ARO Report 87-1

TRANSACTIONS OF THE FOURTH ARMY

CONFERENCE ON APPLIED MATHEMATICS

AND COMPUTING

AD-A183 544

DTIC
S ELECTE D
AUG 2 1 1987

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

87   8 19 033

BLANK PAGES
IN THIS
DOCUMENT
WERE NOT
FILMED

U. S. ARMY RESEARCH OFFICE

Report No. 87-1

February 1987

TRANSACTIONS OF THE FOURTH ARMY CONFERENCE

ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Host

Cornell University

Ithaca, New York

27-31 May 1986

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

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

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

FOREWORD

The Fourth Army Conference on Applied Mathematics and Computing was held 27-30 May 1986 at Cornell University, Ithaca, New York. It coincided with the formal opening of the recently established Mathematical Sciences Institute (MSI). This meeting's seven invited speakers addressed the vital areas of combustion, computational fluid dynamics, parallel computation, stochastic analysis, multiple bifurcation, numerical solutions of partial differential equations and problems in many scales of length and time in modern computing enviroments. There were two special sessions that dealt with Stochastic Algorithms and Computational Vision, and Probabilistic Methods in Solid Mechanics. The one hundred and eight contributed technical papers covered nearly the entire spectrum of basic research. During the course of the meeting several synergetic relationships developed, and the feedback from the Army scientists was very positive.

As in previous meetings, this meeting provided its attendees a chance to see the many scientific developments taking place in various Army laboratories. Through these meetings, techniques developed at one installation are brought to the attention of scientists at other places, thus reducing duplication of effort. Another important phase of these meetings is presenting the members of the audience an opportunity to hear nationally known scientists discuss recent developments of their own fields. This year the invited speakers together with the titles of their addresses are listed below. These gentlemen were more than willing to discuss various problems of special interest to scientists in the Army agencies.

SPEAKERS AND AFFILATION	TITLES OF ADDRESS
Professor K. G. Wilson Cornell University	Renormalization Groups and Problems in Many Scales of Length
Professor Richard Ewing University of Wyoming	Numerical Solution of Partial Differential Equations
Professor John Guckenheimer Cornell University	Multiple Bifurcation
Professor Eugene Wong University of California, Berkeley	Stochastic Differencial Forms
Professor Richard Karp University of California, Berkeley	The Complexity of Parallel Computation
Professor A. F. Ghoniem Massachusetts Institute of Technology	Computing Unsteady Reacting Flows Using Vortex Methods
Professor A. J. Majda Princeton University	High Mach Number Combustion

The benefits derived from these conferences depend a great deal on the host's Chairman on Local Arrangements. The attendees at this meeting were fortunate to have Professor G. S. S. Ludford, Director of the MSI, serving in this capacity. He, together with members of his capable staff, provided all those things, such as projection equipment, travel information, etc., needed for an enjoyable and profitable symposium.

The Amy Mathematics Steering Committee is the sponsor of these Army Conferences on Applied Mathematics and Computing. The members of this committee were pleased, not only with the large number of contributed papers, but also with the scientific quality of these papers. They are also pleased to be able to provide the transaction of this conference. It is hoped the scientific ideas contained therein will benefit not only those who were able to attend the symposium, but also many others that did not enjoy that privilege.

TABLE OF CONTENTS

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

Title	Page

x

Title	Page

FOURTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Cornell University, Ithaca, New York

May 27 - 30, 1986

AGENDA

Tuesday May 27, 1986

08:00 - 16:00   Registration - Warren Hall (WH)

08:30 - 09:00   Opening Remarks - WH 131

09:00 - 10:00   General Session I - WH 131

Chairman: Dr. San-Li Pu, Benet Weapons Laboratory,
            Watervliet Arsenal, Watervliet, New York

Renormalization Groups and problems in many scales of length

Kenneth G. Wilson, Cornell University, Ithaca, New York

10:00 - 10:30   Break

10:30 - 12:30   Technical Session TUM1 - Shock Waves - WH 1-145

Chairperson: Dr. Gary Carrafano, Benet Weapons Laboratory
               Watervliet Arsenal, Watervliet, New York

10:30 - 10:50   Interactions, Bifurcations and instabilities of hydrodynam-
                 ic surfaces

C. Fithian, J. Glimm, J. Grove, C. Gardner, O. McBryan,
Courant Institute of Mathematical Sciences, New York, New
York

10:50 - 11:10   Developments of shock fronts in viscoelastic materials

John A. Nohel, Mathematics Research Center, University of
Wisconsin, Madison, Wisconsin

11:10 - 11:30   Recent Developments in Nonstrictly Hyperbolic Conservation
                 Laws

Michael Shearer, North Carolina State University, Raleigh,
North Carolina and David G. Shaeffer, Duke University,
Durham, North Carolina

11:30 - 11:50   Applications of Matrix Factorization in Hydrodynamic Stabil-
                 ity

P. J. Morris, Pennsylvania State University, University Park, Pennsylvania

11:50 - 12:10   A numerical study of the effect of curvature on detonation speed

B. Bukiet and J. Jones, Courant Institute of Mathematical Sciences, New York, New York

12:10 - 12:30   Pressure transients in a cavity due to impulsive loads

C. Helleur, Valcartier, Quebec; B. Tabarrok and R. G. Fenton University of Toronto, Toronto, Ontario, Canada

12:30 - 14:00   LUNCH

10:30 - 12:30   Technical Session TUM2 - Solid Mechanics A - WH 1-101

Chairperson: Dr. S. C. (Mike) Chu, ARDC, Dover, New Jersey

10:30 - 10:50   Singularities in three-dimensional potential theory and solid mechanics and finite element methods for their treatment

J. R. Whiteman, Brunel University, Uxbridge, U.K.

10:50 - 11:10   Computation of weight functions

T. -L. Sham, Rensselaer Polytechnic Institute, Troy, New York

11:10 - 11:30   Micromechanics of Shearbanding in High Strength Steel

Dennis M. Tracey, Colin E. Freese and Paul J. Perron Army Materials Technology Laboratory, Watertown, Massachusetts

11:30 - 11:50   Fractals, fragmentation and failure

D. L. Turcotte, Cornell University, Ithaca, New York

11:50 - 12:10   Crack tip fields for anisotropic material: A finite element approach

Roshdy Barsoum, Army Materials Technology Laboratory Watertown, Massachusetts

12:10 - 12:30   Numerical Solution of Random Volterra Integral Equations

Nagash Medin and M. Sambandham, Atlanta University, Atlanta, Georgia

12:30 - 14:00   LUNCH

10:30 - 12:30	Technical Session TUM3 - Stochastic Analysis - WH 1-201
	Chairperson: Dr. Vincent Mirelli, Night Vision and Electro-Optics Laboratory, Fort Belvoir, Virginia
10:30 - 10:50	Applications of non-Gaussian Processes
	Mircea Grigoriu, Cornell University, Ithaca, New York
10:50 - 11:10	The method of random characteristics
	Marc A. Berger and Alan D. Sloan, Georgia Institute of Technology, Atlanta, Georgia
11:10 - 11:30	Characteristic Functions of a Class of Probability Distributions
	Siegfried Lehnigk, Army Missile Command, Redstone Arsenal, Alabama
11:30 - 11:50	Poission and extreme value limit theorems for Markov Random Fields
	Simeon M. Berman, Courant Institute of Mathematical Sciences, New York, New York
11:50 - 12:10	A bound on the variation between two probability measures in terms of the intensities of a discrete point process relative to these probabilities
	G. R. Andersen, Ballistics Research Laboratory Aberdeen Proving Ground, Maryland
12:10 - 12:30	Estimating the compensator from Poisson type counting processes
	Michael Phelan, Cornell University, Ithaca, New York
12:30 - 14:00	Lunch
14:00 - 17:15	Special Session I - Stochastic Algorithms, Image Processing, and Computational Vision - WH 1-145
	Chairperson: Professor Sanjoy K. Mitter, Massachusetts Institute of Technology, Cambridge, Massachusetts
14:00 - 14:15	Opening Remarks
14:15 - 15:00	Global optimization via the cooling algorithm and computational complexity

B. Gidas, Brown University, Providence, Rhode Island

15:00 - 15:45 On the complexity of some stochastic search algorithms

 B. Hajek, University of Illinois, Urbana, Illinois and Massachusetts Institute of Technology, Cambridge, Massachusetts

15:45 - 16:15 Break

16.15 - 16:45 Maximum Entropy methods for expert system construction

 A. Lippman, Brown University, Providence, Rhode Island

16:45 - 17:15 Parameter estimation in probabilistic models of images

 J. Marroquin, PEM, Mexico City, Mexico

14:00 - 17:15 Special Session II - Probabilistic Methods in Solid
             Mechanics - WH 1-101

 Chairperson: Professor N. U. Prabhu, Cornell University
           Ithaca, New York

14:00 - 14:15 Opening Remarks

14:15 - 14:45 Probabilistic finite elements and potential applications to fracture

 Wing-Kam Liu and Ted Belytschko, Northwestern University, Evanston, Illinois

14:45 - 15:15 On computing stress intensity factors with uncertainty

 Ram Srivastav, Army Research Office, Durham, North Carolina

15:15 - 15:45 Limit theorems for the size effect in the lifetime distribution of a Fibrous Composite

 S. L. Phoenix, Cornell University, Ithaca, New York

15:45 - 16:15 Break

16:15 - 16:45 Approximate Methods for Structural Reliability

 Mircea Grigoriu and Arnold Buss, Cornell University Ithaca, New York

16:45 - 17:15 Critical strains for adiabatic shear

 Gerald Moss, Ballistics Research Laboratory, Aberdeen Proving Ground, Maryland

19:00 - 22:00　　Banquet (Prepaid participants only)
　　　　　　　　Speaker: Dr. Herbert Hauptman. President and Research
　　　　　　　　　　　　Director, Medical Foundation of Buffalo

Wednesday May 28, 1986

08:00 - 16:00　　Registration - Warren Hall (WH)

08:30 - 10:30　　Technical Session WM1 - Electromagnetics - WH 1-201

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

08:30 - 08:50　　Numerical computation of path integral representations of
　　　　　　　　scalar wave field propagators

　　　　　　　　Louis Fishman, Catholic University, Washington, D. C.

08:50 - 09:10　　Scale invariant equations for Relativistic Waves

　　　　　　　　Richard Weiss, Army Corps of Engineers, Waterways Experi-
　　　　　　　　mental Station, Vicksburg, Mississippi

09:10 - 09:30　　Relativistic wave equations for Real Gases

　　　　　　　　Richard Weiss, Army Corps of Engineers, Waterways Experi-
　　　　　　　　mental Station, Vicksburg, Mississipi

09:30 - 09:50　　Hamiltonian perturbations of the nonlinear Schroedinger
　　　　　　　　Equation

　　　　　　　　Curtis R. Menyuk, University of Maryland, College Park,
　　　　　　　　Maryland

09:50 - 10:10　　Solutions of a non-integrable Hamiltonian system

　　　　　　　　P. K. A. Wai, C. R. Menyuk, H. H. Chen, and Y. C. Lee
　　　　　　　　University of Maryland, College Park, Maryland

10:10 - 10:30　　The effects of boundary conditions on Electromagnetic
　　　　　　　　pulses

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

10:30 - 11:00　　Break

08:30 - 10:30　　Technical Session WM2 - Solid Mechanics B - WH 1-101

　　　　　　　　Chairperson: Dr. T. W. Wright, Ballistics Research
　　　　　　　　　　　　　　Laboratory, Aberdeen Proving Ground, Maryland

08:30 - 08:50	Discovery of the elastic parameters of a layered half-space
	Paul Sacks, Iowa State University, Ames, Iowa
08:50 - 09:10	Stability of free-free columns
	Julian J. Wu, Army European Research Office, London, UK
	John D. Vasilakis, ARDC Benet Weapons Laboratory Watervliet Arsenal, Watervliet, New York
09:10 - 09:30	On fatigue life prediction in thick-walled cylinders
	S. L. Pu and P. C. T. Chen, ARDC Benet Weapons Laboratory Watervliet Arsenal, Watervliet, New York
09:30 - 09:50	Analysis of composite shrink fits-Tresca Material
	Peter C. T. Chen, ARDC Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York
09:50 - 10:10	A shallowly curved shear-deformable beam element
	A. Tessler and L. Spiridigliozzi, Army Materials Technology Laboratory, Watertown, Massachusetts
10:10 - 10:30	Admissible elastic energy density functions for rubber-like solids
	I. Fried, Boston University; A. R. Johnson and C. J. Quigley, Army Materials Technology Laboratory Watertown, Massachusetts
10:30 - 11:00	Break
08:30 - 10:30	Technical Session - WM3 - Transonic Flow - WH 1-145
	Chairperson: Dr. John Polk, Ballistics Research Laboratory, Aberdeen Proving Ground Maryland
08:30 - 08:50	Solutions of the Transonic Flow Equations by Spectral Methods
	P. Hanley, C. Mavriplis, Massachusetts Institute of Technology, Cambridge, Massachusetts and W. L. Harris University of Connecticut, Storrs, Connecticut
08:50 - 09:10	A toolkit of symbol manipulation programs for variational grid generation
	Stanly Steinberg, University of New Mexico, Albuquerque and Patrick J. Roache, Ecodynamics Research Associates, Albuquerque, New Mexico

09:10 - 09:30	A self-adaptive gridding for inviscid transonic projectile aerodynamic computation
	Chen-chi Hsu and Chyuan-Gen Tu, University of Florida Gainesville, Florida
09:30 - 09:50	On computation of transonic projectile aerodynamics
	Chen-chi Hsu and Nae-Hauer Shiau, University of Florida Gainesville, Florida
09:50 - 10:10	Numerical simulation of supersonic flow over a rotating band
	J. Sahu, Ballistics Research Laboratory, Launch and Flight Division, Aberdeen Proving Ground, Maryland
10:10 - 10:30	Improved numerical prediction of transonic flow
	J. Sahu and C. J. Nietubicz, Ballistics Research Laboratory, Launch and Flight Division, Aberdeen Proving Ground, Maryland
10:30 - 11:00	Break
11:00 - 12:00	General Session II - WH 131
	Chairperson: Dr. Billy Jenkins, Army Missile Command, Redstone Arsenal, Alabama
	Numerical Solution of Partial Differential Equations
	Richard Ewing, University of Wyoming, Laramie, Wyoming
12:00 - 13:30	Lunch
13:30 - 15:50	Technical Session WA1 - Nonlinear Analysis and Control A - WH 1-145
	Chairperson: Dr. William Jackson, Army Tank Automotive Command, Warren, Michigan
13:30 - 13:50	Some stability results for advection- diffusion equations
	F. A. Howes, Lawrence Livermore Laboratory, Livermore, California
13:50 - 14:10	Spatial structure of time-periodic solutions of the Ginzburg-Landau equation
	Philip Holmes, Cornell University, Ithaca, New York
14:10 - 14:30	Dissipation in conservative systems

Mark Levi, Boston University, Boston, Massachusetts

14:30 - 14:50   An analysis of the Duffing's equation through examination of initial condition maps and Liapunov exponents

Charles Pezeshki and Earl Dowell, Duke University, Durham, North Carolina

14:50 - 15:10   Poincare maps of a Journal bearing

P. J. Hollis and D. L. Taylor, Cornell University, Ithaca, New York

15:10 - 15:30   The propagation of information and uncertainty in dynamical systems

David F. Delchamps, Cornell University, Ithaca, New York

15:30 - 15:50   Extension of Sarkovskii's theorem

Walter Egerland, Ballistics Research Laboratory, Aberdeen Proving Ground, Maryland

15:50 - 16:15   Break

13:30 - 15:50   Technical Session - WA2 - Fluid Mechanics - WH 1-101

Chairperson: Dr. Miles Miller, Chemical Research and Development Center, Edgewood Arsenal, Maryland

13:30 - 13:50   Fluid Motion in Liquid-filled shells

T. Herbert, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

13:50 - 14:10   The evolution of subharmonic edge wavepackets on a sloping beach

T. R. Akylas and S. Knopping, Massachusetts Institute of Technology, Cambridge, Massachusetts

14:10 - 14:30   Theoretical and Simulation studies of surfactants at liquid interfaces

J. H. Thurtell and K. E. Gubbins, Cornell University Ithaca, New York

14:30 - 14:50   Static capillary bridges: Global stability results for symmetrization methods

Paul H. Steen, Cornell University, Ithaca, New York

15:30 - 15:50	A model of decision-making with sequential information-acquisition with application to the file search problem
	James C. Moore and Andrew B. Whinston, Purdue University, West Lafayette, Indiana
15:50 - 16:15	Break
16:15 - 17:15	General Session III - WH 131
	Chairman: Dr. Gary Anderson, Army Research Office Durham, North Carolina
	Multiple Bifurcation
	John Guckenheimer, Cornell University, Ithaca, New York

Thursday May 29, 1986

08:00 - 16:00	Registration - Warren Hal (WH)
08:30 - 10:30	Technical Session - THM1 - Statistics and Data Analysis - WH 1-201
	Chairperson: Major Rickey Kolb, United States Military Academy, West Point, New York
08:30 - 08:50	On a measure of block design efficiency recovering inter-block information
	Walter T. Federer, Cornell University, Ithaca, New York
08:50 - 09:10	Computing asymptotic confidence bands for Nonlinear regression models
	John J. Peterson, Syracuse University, Syracuse, New York
09:10 - 09:30	Applying statistical graphics to multivariate data
	Steven J. Schwagger, Cornell University, Ithaca, New York
09:30 - 09:50	Unimodular dynamics of SF_6 under coherent excitation
	John C. England, Frederic A. Hopf and Charles M. Bowden, U. S. Army Missile Command, Redstone Arsenal, Alabama
09:50 - 10:10	Testing Curve Fit
	Royce Soanes, Benet Weapons Laboratory, Watervliet Arsenal Watervliet, New York

10:10 - 10:30	The estimation of some network parameters in the pert model of activity networks: Review and critique
	Salah E. Elmaghraby, North Carolina State University, Raleigh, North Carolina
10:30 - 11:00	Break
08:30 - 10:30	Technical Session - THM2 - Multiphase Flow - WH 1-101
	Chairperson: Dr. Csaba Zoltani, Ballistics Research Laboratory, Aberdeen Proving Ground, Maryland
08:30 - 08:50	On the two-phase Stefan problem with interfacial energy and entropy
	Morton E. Gurtin, Carnegie-Mellon University, Pittsburgh, Pennsylvania
08:50 - 09:10	Stefan's Problem in a Finite Domain with constant boundary and initial conditions
	Shunsuke Takagi, Cold Region Research and Engineering Laboratory, Hanover, New Hampshire
09:10 - 09:30	Thin film conductive coating for surface heating and decontamination
	S. S. Sadhal, University of Southern California, Los Angeles, California, P. S. Ayyaswamy, University of Pennsylvania, Philadelphia, Pennsylvania and Arthur K. Stuempfle, Chemical Research and Development Center, Edgewood Arsenal, Maryland
09:30 - 09:50	The Poiseuille Flow of a particle-fluid mixture effective viscosity
	Donald A. Drew, Rensselaer Polytechnic Institute, Troy, New York
09:30 - 10:10	Fluids in Narrow Pores: Computer Simulation and Mean Field Theory
	B. K. Peterson and K. E. Gubbins, Cornell University and J. P. R. B. Walton, B. P. Research Centre, United Kingdom
10:10 - 10:30	Macroscopic and microscopic modelling of mushy regions
	S. D. Howison, Oxford University, United Kingdom
10:30 - 11:00	Break

08:30 - 10:30	Technical Session THM3 - Optimization - WAH 1-145

Chairperson: Dr. Ervin Atzinger, Army Materiel Systems
              Analysis Activity, Aberdeen Proving Ground,
              Maryland

08:30 - 08:50   Global Optimization using Automatic Differentiation and
                 Interval Iteration

                 Louis B. Rall, Mathematics Research Center, University of
                 Wisconsin, Madison, Wisconsin

08:50 - 09:10   Computing disjoint products efficiently

                 Michael O. Ball, University of Maryland, College Park,
                 Maryland and J. Scott Provan, University of North Carolina,
                 Chapel Hill, North Carolina

09:10 - 09:30   Classes of Greedy Matching Heuristics

                 M. D. Grigoriadis and B. Kalantari, Rutgers University, New
                 Brunswick, New Jersey

09:30 - 09:50   Optimization of stochastic systems via Monte Carlo simula-
                 tion

                 Peter W. Glynn, Mathematics Research Center, University of
                 Wisconsin, Madison, Wisconsin

09:50 - 10:10   Efficient sequencing of a four-circle diffractometer

                 Robert G. Bland and David F. Shallcross, Cornell Univer-
                 sity, Ithaca, New York

10:10 - 10:30   Optimal procedure for dynamic programs with complex loop
                 structures

                 A. O. Esogbue, Georgia Institute of Technology and C. Y.
                 Lee, Institute of Technology, Taejon, Korea

10:30 - 11:00   Break

11:00 - 12:00   General Session IV - WH 131

                 Chairman: Dr. Stephen Wolff, Ballistics Research Laboratory
                            Aberdeen Proving Ground, Maryland

                 Stochastic Differential Forms

                 Eugene Wong, University of California, Berkeley

12:00 - 13:30   Lunch

13:30 - 15:50	Technical Session - THA1 - Nonlinear Analysis and Control B - WH 1-145
	Chairperson: Dr. Norman Coleman, Armament Research and Development Command, Dover, New Jersey
13:30 - 13:50	Stochastic filtering and control with wide bandwidth observation noise
	Harold J. Kushner, Brown University, Providence, Rhode Island
13:50 - 14:10	Adaptive Kalman Filtering for Instrumentation Radar
	Charles K. Chui, Texas A&M University, College Station, Texas and Robert E. Green, White Sands Missiles Range, New Mexico
14:10 - 14:30	Optimal impulse correction of a random linear operator
	P. L. Chow and J. L. Menaldi, Wayne State University, Detroit, Michigan
14:30 - 14:50	Pulse arrival times for waves in turbulent media
	P. L. Chow and J. L. Menaldi, Wayne State University, Detroit, Michigan
14:50 - 15:10	Efficient Parallel Algorithms for controllability and eigenvalue assignment problems
	B. N. Datta and Karbi Datta, Northern Illinois University DeKalb, Illinois
15:10 - 15:30	The transition from phase-locking to Drift in a system of Two weakly coupled Van der Pol Oscillators
	Tapesh Chakraborti and Richard H. Rand Cornell University, Ithaca, New York
15:30 - 15:50	Design and implementation of a Multivariable control system for Aircraft/Weapon Applications
	Pak T. Yip and David Ngo, SMCAR-FSF-RC, ARDC Dover, New Jersey
15:50 - 16:15	Break
13:30 -15:50	Technical Session THA2 - NUMERICAL PDE - WH 1-101
	Chairperson: Dr. Nisheeth Patel, Ballistics Research Laboratory, Aberdeen Proving Ground, Maryland

13:30 - 13:50	Domain Contractions around three dimensional anamolies in spherical finite difference computations of Poission's Equation
	A. H. Zemanian and T. S. Zemanian, State University of New York at Stony Brook, Stony Brook, New York
13:50 - 14:10	Upwind schemes and numerical solutions to the MHD Riemann problem
	M. Brio, C. C. Wu, S. J. Osher, A. Harten University of California, Los Angeles, California
14:10 - 14:30	Finite-difference methods for polar coordinate systems
	John C. Strikwerda and Yvonne Nagel Mathematics Research Center, University of Wisconsin, Madison, Wisconsin
14:30 - 14:50	Adaptive Finite Element Methods for Parabolic systems in one- and two - space dimensions
	Slimane Adjerid, Rensselaer Polytechnic Institute, Troy and Joseph E. Flaherty, Rensselaer Polytechnic Institute and Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York
14:50 - 15:10	Fast parallel algorithms via domain decomposition for elliptic problems
	J. H. Bramble, Cornell University, Ithaca, New York, J. Pasciac, Brookhaven National Laboratory, Upton, New York and A. H. Schatz, Cornell University, Ithaca, New York
15:10 - 15:30	A posteriori error estimation in a finite element method for parabolic partial differential equation
	J. M. Coyle and J. E. Flaherty, Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York
15:30 - 15:50	An adaptive method with mesh moving and local mesh refinement for time dependent partial differential equations
	David C. Arney, United States Military Academy, West Point, New York and J. C. Flaherty, Rensselaer Polytechnic Institute, Troy, New York
15:50 - 16:15	Break
13:30 - 15:50	Technical Session THA3 - Vortex Flow and Reaction-Diffusion - WH 1-201

Chairperson: Mr. Arthur Stuempfle, Chemical Research and
Development Center, Edgewood Arsenal.
Edgewood, Maryland

13:30 - 13:50   Vortex Fission and Fusion

Karl Gustafson, University of Colorado, Boulder. Colorado

13:50 - 14:10   Incipient singularities in the Navier-Stokes equations

Alain Pumir and E. Siggia, Cornell University, Ithaca, New
York

14:10 - 14:30   Numerical experiments for a convective reaction diffusion
equation

Tsu-Fen Chen, Howard A. Levine and Paul E. Sacks, Iowa
State University, Ames, Iowa

14:30 - 14:50   Finite element approximation of a reaction-diffusion equa-
tion

Satnam S. Khalsa, Iowa State University, Ames, Iowa

14:50 - 15:10   Interactive diagnostics and graphics for 2D vortex dynamics

C. Serin, M. Melander and N. Zabusky, University of
Pittsburgh, Pittsburgh, Pennsylvania

15:10 - 15:30   Vortex dynamics described by high order moment models

M. Melander, University of Pittsburgh, Pittsburgh, PA

15:30 - 15:50   High resolution, minimal storage algorithms for convection
dominated convection-diffusion processes

V. Ervin and W. Layton, Carnegie-Mellon University,
Pittsburgh, Pennsylvania

15:50 - 16:15   Break

16:15 - 17:15   General Session V - WH 131

Chairperson: Dr. Arthur Wouk, U. S. Army Research Office
Durham, North Carolina

The complexity of parallel computation

Richard Karp, University of California, Berkeley,
California

Friday   May 30, 1986

08:00 - 11:00   Registration - Warren Hall (WH)

08:30 - 10:30   Technical Session FM1 - Combustion - WH 1-201

                 Chairperson: Dr. Norman Slagg, Armament Research and
                               Development Command, Dover, New Jersey

08:30 - 08:50   The thermal explosion solution revisited

                 D. R. Kassoy and J. Beberness, University of Colorado,
                 Boulder, Colorado and J. F. Clarke, Cranfield Institute of
                 Technology, Cranfield, England

08:50 - 09:10   Detonation wave initiation by rapid energy deposition at a
                 confining boundary

                 D. R. Kassoy, University of Colorado, Boulder, Colorado, J.
                 F. Clarke, Cranfield Institute of Technology, England and
                 N. Riley, University of East Anglia, Norwich, England

09:10 - 09:30   A plane premixed flame problem with kinetics: existence
                 and stability results

                 Cl. Laine-Schmidt CNRS Ecole Centrale de Lyon, France and
                 Cornell University, Ithaca, New York

09:30 - 09:50   The effect of structure on the stability of detonations

                 J. D. Buckmaster, University of Illinois, Urbana, Illinois
                 and G. S. S. Ludford, Cornell University, Ithaca, New York

09:50 - 10:10   Limiting thermal jumps in temperature controlled exothermal
                 reactions

                 Paul W. Davis, Worcester Polytechnic Institute, Worcester,
                 Massachusetts, and Jagdish Chandra, U. S. Army Research
                 Office, Research Triangle Park, North Carolina

10:10 - 10:30   Plane propagation through a nonuniform mixture

                 A. Kapila and G. Ledder, Rensselaer Polytechnic Institute,
                 Troy, New York

10:30 - 11:00   Break

08:30 - 10:30   Technical Session FM2 - Computer Vision and Image Processing
                 - WH 1-145

                 Chairperson: Dr. Benjamin E. Cummings, Human Engineering
                                Laboratory, Aberdeen Proving Ground,
                                Maryland

08:30 - 08:50   2- Dimensional Chandrasekhar Filter for Image Restoration

A. K. Mahalanabis, Pennsylvania State University, University Park, Pennsylvania

08:50 - 09:10   A massively parallel architechture for a self-organizing neutral pattern recognition machine

Gail A. Carpenter and Stephen Grossberg, Boston University, Boston, Massachusetts

09:10 - 09:30   Cortical dynamics of three-dimensional form, color, and brightness perception, a predictive synthesis

Stephen Grossberg, Boston University, Boston, Massachusetts

09:30 - 09:50   Object tracking using sensor fusion

Firooz Sadjadi and Mike Bazakos, Honeywell Systems and Research Center, Minneapolis, Minnesota

09:50 - 10:10   Random field identification from samples

M. Rosenblatt-Roth, University of Maryland, College Park, Maryland

10:10 - 10:30   Approximation of two-dimensional fields with Markov meshes

M. Rosenblatt-Roth, University of Maryland, College Park, Maryland

10:30 - 11:00   Break

08:30 - 10:30   Technical Session FM3 - Approximation and Computational Complexity - WH 1-10!

