \{0, \ldots, N\} , \quad j \in \{1, \ldots, N-1\} .$$

In (30) the terminal temperatures $T_0^I(\xi) = T_A$ and $T_N(\xi) = T_F$ must be employed when needed.

To simplify the integration in the first summand of (28), we introduce the epsilon notation, defined by

$$\begin{aligned} \varepsilon(j) &= 1 \qquad \text{when} \quad j = 0 , \\ &= 0 \qquad \text{when} \quad j \neq 0 . \end{aligned} \tag{31}$$

Then letting $\lambda = (x - x_k)/\Delta x$, the entries in $B_i(x)$ in (9) may be rewritten to

$$\left((x - x_{i-1})/\Delta x \right)_{x \in (x_{i-1}, \, x_i)} = \lambda \cdot \varepsilon(k-i+1) ,$$

$$\left((x_{i+1} - x)/\Delta x \right)_{x \in (x_i, \, x_{i+1})} = (1-\lambda) \cdot \varepsilon(k-i) ,$$

$$p^i(x \in [x_h, \, x_{h+1}]) = - \frac{i}{6} \varepsilon(k-h) \left\{ \lambda(2 - 3\lambda + \lambda^2) a_k^i + (\lambda - \lambda^3) a_{k+1}^i \right\} ,$$

where

$$h \in \{0, \ldots, N-1\} .$$

Carrying out the integration, we find

$$\frac{1}{\Delta x} \sum_{k=0}^{N-1} \int_{x_k}^{x_{k+1}} B^i(x) \, B^j(x) \, dx , \qquad i \in \{0, \ldots, N\}, \; j \in \{1, \ldots, N-1\} ,$$

$$= \sum_{k=0}^{N-1} \left\{ \frac{1}{3} \left[\varepsilon(k-i)\varepsilon(k-j) + \varepsilon(k-i+1)\varepsilon(k-j+1) \right] + \frac{1}{6} \left[\varepsilon(k-i)\varepsilon(k-j+1) + \right. \right.$$

$$\left. \left. + \varepsilon(k-i+1)\varepsilon(k-j) \right] - \frac{1}{360} \left[(7a_{i-1}^j + 16a_i^j + 7a_{i+1}^j) + (7a_{j-1}^i + 16a_j^i + 7a_{j+1}^i) \right] + \right.$$

$$+ \frac{1}{36.21} \sum_{k=0}^{N-1} \left[\frac{8}{5} \left(a_k^i a_k^j + a_{k+1}^i a_{k+1}^j\right) + \frac{31}{20} \left(a_k^i a_{k+1}^j + a_{k+1}^i a_k^j\right)\right]\}$$

$$= \Omega_I^{j,i} \quad . \tag{32}$$

Evaluated at $i = j-1, j, j+1$, and $i \notin \{j-1, j, j+1\}$, $\Omega_I^{j,i}$ yields

$$\Omega_I^{j,i=j-1} =$$

$$= \frac{1}{6} - \frac{1}{360} \left[\left(7a_{j-2}^j + 16a_{j-1}^j + 7a_j^j\right) + \left(7a_{j-1}^{j-1} + 16a_j^{j-1} + 7a_{j+1}^{j-1}\right)\right] +$$

$$+ \frac{1}{36.21} \sum_{k=0}^{N-1} \left[\frac{8}{5} \left(a_k^{j-1} a_k^j + a_{k+1}^{j-1} a_{k+i}^j\right) + \frac{31}{20} \left(a_k^{j-1} a_{k+1}^j + a_{k+1}^{j-1} a_k^j\right)\right] , \tag{33.1}$$

$$\Omega_I^{j,i=j} = \frac{2}{3} - \frac{1}{180} \left(7a_{j-1}^j + 16a_j^j + 7a_{j+1}^j\right) +$$

$$+ \frac{1}{36.21} \sum_{k=0}^{N-1} \left[\frac{8}{5} \left((a_k^j)^2 + (a_{k+1}^j)^2\right) + \frac{31}{10} a_k^j a_{k+1}^j\right] , \tag{33.2}$$

$$\Omega_I^{j,i=j+1} =$$

$$= \frac{1}{6} - \frac{1}{360} \left[\left(7a_j^j + 16a_{j+1}^j + 7a_{j+2}^j\right) + \left(7a_{j-1}^{j+1} + 16a_j^{j+1} + 7a_{j+1}^{j+1}\right)\right] +$$

$$+ \frac{1}{36.21} \sum_{k=0}^{N-1} \left[\frac{8}{5} \left(a_k^{j+1} a_k^j + a_{k+1}^{j+1} a_{k+i}^j\right) + \frac{31}{20} \left(a_k^{j+1} a_{k+1}^j + a_{k+1}^{j+1} a_k^j\right)\right] , \tag{33.3}$$

$$\Omega_I^{j,i \notin \{j-1, j, j+1\}} =$$

$$= - \frac{1}{360} \left[\left(7a_{i-1}^j + 16a_i^j + 7a_{i+1}^j\right) + \left(7a_{j-1}^i + 16a_j^i + 7a_{j+1}^i\right)\right] +$$

$$+ \frac{1}{36.21} \sum_{k=0}^{N-1} \left[\frac{8}{5} \left(a_k^i a_k^j + a_{k+1}^i a_{k+i}^j\right) + \frac{31}{20} \left(a_k^i a_{k+1}^j + a_{k+1}^i a_k^j\right)\right] . \tag{33.4}$$

Thus (18.1) becomes

$$\Delta x \frac{d\xi}{dt} \sum_{i=0}^{N} \frac{dT_i^I}{d\xi} \Omega_I^{j,i} + \frac{1}{\Delta x} \{- T_{j-1}^I(\xi) + 2T_j^I(\xi) - T_{j+1}^I(\xi)\} +$$

$$+ \frac{1}{36\Delta x} \sum_{i=0}^{N} T_i^I(\xi) \Lambda_I^{ij} = 0 \quad ,$$

where

$$\Delta x = \xi/N \ .$$

Taking the difference and mean at $\xi + \frac{1}{2} \Delta\xi$, and using the temperature notations defined in Sections IV and V, i.e., letting

$$\frac{dT_i^I}{d\xi} = \frac{y_i - T_i}{\Delta\xi} \quad ,$$

and changing $T_i(\xi)$ to

$$T_i(\xi + \tfrac{1}{2} \Delta\xi) = \tfrac{1}{2} (y_i + T_i) \ ,$$

(18.1) becomes finally

$$\left(\frac{\xi}{N}\right)^2 \frac{\Delta\xi}{\Delta t} \sum_{h=1}^{n} \frac{y_h - T_h}{\Delta\xi} \Omega_I^{j,h} -$$

$$- \frac{y_{j-1} + T_{j-1}}{2} + (y_j + T_j) - \frac{y_{j+1} + T_{j+1}}{2} +$$

$$+ \frac{1}{36} \sum_{h=1}^{n} \frac{y_h + T_h}{2} \Lambda_I^{hj} + \frac{1}{36} T_A \Lambda_I^{Aj} + \frac{1}{36} \Lambda_I^{Fj} = 0 \quad , \qquad (34)$$

where $j \in \{1, \ldots, n\}$. Λ_I^{Aj} and Λ_I^{Fj} are used instead of Λ_I^{0j} and Λ_I^{nj}, respectively, to avoid the possible confusion in the ice domain.

Following the similar procedure, (18.2) yields

$$\frac{1}{\beta} \left(\frac{1-\xi}{M}\right)^2 \frac{\Delta\xi}{\Delta t} \sum_{h=n+1}^{n+m} \frac{y_h - T_h}{\Delta\xi} \Omega_w^{j,h} -$$

$$- \frac{y_{j-1} + T_{j-1}}{2} + (y_j + T_j) - \frac{y_{j+1} + T_{j+1}}{2} +$$

$$+ \frac{1}{36} \sum_{h=n+1}^{n+m} \frac{y_h + T_h}{2} A_w^{hj} + \frac{1}{36} T_F A_w^{Fj} + \frac{1}{36} T_B A_w^{Bj} = 0 \quad , \tag{35}$$

where $j \in \{n+1, \ldots, n+m\}$. A_w^{Fj} and A_w^{Bj} are used to avoid possible confusion in the water domain.

Changing $d\xi/d\tau$ to $\Delta\xi/\Delta\tau$, and taking the difference and mean at $\xi + \frac{1}{2}\Delta\xi$, we rewrite (17) to

$$\frac{\Delta\xi}{\Delta\tau} = \sum_{h=1}^{m+n} q^k y_k + r \quad , \tag{36}$$

where

$$q^k = (C_I/L) \frac{N}{\xi} \frac{1}{12} (2a_N^k + a_n^k) \quad \text{for} \quad k \in \{1, \ldots, n-1\} \quad ,$$

$$q^n = (C_I/L) \frac{N}{\xi} \left[-\frac{1}{2} + \frac{1}{12} (2a_N^n + a_{N-1}^n) \right] \quad ,$$

$$q^{n+1} = (\beta c_w/L) \frac{M}{1-\xi} \left[-\frac{1}{2} + \frac{1}{12} (2b_0^1 + b_1^1 \right] \quad ,$$

$$q^k = (\beta c_w/L) \frac{M}{1-\xi} \frac{1}{12} (2b_0^{k-n} + b_1^{k-n}) \quad \text{for} \quad k \in \{n+1, \ldots, n+m\} \quad ,$$

$$r = (c_I/L) \frac{N}{\xi} \left\{ T_F - \frac{T_n}{2} + \frac{1}{6} (2a_N^0 + a_n^0) T_A + \frac{1}{6} (2a_N^N + a_n^N) T_F + \right.$$

$$\left. + \frac{1}{12} \sum_{i=1}^{n} (2a_N^i + a_n^i) T_i \right\} + \frac{\beta c_w}{L} \frac{L}{1-\xi} \left\{ T_F - \frac{T_{n+1}}{2} + \frac{1}{6} (2b_0^0 + b_1^0) T_F + \right.$$

$$\left. + \frac{1}{6} (2b_0^M + b_1^M) T_B + \frac{1}{12} \sum_{i=1}^{m} (2b_0^i + b_1^i) T_{n+1} \right\} \quad .$$

Substituting (36) into (34) and (35), we find (19), whose coefficients are shown below. Derived from (34), the coefficients for $j \in \{1, \ldots, N\}$ are as follows,

$$\alpha_j^{kh} = \frac{1}{\Delta\xi} \left(\frac{\xi}{N}\right)^2 q^k \Omega_I^{j,h} \quad , \qquad k \in \{1, \ldots, n+m\}, \; h \in \{1, \ldots, n\} \quad .$$

Let λ_j^k for $k \in \{1, \ldots, n\}$ be defined by

$$\lambda_j^k = \frac{1}{\Delta\xi} \left(\frac{\varepsilon}{N}\right)^2 \left\{ - p^k \sum_{h=1}^{n} T_h \, \Omega_I^{j,h} + r \, \Omega_I^{j,k} \right\} + \frac{1}{72} \Omega_I^{j,k} \quad ,$$

then

$$\begin{aligned}
\beta_j^k &= \lambda_j^k && \text{for} \quad k \in \{1, \ldots, j-2\} \\
&= \lambda_j^{j-1} - \frac{1}{2} && \text{for} \quad k = j-1 \\
&= \lambda_j^j + 1 && \text{for} \quad k = j \\
&= \lambda_j^{j+1} - \frac{1}{2} && \text{for} \quad k = j+1 \\
&= \lambda_j^k && \text{for} \quad k \in \{j+2, \ldots, n\} \\
&= - \frac{1}{\Delta\xi} \left(\frac{\xi}{N}\right)^2 p^k \sum_{h=1}^{n} T_h \, \Omega_w^{j,h} && \text{for} \quad k \in \{n+1, \ldots, n+m\} \, .
\end{aligned}$$

Let μ_j be defined by

$$\mu_j = - \frac{1}{\Delta\xi} \left(\frac{\xi}{N}\right)^2 r \sum_{h=1}^{n} T_h \, \Omega_w^{j,h} - \frac{1}{2} T_{j-1} + T_j - \frac{1}{2} T_{j+1} +$$

$$+ \frac{1}{72} \sum_{h=1}^{n} T_h \, \Lambda_I^{jh} + \frac{1}{36} T_A \, \Lambda_I^{Aj} + \frac{1}{36} T_F \, \Lambda_I^{Fj} \quad ,$$

then, we find

$$\gamma_1 = \mu_1 - \frac{1}{2} T_A \quad ,$$

$$\gamma_j = \mu_j \qquad \text{for} \quad j \in \{2, \ldots, n-1\} \quad ,$$

$$\gamma_n = \mu_n - \frac{1}{2} T_F \quad .$$

Derived from (35), the coefficients for $j \in \{n+1, .., n+m\}$ are as follows:

$$\alpha_j^{kh} = \frac{1}{\beta}\frac{1}{\Delta\xi}\left(\frac{1-\xi}{M}\right)^2 q^k \Omega_w^{j,h}, \quad k \in \{1,\ldots, n+m\}, \quad h \in \{n+1,\ldots, n+m\}.$$

Let λ_j^k for $k \in \{n+1,\ldots, n+m\}$ be defined by

$$\lambda_j^k = \frac{1}{\beta}\frac{1}{\Delta\xi}\left(\frac{1-\xi}{M}\right)^2 \{-p^k \sum_{h=n+1}^{n+m} T_h \Omega_w^{j,h} + r\Omega_w^{j,k}\} + \frac{1}{72}\Omega_w^{j,k},$$

then we find

$$\begin{aligned}
\beta_j^k &= -\frac{1}{\beta}\frac{1}{\Delta\xi}\left(\frac{1-\xi}{M}\right)^2 p^k \sum_{h=n+1}^{n+m} T_h \Omega_w^{j,h} && \text{for} \quad k \in \{1,\ldots,n\}, \\
&= \lambda_j^k && \text{for} \quad k \in \{n+1,\ldots, j-2\}, \\
&= \lambda_j^{j-1} - \frac{1}{2} && \text{for} \quad k = j-1, \\
&= \lambda_j^j + 1 && \text{for} \quad k = j, \\
&= \lambda_j^k - \frac{1}{2} && \text{for} \quad k = j+1, \\
&= \lambda_j^k && \text{for} \quad k \in \{n+2,\ldots, n+m\}.
\end{aligned}$$

Let μ_j be defined by

$$\mu_j = -\frac{1}{\beta}\frac{1}{\Delta\xi}\left(\frac{1-\xi}{M}\right)^2 r \sum_{h=h+1}^{n+m} T_h \Omega_w^{j,h} - \frac{1}{2}T_{j-1} + T_j - \frac{1}{2}T_{j+1}$$

$$+ \frac{1}{72}\sum_{h=n+1}^{n+m} T_h \Lambda_w^{jh} + \frac{1}{36}T_F \Lambda_w^{Fj} + \frac{1}{36}T_B \Lambda_w^{Bj},$$

then we find

$$\beta_j^k = -\frac{1}{\beta}\frac{1}{\Delta\xi}\left(\frac{1-\xi}{M}\right)^2 p^k \sum_{h=n+1}^{n+m} T_h \Omega_w^{j,h} \qquad \text{for} \quad k \in \{1,\ldots, n\} ,$$

$$= \lambda_j^k \qquad \text{for} \quad k \in \{n+1,\ldots, j-2\} ,$$

$$= \lambda_j^{j-1} - \frac{1}{2} \qquad \text{for} \quad k = j-1 ,$$

$$= \lambda_j^{j+1} \qquad \text{for} \quad k = j ,$$

$$= \lambda_j^k - \frac{1}{2} \qquad \text{for} \quad k = j+1 ,$$

$$= \lambda_j^k \qquad \text{for} \quad k \in \{j+2,\ldots, n+m\} .$$

VIII. REFERENCES.

1. S. Takagi: Approximate analytical solutions of a Stefan's problem in a finite domain. Quarterly of Applied Mathematics (in press).

2. J. Dieudonné: Foundations of Modern Analysis, Academic Press, New York, 1969.

3. S.T. Lin and Y. Lin: Set Theory with Applications, Mariner Publishing Co., Inc., Tampa, Florida, 1981.

4. J.H. Ahlberg, E.N. Nilson and J.L. Walsh: The Theory of Splines and Their Applications, Academic Press, New York, 1967.

APPENDIX A. Matrix a_i^k in (4)

Case N = 3:

$$(\Delta x)^2 \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 2 & -5 & 4 & -1 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ -1 & 4 & -5 & 2 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Case N = 4:

$$(\Delta x)^2 \cdot 4 \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 7 & -24 & 22 & -8 & 1 \\ 4 & -8 & 4 & 0 & 0 \\ -1 & +8 & -14 & 8 & -1 \\ 0 & 0 & 4 & -8 & 4 \\ 1 & 22 & 22 & -14 & 9 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Case N = 5:

$$(\Delta x)^2 \cdot 15 \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix} = \begin{bmatrix} 34 & -92 & 88 & -37 & 8 & -1 \\ 15 & -30 & 15 & 0 & 0 & 0 \\ -4 & 32 & -58 & 37 & -8 & 1 \\ 1 & -8 & 37 & -58 & 32 & -4 \\ 0 & 0 & 0 & 15 & -30 & 15 \\ -1 & 8 & 37 & 88 & -92 & 34 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}$$

For $N \geq 6$, given $r = [N/2]$, every entry a_j^i can be found by dividing the coefficient of y_i on the right-hand side of an appropriate formula in the following by the factor in front of the summation symbol on the left-hand side, where i and j are column and row numbers of the matrix, respectively. The order below must be strictly followed in the computation. It is demonstrated below that the least number of r is 3.

Case $N = 2r$, where $r \geq 3$.

Row r:

$$\frac{1}{3}(2f_r - f_{r-1}) \sum_0^N a_r^k y_k$$

$$= -(-1)^r f_1 y_0 + (-1)^r(2f_1 + f_2)y_1 - (-1)^r(f_1 + 2f_2 + f_3)y_2 -$$

$$- 6 \sum_3^{r-1} (-1)^{r-k} f_k y_k - 2(f_{r-1} + f_r)y_r -$$

$$- 6 \sum_3^{r-1} (-1)^{r-k} f_k y_{N-k} - (-1)^r(f_1 + 2f_2 + f_3)y_{N-2} +$$

$$+ (-1)^r(2f_1 + f_2)y_{N-1} - (-1)^r f_1 y_N ,$$

where, if $r = 3$, summations $6 \sum_3^{r-1} (-1)^{r-k} f_k y_k$ and $6 \sum_3^{r-1} (-1)^{r-k} f_k y_{N-k}$ must be skipped.

Row r-1, r-2,..., 2:

Letting $i = r-1, r-2, \ldots, 2$ successively in the following, entries a_i^k in row i can be found:

$$f_{i+1} \sum_0^N a_i^k y_k = -f_i \sum_0^N a^k{}_{i+1} y_k - 6(-1)^i f_1 y_0 +$$

$$+ 6(-1)^i(2f_1 + f_2)y_1 - 6(-1)^i(f_1 + 2f_2 + f_3)y_2 -$$

$$- 36 \sum_3^{i-1} (-1)^{i-j} f_j y_j - 6(f_{i-1} + 2f_i)y_i +$$

$$+ 6f_i y_{i+1} ,$$

where $6(-1)^i(f_1 + 2f_2 + f_3)y_2$ must be skipped for $i = 2$, and $36 \sum_3^{i-1} (-1)^{i-j} f_j y_j$

for i = 2 and 3.

Row 1:

$$\sum_{1}^{N} a_1^k y_k = y_0 - 2y_1 + y_2 \quad .$$

Row 0:

$$\sum_{0}^{N} a_0^k y_k = \sum_{0}^{N} (2a_1^k - a_2^k) y_k \quad .$$

The rest of the elements can be found by using the centrosymmetric relation,

$$a_i^k = a_{N-i}^{N-k} \quad , \qquad \text{(A.1)}$$

where k, i = 0,..., N .

Case N = 2r+1, where $r \geq 3$.

Row r:

$$\frac{1}{6}\left(f_{r+1}^2 - f_r^2\right) \sum_{0}^{N} a_r^k y_k =$$

$$= (-1)^r f_{r+1} \left[- f_1 y_0 + (2f_1 + f_2) y_1 - (f_1 + 2f_2 + f_3) y_2 - \right.$$

$$\left. - 6 \sum_{3}^{r-1} (-1)^k f_k y_k \right] - (f_r^2 + 2f_r f_{r+1} + f_{r-1} f_{r+1}) y_r +$$

$$+ 6f_r^2\, y_{r+1} - (-1)^r f_r \Big[6 \sum_{r+2}^{N-3} (-1)^k f_{N-k}\, y_k -$$

$$- (f_1 + 2f_2 + f_3) y_{N-2} + (2f_1 + f_2) y_{n-1} - f_1\, y_N \Big] ,$$

where, if r = 3, summations $6 \sum_{3}^{r-1} (-1)^j f_k\, y_k$ and $6 \sum_{r+2}^{N-3} (+1)^j f_{N-k}\, y_k$ must be skipped.

Row r-1, r-2,..., 2:

Letting i = r-1, r-2,..., 2 successively in the following, elements a_i^k in row i can be found.

$$f_{i+1} \sum_{0}^{N} a_i^k\, y_k =$$

$$= - f_i \sum_{0}^{N} a_{i+1}^k\, y_k - 6(-1)^i f_1\, y_0 + 6(-1)^i (2f_1 + f_2) y_1 -$$

$$- 6(-1)^i (f_1 + 2f_2 + f_3) y_2 - 36 \sum_{j=3}^{i-1} (-1)^{i-j} f_j\, y_j -$$

$$- 6(f_{i-1} + 2f_i) y_i + 6f_i\, y_{i+1} ,$$

where $6(-1)^i(f_1 + 2f_2 + f_3)y_2$ must be skipped for i = 2, and $36 \sum_{j=3}^{i-1} (-1)^{i-j} f_j\, y_j$ for i = 2 and 3 .

Row 1:

$$\sum_{0}^{N} a_1^i\, y_i = y_0 - 2y_1 + y_2 .$$

Row 0:

$$\sum_{0}^{N} a_0^i y_i = \sum_{0}^{N} (2a_1^i - a_2^i) y_i \quad .$$

The rest of the elements can be found by use of the centrosymmetric relations (A.1). The validity of the centrosymmetric relation is obvious in cases N = 3, 4 and 5. For the cases $N \geq 6$, it can be demonstrated by actually writing the expressions, which is avoided in this presentation.

Stipulated by the equations [1], [2],..., [N], elements a_i^k satisfy the relations shown below. Substituting (1) and (4) into the equations, equating the coefficients of y_i, i = 0,..., N, and eliminating the duplicated relations by use of the centrosymmetric relation (A.1), three groups are found.

Group 1: Equations [0] and [N] yield

$$a_0^k - 2a_1^k + a_2^k = 0 \quad , \tag{A.2}$$

where $k \in \{0,\ldots, N\}$.

Group 2: For $k \in \{i-1, i, i+1\}$, equations [1],..., [N-1] yield three relations,

$$a_{i-1}^{i-1} + 4a_i^{i-1} + a_{i+1}^{i-1} = 6$$

$$a_{i-1}^{i} + 4a_i^{i} + a_{i-1}^{i} = -12 \tag{A.3}$$

and

$$a_{i-1}^{i+1} + 4a_i^{i+1} + a_{i+1}^{i+1} = 6 \quad ,$$

where

$$i \in \{1,\ldots, N-1\} \quad .$$

Group 3: For k $\{0,\ldots, i-2\}\bigcup\{i+2,\ldots, N\}$, equations [1],...,[N-1] yield

$$a_{i-1}^{k} + 4a_{i}^{k} + a_{i+1}^{k} = 0 \quad . \tag{A.4}$$

Conventional set-theoretic notations (Ref. 2,3) are employed in the above to describe the subset of concerned equation numbers.

ON SOME FINITE ELEMENT ERROR ESTIMATES FOR STRESS INTENSITY FACTORS IN MODE I LINEAR ELASTIC FRACTURE PROBLEMS

J.R. Whiteman and G. Goodsell

Institute of Computational Mathematics
Department of Mathematics and Statistics
Brunel University, Uxbridge, England.

ABSTRACT

The use of the superconvergence phenomenon for the retrieved gradients of piecewise linear approximations on triangular meshes to the solutions of problems of planar linear elasticity is discussed. In particular results for problems involving singularities are presented, with particular reference to the apparent shortcomings for the case of linear elastic fracture.

I. INTRODUCTION

Finite element methods are now used routinely for problems of linear elastic fracture, see e.g. Owen and Fawkes [5], and theoretical error estimates for finite element approximations to stress intensity factors have been derived, see e.g. Destuynder, Djaoua and Lescure [1]. However, there often remains the hope with finite element methods that further research will produce better error estimates and improved rates of convergence. To this end we consider here the phenomenon of superconvergence in finite element methods and its possible use in the treatment of linear elastic fracture.

The phenomenon of superconvergence in finite element methods, whereby the rate of convergence with decreasing mesh size of the finite element approximation to the true solution of the problem is at certain points of the problem domain superior to that found globally, is now well known and has been extensively researched; see e.g. the review paper of Krizek and Neittaanmäki [4] which contains two hundred references. However, although the superconvergence effect has been extensively exploited by engineers in stress analysis, and even in linear elastic fracture by the use of contour integrals using calculated stress values at Gauss points with quadrilateral elements, it is true to say that the mathematical analysis of superconvergence has lagged behind the practice; largely on account of the high levels of regularity of the problem solutions required to produce meaningful superconvergent error estimates. These have effectively precluded the analysis of methods for realistic problems. This situation motivated the work of Wheeler and Whiteman [8], who derived superconvergent estimates for recovered

gradients on subdomains from piecewise linear finite element approximations to the solutions of two-dimensional Poisson problems. This work was extended to problems of planar linear elasticity by Whiteman and Goodsell [9]. The result is that, for problems of the above types involving boundary singularities, superconvergence estimates are available for the approximations to the gradients of primary variables on subdomains.

However, for linear elastic fracture the main quantity of interest is the stress intensity factor and its approximation. In this paper we consider the finite element approximation of the stress intensity factor for a simple Mode I problem, through the use of the J-integral of Eshelby [2] and Rice [6]. Theoretical error estimates for methods involving recovered gradients are presented, which have lower rates of convergence than those of [1], although the current approximations appear numerically to have the same rate of convergence. However, we feel it worth-while to demonstrate these present limitations, particularly as, for methods involving recovered gradients, one might expect that both the theoretical and numerical rates of convergence would be better than those obtained in the standard manner.

II. LINEAR ELASTICITY AND LINEAR ELASTIC FRACTURE

II.1 Linear Elastic Problem and Weak Formulation

The linear elastic problem is defined in the region $\Omega \subset \mathbb{R}^2$ with polygonal boundary $\partial\Omega \equiv \partial\Omega_c \cup \partial\Omega_T$. The displacement $\underline{u}(\underline{x}) \equiv (u_1,u_2)^t$ at any point $\underline{x} \equiv (x_1,x_2)^t \in \Omega$ satisfies the Lamé equation

$$-\mu\Delta\underline{u} - (\lambda+\mu)\,\mathrm{grad}\,\mathrm{div}\,\underline{u} = \underline{f} \text{ in } \Omega\ , \tag{2.1}$$

and on $\partial\Omega$ the boundary conditions

$$\underline{u} = \underline{0} \quad \text{on} \quad \partial\Omega_c\ , \tag{2.2}$$

$$\sum_{j=1}^{2} \sigma_{ij}(\underline{u})n_j = g_i \quad \text{on} \quad \partial\Omega_T\ , \quad 1 \le i \le 2\ , \tag{2.3}$$

where $\underline{f}$ are given body forces, $\underline{g}$ are boundary tractions and λ and μ, $\lambda,\mu > 0$, are the Lamé constants of the material.

The admissible displacement vectors are $\underline{v} \equiv (v_1,v_2)^t \in (H^1(\Omega))^2$ and we define

$$V \equiv \left\{ \underline{v} : \underline{v} \in (H^1(\Omega))^2 \ , \ v_i\Big|_{\partial\Omega_c} = 0 \ , \ i = 1,2 \right\} . \tag{2.4}$$

The weak form of problem (2.1) - (2.3) is

$$\text{find } \underline{u} \in V \ni a(\underline{u},\underline{v}) = F(\underline{v}) \quad \forall \ \underline{v} \in V \ , \tag{2.5}$$

in which

$$a(\underline{u},\underline{v}) \equiv \int_\Omega \left\{ \lambda \, \mathrm{div}\, \underline{u} \, \mathrm{div}\, \underline{v} + 2\mu \sum_{i,j=1}^{2} \varepsilon_{ij}(\underline{u}) \varepsilon_{ij}(\underline{v}) \right\} d\underline{x} \ , \tag{2.6}$$

$$F(\underline{v}) \equiv \int_\Omega \underline{f}^t \cdot \underline{v} d\underline{x} + \int_{\partial\Omega_T} \sum_{i=1}^{2} g_i v_i \, ds \ . \tag{2.7}$$

For linear elastic fracture we limit the consideration here to a Mode I plane stress problem with symmetric loading as in Fig. 1. This problem is of type (2.1) - (2.3) and we note that the faces of the crack are stress free.

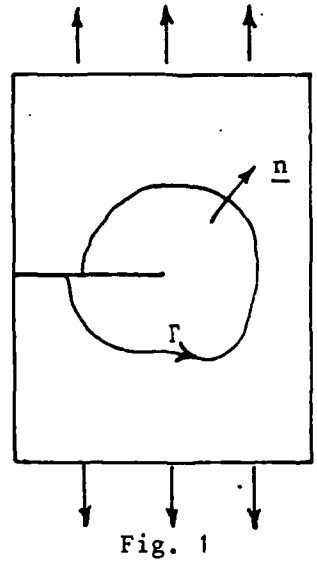


Fig. 1

The near-tip crack displacement field has the form, see [6],

$$\underline{u}(\underline{x}) = \frac{K_I r^{\frac{1}{2}}}{2\mu\sqrt{2\pi}} \begin{Bmatrix} \cos\theta/2\,(\kappa - 1 + 2\sin^2\theta/2) \ , \\ \\ \sin\theta/2\,(\kappa + 1 - 2\cos^2\theta/2) \ , \end{Bmatrix} \tag{2.8}$$

where the constant K_I is the (Mode I) stress intensity factor which has to be determined, and for plane stress $\kappa = (3-\nu)/(1+\nu)$. The stress intensity factor is important as it is a fracture criterion.

One way of calculating K_I for this problem is to use the J-integral, see Eshelby [2] and Rice [6], defined, see Fig. 1, as

$$J = \int_\Gamma \left\{ W dx_2 - \sum_i T_i \frac{\partial u_i}{\partial x_1} ds \right\} , \tag{2.9}$$

where $\Gamma \subset \Omega$ is any closed curve joining the lower and upper faces of the crack, $W \equiv \frac{1}{2}\sigma_{ij}\varepsilon_{ij}$ is the strain energy density and $T_i = \sigma_{ij}n_j$, $\underline{n}$ being the outward normal unit vector to Γ. For the plane stress problem of Fig. 1,

$$J = K_I^2/E , \tag{2.10}$$

where E is the Young's modulus of the linear elastic material of the problem (2.1) - (2.3).

II.2 Finite Element Method and Recovered Gradients

We adopt the same notation and assumptions as were used by Whiteman and Goodsell in [9]. The region Ω is assumed to have subdomains $\Omega_0, \Omega_1, \Omega_2$ such that $\Omega_0 \subset\subset \Omega_1 \subset\subset \Omega_2 \subseteq \Omega$ and is partitioned into triangular elements. The regions Ω_0 and Ω_2 are assumed to be rectangular and such that each is the union of a finite number of squares of side h. Each square is subdivided into two triangles using the diagonal of positive slope, so that Ω_0 and Ω_2 are each meshed completely with uniform isosceles right angled triangles. The mesh in the remainder of $\overline{\Omega} - \overline{\Omega}_2$ is compatible with that of $\overline{\Omega}_2$ and consists of triangles of general shape.

A finite dimensional subspace $S^h \subset V$ consisting of continuous piecewise linear functions is defined over the partition of Ω and the finite element method is applied to (2.5) with trial function $\underline{u}_h$ and test functions $\underline{v}_h$ from S^h. For $\underline{v}_h \in S^h$ we define the recovered mid-point gradient

$$\underline{D}_k(\underline{v}_h) \equiv \frac{1}{2}\left\{ [\underline{\nabla} \underline{v}_h]_{T_k} + [\underline{\nabla} \underline{v}_h]_{T_k'} \right\} , \tag{2.11}$$

for k the mid-points of element sides in Ω_0 and T_k and T_k' any pair of adjacent elements in Ω_0. For element side mid-points M on $\partial\Omega_0$ the recovered gradient is defined as

$$\underline{D}_k(\underline{v}_h(M)) = \sum_i q_i [\underline{\nabla} \underline{v}_h]_{T_k^i} , \tag{2.12}$$

where the q_i are simple numerical coefficients and the summation is over

a small number of triangles involving and near to the point M, see [9].

For any element of Ω_0 we take the linear interpolant to the recovered gradients of $\underline{v}_h$ at the three side midpoints. Over the whole of Ω_0 these linear interpolants form the discontinuous piecewise linear interpolant $\underline{\nabla}^*\underline{v}_h$ to the recovered gradients of $\underline{v}_h$. We define the seminorm

$$\left|\underline{u} - \underline{v}_h\right|^*_{1,\Omega_0} \equiv \left|\underline{\nabla u} - \underline{\nabla}^*\underline{v}_h\right|_{0,\Omega_0} . \tag{2.13}$$

It has been shown by Whiteman and Goodsell [9] that, if $\underline{u} \in V \cap (H^3(\Omega_2))^2$ is the solution of problem (2.5) and $\underline{v}_h \in S^h$, then

$$\left|\underline{u} - \underline{v}_h\right|^*_{1,\Omega_0} \leq C\left\{\left|\underline{u}_I - \underline{v}_h\right|_{1,\Omega_0} + h^2\left|\underline{u}\right|_{3,\Omega_0}\right\} \tag{2.14}$$

where $\underline{u}_I \in S^h$ is the piecewise linear interpolant to $\underline{u}$ at the element vertices.

In order to obtain an estimate for $\left|\underline{u} - \underline{u}_h\right|^*_{1,\Omega_0}$, it is thus necessary to bound $\left|\underline{u}_I - \underline{u}_h\right|_{1,\Omega_0}$ in (2.14). For the problem (2.5) with the configuration as in Fig. 1 it has been shown in [9] that

$$\begin{aligned}\left|\underline{u}_I - \underline{u}_h\right|_{1,\Omega_0} &\leq C\left\{h\left|\underline{u}_I - \underline{u}_h\right|_{1,\Omega_2} + \left|\underline{u}_I - \underline{u}_h\right|_{0,\Omega_2} + h^2\left|\underline{u}\right|_{3,\Omega_2}\right\} \\ &\leq C\left\{h^{1-2\varepsilon}\left|\underline{u}\right|_{2,4/3-\varepsilon,\Omega} + h^2\left|\underline{u}\right|_{3,\Omega_2}\right\} ,\end{aligned} \tag{2.15}$$

where $\varepsilon > 0$ is an arbitrary constant and $\underline{u} \in (W^2_{4/3-\varepsilon}(\Omega))^2$. Combining inequalities (2.14) and (2.15) we have that

$$\left|\underline{u} - \underline{u}_h\right|^*_{1,\Omega_0} \leq C\left\{h^{1-2\varepsilon}\left|\underline{u}\right|_{2,4/3-\varepsilon,\Omega} + h^2\left|\underline{u}\right|_{3,\Omega_2}\right\} , \tag{2.16}$$

showing the $O(h^{1-2\varepsilon})$ convergence on Ω_0 of the recovered gradient function $\underline{\nabla}^*\underline{u}_h$ to $\underline{\nabla u}$.

III. PATH INTEGRAL ERROR ESTIMATES

For the case just considered where problem (2.5) is defined as in Fig. 1 and thus contains a boundary singularity due to the crack and the boundary conditions, the error estimate (2.16) holds only in an interior subdomain Ω_0, because it demands that the solution $\underline{u} \in (H^3(\Omega_2))^2$ where $\Omega_2 \subseteq \Omega$. The J-integral (2.9) is defined over the path Γ, which joins the

lower to the upper face of the crack as in Fig. 1 and thus contains points of $\partial\Omega$. This effectively precludes the use of estimates of the type (2.16) for the estimation of errors in calculated approximations J^* to J. However, we believe that it is of interest to estimate errors in path integrals of this type over interior contours, and this we shall now do for a representative case.

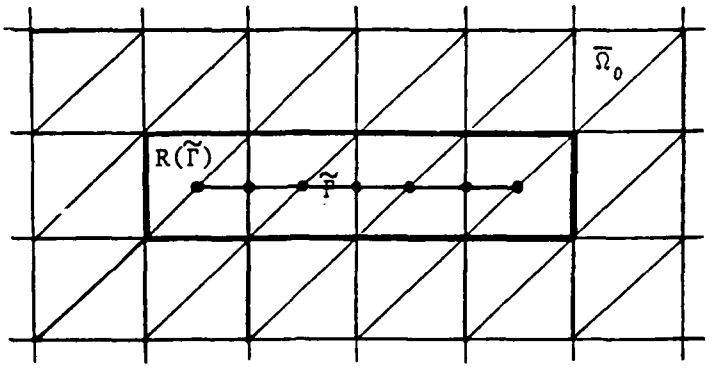


Fig. 2

We consider the fixed contour $\tilde{\Gamma} \subset \bar{\Omega}_0$, which for simplicity we take as a straight line parallel to the x_1-axis through certain element side midpoints of $\bar{\Omega}_0$ as in Fig. 2. Along the contour $\tilde{\Gamma}$ the recovered gradient function $\underline{\nabla}^* \underline{v}_h$, $\underline{v}_h \subset S_h$ is continuous and piecewise linear.

It has been shown by Goodsell and Whiteman [3], that

$$\left|\underline{u} - \underline{u}_h\right|^*_{1,\tilde{\Gamma}} \equiv \left\{ \int_{\tilde{\Gamma}} \left| \underline{\nabla u} - \underline{\nabla}^* \underline{u}_h \right|^2 dx_1 \right\}^{\frac{1}{2}}$$

$$\leq C\left\{ h^{\frac{1}{2}-2\varepsilon} \left|\underline{u}\right|_{2,4/3-\varepsilon,\Omega} + h^{3/2} \left|\underline{u}\right|_{3,\Omega_2} \right\} , \tag{3.1}$$

this being an estimate for the H^{1*} error seminorm on the contour $\tilde{\Gamma}$.

The J-integral (2.9) contains terms of the type

$$\tilde{J}_{ij} \equiv \int_{\tilde{\Gamma}} (\underline{\nabla u})_i (\underline{\nabla u})_j dx_1 , \tag{3.2}$$

which we approximate with terms of the type

$$\tilde{J}^*_{ij} \equiv \int_{\tilde{\Gamma}} (\underline{\nabla}^* \underline{u}_h)_i (\underline{\nabla}^* \underline{u}_h)_j dx_1 . \tag{3.3}$$

We are now able to bound $\left|\tilde{J}_{ij} - \tilde{J}^*_{ij}\right|$, and the estimate and the proof are given in the following theorem.

Theorem If $\underline{u} \in V \cap (H^3(\Omega_2))^2$ is the solution of problem (2.5) defined in the region of Fig. 1 and $\tilde{J}_{ij}$ and $\tilde{J}^*_{ij}$ are defined respectively as in (3.2) and (3.3), then

$$\left|\tilde{J}_{ij} - \tilde{J}^*_{ij}\right| \leq C\left\{\left|\underline{\nabla u}\right|_{0,\tilde{\Gamma}}\left[h^{\frac{1}{2}-2\varepsilon}\left|\underline{u}\right|_{2,4/3-\varepsilon,\Omega} + h^{3/2}\left|\underline{u}\right|_{3,\Omega_2}\right]\right.$$

$$\left. + h^{1-4\varepsilon}\left|\underline{u}\right|^2_{2,4/3-\varepsilon,\Omega} + h^3\left|\underline{u}\right|^2_{3,\Omega_2}\right\} . \qquad (3.4)$$

Proof From (3.2) and (3.3) we have that

$$\left|\tilde{J}_{ij} - \tilde{J}^*_{ij}\right| = \left|\int_{\tilde{\Gamma}}\left\{(\underline{\nabla^* u_h})_i(\underline{\nabla^* u_h})_j - (\underline{\nabla u})_i(\underline{\nabla u})_j\right\}dx_1\right|$$

$$\leq \left|\int_{\tilde{\Gamma}}(\underline{\nabla^* u_h})_i\left\{(\underline{\nabla^* u_h})_j - (\underline{\nabla u})_j\right\}dx_1\right| + \left|\int_{\tilde{\Gamma}}(\underline{\nabla u})_j\left\{(\underline{\nabla^* u_h})_i - (\underline{\nabla u})_i\right\}dx_1\right|$$

$$\leq \int_{\tilde{\Gamma}}\left|\underline{\nabla^* u_h}\right|\left|(\underline{\nabla^* u_h} - \underline{\nabla u})\right|dx_1 + \int_{\tilde{\Gamma}}\left|\underline{\nabla u}\right|\left|(\underline{\nabla^* u_h} - \underline{\nabla u})\right|dx_1$$

$$= \int_{\tilde{\Gamma}}\left(\left|\underline{\nabla^* u_h}\right| + \left|\underline{\nabla u}\right|\right)\left(\left|\underline{\nabla^* u_h} - \underline{\nabla u}\right|\right)dx_1$$

$$\leq \int_{\tilde{\Gamma}}\left\{2\left|\underline{\nabla u}\right| + \left|\underline{\nabla^* u_h} - \underline{\nabla u}\right|\right\}\left\{\left|\underline{\nabla^* u_h} - \underline{\nabla u}\right|\right\}dx_1$$

$$\leq 2\left\{\int_{\tilde{\Gamma}}\left|\underline{\nabla u}\right|^2 dx_1\right\}^{\frac{1}{2}}\left\{\int_{\tilde{\Gamma}}\left|\underline{\nabla^* u_h} - \underline{\nabla u}\right|^2 dx_1\right\}^{\frac{1}{2}} + \int_{\tilde{\Gamma}}\left|\underline{\nabla^* u_h} - \underline{\nabla u}\right|^2 dx_1$$

$$= 2\left|\underline{\nabla u}\right|_{0,\tilde{\Gamma}}\left|\underline{u} - \underline{u_h}\right|^*_{1,\tilde{\Gamma}} + \left|\underline{u} - \underline{u_h}\right|^{*2}_{1,\tilde{\Gamma}} ,$$

so that result (3.4) follows immediately using (3.1). □

If we now define

$$\tilde{K}_{ij} \equiv \left|\tilde{J}_{ij}\right|^{\frac{1}{2}} , \quad \tilde{K}^*_{ij} \equiv \left|\tilde{J}^*_{ij}\right|^{\frac{1}{2}} ,$$

then, using (3.4), it follows that

$$\left|\tilde{K}_{ij} - \tilde{K}^*_{ij}\right| = \left|\left|\tilde{J}_{ij}\right|^{\frac{1}{2}} - \left|\tilde{J}^*_{ij}\right|^{\frac{1}{2}}\right| = \frac{\left|\left|\tilde{J}_{ij}\right| - \left|\tilde{J}^*_{ij}\right|\right|}{\left|\tilde{J}_{ij}\right|^{\frac{1}{2}} + \left|\tilde{J}^*_{ij}\right|^{\frac{1}{2}}}$$

$$\leq \left|\tilde{J}_{ij} - \tilde{J}^*_{ij}\right| \Big/ \tilde{K}_{ij} = O(h^{\frac{1}{2}-\varepsilon}) . \qquad (3.5)$$

The result (3.5) is of course not an estimate for the error in the approximation K_I^* to K_I, which would be derived first by using gradients recovered from the finite element approximation $\underline{u}_h$ at element side midpoints in the computation of J^* approximating the J of (2.9), and then applying (2.10).

Further, even if a result of the type (3.5) were true for K_I^*, it does not have as good a rate of convergence as that derived in [1], where no recovery of gradients is employed. However, numerical evidence, see [3], indicates that in fact the convergence of $\widetilde{K}_{ij}^*$ to $\widetilde{K}_{ij}$ is O(h), thus suggesting that if the method were applied to obtain K_I^* the convergence would be O(h). The disappointing fact that the rate of convergence would not even then be better than O(h) is due to the lack of smoothness of the solution of the fracture problem, see [9].

REFERENCES

1. Destuynder, Ph., Djaoua, M. and Lescure, S., On a numerical method for fracture mechanics. pp. 69-84 of P. Grisvard, W. Wendland and J.R. Whiteman (eds.) Singularities and Constructive Methods for Their Treatment. Lecture Notes in Mathematics 1121. Springer Verlag, Berlin, 1984.
2. Eshelby, J.D., The continuum theory of lattice defects. pp. 79-144 of F. Seitz and D. Turnbull (eds.) Solid State Physics, Vol.3. Academic Press, New York, 1956.
3. Goodsell, G. and Whiteman, J.R., Superconvergence of recovered gradients of finite element approximations for elliptic boundary value problems. (to appear)
4. Krizek, M. and Neittaanmäki, P., On superconvergence techniques. Acta Applicandae Mathematicae 8, 1987.
5. Owen, D.R.J. and Fawkes, A.J., Engineering Fracture Mechanics. Pineridge Press, Swansea, 1983.
6. Rice, J.R., A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. 34, 379-386, 1968.
7. Rice, J.R., Mathematical analysis in the mechanics of fracture. pp.191-311 of H. Liebowitz (ed.) Fracture, Vol.II. Academic Press, New York, 1968.
8. Wheeler, M.F. and Whiteman, J.R., Superconvergent recovery of gradients on subdomains for piecewise linear finite element approximations. Numer. Meth. Partial Differential Equations 3, 65-87, 1987.
9. Whiteman, J.R. and Goodsell,G., Superconvergent recovery for stresses from finite element approximations on subdomains for planar problems of linear elasticity. pp. 29-53 of J.R. Whiteman (ed.) The Mathematics of Finite Elements and Applications VI, MAFELAP 1987. Academic Press, London, 1988.

SOME ISSUES OF NUMERICAL INTEGRATION AND PENALTY RELAXATION IN ANISOPARAMETRIC SHELL ELEMENT FORMULATION

Alexander TESSLER and Luciano SPIRIDIGLIOZZI

Mechanics and Structures Division
U.S. Army Materials Technology Laboratory
Watertown, MA 02172-0001, U.S.A.

ABSTRACT. A simple two-node axisymmetric shell element with the shallowly curved meridian, shear deformation, and rotary inertia is developed. The major aspects include: (a) anisoparametric interpolations of the displacement variables to design out excessive stiffening due to membrane and shear 'locking'; (b) consistent shear relaxation to further upgrade the element strain energy; (c) low-order quadrature evaluation. The resulting element possesses an improved condition of the stiffness matrix, increased efficiency in explicit time integration, and enhanced accuracy in coarse discretizations. Comprehensive vibration examples are carried out to assess the element performance. The numerical results demonstrate a wide applicability range with respect to element slenderness and curvature properties.

I. INTRODUCTION. Shear-deformable curved beam and shell finite elements formulated by the displacement approach present a number of conceptual difficulties [1,2]. The major issue is that of properly approximating so-called penalty strains. These are membrane strains that, due to initial geometry curvatures, couple membrane and bending deformations; and transverse shear strains which couple the transverse displacement and normal rotation kinematic variables. The computational difficulties arise when the element geometry is very thin, in which case the states of inextensional and shearless deformations are enforced by the presence of large multipliers (or penalty parameters) of the penalty strain energies. The enforcement of these deformation states at the element level implies that each polynomial coefficient of the penalty strain vanishes in the limit as the element becomes infinitely thin. The resulting constraint equations (known as penalty modes) are either properly coupled (involving contributions from all kinematic variables of the penalty strain) or spuriously uncoupled (having degrees of freedom (d.o.f.) from a single kinematic variable). It is the latter type of penalty modes that produces either nearly vanishing kinematic response (the phenomenon known as 'locking') or yields excessively stiff solutions. Thus, having properly coupled penalty strains (in all modes) is paramount in achieving practical convergence in the thin regime.

Although requiring properly coupled penalty strains is necessary, it is often not sufficient to ensure adequate thin-regime behavior. For instance, reduced integration and an analogous (and often equivalent) "field-consistent" approach [21,22], which produce properly coupled penalty strains, have been shown [11] to yield inconsistent force vectors (due to distributed loads) and mass matrices; thereby, producing dramatically inferior results in problems involving higher vibrational

modes or distributed loading. Moreover, simple (low-order) elements often experience undesirable overconstraining at coarse discretization levels, and lock severely in cases of overly restrained boundaries [3,4].

To avoid locking and/or excessive constraining entirely, 'relaxation' of element penalty constraints proved effective. The concept of relaxing shear constraints, advocated by Fried [5] and MacNeal [6] (though interpreted somewhat differently), introduces an element relaxation parameter (correction factor [7]) in the shear stress-resultant(s), hence appearing as a multiplier to the penalty parameter. As the element approaches its thin limit, the relaxation parameter diminishes, reducing the penalty value. The remarkable aspect of this approach is that at the global (whole discretization) level the penalty constraints are enforced in a superior fashion [2]. Furthermore, the penalty relaxation provides practical benefits such as enhanced accuracy in coarsely discretized models, a well-conditioned stiffness matrix over the whole range of the element slenderness, and a reduced value of the highest element frequency. The latter aspect allows larger time steps in the explicit time integration procedures.

In this paper we will develop a simple yet extremely effective shallowly-curved, axisymmetric, displacement-type shell element in which the effects of shear deformation and, in dynamics, rotary inertia are included. The element is an extension of the conical shell proposed in [8]. This effort will lay the groundwork for a three-dimensional shallow shell model.

The axisymmetric shell is an analog of a curved beam element discussed in [2]. From the interpolation standpoint the two elements are identical. They employ so-called anisoparametric (i.e., distinct degree) kinematic polynomials, which yield proper polynomial representations for the membrane and shear penalty strains. By enforcing the higher-degree membrane and shear penalty modes explicitly (i.e., by insisting upon constant variation of these strains along the element meridian), a two-node configuration having three d.o.f at each node is obtained. This, of course, implies that the lowest possible integration order (single-point Guass quadrature) exactly evaluates the respective strain energy contributions. However, normal Gauss quadrature rules are used throughout so that the kinematic reliability of the element is ensured (i.e., the only zero energy modes are those due to the rigid-body motion) along with the variational consistency of the load vector and the mass matrix.

Several penalty relaxation ideas are discussed. It is concluded that a single penalty relaxation parameter on the shear stress resultant (and, hence, the shear strain energy) can effectively be employed to enhance the strain energy approximation. The parameter is found analytically by a strain energy matching procedure.

The numerical experiments focus on the dynamic behavior of the element; specifically, on its performance in free vibration problems. Results are presented for a wide range of shell geometries including very thin, moderately thick, shallow, and deep axisymmetric shells.

II. SHALLOW SHELL EQUATIONS. To present the finite element approach in a clear fashion, we shall focus exclusively upon the axisymmetric linearly elastic shell equations in which the effects of shear deformation and rotary inertia are included in the manner of Naghdi [9] and, furthermore, the meridian curvature effect is accounted for using the shallowness approximations of Marguerre [10]. The methodology, however, is general enough to be applicable to an asymmetric shell-of-revolution response once the appropriate shell equations are invoked.

Consider the shallow axisymmetric shell element depicted in Figure 1. The kinematic variables describing the axisymmetric response are the middle-surface membrane displacement, $u(s,t)$ (henceforth, t denotes time), transverse displacement, $w(s,t)$, and meridian cross-sectional rotation, $\theta(s,t)$. Note that due to the shallowness assumption [10], which in effect is a perturbation from a conical (straight meridian) shell, these variables are attributed to the conical surface rather than the actual curved one. As a consequence of this simplifying assumption, all energy integrals are carried out over the conical surface.

The strain and curvature components may be written as:

Membrane strains

$$\varepsilon_s = u_{,s} + w_{I,s}\, w_{,s}, \quad \varepsilon_\phi = (u \sin\phi + w \cos\phi)/r \tag{2.1}$$

Bending curvatures

$$\kappa_s = -\theta_{,s}, \quad \kappa_\phi = -\theta \sin\phi/r \tag{2.2}$$

Shear strain

$$\gamma = w_{,s} - \theta \tag{2.3}$$

where s, ϕ, and r denote the shell coordinates, and w_I describes a shallow meridian shape of the shell ($w_{I,s}{}^2 \ll 1$). Note that when the meridian is straight (i.e., $w_I=0$, a conical shell), all strains (2.1)-(2.3) are those according to Naghdi theory [9].

The corresponding stress resultants are related to the strains through the constitutive relations:

$$\underline{N} = \underline{D}_m\, \underline{\varepsilon}, \quad \underline{M} = \underline{D}_b\, \underline{\kappa}, \quad Q = D_s\, \gamma \tag{2.4}$$

where

$$\underline{N}^T = \{N_s, N_\phi\}, \quad \underline{M}^T = \{M_s, M_\phi\},$$

$$\underline{\varepsilon}^T = \{\varepsilon_s, \varepsilon_\phi\}, \quad \underline{\kappa}^T = \{\kappa_s, \kappa_\phi\}.$$

For a homogeneous isotropic shell of constant thickness h the constitutive matrices are:

$$\underline{D}_m = \frac{12D}{h^2}\, \underline{I}_\nu, \quad \underline{D}_b = D\, \underline{I}_\nu, \quad D_s = k^2 G h \quad (k^2 = \pi^2/12),$$

$$\underline{I}_\nu = \begin{bmatrix} 1 & \nu \\ \nu & 1 \end{bmatrix}, \qquad D = \frac{Eh^3}{12(1-\nu^2)}, \tag{2.5}$$

where E, G, and ν denote Young's modulus, shear modulus, and Poisson's ratio, respectively; k^2 is the shear correction factor; and h is the shell thickness.

The equations of motion are readily derived from Hamilton's variational principle:

$$\delta\int_{t_0}^{t_1} L\,dt = \delta \int_{t_0}^{t_1} \{\tfrac{1}{2} \int [\rho h(\dot{u}^2 + \dot{w}^2) + \rho h^2 \dot{\theta}^2/12]\, 2\pi r\,ds$$

$$- \tfrac{1}{2}\int [\, \underline{M}^T \underline{\kappa} + \underline{N}^T \underline{\varepsilon} + Q\gamma]\, 2\pi r\,ds$$

$$+ \int wq\, 2\pi r\,ds\}dt = 0 \tag{2.6}$$

where a superior dot denotes differentiation with respect to t, ρ is the mass density, and q is the distributed transverse loading.

III. PENALTY STRAIN ISSUES AND INTERPOLATION CONSEQUENCES. The pivotal issue in formulating an effective finite element based on this theory is the resolution of a penalty effect engendered in the thin shell-element regime. For the present case, we distinguish two types of penalized strains that control thin-regime behavior — the membrane meridian strain and the transverse shear strain. With ℓ being fixed and $h \to 0$, the thin limits of membrane inextensibility and shearless (Poisson-Kirchhoff) deformation are enforced at the element level. The pivotal constraints take the form:

Meridian inextensibility: $\varepsilon_s = u_{,s} + w_{I,s}\, w_{,s} \to 0$ (3.3)

Poisson-Kirchhoff: $\gamma_s = w_{,s} - \theta \to 0$ (3.4)

It follows that this constraining action reduces the number of independent d.o.f. by at least two. When standard isoparametric schemes are used (i.e., uniform kinematic interpolation), spurious 'locking' constraints take precedence, making an element extremely stiff [3]. Clearly, the lower-order elements are most susceptible to 'locking'.

The desired interpolation requirements are that:
(1) the polynomial degrees of u, w, and θ should accommodate consistent coupling within the vanishing strain coefficients (penalty modes); (2) the number of penalty modes should be small to further reduce the possibility of excessive kinematic constraining.

Having a simple two-node element as our goal, the C^0 interpolation strategy developed in [2] is invoked. Considering (3.4), it is clear that if $\theta=O(s)$ (linear) then w should be $O(s^2)$. The interpolation for u can then be derived from the requirement posed by the penalty constraints of (3.3). Assuming w_I is cubic (refer to Fig. 1),

$$w_I(x) = \beta_0 \ell(\eta - 2\eta^2 + \eta^3) + \beta_1 \ell(\eta^3 - \eta^2), \tag{3.5}$$

where

$$\beta_i = w_{I,s}(\eta_i),\ i=0,1 \quad (\eta = s/\ell \subset [0,1])$$

it follows that $u=O(s^4)$. By explicitly enforcing the shear and membrane meridian strains to be constant within the element, in a manner coincident with the curved beam formulation [2], the desired two-node (six d.o.f.) kinematic field is derived:

$$\theta = \sum_{i=0}^{1} N_i(\eta)\ \theta_i, \quad w = \sum_{i=0}^{1} [N_i(\eta)\ w_i + K_i(\eta^2)\ \theta_i]$$

$$u = \sum_{i=0}^{1} [N_i(\eta)\ u_i + L_i(\eta^3)\ w_i + M_i(\eta^4)\ \theta_i]. \tag{3.6}$$

where expressions for the shape functions can be found in [2]. The resulting constant ε_s and γ strains are:

$$\varepsilon_s = \frac{1}{\ell}(u_1 - u_0) + \beta(\theta_1 - \theta_0), \tag{a}$$

$$\gamma = \frac{1}{\ell}(w_1 - w_0) - (\theta_1 + \theta_0)/2. \tag{b}$$

(3.7)

where

$$\beta = (\beta_1 - \beta_0)/12.$$

In the thin limit ($\varepsilon_s \to 0$, $\gamma \to 0$), these penalty modes ensure the desired kinematic coupling.

IV. STRAIN ENERGY UPGRADING VIA PENALTY RELAXATION. The preceding formulation, utilizing analytic shell equations directly into a finite element variational scheme, may be regarded as conventional. This common approach renders the membrane and shear penalty constraints (3.3) and (3.4) enforceable at the element level, consequently requiring consistent interpolations [2,12] or related strategies [13,14] to overcome the thinness limitations. Although in one-dimensional interpolation models such strategies are generally successful, they are, in fact, insufficient in three-dimensional plate/shell models, where boundary restraints often produce shear locking [3,4].

Another deficiency of the conventional approach is that in the thin regime the stiffness matrix is ill-conditioned. In addition to requiring high-precision computations, the ill-conditioning causes prohibi-

tively small time steps in explicit transient integrations [15] and, as we shall see further, unrealistically large errors in the higher natural frequencies and corresponding mode shapes.

We therefore view the conventional approach as too prohibitive for generating simple and effective elements. To produce well behaved thin-regime elements, an element level relaxation of penalty constraints is undertaken. The idea is to introduce a correction parameter in the element constitutive relations that would account for the limited kinematic freedoms of penalty strains by reducing the penalty parameter in the thin limit. In this circumstance, the penalty constraints are said to be relaxed at the element level. The problems of locking, excessive stiffening, and ill-conditioning would then be eliminated.

Shear relaxation. For clarity, it suffices to consider the shallowly curved beam [2], possessing the same basic penalty features as the present shell. To illustrate the concept, consider a curved cantilever beam, with an initial cubic shape, loaded at its free end by both membrane, N_1, and shear, Q_1, forces. The shear relaxation (correction [7]) is introduced via a positive parameter ϕ^2_s, which appears as a multiplier in the transverse shear constitutive relations of the element:

$$Q^{\phi} = \phi_s^2 \, Q = \phi_s^2 \, k^2 GA\gamma \tag{4.1}$$

Equating the strain energy captured by a single anisoparametric curved beam element of the lowest order, (p=1) [2], with the exact strain energy for this problem, and solving for ϕ^2_s results in the same expression as in the straight beam case [7]:

$$\phi_s^2 = (1 + C_s \alpha_s)^{-1} \tag{4.2}$$

in which

$$\alpha_s \equiv 3k^2 \frac{G}{E} (\ell/h)^2 \tag{4.2-a}$$

is the shear penalty parameter, and $C_s = 1/3$.

An important consequence of the above result is that the modified element has a new penalty parameter

$$\alpha_s^{\phi} = \alpha_s \, \phi_s^2 = \frac{\alpha_s}{1 + C_s \alpha_s} \tag{4.3}$$

with the following desirable properties:

$$\alpha_s^{\phi} \dashrightarrow \begin{cases} C_s^{-1} & \text{if } h \to 0 \text{ (with fixed } \ell\text{), i.e. thin regime} \\ \alpha_s & \text{if } \ell \to 0 \text{ (with fixed h), i.e. thick regime} \end{cases}$$

In the thin limit, the new penalty parameter approaches a finite value. This implies that the Poisson-Kirchhoff constraint is relaxed at the element level. (By contrast, the conventional penalty parameter approaches infinity in this case).

It is apparent from this analysis that the new element is upgraded in the energy sense to the level of a higher-order element, namely, the second order element (p=2), which happens to model this cantilever beam problem exactly.

Membrane relaxation. In [2], in addition to the shear relaxation parameter ϕ^2_s, we also employed the membrane relaxation parameter, ϕ^2_m, which served as a multiplier in the membrane constitutive relations

$$N^{\phi} = \phi_m^2 N = \phi_m^2 AE\varepsilon \tag{4.4}$$

having the form analogous to that of ϕ_s^2:

$$\phi_m^2 = (1 + C_m \alpha_m)^{-1} \tag{4.5}$$

in which α_m is the conventional membrane penalty parameter

$$\alpha_m = \frac{k_{m\theta\theta}}{k_{b\theta\theta}} = 12\,(\beta\ell/h)^2 \tag{4.6}$$

The constant $C_m = 1/4$ was established on the basis of numerical tests to yield an overall best element performance. (Although the results reported in [2] were based on the correct α_m shown in (4.6), the β contribution in α_m was typographically omitted in the text). It was found that ϕ_m^2 produced only minor solution improvements; the major enhancement was due to the shear relaxation, ϕ_s^2.

This outcome can be predicted by assessing the relative strength of the two conventional penalty parameters, which can be defined by the ratio:

$$R \equiv \frac{\alpha_m}{\alpha_s} = \frac{4\,E}{k^2 G}\,\beta^2 \tag{4.7}$$

For a typical isotropic shallow element $R \leq 0.01$, and thus α_m is at least two orders of magnitude weaker than its shear counterpart. This implies that much of the penalty related stiffening action is predominantly due to α_s. Thus, the ϕ^2_m relaxation of the inextensional membrane strain is not essential. We shall further highlight many of these issues by means of numerical examples.

V. NUMERICAL RESULTS. We focus our numerical studies exclusively on natural vibrations of spherical shells, ranging from shallow to deep and from thin to moderately thick. Our motivation is to assess the element on the basis of its dynamic performance by examining a wide range of vibrational modes. The computed frequencies are compared with available analytic and finite element solutions.

In all numerical examples, unless stated otherwise, the values of $E/G=2.6$ and $k^2=\pi^2/12$ were assumed. The calculations were carried out on an Apollo DN3000 in double precision.

Anisoparametric versus Isoparametric element. In this study we establish an appropriate Gaussian quadrature rule for the present element without shear relaxation (labelled ANISO•2) and compare the element performance to that of a two-node linear isoparametric element (labelled ISO•2). The test problem is a deep/thin clamped hemispherical shell (see Figure 2d) which, due to its thinness (R/h=100) and deep curvature, is a challenging 'locking' test for this class of shear-deformable curved elements.

Table 1 summarizes the ten lowest symmetric frequencies obtained with a 16-element ISO•2 discretization using 1-, 2-, and 3-point Gaussian quadrature. The results are compared with the benchmark frequencies from a 256-element ANISO•2 model, fully integrated with 3-point Gaussian quadrature. The 2- and 3-point quadrature solutions agree very closely. The first frequency is sufficiently accurate, however, the higher modes experience severe stiffening (locking) as evidenced by their overestimated frequencies. The results corresponding to the 1-point quadrature, which underintegrates all energy contributions (this curved element is a direct analog of the 1-point quadrature conical shell of Zienkiewicz et. al. [16]), produce frequencies converging from either below or above, confirming its variational inconsistency. Again, the highest frequencies are noticeably overestimated.

By contrast, all ten frequencies obtained with ANISO•2 (see Table 2) using 2- and 3-point quadratures are highly accurate. The results based on the 1-point quadrature, which exactly integrates strain energy contributions due to constant strains (ε_s,γ) and curvature (κ_s), are slightly even more accurate, converging consistently from above (the convergence study is not shown). However, further studies must be carried out to verify the reliability of ANISO•2 with the 1-point integration. Henceforth, the 2-point quadrature will be used to integrate the ANISO•2 element.

C_s constant. A suitable C_s value can be established by insisting upon monotonic convergence of vibration frequencies from above, a property which is intrinsic to conforming displacement models. For this purpose, taking into account that ϕ^2_s is independent of β_i, it suffices to seek frequencies of vibration of a circular plate (see Figure 2a). In Figure 3, the error of the first symmetric frequency of a clamped circular plate (R/h=100) is plotted versus the number of elements; where the four curves correspond to C_s = 0, 1/20, 1/8, and 1/5. The best results are obtained with C_s=1/8. Henceforth, this value is adopted for the element, labelled ANISO•2ϕ_s. Note that the beneficial effect of C_s is especially pronounced in the coarse models.

Further evidence of the effect of shear relaxation is illustrated in Figure 4, where the first, third, and fifth symmetric frequencies of ANISO•2 (i.e., C_s=0) are normalized with the corresponding ANISO•2ϕ_s results. It is seen that the higher frequencies ben fit the most from the shear relaxation.

Shallow shells. Tables 3 and 4 summarize the ANISO•2 and ANISO•2ϕ_s results, respectively, for the first five symmetric frequencies of a

thin (R/h=100) and moderately thick (R/h=10) 10-degree clamped spherical shell (see Figure 2b). The 256-element benchmark frequencies and those obtained by a modified Holzer method [17] are cited for comparison purposes. The frequencies obtained with the ANISO•2ϕ elements are consistently lower than those of ANISO•2; hence, they are more accurate, since the convergence is from above. The results are also superior to those reported in [17]. Note that the effect of shear relaxation is particularly beneficial in coarsely discretized models and higher vibrational modes. The diminishing influence of the shear relaxation parameter is noticeable as the mesh is further refined.

Deep shells. Tables 5 and 6 contain the ANISO•2 and ANISO•2ϕ results, respectively, for the first five symmetric frequencies of a thin (R/h=100) and moderately thick (R/h=10) clamped hemispherical shell (see Figure 2d). Again, the 256-element benchmark frequencies and those obtained by a modified Holzer method [17] are cited for comparison purposes. The ANISO•2ϕ frequencies are consistently more accurate than those of ANISO•2 and those reported in [17].

To benchmark the element behavior further, we compared ANISO•2ϕ with two commonly used isoparametric axisymmetric shell elements from the ABAQUS finite element program [23], SAX1 (2-node, linear) and SAX2 (3-node, quadratic). Both of the ABAQUS elements use reduced integration on the shear energy, and a shear relaxation parameter of the form somewhat different than the present one. Figure 5 shows the percent error for the first ten symmetric frequencies using a 48-d.o.f. model for the thin, clamped hemispherical shell. Whereas ANISO•2ϕ performs consistently well, SAX1 produces rather poor frequencies thoughout, and SAX2 begins to deteriorate at higher frequencies. In addition, unlike ANISO•2ϕ, SAX1 and SAX2 do not converge monotonically -- some frequencies converge from above, while others converge from below.

Sixty-degree shell. Our motivation for analyzing the clamped 60-degree shell (R/h=20) (see Figure 2c) was to compare the present element results with several others [18-20]. Table 7 contains frequencies for the first eight symmetric modes of vibration. A 24-element model was used. The present elements produce consistently lower frequencies, and because they converge from above, they are of superior accuracy. Of particular interest is the steady progression of improvement as the element is upgraded from ANISO•2 to ANISO•2ϕ.

Explicit integration. In this example we illustrate an often neglected attribute of penalty relaxation. In the explicit conditionally stable transient integration, the critical time step is bounded by the inverse of the largest natural frequency of the individual elements (ω^e_{max}; e.g.,see [15]). Figure 6 depicts the normalized critical time step

$$\Delta t_{crit} = 2c/\ell\omega^e_{max} \quad (c=\sqrt{E/\rho} \text{ — bar-wave velocity})$$

for a single, lumped mass element ($\beta_0=-\beta_1=\pi/64$) plotted versus ℓ/h . While the penalty-relaxed solution (ANISO•2ϕ, $C_s=0.125$) is bounded by a constant ($\Delta t_{crit}=0.257$) as $\ell/h\to\infty$, the standard formulation (ANISO•2, $C_s=0$) exhibits an exponential decline, falling several orders of magnitude below the results for the relaxed case. This example dramatically

illustrates the enormous computational efficiency that can be achieved by shear relaxation. Notably, other methods for enhancing thin-regime behavior (e.g., [21,22]) do nothing to improve on the poor critical time step performance of the standard (unrelaxed) element.

Extreme thinness regime. All shell problems presented herein were solved using an extremely thin shell geometry ($R/h=10^6$). No locking of any type was observed in this extreme-thinness regime.

VI. CONCLUDING SUMMARY. We have presented a shallowly-curved axisymmetric shell element which includes the effects of shear deformation and rotary inertia. In our displacement formulation, we focus particular attention upon the anisoparametric interpolations, shear relaxation, and low-order numerical intergartion. The result is a simple and efficient two-node shell devoid of shear and membrane locking, having no thinness limitations. In addition, shear relaxation (correction) of the shear penalty improved coarse-mesh behavior and produced an element of superior efficiency in explicit time integration.

We regard this element as an excellent candidate for large-scale computations, nonlinear applications, time integration procedures, and microcomputer implementation. Finally, the present methodology appears ideally suited for application to general shell models.

REFERENCES.

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

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

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

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

5. I. Fried, "Finite element analysis of thin elastic shells with residual energy balancing and the role of the rigid body modes," J. Appl. Mech., 6 (1975).

6. R. H. MacNeal, "A simple quadrilateral shell element," Computers & Structures, 8, 175-183 (1978).

7. A. Tessler and T. J. R. Hughes, "An improved treatment of transverse shear in the Mindlin-type four-node quadrilateral element," Comput. Meths. Appl.Mech. Engrg., 39, 311-335 (1983).

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

9. P. M. Naghdi, "Foundations of elastic theory," in Progress in Solid Mechanics, (eds. I.N. Sneddon and R. Hill), North-Holland, Amsterdam, Vol. IV, Chap. 1 (1963).

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

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

12. H. R. Meck, "An accurate polynomial displacement function for finite ring elements," Computers and Structures, 11, 265-269 (1980).

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

14. G. Prathap, "The curved beam/deep arch/finite ring element revisited," Int. J. Numer. Meth. Engng., 21, 389-407 (1985).

15. T.J.R. Hughes, "Analysis of transient algorithms with particular reference to stability behavior," in Computational Methods for Transient Analysis, (eds. T. Belytschko and T.J.R. Hughes), North-Holland, Amsterdam, Vol. 1, Chap. 2 (1983).

16. O.C. Zienkiewicz, J. Bauer, K. Morgan and E. Onate, "A simple and efficient element for axisymmetric shells," Int. J. Numer. Meth. Engng. 11, 1545-1558 (1977).

17. M.S. Zarghamee and A.R. Robinson, "A numerical method for analysis of free vibration of spherical shells," AIAA J., 5, 1256-1261 (1967).

18. A. Kalnins, "Effect of bending on vibrations of spherical shells," J. Acoust. Soc. Am., 36, 74-81 (1964).

19. E.W. Ross, Jr., "Natural frequencies and mode shapes for axisymmetric vibration of deep spherical shells," J. Appl. Mech., 32, 553-561 (1965).

20. D.R. Navaratna, "Natural vibrations of deep spherical shells," AIAA J., 4, 2056-2058 (1966).

21. G. Prathap and C. Ramesh Babu, "A field consistent three-noded quadratic curved axisymmetric shell element," Int. J. Numer. Meth. Engng., 23, 711-723 (1986).

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

23. The ABAQUS Theoretical Manual, Level 4.5, Hibbitt, Karlsson, and Sorenson, Inc., Providence, RI, 4.3.1-19 (1985).

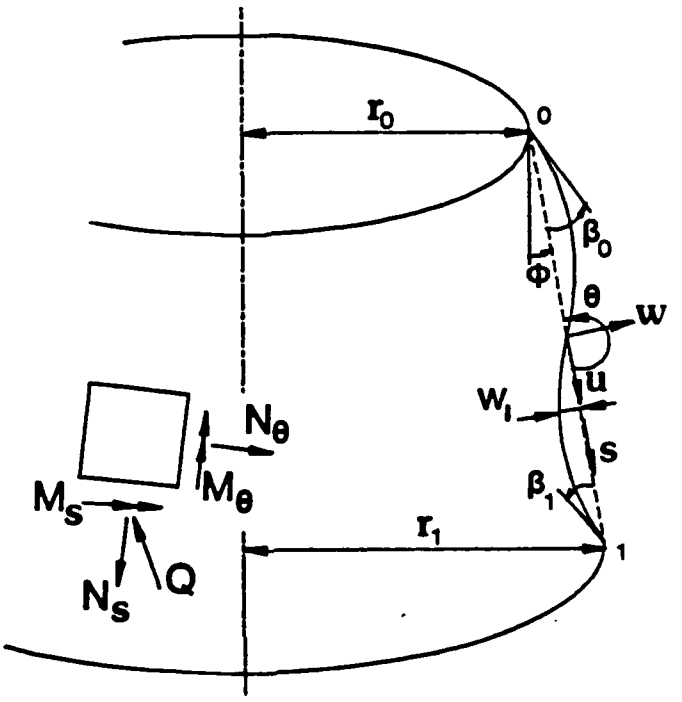


Figure 1. Shallowly curved axisymmetric shell element.

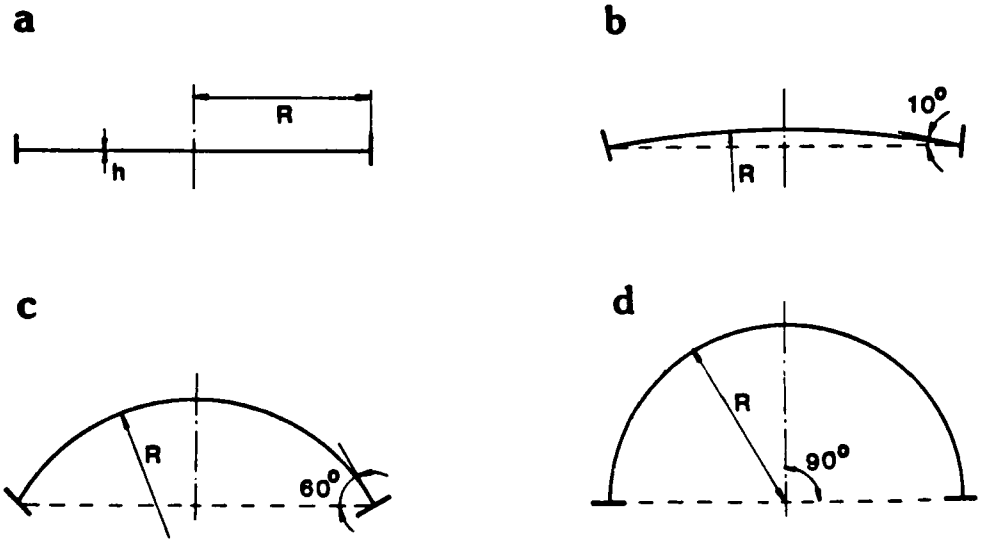


Figure 2. (a) Clamped circular plate; (b) Shallow 10 deg. shell; (c) Deep 60 deg. shell; (d) Deep hemispherical shell.

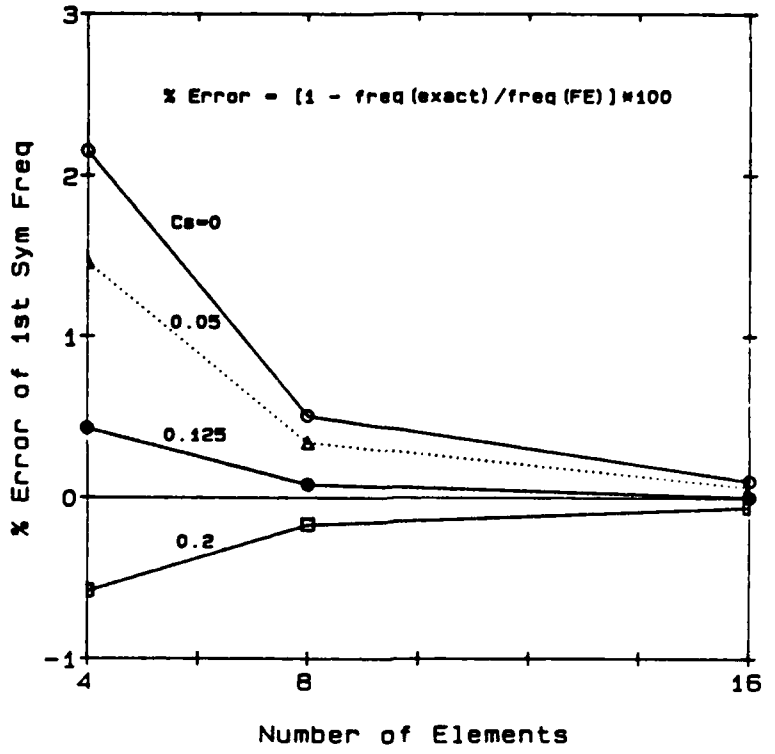


Figure 3. Vibration of clamped circular thin plate (R/h=100); 1st symmetric frequency error for various C_s vs. number of elements.

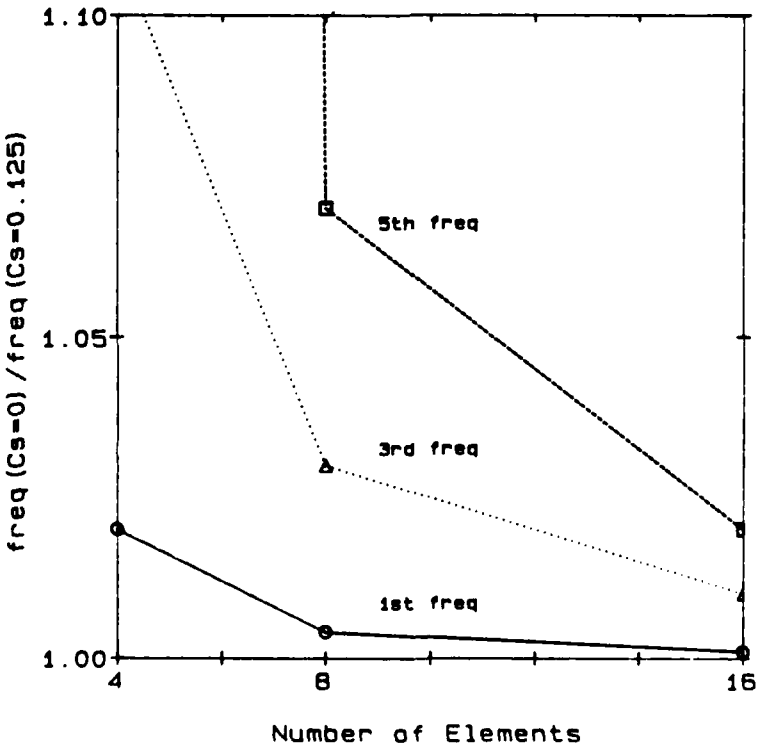


Figure 4. Vibration of clamped circular thin plate (R/h=100); normalized symmetric frequencies versus number of elements.

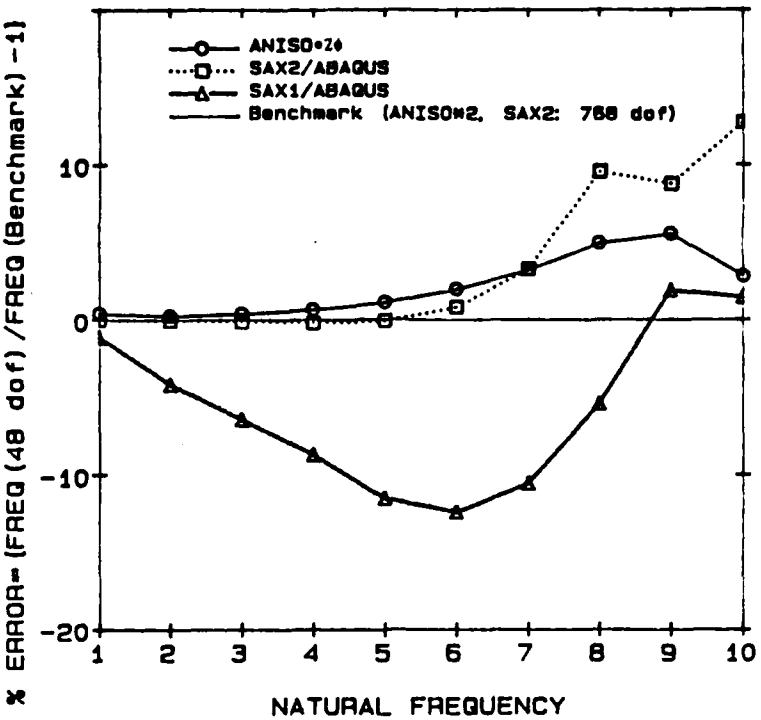


Figure 5. Vibration of clamped, thin hemispherical shell (R/h=100, ν=0.3); comparison of ANISO•2ϕ with SAX1 and SAX2 of ABAQUS for the ten lowest symmetric frequencies.