Chairperson: Dr. Raymond Scanlon, Benet Weapons Laboratory Watervliet Arsenal, Watervliet, New York

08:30 - 08:50   Interpolation by bivariate quadratic splines

C. K. Chui, Texas A. and M. University, College Station, Texas, H. Diamond, West Virginia University, Morgantown, West Virginia, and Louise A. Raphael, Howard University, Washington, District of Columbia

08:50 - 09:10   On the C2 Continuity of piecewise cubic Hermite Polynomials with unequal intervals

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

09:10 - 09:30	Views on the Weierstrass and the generalized Weierstrass functions

M. S. Schlesinger, Office of Naval Research, Arlington,
Virginia, M. A. Hussain, and John Bandler, General Electric
Research and Development Center, Schenectady, New York

09:30 - 09:50	Analytics of period doubling

Paul Phillipson, University of Colorado, Boulder, Colorado

09:50 - 10:10	A fast algorithm for the multiplication of generalized Hilbert Matrices with vectors

A. Gerasoulis, Rutgers University, New Brunswick, New
Jersey

10:10 - 10:30	Complexity in quantum chemical calculations for hypercube processors

George F. Adams and Byron Lengsfield, Ballistics Research
Laboratory, Aberdeen Proving Ground, Maryland

10:30 - 11:00	Break

11:00 - 13:00	General Session VI - WH 131

Chairperson: Dr. Jagdish Chandra, U. S. Army Research
Office, Durham, North Carolina

11:00 - 12:00	Computing unsteady reacting flows using vortex methods

Ahmed F. Ghoniem, Massachusetts Institute of Technology,
Cambridge, Massachusetts

12:00 - 13:00	High Mach Number Combustion

Andrew J. Majda, Princeton University, Princeton, New Jersey

13:00 - 13:15	Concluding Remarks and Adjournment

INTERACTIONS, BIFURCATIONS, AND INSTABILITIES
OF HYDRODYNAMIC SURFACES:
A CONFERENCE REPORT

Craig Fithian [3]
Carl L. Gardner [2]
James Glimm [1,2,3]
John Grove [3]
Oliver McBryan [1,2,4]
John Scheuermann

Courant Institute, New York University
New York, N. Y. 10012

Ralph Menikoff [5]
David H. Sharp [5]

Los Alamos National Laboratory
Los Alamos, N. M. 87545

ABSTRACT

The method of front tracking has been demonstrated to provide high resolution of hydrodynamic interfaces. A basic motive for developing this method was to allow a study of the transition to chaos in the case of interface instability. We also show that interactions of tracked waves and bifurcations of interface topology can in certain cases be computed automatically.

These results are then applied to the study of jets and of fingers formed by the Rayleigh-Taylor and Meshkov instabilities. A statistical model for the chaotic regime, due to J. A. Wheeler and one of the authors (D.H.S.), is presented, and its relation to the above computations is outlined.

We also discuss modifications of the front tracking method due to gravitational and geometrical source terms in the Euler equations, and work in progress concerning use of equations of state for real materials.

1

1. Introduction

Proof of scientific principle for the front tracking method has been established in a series of papers by the authors and coworkers [1,2,3,4,5]. This method has been adapted to a variety of problem areas, including oil reservoirs, shock tube experiments, astrophysics, and detonation waves. Recent improvements of this method allow consideration of more complex and interesting problems. Here we consider the interaction between tracked waves, and interface instabilities in the compressible regime (including the Meshkov, Rayleigh-Taylor, and supersonic jet instabilities). We also discuss modifications of the front tracking method due to gravitational and geometrical source terms in the Euler equations, and report on work in progress to allow for the effects of real equations of state.

2. A Description of the Front Tracking Method

Front tracking is an adaptive computational method for solving a hyperbolic system of nonlinear conservation laws. In two-dimensional problems, a moving one-dimensional grid, called the front, is fitted to and tracks selected waves in the solution. These waves can be sharp discontinuities which exist as mathematical solutions of idealized physical equations (e.g. the Euler equations) or waves for which physical quantities change rapidly but smoothly over a fraction of a mesh length (e.g. chemical reaction fronts). For compressible fluid dynamics, these waves include shock waves, contact discontinuities, material interfaces, phase boundaries, slip lines, and chemical reaction fronts.

The front tracking code employs a finite difference method together with the tracking of selected waves to solve the two-dimensional Euler equations of compressible gas dynamics in conservation form. The Euler equations for a compressible, inviscid gas can be cast in the form of a general hyperbolic system of nonlinear conservation laws

$$\mathbf{w}_t + \nabla \cdot \mathbf{f}(\mathbf{w}) = 0 \tag{2.1}$$

by setting

$$\mathbf{w} = \begin{pmatrix} \rho \\ \mathbf{m} \\ E \end{pmatrix} \quad \text{and} \quad \mathbf{f}(\mathbf{w}) = \begin{pmatrix} \mathbf{m} \\ \dfrac{\mathbf{m} \otimes \mathbf{m}}{\rho} + p \\ \dfrac{\mathbf{m}}{\rho}(E+p) \end{pmatrix}. \tag{2.2}$$

ρ is the mass density, $\mathbf{m} = \rho\mathbf{u}$ is the momentum density, E is the total energy density, and p is the thermodynamic pressure. The total energy can be written as $E = \rho e + \dfrac{|\mathbf{m}|^2}{2\rho}$ where e is the specific internal energy. The pressure, density and specific internal energy are related by a caloric equation of state; thus only two of the three quantities are independent. Eqs. (2.1) and (2.2) express the conservation of mass, momentum, and energy.

The front divides the computational grid into topologically connected interior regions called components. The solution is computed by first propagating the front and then the solution in each component.

The front is advanced in two steps. First the Rankine-Hugoniot equations are used to propagate the front normally by solving a nonlocal Riemann problem. Then tangential waves are propagated along the front using a one-dimensional Lax-Wendroff method. At points (called nodes) where discontinuity curves intersect, the propagation is defined by the solution of shock polar equations, as a first approximation to solving a two-dimensional Riemann problem. The propagation of the front for the Euler equations without source terms is described more thoroughly in Ref. [4], while front tracking and two-dimensional Riemann problems are discussed in Ref. [6].

The interior regions between fronts are treated as initial-boundary value problems and the solutions in these regions are computed using an operator split Lax-Wendroff finite-difference method. The front and interior schemes are coupled in a strip of width one mesh

2

spacing on either side of the front.

3. Shock and Contact Wave Interactions

We distinguish between scalar and vector waves. The scalar waves such as contacts, material interfaces, phase boundaries and concentration waves do not produce reflected waves on interaction whereas the vector waves such as shock waves in gas dynamics do. A fairly general algorithm for resolving the interaction of scalar waves was presented in Ref. [7], in the context of tracked saturation fronts in oil reservoirs. We show the type of complex interface that can develop from a simple one in Fig. 3.1. We are also interested in the interaction of (vector) shock waves with contact waves and with each other. As a model problem for the study of this interaction, we consider a simplified version of the Meshkov instability, in which a shock wave hits a contact having a small (sine wave) perturbation from planar. After passage of the shock wave, this perturbation grows at first exponentially and then linearly in time before coming to rest. The late time behavior is discussed in the next section. Here we describe the sequence of shock interaction problems that take place in the initiation process.

In Fig. 3.2 we show a sequence of shock and contact fronts for a shock wave hitting an interface between warm and cold air, while in Fig. 3.3 we show a similar sequence where the interface separates air and SF_6 at the same temperature. In both cases the shock is incident in the lighter gas (the warmer air in Fig. 3.2 and the air in Fig. 3.3). Each simulation begins shortly before the shock wave collides with the contact discontinuity surface. When the shock wave reaches this surface it is transmitted through and reflected by the contact. The contact discontinuity is in turn deflected by this interaction. We observe diffracted wave patterns propagating away from the original point of collision as the shock continues to propagate into the gas interface. Eventually the shock wave will pass completely through the contact discontinuity, and the reflected and transmitted waves will propagate away from each other on opposite sides of the gas interface. In general this will produce complicated wave interactions, but in our model we only track the transmitted shock and the reflected wave (if it is a shock). This approximation assumes the other waves produced by this interaction are weak enough not to require tracking.

The front tracking code is well suited for the propagation of interior points on tracked curves, but must be extended to handle the complicated wave patterns that occur when two or more waves interact at a single point. In the shock-contact interaction each of the diffraction patterns consists of an incident shock colliding with a contact discontinuity producing reflected and transmitted waves. Such a configuration will be called a diffraction node. The analysis of the interaction between a planar shock wave and a planar contact discontinuity has been discussed in detail in Refs. [8, 9, 10, 11, 6] and here we will only summarize these results as they are applied in the front tracking code. In a neighborhood of a diffraction node we ignore any curvature and replace the two colliding curves by their tangents. We next assume that there exists a reference frame in which the flow near the point of interaction is steady. Finally we restrict our attention to the so called regular reflection case in which the interaction occurs at a single point, the transmitted wave is a shock and the reflected wave is either a shock or a centered rarefaction wave. (More complicated configurations include Mach and multiple Mach type reflections.) This assumption is valid provided the angle between the incident shock and the contact discontinuity is sufficiently small. Since flow does not cross a contact discontinuity, the stream lines on opposite sides of the interface must be parallel. This means that the flow through the incident shock and the reflected wave must be turned by the same amount as the flow through the transmitted shock. If we assume that the states of the gas on both sides of the contact discontinuity ahead of the incident shock are known, together with the strength of the incident shock (say the pressure jump across the shock), then the Rankine-Hugoniot conditions together with this restriction provide a system of algebraic equations from which the pressure behind the reflected and transmitted waves can be found (this solution may be multi-valued). This pressure can then be used to construct the states behind the transmitted shock and reflected wave along with the angles at which these

3

waves meet the point of shock diffraction.

Since we are dealing with curved waves, this calculation is performed at each timestep. The transformation to the steady frame of an individual diffraction node is found by a geometric construction. The incident shock and the ahead contact discontinuity are first propagated separately, ignoring any interaction between the two waves. The intersection between the two propagated curves is found and this is used as the updated node position, from which a node velocity is computed. This velocity defines the transformation to the steady frame of the node. New states and wave angles about the diffraction node are computed and inserted into the tracked wave structures.

The geometrical construction of the node velocity is also important since it provides a method of detecting wave interactions. When the original shock passes through the contact, the ahead curves will both be short segments that will propagate past one another in the finite time Δt. The propagated curves do not intersect and hence a node velocity cannot be computed. At this point control is shifted to routines designed to identify and handle such interactions.

4. Interface Instabilities

We have studied a series of related problems, each of which leads to fingering instabilities or jets, with the penetration of a heavy material into a lighter ambient material. Followed to late time, this leads to a chaotic mixing regime discussed in the next section. The series of problems arise from different procedures to initiate this instability, as an accelerated surface [12], supersonic jet [13], shock-contact collision [14], or Rayleigh-Taylor instability [15]. We have considered a range of density ratios up to 100:1 and accelerating forces, which for the Rayleigh-Taylor problem are in the range of up to 10^6 to 10^9 g depending on the length scale of the perturbation considered. The Mach numbers considered spanned a range of from 0.1 to 6.

The compressible Rayleigh-Taylor problem depends on three dimensionless parameters: the density ratio $D = \dfrac{\rho_b}{\rho_a}$, where ρ_b is the density of the heavy gas just below the interface (we assume gravity points up) and ρ_a is the density of the light gas just above the interface; the polytropic gas constant γ (here we set $\gamma_a = \gamma_b = 1.4$) or other information to set the equation of state for the heavy and light fluids; and a Mach number M defining the ratio of a gravitational time scale to a sound speed time scale. M defines a dimensionless compressibility. We take $M^2 = \dfrac{g\lambda}{c_b^2}$, where λ is the wavelength of the interface perturbation and c_b is the sound speed in the unperturbed heavy fluid. In Fig. 4.1 we show a sequence of interface positions for a compressible heavy gas falling into a lighter gas with $D = 2$ and $M^2 = 0.5$. In this case the terminal Mach number of the bubble and spike is about 0.2. In Fig. 4.2 we show the case of four symmetric bubbles and spikes, for $D = 10$ and $M^2 = 0.89$. In Fig. 4.3 we show a similar sequence for $D = 10$ and $M^2 = 0.89$, in which there is a capture of the smaller side bubbles by the larger central one. (For an interface with multiple modes, we give the maximum value of M^2.) We refer to the cases of single bubble dynamics and of bubble capture as the one and two body problems of bubble dynamics; they are central to the statistical model for the mixing regime discussed in the next section.

Computations of supersonic jets by Norman, Smarr, and Winkler (NSW) [16] have generated a great deal of interest, due to their qualitative agreement with observations and their quantitative predictions. Since the radio telescope observations will become more detailed in the near future, it is of great interest to compare computations of supersonic jets by different methods. To this purpose, our computations using a "surface" front tracking method may be contrasted with the results obtained by NSW using a "volume" front tracking method. We find overall agreement in the wave structure of the computations, but find a marked difference in the details of the contact boundary between the jet and ambient gases. We believe

our method offers a higher degree of resolution of the tracked contact, since our method tracks it as a sharp discontinuity rather than as a "smeared out" interface, and preserves the integrity of the tracked front from step to step.

Fig. 4.4 displays the evolution to late time of a cylindrically symmetric Mach 3 jet. The density ratio of jet gas to ambient gas is 10:1. γ was set equal to 5/3. Note the presence of a bow wave in front of the jet and of a terminal shock near the head of the jet beam, preceded by a rarefaction wave. This terminal shock system may explain the observed hot spots terminating astrophysical jets. The contact shape displays large-scale Kelvin-Helmholtz rollup, and the development of two-dimensional pinch waves.

5. The Mixing Regime

The late stages of a Rayleigh-Taylor unstable interface lead to a chaotic mixing regime. The portion of the mixing layer adjacent to the heavy fluid is dominated by the mechanism of bubble merger or amalgamation. A model for bubble merger due to J. A. Wheeler and one of the authors (D.H.S.) [17] (a brief description is also contained in [18]) has been analyzed numerically. In the model, it is assumed that the interface is piecewise constant and single valued, so that the bubbles are the piecewise constant intervals in the interface. A simple scaling argument shows that the bubble velocity is $\dot{z} = \text{const}\,(gr)^{\frac{1}{2}}$ where r is the bubble radius. The constant is a function of the dimensionless parameters of the problem and can be determined numerically by the solution of the one body problem as discussed in the previous section. When a large bubble moves sufficiently far ahead of a smaller bubble, the two are forced to merge, with a new height set by conservation of mass. The merger height is then determined numerically by a solution of the two body problem as discussed in the previous section. In Fig. 5.1 we show a sequence of successive sample interfaces generated by the numerical solution of this model, and in Fig. 5.2 we plot the average bubble velocity as a function of time, for a specific choice of initial data consisting of a Gaussian distribution about a uniform bubble size. One can see clearly the trend toward merger of bubbles and the growth of larger bubbles at the expense of the smaller ones.

6. Front Tracking with Source Terms

Gravitation and cylindrical symmetry introduce source terms into the conservation form of the Euler equations. In this section, we discuss the modifications necessary in applying the front tracking method to problems with gravity or cylindrical symmetry.

With a gravitational force, Eq. (2.1) is modified by source terms:

$$w_t + \nabla \cdot f(w) = S(w) \tag{6.1}$$

where

$$S = \begin{pmatrix} 0 \\ \rho g \\ m \cdot g \end{pmatrix}. \tag{6.2}$$

In this case, E stands for the internal plus kinetic energy density. The gravitational potential energy density has been shifted from the left-hand side of Eq. (2.1) and appears as $m \cdot g$ in S .

The cylindrically symmetric Euler equations can be written in the form (6.1) with

$$w = \begin{pmatrix} \rho \\ \rho u_r \\ \rho u_z \\ E \end{pmatrix}$$

provided it is understood that $\nabla \cdot f$ is to be interpreted with a flat metric in (r, z) coordinates as $\partial_r f_r + \partial_z f_z$, where

$$f_r = \begin{pmatrix} \rho u_r \\ \rho u_r^2 + p \\ \rho u_r u_z \\ u_r(E+p) \end{pmatrix} \text{ and } f_z = \begin{pmatrix} \rho u_z \\ \rho u_r u_z \\ \rho u_z^2 + p \\ u_z(E+p) \end{pmatrix}.$$

With this interpretation,

$$S = -\frac{u_r}{r}\begin{pmatrix}\rho \\ \rho u_r \\ \rho u_s \\ E+p\end{pmatrix}. \tag{6.3}$$

The interior solver and tangential sweep with source terms are modified only by including S in the finite difference equations. The Lax-Wendroff method remains second-order accurate even with the source term S.

For the normal and tangential sweeps of the front, the Euler Eqs. (6.1) are split into normal and tangential parts:

$$w_t + \hat{n}\cdot[(\hat{n}\cdot\nabla)f(w)] + \hat{s}\cdot[(\hat{s}\cdot\nabla)f(w)] = S_n + S_s, \tag{6.4}$$

where for gravity

$$S_n = \begin{pmatrix}0 \\ \rho g_n \\ m\cdot g_n\end{pmatrix} \quad \text{and} \quad S_s = \begin{pmatrix}0 \\ \rho g_s \\ m\cdot g_s\end{pmatrix},$$

and for cylindrical symmetry

$$S_n = -\frac{n_r}{r}u\cdot\hat{n}\begin{pmatrix}\rho \\ \rho u_r \\ \rho u_s \\ E+p\end{pmatrix} \quad \text{and} \quad S_s = -\frac{s_r}{r}u\cdot\hat{s}\begin{pmatrix}\rho \\ \rho u_r \\ \rho u_s \\ E+p\end{pmatrix}.$$

The splitting method is first to solve the normal equations

$$w_t + \hat{n}\cdot[(\hat{n}\cdot\nabla)f(w)] = S_n, \tag{6.5}$$

and then the tangential equations

$$w_t + \hat{s}\cdot[(\hat{s}\cdot\nabla)f(w)] = S_s. \tag{6.6}$$

Eq. (6.6) is solved by a one-dimensional Lax-Wendroff method. The normal sweep is further modified by an operator splitting method. For tracked shocks, the solution to Eq. (6.5) is found by solving a nonlocal Riemann problem [4] for the homogeneous equation

$$w_t + \hat{n}\cdot[(\hat{n}\cdot\nabla)f(w)] = 0,$$

and then the corrections are added by solving

$$w_t = S_n.$$

For through-flow boundaries, Eq. (6.5) is solved by a one-dimensional Lax-Wendroff method.

For contacts and wall boundaries the solution of the nonlocal Riemann problem in the normal direction is modified to include the effects of source terms in the characteristic equations. If $S = 0$ the states at a contact or wall boundary may be updated by solving the characteristic equations

$$dp \pm \rho c\, du = 0 \tag{6.7}$$

for characteristic wave speeds $u \pm c$. With a non-zero S, Eq. (6.7) becomes

$$dp \pm \rho c\, du = S_n dt, \tag{6.8}$$

where $S_n = \pm\rho c g_n$ for gravity, and $S_n = -\frac{n_r}{r}u\rho c^2$ for cylindrical symmetry. The finite difference form of Eq. (6.8) is

6

$$p - p_0 \pm \rho c(u - u_0) = S_n dt, \qquad (6.9)$$

where the unsubscripted variables indicate the quantities at the head of the characteristic (at time $t + dt$), the subscript 0 indicates the quantities at the foot of the characteristic (at time t), and $u = \hat{n} \cdot \mathbf{u}$.

The node propagation algorithms are modified through Eqs. (6.5) and (6.6); the Rankine-Hugoniot jump relations are unchanged.

7. Real Equations of State

Much work has recently been devoted to the problem of implementing realistic equations of state for gas dynamical calculations. Commonly, scientific studies assume a polytropic or gamma law gas equation of state; our goal is to extend our front tracking hydrodynamics code to handle more general equations of state. Over the past year a considerable effort was made to isolate and modularize the equation of state dependences in our gas dynamics simulation program. This work has now been completed and the equation of state dependences have been isolated to a relatively small number of subprograms such as the calculation of pressures from densities and energies or the calculation of sound speeds. Furthermore these subprograms have been written in such a way that the user may "plug in" additional equations of state as they are developed. We are now in the processes of adding two additional equations of state to our gas code in addition to the currently supported polytropic equation of state. These are the so called stiffened polytropic equation of state and the Los Alamos National Laboratory table look up equation of state SESAME.

An equation of state is a functional relation between the thermodynamic variables that describe the state of a gas. These variables include the density, pressure, temperature, specific internal energy and the specific entropy of the gas. Only two of these variables can be independent and the equation of state describes the remaining quantities when any two are given. For example in the polytropic equation of state the specific internal energy e is given by $e = \dfrac{p}{(\gamma - 1)\rho}$ where p and ρ are the pressure and density of the gas respectively and γ is a dimensionless constant greater than one. The temperature T of a polytropic gas is given by the ideal gas law $RT = \dfrac{p}{\rho}$ where R is a positive constant. The stiffened polytropic equation of state is a generalization of the polytropic equation of state, where $e = \dfrac{p + \gamma p_0}{(\gamma - 1)\rho}$ and $RT = \dfrac{p + p_0}{\rho}$. As in the polytropic model $R > 0$ and $\gamma > 1$ are constants. The additional constant $p_0 \geq 0$ has the dimensions of pressure. If $p_0 = 0$ the stiffened polytropic model reduces to the polytropic case. Stiffened polytropic equations of state have been used to model metals. For instance, tungsten may be modeled with $\gamma \approx 3.2$ and $p_0 \approx 1$ Mbar over a range of pressures from zero to seven Mbar.

Both the polytropic and the more general stiffened polytropic equations of state are examples of simple analytic equations of state. Their implementation into a hydrodynamics code is relatively simple and involves the calculation of various quantities such as the sound speed and shock Hugoniots. Because of the simple nature of these models it is possible to find explicit formulas for these quantities which allow for quick and accurate numerical calculations. Their main limitations are that real materials only approximately satisfy them over a limited range of temperatures and pressures, and they do not include mechanisms for phase transitions. The Los Alamos National Laboratory program SESAME is an attempt to overcome these problems by using a tabular equation of state. Here we are given a rectangular grid of densities and temperatures with the pressure and specific internal energy given at each grid point (ρ, T). Pressures and energies at intermediate densities and temperatures can then be found by interpolation.