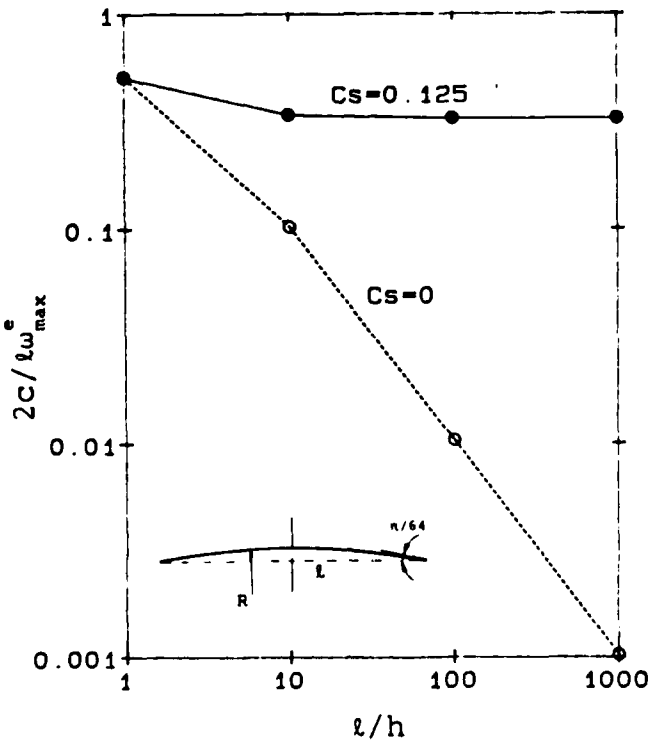


Figure 6. Normalized critical time step vs. (ℓ/h) for a shallow shell; comparison of shear-relaxed (ANISO•2ϕ) and unrelaxed (ANISO•2) models.

TABLE 1. A study of Gaussian integration for ISO•2; symmetric nondimensional natural frequencies $\omega_i [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped, thin hemispherical shell (R/h=100, ν=0.3) discretized with 16 ISO•2 elements.

Mode no.	BENCHMARK*	Gaussian Integration Order (n_{int})		
		1	2	3 or higher
1	0.7262	0.7253	0.7666	0.7666
2	0.8948	0.8905	0.9817	0.9818
3	0.9382	0.9359	1.2389	1.2389
4	0.9730	0.9813	1.5315	1.5315
5	1.021	1.057	1.796	1.796
6	1.090	1.196	2.425	2.425
7	1.187	1.427	2.717	2.717
8	1.313	1.603	3.387	3.387
9	1.465	1.882	4.479	4.479
10	1.582	2.569	4.682	4.682

*BENCHMARK
ANISO•2: n_{el} = 256, n_{dof} = 768, n_{int} = 3
or SAX2/ABAQUS: n_{el} = 128, n_{dof} = 768, n_{int} = 2
(n_{el} = number of elements; n_{dof} = number of DOF; n_{int} = Integration order)

TABLE 2. A study of Gaussian integration for ANISO•2; symmetric nondimensional natural frequencies $\omega_i [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped, thin hemispherical shell (R/h=100, ν=0.3) discretized with 16 ANISO•2 elements.

Mode no.	BENCHMARK*	Gaussian Integration Order (n_{int})		
		1	2	3 or higher
1	0.7262	0.7300	0.7303	0.7303
2	0.8948	0.8959	0.8975	0.8975
3	0.9382	0.9394	0.9423	0.9423
4	0.9730	0.9763	0.9808	0.9808
5	1.021	1.029	1.037	1.036
6	1.090	1.109	1.122	1.122
7	1.187	1.223	1.249	1.246
8	1.313	1.374	1.421	1.416
9	1.465	1.537	1.571	1.569
10	1.582	1.612	1.685	1.672

TABLE 3. Nondimensional symmetric natural frequencies $\omega_i\ [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped shallow (10 deg), thin spherical shell (R/h=100, ν=0.3)

No. of el.	Shear relaxation C_s	ϕ^2_s	Mode number 1	2	3	4	5
4	0	1	1.6511	4.2265	10.3590	22.3270	35.3874
	0.125	0.307	1.6431	4.0191	9.3704	21.2623	22.3745
8	0	1	1.6432	3.9044	8.4852	15.1210	22.0575
	0.125	0.639	1.6412	3.8582	8.2779	14.5438	22.0446
16	0	1	1.6412	3.8342	8.1167	13.9544	21.1207
	0.125	0.876	1.6407	3.8229	8.0675	13.8227	20.8509
64	0	1	1.6406	3.8129	8.0086	13.6263	20.3621
	0.125	0.991	1.6406	3.8122	8.0055	13.6183	20.3458
BENCHMARK							
256	0	1	1.6406	3.8116	8.0019	13.6063	20.3161
	0.125						
MODIFIED HOLZER METHOD [17]			1.6556	-	-	-	-

TABLE 4. Nondimensional symmetric natural frequencies $\omega_i\ [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped shallow (10 deg), moderately thick spherical shell (R/h=10, ν=0.3)

No. of el.	Shear relaxation C_s	ϕ^2_s	Mode number 1	2	3	4	5
4	0	1	6.0209	15.5464	22.3395	27.3698	31.1135
	0.125	0.978	5.9858	15.4274	22.3390	27.1741	30.9162
8	0	1	5.9753	14.9272	22.0599	25.2442	30.1047
	0.125	0.994	5.9666	14.8989	22.0597	25.1942	30.0633
16	0	1	5.9638	14.7707	21.9879	24.6198	29.8900
	0.125	0.999	5.9617	14.7637	21.9878	24.6074	29.8799
64	0	1	5.9602	14.7214	21.9650	24.4178	29.8236
	0.125	1.000	5.9601	14.7210	21.9650	24.4170	29.8229
BENCHMARK							
256	0	1	5.9600	14.7183	21.9636	24.4051	29.8194
	0.125						

TABLE 5. Nondimensional symmetric natural frequencies $\omega_i \, [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped hemispherical thin shell (R/h=100, ν=0.3)

No. of el.	Shear relaxation C_s	ϕ^2_s	Mode number 1	2	3	4	5
4	0	1	0.7826	0.9448	1.4345	2.6343	3.8842
	0.125	0.006	0.7454	0.9270	0.9937	1.0948	1.4494
8	0	1	0.7424	0.9058	0.9551	1.0064	1.0909
	0.125	0.022	0.7352	0.9037	0.9523	0.9982	1.0693
16	0	1	0.7303	0.8975	0.9423	0.9808	1.0364
	0.125	0.081	0.7289	0.8971	0.9418	0.9792	1.0321
64	0	1	0.7264	0.8949	0.9384	0.9734	1.0215
	0.125	0.583	0.7263	0.8949	0.9384	0.9734	1.0213
BENCHMARK							
256	0						
	0.125	1	0.7262	0.8948	0.9382	0.9730	1.0206
MODIFIED HOLZER METHOD [17]			0.7263	0.8948	-	-	-

TABLE 6. Nondimensional symmetric natural frequencies $\omega_i \, [\rho R^2(1-\nu^2)/E]^{\frac{1}{2}}$ for clamped hemispherical moderately thick shell (R/h=10, ν=0.3)

No. of el.	Shear relaxation C_s	ϕ^2_s	Mode number 1	2	3	4	5
4	0	1	0.8514	1.3295	1.6280	2.6573	3.9533
	0.125	0.356	0.8373	1.2772	1.5976	2.4814	2.9115
8	0	1	0.8187	1.2112	1.5464	1.9975	2.6501
	0.125	0.687	0.8159	1.1994	1.5388	1.9480	2.6243
16	0	1	0.8112	1.1842	1.5251	1.8859	2.5394
	0.125	0.898	0.8105	1.1815	1.5229	1.8753	2.5228
64	0	1	0.8089	1.1761	1.5178	1.8555	2.4829
	0.125	0.993	0.8088	1.1760	1.5176	1.8549	2.4817
BENCHMARK							
256	0						
	0.125	1	0.8087	1.1756	1.5173	1.8537	2.4792

Table 7. Nondimensional natural frequencies $\Omega = \omega R (\rho/E)^{\frac{1}{2}}$ for 60 deg. clamped spherical shell (R/h=20, ν=0.3).

Mode No.	Analytic Results		FE Results with 24 Elements		
	Kalnins [18]	Ross [19]	Navaratna [20]	ANISO•2	ANISO•2ϕ
1	1.006	-	1.008	1.0006	1.0003
2	1.391	1.335	1.395	1.370	1.368
3	-	-	1.702	1.675	1.673
4	2.375	2.368	2.387	2.268	2.260
5	3.486	3.478	3.506	3.230	3.213
6	3.991	-	3.996	3.967	3.965
7	4.974	4.970	5.001	4.475	4.442
8	6.690	6.687	6.729	5.829	5.773

A BLOCK QR FACTORIZATION SCHEME FOR LOOSELY COUPLED SYSTEMS OF ARRAY PROCESSORS

Charles Van Loan
Department of Computer Science
Cornell University
Ithaca, New York 14853

Abstract

A statically scheduled parallel block QR factorization procedure is described. It is based on "block" Givens rotations and is modeled after the Gentleman-Kung systolic QR procedure. Independent tasks are associated with each block column. "Tallest possible" subproblems are always solved. The method has been implemented on the IBM Kingston LCAP-I system which consists of ten FPS-164/MAX array processors that can communicate through a large shared bulk memory. The implementation revealed much about the tradeoff between block size and load balancing. Large blocks make load balancing more difficult but give better 164/MAX performance and less shared memory traffic. The results obtained indicate that our approach to parallelizing the QR factorization is competitive for very large problems, e.g., of the order 5000-by-1000.

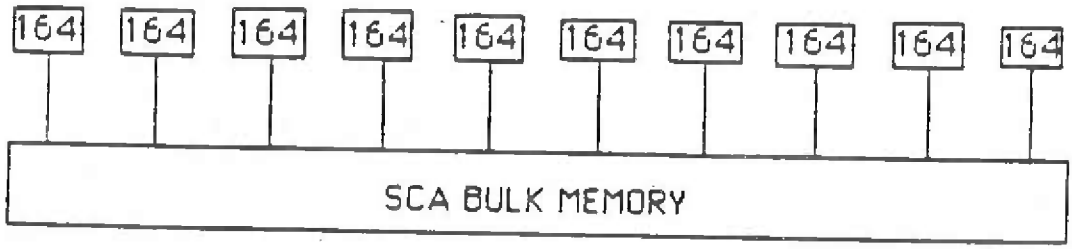
* This work has been supported by ONR contract N00014-83-K-0640, NSF contract CCR86-02310, the Mathematical Sciences Institute at Cornell which is sponsored by the Army Research Office, and by the IBM Corporation. Computations were performed at IBM Kingston and on the Production Supercomputer Facility at Cornell which is supported in part by the National Science Foundation and IBM.

1. Introduction

Computing the QR factorization of a matrix $A \in R^{m\times n}$ involves finding an orthogonal matrix $Q \in R^{m\times m}$ and an upper triangular matrix $R \in R^{m\times n}$ such that $A = QR$. This factorization has a prominent role to play in numerical linear algebra especially because of its bearing on the least square problem. A detailed description of the QR factorization and the various ways that it can be computed may be found in Golub and Van Loan (1983).

Parallel methods for computing the QR factorization have received considerable attention recently. For systolic arrays attention has focussed on methods that rely on Givens rotations. See Gentleman and Kung (1981) or Heller and Ipsen (1983). Dongarra, Sameh, and Sorenson (1986) have implemented both parallel Givens and parallel Householder procedures on the Denelcor Hep.

In this paper we discuss a block version of the Gentleman-Kung method that we have implemented on the IBM Kingston LCAP-1. This system consists of ten FPS-164 array processors (APs) that can communicate through several shared bulk memories. An overview of LCAP-1 is offered in Clementi and Logan (1985). The features of LCAP-1 that figure in the current work are depicted in the following diagram:



There are actually two levels of parallelism here because the APs are each capable of performing twenty parallel dot products. Indeed, the FPS-164/MAX's at Kingston each come equipped with two "MAX boards". The MAX board enhancement enables each AP to perform matrix-matrix multiplication at a peak rate of 55 Mflops if the matrices involved are sufficiently large. Full exploitation of the FPS-164/MAX requires having an algorithm that is rich in matrix multiplication. This is why we have

chosen to develop a parallel block procedure. The blocking of the matrix A is largely a function of the 164/MAX architecture. For example, it turns out to be efficient to have block columns that are a multiple of twenty simply because the LCAP-1 APs can each perform twenty parallel dot products. Further details concerning the FPS-164/MAX architecture may be found in Charlesworth and Gustafson (1986).

The matrix A is stored in a 64 Mword bulk memory unit manufactured by Scientific Computing Associates (SCA). Thus, a dense problem of size 16K-by-4K could potentially be solved. The APs have approximately 600 Kwords of usable memory. This is enough to house, for example, a 1000-by-500 submatrix.

Data between the APs and the bulk memory flows at a rate of 44 Mbytes/sec. However, high latency associated with each transferred message demands that data be moved in fairly good-sized chunks in order to be efficient, e.g., 1000 words.

Additional nuances of the LCAP-1 system as they apply to our QR implementation are detailed later.

This paper is the first of several reports in which we explore the issues associated with parallel matrix computations on the LCAP-1. The parallel block QR factorization scheme that we encoded is derived in §2 and §3. Implementation details are covered in §4 and results in §5. Our current QR code can be improved in several ways as we often opted for the "easy way out" when confronted with an algorithmic dilemma. Despite this we feel that our LCAP-1 experience offers general perspectives on large scale distributed matrix computations.

2. Parallel Givens QR

We say that $G \in R^{m\times m}$ is an adjacent Givens rotation in planes i-1 and i if G is the identity with the following 2-by-2 exception:

$$\begin{bmatrix} g_{i-1,i-1} & g_{i-1,i} \\ g_{i,i-1} & g_{ii} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \qquad 2 \le i \le m$$

Notice that G is orthogonal and that premultiplication by G affects just rows i-1 and i . If $x \in R^m$ then it is not hard to determine $(\cos(\theta),\sin(\theta))$ so that $y_i = 0$ if $y = Gx$. These and other Givens rotations issues are discussed in Golub and Van Loan (1983, pp.43-47).

Adjacent rotations are important because they only combine adjacent rows or columns when applied to a matrix. Moreover, they can be used to compute the QR factorization of a matrix. Assuming $A \in R^{m\times n}$ $(m \ge n)$ we have:

Algorithm 2.1

```
For j = 1:n
    For i = m : -1 : j+1
        Determine an adjacent Givens rotation G_ij such
            that if y = G_ij^T A(:, j:j) then y_i = 0, i.e., zero a_ij .
        A := G_ij^T A
    end i
end j
```

Upon completion A is overwritten by R and

$$Q = (G_{m,1} \cdots G_{21}) \cdots (G_{m,n} \cdots G_{n+1,n})$$

Notice that the algorithm computes R column-by-column and that the zeroing within a column proceeds from the bottom up to the subdiagonal. Here is a depiction of the 4-by-3 case:

$$
\begin{matrix} x\,x\,x \\ x\,x\,x \\ x\,x\,x \\ x\,x\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ x\,x\,x \\ x\,x\,x \\ o\,x\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ x\,x\,x \\ o\,x\,x \\ o\,x\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ o\,x\,x \\ o\,x\,x \\ o\,x\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ o\,x\,x \\ o\,x\,x \\ o\,o\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ o\,x\,x \\ o\,o\,x \\ o\,o\,x \end{matrix}
\rightarrow
\begin{matrix} x\,x\,x \\ o\,x\,x \\ o\,o\,x \\ o\,o\,o \end{matrix}
$$

To indicate the inherent parallelism in this procedure we resort to a slightly larger example and number the a_{ij} in the order that they are zeroed:

$$
\begin{matrix} x & x & x & x \\ 8 & x & x & x \\ 7 & 15 & x & x \\ 6 & 14 & 21 & x \\ 5 & 13 & 20 & 26 \\ 4 & 12 & 19 & 25 \\ 3 & 11 & 18 & 24 \\ 2 & 10 & 17 & 23 \\ 1 & 9 & 16 & 22 \end{matrix}
\qquad m = 9,\ n = 4
$$

Recognize that the computation and application of G_{ij} can begin as soon as $G_{i-1,j-1}$ is applied to A. To illustrate this we tabulate the earliest "time step" that a_{ij} $(i>j)$ can be zeroed:

$$
\begin{matrix} x & x & x & x \\ 8 & x & x & x \\ 7 & 9 & x & x \\ 6 & 8 & 10 & x \\ 5 & 7 & 9 & 11 \\ 4 & 6 & 8 & 10 \\ 3 & 5 & 7 & 9 \\ 2 & 4 & 6 & 8 \\ 1 & 3 & 5 & 7 \end{matrix}
\qquad m = 9,\ n = 4
$$

With this notation we see in the example that four Givens updates can be performed during the seventh time step: G_{31}, G_{52}, G_{73}, and G_{94}. If we had

4 processors then they could each be assigned one of these tasks.

The parallelism that we have exposed in the above example can be formalized by rearranging the loop indexing in Algorithm 2.1 and noting that m+n-2 timesteps are required.

Algorithm 2.2

```
For k = 1: m+n-2
    For All j = 1:n
        i = m-k+1+2(j-1)
        if ( i ≤ m & i ≥ j+1)
            Determine G_ij to zero a_ij
            A := G_ij^T A
        end
    end j
end k
```

The "For All" statement reminds us that all of the updates $A := G_{ij}^T A$ associated with a given time step k are independent and can be performed in parallel.

We point out that G_{ij} can actually be computed "earlier" than we have indicated. For example, in the (m,n) = (9,4) case above, we have assumed that G_{92} is computed as soon as G_{81} has been applied all the way across the matrix. In fact, G_{92} can be computed as soon as G_{81} has been applied to just the second column. For reasons that we give in §4, we have not implemented the "soon as possible" generation of G_{ij}.

Algorithm 2.2 and its natural variants can be mapped nicely onto systolic networks. See Heller and Ipsen (1983).

3. A Parallel Block QR Factorization Method

Some notation is required before a block version of Algorithm 2.2 can be specified. Partition $A \in R^{m \times n}$ as follows:

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{p1} & \cdots & A_{pq} \end{bmatrix} \begin{matrix} m_1 \\ \\ m_p \end{matrix} \qquad (3.1)$$

$$\begin{matrix} n_1 & & n_q \end{matrix}$$

Here, A_{ij} is m_i-by-n_j and we assume that $m_i \geq n_j$ for all i and j. If Q is an orthogonal matrix of dimension $m_{i-1} + m_i$ then we refer to

$$G_i(Q) = \text{diag}(I_{m_1}, \ldots, I_{m_{i-2}}, Q, I_{m_{i+1}}, \ldots, I_{m_p})$$

as an adjacent "block Givens" rotation in block planes i-1 and i.

Algorithm 3.1 (Block Givens QR Factorization)

```
For k = 1: p+q-2
    For All j = 1:q
        i = p-k+1+2(j-1)
        if ( i ≤ p & i ≥ j+1 )
            Determine orthogonal Q_ij such that

                Q_ij^T [ A_i-1,j ]  =  [ R ]   (R upper triangular)
                       [ A_ij    ]     [ 0 ]

            Set G_ij = G_i(Q_ij) and update A := G_ij^T A
        end
    end j
end k
```

This procedure is identical to Algorithm 2.2 except that blocks are zeroed instead of scalars. Upon completion A is overwritten with a block upper triangular matrix R. Unless all the A_{ij} are square, then R will not be upper triangular as a scalar matrix. For example, if the partitioning in (3.1) is defined by $(m_1,m_2) = (3,3)$ and $(n_1,n_2) = (2,2)$ then Algorithm 3.1 overwrites A with

$$R = \left[\begin{array}{cc|cc} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ \hline 0 & 0 & x & x \\ 0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Of course, it is possible to upper triangularize this matrix with further Givens operations, but that is an annoying but necessary follow-up computation.

However, there is a more serious problem associated with rectangular blocks. Consider the example $(m_1,m_2,m_3,m_4) = (2,3,3,8)$, $(n_1,n_2) = (2,2)$. At the beginning of the second time step A looks like

$$\left[\begin{array}{cc:cc} x & x & x & x \\ x & x & x & x \\ \hline x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ \hline x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ \hline 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \\ 0 & 0 & x & x \end{array}\right]$$

At this stage, Algorithm 3.1 specifies that we only upper triangularize the submatrix A(3:8,1:2), i.e., the subproblem defined by blocks A_{21} and A_{31}. However, we see from the figure that a significant amount of zeroing in the second block column can take place concurrently. In particular, we could upper triangularize both A(3:8,1:2) and A(9:16,3:4) .

In general, because the "bottom" submatrix A_{ij} in each subproblem is upper triangular, "taller" submatrices can be upper triangularized throughout Algorithm 3.1. In order to rearrange this algorithm so that "maximally tall" subproblems are solved at each stage, we need to drop the fixed row blocking in (3.1). We continue to assume that A has q block columns with widths n_1 ,..., n_q . However, instead of imposing a fixed blocking of A's rows we have chosen to determine the "height" of the subproblems through an integer parameter m_0 that satisfies $m_0 \geq n_1$. In our scheme, the subproblems in the first block column involve at most m_0 rows. Maximally tall subproblems are then solved in subsequent block columns at each step. To illustrate, consider the case m = 100, m_0 = 20, and (n_1,n_2,n_3,n_4) = (2,3,5,5):

Subproblem Row Ranges

Time Step	Column Ranges			
	1:2	3:5	6:10	11:15
1	81:100	-	-	-
2	63:82	83:100	-	-
3	45:64	65:85	86:100	-
4	27:46	47:67	68:90	91:100
5	9:28	29:49	50:72	73:95
6	1:10	11:31	32:54	55:77
7	-	3:13	14:36	37:59
8	-	-	6:18	19:41
9	-	-	-	11:23

In general four integers rowsrt(t,j), rowend(t,j), colsrt(j), and colend(j) are necessary to describe subproblem (t,j), e.g., 29, 49, 3, and 5 for subproblem (5,2). These index arrays and the total number of time steps

t_f required can be computed as follows:

Algorithm 3.2

Let m, n, m_0, q and the column partitioning $(n_1,\ldots,n_q)$ be given with $m \geq n$ and $m_0 > n_1$. This algorithm determines t_f and the index arrays colsrt(1:q), colend(1:q), rowsrt(1:t_f,1:q), and rowend(1:t_f,1:q).

```
t_f = ceiling( max(0,m-m_0) / (m_0-n_1) ) + q
For t = 1 : t_f
   if t = 1
      For j = 1:q
           if  j = 1
                colsrt(1) = 1
                colend(1) = n_1
                rowsrt(t,j) = max( 1 , m-m_0+1)
           else
                colsrt(j) = colend(j-1) + 1
                colend(j) = colend(j-1) + n_j
                rowsrt(t,j) = m
           end
           rowend(t,j) = m
      end j
   else
      For j = 1:q
           if j = 1
                rowend(t,1) = rowsrt(t-1,1) + n_1 - 1
                rowsrt(t,1) = max(1, rowend(t,1)-m_0+1 )
           else
                rowend(t,j) = min( rowsrt(t-1,j) + n_j - 1 , m )
                rowsrt(t,j) = max( colsrt(j), min( rowend(t,j-1) +1, m))
           end
      end j
   end
end t
```

A couple of comments are in order. In block column 1, the subproblems "climb" at the "rate" $m_0 - n_1$ and so $1 + \text{ceiling}(\max(0,m-m_0)/(m_0-n_1))$ steps are required to complete the processing of block column 1. Thereafter one block column per time step is completed. This explains the formula for t_f and why we must have $m_0 > n_1$.

In block column j, "serious" computation does not begin so long as rowsrt(t,j) = rowend(t,j) = m. After block column j is fully triangularized, rowsrt(t,j) = colsrt(j) and rowend(t,j) = colend(j), conditions that normally signal that there is "nothing to do" in block column j. (An exception occurs when rowsrt(t,j) = colsrt(j) and rowend(t,j) = colend(j) = m.)

With subproblems specified by Algorithm 3.2 we can now describe the overall factorization procedure.

Algorithm 3.3 (Maximally Tall Block Givens QR Factorization)

Given m, n, m_0, q, the column partitioning $(n_1,\ldots,n_q)$ with $m \geq n$ and $m_0 > n_1$, the following algorithm overwrites $A \in R^{m \times n}$ with upper triangular $R = Q^T A$ where Q is orthogonal.

```
Compute t_f , rowsrt(1:t_f ,1:q) , rowend(1:t_f ,1:q),
    colsrt(1:q), and colend(1:q) using Algorithm 3.2
For t = 1 : t_f
    For j = 1:q
        i_1 = rowsrt(t,j)
        i_2 = rowend(t,j)
        j_1 = colsrt(j)
        j_2 = colend(j)
        if ( i_1 = i_2 = m  or  ( i_1 = j_1 & i_2 = j_2 & j_2 ≠ m ) )
            "Nothing to do."
        else
            Compute:  A(i_1:i_2,j_1:j_2)  = QR .
            Apply: A(i_1:i_2,j_1:n)  :=  Q^T A(i_1:i_2,j_1:n)
        end
    end j
end t
```

4. Implementation

In this section we discuss three issues associated with the implementation of Algorithm 3.3 on the LCAP-I system: how A is arranged in shared memory, how the subproblems are solved, and how block column tasks are mapped onto processors.

The Storage of A

At time step t, the relevant row and column delimiters for the j-th subproblem are i_1 = rowsrt(t,j), i_2 = rowend(t,j), j_1 = colsrt(j), and j_2 = colend(j). Here is what the array processor in charge of this subproblem must accomplish:

1. Read $A(i_1{:}i_2, j_1{:}j_2)$ from shared memory.
2. Compute an orthogonal Q such that $Q^T A(i_1{:}i_2, j_1{:}j_2)$ = R is upper triangular.
3. Write the updated $A(i_1{:}i_2, j_1, j_2)$ back into shared memory.
4. Read $A(i_1{:}i_2, j_1{+}1{:}n)$ from shared memory.
5. Apply Q^T to $A(i_1{:}i_2, j_1{+}1{:}n)$.
6. Write the updated $A(i_1{:}i_2, j_2{+}1{:}n)$ back into shared memory.

We assume that $A(i_1{:}i_2, j_1{:}j_2)$ can fit into local memory but that because of its size, the processing of $A(i_1{:}i_2, j_2{+}1{:}n)$ may have to proceed in "chunks". That is, steps 4-5-6 may have to be repeated with a manageable segment of columns from $A(i_1{:}i_2, j_2{+}1{:}n)$ each time. Note that Q stays in the AP during this process. Because one AP is responsible for applying a given Q, there is no need to pass Q on to another AP.

There is an overhead associated with traffic to and from shared memory. Reads and writes to shared memory are accomplished with a "**move**" command and can only involve contiguous portions of memory. Using **move** to transfer n floating point words takes

$$T(n) = (100 + 8n/44)\ \mu sec$$

Note that the 100 μsec startup degrades the 44mb/sec peak transfer rate. Thus, a vector of length 1000 takes 281 μsec to move for an effective

data transfer rate of 28 mb/sec.

From the standpoint of processing the subproblem at hand, it would be ideal if $A(i_1{:}i_2, j_1{:}n)$ was contiguous in shared memory for then a minimum number of **moves** would be required to carry out steps 1,3,4, and 6 above. For example, to read a contiguous 1000-by-500 submatrix from shared memory would require T(500,000) = .09 sec (≡ 44mb/sec). Unfortunately, storing by blocks in Algorithm 3.3 would impose significant buffer requirements and some tedious data manipulation within each AP. The buffer issue is fairly important because the AP's we used have limited local memory (≈ 600 Kwords).

Because we didn't want additional buffer requirements to limit further the size of "working" memory we chose to store A in column major order. This implies that r **moves** are required to move a submatrix with r columns. Thus, to read a 1000-by-500 submatrix requires 500·T(1000) = .14 sec (≡28 mb/sec). This is actually a typical size for a submatrix move in our algorithm. When the overall implementation is considered, we can easily live with a 28 mb/sec data transfer rate.

Subproblem Solution

The basic computation in Algorithm 3.3 consists of computing a QR factorization and then applying the resulting orthogonal matrix to the "rest of A". The normal "Linpack" way to compute a QR factorization of a matrix $C \in R^{m_0 \times n_0}$ is to use Householder matrices. A Householder matrix is an orthogonal transformation of the form

$$P = I - 2vv^T \qquad v \in R^{m_0}, \ \| v \|_2 = 1 .$$

In the Linpack QR procedure Householders $P_1, \ldots, P_{n_0}$ are generated so that $P_{n_0} \cdots P_1 C = R$ is upper triangular. Note that $Q = P_1 \cdots P_{n_0}$.

We now consider the computation $Q^T B$ where B is some matrix. If Q is represented as a product of Householders, then the resulting algorithm is "rich" in matrix-vector multiplications. This is fine for many architectures. However, to exploit fully the FPS-164/MAX architecture, we need an update algorithm that is rich in matrix-matrix multiplication. We could accomplish this by explicitly forming the product $Q = P_1 \cdots P_{n_0}$

before applying it to B. But this would be very costly since $m_0 >> n_0$ usually. An unacceptably large m_0-by-m_0 buffer would also be required by this approach.

Instead, we have chosen to use the "WY" representation for products of Householder matrices that is developed in Bischof and Van Loan (1985). In this scheme m_0-by-n_0 matrices W and Y are generated such that

$$Q = P_1 \cdots P_{n_0} = I + WY^T$$

The ensuing update $B := Q^TB = (I + WY^T)^TB = B + Y(W^TB)$ is then obtained by a pair of matrix-matrix multiplications:

$$\text{(i)} \quad Z = W^TB$$
$$\text{(ii)} \quad B = B + YZ$$

For (i) we used the "MAX" routine **pdot** that can compute twenty parallel dot products. To initiate the parallel dot product the relevant twenty vectors must be placed in the MAX registers using another MAX routine called **ploadd**. We examine this in some detail so that an appreciation of MAX board computing can be obtained. Assume that W and Y are m_0-by-n_0 and that n_0 (for simplicity) is a multiple of twenty. If B is m_0-by- k then here is how the matrix $Z = W^TB$ is formed:

```
For j = 1:20:n0
    Load W(1:m0 , j:j+19)  in to the max registers using
        pload.
    For i = 1:k
        Compute Z(j:j+19,i:i)  = W(1:m0 , j:j+19)  T B(1:m0,i:i)
        using pdot.
    end i
end j
```

The times required for each **pload** and **pdot** are approximately

ploadd: $L(m_0) = 23 + 58.2*m_0$ (μsec)

pdot: $D(m_0) = 29.7 + .738*m_0$ (μsec)

Thus, $Z = W^TB$ is obtained in $(n_0/20)(L(m_0) + k \cdot D(m_0))$ μsec. Since $Z =$ requires $2m_0n_0k$ flops, a calculation shows that the effective performance in megaflops is approximately given by

$$Mflop(W^TB) = \frac{55}{1 + 40/m_0 + 79/k + 31/m_0k}$$

This expression reveals the penalty for short vectors (small m_0) and for low re-use (small k). Here is a table of some representative $Mflop(W^TB)$ values :

	k = 100	k = 500	k = 1000	k = 5000
$m_0 = 100$	25	35	37	39
$m_0 = 500$	29	44	47	50
$m_0 = 1000$	30	46	49	52
$m_0 = 2000$	30	47	50	53

Table 4.1

We mention that because the MAX registers can handle vectors up to length 2047, the subproblem height parameter m_0 should be chosen so that rowend(t,j) - rowsrt(t,j) ≤ 2047 for all t and j .

We now turn our attention to the rank-n_0 update $B \leftarrow B + YZ$ that makes up the second half of the $B \leftarrow (I + WY^T)^TB$ computation. For this calculation the FPS-164/MAX has a parallel saxpy capability that appears well suited. With two MAX boards it is possible to perform nine saxpys of the form $c_i \leftarrow c_i + s_iy$ in parallel. Note that this is a rank-one update: $C \leftarrow C + ys^T$. Here is how the update of B would proceedusing the parallel saxpy routine **pvsma** and the attending load/unload routines **ploadv** and **punldv** . For simplicity, assume that k is a multiple of 9:

For j = 1:9:k

Use **ploadv** to load $B(1:m_0, j:j+8)$ into the max registers.

For i = $1:n_0$

Use **pvsma** to perform the update

$B(1:m_0, j:j+8) \leftarrow B(1:m_0, j:j+8) + Y(1:m_0, i:i)Z(i:i, j:j+8)$

end i

Use **punidv** to write the updated $B(1:m_0, j:j+8)$ back to memory.

end j

Reasoning as we did to determine Mflop(W^TB), it can be shown that

$$\text{Mflop}(B + YZ) = \frac{24}{1 + 34/m_0 + 70/n_0 + 62/m_0n_0}$$

Note that the re-use factor is now n_0 rather than k . This is unfortunate since in our application we typically have $k > m_0 >> n_0$. If we look at some typical values of Mflop(B + YZ), then this is what we find:

	$n_0 = 20$	$n_0 = 40$	$n_0 = 60$	$n_0 = 80$
$m_0 = 100$	4.9	7.7	9.5	10.8
$m_0 = 500$	5.2	8.5	10.7	12.3
$m_0 = 1000$	5.3	8.6	10.9	12.6
$m_0 = 2000$	5.3	8.6	11.0	12.7

Table 4.2

Thus, **pvsma** is ill-suited for the $B \leftarrow B + YZ$ update when compared to the 23-53 Mflop rates sustained by the **pdot** computation of $Z = W^TB$. For this reason we chose to use a new FPS parallel matrix multiply routine called **pmmul** that can perform the update $B \leftarrow B + YZ$ at rates more consistent with the values in Table 4.1.

Two final comments about subproblem solution. The first concerns the recording of the orthogonal matrix Q. This matrix is the product of Householder matrices. Of course, these Householders are clustered and applied in WY form during Algorithm 3.3. But we can save all the Householder vectors by overwriting each zeroed subcolumn of A by the corresponding Householder vector. In particular, whenever a subcolumn $v \in R^d$ of A is zeroed by a Householder matrix $(I + 2uu^T/u^Tu)$, we store u(2:d) in v(2:d) with the convention u(1) = 1. It is then possible to retrieve Q from the final array A so long as the index arrays rowsrt, rowend, colsrt, and colend are available.

Lastly, we mention that the subproblem QR factorizations in Algorithm 3.3 are typically of matrices that have a band structure. Indeed, it is usually the case that A(rowsrt(t,j):rowend(t,j),colsrt(j):colend(j)) has lower bandwidth rowsrt(t-1,j)-rowsrt(t,j). This fact is exploited when the QR factorization is computed and the resulting WY factors found.

Load Balancing and Scheduling

Suppose Algorithm 3.3 is to be implemented on array processors $AP_1,\ldots,AP_p$. At time step t in Algorithm 3.3 there are q independent tasks to perform. Task (t,j) involves

Factoring: A(rowsrt(t,j):rowend(t,j),colsrt(j):colend(j)) = QR

Computing: Q^TA(rowsrt(t,j):rowend(t,j),colsrt(j):n))

Here, t and j satisfy $1 \le t \le t_f$ and $1 \le j \le q$. If p = q then an immediate load balancing problem arises if each block column has the same width because task (t,j) generally has more matrix to update than task (t,j+1). One way around this difficulty is to make each block column wider than its predecessor. We illustrate this for the case q = 2 with block column widths n_1 and n_2. Assuming a subproblem height of m_0 then approximately $2m_0n_1^2 + 2n_1m_0n_2$ flops are required for task (t,1). On the other hand, $2m_0n_2^2$ flops are required for task (t,2) if we again assume a subproblem height of m_0. These two flop counts are approximately equal if $(n_1/n_2) \approx .62$.

For general q it is possible to work out quotients n_j/n_{j+1} for j = 1:q-1 so that approximate load balancing results for the column partitioning $n_1,\ldots,n_q$. Of course, in practice it would make more sense to base column partitioning guidelines upon benchmarks rather than upon flop counts. We have not pursued this.

Instead we make the block column widths narrow enough so that the number of independent tasks q is significantly larger than the number p of assigned APs. Approximate load balancing is then achieved by assigning AP_k to block columns j = k:p:q. For example, if p = 3 and q = 12, then AP_1 works on block columns 1, 4, 7, and 10, AP_2 is assigned to block columns 2, 5, 8, and 11, while AP_3 is applied to block columns 3, 6, 9, and 12. In a typical time step, each AP will work on 4 subproblems with a greater balance of work than if q = 3. This style of distributing tasks has been widely used in parallel matrix factorization work, see George, Heath, and Liu (1985). A fringe benefit of this approach is that we can choose block column widths to be a multiple of twenty. This allows for efficient exploitation of the 164/MAX architecture that permits twenty parallel dot products. In our examples we used uniform block column widths of twenty and thus $q \approx n/20$.

To actually execute Algorithm 3.3 in parallel on LCAP-1 we implemented a lock-step synchronization scheme using "barriers". The blocking arrays rowend, rowend, colsrt, and colend are determined by the host and then downloaded into the p array processors assigned to the computation. The matrix A is also downloaded into the shared memory through the APs. AP_k then executes the following program:

Algorithm 4.1 (Processor k's Share of Algorithm 3.3)

```
For t = 1:t_f
   For j = k:p:q
      Compute  A(rowsrt(t,j):rowend(t,j),colsrt(j):colend(j))  = QR
      Update  A(rowsrt(t,j):rowend(t,j),colsrt(j):n)
   end j
   Barrier
end t
```

When the barrier is encountered, execution is suspended until all the other AP programs reach their barrier. After this is accomplished the processing of the next time step begins.

Further details about the LCAP-1 system software required by our implementation may be found in Chin and Lorenzo(1986).

5. Some Results and Conclusions

In testing our implementation we ran our codes on random matrices $A \in R^{m \times n}$ with the property that $A(1{:}m,1{:}n-1)e = A(1{:}m,n{:}n)$ where e is the vector of all ones. The correctness of R was then confirmed by checking the equations $R(1{:}n-1,1{:}n-1)e = R(1{:}n-1,n{:}n)$ and $R(n,n) = 0$.

We report on two of the several examples that we solved using the parallel QR code. We do not pretend that our results are conclusive. They merely confirm some natural suspicions and point the way to future research.

The first example indicates that we can get away with our lock step, coarse grained approach if A is large enough and suitably blocked. Here is what we found by using one, two, and three APs to solve an (m,n) = (5000,1000) problem with $m_0 = 1000$, $q = 50$, and $n_1 = \cdots = n_{50} = 20$.

Number of Processors	Time (seconds)	Speed-Up	Effective Mflop
1	606	1.00	17
2	310	1.95	33
3	211	2.87	48

Table 5.1

About 25% of the elapsed time is spent on transmitting submatrices to and from the shared memory. To see roughly where this percent comes from consider the update $B \leftarrow (I + WY^T)^T B$ of a 1000-by-500 submatrix B in shared memory where $W,Y \in R^{1000 \times 20}$. If this update is performed at a

rate of 30 Mflops then approximately 1.3 seconds must be devoted to computation. To transfer B to or from shared memory requires about .14 seconds. Thus, the fraction of time spent on communication is approximately $.18 \approx .28/1.58$.

We next discuss an example where the load balancing isn't quite so nice resulting in a degradation of performance. In the example $m = 5040$, $m_0 = 1040$, $n = 500$, $q = 13$, and $n_1 = \cdots = n_{12} = 40$, $n_{13} = 20$. Three APs were used and thus block column tasks are assigned as follows:

$AP_1 \leftarrow (1,4,7,10,13)$ $AP_2 \leftarrow (2,5,8,11)$ $AP_3 \leftarrow (3,6,9,12)$

Because only five steps are required to process each block column, there are never more than five "active" tasks at any one time step. This makes load balancing a little problematical. The following table indicates the time (in seconds) that each AP spends computing at each timestep.

Time Step	AP_1	AP_2	AP_3
1	3.52	0.00	0.00
2	3.64	3.15	0.00
3	3.64	3.39	2.82
4	3.64	3.42	3.16
5	5.85	3.39	3.16
6	3.10	5.31	4.83
7	2.70	2.82	4.83
8	2.70	2.46	2.58
9	3.88	2.46	2.24
10	2.05	3.42	2.24
11	1.76	1.79	3.01
12	1.76	1.52	1.54
13	1.99	1.52	1.30
14	.62	1.52	1.30
15	.43	1.61	1.30
16	.43	0.00	0.13
17	.43	0.00	0.00

Table 5.1

The time required for the entire computation is 51.2 seconds, the sum of the maximum times in each row of the table. If computation was equally shared at each time step then approximately 38.1 seconds would be required for the complete computation.

The somewhat inefficient use of the APs highlighted by the second example could be rectified in several ways:

1. Choose a smaller m_0. This would have the effect of increasing the number of tasks to be shared at each time step.

2. Vary the block column widths so as to even out the update work.

3. Instead of letting the AP that generates a Q be entirely responsible for its application, share the update.

We have not fully explored these possibilities. Note that the first and third suggestions imply smaller matrix multiplications and thereby reduced I64/MAX performance.

A more promising way to address the load balancing issue would be to incorporate a dynamic scheduling of tasks as is discussed, for example, in George, Heath, and Liu (1985) and Dongarra, Sorenson, and Sameh (1986). One way to do this is to order the tasks (t,j) defined in §4 as follows:

$$(1,1),(1,2),\ldots,(1,q),(2,1),(2,2),\ldots,(2,q),\ldots,(t_f,1),\ldots,(t_f,q)$$

After completing a task each AP would go to this list and "grab" the next available task subject to rules that preserve the integrity of the overall procedure. We will report on this elsewhere.

Acknowledgements

The implementation of the algorithm described in this paper would not have been possible without the expert assistance of Dr. Doug Logan at IBM Kingston and my Ph.D. student Chris Bischof at Cornell.

References

C. Bischof and C. Van Loan (1987), "The WY representation for products of Householder matrices", to appear in SIAM J. Scientific and Statistical Computing.

A.E. Charlesworth and J.L. Gustafson (1986), "Introducing replicated VLSI to supercomputing: the FPS-164/MAX scientific computer", IEEE Computer, March, 10-23.

S.Chin and D.Lorenzo (1986), "Parallel computation on the loosely coupled array of processors: tools and guidelines", Report KGN-25, IBM Kingston, Dept 488, Bldg 963, Kingston, NY 12401.

E.Clementi and D. Logan (1985), "Parallel processing with the loosely coupled array processor system", Report KGN-43, IBM Kingston, Dept 488, Bldg 963, Kingston, NY 12401.

J.Dongarra, A.H. Sameh, and D.C.Sorenson (1986), "Implementation of some concurrent algorithms for matrix factorization", Parallel Computing, 3, pp.25-34.

W.M. Gentleman and H.T. Kung (1982), "Matrix triangularization by systolic arrays", Proc. SPIE Vol. 298, Real Time Signal Processing IV, 19-26.

A. George, M.T.Heath, and J.Liu (1985), "Parallel Cholesky factorization on a multiprocessor", Report ORNL 6124, Math.Sci. Div., Oak Ridge National Laboratory, Oak Ridge, TN.

G.H. Golub and C. Van Loan (1983), Matrix Computations, The Johns Hopkins University Press, Baltimore, Md.

D. Heller and I.C.F. Ipsen (1983), "Systolic networks for orthogonal decompositions", Siam J. Scientific and Stat. Computing, 4, 261-269.

NONPARAMETRIC ESTIMATION FROM QUEUES ARISING IN STAGGERED ENTRY CLINICAL TRIALS

Michael J. Phelan and N.U. Prabhu
Mathematical Sciences Institute, Caldwell Hall
Cornell University, Ithaca, NY 14853

Abstract: In clinical trials with staggered entries and fixed duration of study, patients enter at random epochs and are put on test. The objective is to study survival times from a principal cause A, but factors such as end of study or patient withdrawal make it impossible to observe the survival times (censoring). We consider two different situations. (1) For some patients death may actually be from a cause other than A, say B (competing risks). It is desired to study the survival times associated with both causes A and B. (2) A certain number m (≥ 1) treatments are available and each entering patient is diagnosed and assigned to one of these treatments. The objective is to study the survival times from cause A under these treatments. These problems are formulated in terms of queueing models, for which it is desired to obtain nonparametric estimates of service time distributions. We investigate an infinite server model to study case (1) and an m-station model for case (2). The input in both models is a point process. We observe the system over a finite time-interval $[0,t]$, with t fixed. The data collected consist of the arrival epochs, service times of the customers who arrive during $[0,t]$, with some of the service times partially observed, and (in Model 2) delays experienced by them before service. Our estimators are martingale estimators, for which we establish consistency and weak convergence (as $t \to \infty$) of the normalized difference to a Gaussian process. We present the results for Model 1. Work on Model 2 is in progress.

Keywords: Clinical trials, censoring, survival times, queues, counting processes, martingales, product-limit estimators.

1. Introduction

The estimation problems that we consider arise in the context of clinical trials with staggered entries and fixed duration of study. Patients enter the study at random epochs and are put on test (for example, the patient is treated by a drug therapy). Typically, the objective is to study a patients' time of death (survial time) from cause A, such as cancer or AIDS. Factors such as end of study or patient withdrawal make it impossible to observe patients' time of death (censoring). We consider two different situations.

(1) Entering patients are put on test immediately. For some patients, death may actually be from a cause other than A, such as toxicity. Denoting this second cause as B, we say that A and B are competing risks. In such situations it may be important to study the hazard functions associated with both causes A and B, rather than A alone, as is usually done. Thus the observed survival time of each patients is the shorter of the survival times from A and B.

(2) On some occasions there are a certain number m (≥ 1) of treatments available and each entering patient is diagnosed immediately and assigned to one of these treatments depending on factors such as the patient's background and state of health. Furthermore, limited availability of the facilities used in the therapy may cause delay between the patient's time of entry and the actual time he is put on test. The objective is to study the survival times associated with the m treatments.

Situations described above lead to the following queueing models.

Model 1.

(1) Let $\tau_0, \tau_1, \tau_2, \ldots$ denote the arrival epochs of the successive customers. We assume that the point process $\tau = \{\tau_n, n \geq 0\}$ satisfies the following conditions:

(i) $\tau_0 = 0$, (ii) $\tau_n < \infty$ $(n \geq 1)$, and (iii) $\tau_{n+1} > \tau_n$, $\tau_n \uparrow \infty$. Let $N = \{N_t, t \geq 0\}$ denote the counting process generated by the process τ. Then N_t gives the number of arrivals during a time-interval $(0,t]$. Here N has right-continuous sample paths with left limits, jumps of size 1 and $N_0 = 0$. Thus the input into the queueing system is described equivalently by N or τ.

(2) Each customer brings two demands for service. Let (X_n, Y_n) denote the service times of the two demands of the nth customer

$(n \geq 1)$. We assume that $\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$ are independent sequences of mutually independent random variables with common distributions F_1 and F_2, respectively, both concentrated on $(0,\infty)$. We also assume that these service times are independent of the input process.

(3) There are an infinite number of servers, so that there is no waiting line. However, each server meets only the demand that needs the shorter service time for the customer served.

Our objective is to estimate the distributions F_1 and F_2 of the service times of the two demands. For this purpose we observe the system over a time-interval $(0,t]$, with t fixed. The data consist of the arrival epochs and the service times actually received by the first N_t customers, but some of these service times may only be partially observed; namely, those customers with $\tau_n + \min(X_n,Y_n) > t$, for whom we only know that the service time $\min(X_n,Y_n) > t-\tau_n$.

Our estimators for F_1 and F_2 are martingale estimators. The martingale property leads to proofs of their consistency and the weak convergence (as $t \to \infty$) of the normalized differences to a Gaussian process. These results are stated in section 3. Details are given elsewhere (Phelan and Prabhu (1987)). We make only a mild assumption concerning the rate at which N_t goes to infinity. Thus we avoid conditions such as $N_t/t \to \text{constant} > 0$ in some sense, as is often imposed in situations involving random sample sizes. This shows the advantage of our approach based on martingale properties.

Model 2.

(1) Let $\tau_0,\tau_1,\tau_2,\ldots$ denote the arrival epochs of the successive customers, where the point process $\tau = \{\tau_n, n \geq 0\}$ is as in Model 1. We associate with τ_n a random variable Z_n taking values in $E = \{1,2,\ldots,m\}$ in such a way that the marks $Z = \{Z_n, n \geq 0\}$ may depend in an arbitrary manner on the process τ but may also involve some other randomization. Thus the input into this queueing system is the marked point process $\{(\tau_n,Z_n), n \geq 0\}$.

(2) There are m stations $(1 \leq m < \infty)$; each station being a finite-server queueing system. The service times of customers served at the ith station have a distribution F_i concentrated on $(0,\infty)$. Here the distributions $F_1,F_2,\ldots,F_m$ are all distinct. The customer arriving at the epoch τ_n has a service time X_n having distribution F_i whenever $Z_n = i$ $(i \in E, n \geq 0)$. It is assumed that the X_n are mutually independent, and moreover, they are conditionally independent of the τ_n, given the Z_n. Thus the event $\{Z_n = i\}$ indicates that the customer arriving at τ_n is assigned to the ith station.

(3) The queue discipline at each station is first come, first served.

Our objective for this model is to estimate the distributions $F_1, F_2, \ldots, F_m$ of the service times at the m stations. The observation scheme is exactly as in Model 1, but in addition we obtain for each arrival a record of any delay experienced before service, and which of the stations serves this customer. Work on this model is in progress.

In the standard analysis of staggered entry clinical trials it is assumed that the epochs of entry $\{E_n\}$ of the patients are mutually independent random variables and the duration of study is fixed. In order to develop an asymptotic theory it is then assumed that the accrual rate of patient entry increases over this interval. In the terminology of this paper, $\{\tau_n\}$ are the order statistics generated by $\{E_n\}$, so that the input of patients constitutes a special type of a point process. This input model is described by Jennison and Turnbull (1985). The possibility of the patient accrual rate increasing indefinitely over a fixed time-interval can arise in some situations (such as in the explosive pure birth process). However, in other situations accruals increase by virtue of the increasing length of study, which is our approach to the asymptotic theory.

2. The estimators in Model 1

The data described in section 1 can be conveniently summarized as follows. For each $n \geq 1$ define

$$C_n^t = \max(0, t-\tau_n), \quad W_n^t = \min(X_n, Y_n, C_n^t)$$

$$\delta_n^t = 1(C_n^t \geq \min(X_n, Y_n)), \quad \eta_n = 1(X_n \leq Y_n).$$

Here $\eta_n = 1$ iff the first demand of the nth customer is met. In the terminology of survival analysis $\min(X_n, Y_n)$ is the survival time induced by two competing risks, and C_n^t is called a random right-cenoring time due to end of study in staggered entry clinical trials. Thus $1-\delta_n^t$ is the indicator of censoring and W_n^t is the observed randomly right-censored survival time. In the present context W_n^t is the observed service time, $\delta_n^t = 1$ iff the nth customer has arrived and completed his service before time t, and $\delta_n^t \eta_n = 1$ ($\delta_n^t(1-\eta_n) = 1$) iff the server meets this customer's first (second)

demand, so that his service time is $X_n(Y_n)$. Our observation scheme yields the data

$$(2.1) \qquad \{(W_n^t, \delta_n^t, \eta_n),\ n = 1,2,\ldots,N_t\}$$

from which we seek to estimate the distributions F_1 and F_2. The estimators are actually based on the statistics

$$(2.2) \quad \bar{N}(i,t) = \{\bar{N}_s(i,t), s \geq 0\}\ (i = 1,2);\quad \bar{Y}(t) = \{\bar{Y}_s(t),\ s \geq 0\},$$

where

$$(2.2a) \qquad \bar{N}_s(1,t) = \sum_{n=1}^{N^t} 1(W_n^t \leq s,\ \delta_n^t \eta_n = 1)$$

$$(2.2b) \qquad \bar{N}_s(2,t) = \sum_{n=1}^{N^t} 1(W_n^t \leq s,\ \delta_n^t (1-\eta_n) = 1)$$

$$(2.3) \qquad \bar{Y}_s(t) = \sum_{n=1}^{N^t} 1(W_n^t \geq s).$$

Here $\bar{N}_s(i,t)$ is the number of customers whose service times are less than s among those who arrive and complete their service before time t, and whose ith demand is met $(i = 1,2)$. Also, $\bar{Y}_s(t)$ is the number of arrivals in $(0,t]$ whose service times (complete or partial) exceed s.

Our estimation procedure yields estimators of F_1 and F_2, as well as their associated cumulative conditional rate functions b_1 and b_2, defined by

$$(2.4) \qquad b_i(t) = \int_{(0,t]} [1 - F_i(s-)]^{-1} dF_i(s) \quad (i = 1,2,\ t \geq 0).$$

These estimators are provided by the processes $B^t(i) = \{B_s^t(i),\ s \geq 0\}$, and $\hat{F}_i(t) = \{\hat{F}_s^t(i),\ s \geq 0\}$ $(i = 1,2)$ defined by

$$(2.5) \quad B_s^t(i) = \sum_{u \leq s} \bar{Y}_u^{-1}(t) \Delta \bar{N}_u(i,t) = \int_0^s \bar{Y}_u^{-1}(t)\, d\bar{N}_u(i,t) \quad (i = 1,2)$$

(where each term in the sum is interpreted as zero if both factors are zero), and

$$(2.6) \qquad \hat{F}_s^t(i) = 1 - \prod_{u \le s} (1 - \Delta B_u^t(i)) \quad (i = 1,2),$$

where

$$\Delta \bar{N}_u(i,t) = \bar{N}_u(i,t) - \bar{N}_{u-}(i,t)$$

$$\Delta B_u^t(i) = B_u^t(i) - B_{u-}^t(i) \quad (i = 1,2, \ 0 \le u \le s).$$

For $i = 1,2$ in (2.5), each term in the sum is the proportion of completed service times of ith demand equal to u (i.e. $\Delta \bar{N}_u(i,t)$) among those service times (complete or partial) which exceed u (i.e. $\bar{Y}_u(t)$) from among those customers who arrive during the time-interval $(0,t]$. Thus $B_s^t(i)$ estimates the cumulative conditional service rate for the ith demand. The expression (2.6) is an estimator of the product integral of (2.4), which uniquely determines the distributions F_1 and F_2 from b_1 and b_2, respectively. The estimators $\hat{F}_i^t$ are called product-limit estimators, where by virtue of our observation scheme they are defined from a random number of partially observed service times (cf. Gill (1980), who studies this type of estimator from a fixed number of censored survival times).

3. Asymptotic properties of the estimators in Model 1

We state the asymptotic properties of the martingale estimators (2.5) and (2.6). We begin with the problem of consistency. For $s > 0$ define $y(s) = [1 - F_1(s-)][1 - F_2(s-)]$ and $\theta = \sup\{s: y(s) > 0\}$. Suppose $y(\theta) = 0$.

Theorem 3.1 (Consistency). Let θ and the function y be defined above, and suppose that for $s \in [0,\theta)$, $(N_t - N_{t-s})/N_t \xrightarrow{p} 0$ as $t \to \infty$. Then for $u \in [0,\theta)$ we have, as $t \to \infty$,

$$(3.1) \qquad \sup_{s \in [0,u]} |\hat{F}_s^t(i) - F_i(s)| \xrightarrow{p} 0 \quad (i = 1,2),$$

and

$$(3.2) \qquad \sup_{s \in [0,u]} |B_s^t(i) - b_i(s)| \xrightarrow{p} 0 \quad (i = 1,2). \qquad \square$$

According to Theorem 3.1, for large t, $\hat{F}^t(i)$ and $B^t(i)$ are uniformly close estimates of F_i and b_i $(i = 1,2)$, respectively, on subintervals in the intersection of the support of F_1 and F_2. It turns out that, in addition, the normalized differences in (3.1) and (3.2) converge weakly to Gaussian processes. For this purpose let $D(\theta)$ denote the space of right-continuous functions defined on $[0,\theta)$ and having left-limits. Also, let $Z^i = \{Z^i_s,\ s \in [0,\theta)\}$ and $W^i = \{W^i_s,\ s \in [0,\theta)\}$ $(i = 1,2)$ denote mean zero Gaussian processes of independent increments and covariance functions

$$\langle W^i, W^j\rangle(s) = \begin{cases} \int_0^s (1-\Delta b_i(u))(y(u))^{-1} db_i(u), & s \in [0,\theta) \quad \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \tag{3.3}$$

and

$$\langle Z^i, Z^j\rangle(s) = \begin{cases} \int_0^s (1-\Delta b_i(s))^{-2} d\langle W^i, W^i\rangle(s), & s \in [0,\theta) \text{ for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \tag{3.4}$$

Note that Z^1 is independent of Z^2 and that W^1 is independent of W^2. We have the following theorem.

Theorem 3.2 (Asymptotic Normality). Suppose the condition of Theorem 3.1 holds. Consider the normalized processes

$$U_i^t = N_t^{1/2}(\hat{F}^t(i)-F_i)/(1-F_i),\ V_i^t = N_t^{1/2}(B^t(i)-b_i) \quad (i = 1,2). \tag{3.5}$$

Then as $t \to \infty$

$$(U_1^t, U_2^t) \xrightarrow{\mathcal{D}} (Z^1, Z^2) \tag{3.6}$$

and

$$(V_1^t, V_2^t) \xrightarrow{\mathcal{D}} (W^1, W^2) \tag{3.7}$$

in $D(\theta) \times D(\theta)$ endowed with the Skorohod topology. □

References

Gill, R.D. (1980): Censoring and Stochastic Integrals. Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

Jennison, C. and Turnbull, B.W. (1985): Repeated confidence intervals for the median survival time. Biometrika 72 (3), 619-625.

Phelan, M.J. and Prabhu, N.U. (1987): Estimation from an infinite server queueing system with two demands. Mathematical Sciences Institute, Cornell University, Technical Report 87-40.

A Class of Diffusion-Type Probability Distributions

Siegfried H. Lehnigk

Research Directorate
Research, Development, and Engineering Center
U.S. Army Missile Command
Redstone Arsenal, AL 35898-5248

1. The Density Function

Associated with the Markov diffusion equation

$$\frac{1}{2}\left[\alpha_2(x)z\right]_{xx} - \left[\alpha_1(x)z\right]_x - z_t = 0 \, , \; z = z(x,t) \, , \tag{1.1a}$$

with diffusion and drift coefficients

$$\frac{1}{2}\alpha_2(x) = \alpha x^{2-\beta} \, ,$$

$$\alpha_1(x) = \alpha(2-\beta-p)x^{1-\beta} - \tau x \, , \tag{1.1b}$$

and parameters $\alpha > 0$, $\beta > 0$, $p < 1$, $\tau \in R$, is the class of source density functions

$$f(x) = \beta b^{-1} \xi^{-(p-\beta+1)/2} \zeta^{(p+\beta-1)/2} I_q(2(\xi\zeta)^{\beta/2}) \exp -(\xi^\beta + \zeta^\beta), \tag{1.2}$$

$x > 0$, $\xi = xb^{-1}$, $\zeta = yb^{-1} \exp -\tau t_0$, $q = -1 + (1-p)\beta^{-1}$, $I_q(r) =$ modified Bessel function of the first kind of order q. From a

statistical point of view, the parameters in (1.2) are $b > 0$ scale, $p < 1$ initial shape, $\beta > 0$ terminal shape, and $y \geq 0$ source. The restrictions on p and β imply $q > -1$.

The designation of y as a source parameter is based on the fact [1], [2], that the function $f(x)$ given in (1.2) has been derived from the delta function initial condition solution (source solution) of (1.1) with the delta function applied at $t = 0$ and $y > 0$. As $y \downarrow 0$, $f(x)$ reduces to the well-known hyper-Gamma density [2], [3].

2. The Likelihood Function

Application of the source density (1.2) in statistical practice requires a method to determine the numerical values of the components of the parameter vector $P = (b, q, \beta, z)$, $z = y \exp -\tau t_0$, relative to given statistical data (x_ν, f_ν) $(\nu = 1, \ldots, n)$ which represent observations x_ν together with their relative frequencies f_ν. To this end, the likelihood function associated with (1.2) will be established.

In general terms, let $f(x;P)$, $x > 0$, be a density function depending on a parameter vector P. Let X_ν $(\nu = 1, \ldots, N)$ be random sample values of a random variable X which is assumed to be distributed according to $f(x;P)$. The associated likelihood function is

$$L(P) = \prod_{\nu=1}^{N} f(X_\nu;P).$$

If the set $\{X_\nu\}$ contains the distinct elements x_ν ($\nu = 1, \ldots n \le N$) with relative frequencies f_ν, the log-likelihood function is

$$\phi(P) = N^{-1} \log L(P) = \sum_{\nu=1}^{n} f_\nu \log f(x_\nu; P).$$

For the density (1.2) the function $\phi(P)$ takes the form

$$\phi(P) = \log \beta - \beta \log b - \frac{1}{2}(2-2\beta-\beta q)C - \frac{1}{2}\beta q \log z$$

$$- b^{-\beta}B - b^{-\beta}z^{\beta} + \sum_{\nu=1}^{n} f_\nu \log l_q(r_\nu) ,$$

$$r_\nu = 2 z^{\beta/2} b^{-\beta} \exp -(\beta\rho_\nu/2) , \quad \rho_\nu = \log x_\nu ,$$

$$B = B(\beta) = \sum_{\nu=1}^{n} f_\nu \exp \beta\rho_\nu , \quad C = \sum_{\nu=1}^{n} f_\nu \rho_\nu = \text{const.}$$

The objective is to maximize $\phi(P)$ in the interior of the parameter domain $b > 0$, $q > -1$, $\beta > 0$, $z > 0$. The equations $\partial\phi/\partial b = 0$, $\partial\phi/\partial q = 0$, $\partial\phi/\partial\beta = 0$, $\partial\phi/\partial z = 0$ show that the scale parameter can be eliminated. In fact

$$b^{\beta} = (1+q)^{-1}(B - z^{\beta}).$$

Therefore, it is sufficient to maximize the function $\phi^*(q,\beta,z)$ which results from ϕ upon elimination of b .

Numerical solution attempts on the equations $\partial\phi^*/\partial q = 0$, $\partial\phi^*/\partial\beta = 0$, $\partial\phi^*/\partial z = 0$ by means of derivative-based methods have not been satisfactory. Direct optimization techniques are under investigation. Results will be reported upon when they become available.

References

[1] Lehnigk, S. H.: Initial condition solutions of the generalized Feller equation, J. Appl. Math. Phys. (ZAMP), 29 (1978), 273-294.

[2] Lehnigk, S. H.: On a class of probability distributions, accepted, Math. Meth. in the Appl. Sci.

[3] Lehnigk, S. H.: Maximum-likelihood estimation of the parameters of a four-parameter class of probability distributions, accepted. Proc. Edinburgh Math. Soc.

COLUMN MOVEMENT MODEL USED TO SUPPORT AMM

George B. McKinley
U.S. Army Engineer Waterways Experiment Station
Geotechnical Laboratory
Vicksburg, MS 39180-0631

ABSTRACT. For many years mobility maps have been created utilizing the Army Mobility Model (AMM). These maps show the maximum speed which a vehicle can attain in off-road terrain and on-road networks. These maps are useful in comparing the performance of vehicles or as an aid in route selection. Three additional computer models have been developed to support the AMM by predicting the performance of military vehicles over digital terrain units along a specified route. These three models are an Acceleration Model, a Traverse Model, and a Column Movement Model. The Acceleration Model produces a time versus speed curve for a vehicle along a specified route across a terrain unit. The Traverse Model uses the Acceleration Model as a building block and predicts a vehicle's performance along a specified route over a series of terrain units. The Column Movement Model uses the Traverse Model as a building block for predicting performance of a column of vehicles along a specified route. The Column Movement Model maintains vehicle spacing within the column in accordance with military doctrine.

I. TERRAIN DATA SELECTION. The terrain data required by the models may be acquired using one of three basic methods. The first method consists of surveying a traverse to determine slopes and curvatures. The courses are concurrently subdivided into a number of segments (terrain units), each of which should be nominally uniform with respect to values pertinent to mobility including surface roughness (rms elevation), slope, driver recognition distance, radius of curvature, soil type, and soil strength. From these measurements a digital terrain data base is developed for use with AMM (Nuttall, Green, Dean, and Gray 1985).

A second method consists of using the Waterways Experiment Station's (WES) Digital Road Net Data bases which exist for a selected few 1:50,000 scale map sheets in the Federal Republic of Germany. Software has been developed at WES to select "best paths" on this network based on either time or distance. This path selection is accomplished by use of a blind bidirectional search.

The third method involves the manual selection of a path through an areal map. This selection may consist of visually analyzing a speed prediction map and choosing sufficient Universal Transverse Mercator (UTM) coordinates to define the desired path. Software developed at WES will then create the proper terrain file by either assuming linear movement between the specified UTMs or by selecting the "best path" by use of the blind bidirectional search.

II. AMM. The AMM is a comprehensive analytical model designed to evaluate objectively the on- and off-road mobility of vehicles by means of digital computer simulation (Nuttall, Dugoff, and Rula 1974). First developed in 1971, the AMM is the Waterways Experiment Station's living mobility model and is modified as required, based on improved mobility algorithms and customer needs. The AMM is organized as illustrated by the general flow diagram in Figure 1.

For its data base, the AMM requires quantitative input descriptions of terrain, vehicle, and driver attributes as shown in Table 1. In somewhat more detail, Table 2 describes how terrain data are portrayed in the AMM. Driver attributes in AMM characterize the driver according to his ability to perceive and react to visual stimuli affecting his behavior as a vehicle controller and his limiting tolerances to shock and vibration. The influence on vehicle speed of these latter driver attributes is taken into account by the vehicle ride dynamics module of the AMM (Figure 1).

In following the general flow diagram in Figure 1, raw input driver and terrain data first are adjusted to account for the influence of appropriate "scenario" factors, such as season and weather. Terrain data in the AMM are used to describe small patches or segments, each one of which is defined by a set of values of terrain factor classes (Table 2) that is different in at least one terrain factor class value from the sets of class values of all contiguous patches.

As shown in Figure 1, input vehicle, driver, and terrain data are modified by the vehicle data preprocessor, the vehicle ride dynamics module, and the terrain data preprocessor. The vehicle data preprocessor is a part of the main program of AMM, and is used once at the beginning of an AMM run to compute vehicle power train and soil-running gear characteristics that are repeatedly used in making subsequent vehicle mobility predictions for individual areal patches or road segments.

The ride dynamics module operates on a stand-alone basis, and in effect serves as a major preprocessor of input vehicle, driver, and terrain data (Murphy and Ahlvin 1976). Data similar to the output of the ride dynamics module may also be obtained by field testing of vehicles on ride dynamics and obstacle test courses. From input vehicle, driver, and obstacle height data, the ride dynamics module computes vehicle speed values at which a vertical acceleration of 2.5-g's is experienced at the driver's station. The ride dynamics module also computes, as a function of surface microroughness (expressed as the root-mean square elevation (rms) of the effective profile), speed values corresponding to limits of driver tolerance to random vibrations. This tolerance is defined in terms of the vibrational power absorbed by a person at a specific location in the vehicle, often taken as a constant tolerance limit of 6-watts (Lins 1972). Currently, data preprocessing in the ride dynamics module reduces dynamics-based predictions in areal patches and road segments to a rapid table lookup process.

The terrain data preprocessor converts the ranges of values of terrain factor classes stored in the terrain data base to the engineering values used by subsequent AMM modules. Ordinarily, the value assigned for a given terrain or road factor is the best-estimate value of that factor's class range. This preprocessor also accounts for "scenario factors" by adjusting or selecting among stored terrain or road factor values to reflect the influence of variations in season, weather, day or night operation, and other factors.

Continuing in Figure 1, data acted upon by the three preprocessors next are used in the vehicle performance prediction modules that are the heart of the AMM--the areal patch module and the on-road segment module. The general flow of the areal patch module is shown in Figure 2. Input to this module from the vehicle data preprocessor includes the relation between vehicle speed

and tractive force for the vehicle on a smooth, level, firm surface, and the minimum soil strength that the vehicle requires to maintain headway on level, weak soils. Using these data and appropriate data from the terrain data preprocessor, the areal patch module checks for vehicle-obstacle interferences (hangups), determines the total tractive force required to overcome terrain impediments, and computes vehicle speed limited by the total motion-resisting forces. This calculation involves interaction of the soils, slope, obstacle traction, obstacle override, and vegetation impact and override submodels. NOGO is called when vehicle hangup is predicted and when vehicle traction and override force are computed to be insufficient to overcome resistances to motion.

Next, the areal patch module selects the minimum speed among speed limited by the resisting forces; ride-limited speed (obtained from the vehicle ride dynamics module output data array); and visibility-limited speed (from the visibility submodel). This speed is then modified to account for acceleration and deceleration between discrete obstacles and maneuvering to avoid vegetation and other obstacles. This procedure produces a maximum vehicle speed predicted for a particular terrain unit and a particular vegetation override/avoidance option. The procedure is repeated for the vehicle operating up slope, down slope, and across slope, producing three speed predictions.

Compared to areal (off-road) terrain, on-road terrain includes considerably fewer factors that affect vehicle performance . Still, the on-road module (the third AMM performance prediction module) has a computational structure similar to that of the areal module. For the particular road surface material of interest, values of tractive and rolling resistance coefficients are obtained for the given wheeled or tracked vehicle operating straightline and level at maximum speed. Separate speeds are then computed as limited by available traction and countervailing resistances (rolling, grade, and curvature); ride dynamics (absorbed power); visibility and braking; tire load, inflation, and construction; and road curvature. The least of these five speeds is assigned as the maximum for the on-road segment scrutinized. Scenario options and combinatorial procedures to predict vehicle speed are exercised in the on-road module similarly to the method previously described for the areal patch module.

III. ACCELERATION MODEL. The Acceleration Model predicts speed versus time relationships for a vehicle accelerating on a defined surface (road or areal). The vehicle's acceleration is modeled using tractive force versus speed data obtained from the AMM which has been modified to account for slippage of the vehicle's running gear in the soil. The vehicle accelerates using the amount of tractive force available beyond that which is required to overcome the sum of the resisting forces (usually only motion resistance, since acceleration tests are normally run on terrain with no slope or vegetation). The time and distance for acceleration are calculated for each segment of the curve. Two different methods are used to calculate these values. If acceleration is known to not be constant (i.e., the forces that serve as endpoints of the current line segment are unequal) then acceleration is modeled as if it were linear between the two speeds. This is accomplished in two steps. First the time to accelerate between the two speeds is calculated as follows:

```
DELTAT = VMR/A * LOG((A*VX+B)/(A*V1+B))
```

where

```
DELTAT = time necessary to complete acceleration step (seconds)
VMR    = vehicle's mass modified for inertia (slugs)
A      = slope of current line segment of curve
V1     = velocity at start of acceleration step (ft/s)
VX     = velocity at end of acceleration step  (ft/s)
B      = y intercept of current line segment of curve
```

Once the time has been calculated the distance which will be covered during the acceleration step may be calculated as follows:

```
DELTAX = VMR * (A*V1+B)/(A**2)*(EXP(A/VMR*DELTAT)-1.)-B*DELTAT/A)
```

where

```
DELTAX = distance covered during acceleration step (ft)
VMR    = vehicle's mass modified for inertia (slugs)
A      = slope of current line segment of curve
V1     = velocity at start of acceleration step (ft/s)
B      = y intercept of current line segment of curve
DELTAT = time necessary to complete acceleration step (seconds)
```

In the case of constant acceleration, the acceleration is calculated using F=MA and time and distance are calculated using the equations of motion with constant acceleration. In both cases the vehicle's mass is modified by a factor which simulates the inertial mass of the rotating parts which must be accelerated when the entire vehicle is accelerated. The vehicle's mass is modified as follows:

```
VMR = VM * (RMF1+RMF2*(FAVG**2))
```

where

```
VMR    = vehicle's mass modified for inertia (slugs)
VM     = vehicle's unmodified mass (slugs)
RMF1   = 1.14 if there is a tracked assembly on the vehicle
       = 1.06 otherwise
RMF2   = _0.002*(IDIESL*CID)**1.68 /(NCYL*GCW)*TCOR * _ RR/ETA/QMAX **2
```

where

```
IDIESL = 3 if engine is turbine
       = 2 if engine is a two cycle diesel
       = 1 otherwise
CID    = engine displacement in cubic feet (rated horsepower for
         turbine)
NCYL   = number of cylinders (1 is used for turbine)
GCW    = gross combined weight of vehicle
RR     = rolling radius (ft) if wheeled assembly or
         sprocket pitch radius (ft) if it includes a tracked assembly
ETA    = 0.7 if there is a tracked assembly present
       = 0.9 otherwise
QMAX   = maximum engine torque (ft-lbs)
FAVG   = tractive effort at center of acceleration step (ft.-lbs.)
TCOR   = 0.125 if engine is a turbine
       = 1 otherwise
```

These calculations are performed for each segment of the tractive force versus speed curve until the vehicle's maximum predicted speed is attained. Example outputs from the Acceleration Model are shown in Figures 3 through 5.

IV. TRAVERSE MODEL. The WES Traverse Model predicts the time required by a defined vehicle to cross a series of terrain units (AMM road or areal format). The vehicle is first run over the digital terrain using the AMM, thus computing all the values necessary for predicting the vehicle's performance over each terrain unit.

The traverse begins with the vehicle at the start of the first terrain unit at zero velocity. When the vehicle first accelerates and upon entering any other terrain unit, the model finds the corresponding tractive force for the vehicle's current velocity. If this tractive force is found equal to the total of the resisting forces in the current terrain unit then the vehicle will not accelerate. If the vehicle is found to accelerate, then the time and distance for acceleration are calculated using the same algorithms utilized by the Acceleration Model.

Each terrain unit has two speeds associated with it. One speed is the predicted speed, which is the maximum speed which may be reached by acceleration from a lower speed in that terrain unit. The other speed is the maximum speed at which a vehicle may enter the terrain unit. The limit is the lowest speed chosen by the AMM from among the ride, visibility, and curvature (when on-road) limited speeds. The only stipulation for a vehicle's entering speed is that it be less than or equal to the limiting speed. In the case of a soil-strength limited terrain unit a vehicle is allowed to enter at a higher speed than that predicted maximum, but the speed must still be less than or equal to the limited speed. In this situation the vehicle's deceleration will be modeled by moving backwards along the tractive force versus speed curve.

The vehicle's speed at the end of each acceleration step is compared to the limited speed of the next terrain unit. When the vehicle's speed becomes greater than that limit, the distance required to brake from the current speed to that limit is computed. This braking is modeled by allowing the application of the maximum braking force available for that vehicle on the current terrain. The equation $F = MA$ is used to compute this constant deceleration. If the sum of the distance used for acceleration and that required for braking becomes greater than the length of the current terrain unit, then the intersection of the current acceleration step and the braking line is computed. From the time and distance used for both acceleration and braking, an average velocity for the terrain unit can be calculated. If the vehicle reaches the predicted speed for the terrain unit then the time and distance at that speed will also be used to calculate the average speed. If the application of brakes were ever necessary over an entire terrain unit plus portions of a previous unit, the model would revert back to that previous terrain unit and take proper action to correct the exiting speed of that unit to allow for proper braking in the current unit.

The exiting speed of a terrain unit is used as the entering speed for the following terrain unit. The vehicle's time in each terrain unit and the length of the unit are used to compute an average speed for that unit along with an average speed for the distance up to and including that unit.

V. COLUMN MOVEMENT MODEL. The WES Column Movement Model computes the total time required for a selected group of vehicles to traverse a series of terrain units. The vehicles, which constitute the column must follow one of three sets of basic march orders. The first column type is an infiltration in which each vehicle moves at its best speed over the entire route. Vehicles travel together in formation, but vehicles are allowed to pass each other when possible. The vehicles leave the staging area at random intervals of 1 to 10 minutes in duration.

The second column to be modeled is the open column. In the open column vehicles travel in single file over the entire route. The vehicles must maintain a spacing of between 50 and 100 meters. Each vehicle will start 20 seconds after the previous vehicle if this will allow for proper vehicle spacing to be maintained. If the previous vehicle has reached the maximum spacing limit in less than 20 seconds then the next vehicle is allowed to start.

The third type of column to be modeled is the closed column. The closed column is identical to the open column, except that the vehicle spacing must remain between 10 and 50 meters. The optimum start interval is changed to 9 seconds for the closed column and is used as in the open column.

Each vehicle's acceleration and braking are modeled in a manner similar to the acceleration and braking modeled in the traverse model. The major difference is that each vehicle's progress is monitored at a user specified time interval. A small interval (5 seconds or less) is preferred, since it should yield more accurate modeling of vehicle interaction.

Each time interval may be evaluated twice. First each vehicle will traverse the terrain, obeying terrain speed limits, until the time interval is over. Each vehicle's entering speed for each terrain unit and time spent in each terrain unit are saved for possible later modification.

Next the position of each vehicle is checked to insure that the column's unity is maintained. If distances between vehicles are too large or too small, certain vehicles are required to proceed at a slower pace over the time frame.

VI. CONCLUSIONS. The above describes three programs serving as extensions to the AMM, which predict vehicle mobility when a path consisting of known terrain units is specified as input. The Acceleration Model predicts time, speed, and distance relationships for a vehicle over a single terrain unit. The Traverse Model accurately predicts vehicle performance over a series of terrain units. The Column Movement Model adequately represents the movement of groups of vehicles over a sequence of road and areal terrain units in which their interaction with both the terrain and each other's position are modeled. Future plans include applying the methodology utilized by the Column Movement Model to additional unit formations of varying size and composition.

REFERENCES

Lins, W. F. 1972. "Human Vibration Response Measurement," Technical Report 1151, U. S. Army Tank-Automotive Command, Warren, Mich.

Murphy, N. R., Jr., and Ahlvin, R. B. 1976. "AMC-74 Vehicle Dynamics Module," Technical Report M-76-1, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss.

Nuttall, C. J., Jr., Dugoff, H. G., and Rula, A. A. 1974. "Computer Model for the Comprehensive Evaluation of Cross-Country Vehicle Mobility," Society of Automotive Engineers Paper No. 740426, Earthmoving Industry Conference, Peoria, Ill., 23-24 Apr 1974.

Nuttall, C. J., Jr., Green, C. E., Dean, T. C., and Gray, M. W. 1985. "Severity Ranking of Courses Used for High-Mobility Multipurpose Wheeled Vehicle (HMMWV) Tests at Camp Roberts, Fort Hunter Liggett, and Aberdeen Proving Ground," Technical Report GL-85-13, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss.