One advantage of such a program is that it allows one to support a large number of materials using the same basic software. In addition the table for an individual material may

7

be built by combining several different analytic models each with its own range of validity, or by using directly measured experimental information. However such generality and flexibility exact a cost for a hydrodynamics code. Quantities which reduce to simple formulas for the polytropic model must now be found by solving systems of nonlinear equations or differential equations numerically. In particular the calculation of shock Hugoniots and adiabatic (constant entropy) curves can become extremely expensive. Since in any code which involves the solution of Riemann problems (such as our front tracking code) these quantities must be computed hundreds or even thousands of times each timestep, it is easy to see that numerical simulations can be impractical on even the most advanced machines.

We are now in the process of developing an implementation of the SESAME program into our gas dynamics code which will address these inefficiency problems by precomputing as much as possible the quantities which are used repeatedly in the solution of Riemann problems. The original SESAME program already included a facility for inverting the given tables into a format in which the density and specific internal energy were the independent variables. To this we are adding inverted forms with pressure and density or pressure and specific entropy as independent variables. In addition it is possible to precompute various integrals which occur in the solution of the Riemann problem and include them as data in the table. Our hope is that by applying these principles we will be able to achieve rates of solution to Riemann problems which are comparable to those obtained for polytropic or other similar equations of state.

References

1. B. Bukiet, C. L. Gardner, J. Glimm, J. Grove, J. Jones, O. McBryan, R. Menikoff, and D. H. Sharp, "Applications of Front Tracking to Combustion, Surface Instabilities and Two-Dimensional Riemann Problems," *Transactions of the Third Army Conference on Applied Mathematics and Computing*, pp. 223-243, 1986 .

2. James Glimm and D. H. Sharp, "Numerical Analysis and the Scientific Method," *DOE Research and Development Report DOE/ER/03077-270*, 1986.

3. J. Glimm, B. Lindquist, O. McBryan, and L. Padmanabhan, "A Front Tracking Reservoir Simulator I: The Water Coning Problem," in *Frontiers in Applied Mathematics*, vol. 1, SIAM, Philadelphia, 1983.

4. 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.

5. J. Glimm, W. B. Lindquist, O. McBryan, and G. Tryggvason, "Sharp and Diffuse Fronts in Oil Reservoirs: Front Tracking and Capillarity," *SIAM, Proc. Math. and Comp. Methods in Seismic Exploration and Reservoir Modelling, Houston*, Jan, 1985.

6. 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.

7. J. Glimm, J. Grove, B. Lindquist, O. A. McBryan, and G. Tryggvason, "The Bifurcation of Tracked Scalar Waves," *DOE Research and Development Report DOE/ER/03077-272*, 1986.

8. L. F. Henderson , "The Refraction of a Plane Shock Wave at a Gas Interface," *J. Fluid Mech.* , vol. 26, p. 607, 1966 .

9. 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.

10. 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 .

11. 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 .

12. C. L. Gardner, "Supersonic Interface Instabilities of Accelerated Surfaces and Jets," *Phys. Fluids*, vol. 29, p. 690, March 1986.

13. C. Fithian and C. L. Gardner, "Front Tracking and Supersonic Jets: Computations and Methodology," *To Appear*.

14. J. Grove, "Front Tracking and Shock-Contact Interactions," *To Appear*.

15. C. L. Gardner, J. Glimm, O. McBryan, R. Menikoff, and D. H. Sharp, "The Dynamics of Bubble Growth for Rayleigh-Taylor Unstable Interfaces," *To Appear*.

16. L. L. Smarr, M. L. Norman, and K-H.A Winkler, "Shocks, Interfaces, and Patterns in Supersonic Jets," *Physica*, vol. 12D, pp. 83-106, 1984.

17. D. H. Sharp and J. A. Wheeler, "Late Stage of Rayleigh-Taylor Instability," *Institute of Defense Analyses. Unpublished Technical Report*, 1961.

18. D. H. Sharp, "An Overview of Rayleigh-Taylor Instability," *Physica*, vol. 12D, pp. 3-18, 1984.

(a) Time step 0   (b) Time step 40

(c) Time step 80   (d) Time step 240

$10 \, \Delta x = 10 \, \Delta y$

elliptic   hyperbolic
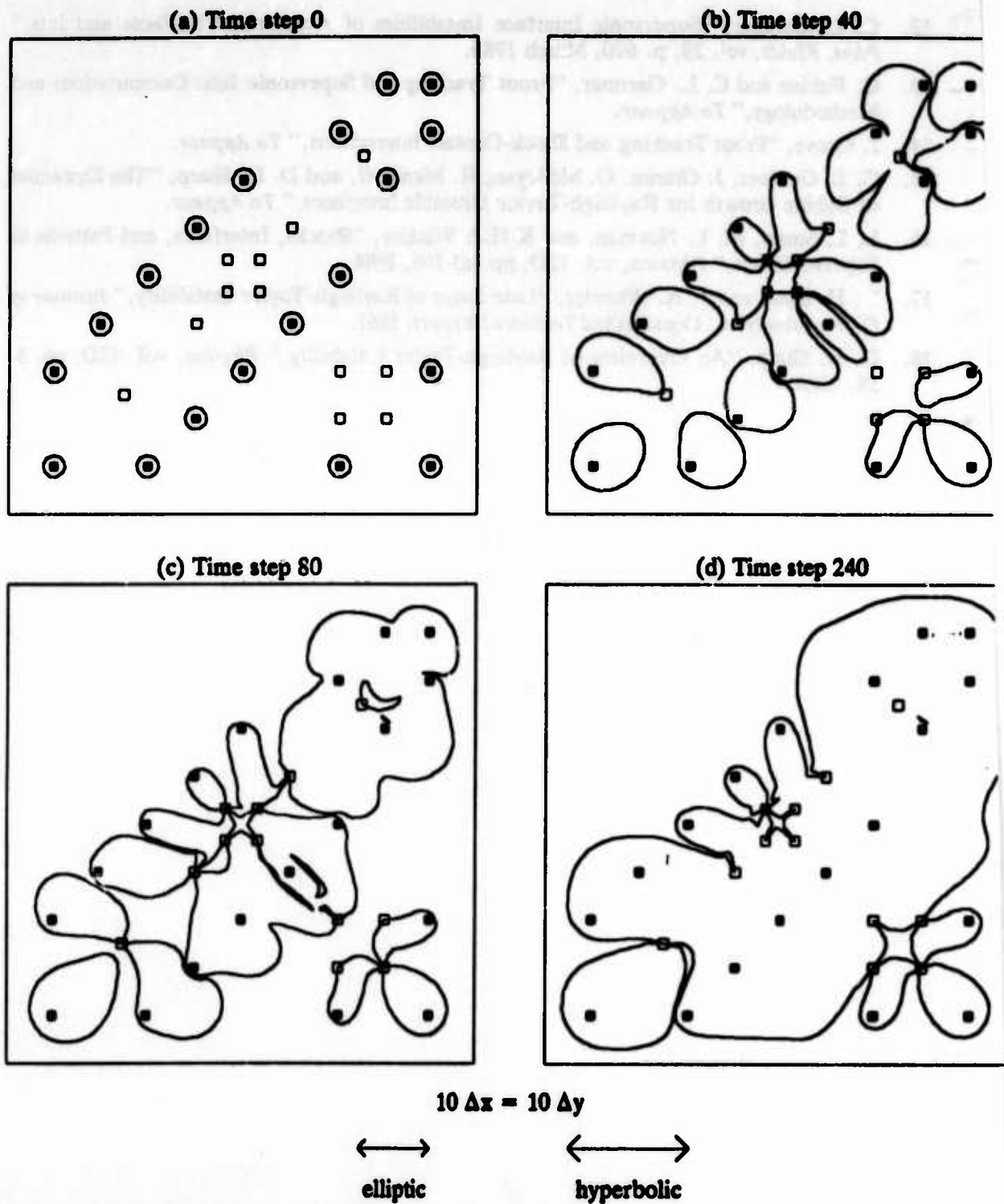
Fig. 3.1 Plots of the oil-water interfaces for a well configuration consisting of 19 injecting wells (crossed squares) and 12 producing wells (open squares).

(a) time 0

(b) time 0.06

$\longleftrightarrow$
$10 \, \Delta x \, = \, 10 \, \Delta y$

(c) time 0.5

(d) time 5
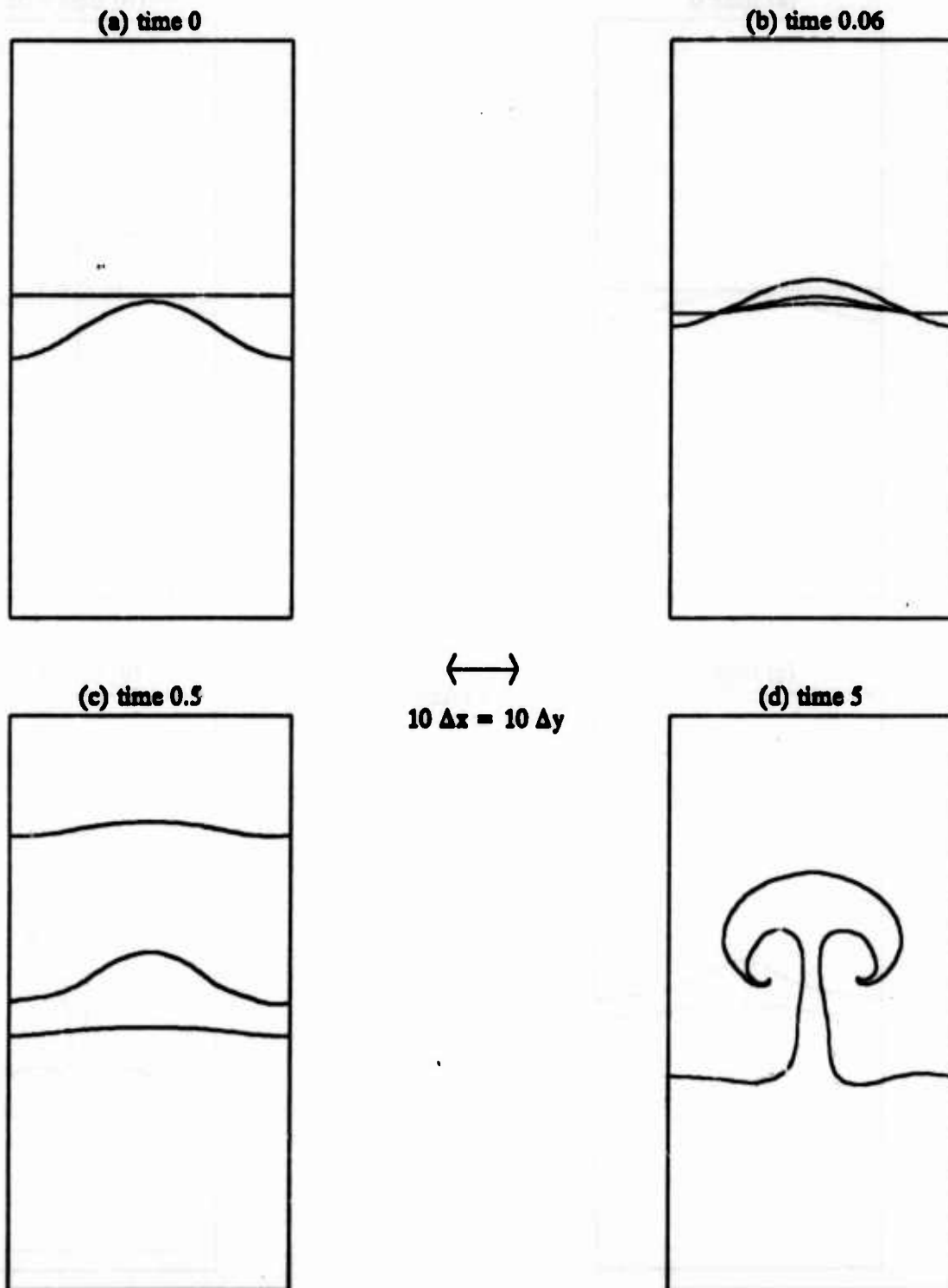
Fig. 3.2 A shock hitting a contact discontinuity separating two masses of air at different temperature. The pressure ratio across the shock is 1000 and the density ratio across the contact discontinuity is approximately 2.86. The shock is incident in the lighter gas.

(a) time 0

(b) time 0.02

(c) time 0.5

$\longleftrightarrow$

$10 \, \Delta x = 10 \, \Delta y$

(d) time 3.5
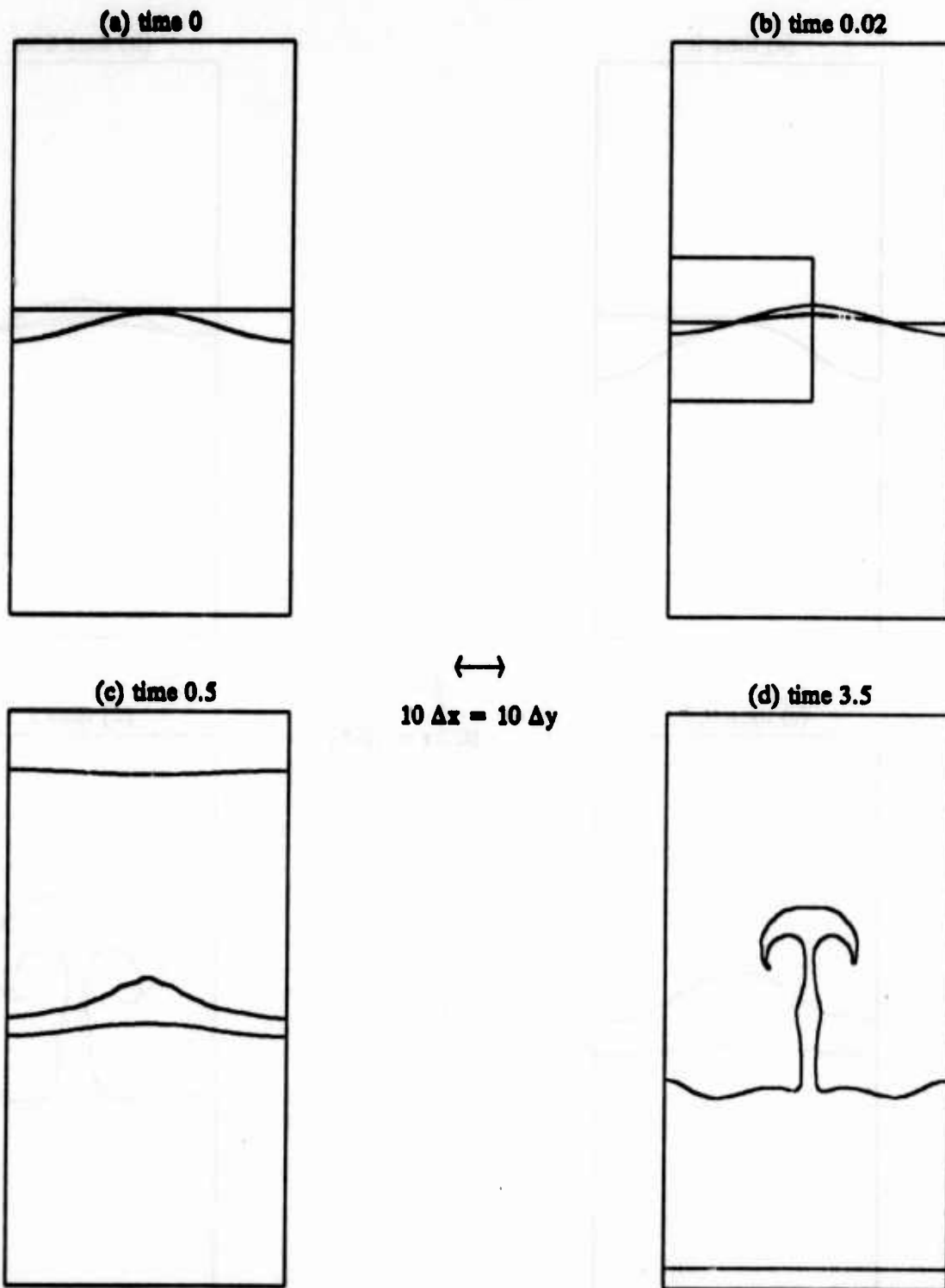
Fig. 3.3 A shock hitting a contact discontinuity separating air from the gas SF_6. The contact discontinuity curve is given an initial shape of a sine curve. The shock is incident from the air and has a pressure ratio of 10. The boxed region in Fig 3.3b is blown up in the next figure.

time 0.02

deflected contact

reflected shock

incident shock

transmitted shock

ahead contact

$\overleftrightarrow{}$
$\Delta x = \Delta y$
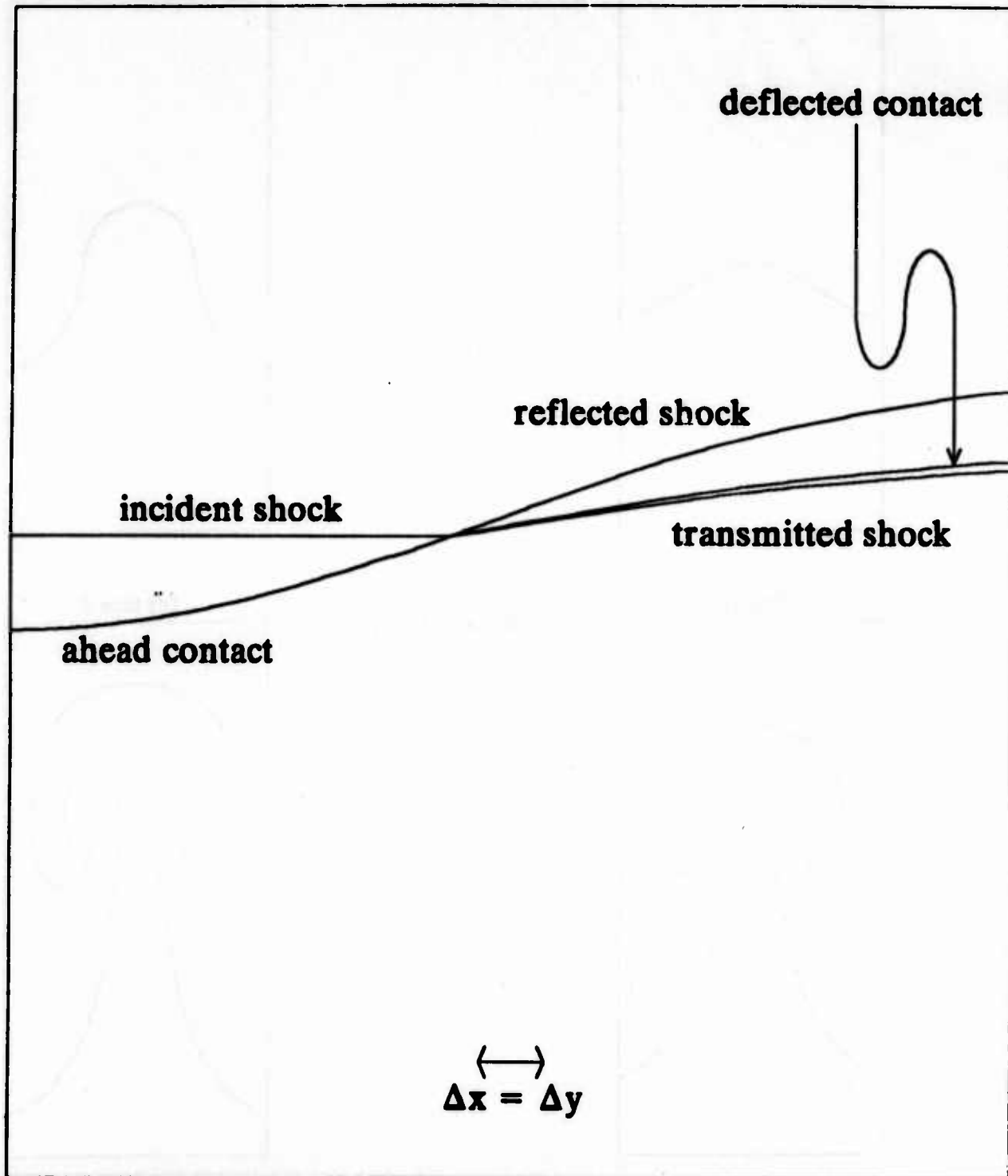
Fig. 3.4 A blowup of a subregion of Fig 3.3b showing the incident shock colliding with the ahead contact discontinuity, producing reflected and transmitted shocks.

13
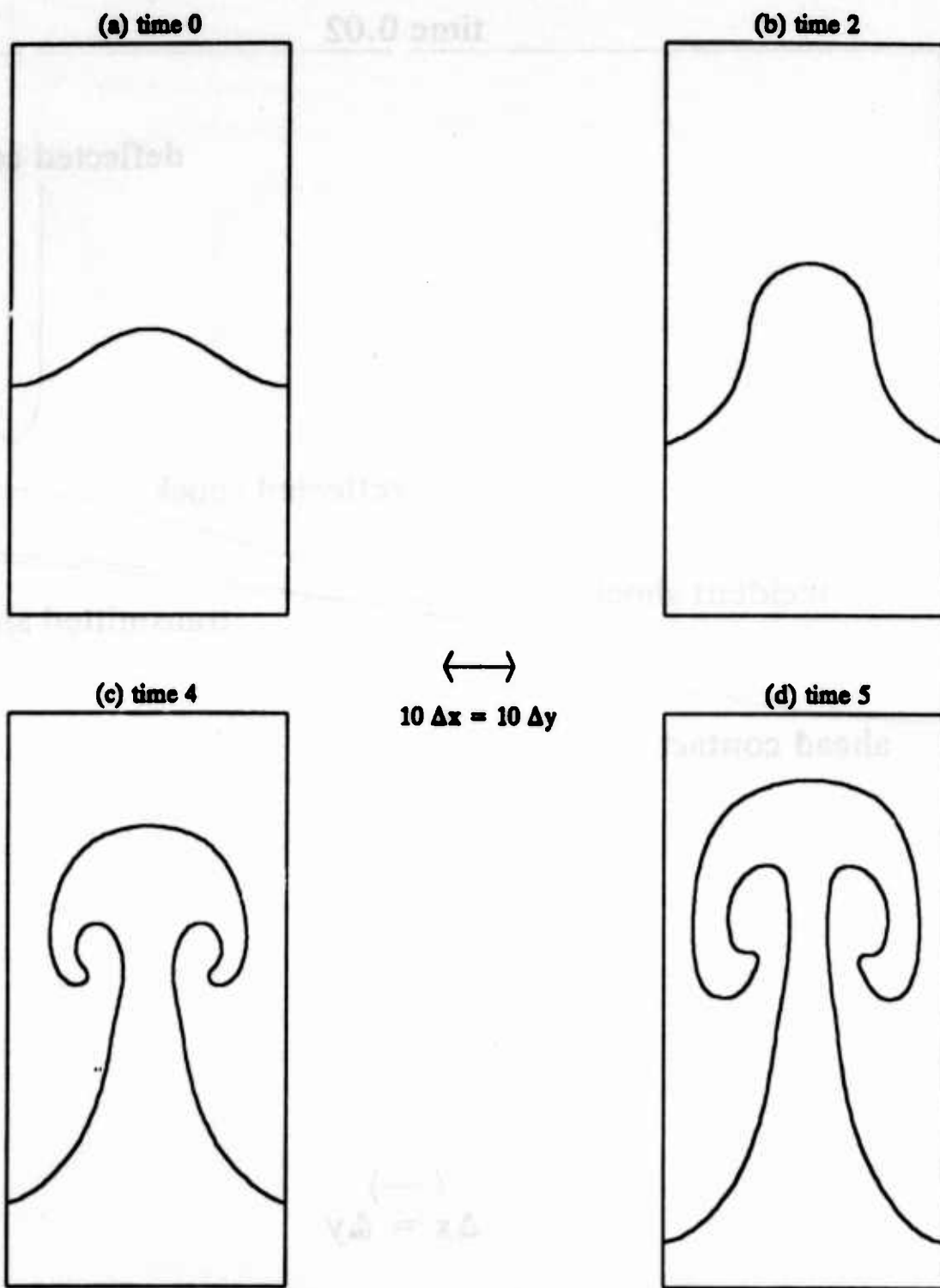
Fig. 4.1 A sequence of interface positions for a compressible heavy gas (below) falling into a lighter gas (above) with a density ratio of 2:1; in this case the terminal Mach number of the bubble and spike is about 0.2. Gravity points upward.
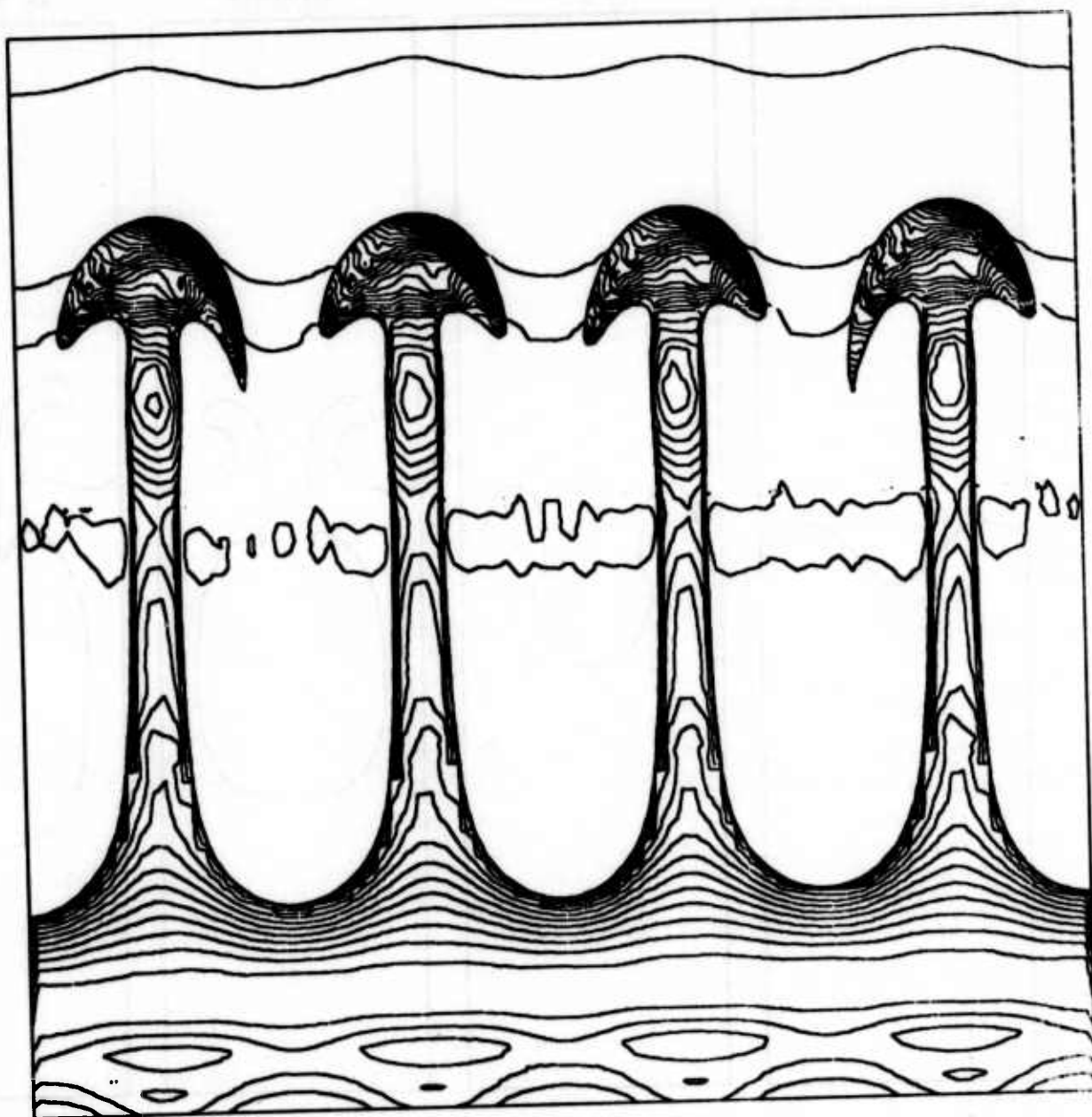
$$10 \, \Delta x = 10 \, \Delta y$$

⟷

Fig. 4.2 Density contours for the Rayleigh-Taylor instability for the case of four symmetric bubbles and spikes with a density ratio of 10:1.

15

(a) time 0 **(b) time 1** **(c) time 2** **(d) time 2.2**

$10\,\Delta x$

$10\,\Delta y$
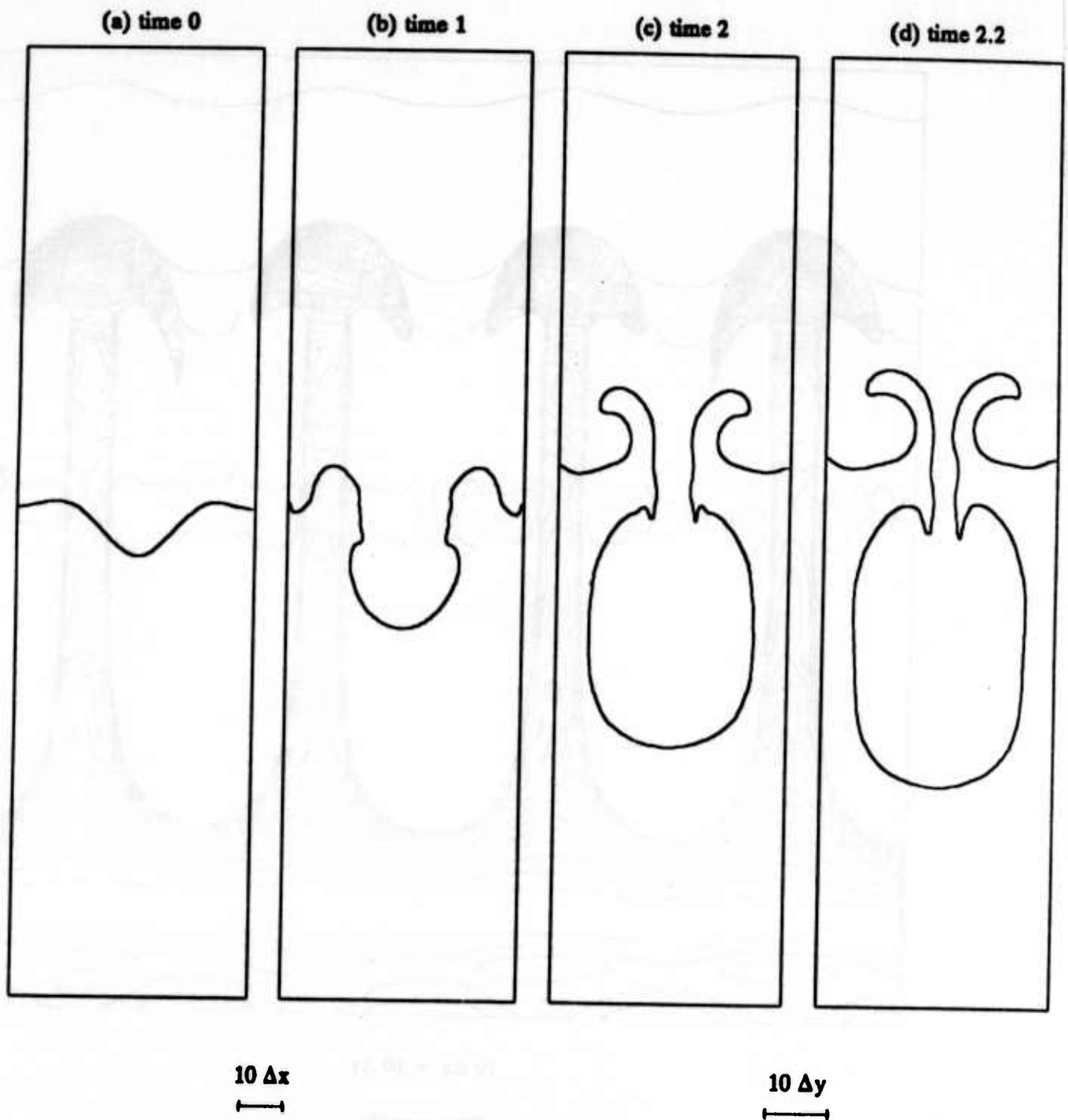
Fig. 4.3 The Rayleigh-Taylor instability showing the capture of smaller side bubbles by a larger central one.

(a) time 0.7 interface plot   (b) time 0.7 pressure contours
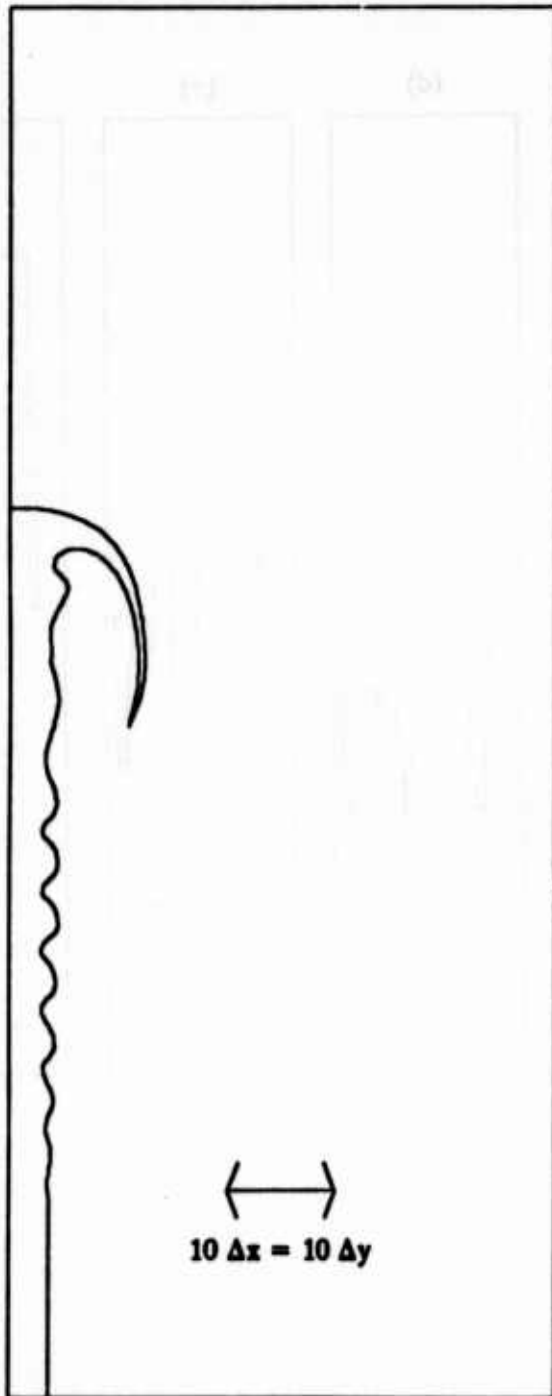
$$\langle \longleftrightarrow \rangle$$
$$10 \, \Delta x = 10 \, \Delta y$$

$$\langle \longleftrightarrow \rangle$$
$$10 \, \Delta x = 10 \, \Delta y$$
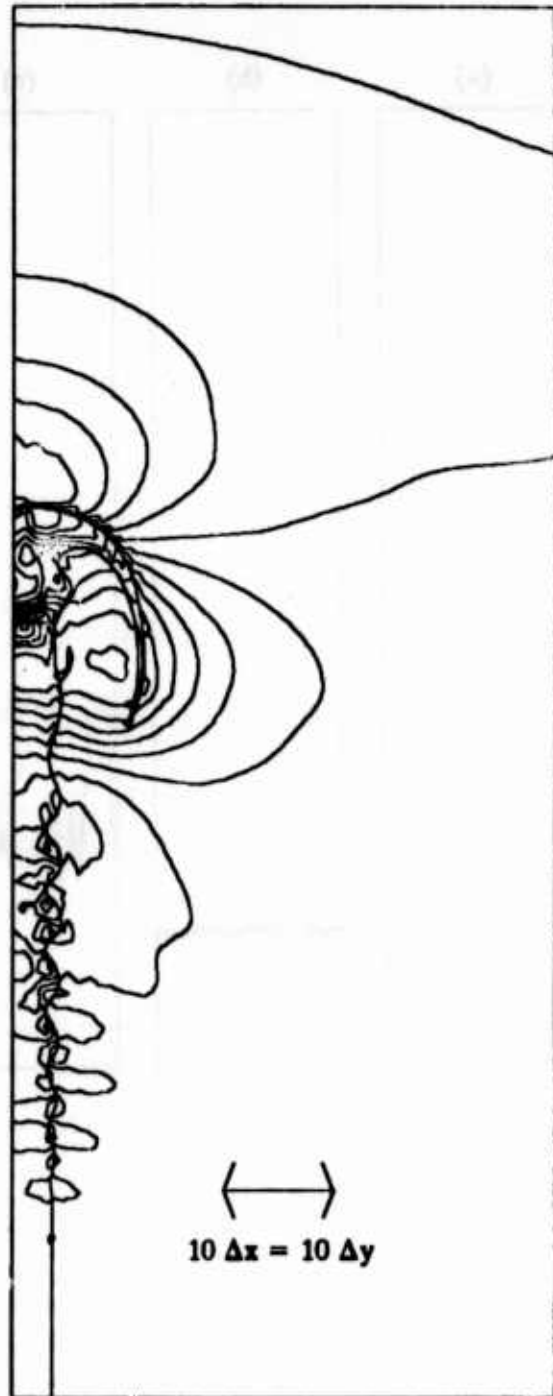
Fig. 4.4 Plots of a cylindrically symmetric Mach 3 jet. The density ratio of jet gas to ambient gas is 10:1.

17
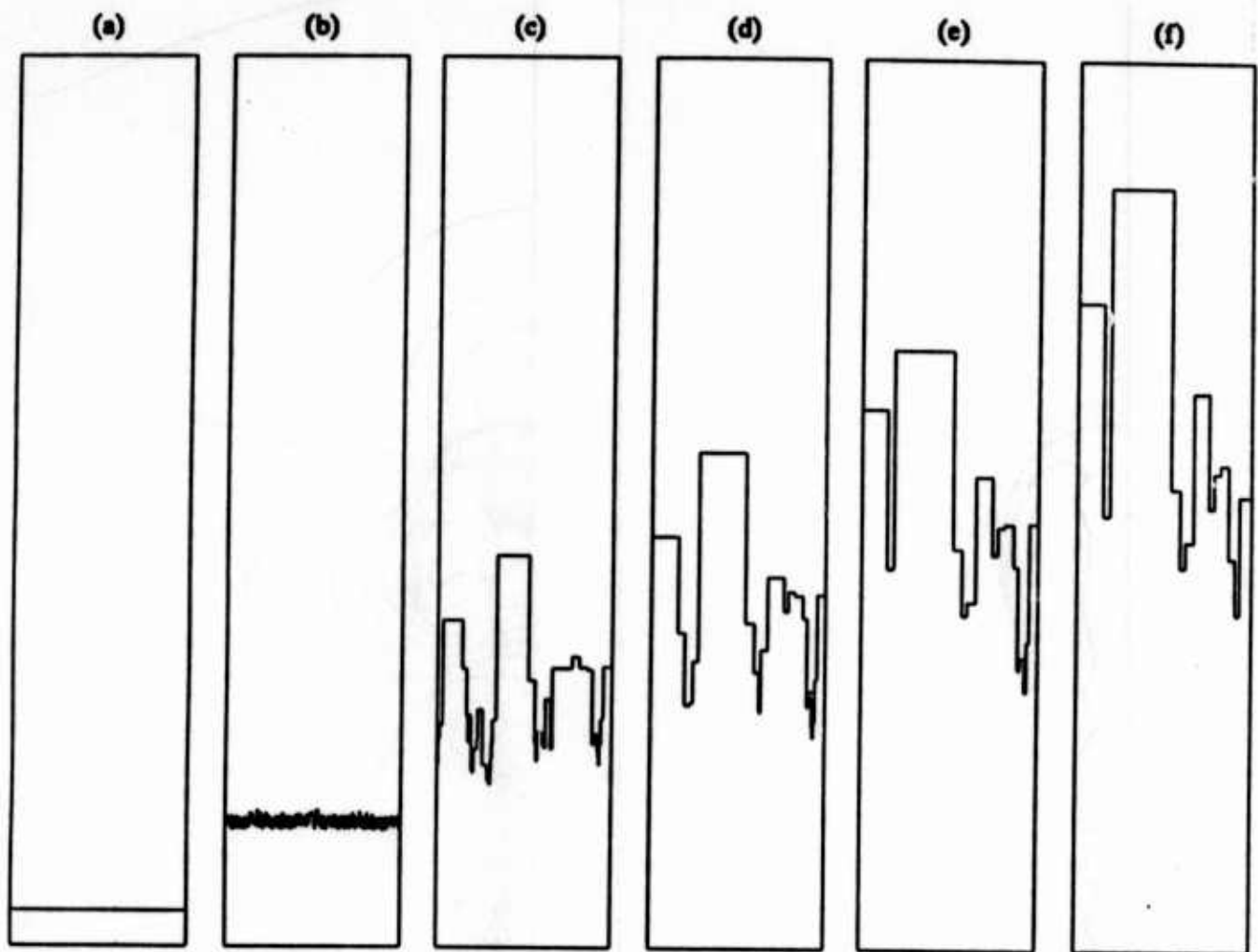
Fig. 5.1 A sequence of successive sample interfaces generated by the numerical solution of a bubble growth model.

18

VELOCITY vs TIME
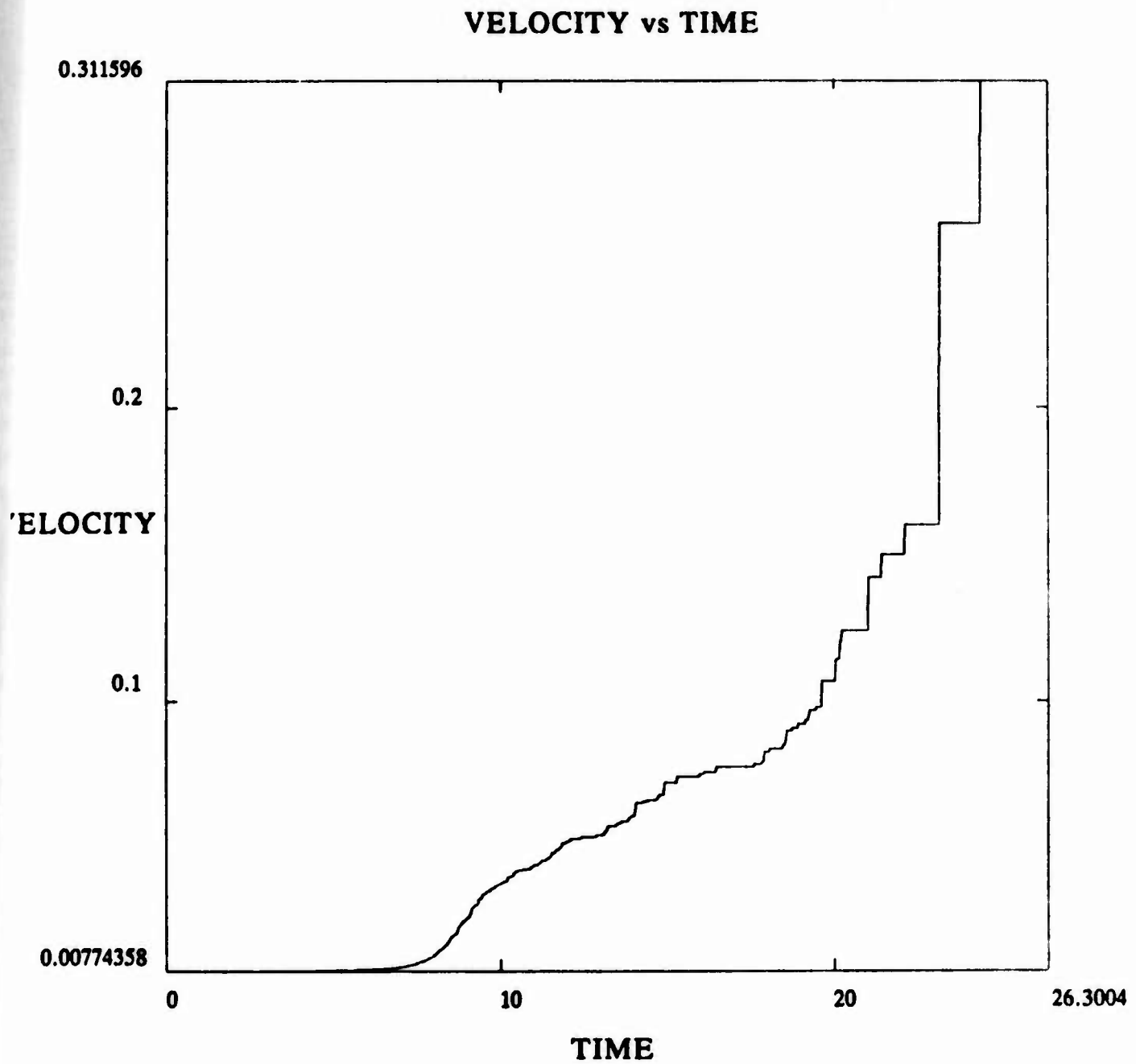
Fig. 5.2 The average bubble velocity as a function of time, for a specific choice of initial data consisting of a Gaussian distribution about a uniform bubble size.

NONLINEAR VISCOELASTIC MATERIALS WITH FADING MEMORY

John A. Nohel[(*)]
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705

Abstract. The equations governing the motion of viscoelastic materials with fading memory incorporate a nonlinear elastic-type response with a natural dissipative mechanism. Our purpose is to discuss the subtle effects of this mechanism in viscoelastic materials of Boltzmann type. Recent results on the global existence and decay of classical solutions for smooth and small data (in one space dimension) are reviewed for smooth and singular memory kernels; for smooth kernels a number of such results can be generalized to several space dimensions. A recent result on the development of singularities in finite time for large data is discussed; several open problems are formulated. A program for a studying weak solutions for such systems, including the development of numerical algorithms, is outlined.

1. **Introduction.** The equations governing the motion of nonlinear elastic bodies are quasilinear hyperbolic systems for which smooth solutions generally lose regularity in finite time due to the formation of shock fronts. Some materials incorporate a nonlinear elastic-type response with a natural dissipative mechanism, and it is important to understand the effects of the dissipation on the behaviour of the solutions of the equations of motion.

The purpose of this lecture is to discuss the effects of the subtle dissipative mechanism due to memory effects in viscoelastic materials of Boltzmann type. This dissipation is more delicate than that exhibited by viscoelastic materials of the rate type for which globally defined smooth solutions exist, even for large smooth data.

The paper is organized as follows. In Section 2 we formulate mathematical models for the motion of nonlinear viscoelastic materials and we motivate the mathematical theory. In Section 3 we survey recent results on the global existence of smooth solutions for smooth and small data. In Section 4 we present a recent result on the breakdown of smooth solutions for large, smooth data and discuss briefly related open questions including those regarding weak solutions and numerical methods (Remarks 4.8). We restrict our attention throughout to one-dimensional problems and provide some references for multidimensional problems. Moreover, we consider only a purely mechanical theory, i.e. we neglect thermal effects.

2. **Mathematical Models and Dynamic Problems.** Consider the longitudinal motion of a homogeneous one-dimensional body (e.g. a bar of uniform cross-

section) occupying an interval B in a reference configuration, which we
assume to be an equilibrium state, and having unit reference density. B may
be bounded or unbounded. Let u(x,t) denote the displacement at time t of
a particle with reference position x (i.e. x + u(x,t) is the position at
time t of the particle at x). The strain which measures local stretching
is defined by $\epsilon := u_x(x,t)$. Let σ denote the stress at time t of the
particle with reference position x (σ measures the contact force per unit
area). The balance of linear momentum yields the equation of motion

$$u_{tt} = \sigma_x + f \ , \quad x \in B, \ t > 0 \ , \tag{2.1}$$

where subscripts denote partial derivatives and where f is an external body
force. In order to characterize the material, (2.1) is supplemented by a
constitutive assumption which relates the stress to the motion. In addition,
initial data, as well as suitable boundary data if B is not R, are adjoined
to (2.1). We remark that in a physical problem the cross-section does not
generally remain uniform as the bar is stretched. More realistic problems can
be treated by similar techniques.

If the body is homogeneous and purely elastic, the stress depends on the
strain through the constitutive relation $\sigma(x,t) = \phi(\epsilon(x,t))$, where φ is a
given smooth function satisfying the assumptions (i) $\phi(0) = 0$, (ii) $\phi'(0) > 0$;
(i) reflects the fact that the reference position is taken as an equilibrium
state, and (ii) that the stress increases with the strain, at least near
equilibrium. The equation of motion (2.1) becomes the familiar, one-
dimensional, quasilinear wave equation

$$u_{tt} = \phi(u_x)_x + f \quad (x \in B, \ t > 0) \ ; \tag{2.2}$$

if B is bounded it is assumed that the assigned boundary data and initial
data are compatible. For (2.2) there is no natural dissipative mechanism.
Indeed, Lax [33], also MacCamy and Mizel [37] and Kleinerman and Majda [31]
have shown that if φ is not linear, the Cauchy problem for (2.2) (f ≡ 0)
does not generally possess globally defined smooth solutions, no matter how
smooth and small one takes the initial data u(x,0) and $u_t(x,0)$.

In a material with memory (such as certain polymers, suspensions, or
emulsions) the stress at a material point x and at time t depends on the
entire history of the strain at x. In 1874 Boltzmann [5] gave the following
linear constitutive law for small deformations in such materials

$$\sigma(x,t) = \beta\epsilon(x,t) + \int_0^\infty m(s)[\epsilon(x,t) - \epsilon(x,t-s)]ds, \ x \in B, \ -\infty < t < \infty \ . \tag{2.3}$$

In (2.3) β > 0 is a given constant and $m : (0,\infty) \to R$ is a given positive,
smooth, nonincreasing function. We limit our discussion to the situation in
which $m \in L^1(0,\infty)$, and we distinguish two cases:

$$\text{(i)} \ 0 < m(0) < \infty, \quad \text{(ii)} \ m(0^+) = +\infty \ . \tag{2.4}$$

The function m is called a memory function. The fact that m > 0 and non-
increasing on (0,∞) means that the stress "relaxes" as t increases and the
memory term in (2.3) fades: deformations which occurred in the distant past
have less influence on the present value of the stress than those which
occurred in the recent past. In the applied literature m is often assumed
to be a finite linear combination of decaying exponentials with positive
coefficients (these expressions result from least squares approximations to

22

experimental data). Such restrictions are neither desirable nor necessary. Moreover, kinetic theories for chain molecules [15,46,53] and certain experiments [32,28] suggest that there are materials for which m is singular as in (2.4)(ii), $m(t) \sim t^{\alpha-1}$ as $t \to 0^+$, $0 < \alpha < 1$, m is positive, nonincreasing on $0 < t < \infty$, and m decays rapidly at infinity. Stronger power singularities at zero ($\alpha < 0$) are also possible, but the resulting mathematical theory for nonlinear materials consistent with our objectives is incomplete at this time.