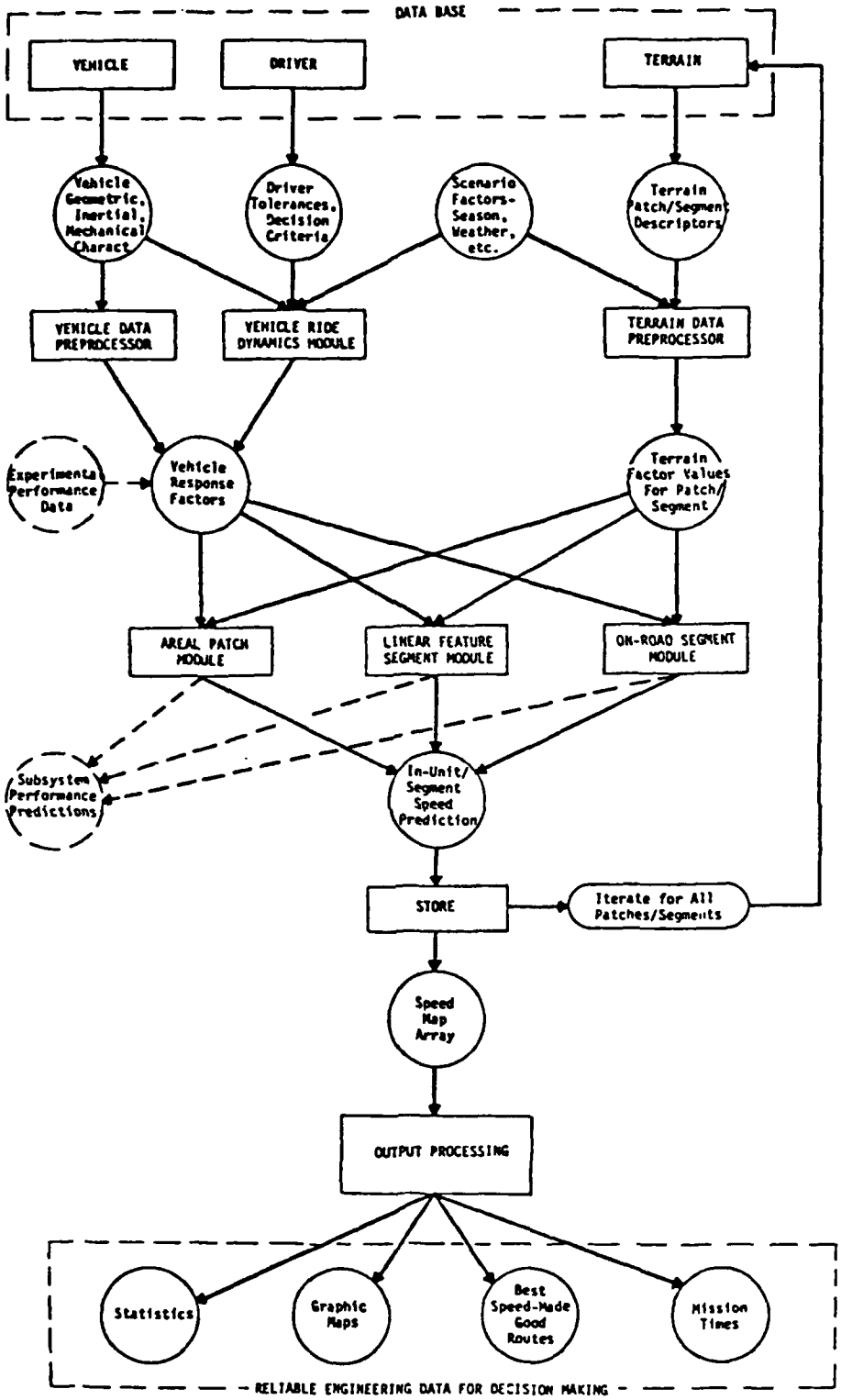


Figure 1. AMM general flow diagram

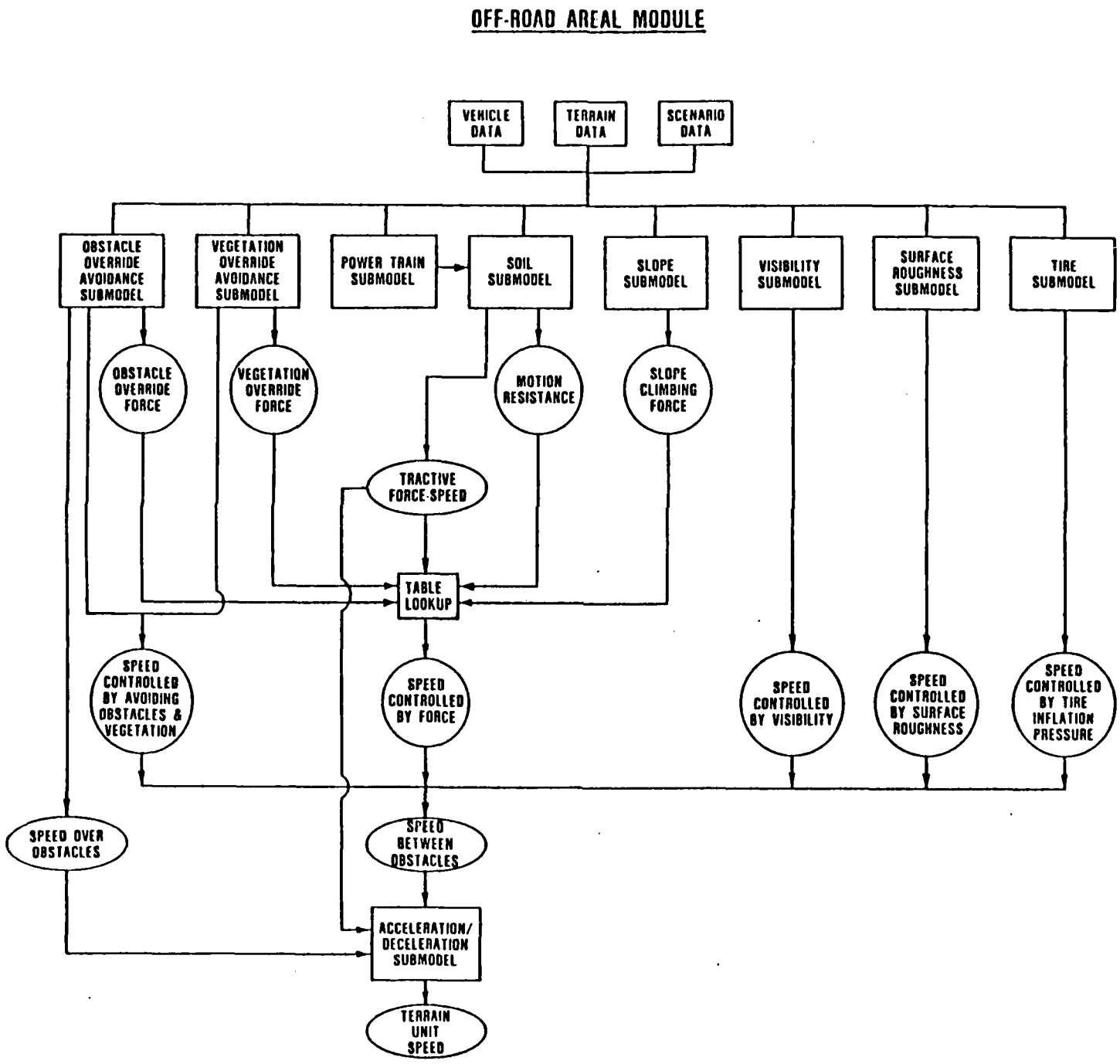


Figure 2. General flow of AMM areal patch module

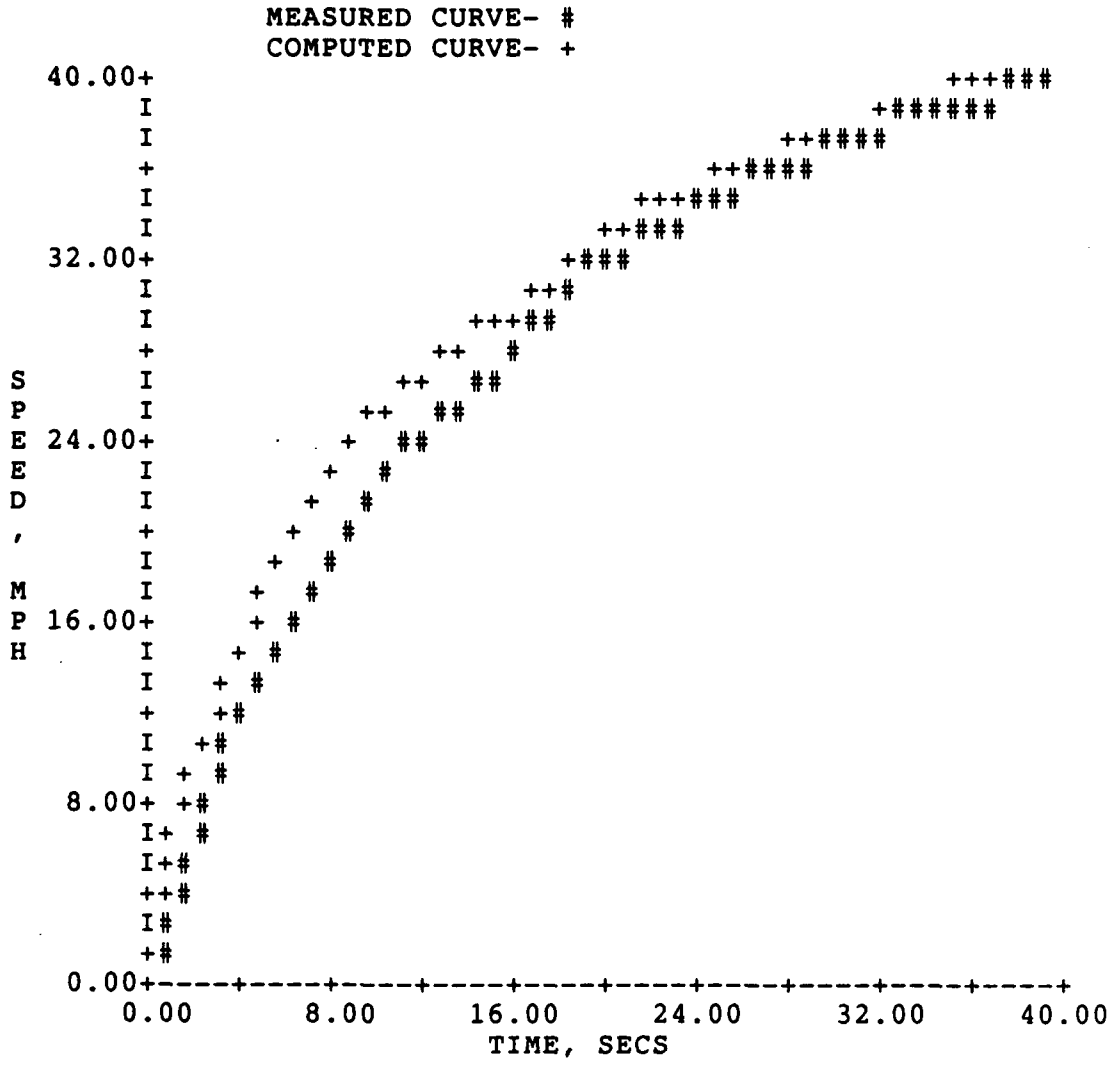


Figure 3. Sample speed versus time plot for M2

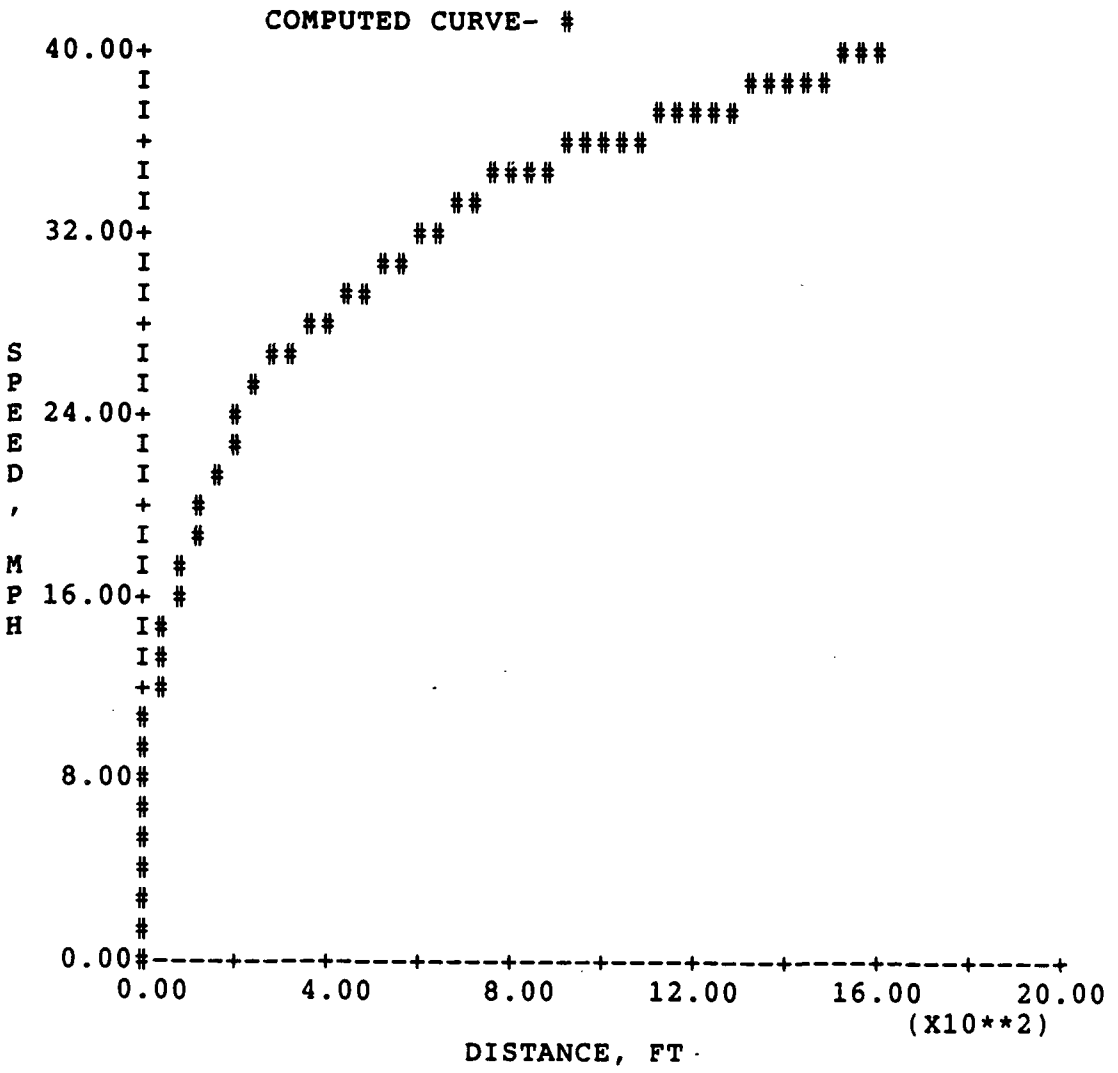


Figure 4. Sample speed versus distance plot for M2

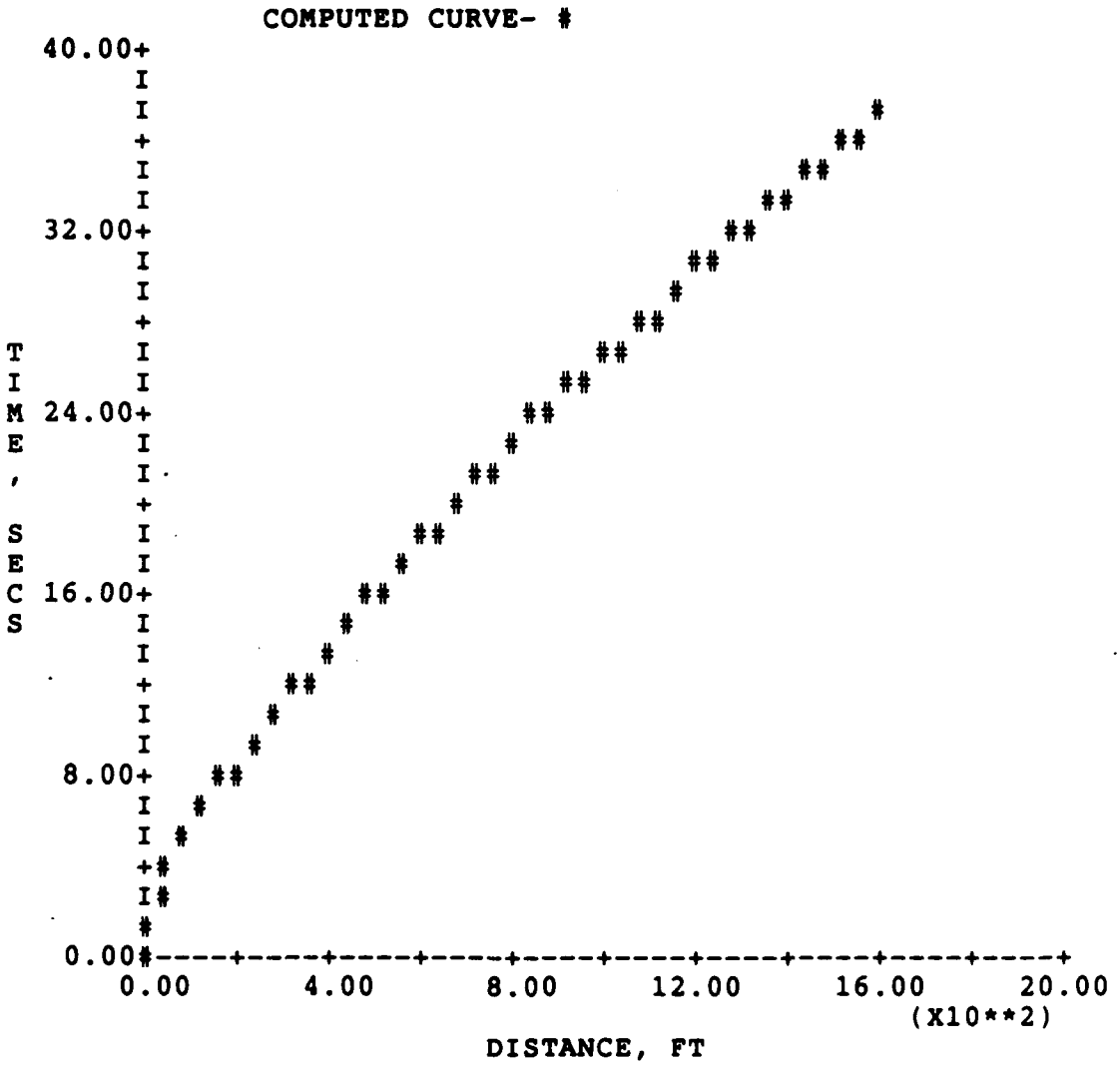


Figure 5. Sample time versus distance plot for M2

TIME			M113 MILES	M113 MILES	M60A3 MILES	M113 MILES	M60A3 MILES	M60A3 MILES
0	0	45	0.220	0.192	0.164	0.157	0.136	0.129
0	1	30	0.447	0.437	0.409	0.402	0.378	0.349
0	2	15	0.707	0.681	0.674	0.655	0.627	0.621
0	3	0	0.920	0.906	0.897	0.890	0.872	0.865
0	3	45	1.167	1.158	1.130	1.123	1.106	1.098
0	4	30	1.382	1.354	1.326	1.313	1.285	1.278
0	5	15	1.503	1.493	1.466	1.454	1.427	1.414
0	6	0	1.646	1.618	1.590	1.575	1.547	1.527
0	6	45	1.846	1.821	1.799	1.776	1.764	1.746
0	7	30	2.057	2.028	2.000	1.979	1.957	1.941
0	8	15	2.234	2.216	2.189	2.182	2.163	2.147
0	9	0	2.350	2.333	2.320	2.302	2.293	2.269
0	9	45	2.534	2.506	2.489	2.466	2.438	2.415
0	10	30	2.646	2.618	2.590	2.562	2.534	2.506
0	11	15	2.860	2.832	2.805	2.783	2.776	2.750
0	12	0	3.067	3.039	3.011	2.983	2.977	2.964
0	12	45	3.307	3.286	3.260	3.232	3.205	3.199
0	13	30	3.579	3.550	3.525	3.497	3.469	3.458
0	14	15	3.818	3.790	3.767	3.756	3.728	3.717
0	15	0	4.028	4.004	3.980	3.960	3.932	3.904
0	15	45	4.241	4.228	4.205	4.182	4.154	4.126
0	16	30	4.443	4.415	4.388	4.369	4.341	4.318
0	17	15	4.659	4.637	4.610	4.598	4.575	4.554
0	18	0	4.835	4.807	4.779	4.765	4.741	4.723
0	18	45	5.036	5.008	4.986	4.964	4.936	4.914
0	19	30	5.233	5.215	5.202	5.179	5.151	5.122
0	20	15	5.473	5.445	5.439	5.429	5.401	5.373
0	21	0	5.666	5.638	5.610	5.597	5.569	5.548
0	21	45	5.867	5.839	5.811	5.784	5.756	5.734
0	22	30	6.069	6.041	6.013	6.006	5.985	5.965
0	23	10	6.094	6.094	6.094	6.094	6.094	6.094

VEHICLE	TIME TO FINISH
M113	0 22 36
M113	0 22 42
M60A3	0 22 51
M113	0 22 52
M60A3	0 23 1
M60A3	0 23 6

Figure 6. Sample output from the WES Column Movement Model

Table 1

Terrain, Vehicle, and Driver Attributes

Characterized in the AMM Data Base

Terrain	Vehicle
Surface composition	Geometric characteristics
Type	Inertial characteristics
Strength	Mechanical characteristics
Surface geometry	
Slope	
Discrete obstacles	
Roughness	
Vegetation	
Stem size and spacing	
Visibility	
Linear geometry	
Stream cross section	
Water velocity and depth	

Driver
Reaction time
Recognition distance
Vertical acceleration limit
Horizontal acceleration limit

Table 2

Terrain Data Required for AMM

Terrain or Road Factor	Description*	Range	No. of Factor Classes
Off-Road			
1. Surface material			
a. Type	USCS/other	NA	4
b. Mass strength	CI or RCI	0 to >280	11
c. Wetness	NA	NA	4
2. Slope	Percent	0 to >70	8
3. Obstacle			
a. Approach angle	Degrees	90 to 270	14
b. Vertical magnitude	cm	0 to >85	7
c. Length	m	0 to >150	7
d. Width	cm	0 to >120	5
e. Spacing	m	0 to >60	8
f. Spacing type	NA	NA	2
4. Surface roughness (× 10)	rms, cm	0 to >7.5	9
5. Stem diameter	cm	0 to >25	8
6. Stem spacing	m	0 to >20	8
7. Visibility	m	0 to >50	9
8. Left approach angle (LA)	Degrees	90 to 270	20
9. Right approach angle (RA)	Degrees	90 to 270	20
10. Differential bank height or differential vertical magnitude (Δ)	m	0 to >4	9
11. Base width or top width	m	0 to >70	21
12. Low bank height or least vertical magnitude (LBH)	m	0 to >6	8

(Continued)

* AMM can accept terrain data in either inch-pound or metric units of measurement. Data preprocessors in AMM convert values of all input data to the inch-pound system before calculations involving these data occur.

Table 2 (Concluded)

Terrain or Road Factor	Description*	Range	No. of Factor Classes
	Off-Road (Continued)		
13. Water depth (D)	m	0 to >5	6
14. Water velocity	mps	0 to >3.5	6
	On-Road		
15. Surface material			
a. Type	NA	NA	4
b. Surface strength	CI or RCI	0 to >280	11
16. Slope	Percent	0 to 50	8
17. Surface roughness (× 10)	rms, cm	0 to 4	9
18. Curvature	Degrees	<140 to 180	9
19. Visibility	m	0 to 91.4	9

INFLUENCE OF REFLECTED SHOCK WAVES ON A HYPERSONIC SHAPED CHARGE JET

H.W. Meyer
J.E. Danberg

Ballistic Research Laboratory
Aberdeen Proving Ground, Md. 21005

Abstract

Considerable research has been devoted to shaped charge jet formation and penetration but little work has been reported on the aerodynamic forces on the jet particles, particularly the interference caused by near by surfaces. The object here was to develop numerical methods and apply them to this problem. A Godunov inviscid technique has been used along with high temperature thermodynamic properties to obtain the flow field in front of a hemisphere. This was used as the initial condition for the computation of the flow field in the annular region between the jet and a surrounding cylindrical tube. Computations were done at Mach number 4 and compared to experimental data. The jet problem (Mach number 20.45) was then solved.

I. INTRODUCTION

The objective of this effort is to study the hypersonic flow field associated with a shaped charge jet. The ultimate concern is the evaluation of how and to what extent aerodynamic effects cause perturbations to the jet. The work to be described here has concentrated on development of numerical techniques applicable to this problem.

A shaped charge warhead consists of a cylindrical explosive charge with a conical cavity in one end. The cavity is typically lined with a hollow conical copper liner of about 2 mm thickness. The shaped charge jet is formed when a detonation wave, traveling through the surrounding explosive, implodes the liner upon itself with such force that a stream of copper is ejected along the axis of the cone. A precision warhead as shown in Figure 1 produces a thin jet traveling at a speed above Mach 20 at standard sea level conditions.

A flash radiograph of a jet from the BRL 3.2 inch precision shaped charge is shown in Figure 2. The break up into many small particles is characteristic of all jets. As long as the particles remain aligned, the jet is highly lethal. If the jet particles are perturbed because of aerodynamic interference between particles or because of wave reflections from near by surfaces, its lethality can be seriously degraded.

While considerable research has been conducted in the fields of jet formation and jet penetration, little effort has been devoted to studying the aerodynamic forces that influence the jet. Many examples have been obtained which show jets disturbed in passing through but not touching various geometries, and it was concluded that aerodynamic forces were most probably

responsible. Experiments were initiated to eliminate the aerodynamic factors by producing a jet in a vacuum. However, the experiments were inconclusive because of the difficulty of maintaining the vacuum during the penetration process.

Yen[1], under contract to the Ballistic Research Laboratory, attempted to developed a better understanding of the interaction of the flow field with the jet. His effort concentrated on the wake behind the particle and provided very limited results.

The approach adopted in this work is to extend and apply techniques developed for solving the ballistic reentry problem to the special situation of the shaped charge jet as it passes near interfering surfaces. The method employed was originally developed by S.K. Godunov[2,3] in 1959-1961 and applied to the hypersonic blunt body by Masson et al[4]. The technique has also been used at BRL to simulate the flow field near the muzzle blast[5,6].

The numerical method will be briefly outlined in the following section. The basic technique for solving the conservation equations has been extended by coupling to it a program to compute the real gas thermodynamic properties appropriate to the hypersonic flow following the methods developed by Hansen[7]. The real gas jump conditions across the shock waves in the flow are evaluated using a method proposed by Colella and Glaz[8].

II. GODUNOV TECHNIQUE

In this section the essential elements of the Godunov method are described as applied to the simplest case of one dimensional flow. The applicable conservation equations for the axisymmetric flow which are used in the computation are then described, followed by the discretization actually employed in the solution algorithm. Finally in this section some details of the real gas method are presented.

1-D GODUNOV-RIEMANN TECHNIQUE

The conservation equations for mass, momentum and energy are written in integral form and applied to the physical domain divided into cells of width Δx as shown in Figure 3. At an initial instant the flow variables in each cell are defined as spatial average values. Discontinuous changes of properties in general occur at the cell boundaries. At the beginning of a time step, the imaginary diaphragm at the cell boundary is ruptured and compression and expansion waves are assumed to propagate into adjacent cells. This is analogous to the classical shock tube problem. The conditions behind these waves are well known from shock tube theory. The fluid between the two waves can be considered as two fluids, one from each cell, separated by the contact discontinuity. The pressure and velocity of the gas is the same on both sides of the contact discontinuity. As the waves propagate the original cell boundary lies in one of four possible flow regions. Both waves may move into the cell to the right, or both waves move to the left. In either of these cases the flow at the boundary is defined by the undisturbed flow in left or right cell respectively. The boundary can be between the two waves whereupon the flow is determined by whether the contact discontinuity moves in the positive or negative direction.

The basic equations to be solved for conditions behind the waves can be written as the following two simultaneous equations:

$$a(u_2 - u_1) + (p_2 - p_1) = 0 \tag{1}$$

$$b(u_3 - u_4) - (p_3 - p_4) = 0 \tag{2}$$

where the subscripts refer to regions defined in Figure 3. Because of continuity across the contact discontinuity $u_2 = u_3 = u_0$ and $p_2 = p_3 = p_0$. The coefficients a and b are mass velocities determined by the respective wave speeds and the density across the waves. In general a and b are functions of the pressure behind the waves which makes the solution for p_0 and u_0 nonlinear. Iterative techniques are used to solve Equations (1) and (2) except when the waves are weak, in which case a linear approximation is valid. With a sufficiently fine grid most of the shock free flow field can be obtained using the simpler linear approximation. The real gas relationships between thermodynamic variables also complicates the calculations and will be discussed later.

The essential element in the technique is that the properties are known and constant at the cell boundary until the arrival of waves developed at neighboring cell boundaries. If the time step of the calculation is kept less than the time required for the waves to cross the cell then the fluxes at the cell boundaries are easily evaluated. The average properties in the cell at the end of the time step are then determined as the initial value plus the fluxes across both boundaries during the time step. Thus a time marching scheme is defined which progresses from an imposed initial condition to a steady state.

In the case of an adaptive cell distribution, the fluxes are calculated taking into account the relative velocity between the fluid and the moving cell boundary.

CONSERVATION EQUATIONS

The fundamental elements of the Godunov method described for the simpler one dimensional problem have been extended by Godunov and many others to two dimensional axisymmetric flows. The flow near the leading element of the shaped charge jet is assumed to be axisymmetric thus the starting point is the conservation equations in cylindrical coordinates, as follows:

$$\frac{\partial}{\partial t}(\rho) + \frac{\partial}{\partial x}(\rho u) + \frac{\partial}{\partial r}(\rho v) + \frac{\rho v}{r} = 0 \tag{3}$$

$$\frac{\partial}{\partial t}(\rho u) + \frac{\partial}{\partial x}(p+\rho u^2) + \frac{\partial}{\partial r}(\rho uv) + \frac{\rho uv}{r} = 0 \tag{4}$$

$$\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}(\rho uv) + \frac{\partial}{\partial r}(p+\rho v^2) + \frac{\rho v^2}{r} = 0 \tag{5}$$

$$\frac{\partial}{\partial t}(\rho \epsilon) + \frac{\partial}{\partial x}\rho u(\epsilon+\frac{p}{\rho}) + \frac{\partial}{\partial r}\rho v(\epsilon+\frac{p}{\rho}) + \frac{\rho v}{r}(\epsilon+\frac{p}{\rho}) = 0 \tag{6}$$

For the blunt nose part of the calculation it is convenient to define the cells in a polar coordinate system because the blunt body shock wave is nearly concentric to the spherical surface in the stagnation region. Figure 4 illustrates how the cell geometry is defined in this region. The shock location is estimated initially and the grid dimensions are allowed to change in the radial direction until the steady state shock position is obtained. The above differential equations of motion are integrated over a cell area as shown in the figure and Table 1 gives the resulting discretized equations. Note that R,U,V and E are the density, x and r components of velocity and total energy respectively, evaluated on the cell boundary. Thus these properties are obtained from the solution of a Riemann problem at that boundary. The subscripts indicate which boundary as defined in Figure 4. Note that on radial boundaries the fluxes are evaluated using the velocity W which is the component of the velocity vector normal to the cell boundary. On the moving circumferential boundaries the flux is determined by the relative velocity component normal to the moving element, (W-q). Pressure forces contribute to the momentum equations depending on the orientation of the cell boundary relative to the cylindrical coordinate system; thus the angles θ and ϕ which specify the orientation of the boundary must be included.

In both the hemisphere and cylindrical computation the downstream boundary was located in a supersonic flow region where the wave system moves downstream. Thus the flux conditions on the boundary of the cell are determined by the properties of the cell, and no special boundary condition is required.

The last term in each equation caused some computational difficulties for cells with the axis of symmetry as a boundary. Although the ratio of v/r is finite on the boundary, its evaluation introduces errors which lead to instabilities in the computation. One source of the difficulty is related to discretization of the shock wave at the axis of symmetry where it should be normal. In the present computations the slope of the shock at the first cell with its boundary on the axis tended to an unrealistically negative value. By arranging the cells so that the center of the first cell falls on the axis, the normality of the shock on the axis is insured. This occurs because all conditions on the two radial boundaries are forced to be symmetrical as the appropriate boundary condition. The line segment representing the stagnation region shock wave is then automatically normal to the axis.

REAL GAS EFFECTS

In the initial phases of the development of the numerical code, a perfect gas equation of state was assumed, ie.:

$$p = \rho RT \tag{7}$$

$$e_{int} = \frac{RT}{\gamma - 1} \tag{8}$$

At Mach 20, however, the relationship between pressure and temperature and the other thermodynamic variables is much more complex. Although the magnitude of the pressures calculated using perfect gas formulas are approximately correct, the shock wave position and thus pressure distribution are strongly affected by real gas density and temperature.

The thermodynamic properties of high temperature air are computed from approximate partition functions for the major components of air using a technique developed by Hansen[7]. The model assumes air to be a mixture of 20 percent oxygen and 80 percent nitrogen, all other components are neglected. Eleven species are considered including three levels of ionization of oxygen and nitrogen.

The method calculates all the thermodynamic variables given the pressure and temperature of the gas. This is inconvenient when coupled to the Godunov code because its primary dependent variables include density and energy. An iteration procedure is required as an intermediate step to search for the correct values of temperature and pressure which correspond to given density and energy.

A second major problem with adding real gas effects is that the above search for the thermodynamic variables makes the computation of even the weak wave Riemann problem nonlinear and iterative. In order to avoid increasing the running time of the computation extensively, a procedure suggested by Colella and Glaz[8] has been adopted. Their method permits evaluation of the pressure and velocity at the contact discontinuity based on conditions in adjacent cells. In the strong shock case, it is still necessary to iterate as in the perfect gas case, but it is not necessary to iterate at the same time to find the thermodynamic variables.

III.RESULTS

The results of the computations are summarized by first considering the hemisphere problem at the relative low speed of Mach 4 where experimental verification can be made. Some results for the flow downstream between concentric cylinders are considered to illustrate the application of the shaped charge jet passing through a cylindrical tube. Finally hemisphere and cylinder computations at Mach number 20.45 are presented.

HEMISPHERE AT MACH NUMBER 4

Figure 5 shows the shock wave stand-off distance plotted around the hemisphere and compared to experimental data[9]. This Mach 4 case was chosen for these initial code verification runs because of the existence of the experimental wind tunnel observations and because this was one of Godunov's original test cases. These results were obtained with a relatively coarse grid of only 8 points radially and 25 cells in the angular direction. Tests with other grid distributions showed only very minor changes. Figure 6 shows the corresponding pressure distribution on the hemisphere again compared to the same experiment. Note that the real gas form of the code was used even though the conditions were essentially those of a perfect gas. The code under predicts the measured stagnation pressure by less than 2.5 per cent. There is a small kink in the pressure curve at about 45 degrees from the stagnation point which appears to be associated with the sonic line in the flow field. It is some what exaggerated because of the coarse grid used. The mass density distribution is shown in Figure 7 and the current calculations are compared to the original results of Godunov and to a well know calculation of Belotserkovskii[10] using the method of integral relations. The Belotserkovskii calculation is about 2

percent higher at the stagnation point but otherwise the agreement is very good. Godunov's calculation is lower than the others and is significantly different at the stagnation point. The reason for the disagreement appears to be because Godunov's grid provided for a grid boundary on the line of symmetry which introduced errors that do not disappear as steady state is approached.

ANNULAR REGION AT MACH NUMBER 4

Once the hemisphere computation was completed, it was used to provide upstream boundary conditions for the computation of the flow between a cylindrical afterbody and a concentric wall. This configuration is meant to simulate a jet passing through a cylindrical tube. Unlike the hemisphere computation, however, the grid or cell distribution was fixed in space with 40 cells radially and 120 axially. Uniform initial conditions were assumed and the code marched in time until steady state was achieved. Figure 8 is an example of a contour plot of the pressure from such a calculation. The outer wall diameter is 1.766 body diameters. The hemisphere shock wave reflects off the outer wall and produces a nearly normal shock standing on the body. Such shocks impinging on the jet could produce strong aerodynamic forces at the higher Mach numbers of interest.

MACH NUMBER 20.45 RESULTS

Figures 9 and 10 show the results of the hemispherical computation for Mach number 20.45. The stagnation pressure in this case is only a few percent above the 539 atmospheres predicted by perfect gas theory. The result of the cylinder computations is shown in figure 11, for a tube to jet diameter ratio of 2.50. The bow shock is plainly visible. The shock angle at the wall is 12.8°, and the reflection occurs at a position three body diameters downstream of the tangent point between the hemisphere and cylinder. The pressure on the wall behind the reflected wave is 71 atm. The reflected wave can be traced back to the body, where it again reflects. The pressure behind this reflection is 34 atm.

IV. SUMMARY AND CONCLUSIONS

This report has presented the current status of an on going program to calculate the aerodynamic forces on a hypersonic shaped charge jet. A numerical technique based on the work of S. K. Godunov has been modified consistent with the blunt jet configuration penetrating a cylindrical tube and extended to include real gas properties. The ability of the code to simulate Mach number 4 conditions for which experimental and other numerical data are available has been used to validate the procedure. Numerical results have been completed for the Mach number 20.45 jet problem. These computations will be compared to experimental measurements of the reflected wave from an actual shaped charge jet going through a cylindrical tube. Results from these experiments should be reported in the near future.

REFERENCES

1. Yen, S.M.,"Interactions Between Multiple Objects in a Hypervelocity Flow Regime," Final Report on Contract DAAK 11-81-C-0011, Aeronautical and

Astronautical Engineering Department, University of Illinois, Urbana Illinois, June 1983.

2. Godunov, S. K., "Finite Difference Method for Numerical Computation of Discontinuous Solutions of the Equations of Fluid Dynamics," Matematicheskii Sbornik, Vol. 47(89) No. 3, p. 271, 1959, Translated by I.O. Bohachevsky.

3. Godunov, S.K., Zabrodyn, A.W. and Prokopov, G.P.,"A Difference Scheme for Two-Dimensional Unsteady Problems of Gas Dynamics and Computation of Flow with a Detached Shock Wave," Zhurnal Vychyslitelnoi Matematiki i Matematicheskoi Fiziki, Vol. I, no. 6, pp 1020-1050, 1961, Translated by I.O. Bohachevsky, Cornell Aeronautical Laboratory, Inc.

4. Masson, B.S., Taylor, T.D. and Foster, R.M., "Application of Godunov's Method to Blunt-Body Calculations," American Institute of Aeronautics and Astronautics Journal, Vol. 7, No. 4, pp.694-698, April 1969.

5. Widhopf, G.F. and Schmidt, E.M., "Time-Dependent Near Muzzle Brake Flow Simulations," American Institute of Aeronautics and Astronautics/American Society of Mechanical Engineers 3rd Joint Thermophysics, Fluids, Plasma and Heat Transfer Conference, paper AIAA-82-0973, St. Louis, Missouri, June 1982.

6. Fansler, K.S.. "Gasdynamic Quantities about a Ramjet Projectile while in the Transitional Ballistics Region," American Institute of Aeronautics and Astronautics 12th Atmospheric Flight Mechanics Conference, paper AIAA-85-1840-CP, Snowmass Colorado, August 1985.

7. Hansen, C.F., "Approximations for the Thermodynamic and Transport Properties of High-Temperature Air," National Aeronautics and Space Administration Technical Report R-50, 1959.

8. Colella, P. and Glaz, H.M., "Efficient Solution Algorithms for the Riemann Problem for Real Gases," Journal of Computational Physics, Vol.59, pp. 264-289, 1985.

9. Belotserkovskii, O.M.,(ed). "Supersonic Gas Flow Past Blunt Bodies: Theoretical and Experimental Studies," Tr. Vychisl. Tsentr. Akad. Nauk SSSR, published by Computing Center AN SSSR, Moscow, (2nd Edition). 1967, US Army Foreign Science and Technology Center Technical Translation, FSTC-HT-23-447-68, 1968.

10. Belotserkovskii, O.M., "On the Computation of Flow around Axisymmetric Bodies with a Detached Shock Wave on an Electronic Computer," Prikl. Mat. i. Mekh, Vol. 24, No. 3, pp 511-517, 1960.

$$\left[\rho A\right]^{n-1/2,m-1/2} - \left[\rho A\right]_{n-1/2,m-1/2} - \tau\Big\{ [RWS]_{n,m-1/2} + [R(W-q)S]_{n-1/2,m-1} + [-RWS]_{n-1,m-1/2} + [R(q-W)S]_{n-1/2,m} + \left[\frac{\rho v}{r}A\right]_{n-1/2,m-1/2} \Big\}$$

$$\left[\rho uA\right]^{n-1/2,m-1/2} - \left[\rho uA\right]_{n-1/2,m-1/2} - \tau\Big\{ [(RUW+P\sin\theta)S]_{n,m-1/2} + [(RU(W-q)-P\sin\phi)S]_{n-1/2,m-1} + [(-RUW-P\sin\theta)S]_{n-1,m-1/2} + [(RU(q-W)+P\sin\phi)S]_{n-1/2,m} + \left[\frac{\rho uv}{r}A\right]_{n-1/2,m-1/2} \Big\}$$

$$\left[\rho vA\right]^{n-1/2,m-1/2} - \left[\rho vA\right]_{n-1/2,m-1/2} - \tau\Big\{ [(RVW+P\cos\theta)S]_{n,m-1/2} + [(RV(W-q)-P\cos\phi)S]_{n-1/2,m-1} + [(-RVW-P\cos\theta)S]_{n-1,m-1/2} + [(RV(q-W)+P\cos\phi)S]_{n-1/2,m} + \left[\frac{\rho v^2}{r}A\right]_{n-1/2,m-1/2} \Big\}$$

$$\left[\rho\epsilon A\right]^{n-1/2,m-1/2} - \left[\rho\epsilon A\right]_{n-1/2,m-1/2} - \tau\Big\{ [(P+R\Xi)WS]_{n,m-1/2} + [(PW+R\Xi(W-q))S]_{n-1/2,m-1} + [-(P+R\Xi)WS]_{n-1,m-1/2} + [(-PW+R\Xi(q-W))S]_{n-1/2,m} + \left[(p+\rho\epsilon)\frac{v}{r}A\right]_{n-1/2,m-1/2} \Big\}$$

Table 1. Finite Difference Form of the Equations of Motion.

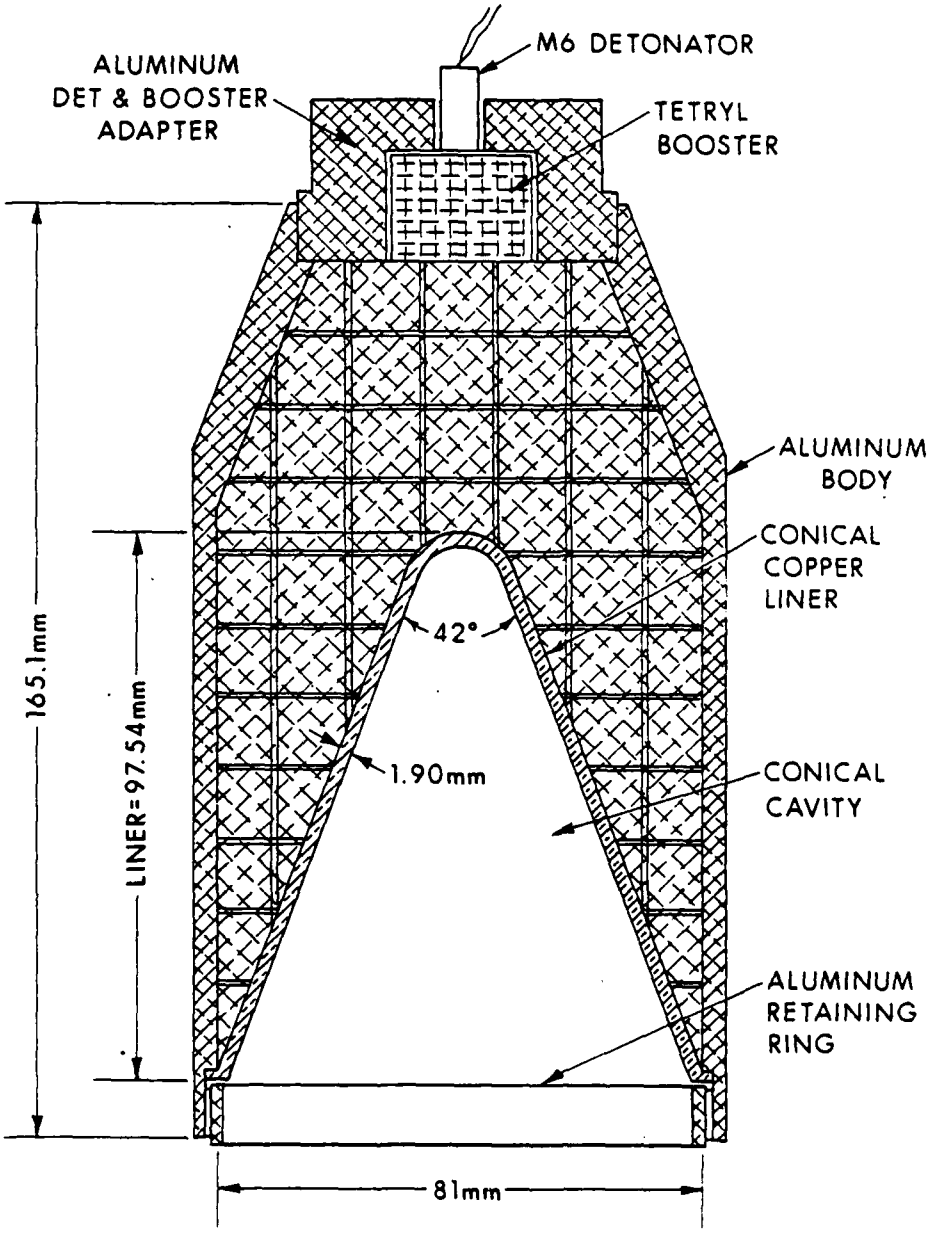


Figure 1. The BRL 3.2 inch Precision Shaped Charge.

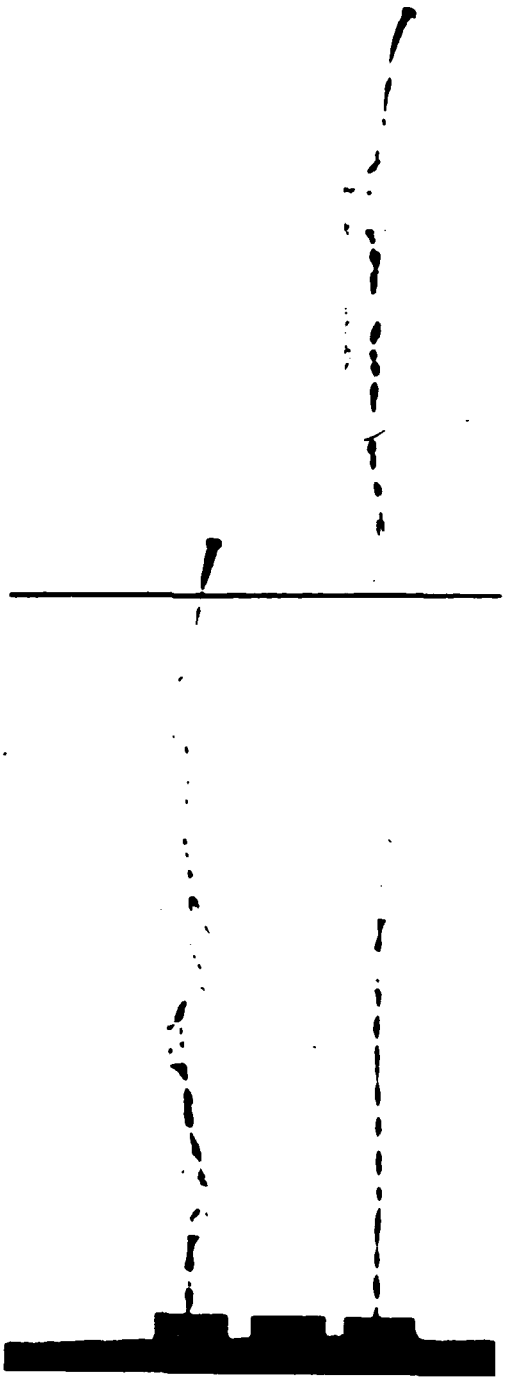


Figure 2. Flash Radiograph of a Perturbed 3.2 inch Jet.

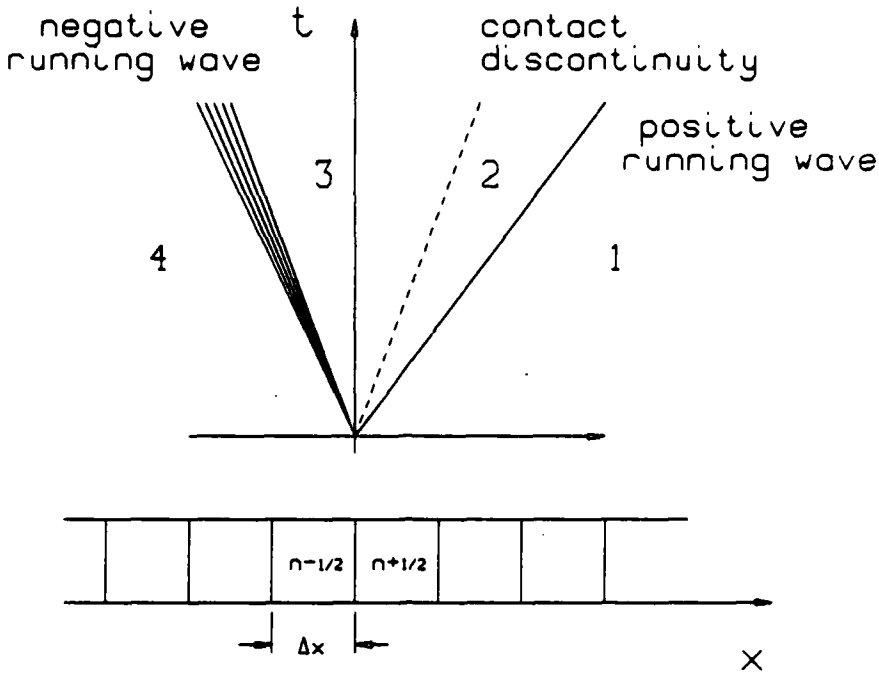


Figure 3. One Dimensional Godunov Technique.

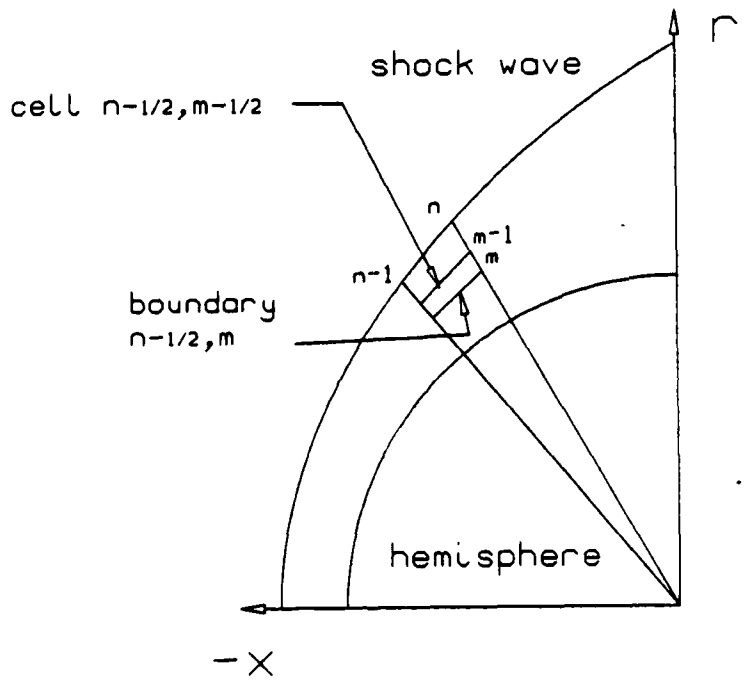


Figure 4. Hemisphere Shock Layer with a Typical Cell.

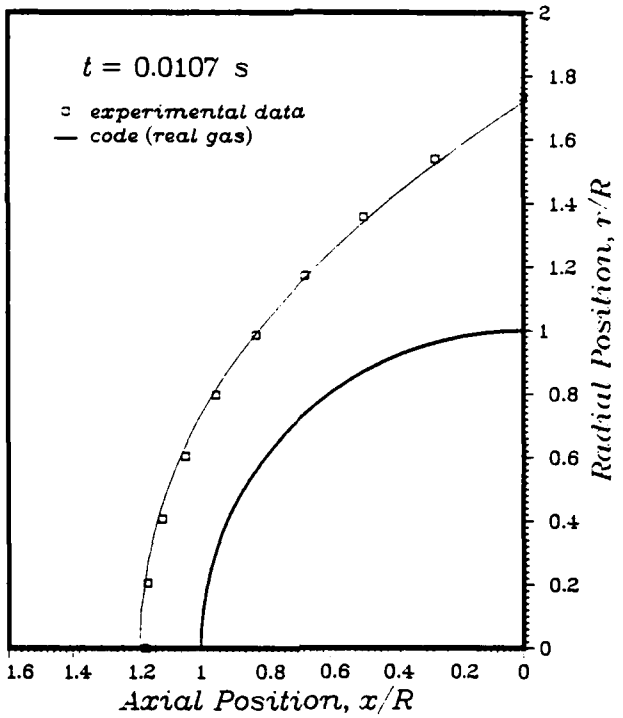


Figure 5. Shock Wave Position on Hemisphere, Mach 4.0.

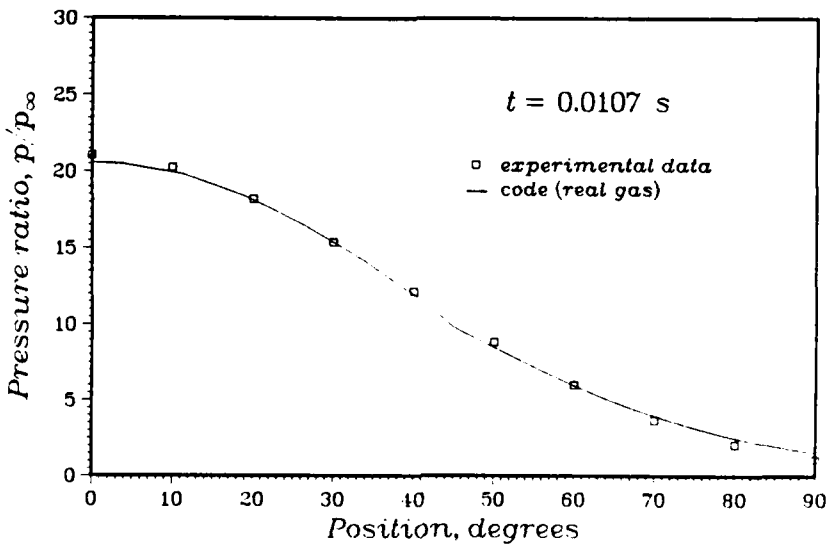


Figure 6. Wall Pressure Distribution on Hemisphere, Mach 4.0.

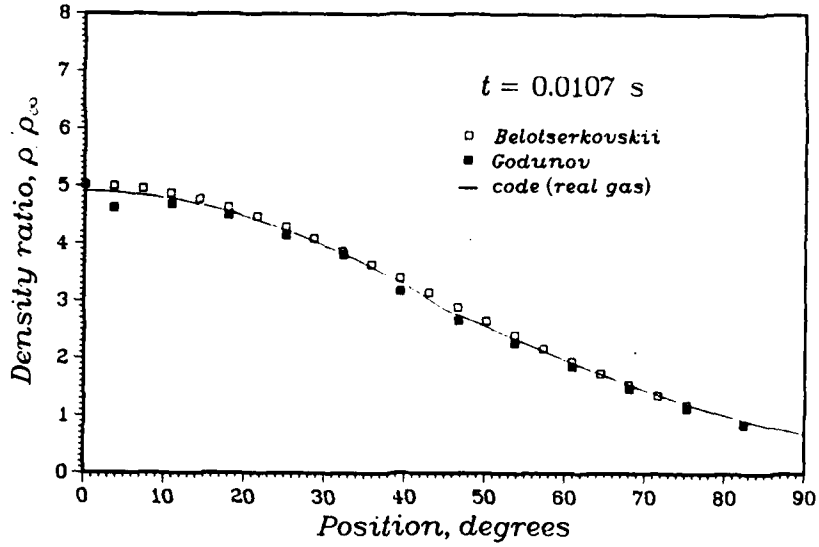


Figure 7. Wall Mass Density Distribution on Hemisphere, Mach 4.0.

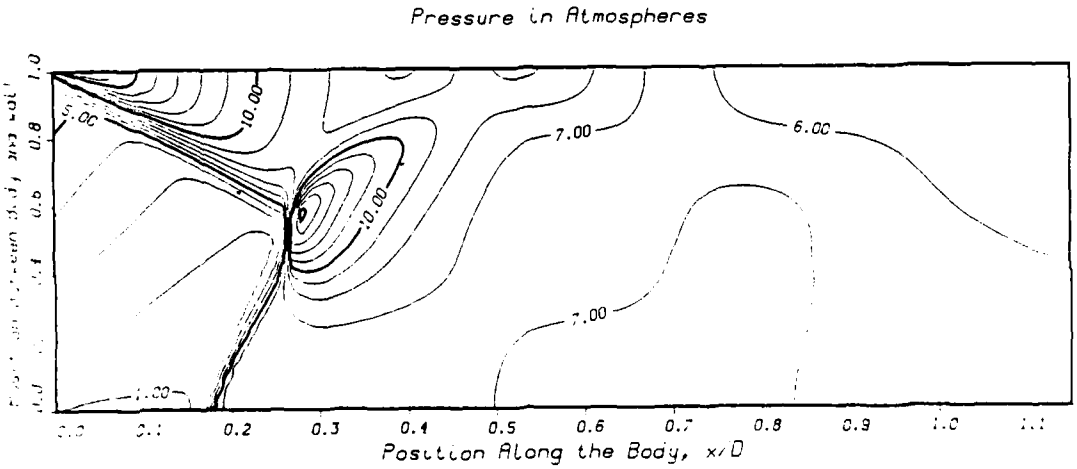


Figure 8. Pressure Contour Plot in Annulus, Mach 4.0.

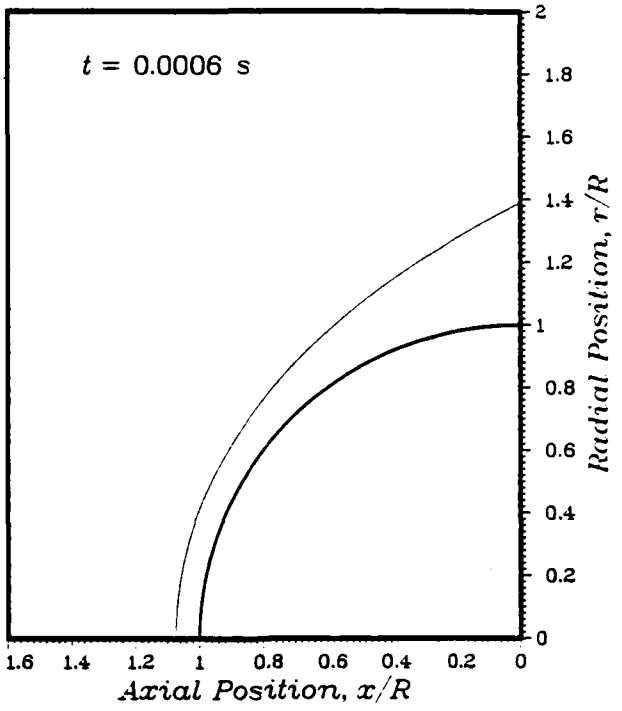


Figure 9. Shock Wave Position on Hemisphere, Mach 20.45.

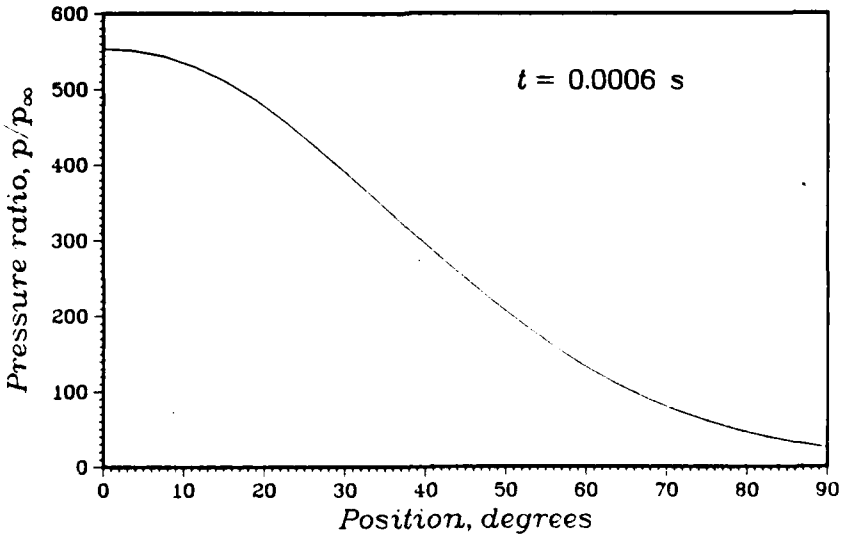


Figure 10. Wall Pressure Distribution on Hemisphere, Mach 20.45.

SKEW GRIDS AND IRROTATIONAL FLOW

Robert S. Bernard
Hydraulics Laboratory
U.S. Army Engineer Waterways Experiment Station
P.O. Box 631, Vicksburg, MS 39180-0631

ABSTRACT. Finite-difference computation of incompressible flow through regions of arbitrary shape often requires the implementation of boundary-fitted coordinates for which the grid lines may be non-orthogonal (skew). When the governing equations are expressed in terms of pressure and velocity, conservation of mass is maintained by the gradient of the pressure. In principle, the gradient is irrotational and should have no effect on the existing circulation in the flow field; but if the grid lines are skew, the discrete representation of the gradient can generate spurious vorticity near the boundaries. In the present work this difficulty is eliminated for uniform skew grids, and markedly reduced for non-uniform skew grids, by adopting a discrete formulation of the pressure gradient that helps maintain irrotationality near boundaries. The procedure is applicable for staggered grids with either Poisson or Chorin equations for pressure.

I. INTRODUCTION. The role of the pressure for incompressible flow is simply to constrain the velocity vector $\underline{u}$ such that

$$\nabla \cdot \underline{u} = 0 \tag{1}$$

which represents conservation of mass. Assuming that the velocity field conserves mass at some time t, let $\underline{u}'$ be a velocity field that would exist at time $t' = t + dt$ in the absence of pressure. Then the corresponding mass-conserving velocity field $\underline{u}$ can always be written as

$$\underline{u} = \underline{u}' - \rho^{-1}\nabla\phi \tag{2}$$

where ρ is the density and the scalar potential ϕ is related to the pressure p by

$$\phi = p\,dt \tag{3}$$

Combining Equations 1 and 2, it follows that ϕ satisfies the Poisson equation,

$$\nabla^2\phi = \rho\nabla \cdot \underline{u}' \tag{4}$$

As long as the primitive variables $\underline{u}$ and p are retained as the unknowns in the governing equations of motion, then it is necessary to solve Equation 4 in order to maintain the constraint given by Equation 1. Even if Chorin's method of pseudo-compressibility [1] is used to add a time derivative of pressure to Equation 1, the end result is equivalent to having solved Equation 4 when the flow reaches steady state.

The purpose herein is not, however, to discuss the pros and cons of pseudo-compressibility versus Poisson equations in the numerous existing algorithms for solving the coupled momentum and continuity equations. It is, rather, to consider the proper formulation of discrete approximations for the pressure gradient and its boundary conditions on non-orthogonal curvilinear grids. Improper treatment of the gradient near computational boundaries can add circulation to the flow, in which case the discrete representation of the gradient is not irrotational as it should be. This sort of error does not arise if the computational grid is orthogonal; but if the grid is non-orthogonal, special treatment of the derivatives in the gradient is necessary to avoid or minimize the creation of spurious vorticity. The objective of the present work is to ascertain what that treatment might be for the case of non-orthogonal, non-uniform, curvilinear finite-difference grids.

II. DISCRETE FORMULATION. Consider a two-dimensional staggered grid of the Marker-and-Cell type [2], with the pressure (and the scalar potential) computed at the cell centers, and the velocity components (u,v) at the midpoints of the cell faces, as shown in Figures 1 and 2. Assuming a unit density and a unit depth normal to the page, the mass-flux components through the right (east) and upper (north) cell faces are denoted by U and V, respectively, and are related to the cartesian velocity components (u,v) by

$$U = y_\eta u - x_\eta v \tag{5}$$

$$V = x_\xi v - y_\xi u \tag{6}$$

The curvilinear coordinates (ξ,η) follow the grid lines shown in Figure 1, and they are functions of the cartesian coordinates (x,y). Conservation of mass for each grid cell demands that

$$U_\xi + V_\eta = 0 \tag{7}$$

Denoting non-conservative velocities and fluxes with a prime, we then have the relations

$$u = u' - \phi_x \tag{8}$$

$$v = v' - \phi_y \tag{9}$$

Using the chain rule [3] to evaluate the x- and y-components of the gradient, we find that

$$\phi_x = J^{-1} (y_\eta \phi_\xi - y_\xi \phi_\eta) \tag{10}$$

$$\phi_y = J^{-1} (x_\xi \phi_\eta - x_\eta \phi_\xi) \tag{11}$$

where J is the Jacobian of the coordinate transformation,

$$J = x_\xi y_\eta - x_\eta y_\xi \tag{12}$$

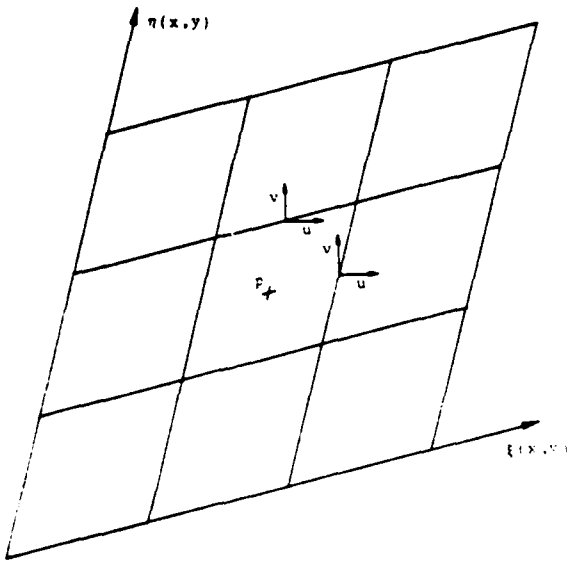


Figure 1. Computational grid in physical (x,y) space.

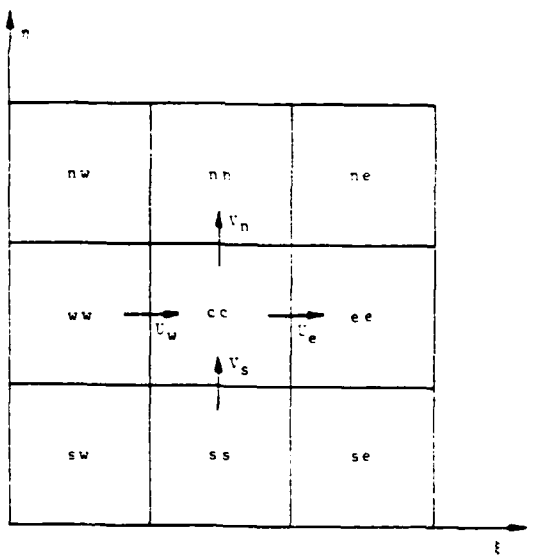


Figure 2. Computational grid in computational (ξ,η) space.

Combining Equations 5 through 11, we obtain the discrete analog of Equation 4,

$$A_e - A_w + B_n - B_s = U'_e - U'_w + V'_n - V'_s \tag{13}$$

The sub/superscripts (e, w, n, s) indicate quantities evaluated on the (east, west, north, south) cell faces, as shown in Figure 2, and

$$A = \alpha\phi_\xi - \gamma\phi_\eta \tag{14}$$

$$B = \beta\phi_\eta - \gamma\phi_\xi \tag{15}$$

$$\alpha = J^{-1}\ (x_\eta^2 + y_\eta^2) \tag{16}$$

$$\beta = J^{-1}\ (x_\xi^2 + y_\xi^2) \tag{17}$$

$$\gamma = J^{-1}\ (x_\xi x_\eta + y_\xi y_\eta) \tag{18}$$

Note that the right-hand side of Equation 13 is simply the imbalance in mass flux produced by the nonconservative fluxes (U',V'), while the left-hand side is the sum of the flux corrections provided by the gradient of ϕ. After Equation 13 has been solved for ϕ, and the flux corrections computed therefrom, the mass-conserving fluxes can be obtained from

$$U = U' - A \tag{19}$$

$$V = V' - B \tag{20}$$

The particular form exhibited by Equation 13 facilitates the elimination of superflous derivatives when one or more of the cell faces coincides with a boundary where there is to be no adjustment to U' or V'. Such is the case for solid boundaries and for open boundaries where the flux normal to the boundary is known or specified. For example, if the east cell face coincides with a boundary, then $A_e = 0$ and Equation 13 reduces to

$$-A_w + B_n - B_s = U'_e - U'_w + V'_n - V'_s \tag{21}$$

Likewise, if the north cell face coincides with a boundary, then $B_n = 0$, and Equation 13 reduces to

$$A_e - A_w - B_s = U'_e - U'_w + V'_n - V'_s \tag{22}$$

If the grid is orthogonal ($\gamma = 0$) there is no question as to how to proceed. In the case of Equation 21, only η-derivatives are needed on the north and south faces, and these can be evaluated from information inside the field. Similarly, for Equation 22, only ξ-derivatives are needed on the east and west

faces. Assuming the grid has unit spacing ($\Delta\xi = \Delta\eta = 1$) in the computational (ξ,η) space [4], then

$$\phi_\xi^e = \phi_{ee} - \phi_{cc} \tag{23}$$

$$\phi_\eta^n = \phi_{nn} - \phi_{cc} \tag{24}$$

The double subscripts (cc, ee, ww, nn, ss, ne, nw, se, sw) indicate quantities at the centers of neighboring cells (central, east, west, etc.) as shown in Figure 2.

If the grid is non-orthogonal ($\gamma \neq 0$) , then ξ-derivatives are needed on the north and south faces, and η-derivatives are needed on east and west faces. On cell faces not touching boundaries, we can approximate these with

$$\phi_\xi^n = \frac{1}{4}\left(\phi_{ne} - \phi_{nw} + \phi_{ee} - \phi_{ww}\right) \tag{25}$$

$$\phi_\eta^e = \frac{1}{4}\left(\phi_{ne} - \phi_{se} + \phi_{nn} - \phi_{ss}\right) \tag{26}$$

But for cell faces with one end touching a boundary, there is an ambiguity: Equations 25 and 26 require information across the boundary in cells lying outside the flow field. The ambiguity arises because there are a number of plausible ways to compute the needed information, but no indication a priori as to which one is best. In order to examine the possibilities, let us focus attention on a cell whose east face coincides with a boundary.

We must find an approximation for ϕ_ξ on the north and south cell faces, and we shall consider only three possibilities although there certainly exist others. The first is simply to replace ϕ_ξ on the north face by its value on the northwest corner of the cell, using the difference expression,

$$\phi_\xi^n = \frac{1}{2}\left(\phi_{nn} - \phi_{nw} + \phi_{cc} - \phi_{ww}\right) \tag{27}$$

Equation 27 imposes a ξ-derivative using information in the flow field, irrespective of what happens on the boundary. For reference we shall call this the "field approximation". The second possibility is to calculate ϕ_ξ on the north face by using the same condition that exists on the east face. Specifically, on the east face we have the boundary condition,

$$\alpha_e\, \phi_\xi^e = \gamma_e\, \phi_\eta^e \tag{28}$$

Applying the same constraint on the north face, we obtain

$$\alpha_n\, \phi_\xi^n = \gamma_n\, \phi_\eta^n \tag{29}$$

and the discrete approximation for ϕ_ξ becomes

$$\phi_\xi^n = \frac{\gamma_n}{\alpha_n}\left(\phi_{nn} - \phi_{cc}\right) \tag{30}$$

Equation 30 imposes a ξ-derivative using only the boundary constraint, irrespective of what happens in the flow field. For reference we shall call it the "boundary approximation". As the third alternative, we approximate ϕ_ξ by a simple average of Equations 27 and 30, which we shall call the "mixed approximation",

$$\phi_\xi^n = \frac{\gamma_n}{2\alpha_n}(\phi_{nn} - \phi_{cc}) + \frac{1}{4}(\phi_{nn} - \phi_{nw} + \phi_{ce} - \phi_{ww}) \tag{31}$$

We now have three discrete alternatives for representing ambiguous derivatives of ϕ adjacent to boundaries. Equations 27, 30, and 31 pertain to ξ-derivatives for north faces touching boundaries of constant ξ. The same principles apply for ambiguous derivatives on other faces touching (but not coincident with) flow field boundaries.

III. TEST CASES AND RESULTS. In order to ascertain which of the three alternatives is best for representing ambiguous derivatives of ϕ, we need test problems that show clearly the adverse affects of grid skewness. Moreover, we should pay special attention to the possible creation of spurious vorticity arising from improper representation of the gradient near boundaries. Thus we propose two classes of tests:

1. Non-orthogonal grids with uniform spacing.

2. Non-orthogonal grids with non-uniform spacing.

The first category allows us to see the effects of skewness alone, while the second adds the possible compounding of error due to non-uniformity.

In all cases the flow field in the physical space will be bounded above and below by solid boundaries, while the left and right boundaries will be open with uniform normal components of velocity: $u = 1$. Inside the flow field we impose the velocity condition: $u' = 1$, $v' = -10$. We then solve Equation 13 subject to the constraint that the flux normal to the boundaries remain fixed and no vorticity be created in the flow field. The large vertical velocity creates a proportionately large violation of continuity at the upper and lower solid boundaries, which is to be eliminated by the gradient of the scalar potential. In all cases the physical boundaries are chosen such that the resulting streamlines should be straight lines, and any deviation therefrom indicates the presence of error.

Results from the first three test cases are presented in Figures 3 through 5, showing computed streamlines for the flow through parallelograms of increasing skewness. Even at 10 degrees, the boundary and mixed approximations exhibit an unacceptable amount of circulation, while the field approximation produces straight lines for each case. The computed solutions were all converged to a maximum residual of less than 0.001 in Equation 13, with the residual ε defined by

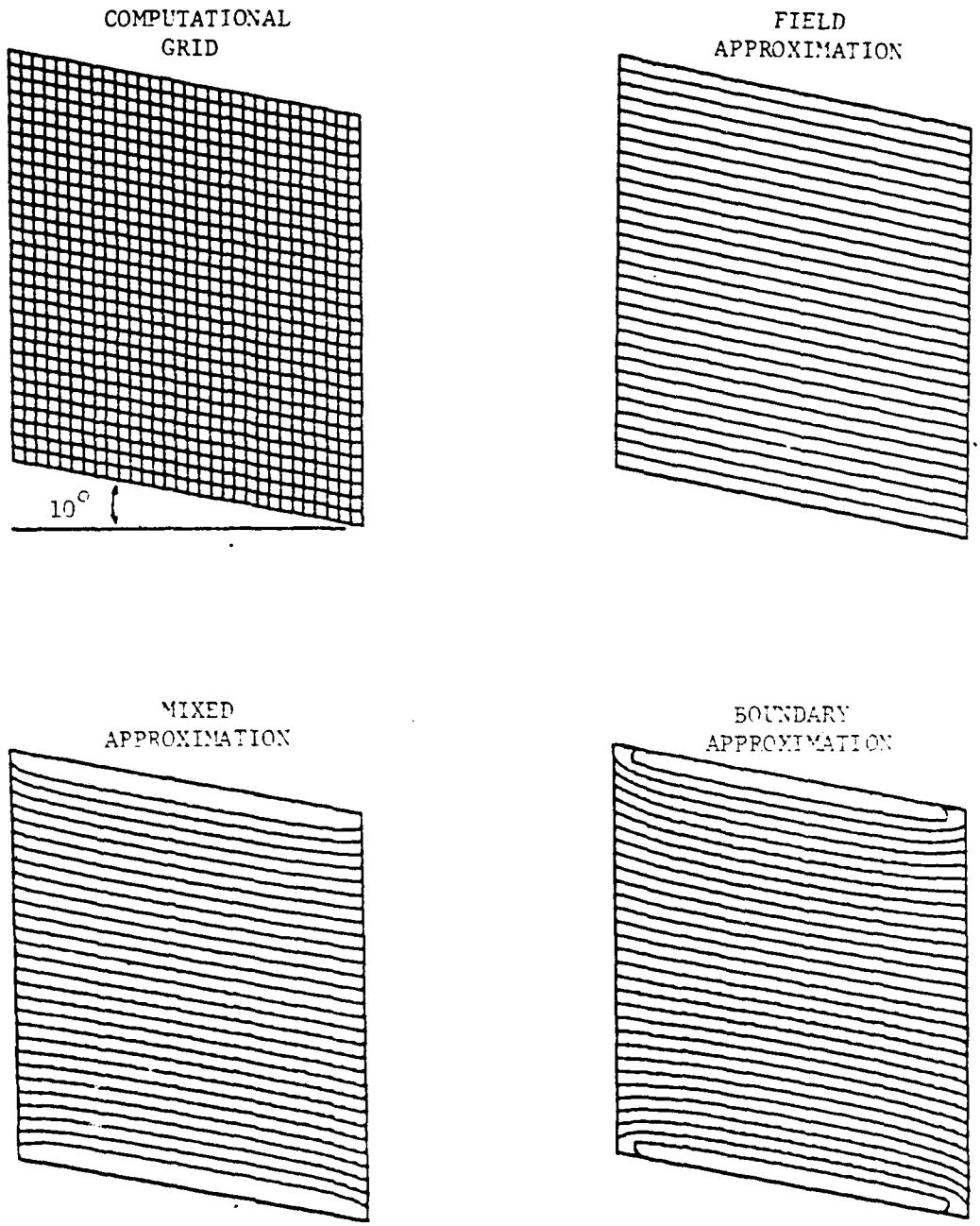


Figure 3. Grid and computed streamlines for uniform irrotational flow through 10-degree parallelogram.

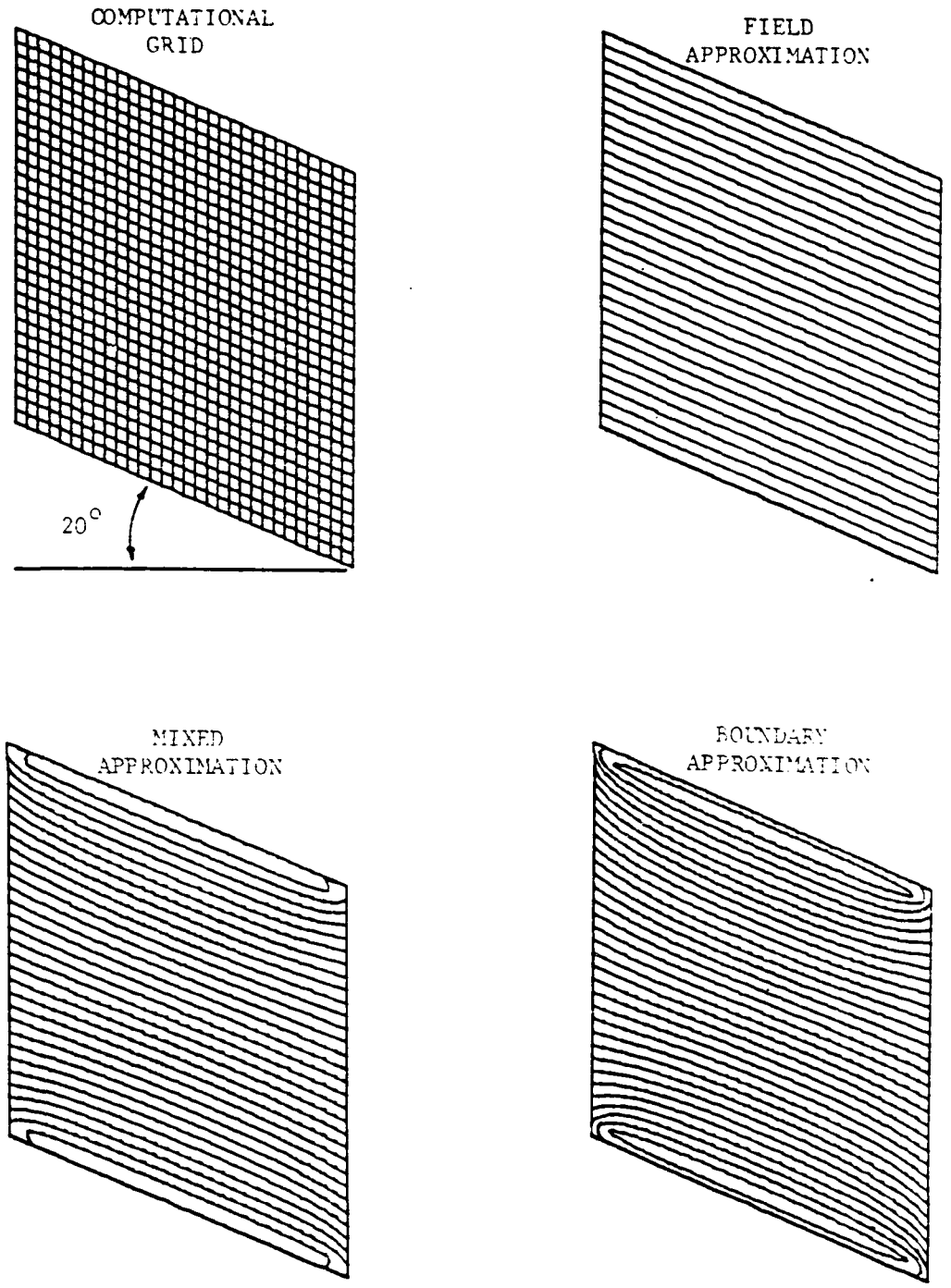


Figure 4. Grid and computed streamlines for uniform irrotational flow through 20-degree parallelogram.

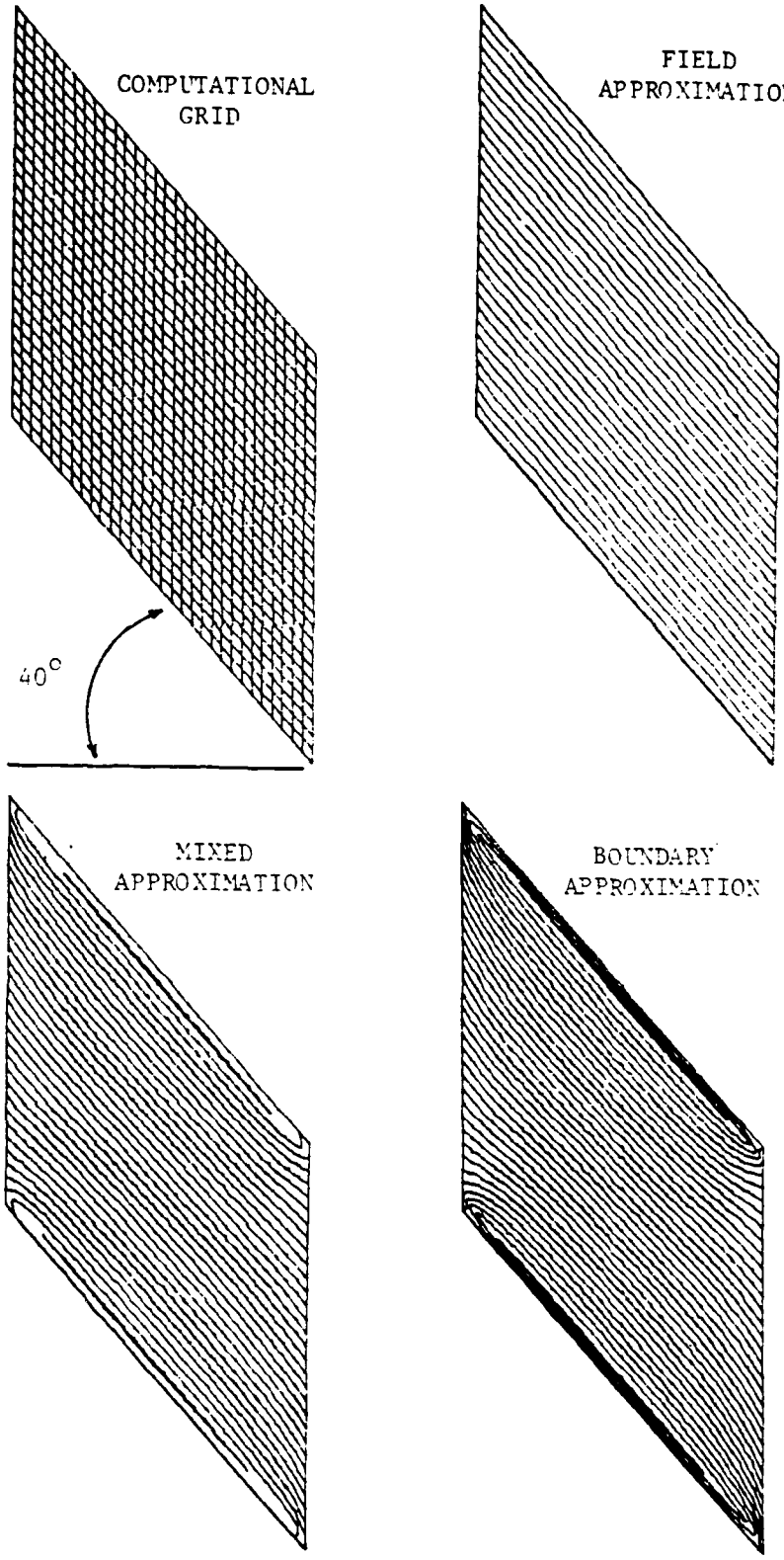


Figure 5. Grid and computed streamlines for uniform irrotational flow through 40-degree parallelogram.