The assumption $m \in L^1(0,\infty)$ implies that (2.3) is equivalent to

$$\sigma(x,t) = c^2 \varepsilon(x,t) - \int_0^\infty m(s)\varepsilon(x,t-s)ds, \quad x \in B, \; -\infty < t < \infty \quad , \quad (2.5)$$

where $c^2 := \beta + \int_0^\infty m(s)ds > 0$ is a constant which measures the instantaneous response of stress to strain; $\beta > 0$ is the equilibrium stress modulus. If $\beta > 0$ the material acts like a solid, while if $\beta = 0$ it acts like a fluid.

A natural generalization of (2.5) to <u>nonlinear materials</u> is the constitutive relation

$$\sigma(x,t) = \phi(\varepsilon(x,t)) - \int_0^\infty m(s)\psi(\varepsilon(x,t-s))ds, \quad x \in B, \; -\infty < t < \infty \quad , \quad (2.6)$$

in which $\phi, \psi : R \to R$ are assigned, smooth material functions which satisfy

$$\phi(0) = \psi(0) = 0, \quad \phi'(0) > 0, \quad \psi'(0) > 0 \quad . \quad (2.7)$$

The memory function m is positive, nonincreasing and integrable on $(0,\infty)$ as above. In the static case $\varepsilon(x,t) = \bar\varepsilon(x)$, $\sigma(x,t) = \bar\sigma(x)$, (2.6) reduces to

$$\bar\sigma(x) = \phi(\bar\varepsilon(x)) - (\int_0^\infty m(s)ds)\psi(\bar\varepsilon(x)), \quad x \in B \quad .$$

A natural assumption, appropriate for viscoelastic solids and crucial in the analysis of global existence results (section 3), is to require that ϕ, ψ also satisfy

$$\phi'(0) - (\int_0^\infty m(s)ds)\psi'(0) > 0 \quad ; \quad (2.8)$$

(2.8) states that the equilibrium stress modulus is positive. The constitutive assumption (2.6) is a particular case of a "simple material" [8] which retains many important qualitative properties of more general material models; moreover, the analysis of the resulting equation of motion is relatively simple and complete.

The balance of linear momentum and (2.6) yield the equation of motion

$$u_{tt} = \phi(u_x)_x - \int_{-\infty}^t m(t-\tau)\psi(u_x(x,\tau))_x d\tau + f, \quad x \in B, \; -\infty < t < \infty \quad , \quad (2.9)$$

where f is a body force and where the change of variable $\tau := t-s$ was made in (2.6). The history of the motion is assumed to be known for $t < 0$ (the history may, but need not satisfy (2.9) for $t < 0$). An appropriate dynamic problem is to find a smooth function $u : B \times (-\infty,\infty) \to R$, satisfying (2.9) for $t > 0$, and such that

$$u(x,t) = \bar u(x,t) \quad , \quad x \in B, \; t < 0 \quad , \quad (2.10)$$

23

where the history $\overline{u} : B \times (-\infty, 0] \to R$ is a given smooth function; (2.9), (2.10) will be referred to as a history value problem. If B is bounded or semibounded compatible boundary conditions are adjoined to (2.9), (2.10). Compatibility of the boundary conditions with the smooth data f and $\overline{u}$ is imposed in order to preclude the propagation of singularities from the boundary into the interior.

If $m \equiv 0$, (2.9) reduces to the quasilinear wave equation (2.2). At the other extreme, if one formally sets $m = -\delta'$, where δ is the Dirac mass at the origin, then (2.9) reduces to the parabolic equation

$$u_{tt} = \psi(u_x)_{xt} + \phi(u_x)_x + f \quad ;$$

the term $\psi(u_x)_{xt}$ represents viscosity of Newtonian type if ψ is smooth and $\psi'(\cdot) > 0$. This equation possesses globally defined smooth solutions even if the data are large [1,34].

Our objective is to discuss the strength of the dissipative mechanism induced by the memory in (2.9) under physically reasonable assumptions by studying the existence and the decay or growth of classical solutions of the history value problem (2.9), (2.10). To motivate the mathematical results, we follow Coleman and Gurtin [6] in their penetrating study on the growth and decay of acceleration waves propagating into a one-dimensional viscoelastic material with memory at rest. An acceleration wave solution u is similar to a shock wave; the difference is that second rather than first derivatives of u experience a jump acrosss the wave front. To apply the results of [6] to (2.9), (2.10), we assume that ϕ, ψ are smooth, satisfy (2.7), $f \equiv 0$, $B = R$, and m is a smooth, regular kernel satisfying (2.4)(i). The wave front is a smooth curve $t = \gamma(x)$, $\gamma(0) = 0$, _and_ $u \equiv 0$ _for_ $t < \gamma(x)$. In [6] the problem of existence of acceleration waves is not discussed. Assuming that they do, an easy but tedious calculation shows that for (2.9) $t = \gamma(x)$ is a straight line, of slope $(\phi'(0))^{-\frac{1}{2}}$, meaning that such waves propagate with constant speed although (2.9) is nonlinear. Let the amplitude of the wave be $q(t) := [u_{tt}]$, where $[u_{tt}]$ is the jump in u_{tt} across the line $t = \gamma(x)$. It follows from the computations in [6] that q evolves in accordance with the Ricatti-Bernoulli equation

$$\frac{d}{dt} q = Aq^2 - Bq \quad , \quad q(0) = q_0 \quad , \tag{2.11}$$

where $\frac{d}{dt} = \frac{\partial}{\partial t} + c \frac{\partial}{\partial x}$, $c^2 = \phi'(0)$, represents differentiation along the wave front and where

$$A = \frac{-\phi''(0)}{2[\phi'(0)]^{3/2}} \quad , \quad B = \frac{m(0)\psi'(0)}{\phi'(0)} \quad .$$

Thus if $\phi''(0) < 0$ (similar results hold for $\phi''(0) > 0$), and $q_0 < B/A$, then every solution of (2.11) tends to zero as $t \to +\infty$. By contrast, if $q_0 > + B/A$, then $q(t) \to +\infty$ as $t \to T_0^-$, where $T_0 = \frac{1}{B} \log \frac{Aq_0}{Aq_0 + B} > 0$. The

24

corresponding jumps in u_{xt} and u_{xx} are given by $[u_{xt}] = -[\phi'(0)]^{1/2} q(t)$ and $[u_{xx}] = [\phi'(0)]^{-1} q(t)$.

This result suggests the following conjectures regarding smooth solutions of the history value problem (2.9), (2.10):

(i) The problem (2.9), (2.10) should have globally defined classical solutions if the history v and the forcing term f are sufficiently smooth and small in appropriate norms. Moreover, such solutions should decay.

(ii) The smooth solutions of (2.9), (2.10) should develop singularities in second derivatives in finite time if the smooth data are chosen sufficiently large.

As will be summarized in Section 3, conjecture (i) has been established rigorously by a number of authors in a number of physically important cases of (2.9), (2.10) for regular kernels $(m(0) < \infty)$, as well as for singular kernels $(m(0^+) = +\infty)$. Conjecture (ii) has only been established for regular kernels (see Section 4). Moreover, based on the discussion in Section 4, Remark 4.6, singular kernels m strengthen the dissipative mechanism of the memory in (2.9) which suggests the possibility that for appropriate classes of singular kernels, global smooth solutions will exist even if the data are arbitrarily large; this interesting question is open.

Most of the results described in Sections 3 and 4 for smooth kernels satisfying (2.4a) apply to more general one-dimensional viscoelastic models with fading memory, e.g. a model for a solid, K-BKZ material [29,2]

$$u_{tt} = \phi(u_x)_x + \int_{-\infty}^{t} m(t-\tau) h(u_x(x,t), u_x(x,\tau))_x \, d\tau + f, \qquad (2.12)$$

$$x \in B, \ -\infty < t < \infty \ .$$

Here ϕ, m, and f are as in (2.9), while $h : R \times R \to R$ is a smooth material function, $h(p,p) = 0$ and the partial derivatives of h satisfy appropriate sign conditions, at least at $(0,0)$. If $\phi \equiv 0$, (2.12) models a K-BKZ fluid. Under suitable assumptions, the energy method for proving existence results in Section 3 and the method of characteristics used to prove blow-up results of Section 4 yield similar results for this case as well. The energy method can also be applied to prove existence for certain multidimensional viscoelastic problems with fading memory (e.g. [13, Sec. 4], [30]). However, to our knowledge, the existence results described in Section 3 for singular kernels satisfying (2.4(ii)) depend crucially on the special form of equation (2.9).

3. Existence of Classical Solutions. For discussion of the mathematical results it is convenient to renormalize the memory function m. Define the relaxation function a by

$$a(t) := \int_{t}^{\infty} m(s)ds, \ 0 < t < \infty \ ; \qquad (3.1)$$

observe that if m is smooth, positive, decreasing and integrable on $[0,\infty)$ then $a'(t) = -m(t)$ and

$$a \text{ is smooth, positive, decreasing and convex on } (0,\infty) \ . \qquad (3.2)$$

Analogous to (2.4) we distinguish two classes of kernels a:

$$(i) \quad 0 < -a'(0^+) < \infty \quad , \quad (ii) \quad -a'(0^+) = +\infty \quad . \tag{3.3}$$

Other normalizations of the memory m are possible; for example, the relaxation function

$$G(t) := \phi'(0) - a(0)\psi'(0) + a(t)\psi'(0) \quad , \quad 0 < t < \infty \quad , \tag{3.4}$$

where ϕ, ψ are the material functions in (2.6), is consistent with the applied literature. Observe that $G(\infty) = \phi'(0) - a(0)\psi'(0)$ and $G(0) = \phi'(0)$.

Returning to the history value problem (2.9), (2.10), let the history $\bar{u}$ be identically zero for $t < 0$. One then seeks a solution of the initial value problem

$$u_{tt} = \phi(u_x)_x + \int_0^t a'(t-\tau)\psi(u_x(x,\tau))_x d\tau + f, \quad x \in B, \ t > 0 \quad , \tag{3.5}$$

$$u(x,0) = u_0(x), \ u_t(x,0) = u_1(x), \ x \in P \quad , \tag{3.6}$$

together with suitable and compatible boundary conditions if B is not R. If the history $\bar{u}$ is not zero for $t < 0$, the part of the integral in (2.9) on $(-\infty,0)$ is incorporated in f.

Global Existence of Classical Solutions. We next discuss global existence and asymptotic behaviour for the Cauchy problem (3.5), (3.6) with B = R, for smooth, small data, and for regular kernels a satisfying (3.2), (3.3)(i). To simplify the exposition, we make the hypothesis

$$a \in C^3[0,\infty), \ (-1)^k a^{(k)}(t) > 0 \ (0 < t < \infty; \ k = 0,1,2,3) \quad ,$$

$$\tag{3.7}$$

$$a' \not\equiv 0, \ \text{and} \ \int_0^\infty t a(t) dt < \infty \quad .$$

The results hold under assumptions on a considerably weaker than (3.7). The interested reader is referred to [13], [22], [24], and the survey paper [23] for the generalizations. The essential point is that kernel a satisfies a, a', a" $\in L^1(0,\infty)$, the moment condition in (3.7), and is "strongly positive" on $[0,\infty)$. The result for B bounded [13] is somewhat simpler than for the Cauchy problem (3.5), (3.6); in particular, the moment condition in (3.7) is only needed for the Cauchy problem (see remarks following Theorem 3.1 and the outline of its proof).

Concerning ϕ, ψ assume

$$\phi, \psi \in C^3(\mathbb{R}), \ \phi(0) = \psi(0) = 0,$$

$$\tag{3.8}$$

$$\phi'(0) > 0, \ \psi'(0) > 0, \ \phi'(0) - a(0)\psi'(0) > 0 \ ;$$

the latter is the analogue of (2.8) in the present normalization. Assume that

$$(i) \ f, \ f_x, \ f_t \in C([0,\infty); \ L^2(\mathbb{R})) \cap L^\infty([0,\infty); \ L^2(\mathbb{R})) \ \text{and}$$

$$\tag{3.9}$$

$$(ii) \ f \in L^1([0,\infty); \ L^2(\mathbb{R})); \ f_x, \ f_t, \ f_{xt} \in L^2([0,\infty); \ L^2(\mathbb{R})) \quad ,$$

26

and let u_0, u_1 satisfy

$$u_0 \in L^2_{\ell oc}(\mathbf{R}), \quad \text{and} \quad u_0', u_1 \in H^2(\mathbf{R}) \quad . \tag{3.10}$$

To measure the size of the data define the quantities

$$U_0(u_0, u_1) := \int_{-\infty}^{\infty} \{u_0'^2 + u_0''^2 + u_0'''^2 + u_1^2 + u_1'^2 + u_1''^2\}(x)dx, \quad \text{and} \tag{3.11}$$

$$F(f) := \sup_{t>0} \int_{-\infty}^{\infty} \{f^2 + f_x^2 + f_t^2\}(x,t)dx + \left(\int_0^{\infty}\left(\int_{-\infty}^{\infty} f^2(x,t)dx\right)^{1/2} dt\right)^2 \tag{3.12}$$

$$+ \int_0^{\infty} \int_{-\infty}^{\infty} \{f_x^2 + f_t^2 + f_{xt}^2\}(x,t)dxdt \quad .$$

The following result is a special case of Theorem 1.1 of [24].

<u>Theorem 3.1.</u> <u>Let assumptions (3.7) – (3.10) be satisfies. There exists a</u> <u>constant</u> $\mu > 0$ <u>such that for each</u> u_0, u_1, f <u>satisfying</u>

$$U(u_0, u_1) + F(f) < \mu^2 \quad , \tag{3.13}$$

<u>the Cauchy problem (3.5), (3.6) has a unique solution</u> $u \in C^2(\mathbf{R} \times [0,\infty))$, <u>and</u>

$$u_x, u_t, u_{xx}, \ldots, u_{ttt} \in C([0,\infty); L^2(\mathbf{R})) \cap L^{\infty}([0,\infty); L^2(\mathbf{R})) \quad . \tag{3.14}$$

<u>Moreover,</u>

$$u_{xx}, u_{xt}, \ldots, u_{ttt} \in L^2([0,\infty); L^2(\mathbf{R}) \quad , \tag{3.15}$$

$$u_{xx}, u_{xt}, u_{tt} \to 0 \quad \text{in} \quad L^2(\mathbf{R}) \quad \underline{\text{as}} \quad t \to \infty \quad , \tag{3.16}$$

$$u_x, u_t, u_{xx}, u_{xt}, u_{tt} \to 0 \quad \underline{\text{uniformly on}} \quad \mathbf{R} \quad \underline{\text{as}} \quad t \to \infty \quad . \tag{3.17}$$

A similar result holds for the history value problem (2.9), (2.10) with $B = \mathbf{R}$. The special case $a(t) = \alpha e^{-\lambda t}$, $\alpha > 0$, $\lambda > 0$, studied by Greenberg [18] for B bounded, is carried out in [23] in the more complicated case when $B = \mathbf{R}$.

Remark 3.2. Theorem 3.1 is a generalization of Theorems 1.1 and 4.1 of [13] establishing small-data global existence results for analogous initial boundary value problems corresponding to motions of **bounded viscoelastic bodies**; Neumann, Dirichlet and mixed boundary conditions are treated. The principal difficulty in proving Theorem 3.1 is that various Poincaré inequalities, not applicable to (3.5), (3.6) when $B = \mathbf{R}$, are used in an essential way in [13] to establish an a priori estimate similar to (3.26) from (3.32) (see outline of proof following Proposition 3.4); the estimate (3.26) is essential for completing the proof. The reader is referred to Hrusa [22] for a discussion of general history value problems on a bounded interval. Although technically extremely complicated, the generalization of the results in [22] to the Cauchy problem is relatively straightforward.

Remark 3.3. If $\psi \equiv \phi$ equations of the form (3.5) have been studied by MacCamy [35], Dafermos and the author [12], and Staffans [49] for bounded and unbounded bodies. If $\psi \equiv \phi$, (3.5) admits certain estimates which do not carry over to the general case $\psi \neq \phi$ (see [23]); there does not appear to be any physical motivation for the restriction $\psi \equiv \phi$ for solids.

Outline of the proof of Theorem 3.1. An essential ingredient of any global result is an appropriate local existence theorem. For regular kernels a satisfying (3.2),(3.3)(i), the idea is to iterate the sequence of linear problems which treat the memory as a lower-order perturbation:

$$u_{tt} = \phi'(w_x)u_{xx} + \int_0^t a'(t-\tau)\psi(w_x(x,\tau))_x d\tau + f, \ x \in \mathbf{R}, \ 0 \leqslant t \leqslant T \ , \qquad (3.18)$$

where $T > 0$, u satisfies the initial conditions (3.6), and where w is an element of a suitably chosen function space X. By using fairly standard enery estimates deduced from (3.18), requiring only very simple estimates of the convolution term which do not use any sign information on the memory, it is shown that the mapping S which carries w into a solution of (3.18) has a unique fixed point for $T > 0$ sufficiently small. The proof is almost identical with that of Theorem 2.1 of [13]. The only significant difference is that the proof in [13] is for $x \in [0,1]$ with Neumann boundary conditions satisfied at $x = 0$ and $x = 1$; thus the Poincaré inequality enables one to deduce estimates for lower order derivatives of u in $L^\infty([0,T]; L^2(0,1))$ from higher order derivative estimates. As far as local existence is concerned when $B = \mathbf{R}$, this causes no serious difficulties. One simply expresses the lower order derivatives of the solution in terms of initial conditions and time integrals of the higher order derivatives, yielding time dependent bounds which, however, cannot be used for obtaining global estimates. The result is:

Proposition 3.4. Let $a, a', a'' \in L^1_{loc}[0,\infty)$ and assume that $\phi, \psi \in C^3(\mathbf{R})$, $\phi'(0) > 0$, and that there exists a number $\underline{\phi}$ such that

$$\phi'(\xi) \geqslant \underline{\phi} \ \text{for every} \ \xi \in \mathbf{R} \ . \qquad (3.19)$$

Concerning the data, let u_0, u_1 satisfy (3.10), f satisfy (3.9)(i) and assume that $f_{xt} \in L^1_{loc}([0,\infty); L^2(\mathbf{R}))$. Then the Cauchy problem (3.5), (3.6) has a unique solution u defined on a maximal time interval $[0,T_0)$ satisfying

$$u_x, \ u_t, \ u_{xx}, \ u_{xt}, \ u_{tt}, \ u_{xxx}, \ u_{xxt}, \ u_{xtt}, \ u_{ttt} \in C([0,T_0); L^2(\mathbf{R})) \ . \ (3.20)$$

Moreover, if

$$\sup_{t \in [0,T_0)} \int_{-\infty}^{\infty} \{u_x^2 + u_t^2 + \cdots + u_{xtt}^2 + u_{ttt}^2\}(x,t)dx < \infty \ , \qquad (3.21)$$

then $T_0 = +\infty$. By Sobolev embedding $u \in C^2(\mathbf{R} \times [0,T_0))$.

In outline, the proof of the global result then proceeds as follows. Define the equilibrium stress χ by

$$\chi(\xi) := \phi(\xi) - a(0)\psi(\xi), \ \forall \xi \in \mathbf{R} \ ; \qquad (3.22)$$

28

observe that $\chi \in C^3(\mathbb{R})$ and that $\chi'(0) > 0$ (by 3.8). Choose a sufficiently small number $\delta > 0$ and modify ϕ, ψ, and χ outside $[-\delta, \delta]$ such that ϕ'', ψ'', χ'' vanish outside $[-2\delta, 2\delta]$, and choose positive constants $\underline{\phi}$, $\underline{\psi}$, $\underline{\chi}$ such that

$$\phi'(\xi) \geq \underline{\phi}, \quad \psi'(\xi) \geq \underline{\psi}, \quad \chi'(\xi) \geq \underline{\chi} \quad \forall \xi \in \mathbb{R} \quad . \tag{3.23}$$

It is shown a posteriori that $|u_x(x,t)| < \delta$ for all $x \in R$, $t \geq 0$. By Proposition 3.4 the Cauchy problem (3.5), (3.6), $B = R$ has a unique solution u on a maximal interval $[0, T_0)$. The objective is to show that if (3.13) holds with $\mu > 0$ sufficiently small, then (3.21) is bounded independent of T_0; a standard continuation procedure implies $T_0 = +\infty$. Define

$$E(t) := \max_{s \in [0,t]} \int_{-\infty}^{\infty} \{u_t^2 + u_x^2 + \cdots + u_{ttt}^2\}(x,s)dx$$

$$\tag{3.24}$$

$$+ \int_0^t \int_{-\infty}^{\infty} \{u_{xx}^2 + u_{xt}^2 + \cdots + u_{ttt}^2\}(x,s)dxds \quad ,$$

where $\cdots$ represent the sum of the second and third derivatives not explicitly written down. It is shown that if (3.13) holds for $\mu > 0$ sufficiently small, then $E(t)$ is bounded. For this purpose define

$$\nu(t) := \sup_{\substack{x \in R \\ s \in [0,t]}} \{u_x^2 + u_{xx}^2 + u_{xt}^2\}^{1/2}(x,s), \quad \forall t \in [0, T_0) \quad . \tag{3.25}$$

To prove the result one establishes the following key estimate

$$E(t) \leq \Gamma(U_0(u_0, u_1) + F(f)) + \Gamma(\nu(t) + \nu^3(t))E(t), \quad 0 \leq t \leq T_0 \quad , \tag{3.26}$$

where here and below Γ is a generic constant, possibly large, independent of u_0, u_1, f, and T_0. We shall comment below only briefly how this is accomplished.

Once (3.26) is established, the conclusions of Theorem 3.1 are obtained as follows. Choose $\bar{E}$, $\mu > 0$ such that

$$\bar{E} \leq \delta^2, \quad \Gamma\{(2\bar{E})^{1/2} + (2\bar{E})^{3/2}\} \leq \frac{1}{2}, \quad \Gamma\mu^2 \leq \frac{1}{4}\bar{E} \quad . \tag{3.27}$$

Select the data u_0, u_1, f such that (3.13) holds for μ chosen in accordance with (3.27). The Sobolev embedding theorem implies that

$$\nu(t) \leq (2E(t))^{1/2} \quad \forall t \in [0, T_0) \quad . \tag{3.28}$$

Therefore, it follows from (3.26), (3.27), (3.28) that for any $t \in [0, T_0)$ with $E(t) \leq \bar{E}$, we actually have $E(t) \leq \frac{1}{2}\bar{E}$. By continuity $E(t) \leq \frac{1}{2}\bar{E}$, $\forall t \in [0, T_0)$, provided $E(0) \leq \frac{1}{2}\bar{E}$; the latter is insured by choosing μ^2 smaller if necessary so that (3.13) will imply $E(0) \leq \frac{1}{2}\bar{E}$. Then $E(t) \leq \frac{1}{2}\bar{E}$, $\forall t \in [0, T_0)$, and (3.24), Proposition 3.1, and a standard continuation method yield $T_0 = +\infty$. One also has that (3.14), (3.15) hold, and conclusions (3.16), (3.17) follow by standard embedding inequalities. Moreover, (3.25), (3.27), (3.28) yield

$$|u_x(x,t)| \leq \nu(t) \leq (2E(t))^{1/2} \leq (\bar{E})^{1/2} \leq \delta, \quad \forall x \in R, \ t \in [0, \infty)$$

and the proof is complete.

Establishing the estimate (3.26) is lengthy, delicate, and relies on the correct sign of the memory [under assumption (3.7) or certain generalizations]. The energy method, combined with relevant properties of Volterra operators and their resolvents, is employed. The estimates of derivatives of u appearing in (3.24) are deduced from energy identities obtained directly from (3.5), (3.6), and from the equation equivalent to (3.5):

$$u_{tt} = \chi(u_x)_x + \int_0^t a(t-\tau)\psi(u_x)_{xt}(x,\tau)d\tau$$

$$+ a(t)\psi(u_0(x)_x)_x + f \quad (x \in \mathbb{R}, \ 0 < t < T) \ , \tag{3.29}$$

where $T < T_0$; (3.29) is obtained from (3.5), (3.6) by an integration by parts and use of (3.22). Useful identities for derivatives of u can only be obtained by multiplying the equations by quantities which make it possible to estimate the memory terms. A crucial role is played by the "quadratic integral form"

$$Q(w,t,b) := \int_0^t \int_{-\infty}^\infty w(x,s) \int_0^s b(s-\tau)w(x,\tau)d\tau dx ds, \ t > 0 \ ,$$

defined for $b \in L^1_{loc}[0,\infty)$ and for every $w \in C([0,t]; L^2(\mathbb{R}))$. In the first energy identity, which is obtained by multiplying (3.29) by $\psi(u_x)_{xt}$ and integrating the equation over $\mathbb{R} \times [0,T]$, Q arises with $w = \psi(u_x)_{xt}$ and $b = a$. It is an important fact that kernels a satisfying (3.7) (indeed much weaker assumptions) are positive definite on $[0,\infty)$. To obtain the second energy identity, one needs to take the forward time-difference of (3.29) and integrate the resulting equation over $\mathbb{R} \times [0,T]$. To estimate the relevant derivatives of u from a combination of the first two identities one needs the following technical estimate: It is shown in [24; Lemma 2.5] that if a satisfies (3.7), there exists a constant $\kappa > 0$ such that

$$\int_0^t \int_{-\infty}^\infty w_t^2(x,t)dxdt < \kappa \int_{-\infty}^\infty w_t(x,0)dx + \kappa Q(w_t,t,a)$$

$$+ \kappa \liminf_{h\downarrow 0} \frac{1}{h^2} Q(\Delta_h w_t,t,a) \ , \quad \forall t \in [0,T] \ , \tag{3.30}$$

where $w \in C^1([0,T]; L^2(\mathbb{R})) \ \forall \ T > 0$, and where the forward difference operator $\Delta_h w$ is defined by $\Delta_h w(x,t) := w(x,t+h) - w(x,t)$. In the application of (3.30), $w = \psi(u_x)_{xt}$ and the forward difference operator Δ_h is applied to equations (3.29). The proof of (3.30) also makes use of a result of Staffans ([49, Lemma 4.2]). Using the two energy identities, and (3.30), it is relatively straightforward to estimate all of the terms and arrive at:

$$\int_{-\infty}^\infty \{u_{xx}^2 + u_{xt}^2 + u_{xxt}^2 + u_{xtt}^2\}(x,t)dx + \int_0^t \int_{-\infty}^\infty u_{xxt}^2(x,s)dxds$$

$$< \Gamma(U_0+F) + \Gamma(\nu(t)+\nu^3(t))E(t) + \Gamma(\sqrt{U_0} + \sqrt{F})\sqrt{E(t)}, \ \forall \ t \in [0,T] \ . \tag{3.31}$$

Estimates of $\int_{-\infty}^\infty u_{tt}^2(x,t)dx, \int_{-\infty}^\infty u_{ttt}^2(x,t)dx, \int_0^t \int_{-\infty}^\infty u_{ttt}^2(x,\tau)dxdt, \ \forall \ t \in [0,T]$

30

in terms of the right side of (3.31) are obtained from (3.5). A bound for $\int_0^t \int_{-\infty}^{\infty} u^2_{xtt}(x,s)dxds$ can then be obtained by interpolation. Using the fact that a certain resolvent kernel of a' in (3.5) is in $L^1[0,\infty)$, Lemma 3.2 of [13] makes it possible to estimate $\int_{-\infty}^{\infty} u^2_{xxx}(x,t)dx$ and $\int_0^t \int_{-\infty}^{\infty} u^2_{xxx}(x,s)dxds$. Combining these with (3.31) yields the estimate

$$\int_{-\infty}^{\infty} \{u^2_{xx} + u^2_{xt} + u^2_{tt} + u^2_{xxx} + u^2_{xxt} + u^2_{xtt} + u^2_{ttt}\}(x,t)dx \qquad (3.32)$$

$$+ \int_0^t \int_{-\infty}^{\infty} \{u^2_{xxx} + u^2_{xxt} + u^2_{xtt} + u^2_{ttt}\}(x,s)ds$$

$$\leq \Gamma(U_0 + F) + \Gamma(v(t) + v^3(t))E(t)$$

$$+ \Gamma(\sqrt{U_0} + \sqrt{F}) \sqrt{E(t)}, \ \forall \ t \in [0,T] \ .$$

The estimate (3.32) is implicit in the argument of [13]. It should be observed that for problems on bounded intervals (Remark 3.2), it is a simple matter to apply the Poincaré inequality to deduce the remaining estimates of derivatives of u appearing in (3.24) and arrive at the final estimate (3.26) directly from (3.32). However, to accomplish this task for (3.5), (3.6) when $B = R$ is quite tricky and involves additional properties of Volterra operators and certain other of their resolvents. The reader is referred to Lemmas 2.3 and 2.4, as well as the argument on pages 405-410 of [24] for details. This part of the proof makes essential use of the assumption $a'' \in L^1[0,\infty)$ which is automatic when a satisfies (3.7), but cannot be satisfied by singular kernels.

For singular kernels satisfying (3.2) and (3.3)(ii), it is simpler to restrict the analysis to the history value problem, (2.9), (2.10), with a defined by (3.1), in which that history $\bar{u}$ satisfies the equation (and the boundary conditions if B is bounded). This ensures that the compatibility conditions between the history and boundary data, as well as compatibility conditions between the derivatives of the history and the solution for $t > 0$ are satisfied. If u is a smooth solution of (2.9) and the kernel a is singular, the integral in (2.9) is also a smooth function, but the integrals $\int_{-\infty}^0$ and $\int_0^t$ have singularities at $t = 0$ which cancel. Thus if formulated as an initial value problem the results would involve a singular forcing term. For reasons explained below, global existence results for singular kernels only hold for B bounded.

The principal difficulty when dealing with singular kernels is establishing a suitable local existence result. In Proposition 3.4 for regular kernels no hypothesis is made concerning the sign of the memory and the size of the data. In the proof the memory is treated as a perturbation of the elastic term $\phi(u_x)_x$ in (3.5). However, the proof makes crucial use of the hypothesis $a'' \in L^1_{loc}(0,\infty)$ which rules out singular kernels a satisfying (3.2), (3.3)(ii).

Hrusa and Renardy [25, Theorem 4.1] recently obtained an elegant extension of Proposition 3.4 for such singular kernels. They consider the history value problem with the history satisfying the equation and the boundary conditions for $t < 0$. The singular kernel a satisfies the

assumptions

$$a, a' \in L^1(0,\infty); \quad a(t) > 0, \quad a'(t) < 0, \quad a''(t) > 0, \quad 0 < t < \infty \qquad (3.33)$$

in the sense of measures, and a" is not a purely singular measure; a certain assumption on the Laplace transform of a is imposed in order to guarantee that the third derivatives of u are continuous with values in $L^2(0,1)$. The material function ψ is also required to satisfy $\psi'(0) > 0$, and the technical assumptions regarding the forcing function f are strengthened. The sign of the memory now plays a crucial role in the local analysis in which one iterates a sequence of linear integrodifferential equations (compare with (3.18))

$$u_{tt} = \phi'(w_x)u_{xx} + \int_{-\infty}^{t} a'(t-\tau)\psi'(w_x)u_{xx}(x,\tau)d\tau + f \qquad (3.34)$$

where $u(x,t) = \bar{u}(x,t)$ for $t < 0$, and where w is an element of an apropriately chosen function space. The singular kernel a satisfying (3.33) is replaced in (3.34) by regular kernels a_δ defined by

$$a_\delta(t) := \int_{-\delta}^{\delta} p_\delta(\tau)a(t+\delta+\tau)d\tau, \quad 0 < t < \infty, \quad \delta > 0 \quad ,$$

where p_δ is a standard mollifier supported in $[-\delta/2,\delta/2]$. The analysis with singular kernels is far more complicated because $a'' \notin L^1_{loc}[0,\infty)$, and $|a_\delta''|_{L^1}$ does not necessarily remain bounded as $\delta \downarrow 0$. The energy estimates are also considerably more delicate and to obtain them certain technical lemmas concerning Volterra operators with kernels a satisfying (3.33) are required (such kernels are known to be strongly positive definite [43]). It is first shown that each linear problem (3.34) has a unique solution having the required regularity by justifying passage to the limit as $\delta \downarrow 0$. Then a contraction mapping argument for (3.34) is used in [25] to obtain the analogue of Proposition 3.4 for w belonging to an appropriate function space. The proof in [25] is carried out for B = [0,1] with Dirichlet boundary conditions satisfied at x = 0 and x = 1; it is straightforward to obtain a similar local result for B = R, because the local existence proof in [25] avoids the use of Poincaré inequalities.

Using their local result, Hrusa and Renardy then obtain an analogue of Theorem 3.1 for the history value problem (2.9), (2.10) and the (singular) kernel a, defined by (3.1), satisfying (3.33) <u>on bounded intervals</u>. They impose the requirement that the history and the solution satisfy Dirichlet boundary conditions at x = 0 and x = 1 and that the history and forcing term be suitably small. Their result ([25, Theorem 5.1]) is then a simple extension of the proof of [13, Theorem 1.1] involving the modification of only one estimate in [13]; the modification uses a refinement of Lemma 4.2 in [49], because $a'' \notin L^1[0,\infty)$ whenever a is singular. The fact that $a'' \notin L^1[0,\infty)$ makes it difficult to prove Theorem 3.1 for singular kernels using the analysis in [25]. It is a challenging open problem to prove such a result for singular kernels on all of space.

4. <u>Development of Singularities and Related Problems</u>. In this section we consider the Cauchy problem (3.5), (3.6) for regular kernels a, and we discuss the development of singularities in smooth solutions in finite time

33

for smooth but large data by using the method of characteristics. To avoid
technical complications we assume that the forcing term $f \equiv 0$ in (3.5), and
we study

$$u_{tt} = \phi(u_x)_x + a' * \psi(u_x)_x, \quad x \in R, \quad t > 0 \quad , \tag{4.1}$$

$$u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x), \quad x \in R \quad , \tag{4.2}$$

where * denotes the time convolution on [0,t]. The following result was
recently established by M. Renardy and the author [42], and independently by
Dafermos [10] for general memory functionals using a somewhat different
proof. The result can also be established by extending techniques of F. John
[27] to quasilinear, first-order hyperbolic systems with lower order source
terms; however, the approach outlined below is more direct.

Theorem 4.1. Let $\phi, \psi \in C^3(\mathbb{R})$ let ϕ satisfy (3.19) and let a be
smooth with $a, a', a'' \in L^1_{loc}[0,\infty)$. In addition, let $\phi''(0) \neq 0$. Then for
every $T_1 > 0$, there exists initial data $u_0, u_1 \in C^2(R) \cap L^\infty(R)$ such that
the maximal interval of existence of the smooth solution u of the Cauchy
problems (4.1), (4.2) cannot exceed T_1. More precisely, if $\sup_{x \in R} |u_0'(x)|$ and
$\sup_{x \in R} |u_1(x)|$ are sufficiently small, while $u_0''(x)$ and $u_1'(x)$ are
sufficiently large (with appropriate signs), then there exists a number
$t^* < T_1$ such that

$$\sup_{R \times [0,t^*)} \{|u_{xx}(x,t)| + |u_{xt}(x,t)|\} = \infty \quad , \tag{4.3}$$

while

$$\sup_{R \times [0,t^*)} \{|u_x(x,t)| + |u_t(x,t)|\} < \infty \quad , \tag{4.4}$$

For the special case $\psi \equiv \phi$, Hattori [21] has shown that if $\phi'' \neq 0$ and
if the body B is bounded, then there exist data u_0, u_1 such that the
initial-boundary value problem (consisting of (4.1), (4.2) and compatible
Dirichlet boundary conditions) does not have a globally defined smooth
solution. However, his method does not enable him to characterize the data.
Ramaha [45] has recently obtained a blow-up result when $\psi \equiv \phi$.

For first-order model problems with fading memory, blow-up results
similar to Theorem 4.1 have been obtained by a number of authors ([38], [36],
[9]) by the method of characteristics. Existence of classical solutions for
small data for such models is discussed in [41]. The elegant method of
Dafermos [9] avoids use of characteristics; instead a maximum principle is
obtained and used.

Remark 4.2. The reader should observe that in Theorem 4.1 only the additional
hypothesis $\phi''(0) \neq 0$ is added to the assumptions guaranteeing the existence
of a local smooth solution of (4.1), (4.2) (Proposition 3.4). No sign
information on the kernel a is required. Assumption (3.19) is not
restrictive because it is shown that the supremum in (4.4) is in fact small.

The proof of Theorem 4.1 generalizes the approach of Lax [33] using the
method of characteristics and generalized Riemann invariants. We transform
(4.1), (4.2) to an equivalent first-order system as follows. Let $w = u_x$,

34

$v = u_t$, define

$$\sigma := \phi(w) - z \quad , \quad z := -a'*\psi(w) \quad , \tag{4.5}$$

and observe that σ is the stress-strain functionals (2.6). Since $\phi'(\cdot) > 0$, equation (4.5) can be solved for w, $w = \phi^{-1}(\sigma+z) := g(\sigma,z)$, and g is a smooth function on $R \times R$. As long as the solution u of (4.1), (4.2) remains smooth, (4.1), (4.2) is equivalent to the system

$$v_t = \sigma_x$$

$$\sigma_t = C^2(\sigma,z)v_x + a'(0)\psi(g(\sigma,z)) + a''*\psi(g(\sigma,z)) \quad , \tag{4.6}$$

$$z_t = -a'(0)\psi(g(\sigma,z)) - a''*\psi(g(\sigma,z)) \quad ,$$

$$v(x,0) = u_1(x), \; \sigma(x,0) = \phi(u_0^{\flat}(x)), \; z(x,0) \equiv 0 \quad , \tag{4.7}$$

where the wave speed $C(\sigma,z) := [\phi'(g(\sigma,z))]^{1/2}$ is a smooth function. The system (4.6) is hyperbolic with eigenvalues $C, -C, 0$. We define generalized Riemann invariants r, s by

$$r = r(v,\sigma,z) := v + \Phi(\sigma,z); \; s = s(v,\sigma,z) := v - \Phi(\sigma,z); \; \Phi(\sigma,z) := \int^{\sigma} \frac{d\xi}{C(\xi,z)} \quad .$$

Thus $v = \dfrac{r+s}{2}$, $\Phi = \dfrac{r-s}{2}$; the correspondence is smoothly invertible because $\Phi_\sigma = C^{-1} > 0$. Observe that if $a' \equiv 0$ in (4.1), $z \equiv 0$ and g, C are independent of z. In this situation r and s reduce to the Riemann invariants for the system

$$v_t = \sigma_x, \; \sigma_t = \phi'(\phi^{-1}(\sigma))v_x$$

which can be transformed to the quasilinear wave equation. In the proof r, s, z are introduced as dependent variables and (4.6) is replaced by an equivalent system obtained by differentiating r, s, z along the characteristics $C, -C, 0$ respectively. One then differentiates the quantities

$$\rho := v_x + \frac{\sigma_x}{C(\sigma,z)} \quad , \quad \tau := v_x - \frac{\sigma_x}{C(\sigma,z)} \quad , \text{ and } z_x$$

along the $C, -C, 0$ characteristics respectively (observe that if $a' \equiv 0$, $\rho = r_x, \tau = s_x$). It is shown (see [42] for details) that to leading order the characteristic derivatives of $\sqrt{C}\rho, \sqrt{C}\tau$ satisfy a coupled system of Ricatti equations in ρ and τ with coefficients which are smooth functions of r, s, z. The differential equation for z_x is linear in ρ, τ, z_x, and it is shown that z_x grows at most logarithmically. Blow-up in finite time is established by showing that r, s, z remain in a neighborhood U of zero up to the blow-up time, if they are small initially (i.e. if $\sup_{R}\{|v(x,0)| + |\sigma(x,0)|\}$ is small), while $v'(x,0)$ and $\sigma'(x,0)$ are sufficiently large (with appropriate signs). Moreover, the hypothesis $\phi''(0) \neq 0$ provides upper and lower nonzero bounds for the coefficients ρ^2 and τ^2 in the Ricatti

equations when r, s, z are in U.

Remark 4.3. A physical interpretation of conclusions (4.3), (4.4), coupled with examples of Coleman, Gurtin, and Herrera [7], is that the strain remains bounded but its first derivatives become infinite as $t \to t_-^*$. Thus Theorem 4.1 suggests, but does not prove, the development of a shock front in finite time.

Remark 4.4. Certain models for shearing flows of viscoelastic fluids can be analyzed by the technique of Theorem 4.1. With $v(x,t)$ denoting the velocity of the fluid in simple shear, Slemrod [48] studies the problem

$$v_t = a*\phi(v_x)_x, \; x \in R, \; t > 0$$

$$v(x,0) = v_0(x) \; , \; x \in R$$

(4.8)

in the special case $a(t) = e^{-t}$. Differentiation of the equation leads to a Cauchy problem of the form (4.1), (4.2). Global existence for smooth, small data follows from [12, Theorem 4.1]; see also Remark 3.3. Development of singularities for large data is an easy application of Theorem 4.1 above. Other popular models for viscoelastic fluids can be discussed by a similar analysis. Slemrod [47] and Gripenberg [20] established similar results for a different model of shearing flows for a viscoelastic fluid. If $a = e^{-t}$, (4.8) as well as the problem studied in [47], can be transformed to the quasi-linear wave equation with linear frictional damping for which finite time blow-up for large data can be established by the method of Lax [33].

We close this section by discussing a number of open problems.

Remark 4.5. The techniques of proof of Theorem 4.1 and that of [10] depend crucially on the hypothesis $\phi''(0) \neq 0$. The physically important situation $\phi''(0) = 0$, permitted in the finite time blow-up result for the quasilinear wave equation (2.2) (with $f \equiv 0$) in [37], constitutes an interesting open problem for (4.1), (4.2).

Remark 4.6. Singular kernels a satisfying (3.2) and (3.3)(ii) $-a'(0^+) = +\infty$ violate the hypothesis $a'' \in L^1_{loc}[0,\infty)$ which is crucial to the technique of proof of Theorem 4.1 and that of the similar result in [10]. Indeed, there is strong evidence based on the following arguments, that there may exist singular kernels a such that (4.1) would have globally defined smooth solutions, even if the data are arbitrarily large. These arguments suggest that singular kernels strengthen the dissipation induced by the memory. Thus far it has not been possible to resolve this important open problem.

First, for smooth kernels with $-a'(0^+)$ finite, it follows from (2.11) and the definition of the constant B that the diameter of the set of points $q_0 > B/A$ for which $q(t) \to +\infty$ in finite time shrinks as $m(0) = -a'(0^+) > 0$ is increased. However, the derivation of (2.11) rests on the assumption that $m(0) = -a'(0^+)$ remains finite. Second, there are interesting results of Hrusa and Renardy [26] in their analysis of wave propagation in linear visco-elasticity. They study the linear history value problem (2.9), (2.10) with $\phi'(\cdot) \equiv c^2 = \beta + \int_0^\infty m(\tau)d\tau$ and $\psi'(\cdot) \equiv 1$, $u(x,t) \equiv 0, t < 0, B = R$, and they adjoin step jump initial data $u(x,0)$, $u_t(x,0), x \in R$. They prove that if the memory m is smooth on $[0,\infty)$, the

solution has discontinuities propagating along characteristics of the linear wave equation $u_{tt} = c^2 u_{xx}$ and a stationary discontinuity of higher order at the initial step-jumps. For singular memory kernels the propagating waves are smoothed out. The degree of smoothing increases as the kernel becomes more singular; the stationary discontinuities remain.

Remark 4.7. There is numerical evidence concerning the development of singularities in finite time for regular kernels a and large smooth data. Markowich and Renardy [39] used the Lax-Wendroff method to discretize the hyperbolic part in (4.1) and the trapezoidal rule to discretize the integral. They show that the method is second-order convergent and stable on any finite time interval on which smooth solutions exist. For spatially periodic and small Cauchy data, and for kernels a which are finite sums of decaying exponentials, they prove second order convergence on $[0,\infty)$. They also carry out numerical experiments in the special case $\psi \equiv \phi$ which exhibit the formation of a singularity in finite time for particularly chosen ϕ, a, and suitably large u_0 and u_1. Their numerical solution exhibits but does not prove the formation of shock fronts in u_x and u_t at the critical time. Other numerical schemes merit investigation.

Remark 4.8. Weak Solutions. Remarks 4.3 and 4.7 motivate the study of weak solutions for equations such as (4.1), (4.2) governing the motion of materials with memory. Except for certain special situations valid for steady visco-elastic fluid flows (Pipkin [44] and Greenberg [17]), there is no rigorous theory for the existence of shock waves and acceleration waves. MacCamy [36], Greenberg and Hsiao [19] have studied several aspects of weak solutions but only for a single first-order conservation law with memory in one space dimension. Dafermos and Hsiao [11] proved the existence of weak solution of one-dimensional first-order quasilinear hyperbolic systems with memory using Glimm's modified random choice method [16] with fractional steps. However, their method requires assumptions of "diagonal dominance" which are not satisfied in the case of the Cauchy problem (4.1), (4.2) modelling a viscoelastic solid. They are satisfied for certain models of heat flow (see [12]) and the specific model (4.8) for viscoelastic fluid flow).

In order to address the problem of weak solutions which would include one-dimensional problems for viscoelastic solids of the form (4.1), (4.2), a program has been initiated involving analytical techniques, the design of numerical algorithms and numerical experiments. We consider the Cauchy problem (4.1), (4.2) in the form of a first-order equivalent system. Let $w = u_x$, $v = u_t$. For classical solutions, (4.1), (4.2) is equivalent to the system

$$w_t = v_x$$
$$v_t = \phi(w)_x + a' * \psi(w)_x \qquad (4.9)$$

satisfying the initial conditions

$$w(x,0) = w_0(x), \quad v(x,0) = v_0(x) \quad . \qquad (4.10)$$

It is easy to show that a weak solution (in the sense of distributions) of (4.1), (4.2) is a weak solution of (4.9), (4.10). It is straightforward that the Rankine-Hugoniot jump conditions for elastic shocks ($a \equiv 0$ in (4.1)) are also necessary for viscoelastic shocks.

The Riemann problem is only partially understood for scalar first-order conservation laws with memory [36], but not at all for the viscoelastic problem (4.9), (4.10). Therefore it is difficult to use the random choice method [16]. If $\psi \neq \phi$, define $z = a' * \psi(w)$. Then (4.9) transforms to the hyperbolic system with lower order source terms:

$$
\begin{aligned}
w_t &= v_x \\
v_t &= \phi(w)_x + z_x \\
z_t &= a'(0)\psi(w) + a'' * (w) \quad ,
\end{aligned}
\tag{4.11}
$$

with $w(x,0)$, $v(x,0)$ satisfying (4.10) and $z(x,0) \equiv 0$. If $\phi'(\cdot) > 0$ (4.11) has the eigenvalues $\pm (\phi'(\cdot))^{1/2}$ and 0. If $\phi'(\cdot) - a(0)\psi'(\cdot) > 0$, (4.11) has a uniquely determined steady state solution. Observe that initially $z_x \equiv 0$; one can solve the first two equations in (4.11) by various techniques for conservation laws on the first time step, update z using the last equation and proceed forward in time. Jointly with B. Plohr we have initiated a study of various numerical algorithms for (4.11) in the special case $a(t) = \sum\limits_{i=1}^{n} a_k \exp(-\lambda_k t)$, $a_k > 0$, $\lambda_k > 0$, including the Glimm scheme with fractional steps. One objective is to establish existence of weak solutions for small BV data. Another is to obtain implementable numerical algorithms which can be tested on concrete problems.

Boldrini [3, 4] used techniques of compensated compactness to study elastic and viscoelastic problems including the system (4.9), (4.10). These techniques were developed by Tartar [50,51,52], Murat [40] and DiPerna [14]; in [14] DiPerna succeeded to extend these techniques and apply them to establish the existence of weak solutions of the purely elastic one-dimensional problem (i.e. (4.9), (4.10) with $a \equiv 0$) on $R \times [0,T]$ for any $T > 0$, without restricting the size of the data. Boldrini [4] assumes that the memory in (4.9) is small in the sense that

$$
a := a(\delta,t), \quad \psi(\cdot) := \phi(\cdot) + \mu g(\cdot) \quad , \tag{4.12}
$$

where $\delta > 0$, $\mu > 0$ are small parameters, g is a smooth function satisfying the growth condition $|g(w)| < \kappa |w|$, $\kappa > 0$, and $a'(\delta,t) = O(\delta)$, $a''(\delta,t) = O(\delta)$ uniformly in t. In place of (4.9) he considers the regularized system

$$
\begin{aligned}
w_t &= v_x \\
v_t &= \phi(w)_x + a'(\delta,\cdot) * (\phi(w) + \mu g(w))_x + \varepsilon v_{xx},
\end{aligned}
\tag{4.13}
$$

with initial data (4.10) (the Newtonian viscosity can be more general than εv_{xx}), where $\varepsilon > 0$ is a small parameter. Let $w_{\varepsilon,\delta,\mu}$, $v_{\varepsilon,\delta,\mu}$ be a solution of (4.9), (4.10) on $R \times [0,T]$ for any $T > 0$. Boldrini gives sufficient conditions which insure that there is a subsequence such that $w_{\varepsilon,\delta,\mu} \to w$, $v_{\varepsilon,\delta,\mu} \to v$ on $R \times [0,T]$ as $\varepsilon, \delta, \mu \to 0^+$, where $\mu = O(\varepsilon^{1/2}\delta^{-1})$. Moreover, w, v is a weak solution of the purely elastic problem on $R \times [0,T]$. The most serious of his assumptions is the crucial hypothesis requiring the solutions $w_{\varepsilon,\delta,\mu}$, $v_{\varepsilon,\delta,\mu}$ of (4.13) to lie in L^∞ uniformly in the parameter ε, δ, μ. Since the memory is a nonlocal operator, this assumption is difficult to verify.

Jointly with W. Rogers and T. Tzavaras, we are using compensated compactness techniques to establish the existence of weak solutions of (4.9), (4.10). The special case $\psi \equiv \phi$, but with the memory not small (i.e. a independent of δ) is tractable by these methods and the case $\psi \neq \phi$ appears doable. However, obtaining an invariant region in order to show that solutions of the relevant regularized system lie in L^∞ is extremely difficult. It is of interest to note that the existence of weak solutions of the Cauchy problem for the model first-order scalar equation with memory

$$u_t + \phi(u)_x + a'*\psi(u)_x = 0, \quad x \in R, \; t > 0$$

$$u(x,0) = u_0(x), \quad x \in R,$$

(4.14)

where a, ϕ, ψ have the same meaning as in (4.9), can be solved completely by using the method of compensated compactness. The maximum principle proved by Dafermos in [9] for classical solution of (4.14) makes it possible to prove the needed L^∞ estimates for solutions of the regularized problem (i.e. (4.14) with εu_{xx} on the right side in place of zero). This problem was recently solved by Dafermos (oral communication). Unfortunately, it does not appear that this approach can be extended to coupled two by two systems with memory.

REFERENCES

[1] Andrews, G., On the existence of solutions to the equations $u_{tt} = u_{xxt} + \sigma(u_x)_x$, J. Diff. Eq. **35** (1980), 200-231.