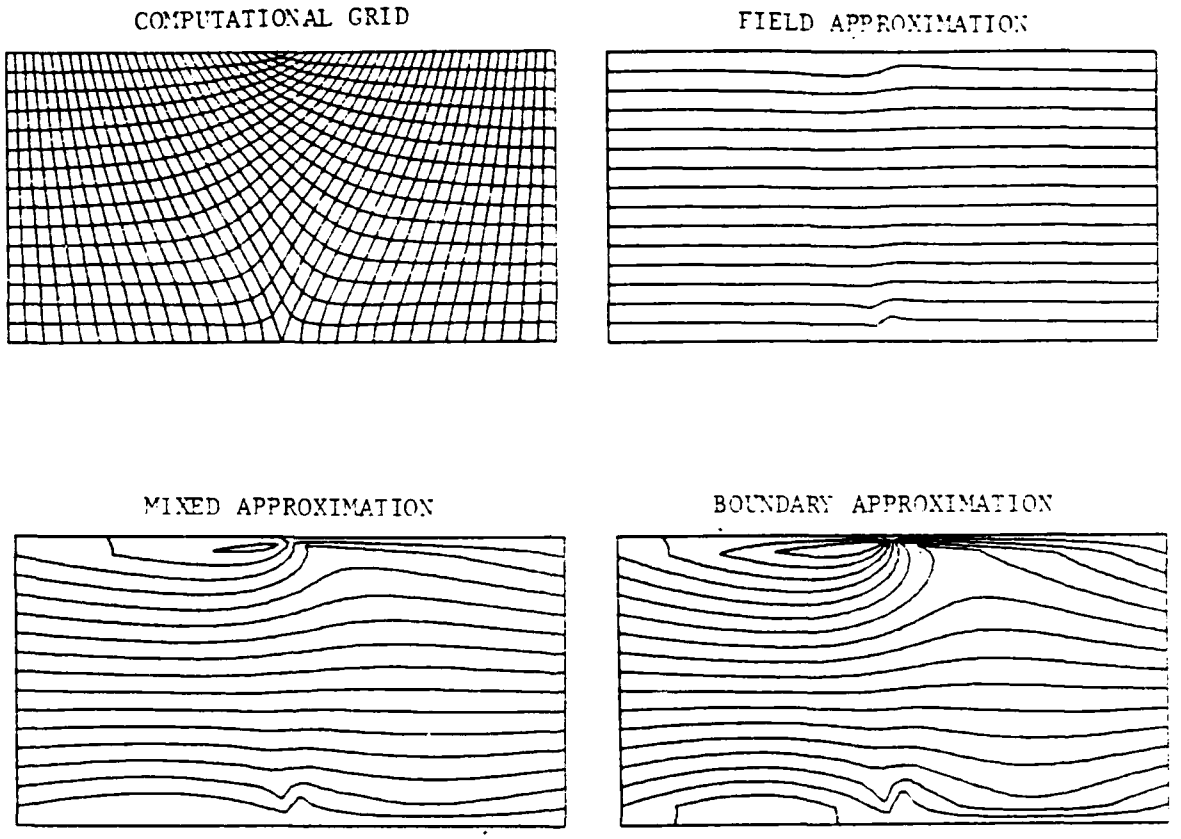


Figure 6. Grid and computed streamlines for uniform irrotational flow through rectangular channel.

$$\varepsilon = \frac{|U_\xi + V_\eta|}{|U| + |V|} \tag{32}$$

The last test involves a grid with non-uniformity as well as skewness, shown in Figure 6. In this case the flow field is rectangular in the physical (x,y) space, but L-shaped in the computational (ξ,η) space. No one would actually use such a distorted grid for serious computation, but it serves our needs in that it allows us to observe directly the grid-induced error in a flow where we know the exact solution in advance (u = 1). Moreover, the large continuity violation associated with the initial vertical velocity (v' = -10) is intended to magnify the error. As in the examples for uniform grids, the field approximation generates far better results than the boundary and mixed approximations, but there is still some distortion of the streamlines even with the field approximation.

IV. CONCLUSION. Three alternatives have been proposed for representing ambiguous derivatives in the pressure gradient on non-orthogonal grid cells adjacent to flow field boundaries. For the test cases presented herein, the best results were obtained with the field approximation; that is, by replacing the ambiguous derivative on a cell face by its value on the adjacent cell corner lying in the field (rather than on the boundary). Using this approach, the discrete pressure gradient remains irrotational for uniform skew grids, creating no spurious vorticity whatsoever. The presence of non-uniformity and skewness together, however, can generate grid-induced vorticity even with the field approximation. Thus, while non-orthogonal staggered grids can indeed be used for computing incompressible flow, it is advisable to keep the grid as smooth as possible in the presence of skewness.

ACKNOWLEDGMENT

This work was funded by the Repair, Evaluation, Maintenance and Rehabilitation (REMR) Research Program sponsored by the Office, Chief of Engineers, US Army.

REFERENCES

[1] Chorin, A. J. 1967. "A Numerical Method for Solving Incompressible Viscous Flow Problems", Journal of Computational Physics, Vol. 2, pp 12-26.

[2] Harlow, F. H., and Welch, J. E. 1965. "Numerical Calculation of Time-Dependent Viscous Incompressible Flow of Fluid with Free Surface", Physics of Fluids, Vol. 8, No. 12, pp 2182-2189.

[3] Anderson, D. A., Tannehill, J. C., and Pletcher, R. H. 1984. Computational Fluid Mechanics and Heat Transfer, pp 252-255. Hemisphere Publishing, McGraw-Hill, New York.

[4] Thompson, J. F., Warsi, Z. U. A., and Mastin, C. W. 1985. Numerical Grid Generation: Foundations and Applications, pp 15-16, Elsevier Science Publishing, North-Holland, New York.

A RANDOMNESS PROPERTY OF m-SEQUENCES

Harold Fredricksen
Mathematics Department
Naval Postgraduate School
Monterey, CA 93940

and

Gary Krahn
Mathematics Department
United States Military Academy
West Point, NY 10996

ABSTRACT. Maximal length linear shift register sequences (m-sequences) are used in a number of communications applications. Their nearly ideal randomness properties are what make these sequences so employable. In this note we discuss an additional randomness property that m-sequences possess.

I. INTRODUCTION. Maximal length shift register sequences (m-sequences) have been used in a myriad of applications for high-speed communications. The nearly ideal randomness properties of m-sequences is the primary reason for their extensive applicability. The balance and run properties are intrinsic in other sequences such as full sequences of length 2^n [1]. However, it is the correlation property (or shift-and-add property) which makes m-sequences useful for spread spectrum [2]-[3], synchronization [4], range-radar [5], error correction [6]-[7], random number generation [8] and other applications. For additional properties of m-sequences, applications and methods for their generation, see [9]-[10].

In this note we discuss another randomness property which is also possessed

by m-sequences. Actually this property is equivalent to the shift-and-add property, though not obviously so.

2. A RANDOMNESS PROPERTY. Suppose a pair of distinct elements (A,B) is drawn at random from the set $T = \{ 1,2,3\ldots,2^n-2 \}$ where $A < B$. We seek the expected value of A and B.

The number of ways of selecting any pair of numbers, without replacement, from a set of (2^n-2) elements is $[(2^n-2)(2^n-3)/2] = C(2^n-2,2)$ i.e. the combination of 2^n-2 objects taken 2 at a time. If A is the smaller of the two integers and is equal to i, then there are (2^n-2-i) possible choices for the larger integer B. Since each pair (A,B) is equally likely to be selected, the probability that A equals the integer i is given by

$$Pr[A=i] = \frac{2^n-2-i}{[(2^n-2)(2^n-3)/2]}$$

The expected value of A is the weighted average of the possible values that A can take on. The expected value of A is thus given by

$$E[A] = \sum_{i=1}^{2^n-3} i * Pr[A=i]$$

$$= \sum_{i=1}^{2^n-3} i * (2^n-2-i)/M$$

where $M = [(2^n-2)(2^n-3)/2]$

$$= 1/M \sum_{i=1}^{2^n-3} i * (2^n-2-i)$$

$$= 1/M \left[(2^n-2) \sum_{i=1}^{2^n-3} i - \sum_{i=1}^{2^n-3} i^2 \right]$$

$$= 1/M\ [(2^n-2)^2\ (2^n-3)/2$$

$$-\ (2^n-3)(2^n-2)(2^{n+1}-5)/6]$$

$$= (2^n-2)\ -\ (2^{n+1}-5)/3$$

$$= (2^n-1)/3.$$

The probability that (B=j) is given by

$$Pr[B=j] = (j-1)/M$$

The expected value of B is then

$$E[B] = \sum_{j=2}^{2^n-2} j\ *\ Pr[B=j]$$

$$= \sum_{j=2}^{2^n-2} j\ *\ (j-1)/M$$

$$= 1/M \sum_{j=2}^{2^n-2} j\ *\ (j-1)$$

$$= 1/M\ [(2^n-2)(2^n-1)(2^{n+1}-3)/6$$

$$-\ 1\ -\ (2^n-2)(2^n-1)/2\ +\ 1]$$

$$= (2^n-1)(2^{n+1}-3)/[3\ *\ (2^n-3)]$$

$$-\ (2^n-1)/(2^n-3)$$

$$= (2^n-1)(2^{n+1}-6)/[3\ *\ (2^n-3)]$$

$$= 2\ *\ (2^n-1)/3.$$

Thus, $E[B] = 2\ *\ E[A] = 2\ *\ (2^n-1)/3$ when selecting a pair of numbers (A,B), where (A < B), at random, without replacement, from the set $T = \{1,2,3,\ldots,2^n-2\}$.

3. THE SHIFT-AND ADD PROPERTY. Let S be a sequence of period 2^n-1 generated by a primitive polynomial over GF(2) of degree n. S^j is the sequence formed by cyclically shifting S by j bits to the left. By the shift-and-add property of m-sequences the modulo 2 sum of S and S^j equals S^k ($S \oplus S^j = S^k$) where the shift $k < 2^n-1$ is uniquely determined by the shift j. Let $\{(j_i,k_i)\}$ be the set of all shift-and-add pairs for a primitive polynomial f(x) of degree n over GF(2) where $j_i < k_i < 2^n-1$. Let the set $\{j_i\} = J$ and the set $\{k_i\} = K$. Note: it can be shown that there are $(2^n-2)/2$ distinct shift-and-add pairs for each primitive polynomial of degree n over GF(2) and J and K partition the set $\{1,2,3,\ldots,2^n-2\}$.

Since f(x) generates the sequence

$$S = S(x) = (s_0) + (s_1)x + (s_2)x^2 + \ldots$$

we see that $f(x) \cdot S(x) = 0$. That is f(x) annihilates S(x). Since f(x) primitive, f(x) is the minimal generator of S(x). If $S \oplus S^j \oplus S^k = S(x) + (x^j * S(x)) + (x^k * S(x)) = 0$, then $(1 + x^j + x^k) \cdot S(x) = 0$ and $(1 + x^j + x^k)$ also annihilates S(x). Therefore, f(x) divides every trinomial defined by a shift-and-add pair. Because f(x) is an irreducible polynomial,

$f(x) \nmid x^t$ for any t. Since $f(x) \mid (1 + x^j + x^k)$ then

$f(x) \mid (1 + x^j + x^k)(x^{-j}) = (x^{-j} + 1 + x^{k-j})$ and

$f(x) \mid (1 + x^{k-j} + x^{2^n-1-j})$. Therefore $(k-j, 2^n-1-j)$ is a shift-and-add pair where $(k-j) < (2^n-1-j)$. Hence, $(k-j)$ is an element of J and (2^n-1-j) is an element of K. As a result, the sets $\{k_i-j_i\}$ and $\{j_i\}$ are equal and

$$\sum (k_i - j_i) = \sum j_i$$

from which it follows that

$$\sum k_i = 2 \sum j_i.$$

This is equivalent to the expected statistical result determined above when pairs of numbers are drawn at random without replacement, from the set $\{1,2,3,\ldots,2^n-2\}$. Thus, sequences generated by a primitive polynomial have the randomness property shown by randomly selecting pairs of numbers from a finite set of size 2^n-2.

REFERENCES

1. Fredricksen, H., "A survey of Full Length Nonlinear Shift Register Cycle Algorithms", SAIM Review 24, #2, April 1982, pp.195 - 221.

2. Simon, M. et.al., Spread Spectrum Communications, Computer Science Press, 1985.

3. Dixon, R. C., Spread Spectrum Systems, Wiley, 1976.

4. Stiffler, J. J., The Theory of Synchronous Communications, Prentice Hall, 1971.

5.Golomb, S. W., Digital Communications with Space Applications, Prentice Hall 1964.

6. Peterson, W. W. and Weldon, E. J. Jr., Error Correcting Codes, MIT Press, 1972.

7. Solomon, G. (in Balakrishnam, A.), Communication Theory, McGraw Hill, 1967.

8. Tausworthe, R. C., "Random Numbers Generated by Linear Recurrences Modulo Two", Math. Computation 19, April 1965.

9. Golomb, S. W., Shift Register Sequences, Aegean Park Press, 1982.

10. Selmer, E. S., Linear Recurrence Relations over Finite Fields, Mathematics Department, U. of Bergen, Bergen, Norway, 1964.

THERMODYNAMIC GAUGE THEORY OF SOLIDS AND QUANTUM LIQUIDS WITH INTERNAL PHASE

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. The local gauge invariance of relativistic thermodynamics under phase rotations suggests that bulk matter systems have density and temperature dependent internal phase angles associated with the state functions. A procedure for determining the internal phase angles associated with energy, pressure, entropy, thermodynamic potentials, and gauge parameters of solids and quantum liquids is presented in terms of the renormalization group equations of relativistic thermodynamics. The calculated magnitudes of the thermodynamic state functions depend on the values of the internal phase angles. It is suggested that the external angular momentum of systems may be coupled to the angular momenta associated with the internal space of thermodynamic phase angles. Applications to mechanical waves in matter with internal phase are considered. These effects are expected to be found in high density and pressure systems such as atomic nuclei, neutron stars, nuclear explosions, and the interaction of directed energy beams with matter.

1. INTRODUCTION. The complete understanding of matter and radiation at high densities requires a locally scale and gauge invariant theory of the forces and fields that determine the properties of a physical system.[1,2] The basic forces in a physical system are associated with a local gauge group, as for example the gauge group of the standard model of the strong and electroweak interactions is $SU(3)_C \times SU(2)_L \times U(1)_Y$. For simple systems, such as electromagnetism, the gauge group is U(1) the group of phase rotations.[3] Local gauge symmetry has unified the interactions of nature, and it is only natural because of such success to attempt a similar synthesis in other areas of physics such as thermodynamics and mechanics.

The vacuum state plays an important role in the development of local gauge theories of the four fundamental interactions. It produces observable effects in quantum electrodynamic calculations of the fermion self-energy, vertex modification, and vacuum polarization as manifested in the Lamb shift.[4,5] In quantum flavordynamics the nonzero expectation values of the vacuum Higgs field produces the spontaneous symmetry breaking that gives rise to the massive intermediate vector bosons that mediate the weak interactions.[1] In quantum chromodynamics the vacuum polarization leads to the concepts of a running coupling constant and asymptotic freedom for the non-Abelian gauge theories.[1-3] The question then arises as to whether vacuum effects appear in systems at the macroscopic level, and whether a synthesis of thermodynamics and continuum mechanics can be based on a locally gauge invariant theory that includes the effects of the vacuum state.

As part of a general program to determine the state equation of systems at high densities, a local gauge theory of matter and radiation has been developed that is based on a gauge and scale invariant relativistic trace equation written to include vacuum effects as follows[6]

$$U + T\left(\frac{dU}{dT}\right)_{PV} - 3V\,\frac{d}{dV}(PV)_U = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{1}$$

where U = relativistic (renormalized) internal energy, P = relativistic pressure, T = absolute temperature, V = volume of substance, and U^a and P^a = corresponding nonrelativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic (unrenormalized) calculations. The trace equation (1) can be rewritten as[7,8]

$$\left(1 - b + T\,\frac{\partial}{\partial T} - bV\,\frac{\partial}{\partial V}\right)E - 3\left(1 + \gamma + V\,\frac{\partial}{\partial V} - \gamma T\,\frac{\partial}{\partial T}\right)P \tag{2}$$

$$= \left(1 - b^a + T\,\frac{\partial}{\partial T} - b^aV\,\frac{\partial}{\partial V}\right)E^a$$

Equation (1) can also be written as

$$E + \frac{T}{V}\,C_V - 3(P - K_T) + (T\,\frac{\partial P}{\partial T} - P)(3\gamma - b) = E^a + \frac{T}{V}\,C_V^a - b^a(T\,\frac{\partial P^a}{\partial T} - P^a) \tag{3}$$

where E = relativistic energy density = U/V , E^a = nonrelativistic energy density, and where[7,8]

$$\gamma = \frac{V}{C_V}\left(\frac{\partial P}{\partial T}\right)_V \tag{4}$$

$$b = \frac{T(\partial P/\partial T)_V}{(P - K_T)} \tag{5}$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \tag{6}$$

where γ = Grüneisen parameter, C_V = relativistic heat capacity at constant volume, and C_V^a = nonrelativistic heat capacity at constant volume, given respectively by

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V \tag{7}$$

$$C_V^a = \left(\frac{\partial U^a}{\partial T}\right)_V \tag{8}$$

and where

$$K_T = -V\left(\frac{\partial P}{\partial V}\right)_T \tag{9}$$

$$K_T^a = -V\left(\frac{\partial P^a}{\partial V}\right)_T \tag{10}$$

are the relativistic and nonrelativistic values of the bulk modulus respectively. The parameters b and γ are the gauge parameters of relativistic thermodynamics. Equation (2) can be decoupled into two independent equations by noting that E and P are related by the Gibbs-Helmholtz relation as follows[9]

$$\left(\frac{\partial U}{\partial V}\right)_T = E + V\left(\frac{\partial E}{\partial V}\right)_T = T\left(\frac{\partial P}{\partial T}\right)_V - P \tag{11}$$

With the introduction of a Lagrange undetermined multiplier η, equation (11) can be rewritten as

$$\eta\left(1 + V\frac{\partial}{\partial V}\right)E + \eta\left(1 - T\frac{\partial}{\partial T}\right)P = 0 \tag{12}$$

Using equation (12) allows equation (2) to be decoupled as follows[8]

$$\left(T\frac{\partial}{\partial T} + fV\frac{\partial}{\partial V} + M\right)E = \psi^a \tag{13}$$

$$\left(T\frac{\partial}{\partial T} + hV\frac{\partial}{\partial V} + N\right)P = 0 \tag{14}$$

where

$$f = \eta - b \tag{15}$$

$$h = \frac{1}{\eta/3 - \gamma} \tag{16}$$

$$M = f + 1 \tag{17}$$

$$N = h - 1 \tag{18}$$

$$\psi^a = (T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a) E^a \tag{19}$$

Equation (13) and (14) are the ground state renormalization group equations of relativistic thermodynamics.

For a solid or low temperature quantum system the nonrelativistic, scalar state equation of the ground state is assumed to have the following form[6-8]

$$E^a = E_o^a + E_j^a T^j + \cdots \tag{20}$$

$$P^a = P_o^a + P_j^a T^j + \cdots \tag{21}$$

where E^a and P^a = nonrelativistic energy density and pressure respectively, E_o^a and P_o^a = nonrelativistic zero-temperature values of the energy density and pressure respectively, E_j^a and P_j^a = nonrelativistic thermal coefficients for the energy density and pressure respectively, T = absolute temperature of the system (°K), and j = numerical index having values characteristic of the type of physical system. Note that $U^a = VE^a$ and $U_o^a = VE_o^a$ where U_o^a = zero-temperature value of the unrenormalized internal energy.

A commonly used descriptor of the thermal state equations given by equations (20) and (21) is the nonrelativistic zero-temperature value of the Grüneisen parameter that is defined by[6-8]

$$\gamma_o^a = \frac{P_j^a}{E_j^a} = \frac{1}{(j-1)} \frac{1}{E_j^a} \frac{d}{dV}(VE_j^a) \tag{22}$$

except for j = 1. Here γ_o^a = nonrelativistic zero-temperature value of the Grüneisen parameter, and V = volume of the material system. When j = 1 , $\gamma_o^a = 2/3$. The zero temperature value of the nonrelativistic bulk modulus is given by $K_o^a = n dP_o^a/dn$, where n = N/V = number of moles per unit volume, and N = number of moles of a substance.

The corresponding relativistic scalar state equation will be written as[6-8]

$$E = E_o + E_j T^j + \cdots \tag{23}$$

$$P = P_o + P_j T^j + \cdots \tag{24}$$

$$\gamma_o = \frac{P_j}{E_j} = \frac{1}{(j-1)} \frac{1}{E_j} \frac{d}{dV}(VE_j) \tag{25}$$

except for $j = 1$, when $E_1 = E_1^a$, where E_o and P_o = relativistic zero-temperature energy density and pressure respectively, E_j and P_j = relativistic thermal coefficients for the energy density and pressure respectively, and γ_o = relativistic zero-temperature Grüneisen parameter. The relativistic value of the zero temperature bulk modulus is given by $K_o = ndP_o/dn$. Note that $U_o = VE_o$, where U_o = zero temperature value of the renormalized internal energy. Combining equation (2) with the state equations (20) through (25) yields the following ground state equations[6]

$$E_o - 3[(1 + \gamma_o)P_o - K_o] = E_o^a \tag{26}$$

$$E_j\left(1 + j + \frac{j\gamma_o P_o}{P_o - K_o} - 3V \frac{d\gamma_o}{dV}\right) = E_j^a\left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a}\right) \tag{27}$$

where the internal energy coefficients are given by

$$\frac{E_j^a}{E_j} = \exp\left[(j - 1) \int^V (\gamma_o^a - \gamma_o) \frac{dV}{V}\right] \tag{28}$$

and where P_o, K_o and γ_o = zero temperature values of the relativistic pressure, incompressibility ($-VdP_o/dV$) and Gruneisen parameter respectively, and P_o^a, K_o^a and γ_o^a are the corresponding nonrelativistic values of these quantities. Eqs. (26) and (27) are a set of coupled nonlinear differential equations for P_o and γ_o. Equation (26) is equivalent to[6]

$$3n^2 \frac{d^2E_o}{dn^2} - 3(1 + \gamma_o)n \frac{dE_o}{dn} + (3\gamma_o + 4)E_o = E_o^a \tag{29}$$

The trace equation for radiation in matter can be derived by a perturbation technique applied to equation (2) with the result[7,9]

$$\left(1 - b + T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V}\right)E_r - \beta_r\left(T \frac{\partial P}{\partial T} - P\right) \tag{30}$$

$$- 3\left[\left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right)P_r - \beta_r\left(T \frac{\partial P}{\partial T} - P\right)\right] = \psi_r^a$$

where E_r and P_r = radiation energy density and pressure respectively, and where[7,8]

$$\beta_r = b_r \frac{P_r - K_{Tr}}{P - K_T} - \frac{T \frac{\partial P}{\partial T} (P_r - K_{Tr})}{(P - K_T)^2} \tag{31}$$

$$b_r = \frac{T \frac{\partial P_r}{\partial T}}{P_r - K_{Tr}} \tag{32}$$

$$\delta_r = \frac{\partial E_r / \partial T}{\partial E / \partial T} (\gamma_r - \gamma) \tag{33}$$

$$\gamma_r = \frac{\partial P_r / \partial T}{\partial E_r / \partial T}$$

$$K_{Tr} = - V \frac{\partial P_r}{\partial V} \tag{35}$$

$$\psi_r^a = \left(T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a \right) E_r^a - \beta_r^a \left(T \frac{\partial P^a}{\partial T} - P^a \right) \tag{36}$$

The parameters γ_r and b_r are the two radiation gauge parameters of the thermal medium.

Equation (30) can be separated into two radiation equations each of which is similar in form to the Callan-Symanzik equation. This is done by using the Gibbs-Helmholtz equation which for radiation becomes[9]

$$\frac{\partial U_r}{\partial V} = E_r + V \frac{\partial E_r}{\partial V} = T \frac{\partial P_r}{\partial T} - P_r \tag{37}$$

Introducing a radiation Lagrange multiplier η_r as follows

$$\eta_r \left[E_r + V \frac{\partial E_r}{\partial V} + P_r - T \frac{\partial P_r}{\partial T} \right] = 0 \tag{38}$$

allows equation (30) to be separated as follows[3]

$$(T\frac{\partial}{\partial T} + f_r V\frac{\partial}{\partial V} + M_r)E_r - \beta_r(T\frac{\partial P}{\partial T} - P) = \psi_r^a \tag{39}$$

$$(T\frac{\partial}{\partial T} + h_r V\frac{\partial}{\partial V} + N_r)P_r - h_r\delta_r(T\frac{\partial P}{\partial T} - P) = 0 \tag{40}$$

where

$$f_r = n_r - b \tag{41}$$

$$h_r = (n_r/3 - \gamma)^{-1} \tag{42}$$

$$M_r = f_r + 1 \tag{43}$$

$$N_r = h_r - 1 \tag{44}$$

Local gauge and scale invariance has unified continuum mechanics and thermodynamics.[6-8] In particular, it has been shown that local scale invariance for thermodynamics requires the introduction of two gauge parameters b and γ which must be determined simultaneously with the energy density. It has been shown that the Lie group $e^{\pm\phi}$ is the scale invariance group of relativistic thermodynamics that is based on a trace equation.[7] The group U(1) of phase rotations $e^{\pm i\phi}$ is the gauge invariance group of this theory.[8]

The invariance of the trace equation under scale transformations of the form $P \to P' = Pe^{\pm\phi}$ and $E \to E' = Ee^{\pm\psi}$, and under phase rotations of the form $\overline{P} \to \overline{P}' = \overline{P}e^{\pm i\phi}$, $\overline{E} \to \overline{E}' = \overline{E}e^{\pm i\psi}$, $\overline{\gamma} \to \overline{\gamma}' = \overline{\gamma}e^{iZ}$, and $\overline{b} \to \overline{b}' = \overline{b}e^{iW}$ leads to the renormalization group equations of relativistic thermodynamics.[7,8] For instance, gauge invariance for phase rotations yields the following renormalization group equations for the complex gauge parameters[8]

$$\left(\frac{d\overline{\gamma}}{d\overline{P}}\right)^{\pm} = \frac{\gamma\frac{T}{\phi}\frac{\partial\phi}{\partial T} - \frac{V}{\phi}\frac{\partial\phi}{\partial V}}{P - T\frac{\partial P}{\partial T} \mp iPT\frac{\partial\phi}{\partial T}} \tag{45}$$

$$\left(\frac{d\overline{b}}{d\overline{E}}\right)^{\pm} = \frac{\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}}{E + V\frac{\partial E}{\partial V} \pm iEV\frac{\partial\psi}{\partial V}} \tag{46}$$

where P and E are the magnitudes of the pressure and energy density respectively. Symmetrization gives the following results for gauge invariance under phase rotations[8]

$$\frac{d\bar{\gamma}}{d\bar{P}} = \frac{1}{2}\left[\left(\frac{d\bar{\gamma}}{d\bar{P}}\right)^{-} + \left(\frac{d\bar{\gamma}}{d\bar{P}}\right)^{+}\right] = \frac{\gamma}{P}\frac{dZ}{d\phi} = \frac{\left(P - T\frac{\partial P}{\partial T}\right)\left(\gamma\frac{T}{\phi}\frac{\partial\phi}{\partial T} - \frac{V}{\phi}\frac{\partial\phi}{\partial V}\right)}{\left(P - T\frac{\partial P}{\partial T}\right)^2 + P^2\left(T\frac{\partial\phi}{\partial T}\right)^2} \tag{47}$$

$$\frac{d\bar{b}}{d\bar{E}} = \frac{1}{2}\left[\left(\frac{d\bar{b}}{d\bar{E}}\right)^{-} + \left(\frac{d\bar{b}}{d\bar{E}}\right)^{+}\right] = \frac{b}{E}\frac{dW}{d\psi} = \frac{\left(E + V\frac{\partial E}{\partial V}\right)\left(\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}\right)}{\left(E + V\frac{\partial E}{\partial V}\right)^2 + E^2\left(V\frac{\partial\psi}{\partial V}\right)^2} \tag{48}$$

Similar equations result from the scale invariance condition which gives $d\gamma/dP$ and db/dE for changes in the magnitudes of the pressure and energy density.[7,8]

These results suggest that the pressure, energy density and the gauge paramaters may themselves be intrinsically complex numbers that are associated with internal phase angles. Accordingly the relativistic trace equation (1) will be written as

$$\bar{U} + T\left(\frac{d\bar{U}}{dT}\right)_{\bar{P}V} - 3V\frac{d}{dV}(\bar{P}V)_{\bar{U}} = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{49}$$

or equivalently as

$$\left(1 - \bar{b} + T\frac{\partial}{\partial T} - \bar{b}V\frac{\partial}{\partial V}\right)\bar{E} - 3\left(1 + \bar{\gamma} + V\frac{\partial}{\partial V} - \bar{\gamma}T\frac{\partial}{\partial T}\right)\bar{P} = \psi^a \tag{50}$$

where $\bar{U}$, $\bar{E}$, $\bar{P}$, $\bar{\gamma}$, and $\bar{b}$ are complex number representations of the internal energy, energy density, pressure, and the gauge parameters. The corresponding equation for radiation in matter with internal phases is derived from equation (50) to be

$$\left(1 - \bar{b} + T\frac{\partial}{\partial T} - \bar{b}V\frac{\partial}{\partial V}\right)\bar{E}_r - \bar{\beta}_r\left(T\frac{\partial\bar{P}}{\partial T} - \bar{P}\right) \tag{51}$$

$$- 3\left[\left(1 + \bar{\gamma} + V\frac{\partial}{\partial V} - \bar{\gamma}T\frac{\partial}{\partial T}\right)\bar{P}_r - \bar{\delta}_r\left(T\frac{\partial\bar{P}}{\partial T} - \bar{P}\right)\right] = \psi^a_r$$

where $\bar{E}_r$, $\bar{P}_r$, $\bar{\beta}_r$, and $\bar{\delta}_r$ are the complex number generalizations of the functions that appear in equations (30) through (36).

This paper presents a theory of the relativistic thermodynamics of solids and quantum liquids with internal phase. The renormalization group equations for systems with internal phase are derived, and a procedure for solving the complex number relativistic trace equation is presented that allows the determination of the internal phase angles associated with the pressure, energy density, and gauge parameters. The non-zero values of the phase angles represent a spontaneously broken symmetry.

2. THERMODYNAMIC STATE FUNCTIONS FOR SYSTEMS WITH INTERNAL PHASE. In order to solve the complex number trace equation (50) it is first necessary to determine the relations between the complex thermodynamic state functions and to determine their connection to the internal phase angles. This will be done using the first and second laws of thermodynamics. The complex number thermodynamic state functions that appear in equations (49) and (50) will be written in terms of their internal phase angles as follows

$$\overline{U} = Ue^{i\theta_u} \tag{52}$$

$$\overline{E} = \overline{U}/V = Ee^{i\theta_u} \tag{53}$$

$$\overline{P} = Pe^{i\theta_p} \tag{54}$$

$$\overline{\gamma} = \gamma e^{i\theta_\gamma} \tag{55}$$

$$\overline{b} = be^{i\theta_b} \tag{56}$$

where θ_u, θ_p, θ_γ, and θ_b = internal phase angles of the internal energy, pressure, Grüneisen parameter, and b gauge parameter respectively. In addition the complex number entropy will be written as

$$\overline{S} = Se^{i\theta_s} \tag{57}$$

where θ_s = internal phase angle of the entropy. In general all of the phase angles are functions of V and T. The quantities U, E, P, γ, b, and S are the magnitudes of the complex thermodynamic state functions, and are also functions of V and T.

The complex number bulk modulus is obtained from equation (54) as follows

$$\overline{K}_T = -V\left(\frac{\partial \overline{P}}{\partial V}\right)_T = K_T e^{i\theta_K} = -e^{i\theta_p}\left(V\frac{\partial P}{\partial V} + iPV\frac{\partial \theta_p}{\partial V}\right) \tag{58}$$

$$= e^{i\theta_p}\left(n\frac{\partial P}{\partial n} + iPn\frac{\partial \theta_p}{\partial n}\right)$$

where

$$K_T = \sqrt{\left(V\frac{\partial P}{\partial V}\right)^2 + P^2\left(V\frac{\partial \theta_p}{\partial V}\right)^2} \tag{59}$$

$$K_T \cos\omega = n\,\partial P/\partial n \tag{59A}$$

$$K_T \sin\omega = Pn\,\partial\theta_p/\partial n \tag{59B}$$

$$\theta_K = \theta_p + \omega \tag{60}$$

$$\tan \omega = \frac{Pn \frac{\partial \theta_p}{\partial n}}{n \frac{\partial P}{\partial n}} \tag{61}$$

Equation (52) immediately gives the complex number heat capacity as

$$\overline{C}_V = \left(\frac{\partial \overline{U}}{\partial T}\right)_V = C_V e^{i\theta C_V} \tag{62}$$

where

$$C_V = \sqrt{\left(\frac{\partial U}{\partial T}\right)^2 + U^2\left(\frac{\partial \theta_u}{\partial T}\right)^2} \tag{63}$$

$$C_V \cos \rho = \partial U/\partial T \tag{63A}$$

$$C_V \sin \rho = U\, \partial \theta_u/\partial T \tag{63B}$$

$$\theta_{C_V} = \theta_u + \rho \tag{64}$$

$$\tan \rho = \frac{U \frac{\partial \theta_u}{\partial T}}{\frac{\partial U}{\partial T}} = \frac{T \frac{\partial \theta_u}{\partial T}}{\frac{T}{U}\frac{\partial U}{\partial T}} \tag{65}$$

Thus the renormalized values of C_V and K_T include the effects of the internal phase angles θ_u and θ_p respectively.

The relationships between the various state functions and their internal phase angles are determined from the First and Second laws of thermodynamics which can be written for matter and radiation with internal phase angles as follows

$$T d\overline{S} = T e^{i\theta_s}(dS + iS d\theta_s) = d\overline{U} + \overline{P} dV \tag{66}$$

or equivalently as

$$T \frac{\partial \overline{S}}{\partial V} = \frac{\partial \overline{U}}{\partial V} + \overline{P} \tag{67}$$

$$T \frac{\partial \overline{S}}{\partial T} = \frac{\partial \overline{U}}{\partial T} \tag{68}$$

Combining equations (52), (53) and (57) with equations (67) and (68), and separating into real and imaginary parts, yields the following equations

$$T\left(\cos\theta_s \frac{\partial S}{\partial V} - S \sin\theta_s \frac{\partial \theta_s}{\partial V}\right) = \cos\theta_u \frac{\partial U}{\partial V} + P\cos\theta_p - U\sin\theta_u \frac{\partial \theta_u}{\partial V} \tag{69}$$

$$T\left(\sin\theta_s \frac{\partial S}{\partial V} + S\cos\theta_s \frac{\partial \theta_s}{\partial V}\right) = \sin\theta_u \frac{\partial U}{\partial V} + P\sin\theta_p + U\cos\theta_u \frac{\partial \theta_u}{\partial V} \tag{70}$$

$$T\left(\cos\theta_s \frac{\partial S}{\partial T} - S \sin\theta_s \frac{\partial \theta_s}{\partial T}\right) = \cos\theta_u \frac{\partial U}{\partial T} - U\sin\theta_u \frac{\partial \theta_u}{\partial T} \tag{71}$$

$$T\left(\sin\theta_s \frac{\partial S}{\partial T} + S\cos\theta_s \frac{\partial \theta_s}{\partial T}\right) = \sin\theta_u \frac{\partial U}{\partial T} + U\cos\theta_u \frac{\partial \theta_u}{\partial T} \tag{72}$$

Squaring and adding equations (69) and (70) gives

$$T^2\left[\left(\frac{\partial S}{\partial V}\right)^2 + S^2\left(\frac{\partial \theta_s}{\partial V}\right)^2\right] = \left(\frac{\partial U}{\partial V}\right)^2 + P^2 + U^2\left(\frac{\partial \theta_u}{\partial V}\right)^2 + 2P\frac{\partial U}{\partial V}\cos(\theta_u - \theta_p) \tag{73}$$

$$+ 2PU\frac{\partial \theta_u}{\partial V}\sin(\theta_p - \theta_u)$$

Squaring and adding equations (71) and (72) gives

$$T^2\left[\left(\frac{\partial S}{\partial T}\right)^2 + S^2\left(\frac{\partial \theta_s}{\partial T}\right)^2\right] = \left(\frac{\partial U}{\partial T}\right)^2 + U^2\left(\frac{\partial \theta_u}{\partial T}\right)^2 \tag{74}$$

The Gibbs-Helmholtz equation for matter with internal phase is written as

$$\frac{\partial \bar{U}}{\partial V} = T\frac{\partial \bar{P}}{\partial T} - \bar{P} \tag{75}$$

Using the Gibbs-Helmholtz equation allows Equation (67) to be rewritten as[9]

$$\frac{\partial \bar{S}}{\partial V} = \frac{\partial \bar{P}}{\partial T} \tag{76}$$

which allows equations (69) and (70) to be rewritten as

$$\cos\theta_s \frac{\partial S}{\partial V} - S\sin\theta_s \frac{\partial \theta_s}{\partial V} = \cos\theta_p \frac{\partial P}{\partial T} - P\sin\theta_p \frac{\partial \theta_p}{\partial T} \tag{77}$$

$$\sin\theta_s \frac{\partial S}{\partial V} + S\cos\theta_s \frac{\partial \theta_s}{\partial V} = \sin\theta_p \frac{\partial P}{\partial T} + P\cos\theta_p \frac{\partial \theta_p}{\partial T} \tag{78}$$

Squaring and adding equations (77) and (78) gives

$$\left(\frac{\partial S}{\partial V}\right)^2 + S^2\left(\frac{\partial \theta_s}{\partial V}\right)^2 = \left(\frac{\partial P}{\partial T}\right)^2 + P^2\left(\frac{\partial \theta_p}{\partial T}\right)^2 \tag{79}$$

Also, from Maxwell's relationship it follows that[9]

$$T = \left(\frac{\partial \bar{U}}{\partial \bar{S}}\right)_V \tag{80}$$

From which it follows that

$$T = \frac{\partial U}{\partial S}\cos\left(\theta_u - \theta_s\right) - U\frac{\partial \theta_u}{\partial S}\sin\left(\theta_u - \theta_s\right) \tag{81}$$

$$TS\frac{\partial \theta_s}{\partial S} = \frac{\partial U}{\partial S}\sin\left(\theta_u - \theta_s\right) + U\frac{\partial \theta_u}{\partial S}\cos\left(\theta_u - \theta_s\right) \tag{82}$$

and

$$T^2 + T^2S^2\left(\frac{\partial \theta_s}{\partial S}\right)^2 = \left(\frac{\partial U}{\partial S}\right)^2 + U^2\left(\frac{\partial \theta_u}{\partial S}\right)^2 \tag{83}$$

The Gibbs-Helmholtz equation (75) can be separated into real and imaginary components as follows

$$\cos\theta_u\frac{\partial U}{\partial V} - U\sin\theta_u\frac{\partial \theta_u}{\partial V} = \cos\theta_p\left(T\frac{\partial P}{\partial T} - P\right) - TP\sin\theta_p\frac{\partial \theta_p}{\partial T} \tag{84}$$

$$\sin\theta_u\frac{\partial U}{\partial V} + U\cos\theta_u\frac{\partial \theta_u}{\partial V} = \sin\theta_p\left(T\frac{\partial P}{\partial T} - P\right) + TP\cos\theta_p\frac{\partial \theta_p}{\partial T} \tag{85}$$

Equations (84) and (85) can be rewritten in terms of the energy density as follows

$$\left(\cos\theta_u - \sin\theta_u V\frac{\partial \theta_u}{\partial V} + \cos\theta_u V\frac{\partial}{\partial V}\right)E + \left(\cos\theta_p + \sin\theta_p T\frac{\partial \theta_p}{\partial T} - \cos\theta_p T\frac{\partial}{\partial T}\right)P = 0 \tag{86}$$

$$\left(\sin\theta_u + \cos\theta_u V\frac{\partial \theta_u}{\partial V} + \sin\theta_u V\frac{\partial}{\partial V}\right)E + \left(\sin\theta_p - \cos\theta_p T\frac{\partial \theta_p}{\partial T} - \sin\theta_p T\frac{\partial}{\partial T}\right)P = 0 \tag{87}$$

Squaring and adding equations (84) and (85) gives

$$\left(\frac{\partial U}{\partial V}\right)^2 + U^2\left(\frac{\partial \theta_u}{\partial V}\right)^2 = \left(T\,\frac{\partial P}{\partial T} - P\right)^2 + P^2\left(T\,\frac{\partial \theta_p}{\partial T}\right)^2 \tag{88}$$

Combining equations (73) and (79) gives

$$T^2\left[\left(\frac{\partial P}{\partial T}\right)^2 + P^2\left(\frac{\partial \theta_p}{\partial T}\right)^2\right] = P^2 + \left(\frac{\partial U}{\partial V}\right)^2 + U^2\left(\frac{\partial \theta_u}{\partial V}\right)^2 + 2P\,\frac{\partial U}{\partial V}\cos\left(\theta_u - \theta_p\right) \tag{89}$$

$$+ 2PU\,\frac{\partial \theta_u}{\partial V}\sin\left(\theta_p - \theta_u\right)$$

Expanding the right hand side of equation (88) and using equation (89) gives the following simple result

$$T\,\frac{\partial P}{\partial T} - P = \frac{\partial U}{\partial V}\cos\left(\theta_p - \theta_u\right) + U\,\frac{\partial \theta_u}{\partial V}\sin\left(\theta_p - \theta_u\right) \tag{90}$$

Similarly

$$P^2\left(T\,\frac{\partial \theta_p}{\partial T}\right)^2 = \left(\frac{\partial U}{\partial V}\right)^2 \sin^2\left(\theta_p - \theta_u\right) + U^2\left(\frac{\partial \theta_u}{\partial V}\right)^2 \cos^2\left(\theta_p - \theta_u\right) \tag{91}$$

$$- 2U\,\frac{\partial U}{\partial V}\,\frac{\partial \theta_u}{\partial V}\cos\left(\theta_p - \theta_u\right)\sin\left(\theta_p - \theta_u\right)$$

The three basic thermodynamic potentials will now be considered.

The enthalpy of a substance with internal phase is written as

$$\overline{H} = He^{i\theta_H} = \overline{U} + \overline{PV} \tag{92}$$

where $\overline{H}$ = complex number enthalpy, H = enthalpy magnitude, and θ_H = internal phase angle of the enthalpy. Combining equation (92) with equations (52) and (53) gives

$$H^2 = \left(U\cos\theta_u + PV\cos\theta_p\right)^2 + \left(U\sin\theta_u + PV\sin\theta_p\right)^2 \tag{93}$$

$$= U^2 + P^2V^2 + 2UPV\cos\left(\theta_u - \theta_p\right)$$

$$\tan \theta_H = \frac{U \sin \theta_u + PV \sin \theta_p}{U \cos \theta_u + PV \cos \theta_p} \tag{94}$$

The differential of the vector enthalpy is

$$d\bar{H} = e^{i\theta_H}(dH + iHd\theta_H) = Td\bar{S} + Vd\bar{P} \tag{95}$$

which yields

$$\cos \theta_H \frac{\partial H}{\partial S} - H \sin \theta_H \frac{\partial \theta_H}{\partial S} = T \cos \theta_s - TS \sin \theta_s \frac{\partial \theta_s}{\partial S} - PV \sin \theta_p \frac{\partial \theta_p}{\partial S} \tag{96}$$

$$\sin \theta_H \frac{\partial H}{\partial S} + H \cos \theta_H \frac{\partial \theta_H}{\partial S} = T \sin \theta_s + TS \cos \theta_s \frac{\partial \theta_s}{\partial S} + PV \cos \theta_p \frac{\partial \theta_p}{\partial S} \tag{97}$$

$$\cos \theta_H \frac{\partial H}{\partial P} - H \sin \theta_H \frac{\partial \theta_H}{\partial P} = V \cos \theta_p - TS \sin \theta_s \frac{\partial \theta_s}{\partial P} - PV \sin \theta_p \frac{\partial \theta_p}{\partial P} \tag{98}$$

$$\sin \theta_H \frac{\partial H}{\partial P} + H \cos \theta_H \frac{\partial \theta_H}{\partial P} = V \sin \theta_p + TS \cos \theta_s \frac{\partial \theta_s}{\partial P} + PV \cos \theta_p \frac{\partial \theta_p}{\partial P} \tag{99}$$

where S and P are taken to be the two independent variables. Combining equations (96) and (97) gives

$$\left(\frac{\partial H}{\partial S}\right)^2 + H^2\left(\frac{\partial \theta_H}{\partial S}\right)^2 = T^2 + T^2S^2\left(\frac{\partial \theta_s}{\partial S}\right)^2 + P^2V^2\left(\frac{\partial \theta_p}{\partial S}\right)^2 \tag{100}$$

$$+ 2PVT \frac{\partial \theta_p}{\partial S} \sin (\theta_s - \theta_p) + 2TSPV \frac{\partial \theta_s}{\partial S} \frac{\partial \theta_p}{\partial S} \cos (\theta_s - \theta_p)$$

while combining equations (98) and (99) gives

$$\left(\frac{\partial H}{\partial P}\right)^2 + H^2\left(\frac{\partial \theta_H}{\partial P}\right)^2 = V^2 + T^2S^2\left(\frac{\partial \theta_s}{\partial P}\right)^2 + P^2V^2\left(\frac{\partial \theta_p}{\partial P}\right)^2 \tag{101}$$

$$+ 2TSV \frac{\partial \theta_s}{\partial P} \sin (\theta_p - \theta_s) + 2TSPV \frac{\partial \theta_s}{\partial P} \frac{\partial \theta_p}{\partial P} \cos (\theta_s - \theta_p)$$

The complex free energy is written as

$$\overline{A} = Ae^{i\theta_A} = \overline{U} - T\overline{S} \tag{102}$$

where $\overline{A}$ = complex number free energy, A = magnitude of the free energy, and θ_A = internal phase angle of the free energy. Combining equations (52) and (57) with equation (102) yields

$$A^2 = (U \cos \theta_u - TS \cos \theta_s)^2 + (U \sin \theta_u - TS \sin \theta_s)^2 \tag{103}$$

$$= U^2 + T^2S^2 - 2UTS \cos (\theta_u - \theta_s)$$

$$\tan \theta_A = \frac{U \sin \theta_u - TS \sin \theta_s}{U \cos \theta_u - TS \cos \theta_s} \tag{104}$$

The differential of the vector free energy in equation (102) is

$$d\overline{A} = e^{i\theta_A}(dA + iAd\theta_A) = - \overline{P}dV - \overline{S}dT \tag{105}$$

from which it follows that

$$\cos \theta_A \frac{\partial A}{\partial V} - A \sin \theta_A \frac{\partial \theta_A}{\partial V} = - P \cos \theta_p \tag{106}$$

$$\sin \theta_A \frac{\partial A}{\partial V} + A \cos \theta_A \frac{\partial \theta_A}{\partial V} = - P \sin \theta_p \tag{107}$$

$$\cos \theta_A \frac{\partial A}{\partial T} - A \sin \theta_A \frac{\partial \theta_A}{\partial T} = - S \cos \theta_s \tag{108}$$

$$\sin \theta_A \frac{\partial A}{\partial T} + A \cos \theta_A \frac{\partial \theta_A}{\partial T} = - S \sin \theta_s \tag{109}$$

and

$$\left(\frac{\partial A}{\partial V}\right)_T^2 + A^2\left(\frac{\partial \theta_A}{\partial V}\right)_T^2 = P^2 \tag{110}$$

$$\left(\frac{\partial A}{\partial T}\right)_V^2 + A^2\left(\frac{\partial \theta_A}{\partial T}\right)_V^2 = S^2 \tag{111}$$

The complex number form of the Gibbs-Helmholtz equation for the free energy is written as[9]

$$\overline{A} - T\left(\frac{\partial \overline{A}}{\partial T}\right)_V = \overline{U} \tag{112}$$

which gives immediately

$$\cos\theta_A\left(A - T\frac{\partial A}{\partial T}\right) + A\sin\theta_A T\frac{\partial \theta_A}{\partial T} = U\cos\theta_u \tag{113}$$

$$\sin\theta_A\left(A - T\frac{\partial A}{\partial T}\right) - A\cos\theta_A T\frac{\partial \theta_A}{\partial T} = U\sin\theta_u \tag{114}$$

and

$$\left(A - T\frac{\partial A}{\partial T}\right)^2 + A^2\left(T\frac{\partial \theta_A}{\partial T}\right)^2 = U^2 \tag{115}$$

Combining equations (103), (111) and (115) gives

$$A\frac{\partial A}{\partial T} = S[TS - U\cos(\theta_u - \theta_s)] \tag{116}$$

The complex number form of the Gibbs free energy is given by

$$\overline{G} = Ge^{i\theta_G} = \overline{U} + \overline{PV} - \overline{TS} = \overline{A} + \overline{PV} \tag{117}$$

where $\overline{G}$ = complex number Gibbs free energy, G = magnitude of the Gibbs free energy, and θ_G = internal phase angle of the Gibbs free energy. It follows immediately from equation (117) that

$$G\cos\theta_G = U\cos\theta_u + PV\cos\theta_p - TS\cos\theta_s \tag{118}$$

$$G\sin\theta_G = U\sin\theta_u + PV\sin\theta_p - TS\sin\theta_s \tag{119}$$

which gives

$$G^2 = U^2 + P^2V^2 + T^2S^2 + 2UPV\cos(\theta_p - \theta_u) - 2UTS\cos(\theta_s - \theta_u) \tag{120}$$

$$- 2PVTS\cos(\theta_s - \theta_p)$$

and

$$\tan \theta_G = \frac{U \sin \theta_u + PV \sin \theta_p - TS \sin \theta_s}{U \cos \theta_u + PV \cos \theta_p - TS \cos \theta_s} \tag{121}$$

The differential of the vector Gibbs function in equation (117) is

$$d\bar{G} = e^{i\theta_G}(dG + iGd\theta_G) = -\bar{S}dT + Vd\bar{P} \tag{122}$$

from which it follows that

$$\cos \theta_G \frac{\partial G}{\partial T} - G \sin \theta_G \frac{\partial \theta_G}{\partial T} = -S \cos \theta_s - PV \sin \theta_p \frac{\partial \theta_p}{\partial T} \tag{123}$$

$$\sin \theta_G \frac{\partial G}{\partial T} + G \cos \theta_G \frac{\partial \theta_G}{\partial T} = -S \sin \theta_s + PV \cos \theta_p \frac{\partial \theta_p}{\partial T} \tag{124}$$

$$\cos \theta_G \frac{\partial G}{\partial P} - G \sin \theta_G \frac{\partial \theta_G}{\partial P} = V \cos \theta_p - PV \sin \theta_p \frac{\partial \theta_p}{\partial P} \tag{125}$$

$$\sin \theta_G \frac{\partial G}{\partial P} + G \cos \theta_G \frac{\partial \theta_G}{\partial P} = V \sin \theta_p + PV \cos \theta_p \frac{\partial \theta_p}{\partial P} \tag{126}$$

Combining equation (123) and (124) gives

$$\left(\frac{\partial G}{\partial T}\right)^2 + G^2\left(\frac{\partial \theta_G}{\partial T}\right)^2 = S^2 + 2PVS \sin (\theta_p - \theta_s) \frac{\partial \theta_p}{\partial T} + P^2V^2\left(\frac{\partial \theta_p}{\partial T}\right)^2 \tag{127}$$

while equations (125) and (126) give

$$\left(\frac{\partial G}{\partial P}\right)^2 + G^2\left(\frac{\partial \theta_G}{\partial P}\right)^2 = V^2 + P^2V^2\left(\frac{\partial \theta_p}{\partial P}\right)^2 \tag{128}$$

Further relationships can be obtained from the vector form of the Gibbs-Helmholtz equation for the Gibbs function which is written as

$$\bar{A} = \bar{G} - \bar{P}V = \bar{G} - \bar{P}\left(\frac{\partial \bar{G}}{\partial \bar{P}}\right)_T \tag{129}$$

When the T = 0 limit exists (as in the case of solids and quantum liquids) for matter with internal phase, the following equations corresponding to equations (52) through (57) can be written

$$\overline{U}_o = U_o e^{i\theta_u^o} \tag{130}$$

$$\overline{E}_o = \overline{U}_o/V = E_o e^{i\theta_u^o} \tag{131}$$

$$\overline{P}_o = P_o e^{i\theta_p^o} \tag{132}$$

$$\overline{\gamma}_o = \gamma_o e^{i\theta_\gamma^o} \tag{133}$$

$$\overline{b}_o = 0 \tag{134}$$

$$\overline{S}_o = 0 \tag{135}$$

where θ_u^o , θ_p^o , and θ_γ^o are the T = 0 values of θ_u , θ_p , and θ_γ respectively. Note from the definition of b in equation (5) it follows that $b_o = 0$ so that $\overline{b}_o = 0$ also, however it will be shown later that $\theta_b^o \neq 0$. From the Third law of thermodynamics it follows that $S_o = 0$ and $\overline{S}_o = 0$.

For solids and quantum liquids one can take the T = 0 limit of equations (84) and (85) and get

$$\cos\theta_u^o \frac{dU_o}{dV} - U_o \frac{d\theta_u^o}{dV} \sin\theta_u^o = -P_o \cos\theta_p^o \tag{136}$$

$$\sin\theta_u^o \frac{dU_o}{dV} + U_o \frac{d\theta_u^o}{dV} \cos\theta_u^o = -P_o \sin\theta_p^o \tag{137}$$

From equations (136) and (137) one gets immediately

$$\tan\theta_p^o = \frac{\sin\theta_u^o \frac{dU_o}{dV} + U_o \frac{d\theta_u^o}{dV} \cos\theta_u^o}{\cos\theta_u^o \frac{dU_o}{dV} - U_o \frac{d\theta_u^o}{dV} \sin\theta_u^o} \tag{138}$$

$$P_o = \sqrt{\left(\frac{dU_o}{dV}\right)^2 + U_o^2\left(\frac{d\theta_u^o}{dV}\right)^2} \tag{139}$$

$$= \sqrt{\left(n\,\frac{dE_o}{dn} - E_o\right)^2 + E_o^2\left(n\,\frac{d\theta_u^o}{dn}\right)^2}$$

Using the trigonometrical formula for the tangent of the sum of two angles allows equation (138) to be rewritten as

$$\theta_p^o = \theta_u^o + \phi_o \tag{140}$$

where

$$\tan\phi^o = \frac{U_o\,\frac{d\theta_u^o}{dV}}{\frac{dU_o}{dV}} = \frac{U_o\,\frac{d\theta_u^o}{dn}}{\frac{dU_o}{dn}} = \frac{E_o n\,\frac{d\theta_u^o}{dn}}{n\,\frac{dE_o}{dn} - E_o} \tag{141}$$

$$P_o \cos\phi_o = n\,dE_o/dn - E_o \tag{141A}$$

$$P_o \sin\phi_o = E_o n\,d\theta_u^o/dn \tag{141B}$$

From equation (141) it follows that

$$\phi^o = 0 \qquad \text{when } \frac{d\theta_u^o}{dn} = 0 \tag{142A}$$

$$\phi^o = \pm\frac{\pi}{2} \qquad \text{when } \frac{dU_o}{dn} = 0 \tag{142B}$$

In general for a T = 0 system

$$\theta_p^o = \theta_u^o + \tan^{-1}\left(\frac{E_o n\,\frac{d\theta_u^o}{dn}}{n\,\frac{dE_o}{dn} - E_o}\right) \tag{143}$$

Figure 1 shows the density dependence of θ_p^o and θ_u^o for an unbound interacting system such as a neutron gas. The two possible signs that occur in equation (142B) arise in systems having a saturation density at which $dU_o/dn = 0$, according to the signs of dU_o/dn and $d\theta_u^o/dn$ as $dU_o/dn \to 0$ as seen in Figure 2. Figure 2 shows the density dependence of θ_p^o and θ_u^o for a system such as N = Z infinite nuclear matter which is bound at a saturation density. These figures

show the effects of equations (140) through (143). In general $\theta_p^o < \theta_u^o$ in the high density limit of bound or unbound quantum systems. From equations (58) and (59) it follows that

$$\overline{K}_o = n \frac{d\overline{P}_o}{dn} = K_o e^{i(\theta_p^o + \omega_o)} \tag{144}$$

where

$$K_o = \sqrt{\left(n \frac{dP_o}{dn}\right)^2 + P_o^2\left(n \frac{d\theta_p^o}{dn}\right)^2} \tag{145}$$

$$\tan \omega_o = P_o(d\theta_p^o/dn)/(dP_o/dn) \tag{146}$$

$$K_o \cos \omega_o = n\, dP_o/dn \tag{146A}$$

$$K_o \sin \omega_o = P_o n\, d\theta_p^o/dn \tag{146B}$$

The complex number analogs of the scalar thermal state equations given in equations (23) and (24) are

$$\overline{E} = \overline{E}_o + \overline{E}_j T^j = E_o e^{i\theta_u^o} + E_j e^{i\theta_u^j} T^j = E e^{i\theta_u} \tag{147}$$

$$\overline{P} = \overline{P}_o + \overline{P}_j T^j = P_o e^{i\theta_p^o} + P_j e^{i\theta_p^j} T^j = P e^{i\theta_p} \tag{148}$$

where E_j and P_j = magnitudes of the thermal components of the energy and pressure respectively, and θ_u^j and θ_p^j = phase angles of the thermal components of the energy density and pressure respectively. From equations (147) and (148) it follows immediately that

$$E^2 = E_o^2 + 2E_oE_j \cos (\theta_u^o - \theta_u^j)T^j + E_j^2 T^{2j} \tag{149}$$

$$P^2 = P_o^2 + 2P_oP_j \cos (\theta_p^o - \theta_p^j)T^j + P_j^2 T^{2j} \tag{150}$$

$$\tan \theta_u = \frac{E_o \sin \theta_u^o + E_j \sin \theta_u^j T^j}{E_o \cos \theta_u^o + E_j \cos \theta_u^j T^j} \tag{151}$$

$$\tan \theta_p = \frac{P_o \sin \theta_p^o + P_j \sin \theta_p^j T^j}{P_o \cos \theta_p^o + P_j \cos \theta_p^j T^j} \tag{152}$$

Note that $\bar{U}_o = V\bar{E}_o$ and $\bar{U}_j = V\bar{E}_j$.

3. GAUGE PARAMETERS FOR SYSTEMS WITH INTERNAL PHASE. The two gauge parameters that appear in the basic trace equation (50) are $\bar{\gamma}$ and $\bar{b}$. The complex number Grüneisen parameter is defined as

$$\bar{\gamma} = \frac{V}{\bar{C}_V} \frac{\partial \bar{P}}{\partial T} = \frac{\partial \bar{P}/\partial T}{\partial \bar{E}/\partial T} = \gamma e^{i\theta_\gamma} \tag{153}$$

where γ and θ_γ = magnitude and phase of the Grüneisen parameter respectively. Combining equations (53) and (54) with equation (153) gives

$$\bar{\gamma} = e^{i(\theta_p - \theta_u)} \left(\frac{\frac{\partial P}{\partial T} + iP \frac{\partial \theta_p}{\partial T}}{\frac{\partial E}{\partial T} + iE \frac{\partial \theta_u}{\partial T}} \right) \tag{154}$$

$$= \sqrt{\frac{\left(\frac{\partial P}{\partial T}\right)^2 + P^2\left(\frac{\partial \theta_p}{\partial T}\right)^2}{\left(\frac{\partial E}{\partial T}\right)^2 + E^2\left(\frac{\partial \theta_u}{\partial T}\right)^2}} \; e^{i(\theta_p - \theta_u + \mu - \rho)} \tag{155}$$

where ρ and μ are given by

$$\tan \rho = \frac{E \frac{\partial \theta_u}{\partial T}}{\frac{\partial E}{\partial T}} \tag{156}$$

$$\tan \mu = \frac{P \frac{\partial \theta_p}{\partial T}}{\frac{\partial P}{\partial T}} \tag{157}$$

Comparing equations (153) and (155) gives

$$\theta_\gamma = \theta_p - \theta_u + \mu - \rho \tag{158}$$

$$\gamma = \frac{\partial P/\partial T}{\partial E/\partial T}\frac{\cos\rho}{\cos\mu} = \sqrt{\frac{(\partial P/\partial T)^2 + P^2(\partial\theta_p/\partial T)^2}{(\partial E/\partial T)^2 + E^2(\partial\theta_u/\partial T)^2}} \tag{159}$$

The T = 0 limit of equations (156) and (157) can be obtained by noting that from equations (149) and (150) it follows that

$$\left(\frac{\partial E}{\partial T}\right)_{T\to o} = jE_j \cos(\theta_u^j - \theta_u^o)\, T^{j-1} \tag{160}$$

$$\left(\frac{\partial P}{\partial T}\right)_{T\to o} = jP_j \cos(\theta_p^j - \theta_p^o)\, T^{j-1} \tag{161}$$

and from equation (151) and (152) it follows that

$$\left(E\frac{\partial\theta_u}{\partial T}\right)_{T\to o} = jE_j \sin(\theta_u^j - \theta_u^o)\, T^{j-1} \tag{162}$$

$$\left(P\frac{\partial\theta_p}{\partial T}\right)_{T\to o} = jP_j \sin(\theta_p^j - \theta_p^o)\, T^{j-1} \tag{163}$$

It then follows from equations (156) and (157) and (160) through (163) that

$$\rho_o = \theta_u^j - \theta_u^o \tag{164}$$

$$\mu_o = \theta_p^j - \theta_p^o \tag{165}$$

and therefore from equation (158) it follows that

$$\theta_\gamma^o = \theta_p^j - \theta_u^j \tag{166}$$

Finally combining equation (159) with (160) through (163) gives

$$\gamma_o = \frac{P_j}{E_j} \tag{167}$$

which is the same form as in equation (25) for the scalar thermal state equation. The results in equations (166) and (167) can be obtained directly from

the T = 0 limit of equation (153) by using the complex number thermal state equation (147) to get

$$\bar{\gamma}_o = \frac{\bar{P}_j}{\bar{E}_j} = \frac{P_j}{E_j} e^{i(\theta_p^j - \theta_u^j)} \tag{168}$$

The gauge parameter $\bar{b}$ is defined as follows

$$\bar{b} = \frac{T \frac{\partial \bar{P}}{\partial T}}{\bar{P} - \bar{K}_T} = be^{i\theta_b} \tag{169}$$

where $\bar{K}_T$ is defined in equation (58). Equation (169) can be rewritten as

$$\bar{b} = \frac{T \frac{\partial P}{\partial T} + iPT \frac{\partial \theta_p}{\partial T}}{P + V \frac{\partial P}{\partial V} + iPV \frac{\partial \theta_p}{\partial V}} = \frac{T \frac{\partial P}{\partial T} + iPT \frac{\partial \theta_p}{\partial T}}{P - n \frac{\partial P}{\partial n} - iPn \frac{\partial \theta_p}{\partial n}} \tag{170}$$

$$= \sqrt{\frac{\left(T \frac{\partial P}{\partial T}\right)^2 + P^2 \left(T \frac{\partial \theta_p}{\partial T}\right)^2}{\left(P + V \frac{\partial P}{\partial V}\right)^2 + P^2 \left(V \frac{\partial \theta_p}{\partial V}\right)^2}} e^{i(\mu+\chi)} \tag{171}$$

where μ is given by equation (157) and χ is given by

$$\tan \chi = \frac{Pn \frac{\partial \theta_p}{\partial n}}{P - n \frac{\partial P}{\partial n}} \tag{172}$$

Comparing equations (169) and (171) gives

$$\theta_b = \mu + \chi \tag{173}$$

$$b = \sqrt{\frac{(T\partial P/\partial T)^2 + P^2 (T\partial \theta_p/\partial T)^2}{(P - n\partial P/\partial n)^2 + P^2 (n\partial \theta_p/\partial n)^2}} = \left| \frac{T\partial P/\partial T}{P - n\partial P/\partial n} \right| \frac{\sec \mu}{\sec \chi} \tag{174}$$

The T = 0 limit of equation (173) is

$$\theta_b^o = \mu^o + \chi^o = \theta_p^j - \theta_p^o + \chi^o \tag{175}$$

where

$$\tan \chi^o = \frac{P_o n \frac{d\theta_p^o}{dn}}{P_o - n \frac{dP_p}{dn}} \tag{176}$$

while the T = 0 limit of b is obtained from equation (174) to be b = 0.

In the past two sections the relationships between the various phase angles and amplitudes of the thermodynamic state functions have been presented. In the next section a method of calculating the phase angles and amplitudes will be presented.

4. RENORMALIZATION GROUP EQUATIONS FOR THE GROUND STATE OF PHASE MATTER.

The phase angles and magnitudes of the complex number thermodynamic state functions are calculated from the solution of the vector renormalization group equation (50). Combining equation (50) with equations (52) through (56) gives

$$\left(1 - be^{i\theta_b} + T\frac{\partial}{\partial T} - be^{i\theta_b} V\frac{\partial}{\partial V}\right)Ee^{i\theta_u} - 3\left(1 + \gamma e^{i\theta_\gamma} + V\frac{\partial}{\partial V} - \gamma e^{i\theta_\gamma} T\frac{\partial}{\partial T}\right)Pe^{i\theta_p} = \psi^a \tag{177}$$

where

$$\psi^a = \left(1 - b^a + T\frac{\partial}{\partial T} - b^a V\frac{\partial}{\partial V}\right)E^a \tag{178}$$

Equation (177) can be separated into real and imaginary parts. The real part is given by

$$\cos\theta_u\left(1 - b\cos\theta_b + T\frac{\partial}{\partial T} - b\cos\theta_b V\frac{\partial}{\partial V} + b\sin\theta_b V\frac{\partial\theta_u}{\partial V}\right)E \tag{179}$$

$$- \sin\theta_u\left(-b\sin\theta_b + T\frac{\partial\theta_u}{\partial T} - b\sin\theta_b V\frac{\partial}{\partial V} - b\cos\theta_b V\frac{\partial\theta_u}{\partial V}\right)E$$

$$- 3\cos\theta_p\left(1 + \gamma\cos\theta_\gamma + V\frac{\partial}{\partial V} - \gamma\cos\theta_\gamma T\frac{\partial}{\partial T} + \gamma\sin\theta_\gamma T\frac{\partial\theta_p}{\partial T}\right)P$$

$$+ 3\sin\theta_p\left(\gamma\sin\theta_\gamma + V\frac{\partial\theta_p}{\partial V} - \gamma\sin\theta_\gamma T\frac{\partial}{\partial T} - \gamma\cos\theta_\gamma T\frac{\partial\theta_p}{\partial T}\right)P = \psi^a$$

The imaginary part of equation (177) is written as

$$\sin\theta_u\left(1 - b\cos\theta_b + T\frac{\partial}{\partial T} - b\cos\theta_b V\frac{\partial}{\partial V} + b\sin\theta_b V\frac{\partial\theta_u}{\partial V}\right)E \tag{180}$$

$$+ \cos\theta_u\left(-b\sin\theta_b + T\frac{\partial\theta_u}{\partial T} - b\sin\theta_b V\frac{\partial}{\partial V} - b\cos\theta_b V\frac{\partial\theta_u}{\partial V}\right)E$$

$$- 3\sin\theta_p\left(1 + \gamma\cos\theta_\gamma + V\frac{\partial}{\partial V} - \gamma\cos\theta_\gamma T\frac{\partial}{\partial T} + \gamma\sin\theta_\gamma T\frac{\partial\theta_p}{\partial T}\right)P$$

$$- 3\cos\theta_p\left(\gamma\sin\theta_\gamma + V\frac{\partial\theta_p}{\partial V} - \gamma\sin\theta_\gamma T\frac{\partial}{\partial T} - \gamma\cos\theta_\gamma T\frac{\partial\theta_p}{\partial T}\right)P = 0$$

Equations (179) and (180) can be further simplified by introducing equations (86) and (87) and their two corresponding Lagrange indeterminate multipliers η and τ to get

$$\eta\left(\cos\theta_u - \sin\theta_u V\frac{\partial\theta_u}{\partial V} + \cos\theta_u V\frac{\partial}{\partial V}\right)E \tag{181}$$

$$+ \eta\left(\cos\theta_p + \sin\theta_p T\frac{\partial\theta_p}{\partial T} - \cos\theta_p T\frac{\partial}{\partial T}\right)P = 0$$

$$\tau\left(\sin\theta_u + \cos\theta_u V\frac{\partial\theta_u}{\partial V} + \sin\theta_u V\frac{\partial}{\partial V}\right)E \tag{182}$$

$$+ \tau\left(\sin\theta_p - \cos\theta_p T\frac{\partial\theta_p}{\partial T} - \sin\theta_p T\frac{\partial}{\partial T}\right)P = 0$$

Combining equation (179) and (180) with the constraints in equations (181) and (182) gives the following four independent partial differential equations

$$(\ell T\frac{\partial}{\partial T} + fV\frac{\partial}{\partial V} + M)E = \psi^a \tag{183}$$

$$(mT\frac{\partial}{\partial T} + qV\frac{\partial}{\partial V} + R)E = 0 \tag{184}$$

$$(wT\frac{\partial}{\partial T} + xV\frac{\partial}{\partial V} + Y)P = 0 \tag{185}$$

$$(sT\frac{\partial}{\partial T} + zV\frac{\partial}{\partial V} + I)P = 0 \tag{186}$$

where

$$\ell = \cos \theta_u \tag{187}$$

$$f = \cos \theta_u(\eta - b \cos \theta_b) + b \sin \theta_u \sin \theta_b \tag{188}$$

$$M = \cos \theta_u\left[1 + \eta - b \cos \theta_b + b \sin \theta_b V\frac{\partial \theta_u}{\partial V}\right] \tag{189}$$

$$- \sin \theta_u\left[- b \sin \theta_b + T \frac{\partial \theta_u}{\partial T} + (\eta - b \cos \theta_b)V \frac{\partial \theta_u}{\partial V}\right]$$

$$m = \sin \theta_u \tag{190}$$

$$q = \sin \theta_u(\tau - b \cos \theta_b) - b \cos \theta_u \sin \theta_b \tag{191}$$

$$R = \sin \theta_u\left[1 + \tau - b \cos \theta_b + b \sin \theta_b V\frac{\partial \theta_u}{\partial V}\right] \tag{192}$$

$$+ \cos \theta_u\left[- b \sin \theta_b + T \frac{\partial \theta_u}{\partial T} + (\tau - b \cos \theta_b)V \frac{\partial \theta_u}{\partial V}\right]$$

$$w = \cos \theta_p\left(\frac{\eta}{3} - \gamma \cos \theta_\gamma\right) + \gamma \sin \theta_p \sin \theta_\gamma \tag{193}$$

$$x = \cos \theta_p$$

$$Y = \cos \theta_p\left[1 - \frac{\eta}{3} + \gamma \cos \theta_\gamma + \gamma \sin \theta_\gamma T\frac{\partial \theta_p}{\partial T}\right] \tag{195}$$

$$- \sin \theta_p\left[\gamma \sin \theta_\gamma + V \frac{\partial \theta_p}{\partial V} + \left(\frac{\eta}{3} - \gamma \cos \theta_\gamma\right)T \frac{\partial \theta_p}{\partial T}\right]$$

$$s = \sin \theta_p\left(\frac{\tau}{3} - \gamma \cos \theta_\gamma\right) - \gamma \cos \theta_p \sin \theta_\gamma \tag{196}$$

$$z = \sin \theta_p \tag{197}$$

$$I = \sin \theta_p\left[1 - \frac{\tau}{3} + \gamma \cos \theta_\gamma + \gamma \sin \theta_\gamma T\frac{\partial \theta_p}{\partial T}\right] \tag{198}$$

$$+ \cos \theta_p\left[\gamma \sin \theta_\gamma + V \frac{\partial \theta_p}{\partial V} + \left(\frac{\tau}{3} - \gamma \cos \theta_\gamma\right)T \frac{\partial \theta_p}{\partial T}\right]$$

In the limit of $\theta_i \to 0$ one has

$$\ell = 1 \qquad m = 0 \qquad w = \frac{\eta}{3} - \gamma \qquad s = 0 \tag{199}$$

$$f = \eta - b \qquad q = 0 \qquad x = 1 \qquad z = 0$$

$$M = 1 + \eta - b \qquad R = 0 \qquad Y = 1 - \frac{\eta}{3} + \gamma \qquad I = 0$$

which agree with equations (13) through (18).

Equations (183) through (186) are the renormalization group equations that describe relativistic thermodynamic systems having internal phase angles. There are ten unknown quantities in equations (183) through (186): E , θ_u , P , θ_p , γ , θ_γ , b , θ_b , η , τ . The ten equations required to determine these quantities are: the four renormalization group equations (183) through (186), the two equations (158) and (159) that define $\overline{\gamma}$, the two equations (173) and (174) that define $\overline{b}$, and the two constraint equations (181) and (182). Equations (183) through (186) can be derived from a Lagrangian formalism in a manner similar to that in the accompanying paper.

5. RENORMALIZATION GROUP EQUATIONS FOR RADIATION IN PHASE MATTER. In a manner similar to equations (52) through (56) the state functions and gauge parameters for radiation that appear in equation (51) are written as

$$\overline{U}_r = U_r e^{i\theta_{ur}} \tag{200}$$

$$\overline{E}_r = \overline{U}_r/V = E_r e^{i\theta_{ur}} \tag{201}$$

$$\overline{P}_r = P_r e^{i\theta_{pr}} \tag{202}$$

$$\overline{\gamma}_r = \gamma_r e^{i\theta_{\gamma r}} \tag{203}$$

$$\overline{b}_r = b_r e^{i\theta_{br}} \tag{204}$$

$$\overline{\delta}_r = \delta_r e^{i\theta_{\delta r}} \tag{205}$$

$$\overline{\beta}_r = \beta_r e^{i\theta_{\beta r}} \tag{206}$$

where θ_{ur} = internal phase angle of the radiation internal energy

θ_{pr} = internal phase angle of the radiation pressure

$\theta_{\gamma r}$ = internal phase angle of the radiation Grüneisen gauge parameter

θ_{br} = internal phase angle of the radiation b_r gauge parameter

$\theta_{\delta r}$ = internal phase angle of δ_r gauge function

$\theta_{\beta r}$ = internal phase angle of β_r gauge function

In general all of the phase angles and magnitudes that appear in equations (200) through (206) are functions of V and T. Also all of the equations that appear in Sections 2. and 3. are also valid for radiation and can be carried over into the present calculation by adding a subscript "r".

The complex number form of the functions that appear in equation (51) can be written in a form analogous to that in equations (31) through (35) as follows

$$\bar{\beta}_r = \bar{b}_r \frac{\bar{P}_r - \bar{K}_{Tr}}{\bar{P} - \bar{K}_T} - \frac{T \frac{\partial \bar{P}}{\partial T} (\bar{P}_r - \bar{K}_{Tr})}{(\bar{P} - \bar{K}_T)^2} \tag{207}$$

$$\bar{b}_r = \frac{T \frac{\partial \bar{P}_r}{\partial T}}{\bar{P}_r - \bar{K}_{Tr}} \tag{208}$$

$$\bar{\delta}_r = \frac{\partial \bar{E}_r / \partial T}{\partial \bar{E} / \partial T} (\bar{\gamma}_r - \bar{\gamma}) \tag{209}$$

$$\bar{\gamma}_r = \frac{\partial \bar{P}_r / \partial T}{\partial \bar{E}_r / \partial T} \tag{210}$$

$$\bar{K}_{Tr} = -V \frac{\partial \bar{P}_r}{\partial V} \tag{211}$$

Combining equations (210) and (201) through (203) gives

$$\gamma_r = \sqrt{\frac{\left(\frac{\partial P_r}{\partial T}\right)^2 + P_r^2 \left(\frac{\partial \theta_{pr}}{\partial T}\right)^2}{\left(\frac{\partial E_r}{\partial T}\right)^2 + E_r^2 \left(\frac{\partial \theta_{ur}}{\partial T}\right)^2}} = \frac{\partial P_r / \partial T}{\partial E_r / \partial T} \frac{\sec \mu_r}{\sec \rho_r} \tag{212}$$

$$\theta_{\gamma r} = \theta_{pr} - \theta_{ur} + \mu_r - \rho_r \tag{213}$$

where

$$\tan \mu_r = \frac{P_r \dfrac{\partial \theta_{pr}}{\partial T}}{\dfrac{\partial P_r}{\partial T}} \tag{214}$$

$$\tan \rho_r = \frac{E_r \dfrac{\partial \theta_{ur}}{\partial T}}{\dfrac{\partial E_r}{\partial T}} \tag{215}$$

Combining equations (174), (204), and (208) gives

$$b_r = \sqrt{\frac{\left(T \dfrac{\partial P_r}{\partial T}\right)^2 + P_r^2 \left(T \dfrac{\partial \theta_{pr}}{\partial T}\right)^2}{\left(P_r + V \dfrac{\partial P_r}{\partial V}\right)^2 + P_r^2 \left(V \dfrac{\partial \theta_{pr}}{\partial V}\right)^2}} = \left| \frac{T \dfrac{\partial P_r}{\partial T}}{P_r + V \dfrac{\partial P_r}{\partial V}} \right| \frac{\sec \mu_r}{\sec \chi_r} \tag{216}$$

$$\theta_{br} = \mu_r + \chi_r \tag{217}$$

where

$$\tan \chi_r = - \frac{P_r V \dfrac{\partial \theta_{pr}}{\partial V}}{P_r + V \dfrac{\partial P_r}{\partial V}} = \frac{P_r n \dfrac{\partial \theta_{pr}}{\partial n}}{P_r - n \dfrac{\partial P_r}{\partial n}} \tag{218}$$

Combining equations (201), (53), (203) with (209) gives

$$\delta_r = \sqrt{\frac{\left(\dfrac{\partial E_r}{\partial T}\right)^2 + E_r^2 \left(\dfrac{\partial \theta_{ur}}{\partial T}\right)^2}{\left(\dfrac{\partial E}{\partial T}\right)^2 + E^2 \left(\dfrac{\partial \theta_u}{\partial T}\right)^2}} \sqrt{\gamma_r^2 + \gamma^2 - 2\gamma\gamma_r \cos(\theta_{\gamma r} - \theta_\gamma)} \tag{219}$$

$$\theta_{\delta r} = \theta_{ur} - \theta_u + \rho_r - \rho + \pi_r \tag{220}$$

where

$$\tan \pi_r = \frac{\gamma_r \sin \theta_{\gamma r} - \gamma \sin \theta_\gamma}{\gamma_r \cos \theta_{\gamma r} - \gamma \cos \theta_\gamma} \tag{221}$$

and where ρ is defined in equation (156). Combining equation (206) and (207) gives

$$\bar{\beta}_r = A_r e^{i\psi_r} - B_r e^{i\phi_r} \tag{222}$$

where β_r and $\theta_{\beta r}$ of equation (206) are given as

$$\beta_r = \sqrt{A_r^2 + B_r^2 - 2A_r B_r \cos(\psi_r - \phi_r)} \tag{223}$$

$$\tan \theta_{\beta r} = \frac{A_r \sin \psi_r - B_r \sin \phi_r}{A_r \cos \psi_r - B_r \cos \phi_r} \tag{224}$$

where

$$A_r = b_r \sqrt{\frac{\left(P_r + V \frac{\partial P_r}{\partial V}\right)^2 + P_r^2 \left(V \frac{\partial \theta_{pr}}{\partial V}\right)^2}{\left(P + V \frac{\partial P}{\partial V}\right)^2 + P^2 \left(V \frac{\partial \theta_p}{\partial V}\right)^2}} \tag{225}$$

$$B_r = \frac{\sqrt{\left(T \frac{\partial P}{\partial T}\right)^2 + P^2 \left(T \frac{\partial \theta_p}{\partial T}\right)^2} \sqrt{\left(P_r + V \frac{\partial P_r}{\partial V}\right)^2 + P_r^2 \left(V \frac{\partial \theta_{pr}}{\partial V}\right)^2}}{\left(P + V \frac{\partial P}{\partial V}\right)^2 + P^2 \left(V \frac{\partial \theta_p}{\partial V}\right)^2} \tag{226}$$

$$\psi_r = \theta_{br} + \theta_{pr} - \theta_p - \chi_r + \chi = \theta_{pr} - \theta_p + \mu_r + \chi \tag{227}$$

$$\phi_r = \theta_{pr} - \theta_p + \mu - \chi_r + 2\chi \tag{228}$$

Finally the radiation bulk modulus in equation (211) is given by

$$\bar{K}_{Tr} = e^{i\theta_{pr}} \left(n \frac{\partial P_r}{\partial n} + i P_r n \frac{\partial \theta_{pr}}{\partial n}\right) \tag{229}$$

$$= K_{Tr} e^{i(\theta_{pr} + \varphi_r)}$$

where

$$\tan \omega_r = \frac{P_r n \frac{\partial \theta_{pr}}{\partial n}}{n \frac{\partial P_r}{\partial n}} \tag{230}$$

It remains to show how the radiation equation (51) can be decomposed into four radiation equations which combined with the defining relations in equations (200) through (228) can be used to calculate the eight quantities E_r, θ_{ur}, P_r, θ_{pr}, γ_r, $\theta_{\gamma r}$, b_r, and θ_{br}. First note that equation (51) can be rewritten as

$$\left(1 - be^{i\theta_b} + T\frac{\partial}{\partial T} - be^{i\theta_b} V\frac{\partial}{\partial V}\right)E_r e^{i\theta_{ur}} - \beta_r e^{i\theta_{\beta r}}\left[T\frac{\partial}{\partial T}(Pe^{i\theta_p}) - Pe^{i\theta_p}\right] \tag{231}$$

$$- 3\left\{\left(1 + \gamma e^{i\theta_\gamma} + V\frac{\partial}{\partial V} - \gamma e^{i\theta_\gamma} T\frac{\partial}{\partial T}\right)P_r e^{i\theta_{pr}} - \delta_r e^{i\theta_{\delta r}}\left[T\frac{\partial}{\partial T}(Pe^{i\theta_p}) - Pe^{i\theta_p}\right]\right\} = \gamma_r^a$$

The simplification of equation (231) can be realized by noting that the complex Gibbs-Helmholtz equation for radiation

$$\frac{\partial \overline{U}_r}{\partial V} = T\frac{\partial \overline{P}_r}{\partial T} - \overline{P}_r \tag{232}$$

yields the following two constraint equations similar to equations (181) and (182)

$$\eta_r\left(\cos \theta_{ur} - \sin \theta_{ur} V\frac{\partial \theta_{ur}}{\partial V} + \cos \theta_{ur} V\frac{\partial}{\partial V}\right)E_r \tag{233}$$

$$+ \eta_r\left(\cos \theta_{pr} + \sin \theta_{pr} T\frac{\partial \theta_{pr}}{\partial T} - \cos \theta_{pr} T\frac{\partial}{\partial T}\right)P_r = 0$$

$$\tau_r\left(\sin \theta_{ur} + \cos \theta_{ur} V\frac{\partial \theta_{ur}}{\partial V} + \sin \theta_{ur} V\frac{\partial}{\partial V}\right)E_r \tag{234}$$

$$+ \tau_r\left(\sin \theta_{pr} - \cos \theta_{pr} T\frac{\partial \theta_{pr}}{\partial T} - \sin \theta_{pr} T\frac{\partial}{\partial T}\right)P_r = 0$$

where two radiation Lagrange indeterminate multipliers, η_r and τ_r, are introduced. Separating equation (231) into real and imaginary parts and using the constraint equations (233) and (234) yields the following four independent partial differential equations

$$(\ell_r T \frac{\partial}{\partial T} + f_r V \frac{\partial}{\partial V} + M_r)E_r - \beta_r J \cos(\theta_{\beta r} + \theta_p + \lambda) = \psi_r^a \tag{235}$$

$$(m_r T \frac{\partial}{\partial T} + q_r V \frac{\partial}{\partial V} + R_r)E_r - \beta_r J \sin(\theta_{\beta r} + \theta_p + \lambda) = 0 \tag{236}$$

$$(w_r T \frac{\partial}{\partial T} + x_r V \frac{\partial}{\partial V} + Y_r)P_r - \delta_r J \cos(\theta_{\delta r} + \theta_p + \lambda) = 0 \tag{237}$$

$$(s_r T \frac{\partial}{\partial T} + z_r V \frac{\partial}{\partial V} + I_r)P_r - \delta_r J \sin(\theta_{\delta r} + \theta_p + \lambda) = 0 \tag{238}$$

where β_r , $\theta_{\beta r}$, δ_r , and $\theta_{\delta r}$ are given by equations (223), (224), (219), and (220), and where

$$\tan \lambda = \frac{PT \frac{\partial \theta_p}{\partial T}}{T \frac{\partial P}{\partial T} - P} \tag{239}$$

$$J = \sqrt{(T \frac{\partial P}{\partial T} - P)^2 + P^2\left(T \frac{\partial \theta_p}{\partial T}\right)^2} \tag{239A}$$

and where

$$\ell_r = \cos \theta_{ur} \tag{240}$$

$$f_r = \cos \theta_{ur}(\eta_r - b \cos \theta_b) + b \sin \theta_{ur} \sin \theta_b \tag{241}$$

$$M_r = \cos \theta_{ur}\left(1 + \eta_r - b \cos \theta_b + b \sin \theta_b V \frac{\partial \theta_{ur}}{\partial V}\right) \tag{242}$$

$$- \sin \theta_{ur}\left[- b \sin \theta_b + T \frac{\partial \theta_{ur}}{\partial T} + (\eta_r - b \cos \theta_b)V \frac{\partial \theta_{ur}}{\partial V}\right]$$

$$m_r = \sin \theta_{ur} \tag{243}$$

$$q_r = \sin \theta_{ur}(\tau_r - b \cos \theta_b) - b \cos \theta_{ur} \sin \theta_b \tag{244}$$

$$R_r = \sin\theta_{ur}\left(1 + \tau_r - b\cos\theta_b + b\sin\theta_b V\frac{\partial\theta_{ur}}{\partial V}\right) \tag{245}$$

$$+ \cos\theta_{ur}\left[-b\sin\theta_b + T\frac{\partial\theta_{ur}}{\partial T} + (\tau_r - b\cos\theta_b)V\frac{\partial\theta_{ur}}{\partial V}\right]$$

$$w_r = \cos\theta_{pr}\left(\frac{\eta_r}{3} - \gamma\cos\theta_\gamma\right) + \gamma\sin\theta_{pr}\sin\theta_\gamma \tag{246}$$

$$x_r = \cos\theta_{pr} \tag{247}$$

$$Y_r = \cos\theta_{pr}\left(1 - \frac{\eta_r}{3} + \gamma\cos\theta_\gamma + \gamma\sin\theta_\gamma T\frac{\partial\theta_p}{\partial T}\right) \tag{248}$$

$$- \sin\theta_{pr}\left[\gamma\sin\theta_\gamma + V\frac{\partial\theta_{pr}}{\partial V} + \left(\frac{\eta_r}{3} - \gamma\cos\theta_\gamma\right)T\frac{\partial\theta_{pr}}{\partial T}\right]$$

$$s_r = \sin\theta_{pr}\left(\frac{\tau_r}{3} - \gamma\cos\theta_\gamma\right) - \gamma\cos\theta_{pr}\sin\theta_\gamma \tag{249}$$

$$z_r = \sin\theta_{pr} \tag{250}$$

$$I_r = \sin\theta_{pr}\left(1 - \frac{\tau_r}{3} + \gamma\cos\theta_\gamma + \gamma\sin\theta_\gamma T\frac{\partial\theta_{pr}}{\partial T}\right) \tag{251}$$

$$+ \cos\theta_{pr}\left[\gamma\sin\theta_\gamma + V\frac{\partial\theta_{pr}}{\partial V} + \left(\frac{\tau_r}{3} - \gamma\cos\theta_\gamma\right)T\frac{\partial\theta_{pr}}{\partial T}\right]$$

Equations (235) through (238) are the renormalization group equations for radiation with internal phase angles. Setting $\theta_i = 0$ in equations (240) through (251) reduces equations (235) through (251) to equations (39) through (44). The radiation renormalization group equations (235) through (238) can easily be derived from a Lagrangian formalism in a manner similar to that given in the accompanying paper.