[2] Bernstein, B., E. A. Kearsley and L. J. Zapas, A study of stress relaxation with finite strain, Trans. Soc. Rheology **7** (1963), 391-410.

[3] Boldrini, J. L., Is elasticity the proper asymptotic theory for materials with small viscosity and capillarity? Brown Univ., Ph.D. Thesis '85, LCDS Report #85-8.

[4] Boldrini, J. L., Is elasticity the proper asymptotic theory for materials with memory? Brown Univ., Ph.D. Thesis '85, LCDS Report #85-9.

[5] Boltzmann, L., Zur Theorie der elastischen Nachwirkung, Ann. Physik **7** (1876), Ergänzungsband, 624-625.

[6] Coleman, B. D. and M. E. Gurtin, Waves in materials with memory II. On the growth and decay of one dimensional acceleration waves, Arch. Rational Mech. Anal. **19** (1965), 239-265.

[7] Coleman, B. D., M. E. Gurtin and I. R. Herrera, Waves in materials with memory I, Arch. Rational Mech. Anal. **19** (1965), 1-19.

[8] Coleman, B. D. and W. Noll, An approximation theorem for functionals with applications in continuum mechanics, Arch. Rational Mech. Anal. **6** (1960), 355-370.

[9] Dafermos, C. M., Dissipation in materials with memory, Viscoelasticity and Rheology, Proceedings of Mathematics Research Center Symposium, October 1984, A. S. Lodge, J. A. Nohel and M. Renardy, coeditors, Academic Press, Inc., New York (1985), 221-234.

[10] Dafermos, C. M., Development of singularities in the motion of materials with fading memory, Arch. Rational Mech. Anal. **91** (1986), 193-205.

[11] Dafermos, C. M. and L. Hsiao, Discontinuous motions of materials with fading memory, in preparation

[12] Dafermos, C. M. and J. A. Nohel, Energy methods for nonlinear hyperbolic Volterra integrodifferential equations, Comm. PDE **4** (1979), 219-278.

[13] Dafermos, C. M. and J. A. Nohel, A nonlinear hyperbolic Volterra equation in viscoelasticity, Amer. J. Math. Supplement (1981), 87-116.

[14] DiPerna, R. J., Convergence of approximate solutions to conservation laws, Arch. Rational Mech. Anal. **82** (1983), 27-70.

[15] Doi, M. and S. F. Edwards, Dynamics of concentrated polymer systems, J. Chem. Soc. Faraday **74** (1978), 1789-1832 and **75** (1979), 38-54.

[16] Glimm, J., Solutions in the large for nonlinear hyperbolic systems of equations, Comm. Pure Appl. Math. **18** (1965), 697-715.

[17] Greenberg, J. M., The existence of steady shock waves in nonlinear materials with memory, Arch. Rat. Mech. Anal. **24** (1967), 1-21.

[18] Greenberg, J. M., A priori estimates for flows in dissipative materials, J. Math. Anal. Appl. **60** (1977), 617-630.

[19] Greenberg, J. M. and L. Hsiao, The Riemann problem for the system $u_t + \sigma_x = 0$, $(\sigma - \hat{\sigma}(u))_t + \frac{1}{\epsilon}(\sigma - \mu\hat{\sigma}(u)) = 0$, Arch. Rational Mech. Anal. **82** (1983), 87-108.

[20] Gripenberg, G., Nonexistence of smooth solutions for shearing flows in a nonlinear viscoelastic fluid, SIAM J. Math. Anal. **13** (1982), 954-961.

[21] Hattori, H., Breakdown of smooth solutions in dissipative nonlinear hyperbolic equations, Quart. Appl. Math. **40** (1982/83), 113-127.

[22] Hrusa, W. J., A nonlinear functional differential equation in Banach space with applications to materials with fading memory, Arch. Rational Mech. Anal. **84** (1983), 99-137.

[23] Hrusa, W. J. and J. A. Nohel, Global existence and asymptotics in one-dimensional nonlinear viscoelasticity, in: P. G. Ciarlet and M. Roseau (ed.), Trends and Applications of Pure Mathematics to Mechanics, Springer Lecture Notes in Physics 195 (1984), 165-187.

[24] Hrusa, W. J. and J. A. Nohel, The Cauchy problem in one-dimensional nonlinear viscoelasticity, J. Diff. Eq. **59** (1985), 388-412.

[25] Hrusa, W. J. and M. Renardy, On a class of quasilinear partial integro-differential equations with singular kernels, J. Diff. Eq., accepted.

[26] Hrusa, W. J. and M. Renardy, On wave propagation in linear visco-elasticity, Quart. Appl. Math. **43** (1985), 237-254.

[27] John, F., Formation of singularities in one-dimensional nonlinear wave propagation, Comm. Pure Appl. Math. **27** (1974), 377-405.

[28] Joseph, D. D., O. Riccius and M. Arney, Shear wave speeds and elastic moduli for different liquids, II. Experiments,

[29] Kaye, A., Non-Newtonian flow in incompressible fluids, College of Aeronautics, Cranfield, England, Tech. Note 134 (1962).

[30] Kim, J. U., Global smooth solutions for the equations of motion of a nonlinear fluid with fading memory, Arch. Rational Mech. Anal. **79** (1982), 97-130.

[31] Klainerman, S. and A. Majda, Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids, Comm. Pure Appl. Math. **34** (1981), 481-524.

[32] Laun, H. M., Description of the non-linear shear behavior of a low density polyethylene melt by means of an experimentally determined strain dependent memory function, Rheol. Acta **17** (1978), 1-15.

[33] Lax, P. D., Development of singularities of solutions of nonlinear hyperbolic partial differential equations, J. Math. Phys. 5 (1964), 611-613.

[34] MacCamy, R. C., Existence, uniqueness and stability of solutions of the equation $u_{tt} = \sigma(u_x)_x + \lambda(u_x)u_{xt}$, Indiana Univ. Math. J. **20** (1970), 231-238.

[35] MacCamy, R. C., A model for one-dimensional nonlinear viscoelasticity, Quart. Appl. Math. **35** (1977), 22-33.

[36] MacCamy, R. C., A model Riemann problem for Volterra equations, Arch. Rational Mech. Anal. **82** (1983), 71-86.

[37] MacCamy, R. C. and V. J. Mizel, Existence and nonexistence of solutions of quasilinear wave equations, Arch. Rational Mech. Anal. **25** (1967), 299-320.

[38] Malek-Madani, R. and J. A. Nohel, Formation of singularities for a conservation law with memory, SIAM J. Math. Anal. **16** (1985), 530-540.

[39] Markowich, P. and M. Renardy, Lax-Wendroff methods for hyperbolic history value problems, SIAM J. Numer. Anal. **21** (1984), 24-51.

[40] Murat, F., L'injection du cone positif de H^{-1} dans $W^{-1,q}$ est compacte pour tont $q < 2$, J. Math. Pures et Appl. **60** (1981), 309-322.

[41] Nohel, J. A., A nonlinear conservation law with memory. Volterra and Functional Differential Equations, Kenneth B. Hannsgen, Terry L. Herdman, Harlan W. Stech. and Robert L. Wheeler, eds., Marcel Dekker, Inc., New York (1982), 91-123.

[42] Nohel, J. A. and M. Renardy, Development of singularities in nonlinear viscoelasticity, Proceedings of Workshop on Amorphous Polymers, Springer Verlag Lecture Notes, to appear.

[43] Nohel, J. A. and D. F. Shea, Frequency domain methods for Volterra equations, Adv. Math. **22** (1976), 278-304.

[44] Pipkin, A. C., Shock structure in a viscoelastic fluid, Quart. Appl. Math. **23** (1965/66), 297-303.

[45] Rammaha, M. A., Development of singularities of smooth solutions of nonlinear hyperbolic Volterra equations, Communications in PDE (submitted).

[46] Rause, P. E., A theory of the linear viscoelastic properties of dilute solutions of coiling polymers, J. Chem. Phys. **21** (1953), 1271-1280.

[47] Slemrod, M., Instability of steady shearing flows in a nonlinear visco-elastic fluid, Arch. Rational Mech. Anal. **68** (1978), 211-225.

[48] Slemrod, M., Breakdown of smooth shearing flow in viscoelastic fluids for two constitutive relations: vortex sheet vs vortex shock, Appendix A to: D. D. Joseph, Hyperbolic phenomena in the flow of viscoelastic fluids, Viscoelasticity and Rheology, Proceedings of Mathematics Research Center Symposium, October 1984, A. S. Lodge, J. A. Nohel and M. Renardy, coeditors, Academic Press Inc., New York (1985), 235-322.

[49] Staffans, O., On a nonlinear hyperbolic Volterra equation, SIAM J. Math. Anal. **11** (1980), 793-812.

[50] Tartar, L., Une nouvelle méthode de resolution d'équations aux dérivées patielles nonlinéaires, Lecture Notes in Math., vol. 665, Springer-Verlag (1977), 228-241.

[51] Tartar, L., Compensated compactness and applications to partial differential equations, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, vol. IV, Research Notes in Mathematics, 39, R. J. Knops, Ed., Pitman Publ. Inc., 1979.

[52] Tartar, L., The compensated compactness method applied to systems of conservation laws, Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., Reidel Publishing Co., Holland (1983), 263-285.

[53] Zimm, B. H., Dynamics of polymer molecules in dilure solutions: Visco-elasticity, flow birefrigence and dilectric loss, J. Chem. Phys. **24** (1956), 269-278.

41

RECENT DEVELOPMENTS IN NONSTRICTLY HYPERBOLIC CONSERVATION LAWS

Michael Shearer

Department of Mathematics

North Carolina State University

Raleigh, North Carolina 27695

David G. Schaeffer

Department of Mathematics

Duke University

Durham, North Carolina 27706

Abstract.   Our continuing study of nonstrictly hyperbolic 2×2 systems of
conservation laws is described.   Preliminary results on shock formation
in a special case are given.   The Riemann initial value problem is
discussed in the context of the four cases arising from the
classification of nonstrictly hyperbolic equations.   The solution is
outlined in one of the cases, with a discussion of some of the new
features.

1. Introduction.   In this paper, we describe recent progress in
understanding systems of nonlinear hyperbolic conservation laws whose
characteristic speeds coincide at some value of the state variable.   Such
nonstrictly hyperbolic equations arise in modelling three phase flow in
porous media (the primary motivation for our work) [14], in studies of
plane elastic waves [17], and in the Lundquist equations of
Magnetohydrodynamics (cf. [2]).   Here we consider only 2×2 systems:

$$U_t + F(u)_x = 0 \quad -\infty < x < \infty, \ t > 0, \tag{1.1}$$

where $U = U(x,t) \in R^2$ and $F: R^2 \longrightarrow R^2$.

System (1.1) is hyperbolic if $dF(U)$ has real eigenvalues $\lambda_1(U) \leq \lambda_2(U)$. Strict hyperbolicity fails at points U^* for which $\lambda_1(U^*) = \lambda_2(U^*)$. As shown in [14], such a point U is generically an umbilic point: i.e., $dF(U^*)$ is a multiple of the identity, and $\lambda_1(U) \neq \lambda_2(U)$ for $U \neq U^*$ near U^*. This situation differs from that in [5,9], where the form of the equations allows the presence of a curve of values of U for which $\lambda_1(U) = \lambda_2(U)$. We remark further that an umbilic point may be regarded as an elliptic region that has been shrunk to a point. Indeed, perturbing the equations near an umbilic point will in general produce a small region in which the eigenvalues of $dF(U)$ are complex (cf. [2,4,14]).

In a neighborhood of an umbilic point, the properties of equation (1.1) are strikingly different from properties of strictly hyperbolic equations. In this paper, we discuss preliminary results on shock formation, and present a sample of solutions of the Riemann initial value problem that is central to numerical front tracking [3].

Properties of equation (1.1) near an umbilic point U^* depend on the form of the quadratic terms in the Taylor series expansion of $F(U)$ about U^*. To focus on these terms, consider purely quadratic nonlinearities Q:

$$U_t + Q(U)_x = 0, \quad -\infty < x < \infty, \ t > 0. \tag{1.2}$$

In order that (1.2) be hyperbolic, we require that $dQ(U)$ has real eigenvalues for all U. Then, up to a linear constant change of variables in U, we may take

$$Q(U) = dC(U) \tag{1.3}$$

where

$$C(u,v) = au^3/3 + bu^2v + uv^2 \tag{1.4}$$

(see [14]). With this result, we can classify variations in the properties of equation (1.1) near an umbilic point in terms of the parameters a,b.

As an indication of the effect of the umbilic point, consider rarefaction curves for system (1.2). These are integral curves of the right eigenvectors of $dQ(U)$:

$$U' = r_k(U), \quad \text{where} \quad dQ(U)r_k(U) = \lambda_k(U)r_k(U), \tag{1.5}$$

k = 1 or 2. There are three patterns of rarefaction curves, depending on (a,b). One of these patterns splits into two cases by considering directions of increasing characteristic speed $\lambda_k(U)$ (indicated by arrows in Figure 2). The four cases are indicated in the (a,b) plane in Figure 1, with the corresponding rarefaction curves shown in Figure 2.
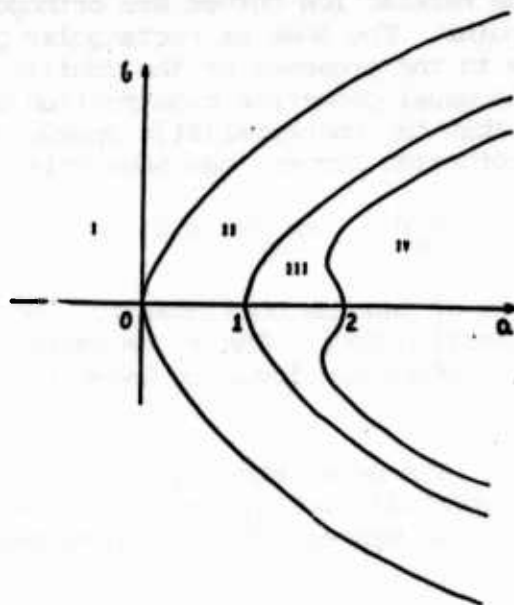
Figure 1. The (a,b) - plane.
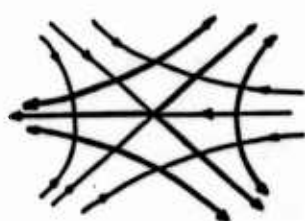
Case I
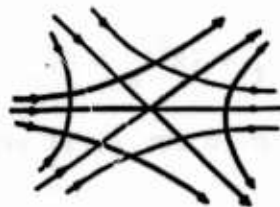
Case II
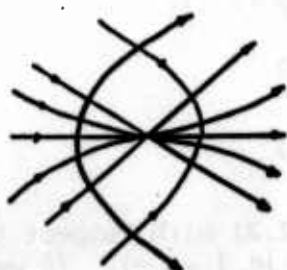
Case III
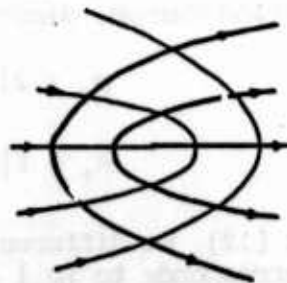
Case IV

_____ slow waves
_____ fast waves

Figure 2. The rarefaction curves.

45

Rarefaction curves give the values of centered piecewise smooth solutions $U(x/t)$, with $x/t = \lambda_k(U)$. These special solutions are called rarefaction waves. Waves associated with λ_1 are called slow waves, while those associated with λ_2 are called fast waves. In §3, we show how rarefaction waves and shock waves are used to solve certain Riemann problems. Note that the rarefaction curves are orthogonal trajectories away from the umbilic point. The loss of rectangular geometry of the rarefaction curves, due to the presence of the umbilic point, has a profound effect upon the usual geometric construction of solutions of the Riemann problem. Note that the characteristic speeds, when restricted to their corresponding rarefaction curves, can have critical points:

$$r_k(U) \cdot \nabla\lambda_k(U) = 0 \qquad (1.6)$$

corresponding to the loss of genuine nonlinearity. We refer to the lines defined by (1.6) as inflection loci. There are three inflection loci in Case I, and there is one inflection locus in Cases II-IV.

2. Formation of shocks. The usual strategy for studying shock formation for strictly hyperbolic 2×2 systems is to use the Riemann invariants to diagonalize the system. Results have been obtained even when the equations are not genuinely nonlinear for all U [10,13].

Riemann invariants for equation (1.2) are known to exist only for $a = -1$, $b = 0$. In this case, we rewrite equation (1.2) using complex notation $z = u+iv$:

$$z_t + (\overline{z}^2)_x = 0, \quad -\infty < x < \infty, \quad t > 0. \qquad (2.1)$$

Riemann invariants ρ, σ for (2.1) are given by

$$w = \rho + i\sigma, \quad w = z^{3/2}. \qquad (2.2)$$

The mapping (2.2) takes the coordinate system of Case I, Figure 1 onto a rectangular grid. This corresponds to ρ, σ being constant on their respective rarefaction curves, diagonalizing (2.1):

$$\rho_t - 2|z|\rho_x = 0 \qquad (2.3)$$

$$\sigma_t + 2|z|\sigma_x = 0 \qquad (2.4)$$

Following Lax [12], we differentiate (2.3) with respect to x (since shock formation corresponds to $|\rho_x| \to \infty$ or $|\sigma_x| \to \infty$). If we introduce a new variable

$$q = |z|^{1/3} \rho_x, \tag{2.5}$$

we find, after a straightforward calculation, that q satisfies the equation

$$\frac{dq}{dt} = c \frac{\rho}{|z|^{5/3}} q^2, \tag{2.6}$$

where $d/dt = \partial_t - 2|z|\partial_x$, and $c > 0$ is a constant. Now ρ = constant in (2.6), so we easily read off that $q \longrightarrow \pm \infty$ in finite time if $\rho q > 0$ at $t = 0$, i.e. if $\rho(x,0) \rho_x(x,0) > 0$ for some x. Similarly, if $\sigma(x,0)\sigma_x(x,0) < 0$ for some x, then (2.1) cannot have globally smooth solutions, due to $\sup_x |\sigma_x(x,t)| \longrightarrow \infty$ in finite time.

These conditions for the nonexistence of globally smooth solutions have the interpretation that the initial data should reverse the orientation of the appropriate rarefaction curve. This is precisely the situation that guarantees shock formation for strictly hyperbolic, genuinely nonlinear equations. The reason the same conditions apply here is that, for equation (2.1), the rarefaction curves of one characteristic family do not encounter inflection loci of the other family.

For $(a,b) \neq (-1,0)$, it is appropriate to use generalized Riemann invariants [8], in order to get a coupled system of Ricatti equations, each equation having the form (2.6). The coefficient of q^2 will not however automatically have a single sign for all $t > 0$, due to the crossing of rarefaction curves and inflection loci of the opposite characteristic family. It is not known how to describe the class of smooth initial data giving rise to finite time shock formation, except in the special case $a = -1$, $b = 0$ considered above.

3. Solution of the Riemann problem. In this section, we present some features of the Riemann initial value problem for equation (1.2), with Q given by (1.3), (1.4). The Riemann problem consists of finding a physical weak solution $U(x/t)$ of (1.2) satisfying the initial condition

$$U(x,0) = \begin{cases} U_L & \text{if } x < 0 \\ U_R & \text{if } x > 0 \end{cases} \tag{3.1}$$

47

The solution consists of rarefaction waves and shock waves. The former are smooth functions, with U taking values along one of the rarefaction curves, while the shock waves are discontinuous solutions, which we take to satisfy the Lax admissibility condition [11]. Generally, the solution of the Riemann problem involves a slow wave and a fast wave, separated by a constant value of U. Each wave may be composite, although for quadratic nonlinearites we have shown that the only physical composite waves are slow rarefaction-shocks and fast shock-rarefactions [15].
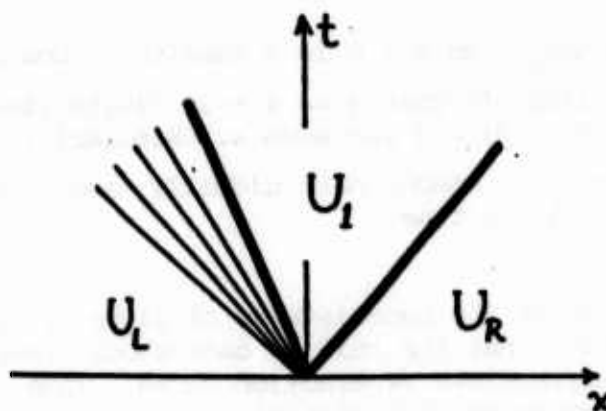
Figure 3. (RS)S solution of the Riemann problem.

A typical solution of the Riemann problem is shown in Figure 3. The solution consists of a composite slow wave (a rarefaction-shock, denoted by RS), and a fast shock, denoted by S. We codify the solution for these values of U_L, U_R by (RS)S. For a fixed U_L, the set of U_R that give rise to (RS)S solutions forms a region in the U_R plane. By considering all possible combinations of waves, for a fixed U_L, we build a picture of regions in the U_R plane. As U_L varies, these regions distort, and coalesce (for example if the strength of one of the waves goes to zero). We thus have U_L sectors: for U_L in each sector, the pattern of U_R regions is qualitatively the same.
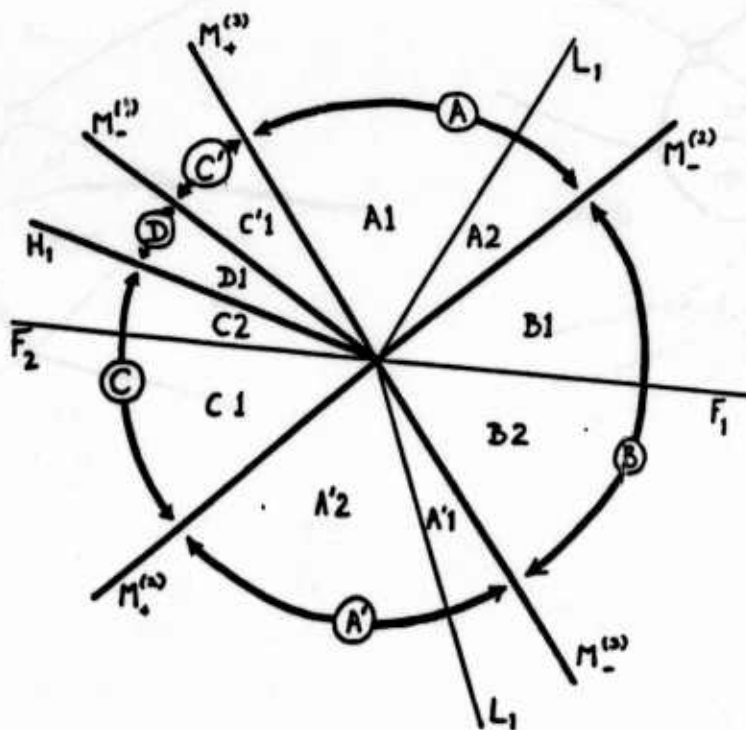
Figure 4. U_L sectors, Case III.

For Case III, the sectors are shown in Figure 4, and representative U_R diagrams are shown in Figure 5. The heavy lines in Figure 4 indicate values of U_L for which comparatively major changes occur in the U_R diagrams, while the fainter lines correspond to minor changes in the U_R diagrams. The intermediate state U_1, between the two waves, lies on one of the heavy lines in Figure 5, corresponding to slow waves. The knotted lines represent overcompressive waves, in which the two waves used to solve the Riemann problem touch, so that there is no intermediate state U_1. Another role of the knotted lines is that the solution of the Riemann problem is discontinuous with respect to U_R across this line. Specifically, the intermediate state U_1 experiences a jump, from one section of the heavy line to another section, as U_R crosses the knotted line. Note however, that the solution is continuous in the L_1 norm, due to the touching of the slow and fast waves in the limit as U_R approaches the knotted line.

49
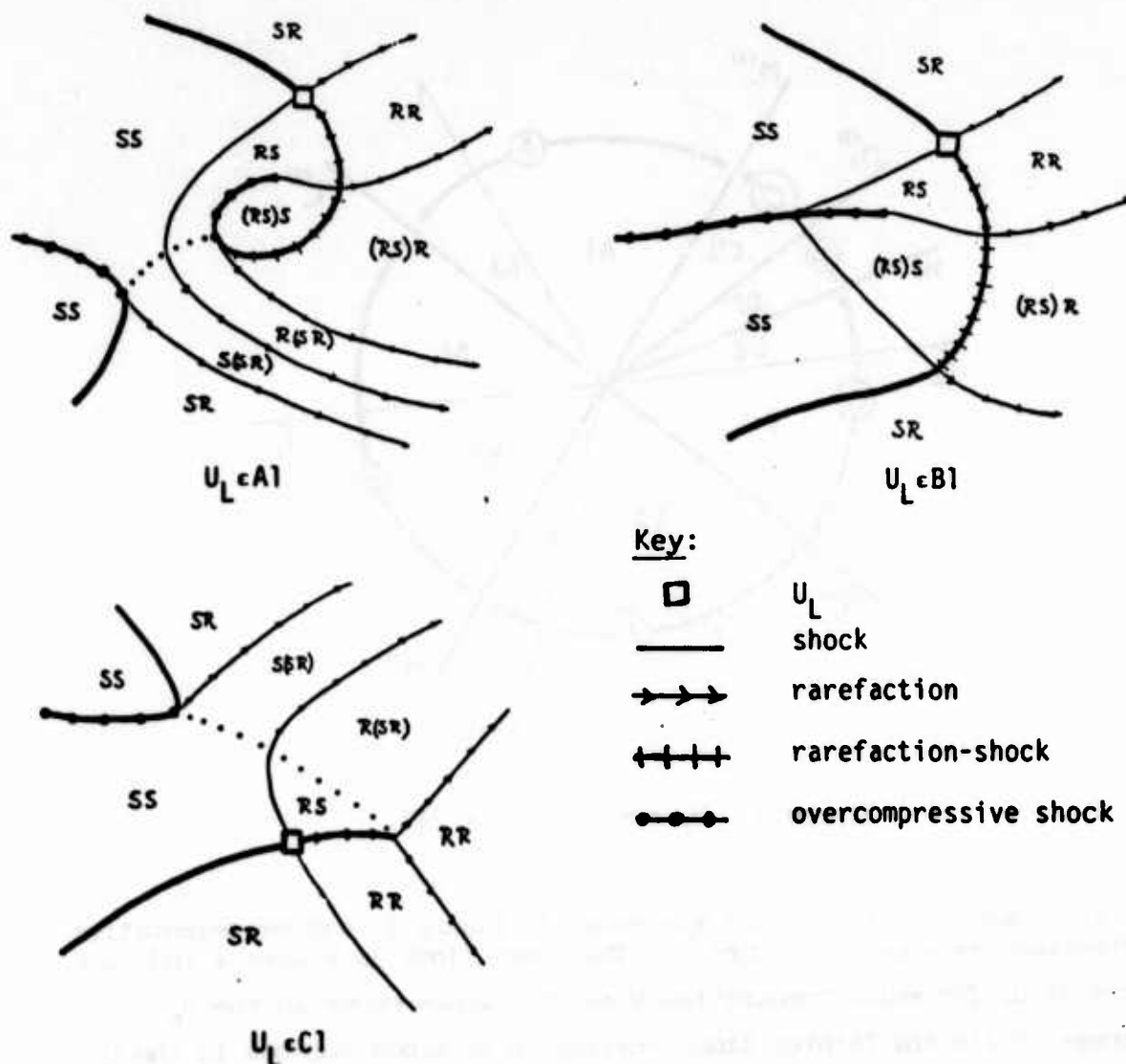
Figure 5. Patterns of U_R regions, Case III.