6. GROUND STATE OF SOLIDS AND LOW TEMPERATURE QUANTUM LIQUIDS WITH INTERNAL PHASE. This section considers the calculation of the energy density, pressure, and internal phase angles associated with the relativistic state equation of the form given in equations (147) and (148). The complex number analogs of equations (26) through (28) are written as

$$\bar{E}_o - 3[(1 + \bar{\gamma}_o)\bar{P}_o - \bar{K}_o] = \bar{\varepsilon}_o^a \tag{252}$$

$$\bar{E}_j\left(1 + j + \frac{j\bar{\gamma}_o\bar{P}_o}{\bar{P}_o - \bar{K}_o} + 3n\frac{d\bar{\gamma}_o}{dn}\right) = E_j^a\left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a}\right) \tag{253}$$

$$\frac{\bar{E}_j}{E_j^a} = \exp\left[(j-1)\int(\gamma_o^a - \bar{\gamma}_o)\frac{dn}{n}\right] = \frac{E_j}{E_j^a}e^{i\theta_u^j} \tag{254}$$

where

$$\bar{K}_o = n\frac{d\bar{P}_o}{dn} = e^{i\theta_p^o}\left(n\frac{dP_o}{dn} + iP_o n\frac{d\theta_p^o}{dn}\right) \tag{255}$$

$$= K_o e^{i(\theta_p^o + \omega_o)}$$

where P_o , K_o , and ω_o are given by equations (139), (145), and (146) respectively. Equation (252) can also be written as

$$3n^2\frac{d^2\bar{E}_o}{dn^2} - 3(1 + \bar{\gamma}_o)n\frac{d\bar{E}_o}{dn} + (3\bar{\gamma}_o + 4)\bar{E}_o = E_o^a \tag{256}$$

Equations (252) and (253) and the constraint equations (136) and (137) must be solved for E_o , θ_u^o , P_o , θ_p^o , γ_o , and θ_γ^o . Combining equation (256) with equations (131) and (132) and taking the real and imaginary parts yields the following two equations

$$F_1\cos\theta_u^o + 3G_1\sin\theta_u^o = E_o^a \tag{257}$$

$$F_1\sin\theta_u^o - 3G_1\cos\theta_u^o = 0 \tag{258}$$

where

$$F_1 = 3n^2\frac{d^2E_o}{dn^2} - 3(1 + \gamma_o\cos\theta_\gamma^o)n\frac{dE_o}{dn} \tag{259}$$

$$+\left[4 + 3\gamma_o\left(\cos\theta_\gamma^o + \sin\theta_\gamma^o n\frac{d\theta_u^o}{dn}\right) - 3\left(n\frac{d\theta_u^o}{dn}\right)^2\right]E_o$$

$$G_1 = \left(\gamma_o \sin \theta_\gamma^o - 2n \frac{d\theta_u^o}{dn}\right) n \frac{dE_o}{dn} \tag{260}$$

$$+ \left[(1 + \gamma_o \cos \theta_\gamma^o) n \frac{d\theta_u^o}{dn} - n^2 \frac{d^2\theta_u^o}{dn^2} - \gamma_o \sin \theta_\gamma^o\right] E_o$$

Equation (253) can be written as the following two equations

$$\frac{U_j}{U_j^a}\left[\cos \theta_u^j (1 + j + A + C) - \sin \theta_u^j (B + D)\right] = 1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a} \tag{261}$$

$$\sin \theta_u^j (1 + j + A + C) + \cos \theta_u^j (B + D) = 0 \tag{262}$$

where

$$A = \frac{j\gamma_o P_o \cos(\theta_\gamma^o + \chi_o)}{\sqrt{\left(P_o - n \frac{dP_o}{dn}\right)^2 + P_o^2 \left(n \frac{d\theta_p^o}{dn}\right)^2}} \tag{263}$$

$$B = \frac{j\gamma_o P_o \sin(\theta_\gamma^o + \chi_o)}{\sqrt{\left(P_o - n \frac{dP_o}{dn}\right)^2 + P_o^2 \left(n \frac{d\theta_p^o}{dn}\right)^2}} \tag{264}$$

$$C = 3\left(\cos \theta_\gamma^o n \frac{d\gamma_o}{dn} - \sin \theta_\gamma^o n \frac{d\theta_\gamma^o}{dn}\right) \tag{265}$$

$$D = 3\left(\sin \theta_\gamma^o n \frac{d\gamma_o}{dn} + \cos \theta_\gamma^o n \frac{d\theta_\gamma^o}{dn}\right) \tag{266}$$

where

$$\tan \chi_o = \frac{P_o n \frac{d\theta_p^o}{dn}}{P_o - n \frac{dP_o}{dn}} \tag{267}$$

$$\theta_u^j = -(j-1)\int \gamma_o \sin\theta_\gamma^o \frac{dn}{n} \tag{268}$$

$$\frac{U_j}{U_j^a} = \exp\left[(j-1)\int(\gamma_o^a - \gamma_o \cos\theta_\gamma^o)\frac{dn}{n}\right] \tag{269}$$

The angle θ_u^j enters the calculations through the relation (254)

$$\frac{\bar{U}_j}{U_j^a} = \frac{U_j}{U_j^a} e^{i\theta_u^j} = \exp\left[(j-1)\int(\gamma_o^a - \gamma_o e^{i\theta_\gamma^o})\frac{dn}{n}\right] \tag{270}$$

from which equations (268) and (269) follow immediately. Equivalently one can use

$$\bar{\gamma}_o = \gamma_o e^{i\theta_\gamma^o} = \frac{1}{(j-1)}\frac{V}{\bar{U}_j}\frac{d\bar{U}_j}{dV} = \frac{V\bar{P}_j}{\bar{U}_j} \tag{271}$$

$$= \frac{1}{j-1}\left(\frac{V}{U_j}\frac{dU_j}{dV} + iV\frac{d\theta_u^j}{dV}\right)$$

to obtain

$$\gamma_o = \frac{1}{(j-1)}\sqrt{\left(\frac{V}{U_j}\frac{dU_j}{dV}\right)^2 + \left(V\frac{d\theta_u^j}{dV}\right)^2} = \frac{VP_j}{U_j} \tag{272}$$

$$\tan\theta_\gamma^o = \frac{V\frac{d\theta_u^j}{dV}}{\frac{V}{U_j}\frac{dU_j}{dV}} \tag{273}$$

Note that $\bar{U}_1 = U_1^a$ when $j = 1$. From equation (271) it follows that

$$\gamma_o \sin\theta_\gamma^o = \frac{1}{j-1} V\frac{d\theta_u^j}{dV} = -\frac{1}{j-1} n\frac{d\theta_u^j}{dn} \tag{274}$$

$$\gamma_o \cos \theta_\gamma^o = \frac{1}{j-1} \frac{V}{U_j} \frac{dU_j}{dV} = - \frac{1}{j-1} \frac{n}{U_j} \frac{dU_j}{dn} \tag{275}$$

which immediately give equations (268) and (269) respectively. From equation (271) it also follows immediately that

$$\theta_\gamma^o = \theta_p^j - \theta_u^j \tag{276}$$

$$P_j = \frac{\gamma_o U_j}{V} = \frac{1}{(j-1)} \sqrt{\left(\frac{dU_j}{dV}\right)^2 + U_j^2 \left(\frac{d\theta_u^j}{dV}\right)^2} \tag{277}$$

Note that when $j = 1$: $U_1 = 3/2\ NR = U_1^a$, and $P_1 = \gamma_o 3/2\ nR$. The simultaneous solution of equations (257), (258), (261), and (262) along with the constraint equations (136) and (137) give E_o , θ_u^o , P_o , θ_p^o , γ_o , and θ_γ^o . Then equations (272) and (273) give U_j and θ_u^j ; P_j is obtained from equation (277), and finally θ_p^j is obtained from equation (276). In this way all of the elements of the renormalized state equations (147) and (148) can be determined.

7. EXCITED STATES OF SOLIDS AND LOW TEMPERATURE QUANTUM LIQUIDS WITH INTERNAL PHASE. The complex number state equations for the excited states of solids and low temperature quantum liquids are written in analogy to the ground state equations (147) and (148) as follows

$$\overline{E}_r = \overline{E}_{or} + \overline{E}_{jr} T^j = E_{or} e^{i\theta_{ur}^o} + E_{jr} e^{i\theta_{ur}^j} T^j = E_r e^{i\theta_{ur}} \tag{278}$$

$$\overline{P}_r = \overline{P}_{or} + \overline{P}_{jr} T^j = P_{or} e^{i\theta_{pr}^o} + P_{jr} e^{i\theta_{pr}^j} T^j = P_r e^{i\theta_{pr}} \tag{279}$$

where the T = 0 equivalents of the quantities in equations (200) through (206) are

$$\overline{U}_{or} = U_{or} e^{i\theta_{ur}^o} \tag{280}$$

$$\overline{E}_{or} = \overline{U}_{or}/V = E_{or} e^{i\theta_{ur}^o} \tag{281}$$

$$\overline{P}_{or} = P_{or} e^{i\theta_{pr}^o} \tag{282}$$

$$\overline{\gamma}_{or} = \gamma_{or} e^{i\theta_{\gamma r}^o} \tag{283}$$

$$\bar{b}_{or} = 0 \tag{284}$$

$$\bar{\delta}_{or} = \delta_{or} e^{i\theta^o_{\delta r}} \tag{285}$$

$$\bar{\beta}_{or} = 0 \tag{286}$$

where θ^o_{ur} , θ^o_{pr} , $\theta^o_{\gamma r}$, and $\theta^o_{\delta r}$ are the T = 0 values of θ_{ur} , θ_{pr} , $\theta_{\gamma r}$, and $\theta_{\delta r}$ respectively; and where θ^j_{ur} and θ^j_{pr} are the phase angles associated with the thermal components of the radiation energy and pressure respectively.

The T = 0 and T^j components of the complex number equation (51) are respectively[8]

$$\bar{E}_{or} - 3[(1 + \bar{\gamma}_o)\bar{P}_{or} - \bar{K}_{or}] - 3\frac{\bar{E}_{jr}}{\bar{E}_j}\bar{P}_o(\bar{\gamma}_{or} - \bar{\gamma}_o) = E^a_{or} \tag{287}$$

$$j\bar{E}_j(\bar{\alpha}\bar{K}_{or} - \bar{\beta}\bar{P}_{or}) + \bar{E}_{jr}\bar{S}_{jr} = jE^a_j\left(\alpha^a K^a_{or} - \beta^a P^a_{or}\right) + E^a_{jr}T^a_{jr} \tag{288}$$

where

$$\bar{P}_{or} = n\frac{d\bar{E}_{or}}{dn} - \bar{E}_{or} = e^{i\theta^o_{ur}}\left(n\frac{dE_{or}}{dn} - E_{or} + iE_{or}n\frac{d\theta^o_{ur}}{dn}\right) \tag{289}$$

$$= P_{or}e^{i(\theta^o_{ur} + \phi^o_r)}$$

where

$$P_{or} = \sqrt{\left(\frac{dU_{or}}{dV}\right)^2 + U^2_{or}\left(\frac{d\theta^o_{ur}}{dV}\right)^2} \tag{290}$$

$$= \sqrt{\left(n\frac{dE_{or}}{dn} - E_{or}\right)^2 + E^2_{or}\left(n\frac{d\theta^o_{ur}}{dn}\right)^2}$$

$$\theta^o_{pr} = \theta^o_{ur} + \phi^o_r \tag{291}$$

$$P_{or}\cos\phi^o_r = n\,dE_{or}/dn - E_{or} \tag{291A}$$

$$P_{or}\sin\phi^o_r = E_{or}n\,d\theta^o_{ur}/dn \tag{291B}$$

where

$$\tan \phi_r^o = \frac{U_{or} \dfrac{d\theta_{ur}^o}{dV}}{\dfrac{dU_{or}}{dV}} = \frac{E_{or} n \dfrac{d\theta_{ur}^o}{dn}}{n \dfrac{dE_{or}}{dn} - E_{or}} \tag{292}$$

The T = 0 radiation bulk modulus is given by

$$\bar{K}_{or} = n \frac{d\bar{P}_{or}}{dn} = e^{i\theta_{pr}^o}\left(n \frac{dP_{or}}{dn} + iP_{or} n \frac{d\theta_{pr}^o}{dn}\right) \tag{293}$$

$$= K_{or} e^{i(\theta_{pr}^o + \omega_r^o)}$$

where

$$K_{or} = \sqrt{\left(n \frac{dP_{or}}{dn}\right)^2 + P_{or}^2\left(n \frac{d\theta_{pr}^o}{dn}\right)^2} \tag{294}$$

$$K_{or} \cos \omega_r^o = n\, dP_{or}/dn \tag{294A}$$

$$K_{or} \sin \omega_r^o = P_{or} n\, d\theta_{pr}^o/dn \tag{294B}$$

$$\tan \omega_r^o = P_{or}(d\theta_{pr}^o/dn)/(dP_{or}/dn) \tag{295}$$

The functions $\bar{S}_{jr}$ and T_{jr}^a that appear in equation (288) are given by[8]

$$\bar{S}_{jr} = 1 + j + \frac{j\bar{P}_o\bar{\gamma}_{or}}{\bar{P}_o - \bar{K}_o} + 3n \frac{d\bar{\gamma}_{or}}{dn} - 3(j-1)(\bar{\gamma}_{or} - \bar{\gamma}_o)^2 \tag{296}$$

$$T_{jr}^a = 1 + j + \frac{jP_o^a\gamma_{or}^a}{P_o^a - K_o^a} \tag{297}$$

and where

$$\bar{\alpha} = \frac{\bar{\gamma}_o\bar{P}_o}{(\bar{P}_o - \bar{K}_o)^2} \qquad \alpha^a = \frac{\gamma_o^a P_o^a}{(P_o^a - K_o^a)^2} \tag{298}$$

$$\bar{\beta} = \frac{\bar{\gamma}_o \bar{K}_o}{(\bar{P}_o - \bar{K}_o)^2} \qquad \beta^a = \frac{\gamma_o^a K_o^a}{(P_o^a - K_o^a)^2} \tag{299}$$

In analogy with equations (271) through (275) one has the following relations

$$\bar{\gamma}_{or} = \frac{\bar{P}_{jr}}{\bar{E}_{jr}} = \frac{1}{(j-1)} \frac{V}{\bar{U}_{jr}} \frac{d\bar{U}_{jr}}{dV} = \gamma_{or} e^{i\theta^o_{\gamma r}} \tag{300}$$

$$= \frac{1}{(j-1)} \left(\frac{V}{U_{jr}} \frac{dU_{jr}}{dV} + iV \frac{d\theta^j_{ur}}{dV} \right)$$

$$\gamma_{or} = \frac{1}{(j-1)} \sqrt{\left(\frac{V}{U_{jr}} \frac{dU_{jr}}{dV} \right)^2 + \left(V \frac{d\theta^j_{ur}}{dV} \right)^2} \tag{301}$$

$$\tan \theta^o_{\gamma r} = \frac{V \dfrac{d\theta^j_{ur}}{dV}}{\dfrac{V}{U_{jr}} \dfrac{dU_{jr}}{dV}} \tag{302}$$

$$\gamma_{or} \sin \theta^o_{\gamma r} = \frac{1}{j-1} V \frac{d\theta^j_{ur}}{dV} \tag{303}$$

$$\gamma_{or} \cos \theta^o_{\gamma r} = \frac{1}{j-1} \frac{V}{U_{jr}} \frac{dU_{jr}}{dV} \tag{304}$$

$$\theta^o_{\gamma r} = \theta^j_{pr} - \theta^j_{ur} \tag{305}$$

$$P_{jr} = \frac{\gamma_{or} U_{jr}}{V} = \frac{1}{j-1} \sqrt{\left(\frac{dU_{jr}}{dV} \right)^2 + U_{jr}^2 \left(\frac{d\theta^j_{ur}}{dV} \right)^2} \tag{306}$$

When $j = 1$: $E_{1r} = E^a_{1r}$ and $P_{1r} = \gamma_{or} E_{1r}$. Integration of (303) and (304) yields

$$\theta_{ur}^{j} = -(j - 1)\int \gamma_{or} \sin \theta_{\gamma r}^{o} \frac{dn}{n} \tag{307}$$

$$\frac{E_{jr}}{E_{jr}^{a}} = \exp\left[(j - 1)\int (\gamma_{or}^{a} - \gamma_{or} \cos \theta_{\gamma r}^{o}) \frac{dn}{n}\right] \tag{308}$$

Using equations (130) through (133) and (280) through (283) in equation (287) and separating into real and imaginary parts yields the following two radiation renormalization group equations

$$F_{1r} \cos \theta_{ur}^{o} + 3G_{1r} \sin \theta_{ur}^{o} - 3 \frac{E_{jr}}{E_{j}} P_{o}(A_{or} \cos \Gamma_{or}^{j} - B_{or}^{j} \sin \Gamma_{or}^{j}) = E_{or}^{a} \tag{309}$$

$$F_{1r} \sin \theta_{ur}^{o} - 3G_{1r} \cos \theta_{ur}^{o} - 3 \frac{E_{jr}}{E_{j}} P_{o}(A_{or} \sin \Gamma_{or}^{j} + B_{or}^{j} \cos \Gamma_{or}^{j}) = 0 \tag{310}$$

where

$$F_{1r} = 3n^{2} \frac{d^{2}E_{or}}{dn^{2}} - 3(1 + \gamma_{o} \cos \theta_{\gamma}^{o}) n \frac{dE_{or}}{dn} \tag{311}$$

$$+ \left[4 + 3\gamma_{o}\left(\cos \theta_{\gamma}^{o} + \sin \theta_{\gamma}^{o} n \frac{d\theta_{ur}^{o}}{dn}\right) - 3\left(n \frac{d\theta_{ur}^{o}}{dn}\right)^{2}\right] E_{or}$$

$$G_{1r} = \left(\gamma_{o} \sin \theta_{\gamma}^{o} - 2n \frac{d\theta_{ur}^{o}}{dn}\right) n \frac{dE_{or}}{dn} \tag{312}$$

$$+ \left[(1 + \gamma_{o} \cos \theta_{\gamma}^{o}) n \frac{d\theta_{ur}^{o}}{dn} - n^{2} \frac{d^{2}\theta_{ur}^{o}}{dn^{2}} - \gamma_{o} \sin \theta_{\gamma}^{o}\right] E_{or}$$

$$A_{or} = \gamma_{or} \cos \theta_{\gamma r}^{o} - \gamma_{o} \cos \theta_{\gamma}^{o} \tag{313}$$

$$B_{or} = \gamma_{or} \sin \theta_{\gamma r}^{o} - \gamma_{o} \sin \theta_{\gamma}^{o} \tag{314}$$

$$\Gamma_{or}^{j} = \theta_{ur}^{j} + \theta_{p}^{o} - \theta_{u}^{o} \tag{315}$$

In order to decouple the complex number equation (288) into two real equations one must first rewrite the expressions for $\bar{\alpha}$, $\bar{\beta}$, and $\bar{S}_{jr}$. From equation (298) it follows that

$$\bar{\alpha} = \frac{\gamma_o P_o}{T_o^2} e^{i(2\chi_o + \theta_\gamma^o - \theta_p^o)} \tag{316}$$

where γ_o , P_o , and χ_o are given by equations (272), (139), and (267) respectively, and T_o is given by

$$T_o^2 = \left(P_o - n \frac{dP_o}{dn}\right)^2 + P_o^2\left(n \frac{d\theta_p^o}{dn}\right)^2 \tag{317}$$

From equation (299) it follows that

$$\bar{\beta} = \frac{\gamma_o K_o}{T_o^2} e^{i(2\chi_o + \theta_\gamma^o + \omega_o - \theta_p^o)} \tag{318}$$

where K_o and ω_o are given by equations (145) and (146) respectively. The expression $\bar{S}_{jr}$ is obtained from equations (296), (132), (133), (144), and (283) to be

$$\bar{S}_{jr} = R_{jr} + iT_{jr} \tag{319}$$

where

$$R_{jr} = 1 + j + \frac{jP_o\gamma_{or}}{T_o^2}\left[\left(P_o - n \frac{dP_o}{dn}\right)\cos\theta_{\gamma r}^o - P_o n \frac{d\theta_p^o}{dn} \sin\theta_{\gamma r}^o\right] \tag{320}$$

$$+ 3\left(n \frac{d\gamma_{or}}{dn} \cos\theta_{\gamma r}^o - \gamma_{or} n \frac{d\theta_{\gamma r}^o}{dn} \sin\theta_{\gamma r}^o\right)$$

$$- 3(j - 1)[\gamma_{or}^2 \cos(2\theta_{\gamma r}^o) - 2\gamma_o\gamma_{or} \cos(\theta_p^o + \theta_{\gamma r}^o) + \gamma_o^2 \cos(2\theta_\gamma^o)]$$

$$T_{jr} = \frac{jP_o\gamma_{or}}{T_o^2}\left[\left(P_o - n\frac{dP_o}{dn}\right)\sin\theta^o_{\gamma r} + P_o n\frac{d\theta^o_p}{dn}\cos\theta^o_{\gamma r}\right] \tag{321}$$

$$+ 3\left(n\frac{d\gamma_{or}}{dn}\sin\theta^o_{\gamma r} + \gamma_{or} n\frac{d\theta^o_{\gamma r}}{dn}\cos\theta^o_{\gamma r}\right)$$

$$- 3(j-1)[\gamma^2_{or}\sin(2\theta^o_{\gamma r}) - 2\gamma_o\gamma_{or}\sin(\theta^o_\gamma + \theta^o_{\gamma r}) + \gamma^2_o\sin(2\theta^o_\gamma)]$$

Using equations (289) through (321) allows the second radiation equation (288) to be separated into real and imaginary parts as follows

$$\frac{jE_j\gamma_o}{T_o^2}(L_{or}\cos\Omega^j_{or} - M_{or}\sin\Omega^j_{or}) + E_{jr}(R_{jr}\cos\theta^j_{ur} - T_{jr}\sin\theta^j_{ur}) \tag{322}$$

$$= jE^a_j(\alpha^a K^a_{or} - \beta^a P^a_{or}) + E^a_{jr}T^a_{jr}$$

$$\frac{jE_j\gamma_o}{T_o^2}(L_{or}\sin\Omega^j_{or} + M_{or}\cos\Omega^j_{or}) + E_{jr}(R_{jr}\sin\theta^j_{ur} + T_{jr}\cos\theta^j_{ur}) = 0 \tag{323}$$

where

$$L_{or} = P_o K_{or}\cos\omega^o_r - K_o P_{or}\cos\omega_o \tag{324}$$

$$M_{or} = P_o K_{or}\sin\omega^o_r - K_o P_{or}\sin\omega_o \tag{325}$$

$$\Omega^j_{or} = \theta^o_{pr} + \theta^j_u + 2\chi_o + \theta^o_\gamma - \theta^o_p \tag{326}$$

Equations (309), (310), (322), and (323) are the four renormalization group equations needed to solve for the four unknown radiation functions E_{or}, θ^o_{ur}, γ_{or}, and $\theta^o_{\gamma r}$. The radiation analogs of equations (136) and (137) relate θ^o_{ur}, θ^o_{pr}, E_{or}, and P_{or}. Equations (307) and (308) relate $\bar{E}_{jr}$ to $\bar{\gamma}_{or}$. These eight equations can be used to solve for the eight quantities $\bar{E}_{or}$, $\bar{P}_{or}$, $\bar{E}_{jr}$, and $\bar{P}_{jr}$ (or $\bar{\gamma}_{or}$).

8. PROCESSES IN PHASE MATTER AND RADIATION. The internal phase angles of matter and radiation allow an extended interpretation of the types of processes that can occur in these systems. Consider for example the change in complex entropy given by equation (57) as

$$d\bar{S} = e^{i\theta_s}(dS + iSd\theta_s) \tag{327}$$

From equation (327) it is clear that $d\bar{S} = 0$ cannot represent a physical process, as the real and imaginary parts both set equal to zero would determine two $V = V(T)$ curves, and these conditions would perhaps hold jointly only at an intersection point. In fact two distinct processes can be obtained from equation (327)

$$dS = 0 \qquad \text{adiabatic process} \tag{328}$$

$$d\theta_s = 0 \qquad \text{entropy isophase process} \tag{329}$$

Thus an adiabatic process corresponds to a rotation of the entropy vector $\bar{S}$ in internal phase angle space. Processes may occur in nature such that changes in volume and temperature cause rotation of the entropy and internal energy vectors. For the case of constant entropy magnitude ($dS = 0$) the heat increment is obtained from equation (327) to be

$$d\bar{Q} = iT\bar{S}d\theta_s \tag{330}$$

For this adiabatic process the conservation of energy is written as

$$iT\bar{S}d\theta_s = d\bar{U} + \bar{P}dV \tag{331}$$

This results in equations (69) through (91) with $\partial S/\partial V = 0$ and $\partial S/\partial T = 0$.

For the case of constant magnitude of the internal energy, $dU = 0$ and the rotation of the internal energy vector is given by

$$d\bar{U} = i\bar{U}d\theta_u \tag{332}$$

and equation (66) gives

$$Td\bar{S} = i\bar{U}d\theta_u + \bar{P}dV \tag{333}$$

This results in equations (69) through (91) with $\partial U/\partial V = 0$ and $\partial U/\partial T = 0$. For the case when both $dS = 0$ and $dU = 0$, corresponding to rotations of both the entropy and internal energy vectors, one has

$$iT\bar{S}d\theta_s = i\bar{U}d\theta_u + \bar{P}dV \tag{334}$$

This results in equations (69) through (91) with $\partial U/\partial V = 0$, $\partial U/\partial T = 0$, $\partial S/\partial V = 0$, and $\partial S/\partial T = 0$. Similar results apply for the thermodynamic potentials $\bar{H}$, $\bar{A}$, and $\bar{G}$.

In general a process will result in a combined stretch and rotation of the

thermodynamic functions $\overline{U}$, $\overline{P}$, $\overline{S}$, $\overline{H}$, $\overline{A}$, and $\overline{G}$, and the time rate of change of a thermodynamic quantity will include an angular velocity of the internal phase angles. For example, the rate of pressure change is given by

$$\frac{d\overline{P}}{dt} = e^{i\theta_p}\left(\frac{dP}{dt} + iP\,\frac{d\theta_p}{dt}\right) \tag{335}$$

Thus, even if the magnitude of the pressure were held fixed the pressure vector can rotate internally.

Another possibility is the transfer of external angular momentum to internal angular momentum and vice versa. Thus external rotation may in some cases be coupled to the rotation of the internal phase angles. In fact, the Lagrangian of a rotating body may be of the form

$$L = \frac{1}{2}\sum_e I_e\omega_e^2 + \frac{1}{2}\sum_i I_i\omega_i^2 + \sum_{ie} a_{ei}\omega_e\omega_i - V(\theta_e,\theta_i) \tag{336}$$

where I_e and ω_e = external moment of inertia and angular velocity respectively, I_i and ω_i = internal moments of inertia and angular velocity respectively, and where θ_e and θ_i = external and internal angles respectively. The θ_i consists of θ_p , θ_u , θ_s , etc, and $\dot{\theta}_i$ includes $\dot{\theta}_p$, $\dot{\theta}_u$, $\dot{\theta}_s$ and so on. The internal moments of inertia I_p , I_u , I_s , etc are associated with the internal angle coordinates of pressure, internal energy, entropy, etc. It is expected that for such a system macroscopic energy transfers would occur between the internal and external dynamical systems. Such transfers may account for the glitches that appear in the spin-down of pulsars. Similar processes have been suggested to occur at the level of fundamental particles.[10-13]

9. CONCLUSION. The local gauge invariance of relativistic thermodynamics suggests the possibility that the thermodynamic state functions can be represented as complex numbers whose imaginary parts are related to phase angles in an internal space associated with all interacting systems of matter and radiation. Due to vacuum interactions, bulk matter solids and quantum liquids are coherent in internal space. The phase angles and magnitudes of the thermodynamic state functions are calculated from a solution of the renormalization group equations which represent the mathematical description of the interaction of matter and radiation in matter with the vacuum state. The internal phase angles are expected to manifest themselves in the state equations of matter and radiation in matter. In some cases a transfer of energy may occur between external rotations and the rotations of the internal phase angles. The internal phase angles are expected to affect the equations of motion of classical and quantum systems, and should affect the equilibrium configurations of atomic nuclei and the stars.

The renormalized ground state of a relativistic thermodynamic solid or quantum liquid is associated with a broken symmetry manifested by the nonzero values of the internal phases θ_p , θ_u , θ_s , etc. that are obtained as solutions

to the relativistic trace equation. A symmetrical ground state would have $\theta_p = 0$, $\theta_u = 0$, $\theta_s = 0$, etc. This broken symmetry should be associated with massive gauge bosons that are connected with the excited states of the internal phases of bulk matter, i.e., internal spin waves of the pressure and entropy. Similar broken symmetries are common in atomic and nuclear systems.[14] As a practical application of these ideas one can conceive of a bulk matter vacuum-induced broken symmetry thermodynamic engine. Such an internal phase engine would utilize the broken symmetry nature of the ground state of bulk matter in a manner analogous to the broken symmetry ferromagnetic state of an iron armature in an electric motor.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Frampton, P., Gauge Field Theories, Benjamin-Cummings, Menlo Park, 1987.

2. Close, F., An Introduction to Quarks and Partons, Academic, New York, 1979.

3. Okun, L., Particle Physics, Harwood, New York, 1985.

4. Akhiezer, A. and Berestetskii, V., Quantum Electrodynamics, Interscience, New York, 1965.

5. Schweber, S., An Introduction to Relativistic Quantum Field Theory, Harper & Row, New York, 1962.

6. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

7. 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.

8. Weiss, R. A., "Scale Invariant Equations for Relativistic Waves," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, ARO 87-1, May 27-30, 1986, p. 307.

9. Rumer, Y. and Ryvkin, M., Thermodynamics, Statistical Physics, and Kinetics, MIR Publishers, Moscow, 1980.

10. Cheng, T. and Li, L., Gauge Theory of Elementary Particle Physics, Clarendon Press, Oxford, 1984.

11. Jackiw, R. and Rebbi, C., "Vacuum Periodicity in a Yang-Mills Quantum Theory," Phys. Rev. Lett., 37, 172, (1976).

12. Hasenfratz, P. and 't Hooft, G., "Fermion-Boson Puzzle in a Gauge Theory," Phys. Rev. Lett., 36, 1119, (1976).

13. Goldhaber, A., "Connection of Spin and Statistics for Charge-Monopole Composites," Phys. Rev. Lett., 36, 1122, (1976).

14. Dasso, C. H. and Vitturi, A., "Reconstructing the Nuclear Profile in Gauge Space," Phys. Rev. Lett., 59, 634, (1987).

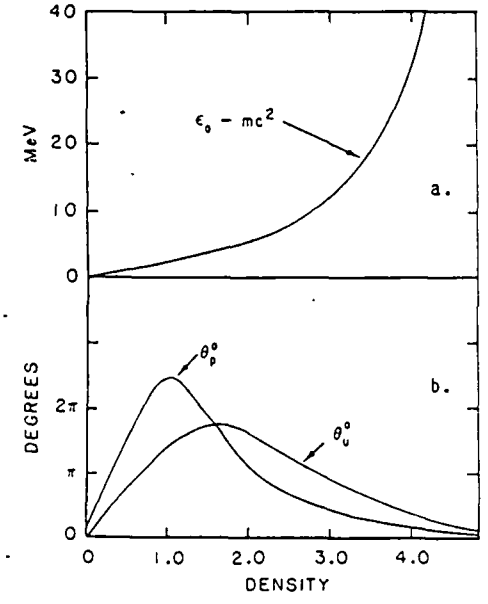


Figure 1. a) Binding energy of T = 0 neutron gas; b) Density dependence of the phase angles for the pressure and internal energy of a T = 0 neutron gas.

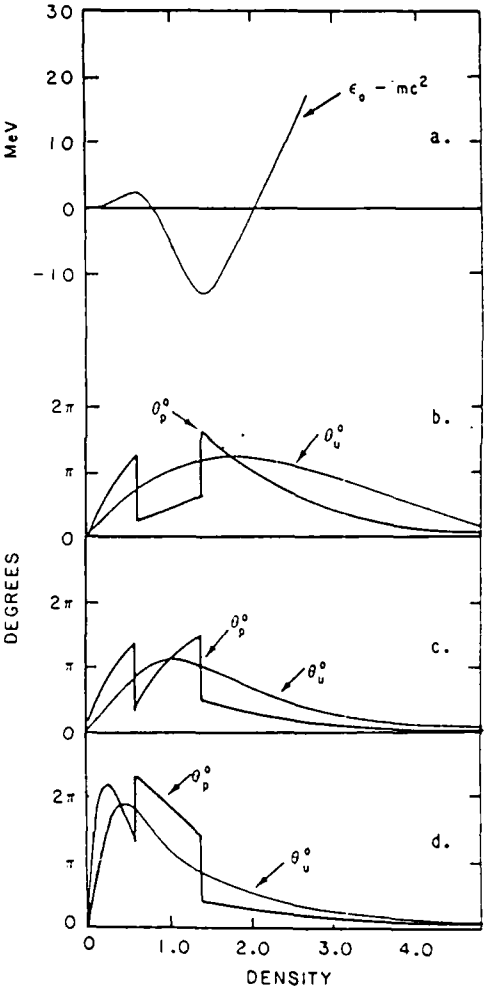


Figure 2. a) Binding energy of T = 0 infinite nuclear matter with N = Z; b), c), d), Three possible cases for the density dependence of the phase angles for the pressure and internal energy of T = 0 infinite nuclear matter with N = Z.

LAGRANGIAN FORMULATION OF RELATIVISTIC THERMODYNAMICS

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. Matter and radiation Lagrangians are developed from which the renormalization group equations of locally gauge invariant relativistic thermodynamics can be obtained by the Euler-Lagrange equations. The Noether current tensor and conservation equations of relativistic thermodynamics can be derived from these Lagrangians. These Lagrangians exhibit a minimum value when expressed in terms of the fractal dimension of a physical system, and are locally symmetric about this minimum value. This suggests that all matter and radiation in matter is fractal in nature. Gases, liquids, solids, quantum liquids, and the mechanical waves that propagate in these systems, have fractal properties. Equations for calculating the fractal dimensions of matter and radiation in matter are presented, and general expressions for the void ratio of gases, condensed matter, and radiation are derived. These results will have applications to matter and radiation at high densities such as occur in neutron stars, nuclear explosions, and the interaction of directed energy beams with matter.

1. INTRODUCTION. Lagrangian formulations of the theory of continuous systems are common in the classical mechanics of particles and fields.[1,2] But it is in the quantum theory of fields that the Lagrangian formalisms have exhibited their unique power to describe new physical effects in addition to yielding the dynamical equations of motion.[3-5] For instance, the properties of a chiral Lagrangian yield the left-right asymmetries of the electroweak force.[6,7] The spontaneously broken symmetry of a Lagrangian gives rise to such diverse phenomena as mass generation of gauge bosons, the existance of Goldstone bosons, the ferromagnetic ground state, the Meissner effect for superconductors, and many other subtle effects.[6,7] These results suggest that other locally gauge invariant systems, such as relativistic thermodynamics, may have a simple Lagrangian description.

A set of relativistic thermodynamic renormalization group equations for the ground and excited states of matter and radiation has been derived using the local scale (gauge) invariance of relativistic thermodynamics.[8,3] These renormalization group equations are partial differential equations for the energy and gauge parameters, and are similar in form to the Callan-Symanzik equations of relativistic quantum field theory.[7] These equations are derived from a relativistic trace equation that accounts for the vacuum interactions of matter and radiation in a four-dimensional formalism.[10] The trace equation is locally gauge invariant under the U(1) group in the sense that the values of the gauge transformation functions depend on the local density and temperature of a system.[9]

This paper develops a Lagrangian formulation of relativistic thermodynamics that is shown to be equivalent to the renormalization group equations. The Lagrangian density can be used to determine the fractal dimension (Hausdorf number) of bulk matter and radiation in bulk matter. Fractal matter systems are discussed extensively in the literature.[11-17] For relativistic thermodynamics, the fractal dimension is related to the state equation of a system and its deviation from the homogeneous case is due to the vacuum interactions of matter and radiation. In this paper the Gibbs-Helmholtz equation is used to estimate the void ratios of fractal matter and radiation.

2. RENORMALIZATION GROUP EQUATIONS FOR A FRACTAL GROUND STATE. The locally gauge invariant interaction of the vacuum state with uniform bulk matter and radiation is described by a relativistic trace equation.[10] The question arises, however, whether the vacuum interaction will produce a uniform system of matter and radiation or whether it will result in a fractal state. This question can be answered by developing the renormalization group equations for the fractal states of matter and radiation. The fractal analog of the relativistic trace equation is written as[10]

$$U + T\left(\frac{dU}{dT}\right)_{PV} - DV\,\frac{d}{dV}(PV)_U = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{1}$$

where D = fractal dimension = Hausdorf number.[11-17] The vacuum state (spacetime) has D = 3 to a very high degree of accuracy.[18] On the other hand, matter and radiation in matter need not have D = 3 , and in general the fractal dimension of a system will depend on volume and temperature, D = D(V,T). In equation (1), U = relativistic internal energy, P = relativistic pressure, T = absolute temperature, V = volume of substance, and U^a and P^a = corresponding nonrelativistic internal energy and pressure. Throughout this paper the index "a" will refer to nonrelativistic calculations.

For a fractal system with Hausdorf number D, equation (1) becomes[9]

$$\left(1 - b + T\,\frac{\partial}{\partial T} - bV\,\frac{\partial}{\partial V}\right)E - D\left(1 + \gamma + V\,\frac{\partial}{\partial V} - \gamma T\,\frac{\partial}{\partial T}\right)P \tag{2}$$

$$= \left(1 - b^a + T\,\frac{\partial}{\partial T} - b^aV\,\frac{\partial}{\partial V}\right)E^a$$

where E = relativistic energy density = U/V , E^a = nonrelativistic energy density, and where[9]

$$\gamma = \frac{V}{C_V}\left(\frac{\partial P}{\partial T}\right)_V \tag{3}$$

$$b = \frac{T(\partial P/\partial T)_V}{(P - K_T)} \tag{4}$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \tag{5}$$

where γ = relativistic Grüneisen parameter; C_V = relativistic heat capacity at constant volume, and C_V^a = nonrelativistic heat capacity at constant volume, given respectively by

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V \tag{6}$$

$$C_V^a = \left(\frac{\partial U^a}{\partial T}\right)_V \tag{7}$$

and where

$$K_T = -V\left(\frac{\partial P}{\partial V}\right)_T \tag{8}$$

$$K_T^a = -V\left(\frac{\partial P^a}{\partial V}\right)_T \tag{9}$$

are the relativistic and nonrelativistic values of the bulk modulus respectively. The parameters b and γ are the gauge parameters of relativistic thermodynamics.

Equation (2) can be rewritten as the following two renormalization group equations[9]

$$(T\frac{\partial}{\partial T} + fV\frac{\partial}{\partial V} + M)E = \psi^a \tag{10}$$

$$(T\frac{\partial}{\partial T} + hV\frac{\partial}{\partial V} + N)P = 0 \tag{11}$$

where

$$f = n - b \tag{12}$$

$$h = \frac{1}{n/D - \gamma} \tag{13}$$

$$M = f + 1 \tag{14}$$

$$N = h - 1 \tag{15}$$

$$\psi^a = \left(T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a\right) E^a \tag{16}$$

and where η = Lagrange multiplier given by[9]

$$\frac{\eta}{D} = \frac{V \frac{\partial P}{\partial V} - \gamma T \frac{\partial P}{\partial T} + (\gamma + 1)P}{P - T \frac{\partial P}{\partial T}} \tag{17}$$

Equations (10) through (17) reduce to equation (25) through (32) of Reference 9 for the case $D = 3$. Thus f , h , M , and N for the fractal ground state are now explicit functions of the fractal dimension D, and therefore E , P , and γ are also explicit functions of D. Finally, it will be assumed that $f \neq 0$ and $h \neq 0$ so that equations (10) and (11) can be rewritten as

$$V \frac{\partial E}{\partial V} + \frac{1}{f} T \frac{\partial E}{\partial T} + \frac{M}{f} E = \psi^a / f \tag{18}$$

$$V \frac{\partial P}{\partial V} + \frac{1}{h} T \frac{\partial P}{\partial T} + \frac{N}{h} P = 0 \tag{19}$$

For a solid or low temperature quantum system the nonrelativistic state equation of the ground state is assumed to have the following form[10]

$$E^a = E^a_o + E^a_j T^j + \cdots \tag{20}$$

$$P^a = P^a_o + P^a_j T^j + \cdots \tag{21}$$

where E^a and P^a = nonrelativistic energy density and pressure respectively, E^a_o and P^a_o = nonrelativistic zero-temperature values of the energy density and pressure respectively, E^a_j and P^a_j = nonrelativistic thermal coefficients for the energy density and pressure respectively, T = absolute temperature of the system (°K), and j = numerical index having values characteristic of the type of physical system. A commonly used descriptor of the thermal state equations given by equations (20) and (21) is the nonrelativistic zero-temperature value of the Grüneisen parameter that is defined by

$$\gamma^a_o = \frac{P^a_j}{E^a_j} = \frac{1}{(j-1)} \frac{1}{E^a_j} \frac{d}{dV}(V E^a_j) \tag{22}$$

except for $j = 1$. When $j = 1$, $\gamma_o^a = 2/3$. The zero temperature value of the nonrelativistic bulk modulus is given by $K_o^a = ndP_o^a/dn$, where $n = N/V$ = number of moles per unit volume, and N = number of moles of a substance.

The corresponding relativistic state equations will be written as[10]

$$E = E_o + E_j T^j + \cdots \tag{23}$$

$$P = P_o + P_j T^j + \cdots \tag{24}$$

$$\gamma_o = \frac{P_j}{E_j} = \frac{1}{(j-1)} \frac{1}{E_j} \frac{d}{dV}(VE_j) \tag{25}$$

except for $j = 1$, when $E_1 = E_1^a$, and where E_o and P_o = relativistic zero-temperature energy density and pressure respectively, E_j and P_j = relativistic thermal coefficients for the energy density and pressure respectively, and γ_o = relativistic zero-temperature Grüneisen parameter. The relativistic value of the zero temperature bulk modulus is given by $K_o = ndP_o/dn$.

Combining equation (2) with the state equations (20) through (25) yields the following ground state equations for fractal solids and low temperature quantum systems[10]

$$E_o - D_o[(1 + \gamma_o)P_o - K_o] = E_o^a \tag{26}$$

$$E_j\left(1 + j + \frac{j\gamma_o P_o}{P_o - K_o} + D_o n \frac{d\gamma_o}{dn}\right) = E_j^a\left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a}\right) \tag{27}$$

where $D_o = D_o(n)$ is the $T = 0$ value of the fractal dimension, and where[10]

$$\frac{E_j}{E_j^a} = \exp\left[-(j-1)\int^n (\gamma_o - \gamma_o^a)\frac{dn}{n}\right] \tag{28}$$

Note that in the derivation of equation (27) it is assumed that any explicit temperature dependence of D in the for $D = D_o + D_j T^j + \cdots$ can be neglected and that essentially $D_j = 0$. If this is not assumed, an additional term $-D_j[(1 + \gamma_o)P_o - K_o]$ has to be inserted into the left hand side of equation (27).

Equation (26) can be rewritten as follows

$$D_o n^2 \frac{d^2 E_o}{dn^2} - D_o(1 + \gamma_o) n \frac{dE_o}{dn} + [D_o(1 + \gamma_o) + 1] E_o = E_o^a \tag{29}$$

Equivalent forms of equation (29) are

$$D_o n^2 \frac{d^2 P_o}{dn^2} - D_o(1 + \gamma_o) n \frac{dP_o}{dn} + \left[D_o\left(\gamma_o - n \frac{d\gamma_o}{dn}\right) + D_o + 1 \right] P_o = P_o^a \tag{30}$$

$$D_o V^2 \frac{d^2 P_o}{dV^2} + D_o(3 + \gamma_o) V \frac{dP_o}{dV} + \left[D_o\left(\gamma_o + V \frac{d\gamma_o}{dV}\right) + D_o + 1 \right] P_o = P_o^a \tag{31}$$

Thus in general $E_o = E_o(E_o^a , \gamma_o^a , D_o , V)$ and $\gamma_o = \gamma_o(E_o^a , \gamma_o^a , D_o , V)$. It is possible that originally the unrenormalized state is fractal in nature so that $E_o^a = E_o^a(D_o^a , V)$ and $\gamma_o^a = \gamma_o^a(D_o^a , V)$. If in equation (29) one takes $E_o \sim n^{\sigma_o}$ and $E_o^a \sim n^{\sigma_o}$, where σ_o = adiabatic index, and γ_o = constant, one gets

$$E_o = \frac{E_o^a}{G(D_o)} \tag{32}$$

where

$$G(D_o) = D_o \sigma_o^2 - D_o \sigma_o (2 + \gamma_o) + D_o(1 + \gamma_o) + 1 \tag{33}$$

It then follows that

$$P_o = \frac{P_o^a}{G(D_o)} \qquad K_o = \frac{K_o^a}{G(D_o)} \tag{34}$$

A general expression for the void ratio in a fractal ground state can be obtained from the Gibbs-Helmholtz equation[10]

$$E + V \frac{\partial E}{\partial V} = T \frac{\partial P}{\partial T} - P \tag{35}$$

which can be rewritten as

$$\frac{dV}{V} = \frac{dE}{T \frac{\partial P}{\partial T} - P - E} \tag{36}$$

Equation (36) will be written in finite difference form corresponding to a change of fractal dimension from the uniform D = 3 case to the general case of arbitrary fractal dimension as follows

$$\frac{\Delta V}{V} \sim \frac{E(D) - E(3)}{P(3) + E(3) - T\,\frac{\partial P(3)}{\partial T}} \tag{37}$$

where the notation, $E(D)$ = energy density associated with a fractal dimension D, and $E(3)$ and $P(3)$ = energy density and pressure respectively for the homogeneous case of D = 3 , is introduced for calculating void ratios. The T = 0 limit of equation (37) is given by

$$\left(\frac{\Delta V}{V}\right)_o = \frac{E_o(D_o) - E_o(3)}{P_o(3) + E_o(3)} = \frac{E_o(D_o) - E_o(3)}{n\,\frac{dE_o(3)}{dn}} \tag{38}$$

Note that the energy density for a fractal ground state is greater than that of the homogeneous state.[19]

3. RENORMALIZATION GROUP EQUATIONS FOR FRACTAL RADIATION IN FRACTAL MATTER. The renormalization group equation for fractal radiation in fractal matter can be written as a simple extension of the corresponding equation for homogeneous matter, using the same notation as in equation (70) of Reference 9, as follows

$$\left(1 - b + T\,\frac{\partial}{\partial T} - bV\,\frac{\partial}{\partial V}\right)E_r - \beta_r\left(T\,\frac{\partial P}{\partial T} - P\right) \tag{39}$$

$$- D\left[\left(1 + \gamma + V\,\frac{\partial}{\partial V} - \gamma T\,\frac{\partial}{\partial T}\right)P_r - \delta_r\left(T\,\frac{\partial P}{\partial T} - P\right)\right]$$

$$+ d_r\left(1 + \gamma + V\,\frac{\partial}{\partial V} - \gamma T\,\frac{\partial}{\partial T}\right)P$$

$$= \left(1 - b^a + T\,\frac{\partial}{\partial T} - b^a V\,\frac{\partial}{\partial V}\right)E_r^a - \beta_r^a\left(T\,\frac{\partial P^a}{\partial T} - P^a\right)$$

where the fractal dimension of the radiation in matter is $D_r = D - d_r$, and where $d_r > 0$ is the incremental change in the fractal dimension due to the presence of radiation in the system, and where[9]

$$\beta_r = b_r\,\frac{P_r - K_{Tr}}{P - K_T} - \frac{T\,\frac{\partial P}{\partial T}\,(P_r - K_{Tr})}{(P - K_T)^2} \tag{40}$$

$$b_r = \frac{T\partial P_r/\partial T}{P_r - K_{Tr}} \tag{41}$$

$$K_{Tr} = -V \frac{\partial P_r}{\partial V} \tag{42}$$

$$\delta_r = \frac{\partial E_r/\partial T}{\partial E/\partial T} (\gamma_r - \gamma) \tag{43}$$

$$\gamma_r = \frac{\partial P_r/\partial T}{\partial E_r/\partial T} \tag{44}$$

The functions γ_r and b_r are the radiation gauge parameters, and K_{Tr} = radiation bulk modulus. Note that β_r , δ_r , and d_r are generally small quantities, while γ_r , b_r , and D_r refer to the radiation itself and are not small quantities. For D = 3 and d_r = 0 , equation (39) reduces to equation (70) of Reference 9.

Equation (39) can be decoupled into two independent radiation renormalization group equations as follows

$$(T \frac{\partial}{\partial T} + f_r V \frac{\partial}{\partial V} + M_r)E_r - \beta_r(T \frac{\partial P}{\partial T} - P) = \psi_r^a \tag{45}$$

$$(T \frac{\partial}{\partial T} + h_r V \frac{\partial}{\partial V} + N_r)P_r - h_r \delta_r (T \frac{\partial P}{\partial T} - P) \tag{46}$$

$$- \frac{h_r d_r}{D} \left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T} \right) P = 0$$

$$f_r = \eta_r - b \tag{47}$$

$$h_r = (\eta_r/D - \gamma)^{-1} \tag{48}$$

$$M_r = f_r + 1 \tag{49}$$

$$N_r = h_r - 1 \tag{50}$$

$$\psi_r^a = (T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a)E_r^a - \beta_r^a (T \frac{\partial P^a}{\partial T} - P^a) \tag{51}$$

where η_r = Lagrange multiplier. For the case D = 3 and $d_r = 0$, equations (45) through (51) reduce to equations (74) through (80) of Reference 9. It will be assumed that $f_r \neq 0$ and $h_r \neq 0$ so that equations (45) and (46) can be written as

$$V \frac{\partial E_r}{\partial V} + \frac{T}{f_r} \frac{\partial E_r}{\partial T} + \frac{M_r}{f_r} E_r - \frac{\beta_r}{f_r} \left(T \frac{\partial P}{\partial T} - P\right) = \psi_r^a / f_r \tag{52}$$

$$V \frac{\partial P_r}{\partial V} + \frac{T}{h_r} \frac{\partial P_r}{\partial T} + \frac{N_r}{h_r} P_r - \delta_r \left(T \frac{\partial P}{\partial T} - P\right) \tag{53}$$

$$- \frac{d_r}{D} \left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right) P = 0$$

The energy density and pressure for radiation in solids and quantum liquids is written as[9]

$$E_r^a = E_{or}^a + E_{jr}^a T^j + \cdots \tag{54}$$

$$P_r^a = P_{or}^a + P_{jr}^a T^j + \cdots \tag{55}$$

and

$$E_r = E_{or} + E_{jr} T^j + \cdots \tag{56}$$

$$P_r = P_{or} + P_{jr} T^j + \cdots \tag{57}$$

where

E_{or}^a and P_{or}^a = nonrelativistic zero-temperature radiation energy density and pressure respectively

E_{jr}^a and P_{jr}^a = nonrelativistic thermal coefficients for the radiation energy density and pressure respectively

E_{or} and P_{or} = relativistic zero-temperature radiation energy density and pressure respectively

E_{jr} and P_{jr} = relativistic thermal coefficients for the radiation energy density and pressure respectively

The zero temperature value of the radiation Grüneisen parameter is obtained from equations (44) and (54) through (57) to be

$$\gamma_{or}^{a} = \frac{P_{jr}^{a}}{E_{jr}^{a}} \qquad \gamma_{or} = \frac{P_{jr}}{E_{jr}} \tag{58}$$

The zero temperature values of the nonrelativistic and relativistic radiation bulk modulus is written as $K_{or}^{a} = ndP_{or}^{a}/dn$ and $K_{or} = ndP_{or}/dn$ respectively.

When radiation is present in a fractal solid or low temperature quantum liquid, the T = 0 fractal dimension of the radiation will be written as $D_{or} = D_o - d_{or}$ where $d_{or} > 0$ is the small change in fractal dimension associated with the addition of radiation to the material system. The excitation equations for such a system are obtained from equations (45) and (46) and are an extension of equations (104) and (105) of Reference 9. They are written as

$$E_{or} - D_o[(1 + \gamma_o)P_{or} - K_{or}] - D_o \frac{E_{jr}}{E_j} P_o(\gamma_{or} - \gamma_o) + d_{or}[(1 + \gamma_o)P_o - K_o] = E_{or}^{a} \tag{59}$$

$$jE_j(\alpha K_{or} - \beta P_{or}) + E_{jr}S_{jr} - d_{or}E_j n \frac{d\gamma_o}{dn} = jE_j^{a}(\alpha^{a}K_{or}^{a} - \beta^{a}P_{or}^{a}) + E_{jr}^{a}T_{jr}^{a} \tag{60}$$

where

$$S_{jr} = 1 + j + \frac{jP_o\gamma_{or}}{P_o - K_o} + D_o n \frac{d\gamma_{or}}{dn} - D_o(j - 1)(\gamma_{or} - \gamma_o)^2 \tag{61}$$

$$T_{jr}^{a} = 1 + j + jP_o^{a}\gamma_{or}^{a}/(P_o^{a} - K_o^{a}) \tag{62}$$

$$\alpha = \gamma_o P_o/(P_o - K_o)^2 \qquad \alpha^{a} = \gamma_o^{a}P_o^{a}/(P_o^{a} - K_o^{a})^2 \tag{63}$$

$$\beta = \gamma_o K_o/(P_o - K_o)^2 \qquad \beta^{a} = \gamma_o^{a}K_o^{a}/(P_o^{a} - K_o^{a})^2 \tag{64}$$

In the derivation of equation (60) it is assumed that one can neglect a temperature dependence of D_r of the form $D_r = D_{or} + D_{jr}T^j + \cdots$ (or equivalently, $d_r = D - D_r = D_o - D_{or} + (D_j - D_{jr})T^j + \cdots = d_{or} + d_{jr}T^j + \cdots$) and that $D_j = 0$, $D_{jr} = 0$, and $d_{jr} = 0$. If this is not the case and $D_j \neq 0$, then an additional term $+d_{jr}[(1 + \gamma_o)P_o - K_o]$ has to be inserted into the left hand side of equation (60). For this case both d_{or} and d_{jr} must be determined. Equation (59) can be rewritten as

$$D_o n^2 \frac{d^2 E_{or}}{dn^2} - D_o(1 + \gamma_o) n \frac{dE_{or}}{dn} + [D_o(1 + \gamma_o) + 1] E_{or} \tag{65}$$

$$+ D_o \frac{E_{jr}}{E_j} P_o(\gamma_o - \gamma_{or}) + d_{or}[(1 + \gamma_o)P_o - K_o] = E^a_{or}$$

Equations (59) through (65) reduce to equations (104) through (113) of Reference 9 for the case $D_o = 3$ and $d_{or} = 0$. The values of d_{or} and d_{jr} can be obtained from an appropriate Lagrangian formalism.

From the Gibbs-Helmholtz equation for a radiation system

$$E_r + V \frac{\partial E_r}{\partial V} = T \frac{\partial P_r}{\partial T} - P_r \tag{66}$$

one obtains the following estimate for the void ratio of a fractal mechanical radiation system in matter

$$\left(\frac{\Delta V}{V}\right)_r = \frac{E_r(D,d_r) - E_r(3,0)}{P_r(3,0) + E_r(3,0) - T \frac{\partial P_r(3,0)}{\partial T}} \tag{67}$$

where the notation $E_r(D,d_r)$ = fractal radiation energy density, and $E_r(3,0)$ and $P_r(3,0)$ = homogeneous radiation energy density and pressure respectively, will be introduced for calculating the void ratios of the fractal radiation field. The $T = 0$ limit of equation (67) is given by

$$\left(\frac{\Delta V}{V}\right)_{or} = \frac{E_{or}(D_o,d_{or}) - E_{or}(3,0)}{P_{or}(3,0) + E_{or}(3,0)} \tag{68}$$

Specific examples of the use of these equations for radiation in gases and condensed matter are given in Sections 7 and 8.

4. GROUND STATE LAGRANGIAN. Lagrangian formulations of nonlocal gauge field theories have been used to describe the four basic interactions that occur in nature.[6,7] One is tempted to write a similar Lagrangian formulation of the effects of the vacuum state on bulk matter. This section develops a Lagrangian description of a nonlocal gauge theory of relativistic thermodynamics. Let the Lagrangian function of a relativistic thermodynamic system be written as

$$L = \int \mathcal{L}\left(\frac{\partial \phi}{\partial v}, \frac{\partial \phi}{\partial t}, \phi, v, t\right) dv \tag{69}$$

and the thermodynamic action I be written as

$$I = \int \mathcal{L}\left(\frac{\partial \phi}{\partial v}, \frac{\partial \phi}{\partial t}, \phi, v, t\right) dv dt = \int L dt \tag{69A}$$

where $\mathcal{L}$ = Lagrangian density, $\phi = \phi(v,t)$ is an appropriately selected field, and where

$$v = \ell n\ V \tag{70}$$

$$t = \ell n\ T \tag{71}$$

The introduction of the variables in equations (70) and (71) is made because it simplifies the ground state renormalization group equations (18) and (19) which now become

$$\frac{\partial E}{\partial v} + \frac{1}{f}\frac{\partial E}{\partial t} + \frac{M}{f}E = \psi^a/f \tag{72}$$

$$\frac{\partial P}{\partial v} + \frac{1}{h}\frac{\partial P}{\partial t} + \frac{N}{h}P = 0 \tag{73}$$

The Euler-Lagrange field equations derived from $\delta I = 0$ are[1,3]

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \phi_{,t}}\right) + \frac{d}{dv}\left(\frac{\partial \mathcal{L}}{\partial \phi_{,v}}\right) = \frac{\partial \mathcal{L}}{\partial \phi} \tag{74}$$

where the following notation was introduced

$$\phi_{,t} = \frac{\partial \phi}{\partial t} \qquad \phi_{,v} = \frac{\partial \phi}{\partial v} \tag{75}$$

The Lagrangian density $\mathcal{L}(\phi_{,t}\ ,\ \phi_{,v}\ ,\phi,v,t)$ and the field $\phi(v,t)$ are selected in an appropriate way for relativistic thermodynamics so that the Euler-Lagrange equations (74) will reproduce the ground state renormalization group equations (72) and (73). In order to reproduce equation (72) one takes $\phi = \xi$ where

$$\xi = \int E dv = \int E\,\frac{dV}{V} = \xi(V,T,D) \tag{76}$$

The corresponding Lagrangian density is

$$\mathcal{L}_1 = \frac{1}{2}A\xi_{,v}^2 + I\xi_{,t} + B\xi + \frac{1}{2}C\xi^2 \tag{77}$$

where

$$A = 1 + Z \tag{78}$$

$$Z = \int \frac{M}{f}\,dv \tag{79}$$

$$I = \int \frac{1}{f} \frac{\partial E}{\partial t} dt \tag{80}$$

$$B = Z(\xi_{,v} + \xi_{,vv}) + \frac{\psi^a}{f} = Z(E + E_{,v}) + \frac{\psi^a}{f} \tag{81}$$

$$C = \frac{M}{f} \tag{82}$$

where $\xi_{,vv} = E_{,v}$ is treated as a parameter dependent of v and t but independent of ξ, $\xi_{,v}$ or $\xi_{,t}$. In order to see that equation (72) can be derived from $\mathcal{L}_1$ and equation (74) it is noted that

$$\frac{\partial \mathcal{L}_1}{\partial \xi_{,t}} = I = \int \frac{1}{f} \frac{\partial E}{\partial t} dt \tag{83}$$

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}_1}{\partial \xi_{,t}}\right) = \frac{1}{f} \frac{\partial E}{\partial t} \tag{84}$$

$$\frac{\partial \mathcal{L}_1}{\partial \xi_{,v}} = A\xi_{,v} + [illegible] = AE + [illegible] \tag{85}$$

$$\frac{d}{dv}\left(\frac{\partial \mathcal{L}_1}{\partial \xi_{,v}}\right) = AE_{,v} + C(\xi + E) + [illegible] \tag{86}$$

$$\frac{\partial \mathcal{L}_1}{\partial \xi} = B + C\xi \tag{87}$$

Placing these quantities in equation (74) yields equation (72).

In a similar fashion, in order to reproduce equation (73) one takes $\phi = \zeta$

$$\zeta = \int P dv = \int P \frac{dV}{V} = \zeta(V,T,D) \tag{88}$$

The corresponding Lagrangian density is

$$\mathcal{L}_2 = \frac{1}{2} F\zeta_{,v}^2 + J\zeta_{,t} + G\zeta + \frac{1}{2} H\zeta^2 \tag{89}$$

where

$$F = 1 + X \tag{90}$$

$$X = \int \frac{N}{h} dv \tag{91}$$

$$J = \int \frac{1}{h} \frac{\partial P}{\partial t} \, dt \tag{92}$$

$$G = X(\zeta_{,v} + \zeta_{,vv}) = X(P + P_{,v}) \tag{93}$$

$$H = \frac{N}{h} \tag{94}$$

where $\zeta_{,vv} = P_{,v}$ is treated as a parameter dependent on v and t but independent of ζ , $\zeta_{,v}$ or $\zeta_{,t}$. The following relationships hold

$$\frac{\partial \mathcal{L}_2}{\partial \zeta_{,t}} = J = \int \frac{1}{h} \frac{\partial P}{\partial t} \, dt \tag{95}$$

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}_2}{\partial \zeta_{,t}}\right) = \frac{1}{h} \frac{\partial P}{\partial t} \tag{96}$$

$$\frac{\partial \mathcal{L}_2}{\partial \zeta_{,v}} = F\zeta_{,v} + \zeta X = FP + \zeta X \tag{97}$$

$$\frac{d}{dv}\left(\frac{\partial \mathcal{L}_2}{\partial \zeta_{,v}}\right) = FP_{,v} + H(\zeta + P) + XP \tag{98}$$

$$\frac{\partial \mathcal{L}_2}{\partial \zeta} = G + H\zeta \tag{99}$$

Placing equations (95) through (99) into equation (74) shows that $\mathcal{L}_2$ is a proper Lagrangian density for the pressure renormalization group equation (73).

It should be pointed out that the Lagrangian densities $\mathcal{L}_1$ and $\mathcal{L}_2$ have a proper T = 0 limit and are given by

$$\mathcal{L}_1^o = \frac{1}{2} A_o \xi_{o,v}^2 + B_o \xi_o + \frac{1}{2} C_o \zeta_o^2 \tag{100}$$

$$\mathcal{L}_2^o = \frac{1}{2} F_o \zeta_{o,v}^2 + G_o \zeta_o + \frac{1}{2} H_o \zeta_o^2 \tag{101}$$

where

$$\xi_o = \int E_o dv \qquad \zeta_o = \int P_o dv \tag{102}$$

and where

$$A_o = 1 + Z_o \tag{103}$$

$$Z_o = \int \frac{M_o}{f_o}\, dv = \int C_o dv \tag{104}$$

$$B_o = A_o(E_o + E_{o,v}) + \frac{E_o^a}{f_o} \tag{105}$$

$$C_o = \frac{M_o}{f_o} \tag{106}$$

$$F_o = 1 + X_o \tag{107}$$

$$X_o = \int \frac{N_o}{h_o}\, dv = \int H_o dv \tag{108}$$

$$G_o = X_o(P_o + P_{o,v}) \tag{109}$$

$$H_o = \frac{N_o}{h_o} \tag{110}$$

where f_o , h_o , M_o and N_o are defined in Reference 9. Equations (100) and (101) yield the following T = 0 ground state equations

$$\frac{dE_o}{dv} + \frac{M_o}{f_o} E_o = \frac{E_o^a}{f_o} \tag{111}$$

$$\frac{dP_o}{dv} + \frac{N_o}{h_o} P_o = 0 \tag{112}$$

The two potential functions associated with $\mathcal{L}_1$ and $\mathcal{L}_2$ are obtained from equations (77) and (89) to be

$$V_1 = B\xi + \frac{1}{2} C\xi^2 \tag{113}$$

$$V_2 = G\zeta + \frac{1}{2} H\zeta^2 \tag{114}$$

If $B < 0$, $G < 0$ and $C > 0$, $H > 0$ the potentials have a minimum at certain values of ξ and ζ which are determined from

$$\frac{\partial V_1}{\partial \xi} = 0 \qquad \frac{\partial V_2}{\partial \zeta} = 0 \tag{115}$$

where ξ and ζ are varied by changing the fractal dimension D for fixed V and T. The conditions in equation (115) are equivalent to

$$B + C\xi_M = 0 \tag{116}$$

$$G + H\zeta_M = 0 \tag{117}$$

or

$$\xi_M = -\frac{B}{C} = -\frac{1}{C}\left[Z(E + E_{,v}) + \psi^a/f\right] \tag{118}$$

$$\zeta_M = -\frac{G}{H} = -\frac{X}{H}(P + P_{,v}) \tag{119}$$

where $\xi_M(V,T)$ and $\zeta_M(V,T)$ = values of ξ and ζ for which V_1 and V_2 are respectively minimum. Taking the derivatives of equations (118) and (119) with respect to v yields respectively

$$ZE_{,vv} + \left(Z + C - Z\frac{C_{,v}}{C}\right)E_{,v} + \left(2C - Z\frac{C_{,v}}{C}\right)E = \frac{C_{,v}}{C}\frac{\psi^a}{f} - \frac{\partial}{\partial v}\left(\frac{\psi^a}{f}\right) \tag{120}$$

$$XP_{,vv} + \left(X + H - X\frac{H_{,v}}{H}\right)P_{,v} + \left(2H - X\frac{H_{,v}}{H}\right)P = 0 \tag{121}$$

Either equation (120) or (121) can be solved for $D = D_M(V,T)$ that makes the potential V_1 and V_2 have minimum values. About the minimum, the potentials have the form

$$V_1(\xi) - V_1(\xi_M) = \frac{1}{2} C(\Delta\xi)^2 \tag{122}$$

$$V_2(\zeta) - V_2(\zeta_M) = \frac{1}{2} H(\Delta\zeta)^2 \tag{123}$$

where $\xi = \xi_M + \Delta\xi$ and $\zeta = \zeta_M + \Delta\zeta$. Thus the potentials are locally symmetric about the minimum points. The value of D at the minimum points will be designated as $D_M(V,T)$, and in general $D_M < 3$ so that matter will have voids and is fractal in nature for specified values of V and T. The fractal ground state of matter occurs only for limited regions of V and T corresponding to the conditions $B < 0$, $G < 0$ and $C > 0$, $H > 0$.

It should be pointed out that the Lagrangians $\mathcal{L}_1$ and $\mathcal{L}_2$ are not unique in the sense that the following two Lagrangians also yield the desired renormalization group equations

$$\mathcal{L}_1' = \frac{1}{2} A\xi_{,v}^2 + I\xi_{,t} + \xi(\xi_{,v} + Z\xi_{,vv} + \psi^a/f) \tag{124}$$

$$\mathcal{L}_2' = \frac{1}{2} F\zeta_{,v}^2 + J\zeta_{,t} + \zeta(\zeta_{,v} + X\zeta_{,vv}) \tag{125}$$

These Lagrangians are linear in ξ and ζ respectively and do not contain quadratis potential terms as in the cases of equations (77) and (89). The quadratic Lagrangians are chosen instead of the linear potential Lagrangians because they are symmetrical about their minimum values.

5. EXCITED STATE LAGRANGIAN. The radiation renormalization group equations (45) and (46) can also be obtained from a Lagrangian formulation. The Lagrangian density that gives equation (45) is

$$\mathcal{L}_{1r} = \frac{1}{2} A_r\xi_{r,v}^2 + I_r\xi_{r,t} + B_r\xi_r + \frac{1}{2} C_r\xi_r^2 \tag{126}$$

where

$$\xi_r = \int E_r dv = \int E_r \frac{dV}{V} = \xi_r(V,T,D,D_r) \tag{127}$$

and

$$A_r = 1 + Z_r \tag{128}$$

$$Z_r = \int \frac{M_r}{f_r} dv \tag{129}$$

$$I_r = \int \frac{1}{f_r} \frac{\partial E_r}{\partial t} dt \tag{130}$$

$$B_r = Z_r(\xi_{r,v} + \xi_{r,vv}) + \frac{\beta_r}{f_r}\left(T\frac{\partial P}{\partial T} - P\right) + \frac{\psi_r^a}{f_r} \tag{131}$$

$$C_r = \frac{M_r}{f_r} \tag{132}$$

where $\xi_{r,vv} = E_{r,v}$ is treated as parameter which is dependent on v and t but is independent of ξ_r, $\xi_{r,v}$ or $\xi_{r,t}$. To show that the Euler-Lagrange equation (74) yields equation (45) when $\phi = \xi_r$ it is noted that

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,t}}\right) = \frac{1}{f_r}\frac{\partial E_r}{\partial t} \tag{133}$$

$$\frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,v}} = A_r\xi_{r,v} + \xi_r Z_r = A_r E_r + \xi_r Z_r \tag{134}$$

$$\frac{d}{dv}\left(\frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,v}}\right) = A_r E_{r,v} + C_r(\xi_r + E_r) + Z_r E_r \tag{135}$$

$$\frac{\partial \mathcal{L}_{1r}}{\partial \xi_r} = B_r + C_r\xi_r \tag{136}$$

Combining equations (133) through (136) with equation (74) yields equation (45).

The Lagrangian density that yields equation (46) is found by choosing $\phi = \zeta_r$ in equation (74) with $\mathcal{L} = \mathcal{L}_{2r}$ where

$$\mathcal{L}_{2r} = \frac{1}{2}F_r\zeta_{r,v}^2 + J_r\zeta_{r,t} + G_r\zeta_r + \frac{1}{2}H_r\zeta_r^2 \tag{137}$$

where

$$\zeta_r = \int P_r dv = \zeta_r(V,T,D,d_r) \tag{138}$$

and

$$F_r = 1 + X_r \tag{139}$$

$$X_r = \int \frac{N_r}{h_r}\, dv = \int H_r dv \tag{140}$$

$$J_r = \int \frac{1}{h_r} \frac{\partial P_r}{\partial t}\, dt \tag{141}$$

$$G_r = X_r(\zeta_{r,v} + \zeta_{r,vv}) + \delta_r\left(T \frac{\partial P}{\partial T} - P\right) + \frac{d_r}{D}\left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right)P \tag{142}$$

$$H_r = \frac{N_r}{h_r} \tag{143}$$

If $\zeta_{r,vv}$ is taken to be a parameter only dependent on v and t, one has

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,t}}\right) = \frac{1}{h_r} \frac{\partial P_r}{\partial t} \tag{144}$$

$$\frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,v}} = F_r P_r + \zeta_r X_r \tag{145}$$

$$\frac{d}{dv}\left(\frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,v}}\right) = F_r P_{r,v} + H_r(\zeta_r + P_r) + X_r P_r \tag{146}$$

$$\frac{\partial \mathcal{L}_{2r}}{\partial \zeta_r} = G_r + H_r \zeta_r \tag{147}$$

which combine with equation (74) to give equation (46). It should be noted that the Lagrangians $\mathcal{L}_{1r}$ and $\mathcal{L}_{2r}$ have natural extensions to the case T = 0 .

In a form similar to that in equations (113) and (114), the potentials associated with radiation in matter are given by

$$V_{1r} = B_r \xi_r + \frac{1}{2} C_r \xi_r^2 \tag{148}$$

$$V_{2r} = G_r \zeta_r + \frac{1}{2} H_r \zeta_r^2 \tag{149}$$

If $B_r < 0$, $G_r < 0$, and $C_r > 0$, $H_r > 0$ these potentials will have minimum values at specific values of ξ_r and ζ_r given by

$$\frac{\partial V_{1r}}{\partial \xi_r} = B_r + C_r \xi_{rM} = 0 \tag{150}$$

$$\frac{\partial V_{2r}}{\partial \zeta_r} = G_r + H_r \zeta_{rM} = 0 \tag{151}$$

where $\xi_{rM}(V,T,D)$ and $\zeta_{rM}(V,T,D)$ = values of ξ_r and ζ_r for which V_{1r} and V_{2r} are respectively minimum, and where the variation of ξ_r and ζ_r in equations (150) and (151) corresponds to a change in d_r for fixed values of V, T, and D. The value of the fractal dimension of radiation in fractal matter is $D_r = D - d_r$ generally, where $d_r > 0$ so that in general there will be voids in a mechanical radiation field. If the value $d_r = d_{rM}(V,T,D)$ minimizes the potentials V_{1r} and V_{2r} , then the fractal dimension of mechanical radiation in fractal matter is $D_{rM} = D_M - d_{rM}$, and the fractal dimension of radiation in homogeneous matter is $D_{rM} = 3 - d_{rM}$. Mechanical radiation in matter is fractal in nature, and this includes waves in gases, liquids, and solids for limited regions of temperature and density where $B_r < 0$, $G_r < 0$ and $C_r > 0$, $H_r > 0$. Note that in general $D_M < 3$ so that $D_{rM} < 3$. Finally, when V_{1r} and V_{2r} have minimum values they can be expanded about these minimum values by writing $\xi_r = \xi_{rM} + \Delta\xi_r$ and $\zeta_r = \zeta_{rM} + \Delta\zeta_r$ as follows

$$V_{1r}(\xi_r) - V_{1r}(\xi_{rM}) = \frac{1}{2} C_r (\Delta\xi_r)^2 \tag{152}$$

$$V_{2r}(\zeta_r) - V_{2r}(\zeta_{rM}) = \frac{1}{2} H_r (\Delta\zeta_r)^2 \tag{153}$$

Electromagnetic radiation in matter will be treated in another paper.

6. THERMODYNAMIC NOETHER CURRENT TENSOR. Because the renormalization group equations can be derived from a variational principle, there exists a formal procedure for determining the conservation laws as a result of the form invariance of the Lagrangian density under continuous transformations. This procedure is given by Noether's theorem.[20] If the coordinates of a field $\phi(x_\mu)$ undergo a continuous translation of the form

$$x_\mu \rightarrow x'_\mu = x_\mu + \Delta x_\mu \tag{154}$$

then the Noether tensor

$$\Phi_{\mu\nu} = \frac{\partial \mathcal{L}}{\partial \phi_{,\nu}} \phi_{,\mu} - g_{\mu\nu}\mathcal{L} \tag{155}$$

satisfies the conservation law

$$\frac{\partial \phi_{\mu\nu}}{\partial x_{\mu}} = 0 \tag{156}$$

In this paper it will be assumed that the thermodynamic Lagrangian densities are form invariant under continuous changes of volume and temperature of the form

$$t \rightarrow t' = t + \Delta t \tag{157}$$

$$v \rightarrow v' = v + \Delta v \tag{158}$$

This is true for the Lagrangian densities of relativistic thermodynamics because they are ultimately expressed in terms of the energy density and pressure, and these quantities are form invariant under continuous changes of volume and temperature.

The Noether current tensor for the energy density of the ground state of a thermodynamic system is given by

$$\phi^{E}_{vv} = \frac{\partial \mathcal{L}_1}{\partial \xi_{,v}} \xi_{,v} - \mathcal{L}_1 \tag{159}$$

$$\phi^{E}_{tt} = \frac{\partial \mathcal{L}_1}{\partial \xi_{,t}} \xi_{,t} - \mathcal{L}_1 \tag{160}$$

$$\phi^{E}_{vt} = \frac{\partial \mathcal{L}_1}{\partial \xi_{,t}} \xi_{,v} \tag{161}$$

$$\phi^{E}_{tv} = \frac{\partial \mathcal{L}}{\partial \xi_{,v}} \xi_{,t} \tag{162}$$

which satisfy the following conservation equations

$$\frac{\partial \phi^{E}_{tv}}{\partial t} + \frac{\partial \phi^{E}_{vv}}{\partial v} = 0 \tag{163}$$

$$\frac{\partial \phi^{E}_{vt}}{\partial v} + \frac{\partial \phi^{E}_{tt}}{\partial t} = 0 \tag{164}$$

The components of the Noether tensor for the ground state energy density are obtained from equations (159) through (162) by using the expression for $\mathcal{L}_1$ given in equation (77) as follows

$$\phi^{E}_{vv} = \frac{1}{2} A\xi^{2}_{,v} - I\xi_{,t} - \xi(B - Z\xi_{,v}) - \frac{1}{2} C\xi^{2} \tag{165}$$

$$\phi^{E}_{tt} = -\frac{1}{2} A\xi^{2}_{,v} - B\xi - \frac{1}{2} C\xi^{2} \tag{166}$$

$$\phi^{E}_{vt} = I\xi_{,v} \tag{167}$$

$$\phi^{E}_{tv} = \xi_{,t}(A\xi_{,v} + \xi Z) \tag{168}$$

The Noether tensor for the ground state pressure of a thermodynamic system is written as

$$\phi^{P}_{vv} = \frac{\partial \mathcal{L}_2}{\partial \zeta_{,v}} \zeta_{,v} - \mathcal{L}_2 \tag{169}$$

$$\phi^{P}_{tt} = \frac{\partial \mathcal{L}_2}{\partial \zeta_{,t}} \zeta_{,t} - \mathcal{L}_2 \tag{170}$$

$$\phi^{P}_{vt} = \frac{\partial \mathcal{L}_2}{\partial \zeta_{,t}} \zeta_{,v} \tag{171}$$

$$\phi^{P}_{tv} = \frac{\partial \mathcal{L}_2}{\partial \zeta_{,v}} \zeta_{,t} \tag{172}$$

which satisfy the following conservation equations

$$\frac{\partial \phi^{P}_{tv}}{\partial t} + \frac{\partial \phi^{P}_{vv}}{\partial v} = 0 \tag{173}$$

$$\frac{\partial \phi^{P}_{vt}}{\partial v} + \frac{\partial \phi^{P}_{tt}}{\partial t} = 0 \tag{174}$$

where

$$\phi^{P}_{vv} = \frac{1}{2} F\zeta^{2}_{,v} - J\zeta_{,t} - \zeta(G - X\zeta_{,v}) - \frac{1}{2} H\zeta^{2} \tag{175}$$

$$\phi^{P}_{tt} = -\frac{1}{2} F\zeta^{2}_{,v} - G\zeta - \frac{1}{2} H\zeta^{2} \tag{176}$$

$$\phi^{P}_{vt} = J\zeta_{,v} \tag{177}$$

$$\phi^{P}_{tv} = \zeta_{,t}(F\zeta_{,v} + X\zeta) \tag{178}$$

The solution of the conservation equations (163), (164), (173), and (174) yields the conserved quantities of the ground state of renormalized relativistic thermodynamics.

The Noether tensor for the radiation energy density in a thermodynamic system is given by

$$\phi^{Er}_{vv} = \frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,v}} \xi_{r,v} - \mathcal{L}_{1r} \tag{179}$$

$$\phi^{Er}_{tt} = \frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,t}} \xi_{r,t} - \mathcal{L}_{1r} \tag{180}$$

$$\phi^{Er}_{vt} = \frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,t}} \xi_{r,v} \tag{181}$$

$$\phi^{Er}_{tv} = \frac{\partial \mathcal{L}_{1r}}{\partial \xi_{r,v}} \xi_{r,t} \tag{182}$$

which satisfy

$$\frac{\partial \phi^{Er}_{tv}}{\partial t} + \frac{\partial \phi^{Er}_{vv}}{\partial v} = 0 \tag{183}$$

$$\frac{\partial \phi^{Er}_{vt}}{\partial v} + \frac{\partial \phi^{Er}_{tt}}{\partial t} = 0 \tag{184}$$

where

$$\phi^{Er}_{vv} = \frac{1}{2} A_r \xi^2_{r,v} - I_r \xi_{r,t} - \xi_r (B_r - Z_r \xi_{r,v}) - \frac{1}{2} C_r \xi^2_r \tag{185}$$

$$\phi^{Er}_{tt} = -\frac{1}{2} A_r \xi^2_{r,v} - B_r \xi_r - \frac{1}{2} C_r \xi^2_r \tag{186}$$

$$\phi^{Er}_{vt} = I_r \xi_{r,v} \tag{187}$$

$$\phi^{Er}_{tv} = \xi_{r,t} (A_r \xi_{r,v} + Z_r \xi_r) \tag{188}$$

The Noether tensor for the radiation pressure in a thermodynamic system is given by

$$\phi^{Pr}_{vv} = \frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,v}} \zeta_{r,v} - \mathcal{L}_{2r} \tag{189}$$

$$\phi^{Pr}_{tt} = \frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,t}} \zeta_{r,t} - \mathcal{L}_{2r} \tag{190}$$

$$\phi^{Pr}_{vt} = \frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,t}} \zeta_{r,v} \tag{191}$$

$$\phi^{Pr}_{tv} = \frac{\partial \mathcal{L}_{2r}}{\partial \zeta_{r,v}} \zeta_{r,t} \tag{192}$$

while the conservation equations are

$$\frac{\partial \phi^{Pr}_{tv}}{\partial t} + \frac{\partial \phi^{Pr}_{vv}}{\partial v} = 0 \tag{193}$$

$$\frac{\partial \phi^{Pr}_{vt}}{\partial v} + \frac{\partial \phi^{Pr}_{tt}}{\partial t} = 0 \tag{194}$$

where

$$\phi^{Pr}_{vv} = \frac{1}{2} F_r \zeta^2_{r,v} - J_r \zeta_{r,t} - \zeta_r (G_r - X_r \zeta_{r,v}) - \frac{1}{2} H_r \zeta^2_r \tag{195}$$

$$\phi^{Pr}_{tt} = -\frac{1}{2} F_r \zeta^2_{r,v} - G_r \zeta_r - \frac{1}{2} H_r \zeta^2_r \tag{196}$$

$$\phi^{Pr}_{vt} = J_r \zeta_{r,v} \tag{197}$$

$$\phi^{Pr}_{tv} = \zeta_{r,t} (F_r \zeta_{r,v} + X_r \zeta_r) \tag{198}$$

The simultaneous solution of equations (183), (184), (193), and (194) yield the conserved quantities for radiation in matter.

7. VOID RATIOS FOR THE REAL GASES. In order to use the expression for the ground state void ratio given in equation (37) it is necessary to calculate the energy density and pressure of real gases for both the fractal and homogeneous cases. For the homogeneous case with D = 3, the renormalized pressure and energy density are given by [10,21]

$$P(3) = nRT[1 + nB^a + n^2 C(3) + \cdots] \tag{199}$$

$$E(3) = nRT[\frac{3}{2} - nT \frac{\partial B^a}{\partial T} - \frac{1}{2} n^2 T \frac{\partial C(3)}{\partial T} - \cdots] \tag{200}$$

where[10]

$$C(3) = C^a - 3(B^a)^2 \ell n \psi^a \tag{201}$$

where B^a and C^a = unrenormalized second and third virial coefficients respectively, and C(3) = renormalized third virial coefficient for the uniform D = 3 real gas, and where ψ^a [not to be confused with equation (16)] is a function of the second virial coefficient given by[10,21]

$$\psi^a = \frac{T}{T_R} \left| \frac{B^a(T)}{B^a(T_R)} \right|^{2/3} \tag{202}$$

where T_R = species dependent relativity temperature of real gases.[10] The corresponding state equations for a fractal real gas are written as

$$P(D) = nRT[1 + nB^a + n^2 C(D) + \cdots] \tag{203}$$

$$E(D) = nRT\left[\frac{3}{2} - nT\frac{\partial B^a}{\partial T} - \frac{1}{2}n^2T\frac{\partial C(D)}{\partial T} - \cdots\right] \tag{204}$$

where

$$C(D) = C^a - D(B^a)^2 \ell n\ \psi^a \tag{205}$$

In order to obtain equation (205) it is necessary to assume that D is independent of T and n.

Combining equations (199), (200), (203), and (204) with equation (37) gives

$$\frac{\Delta V}{V} = \frac{1}{3}n^2T\frac{\partial}{\partial T}\left[C(3) - C(D)\right] \tag{206}$$

where the following approximation was used in equation (37)

$$E(3) + P(3) - T\frac{\partial P(3)}{\partial T} \sim \frac{3}{2}nRT \tag{207}$$

Combining equations (201), (205), and (206) gives

$$\frac{\Delta V}{V} = -\frac{n^2}{3}(3 - D)T\frac{\partial}{\partial T}\left[(B^a)^2 \ell n\ \psi^a\right] + \cdots \tag{208}$$

$$= \frac{n^2}{9}(3 - D)T\frac{\partial}{\partial T}\left[C(3) - C^a\right] + \cdots$$

Note that in general

$$C(3) - C(D) = \frac{1}{3}(3 - D)\left[C(3) - C^a\right] \tag{209}$$

Thus voids will exist in the ground state of real gases only in the temperature intervals for which $\Delta V/V > 0$ in equation (208). This condition gives

$$\frac{\partial}{\partial T}\left[(B^a)^2 \ell n\ \psi^a\right] < 0 \tag{210}$$

or equivalently

$$\frac{\partial}{\partial T}\left[C(3) - C^a\right] > 0 \tag{211}$$

as shown in Figure 1. From Figure 1 it is clear that the largest size voids will occur at low temperatures. There is also a narrow fractal region just above the Boyle temperature, and a broad fractal region at high temperatures.

Ultimately, the voids that occur in real gases are due to the interaction of these gases with the vacuum state, which manifests itself in the third and higher virial coefficients. The ideal gas is not fractal.

Consider now mechanical radiation in a real gas. For the homogeneous case with $D = 3$ and $d_r = 0$ the renormalized radiation pressure and energy density are[21]

$$P_r(3,0) = nRT\left[\frac{1}{12}k_o^2A_o^2 + nB_r^a + n^2C_r(3,0) + \cdots\right] \tag{212}$$

$$E_r(3,0) = nRT\left[\frac{1}{4}k_o^2A_o^2 - nT\frac{\partial B_r^a}{\partial T} - \frac{1}{2}n^2T\frac{\partial C_r(3,0)}{\partial T} - \cdots\right] \tag{213}$$

where k_o and A_o = wave number and amplitude of waves in an ideal gas, and where B_r^a = nonrelativistic (unrenormalized) second radiation virial coefficient, and $C_r(3,0)$ = relativistic third virial coefficient for a homogeneous ($D = 3$ and $d_r = 0$) mechanical radiation field given by[21]

$$C_r(3,0) = C_r^a - 3\left[2B^aB_r^a + (B_r^a)^2\right]\ell n\,\psi^a - 3(B^a + B_r^a)^2\,\ell n\left(1 + \frac{\psi_r^a}{\psi^a}\right) \tag{214}$$

where ψ^a is given by equation (202) and ψ_r^a [not to be confused with equation (51)] is given by[21]

$$\psi^a + \psi_r^a = \frac{T}{T_R}\left|\frac{B^a(T) + B_r^a(T)}{B^a(T_R) + B_r^a(T_R)}\right|^{2/3} \tag{215}$$

The procedure for calculating B_r^a and C_r^a is given in Reference 21. The corresponding state equations for fractal radiation in a fractal real gas are given by

$$P_r(D,d_r) = nRT\left[\frac{1}{12}k_o^2A_o^2 + nB_r^a + n^2C_r(D,d_r) + \cdots\right] \tag{216}$$

$$E_r(D,d_r) = nRT\left[\frac{1}{4}k_o^2A_o^2 - nT\frac{\partial B_r^a}{\partial T} - \frac{1}{2}n^2T\frac{\partial C_r(D,d_r)}{\partial T} - \cdots\right] \tag{217}$$

where $C_r(D,d_r)$ = relativistic third virial coefficient for a fractal mechanical radiation field in a fractal real gas and is given by

$$C_r(D,d_r) = C_r^a - D[2B^aB_r^a + (B_r^a)^2]\ \ell n\ \psi^a - D(B^a + B_r^a)^2\ \ell n\left(1 + \frac{\psi_r^a}{\psi^a}\right) \quad (218)$$

$$+ d_r(B^a)^2\ \ell n\ \psi^a$$

where the fractal dimension of the mechanical radiation in a real gas is now $D_r = D - d_r$ which is lower than the fractal dimension D of the ground state of a real gas. Note that in order to obtain equation (218) it is assumed that D and d_r are independent of T and V.

The calculation of the void ratio for radiation in a real gas then proceeds from equation (67). Note first the following approximation

$$E_r(3,0) + P_r(3,0) - T\frac{\partial P_r(3,0)}{\partial T} \sim \frac{1}{4}\,nRTk_o^2A_o^2 \quad (219)$$

Combining equations (213) and (217) gives

$$E_r(D,d_r) - E_r(3,0) = \frac{1}{2}\,n^3RT^2\frac{\partial}{\partial T}\left[C_r(3,0) - C_r(D,d_r)\right] \quad (220)$$

Placing equation (219) and (220) into equation (67) yields

$$\left(\frac{\Delta V}{V}\right)_r = \frac{2n^2}{k_o^2A_o^2}\,T\frac{\partial}{\partial T}\left[C_r(3,0) - C_r(D,d_r)\right] + \cdots \quad (221)$$

where

$$C_r(3,0) - C_r(D,d_r) = -(3 - D)\left\{[2B^aB_r^a + (B_r^a)^2]\ \ell n\ \psi^a + (B^a + B_r^a)^2\ \ell n\left(1 + \frac{\psi_r^a}{\psi^a}\right)\right\}$$

$$- d_r(B^a)^2\ \ell n\ \psi^a \quad (222)$$

Placing equations (201) and (214) into equation (222) yields

$$C_r(3,0) - C_r(D,d_r) = \frac{1}{3}\,(3 - D)[C_r(3,0) - C_r^a] + \frac{d_r}{3}[C(3) - C^a] \quad (223)$$

Combining equations (221) and (223) gives the following expression for void ratio of mechanical radiation

$$\left(\frac{\Delta V}{V}\right)_r = \frac{2n^2}{3k_o^2A_o^2}\left\{(3 - D)T\frac{\partial}{\partial T}[C_r(3,0) - C_r^a] + d_rT\frac{\partial}{\partial T}[C(3) - C^a]\right\} + \cdots \quad (224)$$

It has been shown that the coefficients B_r^a , B_r , C_r^a , and $C_r(3,0)$ are all proportional to $k_o^2A_o^2$.[21] In addition, the radiation fractal decrement d_r is proportional to the radiation energy density so that

$$d_r = \tau_r E_r(D,d_r) \quad (225)$$

$$= \frac{1}{4}\tau_r nRTk_o^2A_o^2 + \cdots .$$

where τ_r is independent of k_o and A_o . It then follows that

$$C_r(3,0) = k_o^2A_o^2C_r'(3,0) \quad (226)$$

$$C_r(D,d_r) = k_o^2A_o^2C_r'(D,d_r) \quad (227)$$

$$C_r^a = k_o^2A_o^2C_r^{a'} \quad (228)$$

where $C_r^{a'}$, $C_r'(3,0)$, and $C_r'(D,d_r)$ are independent of k_o and A_o . Therefore equation (224) can be written as

$$\left(\frac{\Delta V}{V}\right)_r = \frac{2}{3}n^2\left\{(3 - D)T\frac{\partial}{\partial T}[C_r'(3,0) - C_r^{a'}] + \frac{1}{4}\tau_r nRT^2\frac{\partial}{\partial T}[C(3) - C^a]\right\} + \cdots \quad (229)$$

The void ratio for mechanical radiation in real gases is independent of wave amplitude and frequency, and depends only on temperature and density. For the case of fractal mechanical radiation in a homogeneous (D = 3) ground state, equation (229) becomes

$$\left(\frac{\Delta V}{V}\right)_r = \frac{1}{6}\tau_r n^3RT^2\frac{\partial}{\partial T}[C(3) - C^a] + \cdots \quad (230)$$

Equation (230) is somewhat similar to the result for the ground state in equation (208).

Even if the ground state is homogeneous (D = 3) the mechanical radiation state can be fractal with fractal dimension $D_r = 3 - d_r$. In general, however, the ground state may be fractal, and the fractal dimension of the radiation

field will be written as $D_r = D - d_r$. Only in limited temperature regions will mechanical radiation in real gases have fractal properties, i.e., those regions for which $(\Delta V/V)_r > 0$ in equation (229). The fractal dimension D_r of mechanical radiation can be calculated from the general radiation Lagrangian formalism given in Section 5. The voids in a mechanical radiation field are due to the interaction of the excitations of a real gas with the vacuum state. Mechanical radiation in an ideal gas is not fractal.

8. VOID RATIOS FOR FRACTAL SOLIDS AND QUANTUM LIQUIDS. In this section the void ratios for the fractal ground state and excited states of solids and quantum liquids are calculated. For simplicity only the T = 0 state will be considered. In order to use equation (38) for the calculation of the ground state void ratio, the T = 0 energy density and pressure must be calculated for both the fractal and homogeneous states. It will be assumed that

$$E_o^a = An^{\sigma_o} \tag{231}$$

$$P_o^a = (\sigma_o - 1)E_o^a \tag{232}$$

where A and σ_o = constants independent of density. Using equation (29) with γ_o and D_o taken to be constants independent of density gives the renormalized energy densities of the fractal and homogeneous systems respectively as

$$E_o(D_o) = \frac{E_o^a}{G(D_o)} \tag{233}$$

$$E_o(3) = \frac{E_o^a}{G(3)} \tag{234}$$

where

$$G(D_o) = D_o\sigma_o^2 - D_o\sigma_o(2 + \gamma_o) + D_o(1 + \gamma_o) + 1 \tag{235}$$

$$G(3) = 3\sigma_o^2 - 3(2 + \gamma_o)\sigma_o + 3\gamma_o + 4 \tag{236}$$

Then the difference in energy densities for the fractal and homogeneous states is given by

$$E_o(D_o) - E_o(3) = \frac{(3 - D_o)F_oE_o^a}{G(3)G(D_o)} \tag{237}$$

where

$$F_o = \sigma_o^2 - (2 + \gamma_o)\sigma_o + \gamma_o + 1 \tag{238}$$

which satisfies

$$G(D_o) = D_o F_o + 1 \tag{239}$$

$$G(3) = 3F_o + 1 \tag{240}$$

Combining equations (231) and (234) gives

$$n \frac{dE_o(3)}{dn} = \frac{\sigma_o E_o^a}{G(3)} \tag{241}$$

and therefore equation (38) gives the following expression for the T = 0 ground state void ratio for solids and quantum liquids

$$\left(\frac{\Delta V}{V}\right)_o = \frac{[E_o(D_o) - E_o(3)]G(3)}{\sigma_o E_o^a} = \frac{(3 - D_o)F_o}{\sigma_o G(D_o)} \tag{242}$$

In addition to the obvious dependence of D_o , the void ratio depends on the constants σ_o and γ_o .

Often in the literature the average energy per particle is introduced as follows[10]

$$\epsilon_o^a = E_o^a/n = An^{\sigma_o - 1} = Ax^{\kappa} \tag{243}$$

where x is defined by $n = x^3$. This gives

$$\sigma_o = \kappa/3 + 1 \tag{244}$$

Using the parameter κ gives[10]

$$G(3) = \frac{\kappa^2}{3} - \kappa\gamma_o + 1 \tag{245}$$

$$G(D_o) = \frac{D_o\kappa^2}{9} - \frac{\kappa D_o\gamma_o}{3} + 1 \tag{246}$$

$$F_o = \frac{\kappa}{3}\left(\frac{\kappa}{3} - \gamma_o\right) \tag{247}$$

These functions are expressed in terms of both κ and γ_o .

Of particular interest are the cases of the non-relativistic and ultra-relativistic non-interacting degenerate T = 0 Fermi gases. The non-relativistic Fermi gas has the following properties

$$\gamma_o = 2/3 \qquad \kappa = 2 \qquad \sigma_o = 5/3 \tag{248}$$

$$G(3) = 1 \qquad G(D_o) = 1 \qquad F_o = 0 \tag{249}$$

For the ultra-relativistic case these quantities are

$$\gamma_o = 1/3 \qquad \kappa = 1 \qquad \sigma_o = 4/3 \tag{250}$$

$$G(3) = 1 \qquad G(D_o) = 1 \qquad F_o = 0$$

Because $F_o = 0$ for the ideal non-relativistic and ultra-relativistic Fermi gases, it is clear from equation (242) that no voids exist for these cases. Ideal non-relativistic and ultra-relativistic Fermi gases are homogeneous with $D_o = 3$.

In general the Grüneisen parameter γ_o is a function of the index κ . For instance, the effective mass approximation for an interacting system gives the following expression for the zero-temperature Grüneisen parameter[10]

$$\gamma_o \sim \frac{\kappa - 4}{3} = \sigma_o - \frac{7}{3} \tag{252}$$

Using this relationship gives[10]

$$G(3) \sim 1 + \frac{4\kappa}{3} = 4\sigma_o - 3 > 0 \tag{253}$$

$$G(D_o) \sim 1 + \frac{4\kappa D_o}{9} = \frac{4}{3} D_o \sigma_o - \left(\frac{4}{3} D_o - 1\right) > 0 \tag{254}$$

$$F_o \sim \frac{4\kappa}{9} = \frac{4}{3}(\sigma_o - 1) > 0 \tag{255}$$

Equation (252) is valid only for $\kappa \geq 5$ or $\sigma_o \geq 8/3$. In general the value of D_o for solids or quantum liquids may be determined experimentally or possibly theoretically from a T = 0 Lagrangian formulation outlined in Section 4. In

general T = 0 solids and quantum liquids are fractal for all physical densities because equation (242) shows that $(\Delta V/V)_o > 0$.

Consider now the void ratio of mechanical waves in a solid or quantum liquid. For simplicity only waves in a T = 0 system will be considered. The wave equation (59) can be simplified by using equations (231) through (234) and the approximation $E_{jr}/E_j = E_{or}/E_o$ with the result that

$$D_o \frac{E_{jr}}{E_j} P_o(\gamma_o - \gamma_{or}) = D_o E_{or}(\sigma_o - 1)(\gamma_o - \gamma_{or}) \tag{256}$$

and therefore equation (65) can be rewritten as

$$D_o n^2 \frac{d^2 E_{or}}{dn^2} - D_o(1 + \gamma_o) n \frac{dE_{or}}{dn} + C_o E_{or} = E^a_{or} \tag{257}$$

where

$$C_o = D_o \sigma_o(\gamma_o - \gamma_{or}) + D_o(1 + \gamma_{or}) + 1 + \tau_{or}[(1 + \gamma_o)P_o - K_o] \tag{258}$$

and where the relation $d_{or} = \tau_{or} E_{or}$ is used which is similar to equation (225) that was used for the real gases. Assume now that[9]

$$E_{or} = \frac{1}{4} K_o k^2 A^2 \qquad E^a_{or} = \frac{1}{4} K^a_o k^2_a A^2_a \tag{259}$$

where k_a and A_a = nonrelativistic wave number and wave amplitude respectively, and k and A = relativistic wave number and wave amplitude respectively. Placing equation (259) into equation (257) gives

$$\frac{D_o}{4} K_o n^2 \frac{d^2}{dn^2}(k^2A^2) + \frac{D_o}{4}\left[2n \frac{dK_o}{dn} - (1 + \gamma_o)K_o\right] n \frac{d}{dn}(k^2A^2) \tag{260}$$

$$+ \frac{1}{4} k^2 A^2 K_o G_r(D_o, \tau_{or}) = \frac{1}{4} k^2_a A^2_a K^a_o$$

where

$$G_r(D_o, \tau_{or}) = D_o \frac{n^2}{K_o} \frac{d^2 K_o}{dn^2} - D_o(1 + \gamma_o) \frac{n}{K_o} \frac{dK_o}{dn} + C_o \tag{261}$$

Using $K_o \sim n^{\sigma_o}$ in equation (261) gives

$$G_r(D_o,\tau_{or}) = D_o\sigma_o^2 - D_o\sigma_o(2 + \gamma_o) + C_o \tag{262}$$

$$= D_o\sigma_o^2 - D_o\sigma_o(2 + \gamma_{or}) + D_o(1 + \gamma_{or}) + 1 - \tau_{or}[K_o - (1 + \gamma_o)P_o]$$

If kA is taken to be independent of density it follows from equation (260) that

$$k^2A^2 = \frac{k_a^2A_a^2K_o^a/K_o}{G_r(D_o,\tau_{or})} \tag{263}$$

Combining equations (34) and (263) gives

$$k^2A^2 = k_a^2A_a^2G(D_o)/G_r(D_o,\tau_{or}) \tag{264}$$

Similarly, from equation (260) it follows that

$$E_{or}(D_o,\tau_{or}) = E_{or}^a/G_r(D_o,\tau_{or}) \tag{265}$$

$$E_{or}(3,0) = E_{or}^a/G_r(3,0) \tag{266}$$

where

$$G_r(3,0) = 3\sigma_o^2 - 3\sigma_o(2 + \gamma_{or}) + 3\gamma_{or} + 4 \tag{267}$$

Taking $D_o = 3$, $\tau_{or} = 0$, and $\gamma_{or} = 1/3$ in equations (235), (262), and (264) yields equation (121) of Reference 9.

The calculation of the void ratio for mechanical radiation in a fractal solid or quantum liquid follows from equation (68). First determine the enhanced energy of the fractal radiation state from equations (265) and (266) as follows

$$E_{or}(D_o,\tau_{or}) - E_{or}(3,0) = \frac{E_{or}^a[G_r(3,0) - G_r(D_o,\tau_{or})]}{G_r(3,0)G_r(D_o,\tau_{or})} \tag{268}$$

where

$$G_r(3,0) - G_r(D_o,\tau_{or}) = (3 - D_o)F_{or} - \tau_{or}[(1 + \gamma_o)P_o - K_o] \tag{269}$$

$$F_{or} = \sigma_o^2 - \sigma_o(2 + \gamma_{or}) + \gamma_{or} + 1 \tag{270}$$

Calculate also the following expression

$$E_{or}(3,0) + P_{or}(3,0) = n\frac{dE_{or}(3,0)}{dn} = \frac{\sigma_o E^a_{or}}{G_r(3,0)} \tag{271}$$

Then equation (68) gives the void ratio for mechanical waves as follows

$$\left(\frac{\Delta V}{V}\right)_{or} = \frac{(3 - D_o)F_{or} + \tau_{or}[K_o - (1 + \gamma_o)P_o]}{\sigma_o G_r(D_o,\tau_{or})} \tag{272}$$

Note that

$$G_r(3,0) = 3F_{or} + 1 \tag{273}$$

$$G_r(D_o,\tau_{or}) = D_o F_{or} + 1 + \tau_{or}[(1 + \gamma_o)P_o - K_o] \tag{274}$$

A homogeneous ground state ($D_o = 3$) can have a fractal radiation excited state described by

$$\left(\frac{\Delta V}{V}\right)_{or} = \frac{\tau_{or}[K_o - (1 + \gamma_o)P_o]}{\sigma_o G_r(3,\tau_{or})} \tag{275}$$

where

$$G_r(3,\tau_{or}) = 3\sigma_o^2 - 3\sigma_o(2 + \gamma_{or}) + 3\gamma_{or} + 4 - \tau_{or}[K_o - (1 + \gamma_o)P_o] \tag{276}$$

Both equations (272) and (275) are valid only for $(\Delta V/V)_{or} > 0$ and this restricts the density regions in which radiation voids are possible. Only in the unphysical regions below equilibrium density are equations (272) or (275) negative. In fact at the equilibrium density of a T = 0 solid or quantum liquid one has $P_o = 0$ so that $(\Delta V/V)_{or} > 0$ at this point. Radiation voids are also possible in the high density regions beyond the equilibrium density of an interacting system. Note that for ideal non-relativistic and ultra-relativistic T = 0 Fermi gases $(1 + \gamma_o)P_o - K_o = 0$, and $D_o = 3$, and from equation (272) the radiation state is homogeneous.

9. CONCLUSION. The renormalization group equations for fractal matter and fractal mechanical radiation are presented and a Lagrangian formulation of these equations is developed. For limited temperature and density regions the equilibrium fractal dimensions of the ground and excited states of real gases, solids, and quantum liquids may possibly be obtained by minimizing the Lagrangian density with respect to the fractal dimension of the system. Ideal non-relativistic and ultra-relativistic quantum thermodynamic systems are not fractal. For interacting thermodynamic systems the ground and excited states are fractal in nature. The ground and excited states of real gases are fractal only in limited temperature regions. Although in general $D \neq 3$ it has not been shown that the voids in matter and mechanical radiation fields are self similar, which is a basic characteristic of fractal systems.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Goldstein, H., Classical Mechanics, Addison-Wesley, New York, 1980.

2. Lanczos, C., The Variational Principles of Mechanics, University of Toronto Press, Toronto, 1960.

3. Yourgrau, W. and Mandelstam, S., Variational Principles in Dynamics and Quantum Theory, Pitman, London, 1955.

4. Mandl, F. and Shaw, G., Quantum Field Theory, John Wiley, New York, 1984.

5. Bogoliubov, N. and Shirkov, D., Introduction to the Theory of Quantized Fields, Interscience, New York, 1959.

6. Lee, T., Particle Physics and Introduction to Field Theory, Harwood, New York, 1981.

7. Aitchison, I. and Hey, A., Gauge Theories in Particle Physics, Adam Hilger, Bristol, 1982.

8. Weiss, R., "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.

9. Weiss, R., "Scale Invariant Equations for Relativistic Waves," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, ARO 87-1, May 27-30, 1986, p. 307.

10. Weiss, R., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

11. Pietronero, L. and Tosatti, E., editors, Fractals in Physics, North-Holland-Elsevier, New York, 1986.

12. Stanley, H. and Ostrowsky, N., editors, On Growth and Form, Martinus Nijhoff, Boston, 1986.

13. Family, F. and Landau, D., editors, Kinetics of Aggregation and Gelation, North-Holland, Amsterdam, 1984.

14. Pynn, R. and Skjeltorp, A., editors, Scaling Phenomena in Disordered Systems, Plenum, New York, 1985.

15. Boccara, N. and Daoud, M., editors, Physics of Finely Divided Matter, Springer, Berlin, 1985.

16. de Gennes, P., Scaling Concepts in Polymer Physics, Cornell Univ. Press, Ithaca, 1979.

17. Mandelbrot, B., The Fractal Geometry of Nature, Freeman, San Francisco, 1983.

18. Müller, B. and Schäfer, A., "Improved Bounds on the Dimension of Space-Time, Phys. Rev. Lett. 56, 1215, 1986.

19. Hirth, J. and Lothe, J., Theory of Dislocations, McGraw-Hill, New York, 1968.

20. Quigg, C., Gauge Theories of the Strong, Weak, and Electromagnetic Interactions, Addison-Wesley, New York, 1983.

21. Weiss, R., "Relativistic Wave Equations for Real Gases," Fourth Army Conference on Applied Mathematics and Computing, Cornell University, ARO 87-1, May 27-30, 1986, p. 341.

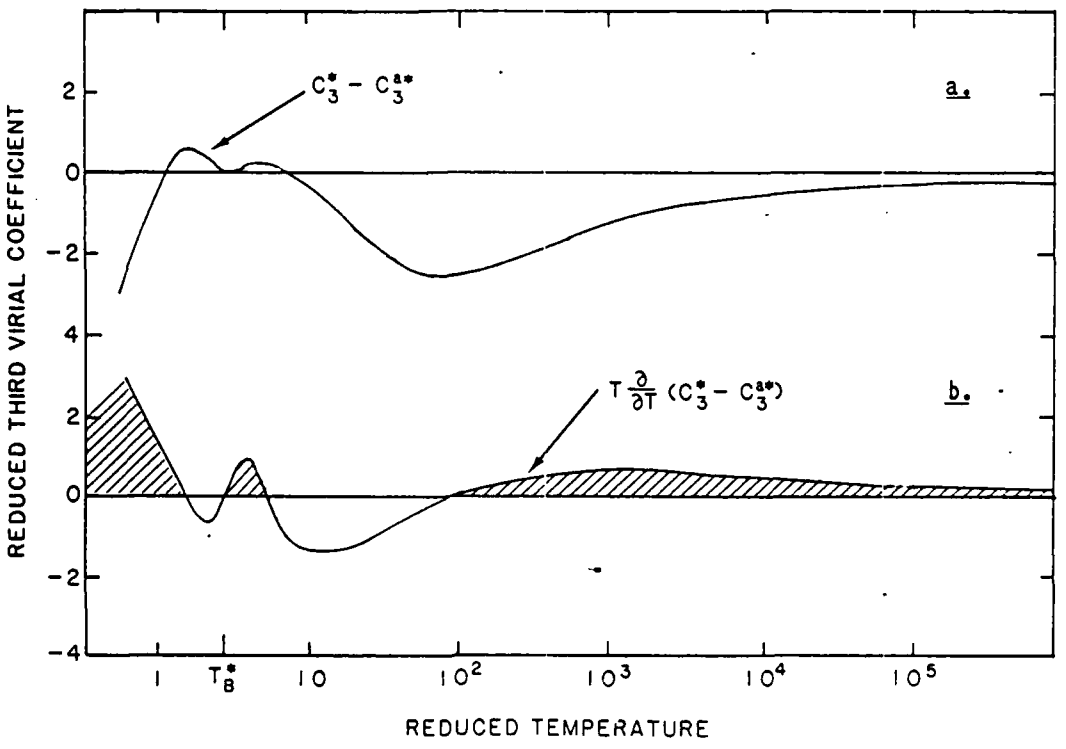


Figure 1. a) $C_3^* - C_3^{a*}$ as a function of reduced temperature; b) $T\partial/\partial T(C_3^* - C_3^{a*})$ versus reduced temperature. When positive, this indicates the regions (shaded areas) where voids are possible for real gases.

MODELLING OF THE LEAN FLAMMMABILITY LIMIT IN FLAME THEORY[1]

Richard Y. Tam
Department of Mathematical Sciences
Purdue University School of Science at Indianapolis
Indianapolis, IN 46223

and

G. S. S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University
Ithaca, NY 14853

ABSTRACT. The phenomenon of the lean flammability limit is modelled by a four-step reaction mechanism. Analytical results, namely that a fuel mixture will not burn if it is too lean, are obtained through use of activation-energy asymptotics.

I. INTRODUCTION

The use of the one-step irreversible reaction in combustion modelling has wielded much success, particularly in the context of activation-energy asymptotics [1]. However, the neglect of radicals, or intermediates as they are sometimes called, has precluded the modelling of some important phenomena. One of the these, the lean flammability limit, is the subject of study in this paper.

By the lean flammability limit, we mean the phenomenon whereby a fuel mixture is incapable of burning when it is too lean. Of course, a mixture may not burn due to other causes, e.g., excessive heat loss, flow divergence, etc., the modelling of which has been successfully done by the use of the one-step model. The lean flammability limit, however, differs from such external effects in that it involves a property of the mixture itself, namely the fuel strength. It has been conjectured that the cause lies in the chemistry, i.e., the reaction mechanism, thus necessitating the use of multi-step kinetics in the modelling of this phenomenon.

The two-step Zeldovich-Liñan mechanism [2] and other simple multi-step schemes studied by Fife and Nicolaenko [3] uncovered many flame phenomena, among which are stretch-resistance ([4], [5]), hysterisis, flame plateaux and kinetic extinction (op. cit.), but they cannot model the lean flammability limit. Peters and Smooke [6] attempted to model the phenomenon by using a four-step model:

$$F + R_1 \longrightarrow R_2$$

$$R_2 + O_2 \longrightarrow 2R_1$$

$$R_2 + O_2 + M \longrightarrow P_1 + M$$

$$2R_1 + M \longrightarrow P_2 + M \quad .$$

---

[1]This work was supported, in part, by the U. S. Army Research Office.

Here F is the fuel, O_2 the oxidant, R_1 and R_2 the radicals, P_1 and P_2 the products, and M a third body. Close examination of their results, however, shows that the phenomenon modelled there was the kinetic extinction phenomenon whereby a mixture, if it burns at all, burns at or above a critical temperature. We shall not discuss the details here, but merely note that they can be found in [7].

We consider here the four-step model

$$F + R_1 \longrightarrow R_2 + C \qquad E_1 \longrightarrow \infty$$

$$R_2 + C \longrightarrow 2R_1 \qquad E_2 \longrightarrow \infty$$

$$R_2 + O_2 \longrightarrow 2P_1 \qquad E_3 \longrightarrow 0$$

$$2R_1 + M \longrightarrow 2P_2 + M \qquad E_4 \longrightarrow 0$$

Here F and O_2 are respectively fuel and oxidant, R_1, R_2 and C are intermediates, P_1 and P_2 are products and M is a third body. The first two reactions have high activation energy while the last two have small (taken as zero here) activation energy.

II. GOVERNING EQUATIONS

To focus on the chemistry, we consider the model in the context of the steady plane flame, which is governed by the following dimensionless system of equations:

$$\frac{dY}{dx} - \mathcal{L}^{-1}\frac{d^2Y}{dx^2} = -\mathcal{D}_1 RYe^{-\theta/T} \tag{1}$$

$$\frac{dR}{dx} - \mathcal{K}^{-1}\frac{d^2R}{dx^2} = -\mathcal{D}_1 RYe^{-\theta/T} + 2\mathcal{D}_2 SCe^{-r\theta/T} - \mathcal{D}_4 R^2 B \tag{2}$$

$$\frac{dS}{dx} - \mathcal{N}^{-1}\frac{d^2S}{dx^2} = \mathcal{D}_1 RYe^{-\theta/T} - \mathcal{D}_2 SCe^{-r\theta/T} - \mathcal{D}_3 SX \tag{3}$$

$$\frac{dC}{dx} - \mathcal{Q}^{-1}\frac{d^2C}{dx^2} = \mathcal{D}_1 RYe^{-\theta/T} - \mathcal{D}_2 SCe^{-r\theta/T} \tag{4}$$

$$\frac{dX}{dx} - \mathcal{P}^{-1}\frac{d^2X}{dx^2} = -\mathcal{D}_3 SX \tag{5}$$

$$\frac{dT}{dx} - \frac{d^2T}{dx^2} = q_1\mathcal{D}_1 RYe^{-\theta/T} + q_2\mathcal{D}_2 SCe^{-r\theta/T} + q_3\mathcal{D}_3 SX$$

$$+ q_4\mathcal{D}_4 R^2 B \qquad (6)$$

Here Y, R, S, C, X, and B, respectively, are the normalized mass fractions of the fuel F, radicals R_1, R_2, and C, oxidant O_2 and third body M, T is the temperature; $\mathcal{L}$, $\mathcal{K}$, $\mathcal{N}$, $\mathcal{Q}$, and $\mathcal{P}$ are the Lewis numbers (assumed constant) of F, R_1, R_2, C, and X respectively. The non-dimensionalized heat releases of the four reactions are q_1, q_2, q_3, and q_4 respectively, whereas $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ are the Damköhler numbers, assumed independent of temperature. The non-dimensionalized activation energy θ is derived from the first reaction and $r = E_2/E_1$, the ratio of the activation energies of the second and first reactions, is taken to be less than $\frac{1}{2}$ for simplicity. These equations are to be solved under the boundary conditions

$$Y, R, S, C, X, T \longrightarrow Y_f, 0, 0, 0, X_f, T_f \quad \text{as} \quad x \longrightarrow -\infty \qquad (7)$$

boundedness is required as $x \longrightarrow +\infty$.

Assuming that the Damköhler numbers are ordered: $\mathcal{D}_3 \ll \mathcal{D}_4 \ll \mathcal{D}_2 \ll \mathcal{D}_1$ and denoting by T_{23} the flame temperature such that

$$\mathcal{D}_2 e^{-r\theta/T_{23}} = \mathcal{D}_3 ,$$

one can show that no flame exists with flame temperature below T_{23}; a detailed discussion can be found in [8]. Apparently, this is the kinetic extinction phenomenon; analysis of the flame structure, however, reveals more. We now focus our attention on temperatures close to T_{23}.

III. ASYMPTOTIC ANALYSIS

Solution to the system (1) - (7) will be sought in the limit of infinite activation energy, whereupon all reactions are confined within the reaction zone where the appropriate coordinate is $\xi = \theta x$. The mass fractions aree ordered

$$Y = \theta^{-1}\tilde{Y}, \quad R = \theta^{-1}\frac{\mathcal{D}_1}{\mathcal{D}_4} e^{-\theta/T_*}\tilde{R}, \quad S = \theta^{-2}\frac{\mathcal{D}_1^2}{\mathcal{D}_2\mathcal{D}_4} e^{\theta(r-2)/T_*}\tilde{S},$$

$$C = C_b + \theta^{-1}\tilde{C}, \quad X = X_b + \theta^{-1}\tilde{X}, \quad T = T_* + \theta^{-1}\tilde{T}$$

where the O(1) variables in the ξ-structure are designated by a tilde; and the subscript b denotes the burnt state. Since for T_* close to T_{23},

$$\mathcal{D}_1 e^{-\theta/T_*} \ll \mathcal{D}_2 e^{-r\theta/T_*} \approx \mathcal{D}_3 \ll \mathcal{D}_4 .$$

R and S are exponentially small; more precisely, the radicals R_1 and R_2 are produced and consumed entirely within the reaction zone, so that on the x-scale,

$$Y = \begin{matrix} Y_f \\ 0 \end{matrix} (1 - e^{\mathcal{Q}x}) , \quad R = S = 0 , \quad C = \begin{matrix} C_b \, e^{\mathcal{Q}x} \\ C_b \end{matrix} . \tag{8}$$

$$X = \begin{matrix} X_b + (X_b - X_f)e^{\mathcal{P}x} \\ X_b \end{matrix} , \quad T = \begin{matrix} T_f + (T_* - T_f)\, e^{x} \\ T_* \end{matrix}$$

for $x \lessgtr 0$ respectively. Also, we find

$$C_b = X_f - X_b . \tag{9}$$

The governing equations in the reaction zone are, to leading order:

$$-\mathcal{L}^{-1} \frac{d^2\tilde{Y}}{d\xi^2} = -D \tilde{R} \tilde{Y} e^{\tilde{T}/T_*^2} \tag{10}$$

$$0 = -D\tilde{R} \tilde{Y} e^{\tilde{T}/T_*^2} + 2D\tilde{S}C_b e^{r\tilde{T}/T_*^2} - D\tilde{R}^2 B \tag{11}$$

$$0 = D\tilde{R} \tilde{Y} e^{\tilde{T}/T_*^2} - D\tilde{S}C_b e^{r\tilde{T}/T_*^2} - D\tilde{S}X_b \tag{12}$$

$$-\mathcal{Q}^{-1} \frac{d^2\tilde{C}}{d\xi^2} = D\tilde{R} \tilde{Y} e^{\tilde{T}/T_*^2} - D\tilde{S}C_b e^{r\tilde{T}/T_*^2} \tag{13}$$

$$-\mathcal{P}^{-1} \frac{d^2\tilde{X}}{d\xi^2} = -D\tilde{S}X_b \tag{14}$$

$$-\frac{d^2\tilde{T}}{d\xi^2} = q_1 D\tilde{R} \tilde{Y} e^{\tilde{T}/T_*^2} + q_2 D\tilde{S}C_b e^{r\tilde{T}/T_*^2} + q_3 D\tilde{S}X_b + q_4 D\tilde{R}^2 B \tag{15}$$

Here

$$D = \frac{\bar{D}}{M_r^2} = \theta^{-3} \frac{\mathcal{D}_1^2}{\mathcal{D}_4} e^{-2\theta/T_*} = \theta^{-3} \frac{D_1}{D_4} e^{-2\theta/T_*/M_r^2}$$

is an O(1) parameter and is to be determined as part of the solution for M_r, the reference mass flux.

Equations (11) and (12) express what may be called the local equilibria of the two radical species R_1 and R_2 : given $\tilde{Y}$, we find

$$\tilde{R} = \tilde{S} = 0 \tag{16}$$

is always possible, whereas there is a second solution

$$\tilde{R} = \frac{C_b e^{r\tilde{T}/T_*^2} - X_b}{C_b e^{r\tilde{T}/T_*^2} + X_b} \cdot \frac{\tilde{Y}_e T/T_*^2}{B}$$

$$\tilde{S} = \frac{\tilde{R}\,\tilde{Y}_e \tilde{T}/T_*^2}{C_b e^{r\tilde{T}/T_*^2} + X_b} \tag{17}$$

for $C_b e^{r\tilde{T}/T_*^2} - X_b > 0$. Since radicals must be produced in the reaction zone, the second solution must hold in some part of the reaction zone; on the other hand, that part cannot lie on the fresh side of the reaction zone since $\tilde{R}$ and $\tilde{S}$ must vanish there and $\tilde{Y}$ does not.

We conclude that the reaction zone is divided at $\xi = \xi^*$ (say) such that the solution (16) holds for $\xi \leq \xi^*$ and (17) holds for $\xi > \xi^*$. At $\xi = \xi^*$, continuity of the solutions require that

$$\tilde{T}(\xi^*) = \frac{T_*^2}{r} \ln\left(\frac{X_b}{C_b}\right) .$$

Since the maximum temperature is attained as $\xi \longrightarrow +\infty$, it follows that

$$X_b < C_b$$

or, equivalently, because of (9),

$$X_b < \frac{1}{2} X_f . \tag{18}$$

This condition in turn implies that Y_f must be greater than a certain value for the solution to exist, as will be shown later.

Substitution of the local equilibria expressions (17) into the fuel, temperature and oxidant equations (10), (15), and (14) yield

$$-\mathcal{L}^{-1}\,\frac{d^2\tilde{Y}}{d\xi^2} = D\,\frac{C_b e^{r\tilde{T}/T_*^2} - X_b}{C_b e^{r\tilde{T}/T_*^2} + X_b}\cdot\frac{\tilde{Y}^2 e^{2\tilde{T}/T_*^2}}{B} \tag{19}$$

$$\frac{d^2\tilde{T}}{d\xi^2} = D\left[(q_1 + q_3 - q_4) + (q_2 - q_3 + 2q_4)\,\frac{C_b e^{r\tilde{T}/T_*^2}}{C_b e^{r\tilde{T}/T_*^2} + X_b}\right] \cdot \frac{C_b e^{r\tilde{T}/T_*^2} - X_b}{C_b e^{r\tilde{T}/T_*^2} + X_b}\cdot\frac{\tilde{Y}^2 e^{2\tilde{T}/T_*^2}}{B} \tag{20}$$

$$-\mathcal{P}^{-1}\,\frac{d^2\tilde{X}}{d\xi^2} = -D\,\frac{C_b e^{r\tilde{T}/T_*^2} - X_b}{\left[C_b e^{r\tilde{T}/T_*^2} + X_b\right]^2}\cdot\frac{\tilde{Y}^2 e^{2\tilde{T}/T_*^2}}{B} \tag{21}$$

Boundary conditions obtained from matching with the outer solution (8) are

$$\begin{aligned} &\frac{d\tilde{Y}}{d\xi} = -\mathcal{L}^{-1}\,Y_f + o(1)\,,\quad \frac{d\tilde{T}}{d\xi} = (T_* - T_f) + o(1)\,,\ \text{and} \\ &\frac{d\tilde{X}}{dx} = \mathcal{P}(X_b - X_f) + o(1) \qquad\qquad \text{as } \xi \downarrow \xi^* \\ &\tilde{Y} = o(1)\,,\quad \tilde{T} = o(1)\,,\ X = o(1)\,,\ \text{as } \xi \longrightarrow +\infty\,. \end{aligned} \tag{22}$$

The above system is then solved for the special case

$$q_1 - q_3 + 2q_4 = 0\,, \qquad q_1 + q_3 - q_4 > 0$$

for which an analytical solution is possible. We shall not give the details here but only quote the final results. We find

$$M_r^2 = \frac{\mathcal{L}^2\,T_*^6\,\bar{D}}{4(q_1 + q_2 + q_4)^3\,Y_f^2\,B}\,I(\tilde{y}^*\,;\,r,\,\tilde{y}^*) \tag{23}$$

$$X_f - X_b = \frac{Y_f}{2} - \frac{rY_f}{4I^{1/2}(\tilde{y}^*\,;\, r,\, \tilde{y}^*)} \int_o^{\tilde{y}^*} \frac{I^{1/2}(y;\, r, \tilde{y}^*) - dy}{\cosh \frac{r}{2}(\tilde{y}^* - y) + 1} \qquad (24)$$

where

$$I(\tilde{y}\,;\, r,\, \tilde{y}^*) = \int_o^{\tilde{y}} \left[1 - \frac{2}{e^{r(\tilde{y}^* - y)/2} + 1}\right] y^2\, e^{-y}\, dy \qquad (25)$$

and

$$\tilde{y}^* = \frac{2}{r} \ln\left(\frac{X_f}{X_b} - 1\right) \qquad (26)$$

(there is also a local structure for the radicals R_1 and R_2 at $\xi = \xi^*$, namely that which connects the trivial solution (16) with the local equilibrated solution (17). Details can be found in [8]).

IV. CONCLUSION

For the special case considered, the relationship between Y_f and X_b is given by (24) and a similar result will be obtained, albeit numerically, in general.

Figure 1 gives a plot of Y_f / X_f vs. X_b / X_f for the parameter value $r = \frac{1}{2}$. We believe it is typical; it shows that

$$X_b < \frac{1}{2} X_f$$

corresponds to

$$Y_f > X_f \equiv Y_{fl}\,.$$

Since this is the condition under which a solution for the flame obtains, we call Y_{fl} the lean flammability limit. A mixture will not burn if its fuel strength is below Y_{fl}.

REFERENCES

1. Buckmaster, J.D., and Ludford, G.S.S., *Theory of Laminar Flames*, Cambridge University Press, 1982.

2. Liñan, A., Instituto Nacional de Tecnica Aeroespacial "Esteban Terradas" (Madrid), USAFOSR Contract No. EOOAR-0031, Technical Report No. 1 (1971).

3. Fife, P.C., and Nicolaenko, B., *Physica D* (1985).

4. Seshadri, K., and Peters, N., *Combust. Sci. Technol.* 33:35 (1983).

5. Tam, R., and Ludford, G.S.S., *Combust. Sci. Technol.* 43:227 (1985).

6. Peters, N., and Smooke, M.D., *Combust. Flame* 60:171 (1985).

7. Tam, R.Y., and Ludford, G.S.S., to appear in *Combust. Flame* (1987a).

8. Tam, R.Y., and Ludford, G.S.S., to appear in *Combust. Flame* (1987b).

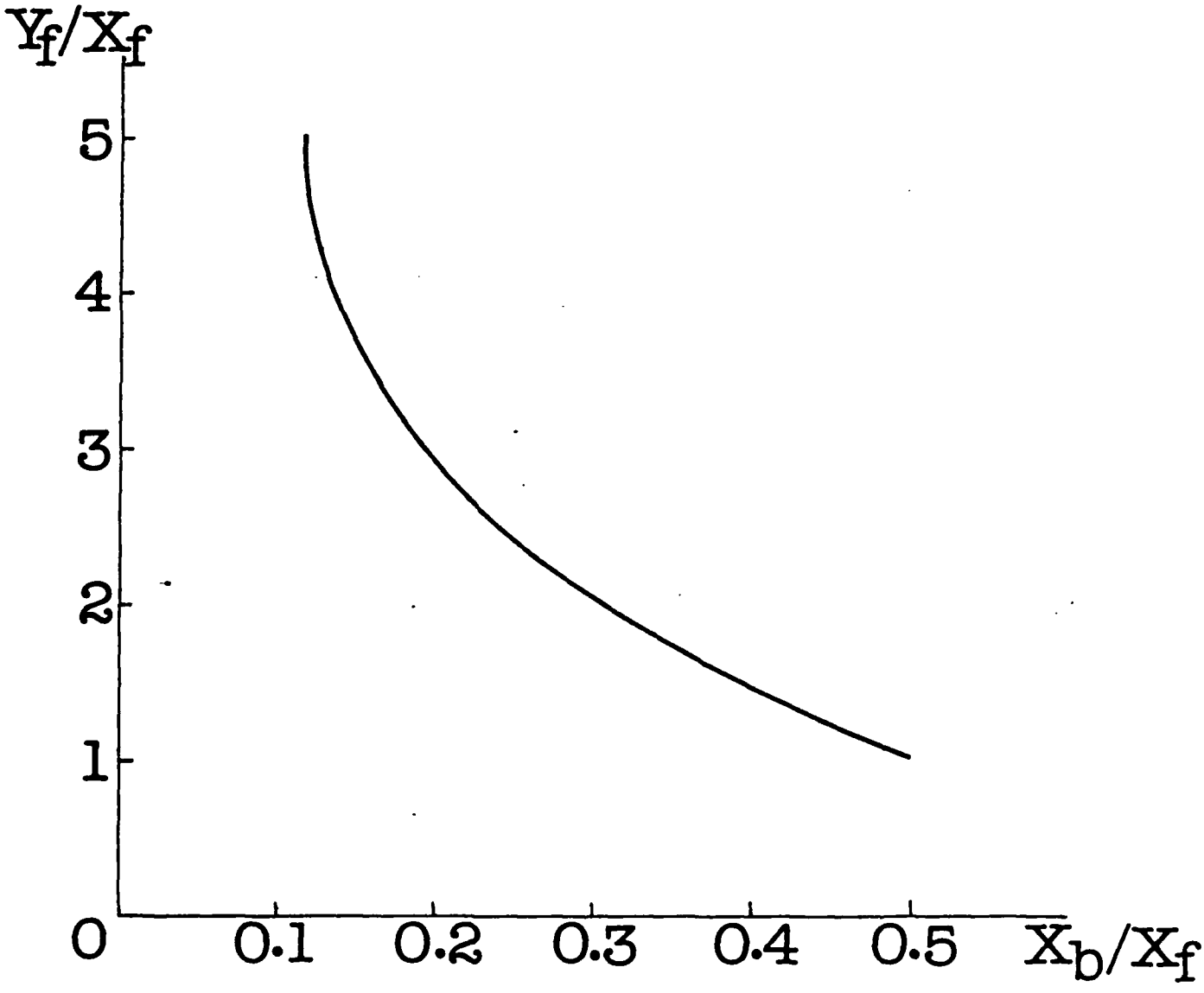


Figure 1. Typical plot of Y_f/X_f versus X_b/X_f .

THE SOLUTION OF THE TYPE-PROBLEM FOR N $\leq$ 4

Nam P. Bhatia
University of Maryland Baltimore County
Catonsville, MD 21228
and
Walter O. Egerland
Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

Introduction

At the second and fourth Army Conference on Applied Mathematics and Computing we introduced the notions of loop and elementary orbit. The implicational strength of loops and elementary orbits enabled us to prove two refinements of Sarkovskii's Theorem, Theorem (SR) and Theorem (SR II) [2,3]. In fact, from Theorem (SR II) follows, as a corollary, a new proof of Sarkovskii's Theorem in a most natural way. If we call each of the $(n-1)!$ different $n-$periodic orbits a continuous function $f : R \to R$ can have a period type, one is naturally led to what we have called the type-problem.

Statement of the Type-Problem. Let $f : R \to R$ be continuous. Given a positive integer n and an $n-$periodic orbit of specified type, find for every positive integer m the types of $m-$periodic orbits that must exist.

At this stage of knowledge the type-problem is an open problem of considerable complexity. Even the restricted (or "little") type-problem appears to be of great difficulty.

Statement of the Restricted Type-Problem. Let $f : R \to R$ be continuous. Given a positive integer n and an $n-$periodic orbit of specified type, find for every positive integer $m \leq n$ the types of $m-$periodic orbits that must exist.

It is well to point out that Sarkovskii's result gives only the (complete) answer to the "typeless" problem: "Given a positive integer n and an $n-$periodic orbit of any type. For which other integers m does there exist an $m-$periodic orbit of any type?"

In the first part of this presentation the solution for the restricted type-problem for $N \leq 4$ is given. It is necessary for this purpose to introduce the notion of a separated loop, a direct generalization of a loop. We shall see that separated loops do not obey a linear order. This is in contrast to loops and elementary orbits that are not only linearly ordered individually, but also when taken together. Accordingly, the various

period types that appear in the solution for the restricted type-problem for $N \leq 4$ are not linearly ordered.

In the second part of this presentation our notions make contact with the notion of turbulence as introduced by Block and Coppel [4]. We show that the notions of turbulence and infinite loop are equivalent, i.e., we prove that "$f : R \to R$ is turbulent if and only if f has an infinite loop."

Unless new, our notation is unaltered from [1], [2], and [3].

This presentation represents only part of our joint work under the U.S. Army Summer Faculty Research and Engineering Program.

1. The Partial Ordering of the Separated Loops.

Definition: A $p-$periodic orbit $(p \geq 2)$ is called a $(m,n)-$separated loop if $p = m+n, \quad m,n \geq 1$, and the points of the orbit satisfy

$$x_{m+n} = x_0 < x_1 < \ldots < x_{m-1} < x_{m+n-1} < \ldots < x_{m+1} < x_m \, .$$

We adopt the notation that $L_{m,n}$ shall mean that $f : R \to R$ has a $(m,n)-$separated loop.

We have the

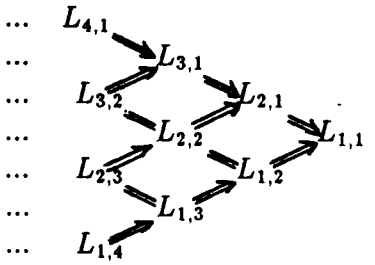
Theorem: Let $f : R \to R$ have a $(m,n)-$separated loop. Then there exist two $(m-1,n)-$separated loops and two $(m,n-1)-$separated loops, except that a $(2,1)$ or a $(1,2)-$separated loop only implies one $(1,1)-$separated loop. In particular,

$$L_{m-1,n} \Longleftarrow L_{m,n} \Longrightarrow L_{m,n-1}.$$

The proof is a direct application of the following well-known lemma.

Lemma. Let $J_1, J_2, \ldots, J_n$ be compact intervals such that $f(J_i) \supset J_{i+1}, i = 1,2,\ldots,n-1$ and $f(J_n) \supset J_1$. Then there is a point $x_0 \in J_1$ with $x_i \in J_{i+1}, i = 1,2,\ldots,n-1$, and $x_0 = x_n$. The point x_0 has period n or n', where n' divides n.

The following diagram displays the partial ordering on all separated loops.



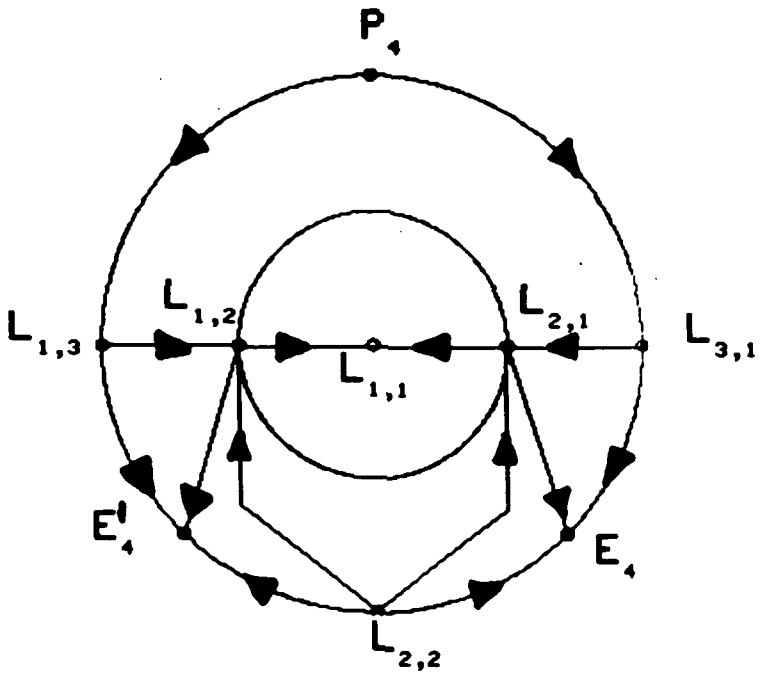
That in general no other implications hold can be demonstrated by examples.

2. The Solution of the Restricted Type-Problem for N $\leq$ 4.

The six types of 4−periodic orbits are given by

$$\begin{array}{llll} P_4: & x_4 = x_0 < x_3 < x_1 < x_2 & L_{2,2}: & x_4 = x_0 < x_1 < x_3 < x_2 \\ L_{3,1}: & x_4 = x_0 < x_1 < x_2 < x_3 & L_{1,3}: & x_4 = x_0 < x_3 < x_2 < x_1 \\ E_4: & x_4 = x_0 < x_2 < x_1 < x_3 & E_4': & x_4 = x_0 < x_2 < x_3 < x_1 \end{array}$$

We present the solution for the restricted type-problm for $N \leq 4$ diagrammatically:



The proofs of the indicated implications follow from the Theorem and Lemma in Section 1.

3. $\mathbf{T} \Longleftrightarrow \mathbf{L}(\infty)$.

Block and Coppel introduced in [4] the concept of turbulence. By definition a continuous function $f : R \to R$ is turbulent if there exist compact intervals J and K such that $J \bigcap K$ is at most a singleton and

$$J \bigcup K \subseteq f(J) \bigcap f(K).$$

To show the equivalence: "$f : R \to R$ is turbulent if and only if f has an infinite loop", we recall a convenient notation and two lemmas.

For intervals J and K we write $J \leq K$ if $x \leq y$ whenever $x \in J$ and $y \in K$. It follows that $J \leq K$ or $K \leq J$ if and only if $J \bigcap K$ is at most a singleton.

If we call the interval J' minimal with respect to the property $f(J') = K$ if no proper subset of J' has this property, the first well-known lemma reads as follows.

Lemma 3.1. If $f : R \to R$ is continuous and J and K are intervals such that K is compact and $f(J) \supset K$, then there is a minimal compact interval $J' \subset J$ such that $f(J') = K$.

We finally recall the following lemma [3].

Lemma 3.2. If f has a critical point c_0 such that $c_0 < c_{-2} < c_{-1}$, then f has an infinite loop satisfying

$$c_0 < \ldots < c_{-n} < \ldots < c_{-2} < c_{-1}.$$

The same statement holds with all inequalities reversed.

If we say that property T holds if f is turbulent, we have the following

Theorem. $T \Longleftrightarrow L(\infty)$.

Proof. Let f be turbulent. We assume without loss of generality $J \leq K$. Then there are minimal compact intervals $J_1 \subset J$ and $J_2 \subset J$ such that $f(J_1) = K$ and $f(J_2) = J$. Since $f(J_1 \bigcap J_2) \subset f(J_1) \bigcap f(J_2) = K \bigcap J$ and $K \bigcap J$ is at most a singleton, we conclude from the minimality of J_1 and J_2 that $J_1 \bigcap J_2$ is at most a singleton and hence we have either $J_1 \leq J_2 \leq K$ or $J_2 \leq J_1 \leq K$. In case $J_1 \leq J_2 \leq K$, we conclude the existence of a critical point $c_0 \in K$ and predecessors $c_{-1} \in J_1$ and $c_{-2} \in J_2$ from the respective conditions $f(K) \supset K$, $f(J_1) = K$, and $f(J_2) = J \supset J_1$. From $J_1 \leq J_2 \leq K$ follows $c_{-1} \leq c_{-2} \leq c_0$. We note now that no equality in the last statement can hold for that would force J_2 to be a singleton which is impossible. Hence f has an infinite loop by Lemma 3.2. The case $J_2 \leq J_1 \leq K$ is similar.

Conversely, if f has an infinite loop, then there is, in particular, a critical point c_0 and predecessors c_{-1} and c_{-2} satisfying $c_0 < c_{-2} < c_{-1}$ (or $c_0 > c_{-2} > c_{-1}$). We let $J = [c_0, c_{-2}]$ and $K = [c_{-2}, c_{-1}]$ and verify that $J \leq K$ and $f(J) \bigcap f(K) \supset J \bigcup K$ hold, i.e., f is turbulent. This completes the proof of the theorem.

We remark that for applications the simple sufficient condition $x_3 < x_2 < x_0 < x_1$ and Lemma 3.2 are excellent means to establish turbulence.

Acknowledgement

We wish to thank A.B.Cooper, M.A.Hirschberg, and W.H.Mermagen, all of SECAD-BRL, for their ample support of our endeavors.

References

1. Nam P. Bhatia and Walter O. Egerland, Non-periodic Conditions for Chaos and Snap-Back Repellers, Transactions of the Second Army Conference on Applied Mathematics and Computing, ARO Report 85-1, pp. 159-164, 1985.

2. Nam P. Bhatia and Walter O. Egerland, Extensions of Sarkovskii's Theorem, Transactions of the Fourth Army Conference on Applied Mathematics and Computing, ARO Report 87-1, pp. 605-610, 1987.

3. Nam P. Bhatia and Walter O. Egerland, A Refinement of Sarkovskii's Theorem, Accepted for publication in the Proceedings of the AMS.

4. L.S.Block and W.A.Coppel, Stratification of Continuous Maps of an Interval, Transactions of the American Mathematical Society, 279 (2) (1986), 587-604.

NEURAL NETWORKS AND THE SYMMETRIC GROUP S_N

John L. Johnson
Research Directorate
Research, Development, and Engineering Center
Redstone Arsenal, AL 35898-5248

ABSTRACT. A permutation of N objects can be implemented by a neural net which recognizes the initial arrangement and then replaces it with the final arrangement. This is the same process found in parallel computer architectures using symbolic substitution principles, and in turn is the same function performed by optical correlator devices. Since every group is isomorphic to a subgroup of S_N, the symmetric group, these interrelationships provide a powerful mathematical base for optical computers and neural networks. Neural nets for the first few symmetric groups are presented. The number of interconnected nodes is shown to be related to the number of strictly parallel inputs and outputs, and a group categorization for any parallel interconnected network is discussed.

1. INTRODUCTION. There exists a functional correspondence among neural networks, optical correlators, symbolic substitution, the symmetric group S_N, and digital computers. It provides a mechanism for the design, implementation, and use of parallel/serial processor architectures. The basic common feature is that of symbolic substitution: identify a given symbol or pattern and replace it with another symbol or pattern. Symbolic substitution was introduced by Huang[1-3] as a method of implementing digital computers with optical techniques. The functional equivalence of optical correlation and symbolic substitution was recently stated by Casasent.[4] The concept of a digital computer as a finite state machine has long been recognized and provides a direct correspondence with the permutation elements of the symmetric group S_N. The ability of neural networks to recognize and replace spatial patterns, and thus perform symbolic substitution, has been shown by Grossberg.[5] A common thread has been woven through these five ideas, and their unification has in principle been achieved as indicated symbolically in Figure 1.

Their equivalence serves as a reversible recipe for solving processor architecture problems, for designing explicit neural networks which implement symbolic substitution, for specifying the optical hardware, and for programming the processor by duplicating the group structure of the problem with the group structure of the processor.

II. CORRESPONDENCES. The following correspondences are summarized in figure 2.

a. Symbolic Substitution. Symbolic substitution consists of identifying a given pattern or symbol, removing it, and replacing it with another pattern or symbol. The act of replacement implies the existence of gain in a physical system.

b. Optical Correlators. Consider the correlation of two functions, given by .

$$C(x,y) = \int_{-\infty}^{\infty}\int dudv\, A(u,v)\, B^*(u-x, v-y) .$$

When A=B in a optical matched filter correlator, the output is a plane wave which can then be brought to a point focus. The point approximates a delta function centered on the location coordinates of B. If, on the other hand, B is initially a delta function, then the correlation is a replica of the pattern A, again centered on the location of $B^{(6)}$. Suppose we have two correlators, with reference images f and g, respectively, and they are joined so that the output of the first one is the input to the second. The first one identifies all occurrences of f in its input scene and supplies a corresponding delta function map to the second. It, in turn, produces an output scene with its reference image g written at every location where f was present in the original scene. This tandem correlator has then performed symbolic substitution f→g. It recognized the subpattern f and substituted in its place the subpattern g.

c. Neural Networks. Next consider the Grossberg neural model of an instar triggering an outstar[5]. The instar recognizes the previously-encoded training distribution f and delivers a recognition signal to the outstar. It, in turn, plays back a second previously-encoded distribution g on its field of neural nodes. Again, this is the function done by symbolic substitution: recognize f and substitute g. As noted earlier, the actual act of substitution implies gain.

d. The Symmetric Group. The symmetric group S_N is the group of all permutations of a collection of N distinct objects. A permutation element consists of an initial arrangement, or pattern, of the N objects, which is first identified, and is then replaced by a second arrangement of the N objects. A single permutation corresponds to one act of symbolic substitution.

e. Digital Computers. A digital computer, viewed as a finite state machine, goes from its initial state of data words and addresses to its final state by executing its program. The initial and final states are the initial and final arrangements of N objects, N being, in this case, a very large number, and the program corresponds to the particular permutation element which was used.

III. GROUP NETWORKS. In this example, neural networks are devised which model group elements and the group product rule. It leads to a characterization of the group in terms of the number of nodes and input connections required to model it. This characterization is then reversed so that a given network with a specified number of nodes and input connections per node can be identified with the group number N. It is important to note that this reverse charaterization is incomplete; it does not take into account the actual connection matrix. What it says is that for the particular number of nodes and interconnects given, it would be possible to form the N^{th} group, but does not necessarily indicate that such networks are actually present.

Cayley's theorem states that all groups are isomorphic to a subgroup of S_N. Neural networks which form all the permutation elements and provide all possible group products (at least once) can accordingly model processes described by a group. All permutations of a collection of N distinct objects comprise the elements of S_N. There are N! elements. The group operation is two successive permutations.[7]

An ideal totally interconnected Grossberg slab (memoryless) with recurrent, shunting, on-center/off-surround subnets will, in the extreme binary limit, choose the single most active node and drive all the others to zero.[8] Consider such a slab with N nodes. It can have N distinct states, each consisting of one active (and normalized) node and N-1 inactive nodes. If the nodal output channels are permuted to N possible connections, then that slab can perform one permutation.

Figure 3 shows the networks for S_1, S_2, and S_3 group elements. The Grossberg subnets require that each node receive an input from all the nodes including itself, plus an input from each of the N input channels.[9] Some of the inputs excite the node and some inhibit it. Here, both types are simply counted as inputs. Outputs are not counted. To form all the elements (which, as yet, are not connected to each other) of S_N requires N nodes per element and 2 N inputs per node. There are N! elements. The total number of nodes is N·N! and the total number of input connections is $2N^2 \cdot N!$. Suppose further we wanted to be able to perform, at least once, all possible group products. Then each element's N outputs must connect to the inputs of every element including itself. This is shown in Figure 4 for an element of the S_3 group. This requires additional input connections per element consisting of N times the number of elements transmitting to it. This is true for every element. The total number of input connections for group products is $N \cdot (N!)^2$. These group and element counts are shown in Figure 5, where the standard gamma function is shown for comparison. The overall number of input connections per node is $(2N^2N! + N(N!)^2)/(N(N!))$, or $2N + N!$. In loose orders of magnitude, and for large N, the number of inputs is roughly the square of the number of nodes.

The number of distinct objects is N. The above group networks can handle them all in parallel. N can be interpreted as the number of parallel inputs which the group system can handle in a fully interconnected manner, and thus is a measure of the parallel capacity of a memoryless S_N neural network. Informally, it represents how many things the networks can do at the same time. Given the parallel capacity N, then the number of nodes, element input connections, and group product input connections can be listed. On the other hand, given the node and connection count for a network, one can estimate the parallel capacity, or group number, for that network. It is known that in the neural cortex there are about J^2 interconnections for J neurons.[10] This agrees with the group construction. For 10^4 interconnections per neuron in the cortex[10] the parallel capacity would be between N=4 and N=5.

Consider the group networks discussed above. If given basis is chosen for the permutation elements, such as the binary N-node slabs, then one can distinguish between element operations and group operations. The element operations can be done in parallel. They can be called up to perform a serial sequence of group operations. The elements are analogous to subroutines and the sequence of group operations is like an overall software program. This gives a guide from group theory to show which operations should be done in parallel and which should be done in series. If "a symmetric group computer" is made capable of calling up a desired sequence of group operations matching those in the problem to be solved, then these comprise the program sequence and are done serially. The individual group elements are the subroutines and can be done in parallel using the neural net designs of Section 3, and can be implemented in optical correlator hardware according to the functional correspondence reviewed in section 2.b. of this paper.

A basic feature of groups is that of closure. Any product of its elements yields another element of the group. In principle, then, serial operations should be unnecessary because one could simply activate the single parallel element representing the entire product of the other relevant elements. However, for large and complex systems, the problem of finding this element may require taking the serial products of the other elements of the first place. It is for these systems that a "group computer" may offer an advantage, not for those which can be handled analytically.

References

1. A. Huang, "Parallel Algorithms for Optical Digital Computers," PROC. IEEE, 10th International Optical Computing Conference, pp. 13-17, Publication No. 83CH1880-4 (1983).

2. A. Huang, "Architectural considerations involved in the design of an optical digital computer," PROC. IEEE 72 (7), pp. 780-786 (1984).

3. K. H. Brenner and A. Huang, "An optical processor based on symbolic substitution," PROC. Topical Meeting on Optical Computing, OSA, pp. WA4.1-WA4.3 (1985).

4. D. Casasent and E. Botha, "Knowledge in optical symbolic pattern recognition processors," Opt. Eng. 26 (1), pp. 034-040 (1987).

5. S. Grossberg, Studies of Mind and Brain, Boston Studies in the Philosophy of Science, Vol. 70, D. Reidel Publishing Company (1982).

6. J. Goodman, Introduction to Fourier Optics, McGraw-Hill Book Company (1968).

7. F. Byron and R. Fuller, Mathematics of Classical and Quantum Physics, Vol. 2, Chapter 10, Addison-Wesley Publishing Company, Inc. (1970).

8. G. Carpenter and S. Grossberg, "A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine," Computer Vision, Graphics, and Image Proc. 37, 54-115 (1987).

9. Reference 5, Chapter 1, Appendix D.

10. H. Szu, "Globally Connected Network Models for Computing Using Fine-Grained Processing Elements," PROC. Int. Conf. on Lasers '85, pp. 91-97 (1985).

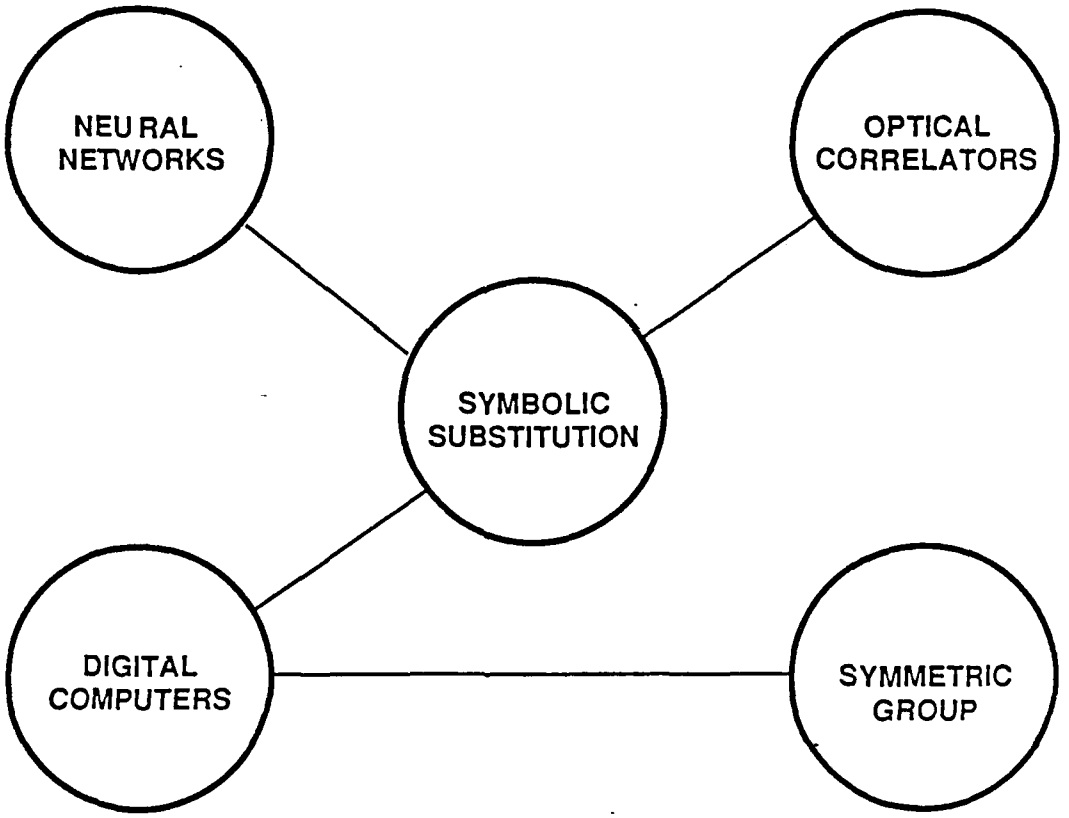


Figure 1. Interrelationships among concepts involving symbolic substitution.

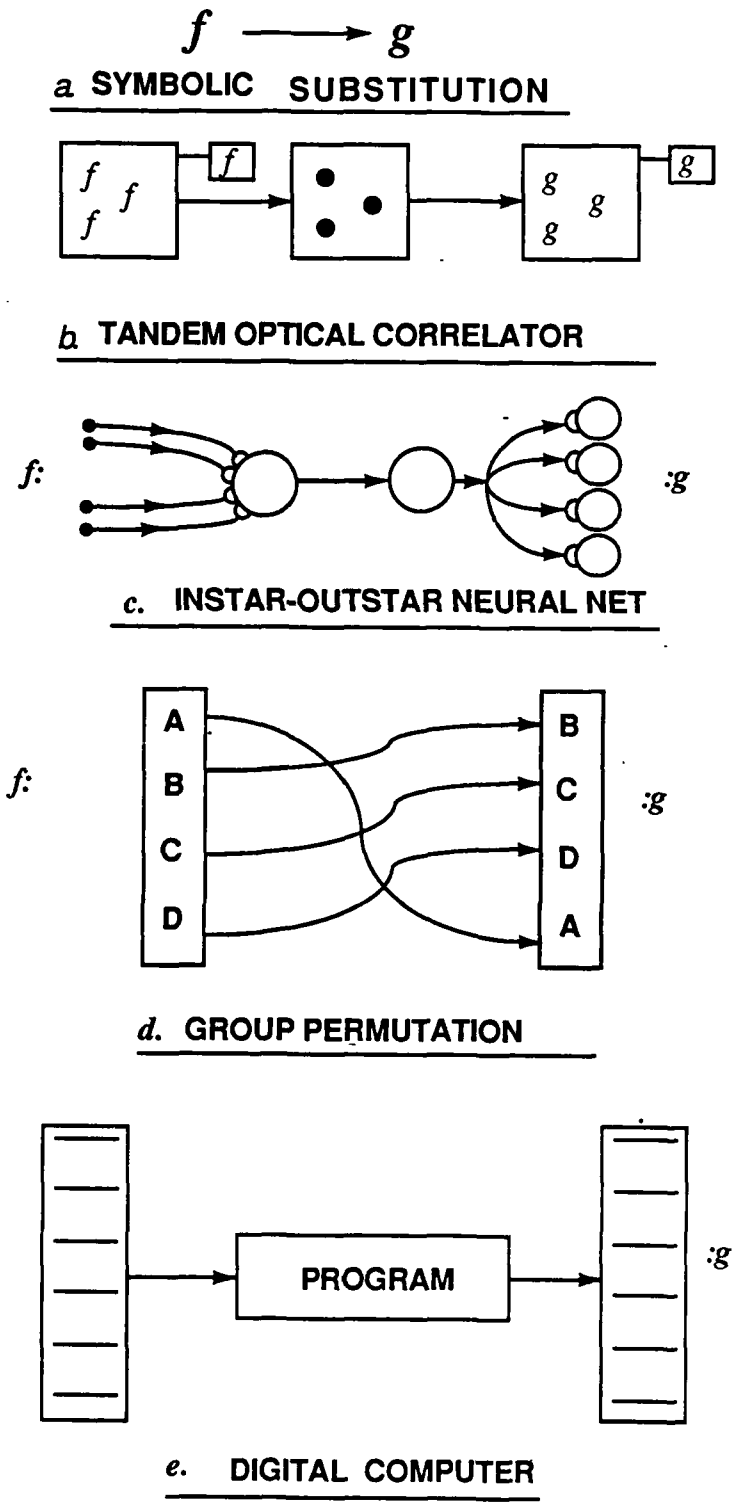


Figure 2. The common functional property f→g is found in optics, neural nets, group theory, and computers.

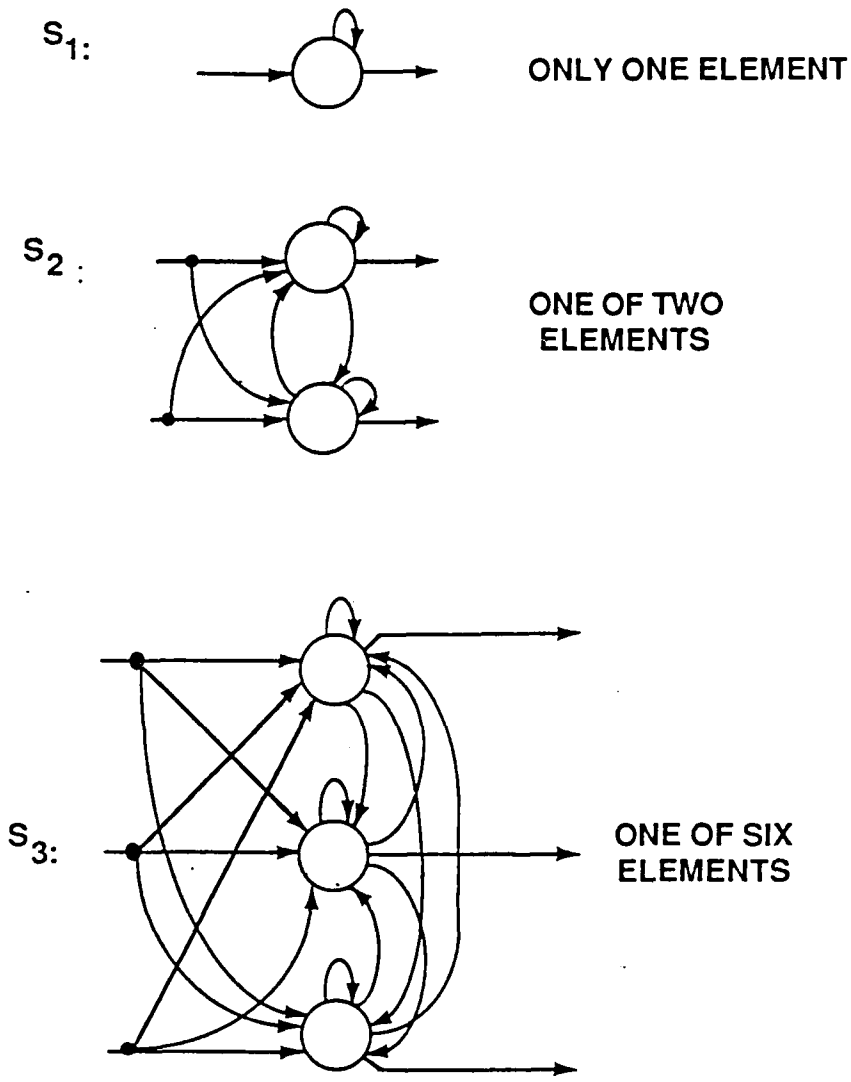


Figure 3. Grossberg subnetwork constructions of group elements for the first three symmetric groups.

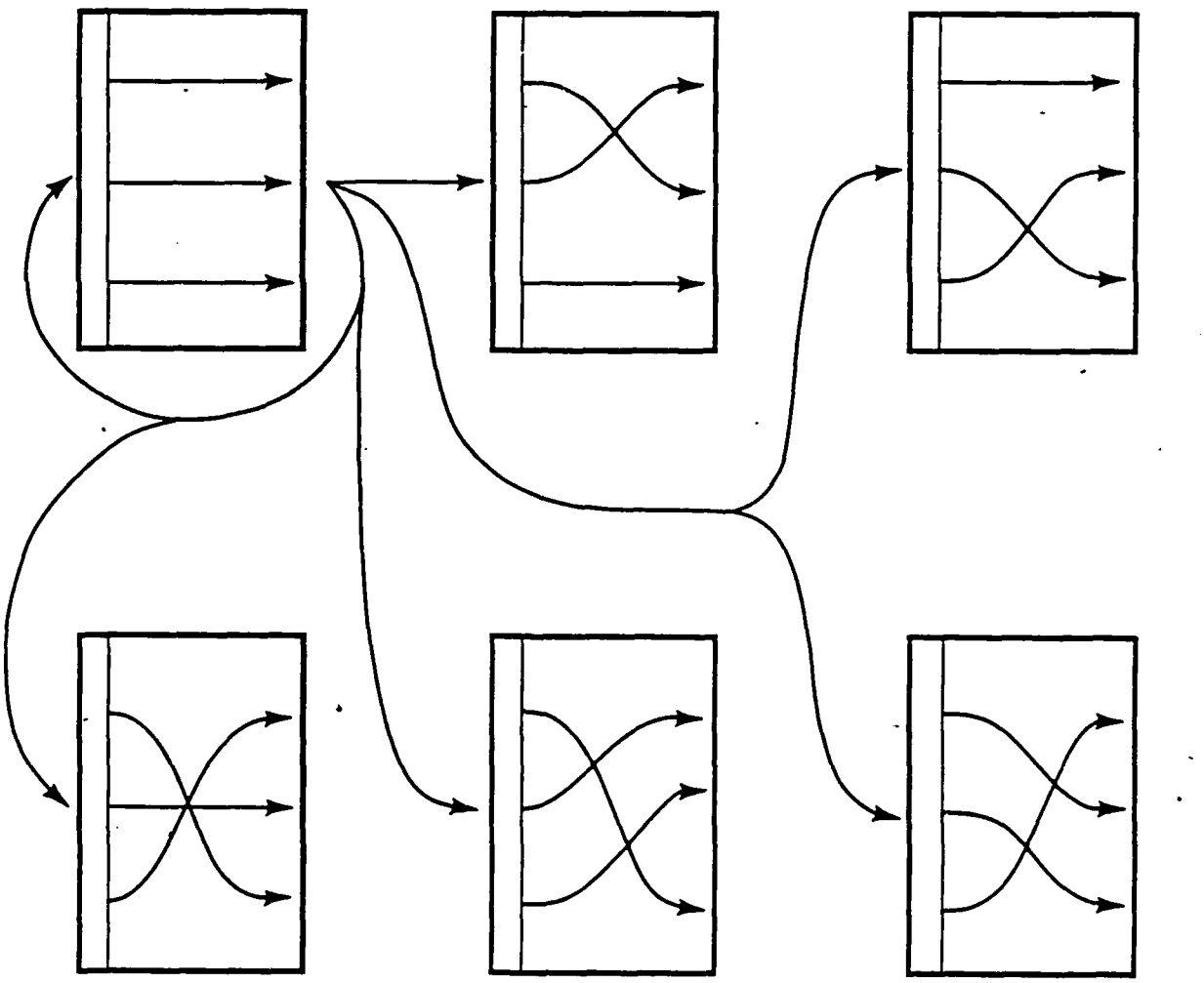


Figure 4. Additional connections among group elements, shown for S_3, which permit application of group product rule between any pair.

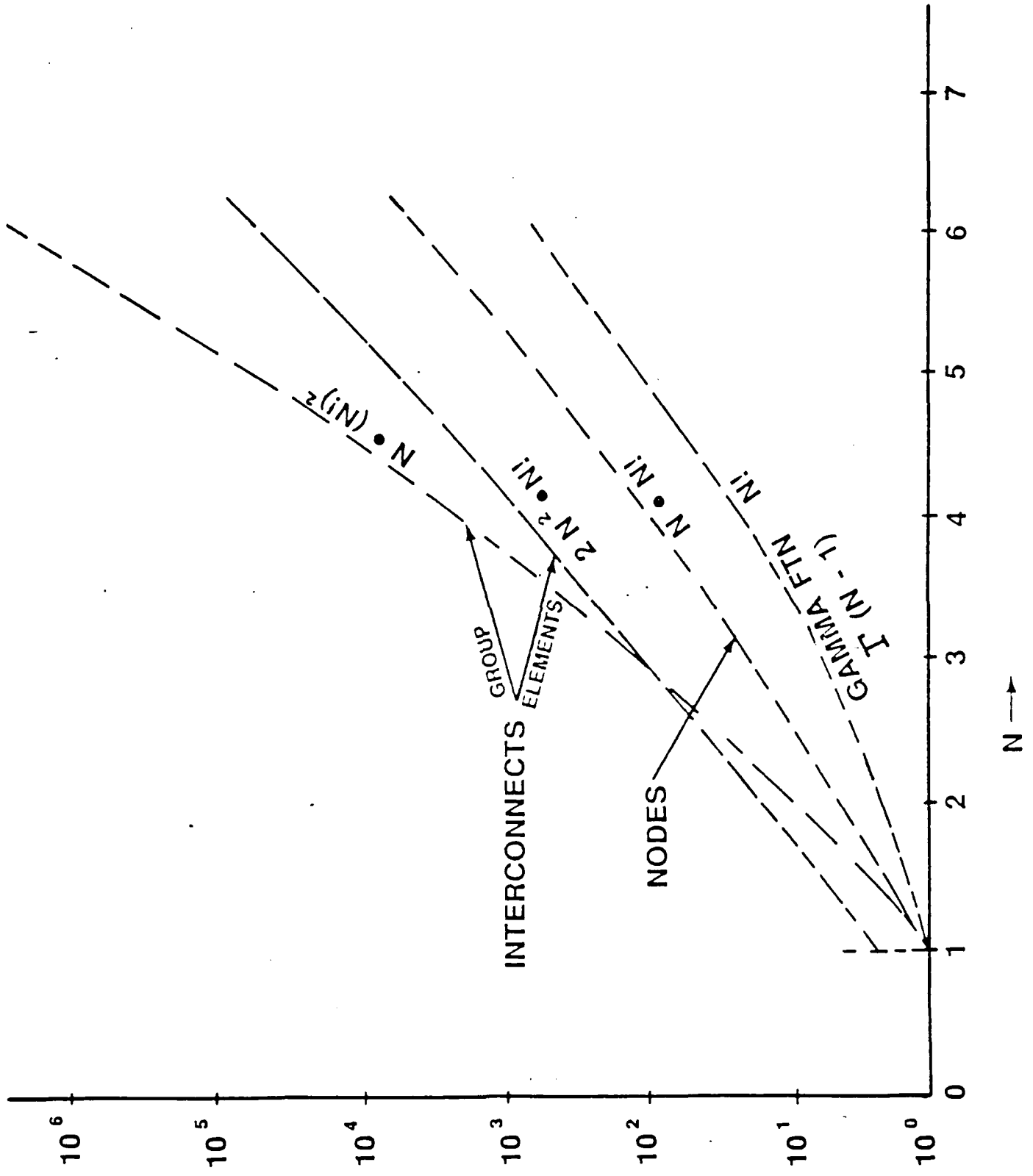


Figure 5. Neural net complexity for symmetric group S_N.

THE INTERACTION OF NONLINEAR HYPERBOLIC WAVES: A CONFERENCE REPORT

James Glimm [1,2]

Courant Institute, New York University
New York, N. Y. 10012

ABSTRACT

Nonlinearities in wave equations lead to focusing and defocusing of solutions. Focusing causes sharply defined wavefronts. The interaction of such sharply defined wavefronts and more generally of nonlinear hyperbolic waves is of fundamental importance and includes such phenomena as Mach triple point formation, shock wave diffraction patterns and the study of Riemann problems in one and higher dimensions.

Recent progress in the study of nonlinear hyperbolic wave interactions has revealed a surprising range of new mathematical phenomena and structures. This mathematical theory should be useful in the design of improved computational algorithms and in part was motivated by such considerations. It is also of considerable interest for its own sake as new mathematical phenomena and is also of interest in terms of the direct insight it provides into physical phenomena.

1. 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
2. Supported in part by the Army Research Office, grant DAAG29-85-K-0188

1. Introduction

1.1. The Problem Formulation.

We consider the nonlinear system of conservation laws

$$U_t + F(U)_x = 0 \tag{1.1}$$

and let

$$\lambda_1 = \lambda_1(U) \leq \cdots \leq \lambda_n(U) \tag{1.2}$$

be the eigenvalues of the Jacobean matrix

$$A(U) = \frac{\partial F(U)}{\partial U}\,, \tag{1.3}$$

while

$$e_1(U),\ \cdots\ , e_n(U) \tag{1.4}$$

are the corresponding right eigenvectors. Equation (1.1) expresses conservation of the components of U :

$$\frac{d}{dt}\int_{-\infty}^{\infty} U(x, t)\, dx = 0\,.$$

The eigenvectors $e_i(U)$ are the normal modes for the propagation of small amplitude signals, linearized about the state U, while $\lambda_j(U)$ are the corresponding wave speeds. The λ_i are assumed to be real, which is equivalent to the system (1.1) being *hyperbolic*. As we will see below, this assumption is not satisfied in all cases. Similarly, we say that (1.1) is *strictly hyperbolic* if the λ_i are real and distinct. Points U_0 of coinciding wave speeds

$$\lambda_i(U_0) = \lambda_{i+1}(U_0)$$

are called *umbilic* points. A remarkable theory has resulted from the careful analysis of these umbilic points.

1.2. Examples.

Conservation laws are basic to physics and the equations (1.1) are often the fundamental or lowest order description of a physical situation. More refined descriptions may arise as modifications to or perturbations of (1.1). For example the right hand side may be replaced by a diffusion term to represent transport effects such as viscosity or heat conduction. It may be replaced by a source term of a geometrical nature, to represent flow through a duct of variable cross section or through a coordinate system (such as radial flow in polar coordinates) in which volumes and densities are not conserved. There may be source terms of a chemical nature to represent stored (chemical) energy, not included in the state U. Some refined descriptions will preserve the same form as (1.1) but will add new variables and equations, for example to represent additional species in a chemical reaction or additional energy partition modes for nonequilibrium thermodynamics.

The specific examples and theories contained within this framework are almost unlimited. The Euler equations for a compressible fluid (gas) are in some sense the prototype example. Here the conservation laws express conservation of mass, momentum and energy while F defines the corresponding fluxes. The fluid can describe several species (multi fluid equations) which can react chemically (chemically reacting equations) or mix (multiphase flow). Continuum equations of elastic and elastic-plastic flow are defined by conservation laws. Magneto-hydrodynamics is a conservation law. The equations for saturation and concentration of fluids in an oil reservoir are of conservation type, as are the equations for adsorption.

1.3. Theory and Computation in Mathematics

Experimental mathematics refers to a working method in which computer experiments play an essential role in the discovery of ideas and the formulation of conjectures. While laboratory experiments, often filtered through the mind of a theoretical physicist, have inspired mathematical thinking for centuries, the direct use of experimental studies by mathematicians from computer simulations is a recent development. There is no doubt that experimental mathematics has played a large role in the

recent progress in our understanding of the interactions of nonlinear hyperbolic waves. A numerical Riemann solver of a very general nature for 2×2 systems was an essential tool in the development of insights and conjectures to guide the mathematical theory [16]. The key tools of an analytic nature have been bifurcation theory, global analysis and geometry.

Equally important has been the connection of nonlinear wave interactions to applications. In fact the wave interaction phenomena has a number of complex aspects. Considered for their own sake, such problems are easily put aside for theories which are more elegant even if less profound. However the firm anchor of these wave interactions to such applications as oil reservoirs, elasticity and chemically reactive flows have allowed the focusing of sufficient talent and energy for significant progress to be made.

1.4. Scale Invariance and Riemann Problems.

The conservation law (1.1) is invariant under the scale transformations

$$x, t \to sx, st, \qquad s > 0 \tag{1.5}$$

in the sense that

$$V(x, t) = U(sx, st) \tag{1.6}$$

is a solution of (1.1) if and only if U is a solution. The restriction to positive $s > 0$ is required to preserve an entropy condition, imposed in addition to (1.1), for weak solutions.

It is natural to look at scale invariant data for (1.1) and the corresponding scale invariant solutions. These are called *Riemann problems* and *Riemann solutions* respectively. The Riemann solution defines the large time asymptotics of a general solution. In this sense and using the language of quantum mechanics, the Riemann solution is the outgoing (W^-) wave operator of a scattering problem [12]. In fact taking $s \to \infty$ in (1.6) gives scale invariant data

$$V_0(x) = V(x, t = 0) = U(\operatorname{sgn} x \cdot \infty, t = 0)$$

and formally the solution $V(x, t)$ is the infinite scaling limit of $U(x, t)$ and thus defines its large time asymptotic behavior. The mathematical proof of these statements as well as the analysis of further terms in the large time behavior has been given by T.-P. Liu [23,24,26].

The limit $s \to 0$ in (1.6) also defines a Riemann problem and its solution. This limit is the instantaneous response to jump discontinuities at the origin in the data $U_0(x) = U(x, t = 0)$. This second interpretation of the Riemann problem allows the following picture of a general solution. It will consist of a number of jump discontinuities (fronts) separated by smooth regions and possibly smaller jumps which cross and interact with one another at isolated points. At an isolated interaction point, the solution behavior is governed by a Riemann solution. In this sense the study of Riemann problems is equivalent to the study of the interaction of nonlinear localized waves.

We define an *elementary wave* to be a scale invariant (Riemann) solution of (1.1) which also moves as a traveling wave:

$$U(x, t) = U(x - ct)$$

for some c. In one space dimension, the elementary waves are the localized waves, i.e. shocks, contact discontinuities, etc., while in two space dimensions they are the intersection points of jump discontinuity surfaces, i.e. Mach triple points, etc.

The elementary waves are the basic building blocks for the solution of Riemann problems; this is true in higher space dimensions as well as in one dimension. Thus elementary waves are of fundamental importance in understanding the solutions of (1.1).

Because of the reduction in the number of independent variables, elementary waves in d dimensions are more or less comparable in difficulty to general solutions in $d - 2$ dimensions and Riemann solutions correspond in approximate difficulty to general solutions in $d - 1$ dimensions.

In particular the $d = 1$ Riemann problem has a lot in common with the theory of ordinary differential equations in the large. Methods such as global analysis,

bifurcation theory and geometry are useful for both classes of problems.

See [11, 12] for a more extensive discussion of the ideas presented in this subsection.

2. Nonlinear Resonance

2.1. Introduction.

The umbilic points U_0 with $\lambda_i(U_0) = \lambda_{i+1}(U_0)$ allow a degree of interaction or nonlinear resonance between distinct modes $e_i(U_0)$ and $e_{i+1}(U_0)$ which is missing in the strictly hyperbolic case. This fact produces a novel and rich range of mathematical phenomena.

Nonlinear theories can be divided into those which are qualitatively linear and those which are essentially or globally nonlinear. The central feature of the phenomena associated with nonlinear resonance and umbilic points is a striking departure from the linear guideposts which have dominated our previous understanding of wave interactions.

Let

$$S \subseteq R^n$$

be the state space in which the solution $U \in S$ takes its values. The waves introduce a type of coordinate geometry in S. An umbilic point is a singularity in this geometry. To be precise, each eigenvector $e_i(U)$ defines a vector field on S. The integral curves, i.e. the solutions of the state space differential equation

$$\frac{dU}{d\xi} = e_i(U) , \qquad (2.1)$$

represent both coordinate lines in S and in their x, t realization, rarefaction waves which contribute to the solution of the one dimensional Riemann problem. The x, t realization comes from setting $\xi = \frac{x}{t} = \lambda_i(U)$ and observing that a solution $U = U(\xi) = U(\frac{x}{t})$ of (2.1) also solves (1.1). These rarefaction waves are called

tered rarefaction waves because they are constant on the characteristic lines $\xi = \frac{x}{t}$ through the origin.

Shock waves are defined by interpreting (1.1) in a weak sense, for example as distributions or measures. Then jump relations are implied between the conserved quantities U_i and the fluxes F_i. We let $[a] = a_- = a_-$ if a is a quantity with a jump discontinuity across a curve in space time and $a_\pm$ represents the values of a on the right (left) of the curve. Then (1.1) is equivalent to

$$s[U] - [F] = 0 , \tag{2.2}$$

as far as jump discontinuities are concerned. In particular, two constant states separated by a jump which satisfies (2.2) define a solution of (1.1).

The solutions to (2.2) lie in families and define a geometry on S, in a manner similar to the rarefaction waves. The resulting curves in S are called *shock curves* or *Hugoniot curves*. They also become singular at an umbilic point.

2.2. The Standard Theory.

The standard theory for a single scalar equation is due to Oleinik [28,29]. It is a theory in the large, but because of the restriction to a single mode, the geometric singularities caused by resonances between distinct modes do not occur. In the Oleinik theory, the solution to the Riemann problem is a *composite wave* formed from a rarefaction wave with embedded jumps (shocks) within it. Considered geometrically from the point of view of state space, it is formed by taking the upper or lower convex envelop of the graph of the flux function. The chords in the envelop correspond to jumps (shocks) in the x, t space solution while the rest of the envelop, which lies on the graph of the flux function itself, corresponds to rarefaction waves in the x, t space picture.

The standard theory for systems is due to Lax [19,20]. It solves the Riemann problem in the small ($U_L \approx U_R$) excluding umbilic points and assuming a convexity condition within each mode

$$<\nabla \cdot \lambda_i(U), e_i(U)> \neq 0 . \tag{2.3}$$

also known as genuine nonlinearity. Under these hypothesis, there are no singularities in the geometry defined by the rarefaction and shock curves. Starting at a given state U_R, there is a unique half of the shock curve which is stable under forward time evolution and a unique half of the rarefaction curve which is realizable in x, t space (because wave speeds must decrease when moving to the left from U_R to U_L in x, t space). These two half curves join smoothly and define the *wave curve* through U_R. There is one such curve for each mode. The solution of the Riemann problem is accomplished by moving a required distance and direction on the n , $n-1$ . . . , 1 wave curves, along a unique path which starts at U_R and ends at U_L. Each segment of this path lies along a wave curve and corresponds to a shock of rarefaction wave in x, t space. In each of the sectors between these waves the Riemann solution is constant.

The Lax theory [19] also allows linearly degenerate families, for which

$$<\nabla \cdot \lambda_i(U), e_i(U)> = 0 . \tag{2.4}$$

For these families, shock and rarefaction waves coincide.

It is clear from these two standard theories that a global theory of the Riemann problem would be built from wave curves which are in general Oleinik composite waves. There was a general (and incorrect) impression that little of interest would occur beyond this.

2.3. The Isolated Umbilic Point.

The theory of an isolated umbilic point is due to Eli Isaacson, D. Marchesin, D. Schaeffer, M. Shearer, P. Paes-Leme and B. Plohr, with recent contributions by H. Holden and C. F. Palmeira. Near an isolated umbilic point U_0 with $\lambda_i (U_0) = \lambda_{i-1} (U_0)$ one can scale and blow up the singularity. This is equivalent to replacing the flux function $F(U)$ by its lowest order nontrivial terms. We assume $n = 2$ and by a Galelean transformation, $\lambda_1 = \lambda_2 = 0$, so the blow up yields generically an $F(U)$ which is a homogeneous quadratic polynomial. There are some inessential scaling parameters in the homogeneous quadratic F and the selection of a

unique F from each equivalence class is the problem of normal forms. It was solved by Isaacson, Plohr and Temple and in a subsequent and more satisfactory form by Schaeffer and Shearer [31]. The classification of the geometry of rarefaction curves and a number of preliminary tools for the analysis of quadratic flux Riemann problems is also presented in [31]. The Riemann problems and normal forms divide into four cases (I, II, III, IV, roughly in order of decreasing difficulty) and in each case there is a symmetric subcase in which one parameter of the normal form is fixed at zero and the resulting Riemann solution is simplified by an extra Z_2 symmetry.

The first Riemann problem of this class was solved by Shearer, Schaeffer, Marchesin and Paes-Leme [34]. It was the symmetric case I. In rapid succession, other cases were solved: the symmetric cases II, III and IV by Eli Isaacson, Marchesin, Plohr and Temple [14] and by Eli Isaacson and Temple [15], the nonsymmetric (general) cases II by Shearer and Schaeffer and cases III and IV by Eli Isaacson and Marchesin [32].

Only the type I nonsymmetric case remains open. The essential ingredients which allowed the rapid progress were analytic ideas from bifurcation theory and experiments from the numerically based Riemann solver. It seems clear that both the analytic and the numerical tools developed will be of considerable importance for the analysis of other Riemann problems.

2.4. New Mathematical Phenomena.

Shock waves can be clearly recognized as belonging to the i^{th} family if i-family characteristics enter from both sides while the j-family, characteristics, $j \neq k$ each cross the shock, entering from one side and leaving from the other. Such shocks are called stable in the sense of Lax or Lax shocks, for short.

It has been found that non-Lax shocks are required to solve the Riemann problem near an umbilic point. The new shocks have the structure of an i_L Lax shock when viewed from the left side and an $i_R \neq i_L$ Lax shock when viewed from the right side. If $i_R > i_L$, the shock is over compressive and fewer than n waves occur in the

Riemann solution. If $i_R < i_L$ the shock is undercompressive and more than n waves occur in the solution of the Riemann problem. The undercompressive shock is also called a crossing shock because it is a bridge, joining two sheets of the Hugoniot surface. Both under and overcompressive shocks arise.

Both the Hugoniot and the wave curves can have disconnected branchs. This means that there are waves which cannot be continuously deformed to zero strength. This fact appears to have been discovered independently by Shearer and Marchesin before they were aware of one another's work. The Hugoniot curves can have loops, or points of self intersection, as was first discovered by Shearer. The singularity of the geometry of the wave curves at an umbilic point has already been noted. Moreover the wave curves may fail to have a continuation.

One can regard the wave curve as an $n+1$ dimensional surface in $S \times S$. In the standard theory, these surfaces are globally distinct. In the presence of umbilic points and especially of undercompressive (also called crossing) shocks and waves, we regard the i^{th} wave family as a single sheet of a global wave surface, as in the case of Riemann surfaces. Locally this surface has n distinct branches, but the branches may join globally and may be deformed onto one another. Thus the distinction between an i wave and a j wave may be well defined locally at $U \approx U_0$ but this distinction is not globally meaningful in all cases.

Various topologically significant surfaces in S and in $S \times S$ have been determined which help to delineate the locations of possible bifurcations of structure in the Riemann solution. These are the *inflection locus* on which (2.3) fails and the convexity of a single mode is reversed, the *bifurcation locus* on which secondary bifurcations of the Hugoniot curves occur, or in other words at which the Hugoniot curves cross. the *2-sided contact locus* across which embedded shocks in the rarefaction fans enter or leave the solution and the *hysteresis locus* across which Hugoniot curves acquire or lose segments of distinct families or types.

2.5. Elliptic Regions.

A small perturbation of the flux F in a neighborhood of an umbilic point can give rise to an *elliptic* region $\mathcal{E} \subseteq S$ for which λ_1 and λ_2 are complex. There have been three major discoveries in connection with elliptic regions. Holden [13] showed that a Riemann problem with an elliptic region still has a satisfactory mathematical solution. Bell, Trangenstein and Shubin [5] solved such a Riemann problem numerically and also obtained satisfactory results. Finally Shearer [35] showed that elliptic regions are almost certainly required on topological grounds for basic problems in petroleum reservoir modeling.

How is this to be reconciled with the idea that initial value problems must be hyperbolic? Evidentially the linear instability guaranteed by the elliptic region $\mathcal{E}$ is only an infinitesimal instability and the problem is stabilized by nonlinear considerations. When the elliptic region $\mathcal{E}$ is bounded, one could expect the linear or infinitesimal instability to cause a solution taking values in $\mathcal{E}$ to grow, until it was forced to exit from $\mathcal{E}$, at which point it would lie in the hyperbolic region $S \setminus \mathcal{E}$, and be stabilized. In fact this is exactly what does occur according to available evidence; the solution, if forced to lie in $\mathcal{E}$ will exit with a shock and not return unless forced to do so. More precisely, it appears that the wave path taken by the Riemann solution will not enter $\mathcal{E}$ unless $U_R \in \mathcal{E}$ or $U_L \in \mathcal{E}$. However the meaning of this elliptic region should be explored more carefully before this or any other explanation is accepted. There are cases, as with the van der Waals equation of state for a compressible fluid with a phase transition where an elliptic region results from an incorrect physical model.

2.6. Open Problems.

Most problems related to uniqueness are open: entropy, admissibility conditions, and the existence of viscous profiles are not satisfactorily understood. The proper formulation of a physically meaningful entropy is open. For the case of a single mode, $n = 1$, in the Buckley-Leverett equation, a physically meaningful entropy has been proposed [4], and this could be the basis of a physically meaningful entropy for the case of systems. Existence of the Riemann problem in the nonsymmetric case

I is open and a solution for the full range of possible elliptic cases is open. It is likely that new phenomena will occur with $n \times n$ systems, $n \geq 3$, but this case has yet to be explored. The effect of these equations on existence theory for general data is not known. The case of an umbilic line has been studied by Keyfitz and Kranser [18] and Eli Isaacson [17]. In this case the existence theory for general data was solved (for small data) by Temple [38] and it turned out that some new ideas were needed. In fact the total variation bounds, which are central to the existence theory, had to be reformulated in this case as they failed when applied to the conserved quantities U. The ability of various finite difference algorithms to solve Riemann problems with umbilic points or lines is not known.

Finally we ask whether these novel structures in Riemann solutions have any counterpart in experimental science.

3. Riemann Problems for Realistic Equations

3.1. Introduction.

Real problems are often not strictly hyperbolic. They may fail to be genuinely nonlinear and usually must be considered in the large. Thus we can expect to encounter the phenomena described in the previous section. Here we explain why some real problems possess special features which limit the solution complexity and others do not. Let us exclude the linearly degenerate waves, which in many problems do not give rise to complex one dimensional wave interactions. For gas dynamics, elasticity and a number of other cases, the state space S is a product

$$S = C \times M \tag{3.1}$$

of a configuration space and a momentum space. The eigenvalues, when expressed in a Lagrangian rest frame, come in pairs $\pm\lambda_i(U)$ and in the interior of C are never zero. Thus umbilic points arise only from coincidence of eigenvalues among the positive or the negative eigenvalue families. It follows that these Riemann problems have a complexity similar to those of a Riemann problem for general systems of $\frac{n}{2}$

equations.

For systems which describe species or concentrations, there is generally no factorization of S and no grouping of eigenvalues into disjoint families. Such systems, which include oil reservoir flow equations as an example, appear to be as complex as the order, n, of the system allows.

3.2. Gas Dynamics.

The main complications of the wave structure in the Euler equations of compressible fluids are those of $n = 1$, scalar, equations, according to the $\frac{n}{2}$ rule and the factorization (3.1) of the state space. Each acoustic mode may develop, through self interactions, composite wave structures containing rarefaction waves and some number of embedded shocks. The details of the allowed composite waves depend on the equation of state, and a comprehensive analysis of this dependence has been prepared by Menikoff and Plohr [27]. This means that qualitative as well as quantitative properties of the solution depend on fluid in which the waves are propagating. Not only are the wave properties of real fluids of interest, but those of artificial or simulated fluids defined by approximate equations of state are important also. In fact, the approximate equations of state are used in numerical simulations of fluids and any anomalous waves implied by the use of such an equation of state will occur in the numerical simulation and so must still be understood.

There appears to be no limit to the allowed number of constituent waves (embedded shocks) in a single composite wave, on the basis of thermodynamical principles. For many materials, the change in convexity occurs at or near a phase transition, and a simple assumption would be that at most two convexity reversals (one convexity reversed region) would be encountered. In other words, a given Hugoniot curve would cross the inflection locus at most twice. However fluid (P wave) modes arise in solids also and a large number of distinct phase transitions occur in real solids so this simple picture would not be universal.

Thus the simplest manifestion of real fluid properties is the splitting of a shock at a phase transition, with a precursor moving ahead in the fluid phase and a second shock following in the vapor phase. Shock splitting of this type is well known in the engineering and applied physics literature.

The Riemann problem for a relativistic fluid with a real equations of state (and a phase transition) was solved by Plohr and Sharp [30]. The phase transition which motivated this study was the condensation of a quark - gluon plasma into a baryon phase in proposed experiments for a new particle accelerator.

In addition to the equilibrium thermodynamical effects considered here, metastable states and non equilibrium thermodynamics are of interest also. T.-P. Liu [25] showed how a nonequilibrium partition of internal energy (vibrations of a diatomic molecule) leads to modified fluid equations. It would be desirable to supplement the van der Waals [36] type of metastable thermodynamics by a more realistic and modern treatment of metastable thermodynamics. Bethe [6] points out that the equilibrium state behind a strong shock in water may be ice, in which case the metastable water would be the preferred solution.

A comprehensive data base of tabulated equations of state has been prepared by the Los Alamos National Laboratory for a wide range of materials [1]. Using this data base, J. Scheuermann has constructed an efficient Riemann solver for real fluids [33]. Colella and Glaz [8] had previously constructed an efficient approximate Riemann solver for real fluids, using a local gamma law gas approximation to reduce the number of calls to the real fluid equation of state. Scheuermann's work contains, in addition, an extensive use of precomputed quantities. In his approach the rarefaction curves are given by a single table look up and are thus faster than the Hugoniot curves to determine.

Phase transitions introduce discontinuities into an equation of state just as shock waves do for solutions of fluid equations. In both situations interpolations and numerical differentiations are required and in both cases these operations cause problems when applied across discontinuities. The front tracking software (see [7, 10] and related papers) designed for x, t space discontinuities works just as well for state

space discontinuities. It provides support for interpolation over irregular regions which result from changes of independent variables, starting with a rectangle in (say) p, e space [33]. It is also appropriate for an accurate representation of phase transition curves in S.

The existence of solutions for the Riemann problem for a compressible fluid with a real equation of state follows from basic physical principles, which give an asymptotic description of the equation of state at large pressures and ensure a solution to the midstate shooting problem. Uniqueness, however, is not properly understood. There are the questions of entropy conditions for complex waves, relaxation limits, viscous profiles or other and distinct physical principles which may be required.

3.3. Real Materials.

We consider here thermo-elastic-plastic materials [41] or viscoelastic materials with a simple relaxation law. Many common materials, including metals, are described by this theory, but it does not have the thermodynamic universality of the real fluids described in the previous section. In order to focus specifically on the question of complex wave structure, we ignore the thermal mode.

Isothermal elasticity has a six dimensional state space with the structure (3.1). There are three coordinates to describe positions and three to describe momenta. There are three types of waves. The first is a pressure, P-wave or longitudinal wave. The other two are S-wave, shear or transverse waves. One of the S-wave modes describes torque or rotation waves. For an isotopic material (which we now assume), there is no elastic energy associated with these rotational waves. For this reason the rotational waves are linearly degenerate and factor out of the problem. This leaves four modes and applying the $\frac{n}{2}$ rule, elasticity has the level of wave structure complication of a general 2×2 system.

A remarkable analysis of the Goursat (half-space) Riemann problem for a third order hyperelastic material has been carried out by Tang and Ting [37]. Related

studies of nonlinear elastic waves can be traced from the literature cited in [37]. The solution is similar in structure to the solution for an isolated umbilic point described in Section 2.3 above. In particular there is an isolated umbilic point located on the zero shear axis in the Tang and Ting solutions. Real materials fail in tension and it would be of considerable interest to determine whether the Tang and Ting umbilic point occurs within the elastic limit and if so for what range of materials and strains.

For metals, a regime of interest is one for which the response is nearly linear in shear and tension, but becomes fluid like (fully nonlinear) in compression. Thus aside from P-wave or fluid nonlinearities already discussed, the most striking nonlinearities of common elastic materials occur at the failure of the elastic theory.

There are at least three common failure modes for an elastic material. These are plastic flow, fracture and collapse. They apply to ductile, brittle and porous materials respectively. The first two are shear wave failures while the third is a P-wave failure.

The theories of elastic failure are complex, phenomenological and incomplete. We discuss the case of plastic failure, which may be the best understood of the three. At a microscopic level, plastic flow results from a breaking and reforming of molecular bonds. This process produces several effects. Stored elastic (potential) energy is converted into thermal energy, with the result that elastic forces (stress) are reduced and the unstrained reference configuration is permanently altered. Also dislocations are produced, which alter the material properties through work hardening. Moreover the thermal energy produces heating and heat softening of the material.

Making the Prandtl-Reuss approximation, we assume plastic relaxation occurs along the normal to the yield surface in stress space, and we represent the degree of plastic flow by a single scalar variable ψ. The resulting equations are given in [41], and their main feature is a new equation and mode for ψ, the amount of plastic deformation. The purely elastic equations together with a nonlinear coupling to describe plastic relaxation and the transfer of energy from elastic to thermal modes complete the system. Plasticity should be contrasted to viscosity, which transfers kinetic energy to thermal modes.

3.4. Oil Reservoirs.

The equations for the saturation of oil, gas, and water in three phase flow in an oil reservoir (porous medium) are to leading order a 2×2 hyperbolic system. The equations for most enhanced oil recovery processes have the same form but usually have more equations. These equations were the motivation which lead to the study of umbilic points as discussed in Section 2.3. In particular the solutions of these equations exhibit the complex wave phenomena mentioned in Section 2.4 and 2.5.

There is no reason to believe that the current catalog of mathematical phenomena in the solution of the Riemann problem is complete, especially as additional processes and more equations are considered.

4. Two Dimensional Wave Interactions

4.1. Elementary Waves.

The elementary waves in two dimensions can be studied by the same global methods which were used for Riemann problems in one dimension. An elementary wave is a scale invariant solution of the conservation law which is stable in time. It consists of angular sectors about a fixed origin. In each sector the solution takes on constant values or is a simple wave. Just as the one dimensional Riemann problem is a shooting problem to connect U_R to U_L through a sequence of elementary waves, the two dimensional elementary wave is a circular problem, to connect some state from one of the sectors to itself through a sequence of one dimensional waves.

Theorem [10]. Generically, the elementary waves for a gamma law gas are one of the following simple types: cross, overtake, Mach triple point, diffraction and transmission.

The proof of the theorem is based on the following considerations. In the steady frame of the elementary wave, one draws a circle about the origin. Because of the scale invariance, all of the analysis can be reduced to this circle. Points of the circle and the one dimensional waves contributing to this two dimensional elementary

wave are now labeled as incoming or outgoing. The incoming waves are in principle unrestricted by the equation (1.1), but too many incoming waves are "coincidental" and nongeneric. The outgoing waves are subject to an analysis similar to a one dimensional Riemann problem, and only a limited number of such outgoing waves can occur.

The considerations of uniqueness, real fluid behavior, real material behaviour etc. as discussed in Section 3 are important here also and are mainly not resolved.

4.2. Two Dimensional Riemann Problems.

The two dimensional Riemann solution could fail to be piecewise smooth if there are too many solution modes ($n \geq 3$) or too many inflection points in a single solution mode [21]. However there should be only a finite number of waves of size greater than any fixed $\epsilon > 0$. To focus on these waves, we suppose for simplicity that the Riemann solution is piecewise smooth. The Riemann solution is built up from elementary waves.

We introduce reduced coordinates

$$x, y, t - \frac{x}{t}, \frac{y}{t}$$

and in the $\frac{x}{t}, \frac{y}{t}$ plane we introduce polar coordinates

$$\frac{x}{t}, \frac{y}{t} = r \sin \theta, r \cos \theta .$$

At large r, the conservation law expressed in terms of r and θ as independent variables, is hyperbolic, with r as the timelike variable. The data at large r is given from the solution of one dimensional Riemann problems. It can be continued inward to smaller r by the solution of this hyperbolic equation until an elliptic region is encountered. Data for the elliptic region is specified across a sonic line or shock.

Scalar Riemann problems have been solved mathematically in two dimensions [18,21,22,39,40]. An interesting set of conjectures have been formulated concerning the solution of certain Riemann problems for isentropic gas dynamics in two

dimensions [42].

Two dimensional Riemann problems arise when one dimensional waves cross or overtake one another or when they reflect off of or interact with walls or boundaries. Generically an interaction will arise when two waves meet or a single wave meets a boundary; it is such simple and generic problems rather than the fully general Riemann problem which should be studied. Two problems which have been studied extensively on the level of experiment and computation are (a) the shock-wedge problem of reflection of a shock wave by a wedge in a shock tube and (b) the shock diffraction problem of reflection and transmission of a shock wave by a contact surface. Representative references for these problems are (a) [2,9] and (b) [3].

There are a series of topologically distinct patterns for the various reflected, transmitted and incident waves, and in some cases it is not known which pattern is correct. It may turn out that on the level of the Euler equation (1.1), the solution is nonunique and will be uniquely specified only by the inclusion of a length scale in a modified theory.

Similar issues apply to the interior interaction of waves. Moreover a two dimensional Riemann problem can also be generated by the self interactions of a single two dimensional elementary wave. In fact the angles and wave strengths in a stable elementary wave configuration will in general deform continuously during time evolution, and at some space time point the elementary wave configuration in question may cease to exist, either by a loss of stability relative to a more favored configuration or by a failure of the shock polar equations to have a real solution. At this point the wave pattern bifurcates. A two dimensional Riemann problem is defined, whose solution gives the bifurcation to a new pattern formed by several elementary waves moving away from the bifurcation point. Both the bifurcation point and the outgoing pattern of elementary waves it produces are also subject to possible nonuniqueness, and again a length scale may be needed to resolve this nonuniqueness.

Length scales come from a variety of sources. In the case of interaction with a boundary, there will in general be a viscous boundary layer. For the interaction of interior waves, the viscous effects again introduce a shock thickness. This thickness is normally very small in gasses and liquids, but is more significant in metals. Relaxation of nonequilibrium thermodynamics also produces a length scale and shock thickness. For chemically reacting flows, the reaction zone defines a length scale, normally considerably larger than a shock thickness. Heterogeneities in a medium or in a background flow such as small scale turbulence also provide a length scale. Two dimensional instabilities of a planar interface may introduce a complicated pseudo-one dimensional traveling wave with an extended thickness.

5. Conclusions

There has been a recent burst of progress in our understanding of the interactions of nonlinear hyperbolic waves. This theory should be important for the light it sheds on physical processes. It also gives analytic or explicit solutions which can be used to check numerical methods. A very direct use for this theory and one of the motivations for developing it has been to embed the theory into enhanced resolution numerical algorithms, such as higher order Godunov methods and front tracking.

References

1. "An Invitation to Participate in the LASL Equation of State Library," Los Alamos Scientific Report LASL-79-62, Los Alamos National Laboratory.
2. P. Woodward and P. Colella, "The Numerical Simulation of Two-Dimensional Fluid Flow with Strong Shocks," *J. Comp. Phys.*, vol. 54, pp. 115-173, 1984.
3. A. M. Abd-El-Fattah, L. F. Henderson, and A. Lozzi, "Precursor Shock Waves at a Slow-Fast Gas Interface," *J Fluid Mech*, vol. 76, pp. 157-176, 1976.
4. I. Avaartsmark, Norsk Hydro Technical Report.
5. J. B. Bell, J. A. Trangenstein, and G. R. Shubin, "Conservation Laws of Mixed Type Describing Three-Phase Flow in Porous Media," *SIAM J Appl Math*, vol.

46. pp. 1000-1017, 1986.

6. H. A. Bethe, "The Theory of Shock Waves for an Arbitrary Equation of State." *Technical Report*, May 1942.

7. I.-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, "Front Tracking for Gas Dynamics," *J. Comp Phys.*, vol. 62, pp. 83-110, 1986.

8. P. Colella and H. Glaz, "Efficient Solution Algorithms for the Riemann Problem for Real Gases," *J. Comp. Phys.*, vol. 59, pp. 264-289, 1985.

9. R. L. Deschambault and I. I. Glass, "An Update on Nonstationary Oblique Shock-Wave Reflections, Actual Isopicnics and Numerical Experiments," *J. Fluid. Mech.*, vol. 131, pp. 27-57, 1983.

10. J. Glimm, C. Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, "Front Tracking and Two Dimensional Riemann Problems," *Advances in Appl. Math.*, vol. 6, pp. 259-290, 1985.

11. J. Glimm, "Elementary Waves and Riemann Solutions: Their Theory and their Role in Science," in *Proceedings of Seminar on Differential Equations*, pp. 65-74, Hsinchu, Taiwan, June 3-7 1985.

12. J. Glimm and D. H. Sharp, "An S Matrix Theory for Classical Nonlinear Physics," *Foundations of Physics*, vol. 16, pp. 125-141, 1986.

13. H. Holden, "On the Riemann Problem for a Prototype of a Mixed Type Conservation Law," *Comm. Pure Appl. Math.*, vol. 40, pp. 229-264, 1987.

14. Eli Isaacson, D. Marchesin, B. Plohr, and B. Temple, "The Classification of Solutions of Quadratic Riemann Problems I," MRC Report, 1985.

15. Eli Isaacson and B. Temple, "The Classification of Solutions of Quadratic Riemann Problems II, III," MRC Report, 1985, 1986.

16. Eli Isaacson, D. Marchesin, and B. Plohr, *Wave Curves for Nonlinear Conservation Laws*, To Appear.

17. Eli Isaacson, "Global Solution of a Riemann Problem for a Nonstrictly Hyperbolic System of Conservation Laws Arising in Enhanced Oil Recovery," *J.*

Comp. Phys., To Appear.

18. B. Keyfitz and H. Kranser, "A System of Non-strictly Hyperbolic Conservation Laws Arising in Elasticity Theory," *Arch Rat. Mech. Anal.*, vol. 72, pp. 219-241, 1980.

19. P. Lax, "Hyperbolic Systems of Conservation Laws II," *Comm. Pure Appl. Math.*, vol. 10, pp. 537-566, 1957.

20. P. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM, Philadelphia, 1973.

21. B. Lindquist, "The Scalar Riemann Problem in Two Spacial Dimensions: Piecewise Smoothness of Solutions and its Breakdown," *SIAM J Anal.* vol. 17, pp 1178-1197, 1986.

22. B. Lindquist, "Construction of Solutions for Two Dimensional Riemann Problems," *Adv Hyperbolic Partial Diff. Eqs. Computers and Math. with Applications*, vol. 12A, pp. 615-630, 1986.

23. T.-P. Liu, "Decay to N-waves of solutions of general system of nonlinear hyperbolic conservation laws," *Comm. Pure Appl. Math.*, vol. 30, pp. 585-610, 1977.

24. T.-P. Liu, "Asymptotic behavior of solutions of general system of nonlinear hyperbolic conservation laws," *Ind. J. Math.*, vol. 27, pp. 211-253, 1978.

25. T.-P. Liu, "Hyperbolic Conservation Laws with Relaxation," *Comm. Math. Phys.*, vol. 108, pp. 153-175, 1987.

26. T.-P. Liu, "Pointwise convergence to N-waves for hyperbolic conservation laws," *Bull. Inst. Acad. Sinica*, 1987.

27. R. Menikoff and B. Plohr, "Riemann Problem for Fluid Flow of Real Materials," Preprint, To Appear.

28. O. A. Oleinik, "Uniqueness and Stability of the Generalized Solution of the Cauchy Problem for a Quasi-linear Equation," *Usp. Math. Nauk.*, vol. 14, pp. 87-158, 1959.

29. O. A. Oleinik. "Uniqueness and Stability of the Generalized Solution of the Cauchy Problem for a Quasi-linear Equation." *English Translation Amer. Math. Soc. Transl. Ser. 2*, vol. 29. pp. 295-381. 1963.

30. B. Plohr and D. H. Sharp. "Riemann problems and their application to ultra-relativistic heavy ion collisions." in *VIII International Congress on Mathematical Physics*, ed. M. Mebkhout and R. Seneor, pp. 708-713, World Scientific Publishing Co. Ltd., Singapore, 1987.

31. D. G. Schaeffer and M. Shearer, "The Classification of 2 × 2 Systems of Non-Strictly Hyperbolic Conservation Laws, with Application to Oil Recovery. With Appendix Including D. Marchesin and P. Paes-Leme.," *Comm Pure Appl Math.* vol. 40, pp. 141-178, 1987.

32. D. G. Schaeffer and M. Shearer, *Riemann Problems for Nonstrictly Hyperbolic 2 × 2* Systems of Conservation Laws. Appendix Including Eli Isaacson and D. Marchesin. Trans. AMS, To Appear.

33. J. Scheuermann, "Efficient Solution of the Riemann Problem Using a Tabular Equation of State," NYU Ph. D. Thesis, In Preparation.

34. M. Shearer, D. G. Schaeffer, D. Marchesin, and P. Paes-Leme, "Solution of the Riemann Problem for a Prototype 2 × 2 System of Non-Strictly Hyperbolic Conservation Laws," *Arch. Rat. Mech.*, vol. 97, pp. 299-320. 1987.

35. M. Shearer, "Loss of Strict Hyperbolicity in the Buckley-Leverett Equations of Three Phase Flow in a Porous Medium.," in *Proceedings of IMA Workshop in Oil Reservoir Simulation*, ed. M. Wheeler, To Appear.

36. M. Slemrod, "Dynamic Phase Transitions in a Van Der Waals Fluid," *J Diff. Eq.*, vol. 52, pp. 1-23, 1984.

37. Z. Tang and T. C. T. Ting, "Wave Curves for the Riemann Problem of Plane Waves in Simple Isotropic Elastic Solids," University of Illinois at Chicago Preprint.

38. B. Temple, "Global Existence of the Caucny Problem for a Class of 2 × 2 Non-strictly Hyperbolic Conservation Laws," *Adv. Appl. Math.*, vol. 3, pp. 355-375,

1982.

39. Chang Tung and Zheng Yu-Xi, "Two Dimensional Riemann Problem for a Single Conservation Law," Report No. 20, Institute of Mathematics, Academia Sinica, Beijing.

40. D. Wagner, "The Riemann Problem in Two Space Dimensions for a Single Conservation Law," *Math. Ann.*, vol. 14, pp. 534-559, 1983.

41. Duane C. Wallace, "Thermoelastic-Plastic Flow in Solids," Technical Report LA-10119, Los Alamos.

42. Tong Zhang, "Riemann Problem and Interaction of Waves in Gasdynamics," Lecture at May 1987 Berkeley Conference.

STABILIZATION OF ZIEGLER'S PENDULUM BY MEANS OF THE METHOD OF VIBRATIONAL CONTROL

G. L. ANDERSON
U.S. Army Research Office
Research Triangle Park, North Carolina 27709-2211

AND

I. G. TADJBAKHSH
Department of Civil Engineering
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

1. INTRODUCTION

A planar, simple pendulum with a stationary point of support has but a single stable equilibrium position, which is along the downward vertical. The dynamic stability of the inverted pendulum can be realized through the parametric loads that arise from the inertia forces induced in the pendulum when its point of support is made to move in an oscillatory manner. The horizontal component of support point motion acts as an ordinary forcing effect on the motion of the pendulum. The vertical motion of the support point is more interesting, however, since it acts in a parametric manner on the rotational motion of the pendulum. Thus, if the motion of the support point is periodic in time, the response problem then becomes mathematically one of the solution of Mathieu's differential equation, which possesses a periodic coefficient.

This means of stabilizing the inverted pendulum was

investigated by Stephenson [1] to [3] shortly after the turn of the century. In particular, he showed that if the pendulum's point of support undergoes a small amplitude but high frequency oscillatory motion in the vertical direction, then the inverted configuration becomes stable, and the pendulum performs small oscillations about this stabilized upward position. Since Stephenson's investigations on the subject of induced stability, numerous other investigators [4] to [46] have studied the possibility of stabilizing the inverted pendulum and other related mechanical systems.

Notable among the early publications that appeared on the possibility of induced stability is a paper by Lowenstern [6]. His approach was based upon the introduction of a new set of coordinates for the purpose of eliminating the effect of the rapid oscillations on those coordinates that do not receive the rapid periodic variations. Then he considered mean values of the pertinent coordinates over a period or a finite multiple of the periods of the imposed oscillations. In Lowenstern's averaged equations, all the coefficients become constants. This represents an early attempt to apply an averaging scheme to the class of problems in which the equations of motion contain periodic coefficients. Bogdanoff [8] employed Lowenstern's cocrdinates and devised a proof of the fact that for certain imposed motions the solutions of the oscillatory coefficient system and the constant coefficient system always remain close to

one another. In Bogdanoff's work, the parameters of the system are subject to rapid stochastic variation in time.

In a somewhat related investigation, Hsu [13] has studied the stability of inverted pendula whose supports are subjected to an oscillatory motion. In the case of systems of several degrees of freedom, he determined conditions under which the equations of motion could be decoupled into a set of independent Mathieu equations by means of a similarity transformation. The method of averaging was not used, and stability conditions were obtained from stability charts.

As mentioned above, the equation of motion of the inverted pendulum whose point of support is driven harmonically in time contains a periodic coefficient. Consequently, the condition that determines the state of stability of the system is frequently derived by means of Floquet theory (see, e.g., Bolotin [47]). This procedure, which is typically complicated and laborious, can be avoided through the application of the method of averaging according to a technique described by Volosov [18], [19]. In essence, the method of averaging serves to transform the original equations of motion containing periodic coefficients into a simpler set of equations of motion that have only constant coefficients. With this simplified system of equations, it becomes possible to apply the well known stability criteria for physical systems subjected to autonomous loading or support conditions. Meerkov [16], [17] has discussed this technique when

applied to the vibrations of mechanical systems with a discrete number of degrees of freedom. This technique has evolved into what is now sometimes called vibrational control theory. It consists of an analysis of the averaged equations of motion, where the objective is to stabilize the equilibrium configuration of a mechanical system (such as an inverted pendulum) through through the application of the appropriate high frequency, small amplitude motion of its point or surface of support.

The stability characteristics of a double pendulum with elastic hinges that is subjected to a constant follower force of magnitude P were studied by Ziegler [48]. He determined that such a non-conservatively loaded system becomes unstable by flutter when the critical value of the load is exceeded. Herrmann and Bungay [49] examined the stability of Ziegler's pendulum but assumed that the direction of the applied load P was determined by a parameter α called the tangency coefficient. When $\alpha = 0$, the load is a purely conservative force, and when $\alpha = 1$ it is a tangential or follower force. If $\alpha < 0$, the force is termed anti-tangential, if $0 < \alpha < 1$ sub-tangential, and if $\alpha > 1$ super-tangential. The system becomes unstable by divergence or flutter depending upon the value of the tangency coefficient. Later, Herrmann and Jong [50] considered the same problem for the case in which the influence of viscous damping in the hinges was also included. Tso and Fung [51] subsequently investigated the parametric instability of Ziegler's pendulum

under the combined actions of a purely tangential force ($\alpha = 1$) and sinusoidal base motion. In particular, they determined the conditions for instability in the non-resonance, parametric resonance, and combination resonance cases.

In the present investigation, Ziegler's pendulum with elastic hinges and subjected to an externally applied force of magnitude P whose orientation is specified by the tangency coefficient α is again considered. However, the base upon which one extremity of the pendulum is elastically restrained is made to undergo a sinusoidal oscillation along the undeformed axis of the system. It will be assumed here that the base motion is of small amplitude and high frequency. The goal is to stabilize the system by means of vibrational control, i.e., high frequency, low amplitude base motion. The equations of motion are linearized relative to the undeformed equilibrium configuration of the system, and the method of averaging is applied to the system of equations containing periodic coefficients in order to generate a simpler and more convenient system of equations with constant coefficients. This latter system serves for the purpose of easily computing the critical flutter or divergence loads for the system as a function of the tangency coefficient.

2. REVIEW OF THE METHOD OF VIBRATIONAL CONTROL

Before the question of determining the possibility of stabilizing Ziegler's pendulum through the action of high frequency, low amplitude base motion is addressed, it is worthwhile to consider the method of vibrational control as it may be applied to the general class of discrete physical systems whose motions are described by a system of second order linear differential equations with periodic coefficients. In this way, the requirements that the system must satisfy in order that the method of vibrational control can be applied will be exposed.

Consider the system of differential equations

$$\ddot{\underset{\sim}{\varphi}}(t) + \epsilon \underset{\sim}{D}(t)\, \dot{\underset{\sim}{\varphi}}(t) + [\epsilon^2 \underset{\sim}{B} + \epsilon \underset{\sim}{q}(t)]\, \underset{\sim}{\varphi}(t) = 0, \qquad (2.1)$$

where ϵ is a small parameter and $\underset{\sim}{D}(t)$ and $\underset{\sim}{q}(t)$ are known periodic n x n matrices. Moreover, it is assumed that

$$\underset{\sim}{D}(t) = \underset{\sim}{D}_o + \underset{\sim}{D}_1(t),$$

and that $\underset{\sim}{D}_1(t)$ and $\underset{\sim}{q}(t)$ have zero mean values, i.e.,

$$M\{\underset{\sim}{D}_1(t)\} = \lim_{T\to\infty} (1/T) \int_o^T \underset{\sim}{D}_1(t)\, dt = 0, \qquad (2.2)$$

$$M\{q(t)\} = \lim_{T\to\infty} (1/T) \int_o^T q(t)\, dt = 0. \qquad (2.3)$$

In order to apply the method of averaging, Equation (2.1) must first be transformed into a suitable form, namely, a

particular type of system of first order ordinary differential equations. For this purpose, it is convenient to introduce the following transformations:

$$\underset{\sim}{\Phi}(t) = \underset{\sim}{y}(t) + \epsilon \underset{\sim}{Q}(t)\underset{\sim}{y}(t), \quad \dot{\underset{\sim}{\Phi}}(t) = \epsilon \underset{\sim}{z}(t) + \epsilon \dot{\underset{\sim}{Q}}(t)\underset{\sim}{y}(t), \tag{2.4}$$

where $\underset{\sim}{Q}(t)$ is an arbitrary n x n matrix that must assure the periodic character of $\underset{\sim}{\Phi}(t)$. A condition that will serve for the determination of $\underset{\sim}{Q}(t)$ will be introduced later. It is also assumed that the mean value of the matrix $\underset{\sim}{Q}(t)$ shall vanish:

$$M\{\underset{\sim}{Q}(t)\} = \lim_{T\to\infty} (1/T) \int_0^T \underset{\sim}{Q}(t)\, dt = 0. \tag{2.5}$$

A differentiation of $\underset{\sim}{\Phi}$ in Equation (2.4) yields

$$\dot{\underset{\sim}{\Phi}} = \dot{\underset{\sim}{y}} + \epsilon \dot{\underset{\sim}{Q}}\underset{\sim}{y} + \epsilon \underset{\sim}{Q}\dot{\underset{\sim}{y}} = \epsilon \underset{\sim}{z} + \epsilon \dot{\underset{\sim}{Q}}\underset{\sim}{y}.$$

whence

$$(\underset{\sim}{I} + \epsilon \underset{\sim}{Q})\dot{\underset{\sim}{y}} = \epsilon \underset{\sim}{z},$$

or, upon rearrangement,

$$\dot{\underset{\sim}{y}} = \epsilon \underset{\sim}{R}\underset{\sim}{z}, \tag{2.6}$$

where

$$\underset{\sim}{R} = (\underset{\sim}{I} + \epsilon \underset{\sim}{Q})^{-1}, \tag{2.7}$$

with $\underset{\sim}{I}$ denoting the identity matrix. The derivative of the $\dot{\underset{\sim}{\Phi}}$ expression in Equation (2.4) is

$$\ddot{\underset{\sim}{\varphi}} = \epsilon \dot{\underset{\sim}{z}} + \epsilon \ddot{\underset{\sim}{Q}} \underset{\sim}{y} + \epsilon \dot{\underset{\sim}{Q}} \dot{\underset{\sim}{y}}. \tag{2.8}$$

Substitution of Equations (2.4) and (2.8) into Equation (2.1) leads to

$$\dot{\underset{\sim}{z}} = -(\ddot{\underset{\sim}{Q}} + \underset{\sim}{q})\underset{\sim}{y} - \dot{\underset{\sim}{Q}} \dot{\underset{\sim}{y}} - \epsilon[\underset{\sim}{D}\dot{\underset{\sim}{Q}} + \underset{\sim}{B} + \underset{\sim}{q}\underset{\sim}{Q} + \epsilon \underset{\sim}{B}\underset{\sim}{Q}]\underset{\sim}{y} - \epsilon \underset{\sim}{D}\underset{\sim}{z}. \tag{2.9}$$

Eliminating $\dot{\underset{\sim}{y}}$ between Equations (2.6) and (2.9), one obtains

$$\dot{\underset{\sim}{z}} = -(\ddot{\underset{\sim}{Q}} + \underset{\sim}{q})\underset{\sim}{y} - \epsilon(\underset{\sim}{B} + \underset{\sim}{q}\underset{\sim}{Q} + \underset{\sim}{D}\dot{\underset{\sim}{Q}} + \epsilon \underset{\sim}{B}\underset{\sim}{Q})\underset{\sim}{y} - \epsilon(\underset{\sim}{D} + \dot{\underset{\sim}{Q}}\underset{\sim}{R})\underset{\sim}{z}. \tag{2.10}$$

To this point, the matrix $\underset{\sim}{Q}(t)$ has been assumed to have zero mean value but has otherwise been left completely arbitrary. At this point, it is now required that

$$\ddot{\underset{\sim}{Q}}(t) + \underset{\sim}{q}(t) = 0, \tag{2.11}$$

so that Equation (2.10) assumes the form

$$\dot{\underset{\sim}{z}} = -\epsilon \underset{\sim}{U}\underset{\sim}{y} - \epsilon \underset{\sim}{V}\underset{\sim}{z}, \tag{2.12}$$

where

$$\underset{\sim}{U} = \underset{\sim}{B} + \underset{\sim}{q}\underset{\sim}{Q} + \underset{\sim}{D}\dot{\underset{\sim}{Q}} + \epsilon \underset{\sim}{B}\underset{\sim}{Q}, \tag{2.13}$$

$$\underset{\sim}{V} = \underset{\sim}{D} + \dot{\underset{\sim}{Q}}\underset{\sim}{R}. \tag{2.14}$$

Equations (2.6) and (2.12) are in the general form to which the method of averaging is applicable, namely,

$$\dot{\underset{\sim}{y}} = \epsilon \underset{\sim}{R}\underset{\sim}{z}, \qquad \dot{\underset{\sim}{z}} = -\epsilon \underset{\sim}{U}\underset{\sim}{y} - \epsilon \underset{\sim}{V}\underset{\sim}{z}. \tag{2.15}$$

The theoretical foundations of this technique are discussed in References [16] to [21] and [52]. The average values of the matrices $\underset{\sim}{R}$, $\underset{\sim}{U}$, and $\underset{\sim}{V}$ are defined to be

$$\underset{\sim}{R}^* = M\{\underset{\sim}{R}\}, \quad \underset{\sim}{U}^* = M\{\underset{\sim}{U}\}, \quad \underset{\sim}{V}^* = M\{\underset{\sim}{V}\}, \tag{2.16}$$

where

$$M\{\underset{\sim}{R}\} = \lim_{T\to\infty} (1/T) \int_o^T \underset{\sim}{R}\, dt, \tag{2.17}$$

etc. Substitution of Equations (2.7), (2.13), and (2.14) into Equation (2.16) yields

$$\underset{\sim}{R}^* = M\{(\underset{\sim}{I} + \epsilon\underset{\sim}{Q})^{-1}\}, \tag{2.18}$$

$$\underset{\sim}{U}^* = \underset{\sim}{B} + M\{\underset{\sim}{q}\underset{\sim}{Q}\} + M\{\underset{\sim}{D}\dot{\underset{\sim}{Q}}\} + \epsilon\underset{\sim}{B}M\{\underset{\sim}{Q}\},$$

$$= \underset{\sim}{B} + M\{\underset{\sim}{q}\underset{\sim}{Q}\} + M\{\underset{\sim}{D}_1\dot{\underset{\sim}{Q}}\}, \tag{2.19}$$

$$\underset{\sim}{V}^* = M\{\underset{\sim}{D}\} + M\{\dot{\underset{\sim}{Q}}\underset{\sim}{R}\}$$

$$= \underset{\sim}{D}_o + M\{\dot{\underset{\sim}{Q}}\underset{\sim}{R}\}, \tag{2.20}$$

where $\underset{\sim}{D} = \underset{\sim}{D}_o + \underset{\sim}{D}_1(t)$, Equation (2.2), and $M\{\underset{\sim}{Q}(t)\} = M\{\dot{\underset{\sim}{Q}}(t)\} = 0$ have been used. Now the averaged forms of the differential equations in Equation (2.15) are

$$\dot{\underset{\sim}{y}} = \epsilon\underset{\sim}{R}^*\underset{\sim}{z}, \quad \dot{\underset{\sim}{z}} = -\epsilon\underset{\sim}{U}^*\underset{\sim}{y} - \epsilon\underset{\sim}{V}^*z.$$

But

$$\ddot{\underset{\sim}{z}} = -\epsilon \underset{\sim}{U}^* \dot{\underset{\sim}{y}} - \epsilon \underset{\sim}{V}^* \dot{\underset{\sim}{z}}$$

$$= -\epsilon^2 \underset{\sim}{U}^* \underset{\sim}{R}^* \underset{\sim}{z} - \epsilon \underset{\sim}{V}^* \dot{\underset{\sim}{z}},$$

whence

$$\ddot{\underset{\sim}{z}} + \epsilon \underset{\sim}{V}^* \dot{\underset{\sim}{z}} + \epsilon^2 \underset{\sim}{U}^* \underset{\sim}{R}^* \underset{\sim}{z} = 0. \tag{2.21}$$

This is the important system of differential equations that the method of averaging generates. It is related to the original system in Equation (2.1) that possesses time dependent (indeed periodic) coefficients. Even though this new system in Equation (2.21) has constant coefficients, it still serves to furnish important information regarding the state of stability of the original physical system under consideration. In the sections that follow, the analysis described will be based upon the consequences of this simplified system of ordinary differential equations.

It is possible to integrate Equation (2.11) twice. Thus, a first integration leads to

$$\dot{\underset{\sim}{Q}}(t) = \dot{\underset{\sim}{Q}}(0) - \int_0^t q(x)\, dx. \tag{2.22}$$

A second integration gives

$$\underset{\sim}{Q}(t) = \underset{\sim}{Q}(0) + \dot{\underset{\sim}{Q}}(0)t - \int_0^t \int_0^y q(x)\, dx\, dy$$

$$= \underset{\sim}{Q}(0) + \dot{\underset{\sim}{Q}}(0)t - \int_0^t (t - x)q(x)\, dx. \tag{2.23}$$

Taking the averages of Equations (2.22) and (2.23), one finds

that

$$\dot{\underset{\sim}{Q}}(0) = M\{\int_0^t \underset{\sim}{q}(x)\,dx\} = M\{(T - t)\underset{\sim}{q}(t)\}$$

$$= \lim_{T\to\infty} (1/T) \int_0^T (T - t)\underset{\sim}{q}(t)\,dt \qquad (2.24)$$

and

$$\underset{\sim}{Q}(0) = -M\{\dot{\underset{\sim}{Q}}(0)t - \int_0^t (t - x)\underset{\sim}{q}(x)\,dx\}$$

$$= -(1/2)\lim_{T\to\infty} (1/T)\{\dot{\underset{\sim}{Q}}(0)T^2 - \int_0^T (T - t)^2\underset{\sim}{q}(t)\,dt\}. \qquad (2.25)$$

Once the constant matrices $\underset{\sim}{Q}(0)$ and $\dot{\underset{\sim}{Q}}(0)$ have been evaluated from Equations (2.24) and (2.25), respectively, the expression for $\underset{\sim}{Q}(t)$ becomes completely known from Equation (2.23).

Example. Let $\underset{\sim}{Q}(t) = [Q_{ij}(t)]$ and $\underset{\sim}{q}(t) = [q_{ij}(t)]$, where, for $i,j = 1(1)n$,

$$q_{ij}(t) = s_{ij} \sin t + r_{ij} \sin(\nu_{ij}t), \quad \nu_{ij} \neq 1, \qquad (2.26)$$

the quantities s_{ij}, r_{ij}, and ν_{ij} being known constants. The goal is to determine $\underset{\sim}{Q}(t)$ from Equation (2.23).

According to Equations (2.23) and (2.26), it follows that

$$Q_{ij}(t) = Q_{ij}(0) + \dot{Q}_{ij}(0)t - \int_0^t (t - x)[s_{ij} \sin x + r_{ij} \sin(\nu_{ij}x)]\,dx. \qquad (2.27)$$

Now

$$\int_0^t (t - x) \sin x\,dx = t - \sin t,$$

$$\int_0^t (t - x) \sin(\nu_{ij} x)\, dx = (1/\nu_{ij}^2)(\nu_{ij} t - \sin \nu_{ij} t),$$

so that Equation (2.24) leads to

$$\dot{Q}_{ij}(0) = \lim_{T\to\infty} (1/T) \int_0^T (T - t)[s_{ij} \sin t + r_{ij} \sin(\nu_{ij} t)]\, dt$$

$$= \lim_{T\to\infty} (1/T)[s_{ij}(T - \sin T) + (r_{ij}/\nu_{ij}^2)(\nu_{ij} T -$$

$$- \sin(\nu_{ij} T)]$$

$$= s_{ij} + r_{ij}/\nu_{ij}. \qquad (2.28)$$

To evaluate $Q_{ij}(0)$ in Equation (2.25), the following integral is required:

$$\int_0^T (T - t)^2 q_{ij}(t)\, dt = s_{ij} \int_0^T (T - t)^2 \sin t\, dt +$$

$$+ r_{ij} \int_0^T (T - t)^2 \sin(\nu_{ij} t)\, dt$$

$$= s_{ij}(T^2 + 2 \cos T - 2) + (r_{ij}/\nu_{ij}^3)[\nu_{ij}^2 T^2 +$$

$$+ 2 \cos(\nu_{ij} T) - 2]. \qquad (2.29)$$

Hence, from Equations (2.25) and (2.29), it is found that

$$Q_{ij}(0) = -\lim_{T\to\infty} (1/T)[s_{ij}(1 - \cos T) + (r_{ij}/\nu_{ij}^3)(1 -$$

$$- \cos(\nu_{ij} T)] = 0. \qquad (2.30)$$

Therefore, inserting Equations (2.28) and (2.30) into Equation (2.23), one has

$$Q_{ij}(t) = (s_{ij} + r_{ij}/\nu_{ij})t - \int_o^t (t - x)[s_{ij} \sin x +$$

$$+ r_{ij} \sin(\nu_{ij}x)] \, dx$$

$$= s_{ij} \sin t + (r_{ij}/\nu_{ij}^2)\sin(\nu_{ij}t). \qquad (2.31)$$

Consequently, given the elements $q_{ij}(t)$ in Equation (2.26), the corresponding elements $Q_{ij}(t)$ are determined to be those stated in Equation (2.31).

3. THE EQUATIONS OF MOTION

Consider now Zielger's double pendulum mounted on a movable base of mass m_1. The base moves as a rigid body in the x- and y-directions as shown in Figure 1. The xy-axes remain stationary, whereas the x'y'-axes translate with the movable base. The coordinates of the support point are $(x_o(t), y_o(t))$ relative to the stationary origin O. The mathematical model to be used here parallels that described by Herrmann [53]. The double pendulum consists of massless rods that carry masses m_2 and m_3 that are concentrated at their extremities. The rods are of length ℓ. Linear elastic hinges are present in the joints of the system. The associated rotational spring constants designated c_1 and c_2. The force of gravity acts in the direction of the negative y-axis. A force P of the follower type is applied at the free extremity of the double pendulum. The parameter α, called the tangency coefficient, measures the degree of deviation of P from the vertical direction. A force $\underline{F}$ defined by

$$\underline{F} = F_x \underline{i} + F_y \underline{j} \tag{3.1}$$

is applied at O' to cause the base of the system to translate in the xy-plane. In Figure 1, g denotes the acceleration of gravity, whereas φ_1 and φ_2 represent the angular displacements of the two rods in the pendulum relative to the vertical axis.

For the purpose of deriving the equations of motion, Kane's method has been used. It consists essentially of employing the

expression

$$K^*_i + K_i = 0, \quad i = 1,2,3, \tag{3.2}$$

where K^*_i and K_i are the generalized inertia forces and generalized active forces, respectively. Since Kane's method is described in considerable detail in References [54] to [56], the various steps in the derivation process will not be reported here.

The equations of motion for the physical system depicted in Figure 1 can be shown to be

$$(m_1 + m_2 + m_3)\ddot{x} - (m_2 + m_3)\ell\dot{\varphi}_1^2 \sin\varphi_1 + (m_2 + m_3)\ell\ddot{\varphi}_1 \cos\varphi_1$$
$$- m_3\ell\dot{\varphi}_2^2 \sin\varphi_2 + m_3\ell\ddot{\varphi}_2 \cos\varphi_2 + P \sin(\alpha\varphi_2) = F_x, \tag{3.3}$$

$$(m_1 + m_2 + m_3)\ddot{y} - (m_2 + m_3)\ell\dot{\varphi}_1^2 \cos\varphi_1 - (m_2 + m_3)\ell\ddot{\varphi}_1 \sin\varphi_1$$
$$- m_3\ell\dot{\varphi}_2^2 \cos\varphi_2 - m_3\ell\ddot{\varphi}_2 \sin\varphi_2 + (m_1 + m_2 + m_3)g +$$
$$+ P \cos(\alpha\varphi_2) = F_y, \tag{3.4}$$

$$(m_2 + m_3)\ell\ddot{x} \cos\varphi_1 - (m_2 + m_3)\ell\ddot{y} \sin\varphi_1 + (m_2 + m_3)\ell^2\ddot{\varphi}_1 +$$
$$m_3\ell^2\dot{\varphi}_2^2 \sin(\varphi_1 - \varphi_2) + m_3\ell^2\ddot{\varphi}_2 \cos(\varphi_1 - \varphi_2) -$$
$$- (m_2 + m_3)\ell g \sin\varphi_1 + (c_1 + c_2)\varphi_1 - c_2\varphi_2 -$$
$$- P\ell \sin(\varphi_1 - \alpha\varphi_2) = 0, \tag{3.5}$$

$$m_3\ell\ddot{x} \cos\varphi_2 - m_3\ell\ddot{y} \sin\varphi_2 - m_3\ell^2\dot{\varphi}_1^2 \sin(\varphi_1 - \varphi_2) +$$

$$+ m_3\ell^2\ddot{\varphi}_1 \cos(\varphi_1 - \varphi_2) + m_3\ell^2\ddot{\varphi}_2 - m_3\ell g \sin\varphi_2 +$$

$$+ c_2(\varphi_2 - \varphi_1) - P\ell \sin(1 - \alpha)\varphi_2 = 0. \qquad (3.6)$$

Suppose that the motion of the double pendulum is such that $|\varphi_j| << 1$ for $j = 1,2$, i.e., it undergoes small oscillations about the translating vertical axis y'. In this situation the non-linear differential equations in Equations (3.3) to (3.6) can be linearized relative to $\varphi_1 = \varphi_2 = 0$. The results of the linearization are the following:

$$(m_1 + m_2 + m_3)\ddot{x} + (m_2 + m_3)\ell\ddot{\varphi}_1 + m_3\ell\ddot{\varphi}_2 + P\alpha\varphi_2 = F_x, \qquad (3.7)$$

$$(m_1 + m_2 + m_3)\ddot{y} + (m_1 + m_2 + m_3)g + P = F_y, \qquad (3.8)$$

$$(m_2 + m_3)\ell\ddot{x} + (m_2 + m_3)\ell^2\ddot{\varphi}_1 + m_3\ell^2\ddot{\varphi}_2 + [c_1 + c_2 - P\ell -$$

$$- (m_2 + m_3)\ell(g + \ddot{y})]\varphi_1 + (P\ell\alpha - c_2)\varphi_2 = 0, \qquad (3.9)$$

$$m_3\ell\ddot{x} + m_3\ell^2\ddot{\varphi}_1 + m_3\ell^2\ddot{\varphi}_2 - c_2\varphi_1 - [c_2 - P\ell(1 - \alpha) -$$

$$- m_3\ell(g + y)]\varphi_2 = 0. \qquad (3.10)$$

Next let

$$m_1 = \sigma m \qquad (3.11)$$

where σ is a parameter and, as was done in Reference [53],

$$m_2 = 2m, \quad m_3 = m, \quad \text{and} \quad c_1 = c_2 = c. \qquad (3.12)$$

Morever, if F_x and F_y are given as explicit functions of time t,

then Equations (3.7) to (3.10) represent four differential equations in the four unknowns x, y, φ_1, and φ_2. Suppose further that the translational displacements of the point of support x(t) and y(t) are known explicitly, namely,

$$x(t) = 0, \quad y(t) = y_o \sin \Omega_1 t. \tag{3.13}$$

Under these assumptions, Equations (3.7) to (3.10) assume the forms

$$F_x = 3m\ell\ddot{\varphi}_1 + m\ell\ddot{\varphi}_2 + P\alpha\varphi_2, \tag{3.14}$$

$$F_y = (3 + \sigma)mg + P - (3 + \sigma)y_o\Omega_1^2 \sin \Omega_1 t, \tag{3.15}$$

and

$$3m\ell^2\ddot{\varphi}_1 + m\ell^2\ddot{\varphi}_2 + [2c - P\ell - 3m\ell(g - y_o\Omega_1^2 \sin \Omega_1 t)]\varphi_1 + (P\ell\alpha - c)\varphi_2 = 0, \tag{3.17}$$

$$m\ell^2\ddot{\varphi}_1 + m\ell^2\ddot{\varphi}_2 - c\varphi_1 + [c - P\ell(1 - \alpha) - m\ell(g - y_o\Omega_1^2 \sin \Omega_1 t)]\varphi_2 = 0. \tag{3.18}$$

The differential equations in Equations (3.17) and (3.18) are linear and homogeneous. It is important to note that each of these equations contains a term in which a periodic coefficient appears. For the determination of the state of stability of the vibrationally controlled system, it suffices to consider only Equations (3.17) and (3.18). These equations are particular

cases of the more general system of equations stated in Equation (2.1). However, if a small parameter can be introduced into the equations stated above, then the method of averaging can be applied, and the averaged system of equations in Equation (2.21) can be employed for the determination of the conditions for stability of the non-conservatively loaded double pendulum shown in Figure 1. It is much more convenient to deal with the averaged equations than with those stated in Equations (3.17) and (3.18).

In the event that $g = y_o = 0$, Equations (3.17) and (3.18) become identical to those considered by Herrmann and Bungay [49]. It may be observed that the dimensionless parameter σ associated with the mass of the movable base does not appear in the central differental equations in Equations (3.17) and (3.18).

It is convenient to put Equations (3.17) and (3.18) into a dimensionless form. For this purpose, the following quantities are introduced:

$$\tau = \Omega_1 t, \quad \epsilon = y_o/\ell, \quad \omega_o^2 = g/\ell,$$
$$\omega = \omega_o \ell/\Omega_1 y_o, \quad r = c/(m\Omega_1^2 y_o^2), \quad Q_o = P\ell/(m\Omega_1^2 y_o^2). \tag{3.19}$$

Using Equation (3.19) and combining Equations (3.17) and (3.18), one can easily express the dimensionless forms of the system of differential equations under consideration in the form required in Equation (2.1), namely,

$$\ddot{\underline{\varphi}}(\tau) + [\epsilon^2 \underline{B} + \epsilon \underline{q}(\tau)]\underline{\varphi}(\tau) = 0, \quad [\dot{\underline{\varphi}} = d\underline{\varphi}/d\tau], \tag{3.20}$$

since $\underline{D}(\tau) = 0$, i.e., no damping terms have been included in the present analysis, where the elements of the $\underline{B}$ and $\underline{q}(\tau)$ matrices are

$$B_{11} = (3r - Q_o - 3\omega^2)/2, \quad B_{12} = (-2r + Q_o + \omega^2)/2,$$
$$B_{21} = (-5r + Q_o + 3\omega^2)/2, \quad B_{22} = [4r + (2\alpha - 3)Q_o - 3\omega^2]/2, \tag{3.21}$$

$$s_{11} = 3/2, \quad s_{12} = -1/2, \quad s_{21} = -3/2, \quad s_{22} = 3/2, \tag{3.22}$$

with

$$\underline{q}(\tau) = \underline{s} \sin\tau. \tag{3.23}$$

Equation (3.20) is in the canonical form, so that the theoretical results developed in Section 2 can now be applied for the specific forms of the quantities stated in Equations (3.21) to (3.23).

The system of differential equations in Equation (3.20) desribes the small motions of a non-conservatively loaded double pendulum whose point of support is driven in a sinusoidal manner along a vertical axis. Restoring moments are exerted at the hinges located at the point of support and at the joint of the double pendulum. The intention here is to show that, when the point of support is driven at small amplitude and high frequency, the critical value of the applied force can be raised, i.e., the system can be stabilized.

4. THE AVERAGED EQUATIONS OF MOTION

In lieu of working with Equation (3.20), which has periodic coefficients, it is desirable to determine the explicit form of the differential equation in Equation (2.21) that has constant coefficients. To determine the quantities $\underset{\sim}{V}^*$ and $\underset{\sim}{U}^*\underset{\sim}{R}^*$, it is first necessary to evaluate $\underline{Q}(\tau)$. This has, in fact, already been accomplished in the Example in Section 2. Equations (3.23) and (2.26) are equivalent if $r_{ij} = 0$ for all i and j considered. It follows from Equation (2.31) that

$$\underline{Q}(\tau) = \underline{s}q(\tau), \tag{4.1}$$

where

$$q(\tau) = \sin\tau. \tag{4.2}$$

It is easily shown from Equations (2.18) and (4.1) that

$$\underset{\sim}{R}^* = M\{(\underline{I} + \epsilon \underline{s}q(\tau))^{-1}\} = M\{(\underline{I} + \epsilon \underset{\sim}{s}^{-1}q(\tau))/G(\tau)\}, \tag{4.3}$$

where

$$G(\tau) = 1 + 3\epsilon q(\tau) + 3\epsilon^2 q^2(\tau)/2. \tag{4.4}$$

Then from Equations (2.20) and (4.1) to (4.4), one finds that

$$\underline{V}^* = \underline{s}M\{\dot{q}(\tau)\underline{R}\} = \underline{s}M\{q(\tau)[\underline{I} + \epsilon s^{-1}q(\tau)]/G(\tau)\}$$

$$= \underline{s}M\{\dot{q}(\tau)\ G(\tau)\} + \epsilon M\{q(\tau)\dot{q}(\tau)/G(\tau)\} = 0 \tag{4.5}$$

since

$$M\{\dot{q}(\tau)/G(\tau)\} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\dot{q}}{1 + 3\epsilon q + 3\epsilon^2 q^2/2}\, d\tau$$

$$= \frac{1}{\epsilon\sqrt{3}} \lim_{T\to\infty} \frac{1}{T} \ln\left[\frac{2 + \epsilon(3 + \sqrt{3})q(T)}{2 + \epsilon(3 - \sqrt{3})q(T)}\right] = 0$$

and

$$M\{q(\tau)\dot{q}(\tau)/G(\tau)\} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{q\dot{q}}{1 + 3\epsilon q + 3\epsilon^2 q^2/2}\, d\tau$$

$$= \frac{1}{3\epsilon^2} \lim_{T\to\infty} \frac{1}{T} \ln[1 + 3\epsilon q(T) + 3\epsilon^2 q^2(T)/2] -$$

$$- M\{\dot{q}(\tau)/G(\tau)\} = 0.$$

Now Equation (4.3) yields

$$\underline{R}^* = \underline{I}M\{1/G(\tau)\} + \epsilon \underline{s}^{-1} M\{q(\tau)/G(\tau)\}$$

$$= \underline{I} - \epsilon(3\underline{I} - \underline{s}^{-1})M\{q(\tau)\} + \epsilon^2[(15/2)\underline{I}$$

$$- 3\underline{s}^{-1}]M\{q^2(\tau)\} - \epsilon^3[18\underline{I} -$$

$$- (15/2)\underline{s}^{-1}]M\{q^3(\tau)\} + \epsilon^4[(279/4)\underline{I} -$$

$$-18\underline{s}^{-1}]M\{q^4(\tau)\} + O(\epsilon^5) \qquad (4.6)$$

in view of Equations (4.2) and (4.3), where the right side is the result of expanding the quantity $1/G(\tau)$ in terms of the small parameter ϵ. But

$$q(\tau) = \sin\tau, \qquad q^2(\tau) = (1 - \cos 2\tau)/2,$$

$$q^3(\tau) = (3\sin\tau - \sin 3\tau)/4, \qquad q^4(\tau) = (3 - 4\cos 2\tau + \cos 4\tau)/8,$$

so that

$$M\{q(\tau)\} = M\{q^3(\tau)\} = 0,$$

$$M\{q^2(\tau)\} = 1/2, \qquad M\{q^4(\tau)\} = 3/8.$$

Therefore, Equation (4.6) is reduced to

$$\underline{R}^* = \underline{I} + \epsilon^2[(15/2)\underline{I} - 3\underline{s}^{-1}]/2 + 3\epsilon^4[(279/4)\underline{I} -$$

$$- 18\underline{s}^{-1}]/8 + O(\epsilon^5). \tag{4.7}$$

Because $\underline{D}_1(\tau) = 0$ and by virtue of Equation (4.1), Equation (2.19) becomes

$$\underline{U}^* = \underline{B} + \underline{s}^2 M\{q^2(\tau)\}$$

$$= \underline{B} + \underline{s}^2/2. \tag{4.8}$$

Therefore, the product $\underline{U}^*\underline{R}^*$ to a first approximation is, in view of Equations (4.7) and (4.8)

$$\underline{U}^*\underline{R}^* = \underline{B} + \underline{s}^2/2 + O(\epsilon^2). \tag{4.9}$$

Neglecting the ϵ^2-term in Equation (4.5), one has as the specific form of Equation (2.21) for the problem under consideration

$$\ddot{\underline{z}} + \epsilon^2\underline{A}\underline{z} = 0, \tag{4.10}$$

where

$$\underline{A} = \underline{B} + \underline{s}^2/2 \tag{4.11}$$

with the matrices $\underline{B}$ and $\underline{s}$ being defined in Equations (3.21) and (3.22). The elements of the matrix $\underline{A}$ are obviously constants; specifically, the A_{ij} are

$$A_{11} = (3 + 3r - Q_o - 3\omega^2)/2, \quad A_{12} = -(3/2 + 2r - Q_o - \omega^2)/2,$$

$$A_{21} = -(9/2 + 5r - Q_o - 3\omega^2)/2, \tag{4.12}$$

$$A_{22} = [3 + 4r + (2\alpha - 3)Q_o - 3\omega^2]/2,$$

by virtue of Equations (3.21) and (3.22).

5. STABILITY ANALYSIS

The determination of the state of stability of the sinusoidally driven Ziegler's pendulum is based upon a careful examination of the eigenvalues associated with the system of differential equations in Equation (4.10).

A solution of Equation (4.10) is sought in the following form:

$$\underline{z} = \underline{a}e^{i\lambda\tau}, \tag{5.1}$$

where $\underline{a}$ is a constant column vector, $i = (-1)^{1/2}$, and λ is the eigenvalue (i.e., the dimensionless natural frequency of the system) to be determined. Substitution of Equation (5.1) into Equation (4.10) yields the system of homogeneous algebraic equations

$$(\lambda^2\underline{I} - \epsilon^2\underline{A})\underline{a} = 0, \tag{5.2}$$

which has a non-trivial solution of and only if

$$\mathrm{Det}(\lambda^2\underline{I} - \epsilon^2\underline{A}) = 0,$$

whence, upon expansion,

$$\lambda^4 - \epsilon^2(A_{11} + A_{22})\lambda^2 + \epsilon^4(A_{11}A_{22} - A_{12}A_{21}) = 0, \tag{5.3}$$

where the A_{ij}'s have been given in Equation (4.12). In particular, Equation (5.3) can be expressed as

$$\lambda^4 - (\epsilon^2/2)[7r + 6(1 - \omega^2) + 2(\alpha - 2)Q_o]\lambda^2 +$$

$$+ (\epsilon^4/4)\{2(1 - \alpha)Q_o^2 + 2[\omega^2 - 3(1 - \alpha)(1 +$$

$$+ r - \omega^2)]Q_o + 2r^2 - 10r\omega^2 + 6\omega^4 + 9r/2 -$$

$$- 9\omega^2 + 9/4\} = 0. \qquad (5.4)$$

The forms of the dimensionless parameters in Equation (3.19) were introduced for the purpose of expressing the system of differential equations in Equations (3.17) and (3.18) in the canonical form of Equation (2.1). Now it is desirable to select a new set of parameters in order to conform more closely with previously published investigations on non-conservatively loaded systems, such as those reported in References [48] to [51] and [53], for example. Therefore, the following definitions are made:

$$Q = P\ell/c, \quad \gamma = mg\ell/c, \quad \xi = \epsilon = y_o/\ell,$$

$$\Omega = \ell\Omega_1(m/c)^{1/2}, \quad \sigma = \xi\Omega, \quad \tilde{\omega} = \sigma\lambda. \qquad (5.5)$$

The quantity Q denotes the dimensionless load parameter, γ the gravity parameter, and $\tilde{\omega}$ the natural frequency parameter. The quantity ξ is a measure of the amplitude, and Ω is the dimensionless frequency of the vertical sinusoidal motion of the point of support of the pendulum. The product of ξ and Ω is designated as σ; it is a measure of the motion of the point of

support. It is easily shown that $r = 1/\sigma^2$, $\omega^2 = \gamma/\sigma^2$, and $Q_o = Q/\sigma^2$. Using the quantities defined in Equation (5.5), one can express the frequency equation in Equation (5.4) as

$$\tilde{\omega}^4 - (\xi^2/2)[7 + 6(\sigma^2 - \gamma) + 2(\alpha - 2)Q]\tilde{\omega}^2 +$$

$$+ (\xi^4/4)\{2(1 - \alpha)Q^2 + 2[\gamma - 3(1 - \alpha)(1 - \gamma +$$

$$+ \sigma^2)]Q + 2 - 10\gamma + 6\gamma^2 + \sigma^2(9/2 - \gamma +$$

$$+ 9\sigma^2/4)\} = 0. \qquad (5.6)$$

The value of the dimensionless critical divergence load Q_d is determined from the condition of vanishing frequency ($\tilde{\omega} = 0$). In this case, Equation (5.6) becomes simply

$$8(1 - \alpha)Q_d^2 + 8[\gamma - 3(1 - \alpha)(1 - \gamma + \sigma^2)]Q_d + 8 -$$

$$- 40\gamma + 24\gamma^2 + \sigma^2(18 - 36\gamma + 9\sigma^2) = 0. \qquad (5.7)$$

This is a quadratic equation in Q_d, so its solution can be determined by elementary methods. It is evident that the value of the critical divergence load depends only upon the tangency coefficient, and the gravity and base motion parameters.

The value of the dimensionless critical flutter load Q_f is determined from the condition of the coalescence of the two natural frequencies of vibration of the system, i.e., $\tilde{\omega}_1 = \tilde{\omega}_2$. This implies that the discriminant of the quadratic equation in

$\tilde{\omega}^2$ in Equation (5.6) must vanish. Expansion of this discriminant leads to the following equation for Q_f:

$$4(2 - 2\alpha + \alpha^2)Q_f^2 + 4(\alpha - 8 + 4\gamma - 6\sigma^2)Q_f + 41 -$$

$$- 44\gamma + 12\gamma^2 + \sigma^2(66 - 36\gamma + 27\sigma^2) = 0. \qquad (5.8)$$

Equation (5.8) is a quadratic equation in Q_f. Obviously, the value of Q_f will depend upon the values of the tangency coefficient α, the gravity coefficient γ, and the support motion parameter σ.

If one sets $\gamma = \sigma = 0$ in Equations (5.7) and (5.8), they become

$$(1 - \alpha)Q_d^2 - 3(1 - \alpha)Q_d + 1 = 0, \qquad (5.9)$$

and

$$4(\alpha^2 - 2\alpha + 2)Q_f^2 + 4(\alpha - 8)Q_f + 41 = 0, \qquad (5.10)$$

respectively, which are the expressions equivalent to those reported by Herrmann and Bungay [49].

6. NUMERICAL RESULTS

In this section, the information regarding the state of stability of the double pendulum contained in Equations (5.7) and (5.8) will be extracted by analytical procedures, in so far as possible, and then by numerical computations. The goal is to plot stability diagrams, i.e., plots of the critical divergence and flutter loads versus the tangency coefficient α.

Equation (5.7) can be expressed as

$$8(1 - \alpha)Q_d^2 - 8(b_1 - \alpha b_2)Q_d + b_3 = 0, \tag{6.1}$$

where

$$b_1 = 3 - 4\gamma + 3\sigma^2, \quad b_2 = 3(1 - \gamma + \sigma^2),$$
$$b_3 = 8(1 - 5\gamma + 3\gamma^2) + 9\sigma^2(2 - 4\gamma + \sigma^2). \tag{6.2}$$

The solutions of Equation (6.1) are easily shown to be

$$Q_{d1} = \{8(b_1 - \alpha b_2) - 4\sqrt{2}[2(b_1 - \alpha b_2)^2 - (1 - \alpha)b_3]^{1/2}\}/16(1 - \alpha),$$
$$Q_{d2} = \{8(b_1 - \alpha b_2) + 4\sqrt{2}[2(b_1 - \alpha b_2)^2 - (1 - \alpha)b_3]^{1/2}\}/16(1 - \alpha). \tag{6.3}$$

Each expression in Equation (6.3) represents a branch of the stability diagram. As will become evident later, it is useful to know for which value or values of α will these two branches merge ($Q_{d1} = Q_{d2}$). This condition implies that

$$2(b_1 - \alpha b_2)^2 - (1 - \alpha)b_3 = 0,$$

which leads to the following quadratic equation in α:

$$2b_2^2\alpha^2 + (b_3 - 4b_1b_2)\alpha + 2b_1^2 - b_3 = 0. \qquad (6.4)$$

The solutions α_m of Equation (6.4) are given by

$$36(1 - \gamma + \sigma^2)^2\alpha_m = 4(7 - 11\gamma + 6\sigma^2) + 3\sigma^2(18 - 16\gamma +$$

$$+ 9\sigma^2) \pm [8(1 - 5\gamma + 3\gamma^2) + 9\sigma^2(2 - 4\gamma + \sigma^2)]^{1/2} \cdot$$

$$\cdot[8(1 - 2\gamma) + 3\sigma^2(6 - 4\gamma + 3\sigma^2)]^{1/2}. \qquad (6.5)$$

Thus, the values of α_m depend upon γ and σ in a rather complicated manner. In the event that gravity is absent ($\gamma = 0$), Equation (6.5) leads to

$$\alpha_{m1} = \frac{10 + 18\sigma^2 + 9\sigma^4}{18(1 + \sigma^2)^2}, \qquad \alpha_{m2} = 1. \qquad (6.6)$$

When $\sigma = 0$ (i.e., there is no vertical motion of the support point) it follows from Equation (6.6) that $\alpha_{m1} = 5/9$ and $\alpha_{m2} = 1$, which are the values reported in Reference [49]. In the other extreme as the value of tends to infinity, Equation (6.6) yields $\alpha_{m1} = 1/2$ and $\alpha_{m2} = 1$.

In analogy with Equation (6.1), it is convenient to express Equation (5.8) in a more compact form:

$$4a_oQ_f^2 + 4(\alpha - b_o)Q_f + c_o = 0, \qquad (6.7)$$

where

$$a_o = 2 - 2\alpha + \alpha^2, \quad b_o = 8 - 4\gamma + 6\sigma^2,$$
$$c_o = 41 - 44\gamma + 12\gamma^2 + \sigma^2(66 - 36\gamma + 27\sigma^2). \tag{6.8}$$

The solutions of Equation (6.7) are

$$Q_f = [b_o - \alpha \pm (\alpha^2 - 2b_o\alpha + b_o^2 - a_o c_o)^{1/2}]/2a_o, \tag{6.9}$$

which provides two branches of the stability diagram in the αQ-plane. These branches can exist only if the radicand in Equation (6.9) is positive. Limiting values of the tangency coefficient can be found when $Q_{f1} = Q_{f2}$, which implies that the radicand vanishes. This condition means that

$$(c_o - 1)\alpha^2 + 2(b_o - c_o)\alpha + 2c_o - b_o^2 = 0 \tag{6.10}$$

must hold. But its solutions are

$$\alpha_t = (33 - 40\gamma + 12\gamma^2 + 3\sigma^2(20 - 12\gamma + 9\sigma^2) \pm$$
$$\pm [41 - 44\gamma + 12\gamma^2 + \sigma^2(66 - 36\gamma + 27\sigma^2)]^{1/2}[9 -$$
$$- 12\gamma + 4\gamma^2 + 3\sigma^2(6 - 4\gamma + 3\sigma^2)]^{1/2})/[40 -$$
$$- 44\gamma + 12\gamma^2 + \sigma^2(66 - 36\gamma + 27\sigma^2)]. \tag{6.11}$$

If gravity is absent from the physical system under consideration, then $\gamma = 0$ and Equation (6.11) becomes

$$\alpha_t = \frac{3(1 + \sigma^2)[11 + 9\sigma^2 \pm (41 + 66\sigma^2 + 27\sigma^4)^{1/2}]}{(4 + 3\sigma^2)(10 + 9\sigma^2)}. \tag{6.12}$$

If the point of support is stationary ($\sigma = 0$), Equation (6.12) yields

$$\alpha_t = 3(11 \pm \sqrt{41})/40 = 0.3448,\ 1.3052,$$

which are the values reported in Reference [49]. As the value of tends to infinity, Equation (6.12) in the limit leads to

$$\alpha_t = 1 \pm \sqrt{3}/3 = 0.4226,\ 1.5773.$$

The expressions for the values of the critical divergence and flutter loads given in Equations (6.3) and (6.9) serve to furnish the stability boundaries in the αQ-plane that are plotted in stability diagrams. As a first set of stability diagrams, suppose that the gravitational force is absent ($\gamma = 0$). Then the values of Q_d and Q_f become functions of only the tangency coefficient α and the support motion parameter σ.

If the point of support remains stationary ($\sigma = 0$), then the stability diagram becomes that reported in Reference [49] and shown in Figure 2. The various regions of the diagram are delineated by boundaries of the divergence, flutter, and stability zones, as labeled in the figure. This plot reveals that the smallest critical load for the system will be a divergence load as long as $\alpha < \alpha_{m1} = 5/9$. Thus, even though the loading is non-conservative when $\alpha < \alpha_{m1}$ and $\alpha \neq 0$, the system becomes unstable by divergence. For $\alpha_{m1} < \alpha < \alpha_t = 3(11 + \sqrt{41})/40$, the smallest positive critical load is a flutter load.

A tensile critical load can lead to divergence when $\alpha > 1$, i.e., when the applied load is super-tangential. If $\alpha > 3(11 + \sqrt{41})/40$, then either a compressive or a tensile load of sufficient magnitude can lead to instability through divergence.

For $\sigma > 0$, it will be seen in Figures 3 to 5 that the stability boundaries will be shifted away from those in the reference stability diagram in Figure 2. In Figures 3 to 5, the stability boundaries have been plotted for $\sigma = 1$, 2, and 3, respectively, along with the boundaries for $\sigma = 0$ for purposes of comparison. For compressive critical loads, the numerical values are shifted higher along the Q-axis, whereas the boundary is displaced downward for tensile super-tangential loads. The effect of the displacement of the stability boundaries becomes more pronounced as the value of the support motion parameter σ is increased. Clearly, for a given value of the tangency coefficient α, the sinusoidal motion of the point of support of the double pendulum increases the value of the critical load, be it a divergence or a flutter load.

To render evident the stabilizing effect of the method of vibrational control when applied to this system, it is convenient to consider a couple of special values of α, namely $\alpha = 0$ and $\alpha = 1$. When the load is conservative ($\alpha = 0$), the system becomes unstable by divergence, and the value of its critical load can be calculated from

$$Q_{dl} = [6(1 + \sigma^2) - (20 + 36\sigma^2 + 18\sigma^4)^{1/2}]/4,$$

which is obtained from Equation (6.3) when $\alpha = \gamma = 0$. The variation of Q_{dl} versus σ as computed from this equation is plotted in Figure 6. It is clear that the value of Q_{dl} increases monotonically in a non-linear fashion as σ is increased. Indeed, it can be shown that

$$Q_{dl} = (3 - \sqrt{5})/2 + (30 - 9\sqrt{5})\sigma^2/20 - 9\sqrt{5}\sigma^4/400 + O(\sigma^6)$$

for small values of σ and that

$$Q_{dl} = 3(2 - \sqrt{2})[\sigma^2 + 1 - (2 + \sqrt{2})/18\sigma^2] + O(1/\sigma^4)$$

for large values of σ. For the critical flutter load in the case of a purely tangential force ($\alpha = 1$), Equation (6.9) leads to the expression

$$Q_{fl} = [7 + 6\sigma^2 - (8 + 18\sigma^2 + 9\sigma^4)^{1/2}]/2,$$

when $\gamma = 0$. The variation of Q_{fl} with σ is shown in Figure 6. For large values of σ, this formula leads to the expansion

$$Q_{fl} = 3\sigma^2/2 + 2 + 1/12\sigma^2 + O(1/\sigma^4).$$

It is clear that both Q_{dl} and Q_{fl} increase with increasing values of σ and that both vary asymptotically as σ^2. Therefore, the method of vibrational control offers the potential of increasing the value of the critical load by a rather significant amount

depending upon the value of the support motion parameter σ.

A plot of the variation of α_t as a function of σ as determined from Equation (6.12) is shown in Figure 7. It reveals that the values of α_{t1} and α_{t2} increase monotonically and very slowly with the increasing values of the support motion parameter.

Scrutiny of the flutter zone in the stability diagram shown in Figures 2 to 5 leads one to conclude that the value of the critical flutter load Q_f attains a minimum value in the αQ_f-plane at which point a horizontal tangent exists. It is of interest to compute min Q_f and the value of α, say α_M, at which this minimum occurs. To accomplish this objective, it will be convenient to use Equations (6.7) and (6.9). Differentiating Equation (6.7) with respect to α and imposing the condition that $dQ_f/d\alpha = 0$, one finds,

$$\min Q_f = 1/2(1 - \alpha_M). \tag{6.13}$$

The branch of Equation (6.9) that is pertinent to the present goal is the expression that contains the negative sign before the radical. If the derivative $dQ_f/d\alpha$ of Equation (6.9) is formed, then the condition $dQ_f/d\alpha = 0$ leads to

$$4(1 - \alpha_M)\min Q_f = 1 + [(1 - c_o)\alpha_M + c_o - b_o]/R^{1/2}, \tag{6.14}$$

where

$$R = (1 - c_o)\alpha_M^2 + 2(c_o - b_o)\alpha_M + b_o^2 - 2c_o. \qquad (6.15)$$

Substitution of Equation (6.13) into (6.14) yields, after rearrangement,

$$R^{1/2} = (1 - c_o)\alpha_M + c_o - b_o.$$

Squaring both sides of this expression and rearranging the result, one finds

$$(c_o - 1)\alpha_M^2 + 2(b_o - c_o)\alpha_M + 2c_o - 2b_o = 0,$$

whence

$$\alpha_M = (33 - 40\gamma + 12\gamma^2 + 3\sigma^2(20 - 12\gamma + 9\sigma^2) \pm$$

$$\pm [(3 - 2\gamma)^2 + 3\sigma^2(6 - 4\gamma + 3\sigma^2)]^{1/2}/(4(10 - 11\gamma +$$

$$+ 3\gamma^2) + 3\sigma^2(22 - 12\gamma + 9\sigma^2)). \qquad (6.16)$$

In the special case of the absence of the gravitational force ($\gamma = 0$), Equation (6.16) becomes simply

$$\alpha_M = \frac{3(1 + \sigma^2)}{4 + 3\sigma^2}, \quad \frac{9(1 + \sigma^2)}{10 + 9\sigma^2}. \qquad (6.17)$$

The first of the values for α_M in Equation (6.17) pertains to the minimum. Therefore, with this value of α_M, Equation (6.13) becomes simply

$$\min_{\alpha} Q_f = 2 + 3\sigma^2/2, \qquad (6.18)$$

a remarkably simple result. Hence, if the support point remains immobile ($\sigma = 0$), then min $Q_f = 2$. Otherwise the value of min Q_f increases with the square of the base motion parameter σ.

It is possible to derive from Equation (6.9) a relatively simple asymptotic expression for Q_f as σ becomes very large. The process is very straightforward, so the details are not repeated here. The result can be shown to be

$$Q_f \sim \frac{3\sigma^2[2 - (6\alpha - 2 - 3\alpha^2)^{1/2}]}{2(2 - 2\alpha + \alpha^2)} \quad \text{as } \sigma \to \infty. \tag{6.19}$$

In particular, when $\alpha = 1$, i.e., the follower force is tangential, $Q_f \sim 3\sigma^2/2$ as σ tends to infinity.

When the value of the dimensionless gravity parameter γ is positive and relatively small, the stability diagrams continue to resemble those in Figures 3 to 5, but now with the divergence boundaries displaced toward the α-axis. A transition in the forms of these boundaries occurs when the value of γ reaches a particular level. This happens when the coefficient b_3 in Equation (6.1) vanishes. Thus, when

$$b_3 = 8(1 - 5\gamma + 3\gamma^2) + 9\sigma^2(2 - 4\gamma + \sigma^2) = 0, \tag{6.20}$$

Equation (6.1) assumes the form

$$Q_d[(1 - \alpha)Q_d - b_1 + \alpha b_2] = 0,$$

whence

$$Q_d = 0 \quad \text{and} \quad Q_d = 3(1 - \gamma + \sigma^2) - \gamma/(1 - \alpha). \tag{6.21}$$

Solving Equation (6.20) for the transition values of γ, say, γ_T, one finds

$$\gamma_T = [10 + 9\sigma^2 \pm (52 + 72\sigma^2 + 27\sigma^4)^{1/2}]/12. \tag{6.22}$$

Therefore, from Equations (6.21) and (6.22), it follows that the second expression for Q_d becomes

$$Q_{dT} = 3(1 + \sigma^2) + (3\alpha - 4)[10 + 9\sigma^2 \pm (52 + 72\sigma^2 + 27\sigma^4)^{1/2}]/12(1 - \alpha) \tag{6.23}$$

for $\alpha \neq 1$. It should be noted from Equation (6.1) that when $\alpha = 1$ then $Q_{dT} = 0$ is the only solution. Now there is a value of the tangency coefficient α, say α_o, at which the value of Q_{dT} will vanish. From Equation (6.21), this is easily shown to be

$$\alpha_o = [3(1 + \sigma^2) - 4\gamma_T]/3(1 + \sigma^2 - \gamma_T). \tag{6.24}$$

As a special case of the results shown in Equations (6.22) to (6.24), one has for $\sigma = 0$ (the point of support is stationary)

$$\gamma_T = (5 \pm \sqrt{13})/6,$$

$$Q_{dT} = 3 + (3\alpha - 4)(5 \pm \sqrt{13})/6(1 - \alpha), \tag{6.25}$$

$$\alpha_o = (9 \pm \sqrt{13})/6.$$

In particular for $\gamma_T = (5 - \sqrt{13})/6 = 0.2324$, it follows from

Equation (6.25) that $\alpha_o = (9 - \sqrt{13})/6 = 0.8991$. In Figure 8, the plotted boundaries of the divergence zones are determined from $Q_{dT} = 0$ and

$$Q_{dT} = 3 + (3\alpha - 4)(5 - \sqrt{13})/6(1 - \alpha). \qquad (6.26)$$

The boundaries for the flutter zone are computed as before. It may be observed from this figure that $Q_d = 0$ is a critical load when $\gamma = (5 - \sqrt{13})/6$ and $\sigma = 0$. This implies that the double pendulum will collapse under the weight of the concentrated masses. A new feature appears in this plot, namely, the aspect that for $\sigma = 0$ the left hand divergence boundaries are characterized by a pair of intersecting curves rather than a pair of branches that terminate in a point at which a vertical tangent exists. However, as soon as the point of support begins to oscillate at small amplitude and high frequency, the intersection feature disppears and the more familiar shape of the boundary is restored. To illustrate this, the divergence and flutter boundaries have also been plotted in Figure 8 for the combination of parameters $\sigma = 1$ and $\gamma = (5 - \sqrt{13})/6$. Again, it is evident that the oscillation of the support point tends to improve the stability properties of the system.

For $\gamma > \gamma_T$, the boundaries of the divergence zones in the stability map change their character still more. For example, the right hand branch will now intersect the vertical line $\alpha = 1$ and will then become asymptotic to it. Thus, for a purely

tangential force ($\alpha = 1$). Equation (6.1) leads to a single root for Q_d, namely,

$$Q_d = -[8(1 - 5\gamma + 3\gamma^2) + 9\sigma^2(2 - 4\gamma + \sigma^2)]/8\gamma. \qquad (6.27)$$

For example, when $\gamma = 1/4$ and $\sigma = 1/2$, Equation (6.27) leads to $Q_d = -37/32 = -1.15625$.

Since the stability boundary curve intersects the line $\alpha = 1$ and eventually becomes asymptotic to it, it follows that this curve must possess a vertical tangent for a value of α that is probably slightly less than unity. An expression for this value of α will now be determined. If Equation (6.1) is differentiated with respect to Q_d and the derivative $d\alpha/dQ_d$ is equated to zero, the result is, after a little rearrangement,

$$Q_d = (b_1 - \alpha b_2)/2(1 - \alpha). \qquad (6.28)$$

Substitution of Equation (6.27) into Equation (6.1) yields

$$2(b_1 - \alpha b_2)^2 = b_3(1 - \alpha),$$

which is a quadratic equation in α. The solutions of this are easily shown to be

$$\alpha = \{4b_1b_2 - b_3 \pm [b_3(b_3 + 8b_2^2 - 8b_1b_2)]^{1/2}\}/4b_2^2,$$

or, more explicitly,

$$\alpha = \{4(7 - 11\gamma + 6\gamma^2) + 3\sigma^2(18 - 16\gamma + 9\sigma^2) +$$

$$\pm [8(1 - 5\gamma + 3\gamma^2) + 9\sigma^2(2 - 4\gamma + \sigma^2)]^{1/2}$$

$$[8(1 - 2\gamma) + 3\sigma^2(6 - 4\gamma + 3\sigma^2)]^{1/2}/$$

$$36(1 - \gamma + \sigma^2)^2. \qquad (6.29)$$

In the special case of $\gamma = 1/4$ and $\sigma = 1/2$, Equation (6.29) yields

$$\alpha = (491 \pm \sqrt{4921})/576 = 0.7306,\ 0.9742.$$

The second of these values is indeed slightly less than unity as was foreseen above. The corresponding values of Q_d can be computed from Equation (6.28). The results are $Q_d = 1.0359$ and $Q_d = -3.3484$, respectively.

With the help of the information assembled in the two preceding paragraphs, it is now possible to plot the stability map for $\gamma > \gamma_T$. As a case in point, Figure 9 has been plotted for $\gamma = 1/4$ and $\sigma = 0$ (the dashed boundaries) and $\sigma = 1/2$ (the solid boundaries). It is to be observed that for $\sigma = .0$ the boundaries of the left hand divergence zone no longer coalesce for some value of α in the domain $0 < \alpha < 1$, as was the case in Figures 2 to 4. Instead of curving upward as the value of α is increased, the lower branch curves abruptly downward and tends rapidly in an asymptotic sense to negative infinity as the value of α tends to unity. The upper branch of the left hand divergence zone continues to show a tendency to decrease with

increasing α. It initially appears to approach the lower branch but then veers away as α approaches unity. Thereafter, for increasing α, the value of Q_d decreases monotonically and tends asymptotically to $Q_d = 0$. The region between these two branches is a zone of divergence. The region below the lower branch is a zone of stability. The right hand branch of the divergence boundary remains very similar to those already seen in Figures 2 to 4. The region above it is a zone of divergence. The flutter boundary is also quite similar to the analogous boundaries shown in Figures 2 to 4. The region bounded by the upper boundary of the left hand divergence branch, the lower boundary of the flutter zone, and the upper boundary of the right hand divergence branch enclose a zone of stability.

When the point of support is made to oscillate harmonically such that σ assumes the value $\sigma = 1/2$, the boundary curves are not only shifted in the same senses as they were in Figures 2 to 4, but the branches of the left hand divergence zone now once again coalesce. The low amplitude, high frequency motion of the point of support obviously overcomes the influence of gravity and stabilizes the system. If the value of σ were increased, the degree of stabilization would be increased accordingly.

7. CONCLUSIONS

The method of vibrational control, as applied here, has been shown to be successful in stabilizing the double pendulum subjected to an external force of the follower type. The pendulum has restoring spring hinges at its point of support and at the point of connection of its two links. Since the induced motion of the point of support is made to oscillate sinusoidally at small amplitude and high frequency, the ordinary differential equations of motion of the system, when expressed in the proper form, can be subjected to the averaging process, which serves to replace the periodic coefficients with constant coefficients. This renders the equations of motion amenable to the elementary techniques of stability analysis and avoids the necessity of dealing with coupled systems of equations of the Mathieu type and Floquet theory.

To convert the original system of differential equations with periodic coefficients into a form suitable for the application of the method of averaging, a convenient linear transformation was introduced in Equation (2.4). The averaging process was shown to lead to a system of equations - see Equation (2.21) - that has only constant coefficients. The averaged system is related to the original system through the transformations given in Equation (2.4) and $\dot{\underline{y}} = \epsilon \underline{R}^{*} \underline{z}$. It has been shown by Sethna [20] that the stability properties determined

from the averaged system are a sub-set of those that can be derived from the original system with periodic coefficients. This observation then permits the drawing of conclusions regarding the stabilization of the double pendulum under consideration based upon the averaged system of equations of motion.

It was shown in Sections 5 and 6 that the shape of the stability boundaries depends upon the gravity parameter γ and the induced support motion parameter σ. It may be recalled from Equation (5.5) that

$$\gamma = mg\ell/c \quad \text{and} \quad \sigma = y_o\Omega_1(m/c)^{1/2}.$$

Thus, σ is essentially the product of the amplitude y_o and the frequency Ω_1 of the motion of the point of support. Stability maps have been drawn for representative values of the parameters γ and σ to illustrate their effects upon the state of stability of the system. All the calculations reported here reveal that the values of the critical divergence and flutter loads for a given value of the tangency coefficient α can be raised significantly by increasing the value of the parameter σ.

REFERENCES

1. A. Stephenson. On induced stability. Philosophical Magazine, 15, pp. 233-236 (1908).

2. A. Stephenson. A new type of dynamical stability. Proceedings of the Manchester Literary and Philosophical Society, 52, pp. (1908).

3. A. Stephenson. On induced stability. Philosophical Magazine, 17, pp. 765-766 (1909).

4. P. Hirsch. Das Pendel mit oszillierendem Aufhaengepunkt. Zeitschrift fuer angewandte Mathematik und Mechanik, 10, pp. 41-52 (1930).

5. K. Klotter. Stabilisierung und Labilisierung durch Schwingungen. Forschung Ingenieur-Wesenheit, 12, pp. 209-225 (1941).

6. E. R. Lowenstern. The stabilizing effect of imposed oscillations of high frequency on a dynamical system. Philosophical Magazine, Ser. 7, 13, pp. 458-486 (1932).

7. J. L. Bogdanoff. Influence on the behavior of a dynamical system of some imposed rapid motions of small amplitude. Journal of the Acoustical Society of America, 34, pp. 1055-1062 (1962).

8. A. D. S. Barr and D. C. McWhannell. Parametric instability in structures under support motion. Journal of Sound and Vibration, 14, pp. 491-509 (1971).

9. D. J. Ness. Small oscillations of a stabilized, inverted pendulum. American Journal of Physics, 35, pp. 964-967 (1967).

10. W. Haacke. Bemerkungen zur Stabilisierung eines physikalischen Pendels. Zeitschrift fuer angewandte Mathematik und Mechanik, 31, pp. 161-169 (1951).

11. W. K. Tso and K. G. Asmis. Parametric excitation of a pendulum with bilinear hysteresis. Journal of Applied Mechanics, 37, pp. 1061-1068 (1970).

12. R. Gradewald and W. Moldenhauer. Untersuchung der Stabilitaet des parametererregten Pendels als speziellen rheonichtlinearen Schwinger. Annalen der Physik, 27, pp. 359-364 (1971).

13. C. S. Hsu. On a restricted class of coupled Hill's equations and some applications. Journal of Applied Mechanics, 28, pp. 551-556 (1961).

14. A. I. Menyailov and A. V. Movchan. Stabilization of a pendulum ring system under conditions of vibration of the base. Izvestia AN SSSR Mekhanika Tverdogo Tela, 19, pp. 32-36 (1984).

15. L. Trave, A. M. Tarras, and A. Tilti. An application of vibrational control to cancel unstable decentralized fixed modes. IEEE Transactions on Automatic Control, AC-30, pp. 283-286 (1985).

16. S. M. Meerkov. Vibrational control theory. Journal of the Franklin Institute, 303, pp. 117-128 (1977).

17. S. M. Meerkov. Principle of vibrational control: theory and applications. IEEE Transactions on Automatic Control, AC-25, pp. 755-762 (1980).

18. V. M. Volosov. The method of averaging. Soviet Mathematics Doklady, 2, pp. 221-224 (1961).

19. V. M. Volosov. Higher approximations in averaging. Soviet Mathematics Doklady, 2, pp. 382-385 (1961).

20. P. R. Sethna. An extension of the method of averaging. Quarterly of Applied Mathematics, 25, pp. 205-211 (1967).

21. N. N. Bogoliubov and Yu. A. Mitropolski. **Asymptotic Methods in the Theory of Non-Linear Oscillations**. New York: Gordon and Breach (1967).

22. K. G. Valeev. Dynamic stabilization of unstable systems. Izvestia AN SSSR Mekhanika Tverdogo Tela, 16, pp. 9-17 (1971).

23. R. Mitchell. Stability of the inverted pendulum subjected to almost periodic and stochastic base motion - an application of the method of averaging. International Journal of Non-Linear Mechanics, 7, pp. 101-123 (1972).

24. C. S. Hsu. On the parametric excitation of a dynamic system having multiple degrees of freedom. Journal of Applied Mechanics, 30, pp. 367-372 (1963).

25. M. S. Howe. The mean square stability of an inverted pendulum subject to random parametric excitation. Journal of Sound and Vibration, 32, pp. 407-421 (1974).

26. B. Van der Pol. Stabiliseering door kleine trillingen. Physica: Nederlandsch Tijdschrift voor Natuurkunde, 5, pp. 157-162 (1925).

27. M. J. O. Strutt. Stabiliseering en labiliseering door trillingen. Physica: Nederlandsch Tijdschrift voor Natuurkunde, 7, pp. 265-271 (1927).

28. P. R. Sethna and G. W. Hemp. Non-linear oscillations of a gyroscopic pendulum with an oscillating point of suspension. Proc. Colloq. Internat. du Centre National de la Recherche Scientifique N-148, Les Vibrations Forcees dans les Systèmes Non-Lineaires, pp. 375-392 (1964).

29. G. W. Hemp and P. R. Sethna. On dynamical systems with high frequency parametric excitation. International Journal of Non-Linear Mechanics, 3, pp. 351-365 (1968).

30. F. M. Phelps and J. H. Hunter. An analytical solution of the inverted pendulum. American Journal of Physics, 33, pp. 285-295 (1965).

31. S. S. Joshi. Inverted pendulum with damping. American Journal of Physics, 34, p. 533 (1966).

32. F. M. Phelps and J. H. Hunter. Reply to Joshi's comments on a damping term in the equations of motion of the inverted pendulum. American journal of Physics, 34, pp. 533-535 (1966).

33. J. Dugundji and C. K. Chhatpar. Dynamic stability of a pendulum under parametric excitation. Rev. Roum. Sci. Techn.-Mec. Appl., 15, pp. 741-763 (1970).

34. T. J. Moran. Transient motions in dynamical systems with high frequency parametric excitation. International Journal of Non-Linear Mechanics, 5, pp. 633-644 (1970).

35. I. I. Blekhman. The development of the concept of direct separation of motions in non-linear mechanics. In **Advances in Theoretical and Applied Mechanics**, edited by A. Ishlinkshy and F. Chernousko, pp. 127-147 (1981).

36. S. M. Meerkov. Vibrational control. Automation and Remote Control, 34, pp. 201-209 (1973).

37. S. M. Meerkov. Averaging of trajectories of slow dynamic systems. Differential Equations, 9, pp. 1239-1245 (1973).

38. S. M. Meerkov and M. Yu. Tsitkin. The effectiveness of the method of vibrational control for dynamic systems described by a differential equation of order n. Automation and Remote Control, 36, pp. 525-529 (1975).

39. R. Bellman, J. Bentsman, and S. M. Meerkov. Non-linear systems with fast parametric oscillations. Journal of Mathematical Analysis and Applications, 97, pp. 572-589 (1983).

40. S. M. Meerkov. Vibrational stabilizability of distributed parameter systems. Journal of Mathematical Analysis and Applications, **98**, pp. 408-418 (1984).

41. R. Bellman, J. Bentsman, and S. M. Meerkov. Stability of fast periodic systems. IEEE Transactions on Automatic Control, **AC-30**, pp. 289-291 (1985).

42. R. Bellman, J. Bentsman, and S. M. Meerkov. On vibrational stabilizability of non-linear systems. Journal of Optimization Theory and Applications, **46**, pp. 421-430 (1985).

43. J. L. Bogdanoff and S. J. Citron. On the stabilization of the inverted pendulum. In **Developments in Mechanics**, Vol. 3, Proceedings of the Ninth Midwestern Mechanics Conference. New York: John Wiley, pp. 3-15 (1965).

44. L. D. Akulenko. Control of motion of a non-linear vibratory system by displacement of the equilibrium point. Isvestia AN SSSR Mekhanika Tverdogo Tela, **13**, pp. 1-9 (1978).

45. P. L. Kapitsa. The dynamic stability of a pendulum with a vibrating point of suspension. Journal of Experimental and Theoretical Physics, **21**, pp. 588-598 (1951).

46. V. N. Chelomei. On the possibility of enhancing the stability of elastic systems by means of vibrations. Doklady of the USSR Academy of Sciences, **110**, pp. 345-347 (1956).

47. V. V. Bolotin. **Dynamic Stability of Elastic Systems**. San Francisco: Holden-Day, Inc. (1964).

48. H. Ziegler. Die Stabilitaetskriterien der Elastomechanik. Ingenieur-Archiv, **20**, pp. 49-56 (1952).

49. G. Herrmann and R. W. Bungay. On the stability of elastic systems subjected to non-conservative forces. Journal of Applied Mechanics, **31**, pp. 435-440 (1964).

50. G. Herrmann and I. C. Jong. On non-conservative stability problems of elastic systems with slight damping. Journal of Applied Mechanics, **33**, pp. 125-132 (1966).

51. W. K. Tso and D. P. K. Fung. Dynamic instability under the combined actions of non-conservative loading and base motion. Journal of Applied Mechanics, **38**, pp. 1074-1076 (1971).

52. J. A. Sanders and F. Verlust. **Averaging Methods in Non-Linear Dynamical Systems**. New York: Springer-Verlag (1985).

53. G. Herrmann. **Dynamics and Stability of Mechanical Systems with Follower Forces**. NASA Contractor's Report CR-1782 (1971).

54. T. R. Kane and D. A. Levinson. Formulation of equations of motion for complex spacecraft. Journal of Guidance and Control, 3, pp. 99-112 (1980).

55. T. R. Kane, P. W. Likins, and D. A. Levinson. **Spacecraft Dynamics**. New York: McGraw-Hill Book Company (1983).

56. T. R. Kane and D. A. Levinson. **Dynamics: Theory and Applications**. New York: McGraw-Hill Book Company (1985).

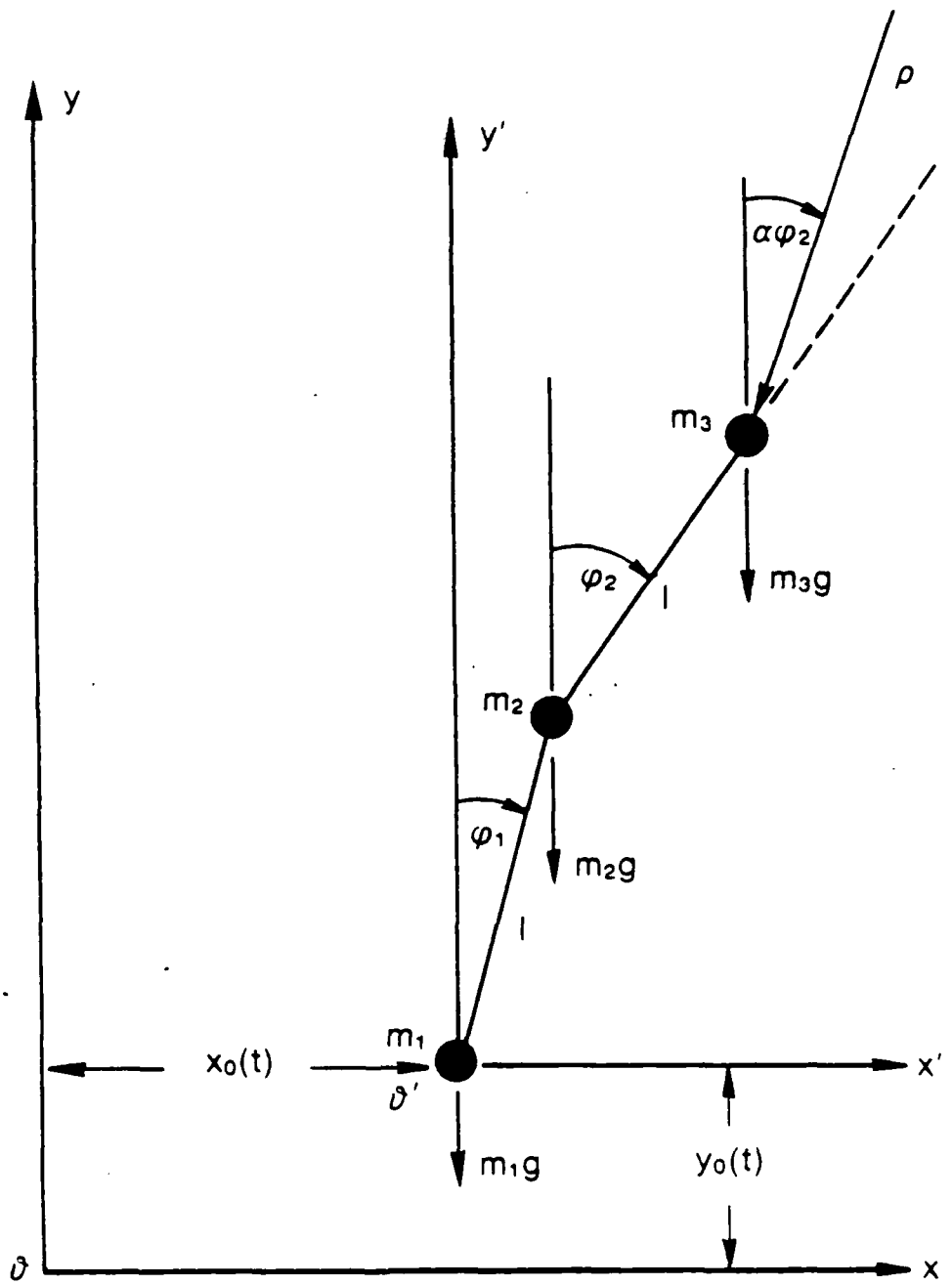


Figure 1. Coordinate systems for the double pendulum supported on a movable base.

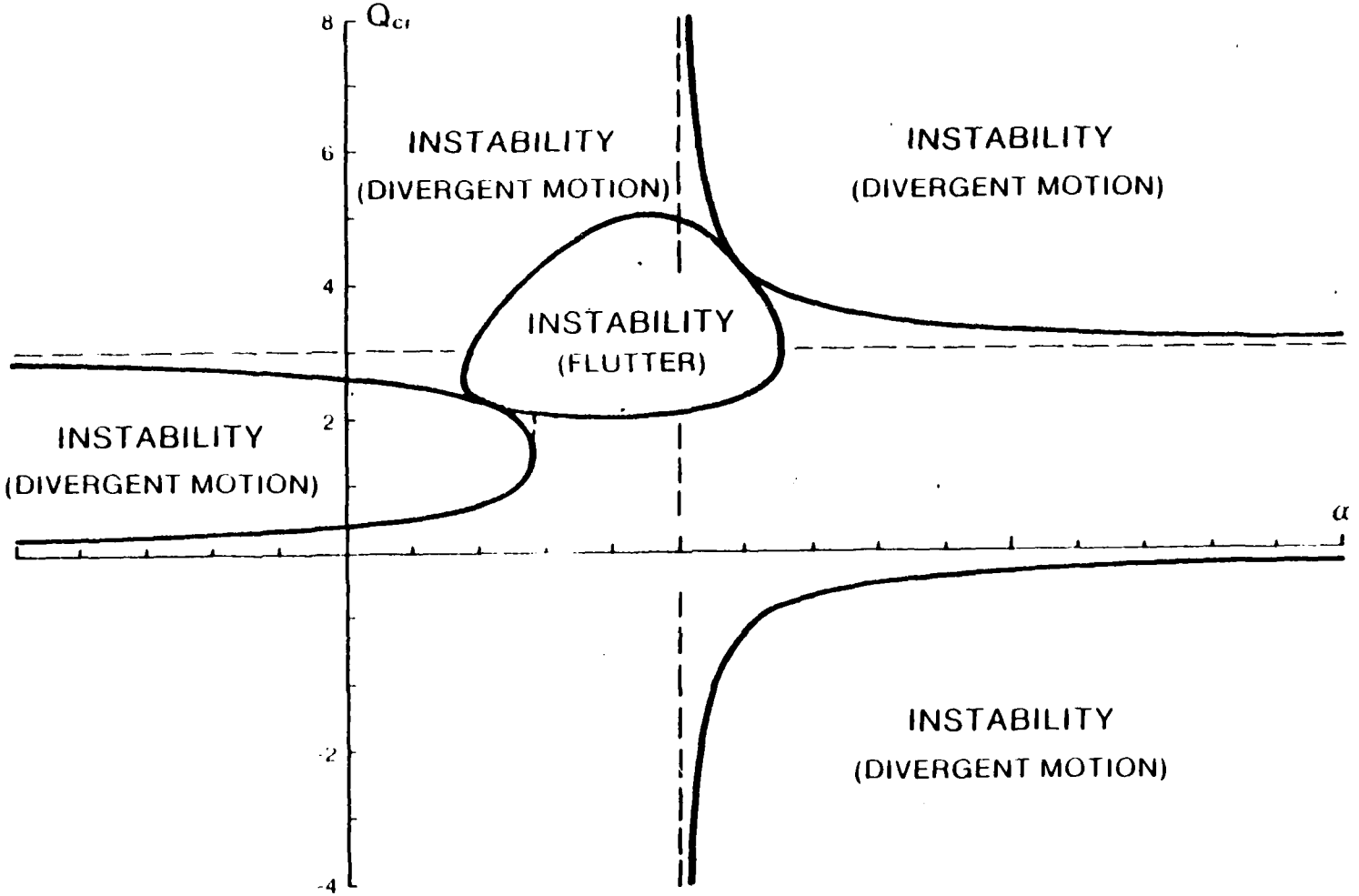


Figure 2. Stability diagram for $\sigma = 0$, $\gamma = 0$.

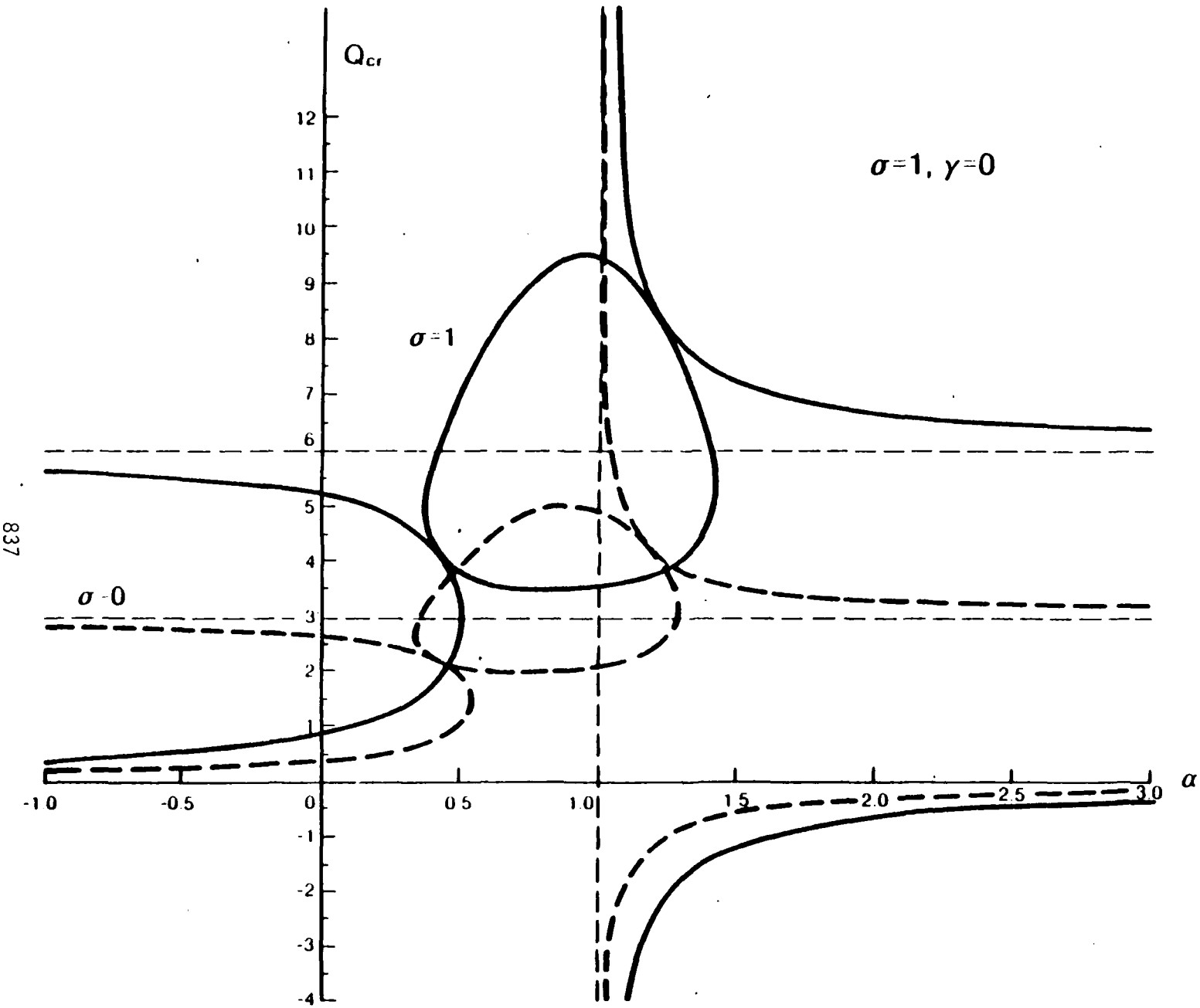


Figure 3. Stability diagrams for σ = 1 —— and σ = 0 ---, with γ = 0.

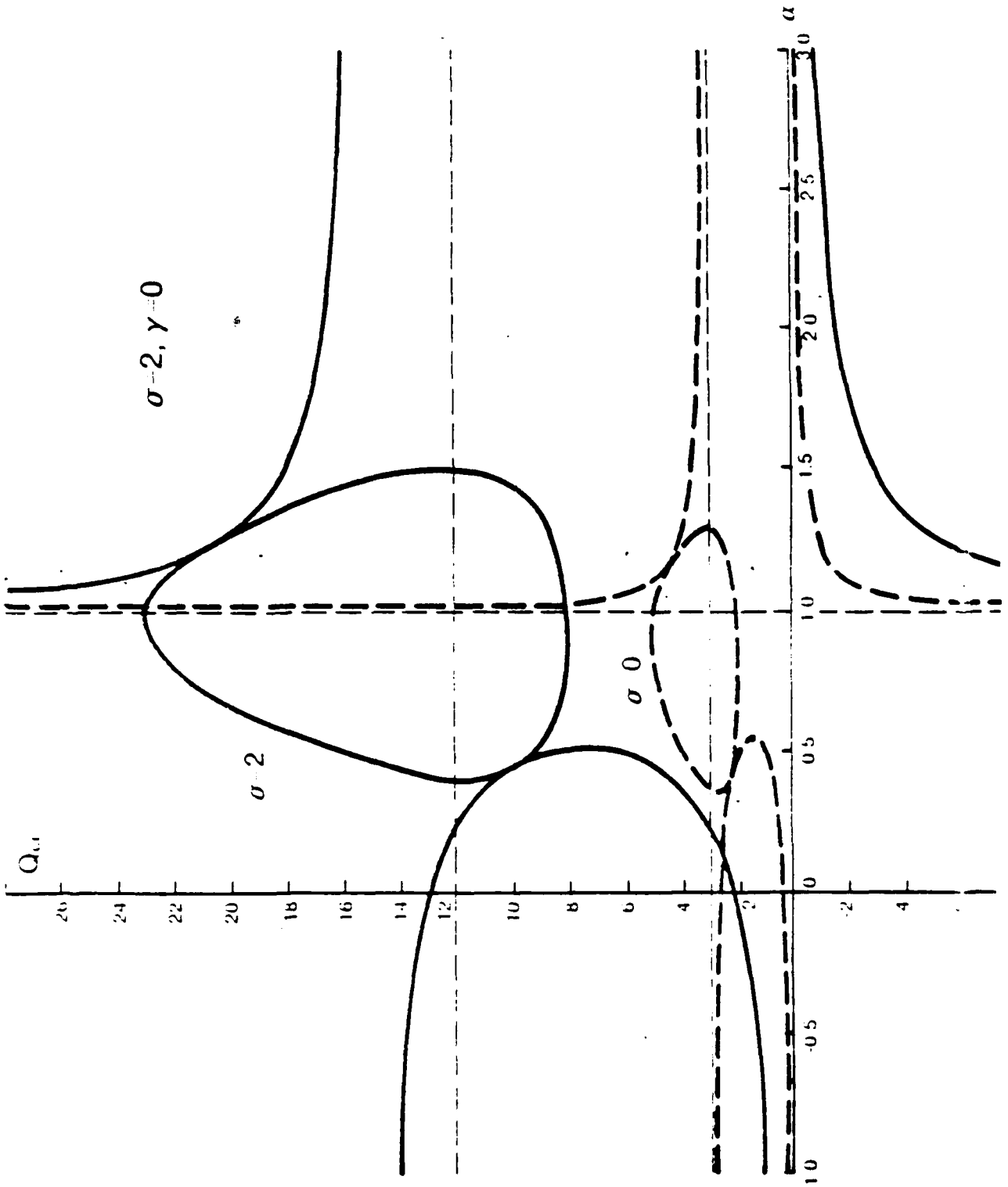


Figure 4. Stability diagrams for $\sigma = 2$ — and $\sigma = 0$ ---, with $\gamma = 0$.

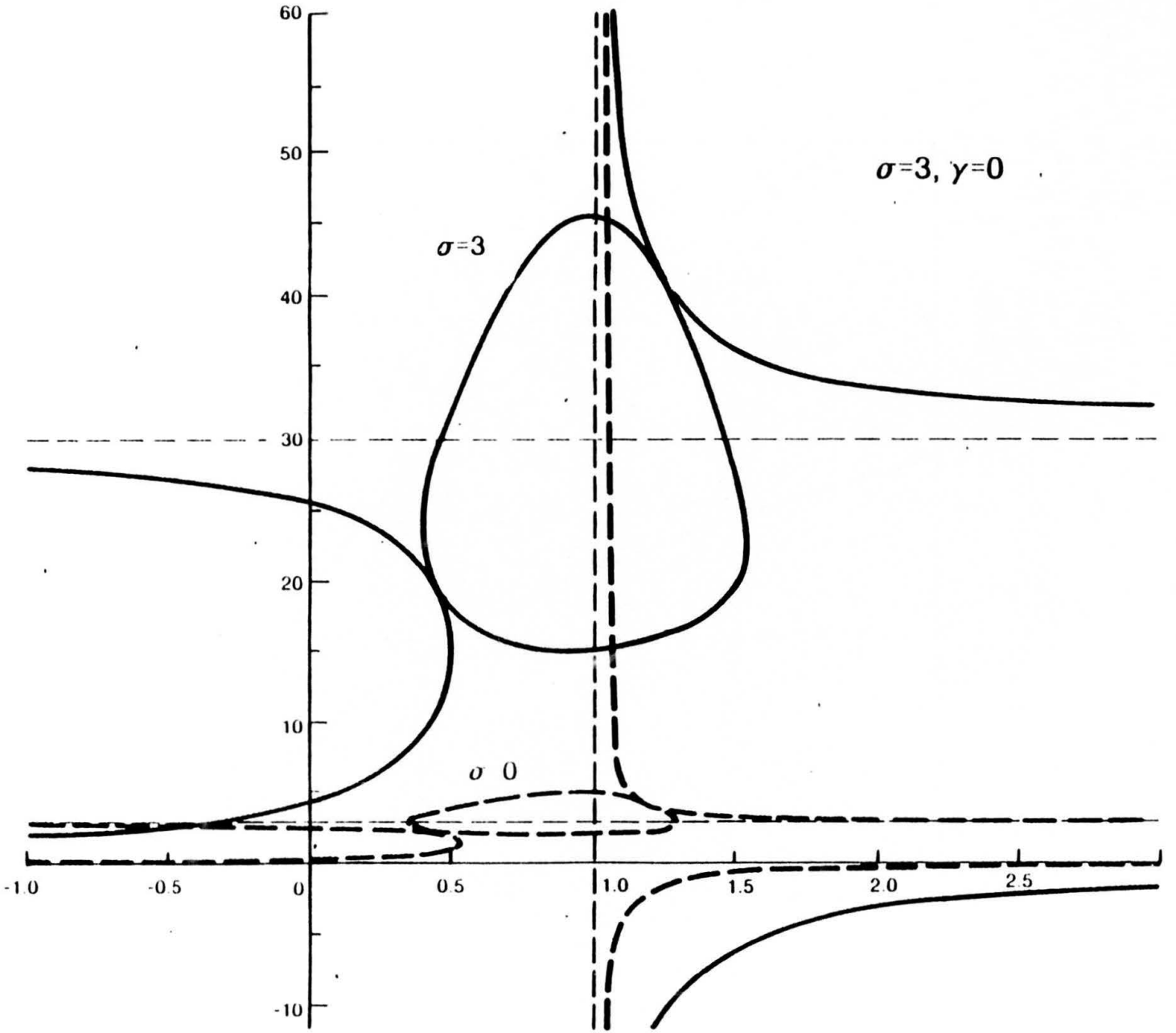


Figure 5. Stability diagrams for $\sigma = 3$ —— and $\sigma = 0$ ---, with $\gamma = 0$.

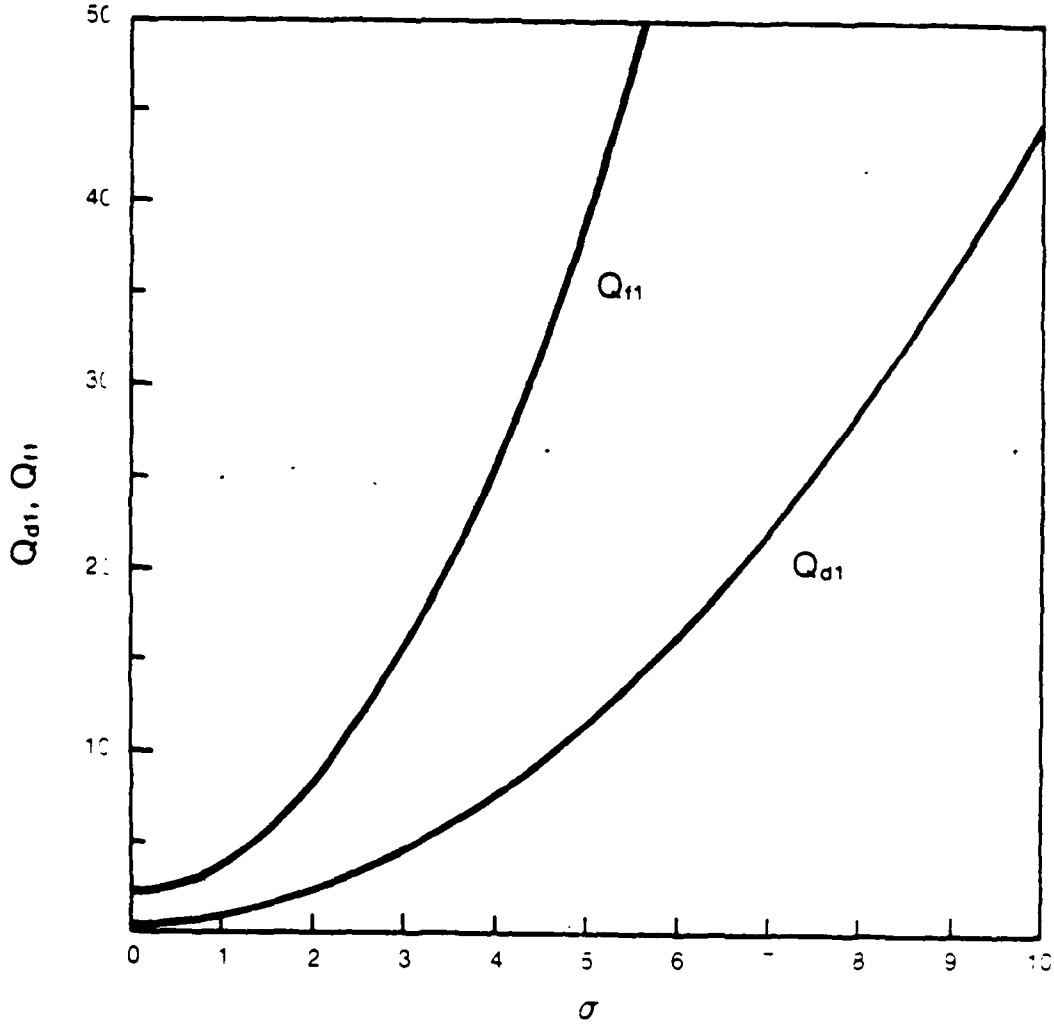


Figure 6. Variations of Q_{f1} for $\alpha = 1$ and Q_{d1} for $\alpha = 0$ versus σ with $\gamma = 0$.

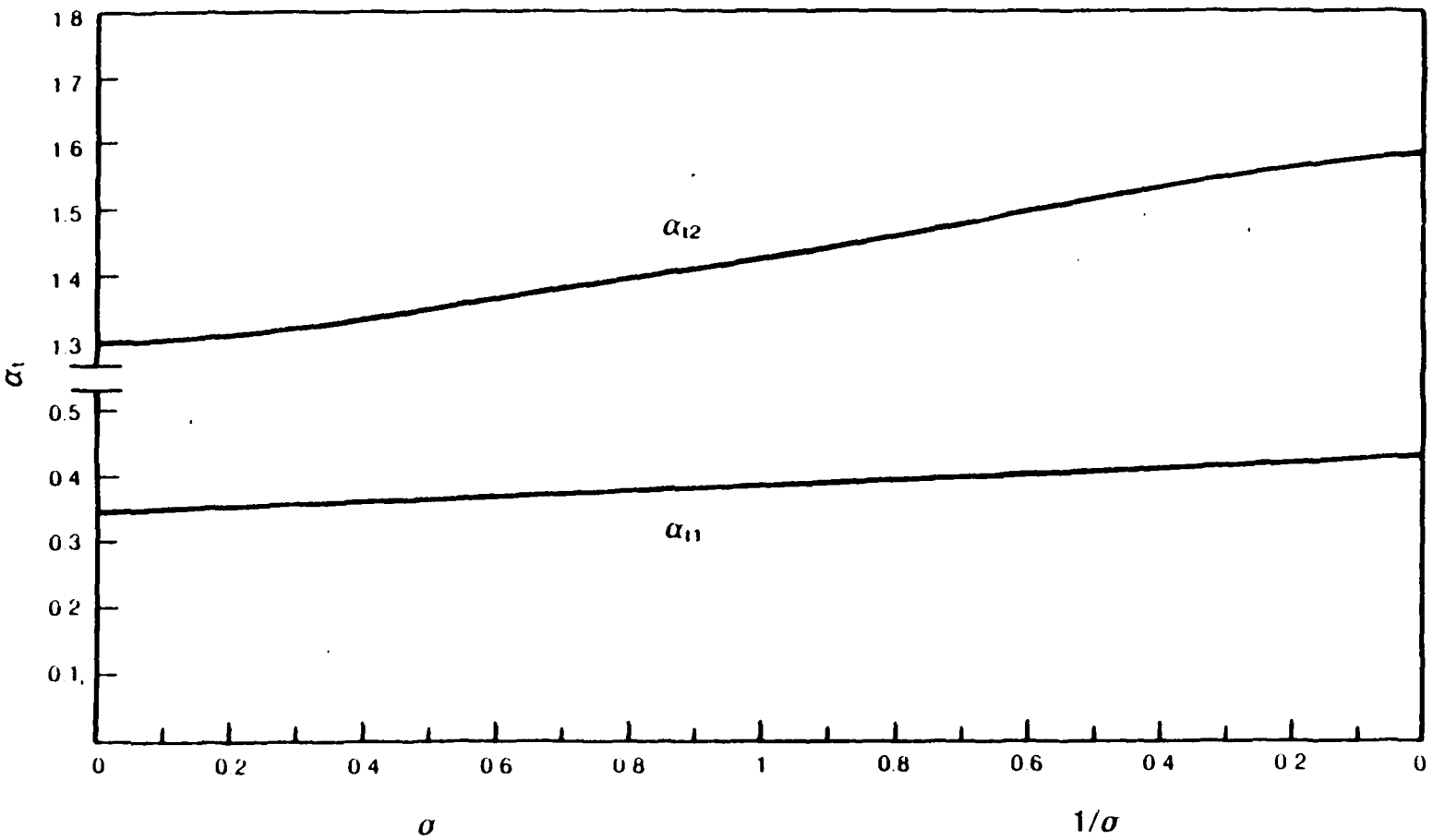


Figure 7. Variations of α_{t1} and α_{t2} versus the support motion parameter σ for $\gamma = 0$.

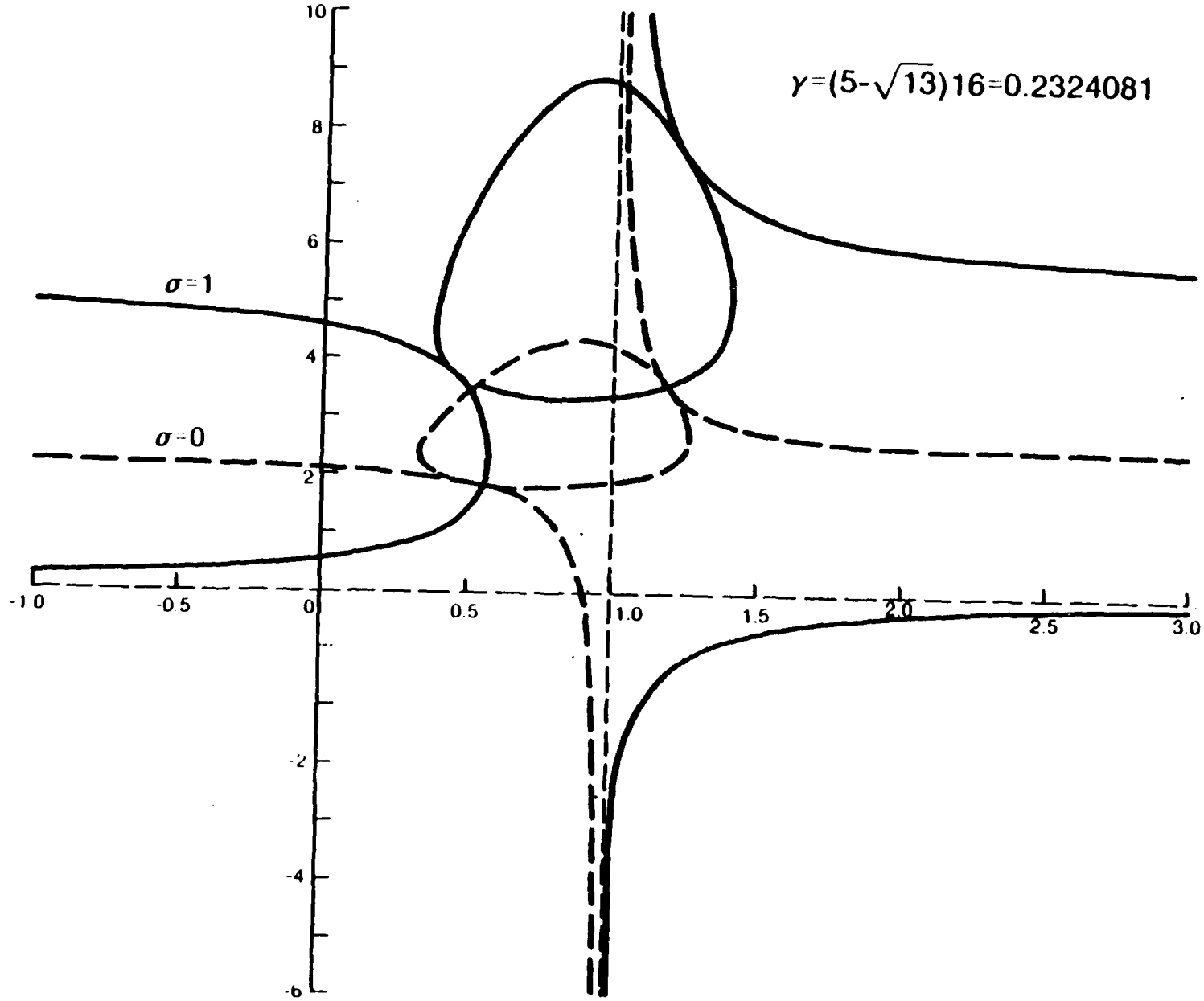


Figure 8. Stability diagrams for $\sigma = 1$ —— and $\sigma = 0$ --- with $\gamma = (5 - \sqrt{13})/6$.

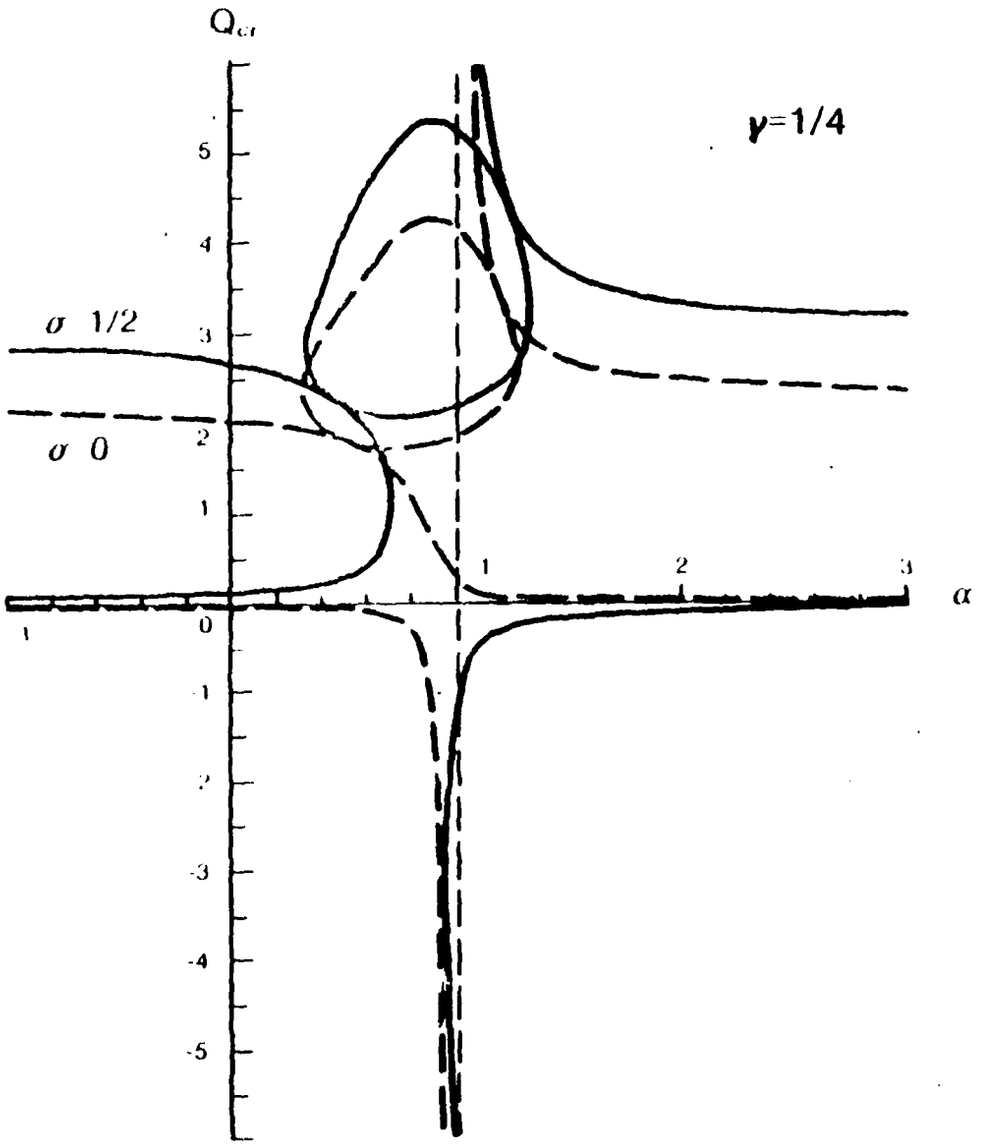


Figure 9. Stability diagram for σ = 1/2 —— and σ = 0 -- with γ = 1/4.

FIFTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING
FINAL ATTENDANCE LIST

NAME/ADDRESS

ANDERSEN, GERALD R., Math Sci Div, USARO, Research Triangle Park, NC
ARNEY, DAVID C., Dept of Math, USMA, West Point, NY
ASHRAFIUON, HASHEM, Dept of Aerospace Mechanics, SUNY, Buffalo, NY
BAHETI, RADHAKISAN S., GE R&D Ctr, Schenectady, NY
BARSOUM, ROSHDY S., USA Materials Tech Lab, Watertown, MA
BENTON, THOMAS, USA Materials Tech Lab, Watertown, MA
BERNARD, ROBERT S., Hydr. Lab, Engr Waterways Expmnt Stn, Vicksburg, MS
BISWAS, RUPAK, Dept of Computer Sciences, RPI, Troy, NY
BURNS, TIMOTHY J., Nat'l Bureau of Stds, Sci Comp Div, Gaithersburg, MD
CAMERON, DAVID H., Dept of Mathematics, USMA, West Point, NY
CHAKRAVARTHY, SUKUMAR, Rockwell Int'l Science Ctr, Thousand Oaks, CA
CHANDRA, JAGDISH, Math Sciences Div, US ARO, Research Triangle Park, NC
CHEN, PETER C.T., Benet Wpns Lab, Watervliet, NY
CHU, SHIH C., ARDEC, Picatinny Arsenal, NJ
CHUI, CHARLES K., Dept of Math, Texas A&M Univ, College Station, TX
COOPER, GENE R., Ballistic Rsch Lab, Aberdeen Proving Grounds, MD
DANBERS, JAMES E., Ballistic Rsch Lab, Aberdeen Proving Grounds, MD
DIAMOND, HARVEY, Dept of Math, WVa Univ, Morgantown, WV
DONOVAN, WILLIAM F., Ballistic Rsch Lab, Aberdeen Proving Ground, MD
DREW, DONALD, Dept of Math Sci, RPI, Troy, NY
DVORAK, GEORGE J., Dept of Civil Engineering, RPI, Troy, NY
EGERLAND, WALTER O., USA Ballistic Rsch Lab, Aberdeen Proving Ground, MD
FALLS, T. C., USA Engineer Waterways Experiment Station, Vicksburg, MS
FARCUKI, RIDA T., IBM Thomas J. Watson Rsch Ctr, Yorktown Heights, NY
FLAHERTY, JOSEPH E., Dept of Computer Sciences, RPI, Troy, NY
FRIED, ISAAC, Dept of Math, Boston Univ, Boston, MA
GLIMM, JAMES, Courant Institute, NYU, New York, NY
GLOWINSKI, ROLAND, Dept of Math, Univ of Houston, Houston, TX
GROVE, JOHN W., Dept of Math, Courant Inst, NYU, New York, NY
GURTIN, MORTON E., Dept of Math, Carnegie-Mellon Univ, Pittsburgh, PA
HALL, CHARLES E. JR, R,D&Engr Ctr, USA Missile Cmd, Redstone Arsenal, AL
HEATON, KENNETH, Defence Rsch Establishment, Valcartier, Quebec, Canada
HOELTZEL, DAVID A., Dept of Mech Eng, Columbia Univ, New York, NY
HOLLIG, KLAUS, Comp Sci Dept, Univ of Wisconsin, Madison, WI
HUSSAIN, MAQYYED A., GE Corporate R&D Ctr, Schenectady, NY
JACKSON, WILLIAM D., USA TACOM, AMSTA-RSA, Warren, MI
JAMESON, ANTONY, Dept of Math, Princeton Univ, Princeton, NJ
JOHNSON, ARTHUR R., USA Materials Tech Lab, Watertown, MA
JOHNSON, JOHN L., USA Missile Cmd, Redstone Arsenal, AL
JOHNSON, MARK, Benet Wpns Lab, Watervliet, NY
KASCAK, ALBERT F., Propulsion Directorate, Lewis Rsch Ctr, Cleveland, OH
KAYS, JAMES L., Dept of Math, USMA, West Point, NY
KIM, DONG S., USA Concepts Analysis Agency, Bethesda, MD
KRAHN, GARY W., Dept of Math, USMA, West Point, NY
LAGOUDAS, DIMITRIS C., Dept of T & A Mech, Cornell Univ, Ithaca, NY
LEHNIGK, SIEGFRIED H., USA Missile Cmd, Redstone Arsenal, AL
LEN, JON, Dept of Math, Cornell Univ, Ithaca, NY
LIN, JERRY, Dept of Civil Engr, RPI, Troy, NY
LIPTON, ROBERT, Math Sci Inst, Cornell Univ, Ithaca, NY
LUDWIG, RAY, Dept of Comp Sci, RPI, Troy, NY
MCKINLEY, GEORGE B., Engr Waterways Expt Stn, Geotech Lab, Vicksburg, MS

MCMAHON, JOHN R., Dept of Math, USMA, West Point, NY
MEDLER, LEON, USA Belvoir R & D Ctr, Ft Belvoir, VA
MEYER, HUBERT W. JR., DA Ballistic Rsch Lab, Aberdeen Proving Ground, MD
MILLER, MILES C., Chemical R&D Center, Aberdeen Proving Ground, MD
MITTER, SANJOY K., Ctr for Intelligent Control Sys, MIT, Cambridge, MA
MOORE, PETER K., Dept of Computer Science, RPI, Troy, NY
MUMFORD, DAVID B., Dept of Math, Harvard Univ, Cambridge, MA
PERRONE, PAUL J., Materials Technology Lab, Watertown, MA
PRABHU, NARAHARI U., OpnsRsch & IndustEngr, Cornell Univ, Ithaca, NY
RAMNATH, R. V., Sparta Systems, Inc., Lexington, MA
RAPHAEL, LOUISE A., Dept of Math, NSF, SEE/DTPE , Washington, DC
REDDY, J. N., Dept of ESM, Virginia Tech, Blacksburg, VA
REINIG, KARL D., DA, Harry Diamond Lab, Adelphi, MD
SCANLON, RAY D., Benet Wpns Lab, Watervliet, NY
SCHEFFNER, NORMAN W., USA Engr Waterways Expmnt Stn, Vicksburg, MS
SEDNEY, RAYMOND, Ballistic Rsch Lab, Aberdeen Proving Ground, MD
SEGUR, HARVEY, ARAP Div of CRT, Inc., Princeton, NJ
SHEPHARD, MARK S., Ctr for Interactive Computer Graphics, RPI, Troy, NY
SOANES, ROYCE, Benet Wpns Lab, Watervliet, NY
SPIRIDIGLIOZZI, LOU, USA Materials Tech Lab, Watertown, MA
SRIVASTAV, RAM P., Dept of Appl Math & Stats, SUNY, Stony Brook, NY
STRANG, GILBERT, Dept of Math, MIT, Cambridge, MA
TADJBAKHSH, IRADJ G., Dept of Engr, US ARO, Research Triangle Park, NC
TAKAGI, SHUNSUKE, DA Cold Regions Rsch Lab, Hanover, NH
TAM, RICHARD Y., Dept of Math, Purdue Univ, Indianapolis, IN
TAYLOR, ROBERT L., Dept of Civil Engr, Univ of Cal, Berkeley, CA
TEPLY, JAN L., Product Eng. Div, Alcoa Center, PA
TRACEY, DENNIS M., USA Materials Technology Lab, Watertown, MA
TURNER, JOSHUA U., IBM, Poughkeepsie, NY
VAN LOAN, CHARLES F., Dept of CompSci, Cornell Univ, Ithaca, NY
VASILAKIS, JOHN D., Benet Weapons Lab, Watervliet, NY
WANG, YUN, Dept of Math, RPI, Troy, NY
WEHAGE, ROGER A., USA Tank-Automotive Cmd, Warren, MI
WEILER, KEVIN., GE Co. Corporate R&D, Schenectady, NY
WEISS, RICHARD A., Waterways Expmnt Stn, Vicksburg, MS
WHITEMAN, JOHN, Inst of Comput Math, Brunel Univ, Uxbridge, England
WOLFF, STEPHEN S., Nat'l Sci Foundation, Washington, DC
WOUK, ARTHUR, US ARO, Research Triangle Park, NC
WU, JULIAN J., USARDevelopmentStandardizationGp-UK, England
ZURLINDEN, JOSEPH E., STEWS-DE-OD, White Sands Missile Range, NM

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 88-1	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION: Army Research Office; 6b OFFICE SYMBOL (if applicable): SLCRO-ARO	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS (City, State, and ZIP Code): Research Triangle Park, NC 27709	7b ADDRESS (City, State, and ZIP Code)
8a NAME OF FUNDING/SPONSORING ORGANIZATION; 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO; PROJECT NO; TASK NO; WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)

Transactions of the Fifth Army Conference on Applied Mathematics and Computing (U)

12 PERSONAL AUTHOR(S)

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final Technical Report	FROM 6/15/87 TO 6/18/87	1988 March	846

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Fluid and solid mechanics, shock waves, control theroy, statistical analysis, reaction-diffusion, bifurcation, approximation.

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

(U) This is a technical report resulting from the Fifth Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat various Army applied mathematical problems.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION
22a NAME OF RESPONSIBLE INDIVIDUAL: Dr. Francis G. Dressel	22b TELEPHONE (Include Area Code): (919) 549-0641; 22c OFFICE SYMBOL: SLCRO-MA

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.