A detailed interpretation of the diagrams, together with diagrams for Cases II and IV, is given in [15]. The main result here is that in Cases II-IV, the Riemann problem for equation (1.3) has a unique physical solution that can be constructed graphically. A computer program to automate this solution is being developed by E. Isaacson, D. Marchesin and B. Plohr. Our work on these Riemann problems involves a combination of computer graphics and Mathematical analysis, and owes much to a study of the symmetric cases (b = 0 in (1.4)), given in [6,7] (see also [17]). Case I presents special problems because the Lax admissibility condition on shocks is too restrictive. As shown in [16], the solution of the Riemann problem in Case I will in general require the admissibility of certain undercompressive shocks. It is at present unknown how to characterize these, except in the special symmetric case of equation (2.1), for which the Riemann problem is solved in [16].

REFERENCES

[1] J.B. Bell, J.A. Trangenstein, an G.R. Shubin, Conservation laws of mixed type describing three-phase flow in porous media, Exxon Production Research, preprint, 1985.

[2] K.O. Friedrichs and H.C. Kranzer, Notes on Magnetohydrodynamics VIII; nonlinear wave motion, New York University preprint, 1958.

[3] J. Glimm, E. Isaacson, D. Marchesin and O. McBryan, Front tracking for hyperbolic systems. Adv. Appl. Math. 2 (1981), 91-119.

[4] H. Holden, On the Riemann problem for a prototype of a mixed type conservation law, preprint, New York University, 1986.

[5] E. Isaacson, Global solution of a Riemann problem for a nonstrictly hyperbolic system of conservation laws arising in enhance oil recovery, J. Comp. Phys., to appear.

[6] E. Isaacson, D. Marchesin, B. Plohr and B. Temple, The classification of solutions of quadratic Riemann problems I. MRC report, 1985.

[7] E. Isaacson and B. Temple, The classification of solutions of quadratic Riemann problems II. MRC report, 1985.

[8] F. John, Formation of singularities in one-dimensional nonlinear wave propagation, Comm. Pure Appl. Math. 27(1974), 377-405.

[9] B.L. Keyfitz and H.C. Kranzer, A system of hyperbolic conservation laws arising in elasticity theory, Arch. Rat. Mech. Anal. 72(1980), 219-241.

[10] S. Klainermann and A. Majda, Formation of singularities for wave equations including the nonlinear vibrating string, Comm. Pure Appl. Math. 33(1980), 241-263.

[11] P.D. Lax, Hyperbolic systems of conservation laws II, Comm. Pure Appl. Math. 10(1957), 537-566.

[12] P.D. Lax, Development of singularities of solutions of nonlinear hyperbolic partial differential equations, J. Math. Phys. 5(1964), 611-613.

[13] Li Da-tsin and Shi Jia-hong, Singularities of solutions for first order quasilinear hyperbolic systems, Proc. Roy. Soc. Edinburgh 94A (1983), 137-147.

[14] D.G. Schaeffer and M. Shearer, The classification of 2×2 systems of nonstrictly hyperbolic conservation laws, with application to oil recovery; Appendix with D. Marchesin, P.J. Paes-Leme, Comm. Pure Appl. Math., to appear.

[15] D.G. Schaeffer and M. Shearer, Riemann problems for nonstrictly hyperbolic 2×2 systems of conservation laws, to appear.

[16] M. Shearer, D.G. Schaeffer, D. Marchesin and P.J. Paes-Leme, Solution of the Riemann problem for a prototype 2×2 system of nonstrictly hyperbolic conservation laws, Arch. Rat. Mech. Anal., to appear.

[17] Z. Tang and T.C.T. Ting, Wave curves for the Riemann problem of plane waves in simple isotropic elastic solids, preprint, University of Ill., Chicago, 1985.

APPLICATIONS OF MATRIX FACTORIZATION IN HYDRODYNAMIC STABILITY

Philip J. Morris
Department of Aerospace Engineering
The Pennsylvania State University
University Park, PA 16802

ABSTRACT. The matrix discretization of boundary value problems that occur in hydrodynamic stability contain the frequency ω and the wavenumber α of the normal modes as well as other parameters. For most temporal stability calculations an algebraic eigenvalue problem may be posed since ω appears linearly. Spatial stability problems are more complicated since the eigenvalue α appears nonlinearly. Problems of this type are examined in this paper. The stability of a laminar boundary layer over a compliant wall is considered. In this case the wavenumber appears to power four in the differential equation, the Orr-Sommerfeld equation, and to power five in the wall boundary condition. A model for the compliant surface is developed and the differential problem is defined. The matrix methods applied to the solution of this problem are demonstrated on a model problem. Eigenvalue spectra are calculated for the model problem and the boundary layer stability problem. The methods for obtaining the eigenvalue efficiently depend on the factorization of matrix polynomials. Various factorization schemes are considered including Bernoulli and Traub iteration and Newton's method.

1. INTRODUCTION. This paper is concerned with the application of matrix factorization techniques to problems in hydrodynamic stability. A spectral method is used to discretize the boundary value problem. Orszag [1] used a spectral approach to obtain the eigenvalue spectrum of the Orr-Sommerfeld equation for Poiseuille flow. He considered temporal stability in which the eigenvalue, which is the frequency of the normal mode, appears linearly. Thus the problem becomes an algebraic eigenvalue problem which may be solved by a number of standard algorithms. The spatial stability problem is more complicated since the eigenvalue, which in this case is the wavenumber of the normal mode, appears to power four in the differential equation. However it is this problem that is physically realistic in which fixed real frequency disturbances amplify convectively. For the rigid wall boundary conditions of Poiseuille flow the boundary conditions are independent of the eigenvalue. In this case the spectral discretization of the boundary value problem yields an eigenvalue problem of the form:

$$\left[\sum_{k=0}^{4} A_k \alpha^{4-k}\right] \mathbf{a} = 0 \qquad (1)$$

where **a** is the eigenvector of Chebyshev coefficients and α is the wavenumber. To recast this problem as an algebraic eigenvalue problem Benney and Orszag [2] used the Companion Matrix Method which is described briefly below. This approach yields matrices which

are four times the size of the original matrices A_i. Since the operation count of standard eigenvalue algorithms, such as the QR algorithm, are of the order of N^3, this approach is computationally expensive. For this reason Benney and Orszag [2] chose to solve the temporal eigenvalue problem and then used transformations, which are valid for small growth rates to convert to the spatial stability case. It should be emphasized that if a local iteration method is used to find the eigenvalue the spatial problem is no more complicated than the temporal problem. However a good first approximation for the eigenvalue is required and there is no guarantee that all unstable or critical eigenvalues will be found. Bridges and Morris [3] applied methods based on matrix factorization to the eigenvalue problem in Eq. (1). This technique converts the problem in which the eigenvalue appears nonlinearly to one in which it appears linearly and is readily obtained by standard algorithms. The resulting algebraic eigenvalue problem is of the same size as the original matrices so that this technique is much more efficient than the Companion Matrix Method. The eigenvalues yielded by this approach represent a subset of the eigenvalues of the entire problem Eq. (1). They may be the subset of eigenvalues with either the greatest or smallest absolute values. Bridges and Morris [4] examined the stability of the Blasius boundary layer with this technique. However the only successful globally convergent scheme for this problem was found to be the Companion Matrix Method. The reasons for this are discussed in this paper and a successful application of the matrix factorization scheme is provided. Carpenter and Morris [5] applied the matrix factorization scheme to the problem of the stability of a laminar boundary layer over a non-isotropic compliant surface. Using the matrix factorization scheme they were able to identify various modes of instability simultaneously. However, it should be noted that the accuracy of any eigenvalue is improved if its location is known approximately in the complex plane.

In this paper the formulation of the boundary layer stability problem over a compliant surface is reformulated. In the new form there is no restriction on the degree or nature of non-isotropy of the compliant surface. However, the major emphasis of this paper is not this particular problem, but general problems of the same type. The interesting feature of the compliant wall stability problem is that the eigenvalue appears to a higher power in the boundary conditions than in the differential equation. Also that the domain of the independent variable is unbounded so that the problem exhibits a continuous as well as a discrete spectrum.

In the subsequent sections the boundary value problem for the stability of a laminar boundary layer on a non-isotropic compliant surface will be developed. A model problem with many of the features of the real problem will be introduced. This problem is solved by various methods including the Companion Matrix Method and by matrix factorization. Various schemes for the factorization of matrix polynomials are examined. Calculations of the eigenvalue spectrum and its subset, obtained from the matrix factorization approach, are given for both the model problem and the compliant boundary layer stability problem.

2. PROBLEM FORMULATION.

The efficiency and quietness of underwater vehicles is affected by the nature of their boundary layers: a fully-laminar boundary layer providing the least drag and noise. A passive method for delaying boundary layer transition to turbulence involves the use of a compliant surface. Early experiments by Kramer [6,7]

54

indicated a significant reduction in the drag on a vehicle with a compliant coating but until recently experiments had failed to reproduce these results. However Gaster and Daniel [8] showed that the growth of Tollmien-Schlichting instabilities can be reduced dramatically by an appropriate choice of compliant surface. The compliant surface used was a silicone rubber-based substrate with a latex skin. Since such a simple surface provided a reduction in the growth of instabilities and gave good agreement with the predictions of linear theory it is reasonable to examine other surfaces theoretically that could give further reduction in the wave growth.

The production rate of fluctuation energy, whether in the early stages of transition or for a turbulent flow, depends on the product of the Reynolds stress and the strain rate of the basic flow. If these quantities have unequal signs there is production and if they have equal signs there is negative production or decay of the unsteadiness. In a boundary layer production occurs close to the wall. Grosskreutz [9] proposed a nonisotropic compliant surface that would force the production at the wall to be negative. Some stability calculations for a model of this surface were performed by Carpenter and Morris [5]. A revised formulation of this problem forms the boundary value problem discussed below.

A simple model for the surface is shown in Fig. 1. The nondimensional displacements of the surface η and ξ in the normal and streamwise directions respectively are related to the angular displacement of the swivel arms $\delta\theta$, by

$$\xi\delta^* = \ell\,\delta\theta\,\sin\theta \qquad \text{and} \qquad \eta\delta^* = \ell\,\delta\theta\,\cos\theta. \tag{2}$$

where δ^* is the displacement thickness of the boundary layer. These relationships show that the production term will be negative if the the swivel arms are directed towards the flow direction and positive if the arms point downstream. The equation of motion for an element of the surface in the direction normal to the swivel arm may be written

$$\rho_m b\frac{\partial^2(\ell\delta\theta)}{\partial t^2} = -B\frac{\partial^4\eta}{\partial x^4}\cos\theta - K\ell\delta\theta + Eb\frac{\partial^2\xi}{\partial x^2}\sin\theta \\ - p_0\cos\theta + \sigma_0\cos\theta + \tau_0\sin\theta; \tag{3}$$

x and y are the coordinates in and normal to the streamwise direction; ρ_m and b are the density and thickness of the plate; p_0, σ_0 and τ_0 are the pressure and the normal and shear viscous stresses at the wall; B and E are the flexural rigidity and elastic modulus of the plate; and K is the spring stiffness. Let the velocity fluctuations in the (x,y) directions be (u,v) and seek a solution for the surface displacement in the form:

$$\eta = \hat{\eta}\delta^*\exp[i(\alpha x - \omega t)]. \tag{4}$$

Continuity of normal and tangential motion at the wall then yields:

$$\bar{\omega}\hat{\eta} = i\hat{v}(0) \tag{5}$$

and

$$-i\bar{\alpha}\bar{\omega}\sin\theta\hat{v}(0) = \bar{\alpha}\cos\theta U'(0)\hat{v}(0) + \bar{\omega}\cos\theta\hat{v}'(0). \tag{6}$$

55

U is the mean velocity of the boundary layer, primes denote differentiation with respect to y and all quantities in eqns. (5) and (6) have been nondimensionalized with respect to the freestream velocity U_∞ and the displacement thickness δ^*. The fluctuating stresses at the wall may be related to the normal velocity fluctuation in the fluid using the linearized continuity and momentum equations. If $\varsigma = \hat{v}'$ then η and ξ may be eliminated and the wall boundary conditions may be written in terms of ς and $\hat{v}$ alone:

$$\bar{\alpha}^5\left[\frac{C_B}{C_M}\cos^2\theta\varsigma(0)\right] + \bar{\alpha}^3\left[\frac{C_T}{C_M}\sin^2\theta\varsigma(0)\right]$$

$$+ \bar{\alpha}^2\left[(2\bar{\omega}\sin\theta - 3iU'(0)\cos\theta)\frac{\cos\theta}{C_M R}\varsigma(0)\right]$$

$$+ \bar{\alpha}\left[\left(\frac{C_K}{C_M} - \bar{\omega}^2\right)\varsigma(0) + (\cos\theta U'(0) + i\bar{\omega}\sin\theta)\frac{\sin\theta}{C_M R}\varsigma'(0)\right]$$

$$+ \left[i(\cos\theta U'(0) + i\bar{\omega}\sin\theta)\frac{\cos\theta}{C_M R}\varsigma''(0)\right.$$

$$\left. - \frac{i\bar{\omega}^2}{C_M}\sin\theta\cos\theta\varsigma(0)\right] = 0, \tag{7}$$

and

$$\bar{\alpha}\left[\cos\theta U'(0) + i\bar{\omega}\sin\theta\right]\hat{v}(0) + \bar{\omega}\cos\theta\varsigma(0) = 0. \tag{8}$$

where

$$C_M = \frac{\rho_m b}{\rho_0\delta^*}, \qquad C_B = \frac{B}{\rho_0 U_\infty^2 \delta^*}, \qquad C_K = \frac{K\delta^*}{\rho_0 U_\infty^2} \quad \text{and} \quad C_T = \frac{Eb}{\rho_0 U_\infty^2 \delta^*}.$$

R is the Reynolds number $U_\infty\delta^*/\nu$. It should be noted that $\bar{\alpha}$ appears to power five in eqn. (7). In ref. [5] a different form of the condition contained $\bar{\alpha}$ to power six. Also eqn. (7) is valid for all values of θ. The velocity fluctuations in the boundary layer satisfy the Orr-Sommerfeld equation which may be written in terms of $\hat{v}$ and ς as,

$$\varsigma''' + A(y)\varsigma' + B(y)\hat{v} = 0,$$
$$\hat{v}' - \varsigma = 0, \tag{9}$$

where

$$A(y) = -iR(\bar{\alpha}U - \bar{\omega}) - 2\bar{\alpha}^2$$

and

$$B(y) = iR(\bar{\alpha}U - \bar{\omega})\bar{\alpha}^2 + i\bar{\alpha}RU'' + \bar{\alpha}^4.$$

In addition the fluctuations are required to vanish at infinity:

$$\hat{v}(y) = \hat{v}'(y) \to 0 \quad \text{as} \quad y \to \infty \tag{10}$$

In order to demonstrate the numerical methods without the complexity of the algebra involved in the problem given by eqns. (7)–(10) a model problem will be introduced.

56

Though this problem does not have the stiffness of the Orr-Sommerfeld equation it does have the correct nonlinearity of the eigenvalue in the equation and boundary conditions.

Consider the model problem:

$$\frac{d^2\phi}{dx^2} - 2\alpha\omega\frac{d\phi}{dx} + \alpha^2\phi = 0 \qquad x \in [-1,1] \tag{11}$$

$$\phi(1) = 0 \tag{12.a}$$

$$\alpha^3\phi(-1) + \frac{d\phi}{dx}(-1) = 0 \tag{12.b}$$

The exact solution to this problem is given by,

$$\phi(x) = A\exp(\alpha\omega x)\,[\sin\gamma\cos\gamma x - \cos\gamma\sin\gamma x], \tag{13}$$

where

$$\gamma = \alpha\sqrt{1 - \omega^2},$$

with

$$\tan\left[2\alpha\sqrt{1 - \omega^2}\right] = \sqrt{1 - \omega^2}/(\omega + \alpha^2). \tag{14}$$

$\phi(x)$ is approximated by a finite series of Chebyshev polynomials:

$$\phi(x) = \sum_{r=0}^{N}{}' a_r T_r(x). \tag{15}$$

The formulae for the integrals of Chebyshev polynomials are much simpler than those for the derivatives. Thus eqn.(11) is first integrated twice indefinitely with respect to x. Prior to substitution of the series approximation this equation is perturbed by two additional Chebyshev polynomials to prevent a trivial solution (see ref. [10]). Thus the actual equation solved is,

$$\phi(x) - 2\alpha\omega\int\phi + \alpha^2\int\int\phi$$
$$= C_1 x + C_2 + \tau_{N+1}T_{N+1}(x) + \tau_{N+2}T_{N+2}(x). \tag{16}$$

When the series (15) is substituted into eqn. (16), and the boundary conditions (12) and the coefficients of equal orders of Chebyshev coefficient are set to zero a matrix eigenvalue problem is obtained.

$$\left\{C_0\alpha^3 + C_1\alpha^2 + C_2\alpha + C_3\right\}\,a = 0. \tag{17}$$

a is the vector of unknown Chebyshev coefficients. The equations involving the zero-th and first order Chebyshev polynomials, which would involve the unknown integration constants C_1 and C_2, are replaced by the series approximation to the boundary conditions in rows N and $N + 1$ of the matrix equation. It should be noted that the leading coefficient matrix

C_0 is singular since the only terms invoving α^3 occur in one boundary condition. Thus the elements of C_0 may be written,

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ c_{N,0} & c_{N,1} & \ldots & c_{N,N+1} \\ 0 & 0 & \ldots & 0 \end{pmatrix}. \tag{18}$$

In the next section several procedures for solving the matrix eigenvalue problem (17) will be described.

3. NUMERICAL METHODS.

The Companion Matrix Method that was used by Benney and Orszag [2] involves the definition of two new vectors,

$$a_1 = \alpha a \qquad \text{and} \qquad a_2 = \alpha a_1. \tag{19}$$

With these definitions the matrix eigenvalue problem may be written in block matrix form,

$$\left[\begin{pmatrix} C_1 & C_2 & C_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{pmatrix} - \alpha \begin{pmatrix} -C_0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \right] \begin{pmatrix} a_2 \\ a_1 \\ a \end{pmatrix} = 0. \tag{20}$$

Since C_0 is singular this cannot be changed to an algebraic eigenvalue problem without first introducing a transformation,

$$\lambda = 1/(\alpha - s). \tag{21}$$

Then the problem is readily written as,

$$\left[\begin{pmatrix} A_1 & A_2 & A_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{pmatrix} - \lambda I \right] \begin{pmatrix} a_2 \\ a_1 \\ a \end{pmatrix} = 0. \tag{22}$$

Since the dimension of the block matrix is $3(N+1) \times 3(N+1)$ the computation time required to find the eigenvalue spectrum is increased by a factor 27 over the linear problem. However eqn. (17) may be factorized so that only a specific subset of the eigenvalue spectrum is calculated.

Let eqn.(17), after the use of the transformation (21), be written,

$$\{D_3(\lambda)\} \quad a = 0. \tag{23}$$

If the matrix equivalent of synthetic division is employed on eqn. (23), then the factored form of D_3 is,

$$D_3(\lambda) = \{Q_2(\lambda)\}(\lambda I - Y). \tag{24}$$

58

where Y is a solvent or factor of D_3 and Q_2 is quadratic in λ. It is readily shown that eqn. (24) will only be satisfied if Y is a root of the *matrix polynomial*,

$$Y^3 + A_1 Y^2 + A_2 Y + A_3 = 0. \tag{25}$$

Thus the eigenvalue problem reduces first to finding the roots of this matrix polynomial.

The method used by Bridges and Morris [3] that was proposed by Gohberg et al [11] involves the use of Bernoulli iteration. This method which is an extension of the standard algorithm for scalar polynomials consists of the iterative sequence:

$$X_{i+1} + A_1 X_i + A_2 X_{i-1} + A_3 X_{i-2} = 0, \tag{26}$$

with

$$X_0 = X_1 = 0 \qquad \text{and} \qquad X_2 = I. \tag{27}$$

Then,

$$\lim_{n \to \infty} X_n [X_{n-1}]^{-1} = S_1. \tag{28}$$

S_1 is the dominant solvent of D_3, that is, the solvent that contains the eigenvalues with the maximum modulus. The convergence of this algorithm is slow, though it can be improved dramatically if an appropriate choice is made for the factor s in eqn. (21).

It is reasonable to seek quadratically convergent schemes to find the roots of the matrix polynomial. However such standard scalar algorithms as Newton's method are not readily extended to the matrix polynomial. Consider the matrix polynomial,

$$Y^2 + C_1 Y + C_2 = 0. \tag{29}$$

If an iteration sequence is developed of the form,

$$Y_{i+1} = Y_i + \Delta_i, \tag{30}$$

then it is readily shown that Δ_i satisfies an equation of the form,

$$A_i \Delta_i + \Delta_i B_i = C_i. \tag{31}$$

Bartels and Stewart [12] developed an algorithm to solve for Δ_i in $O(N^3)$ operations by triangularizing the matrices A_i and B_i. A more efficient scheme was developed by Golub et al [13] still requiring $O(N^3)$ operations. However if a higher order matrix polynomial is considered such as given by eqn. (25) then the equation for Δ_i is,

$$A_i \Delta_i + B_i \Delta_i C_i + \Delta_i D_i = E_i. \tag{32}$$

There does not appear to be a particular algorithm for solving this equation. Clearly higher order matrix polynomials will lead to more complicated equations. However it is always possible to construct a system of equations for the N^2 unknown elements of Δ_i. Since Δ_i is only a small correction in the Newton's method it should not have to be evaluated with

59

a high degree of accuracy. Thus with suitable preconditioning an iterative solution of eqn. (32) might not be too lengthy. This possibility is being considered by the author.

The last algorithm to be considered is that developed by Dennis et al, [14]. This is a two-stage algorithm based on the algorithm by Traub for scalar polynomials, ref. [15]. The algorithm consists of the construction of the equivalent of the G-polynomials,

$$G_0(Y) = I$$
$$G_{n+1}(Y) = G_n(Y)Y - \Gamma_1^{(n)}D_3(Y) \tag{33}$$

where

$$G_n(Y) = \Gamma_1^{(n)}Y^2 + \Gamma_2^{(n)}Y + \Gamma_3^{(n)}. \tag{34}$$

The second stage of the algorithm consists of constructing the iterative sequence,

$$Y_0 = (\Gamma_1^{(L)})(\Gamma_1^{(L-1)})^{-1}. \tag{35a}$$

and,

$$Y_{i+1} = G_L(Y_i)G_{L-1}^{-1}(Y_i). \tag{?5b}$$

The first stage of the algorithm with Y given by eqn. (35 a) is equivalent to Bernoulli iteration. The use of the second stage of the algorithm does not change the linear convergence of the iteration but the asymptotic error constant may be made as small as desired by increasing the number of first stage iterations. It would appear to be very desirable to extend the iterative schemes based on the generalized G-polynomials of ref. [15] to the matrix case. However this extension would require properties of matrix derivatives that do not appear to be available.

In the next section some numerical examples of the application of these algorithms will be given.

4. CALCULATIONS. First the model problem given by eqns. (11) and (12) will be considered. Table I shows the eigenvalue spectrum given by the Companion Matrix Method with $\lambda = 0.5$ and $N = 12$. It can be seen that the spectrum contains $N - 1$ "infinite" eigenvalues. This corresponds to the fact that the leading coefficient matrix C_0 has rank unity. The corresponding behavior for scalar polynomials is given by an "infinite" root when the leading coefficient of the polynomial tends to zero. The roots on the real axis are close to multiples of π as could be inferred from the eigenvalue relationship given by eqn. (14). The roots away from the real axis occur in complex conjugate pairs.

Figure 2 shows the finite roots given in Table I as well as the roots obtained using Traub iteration. The number of first and second stage iterations was 10 and 5 respectively. The four eigenvalues closest to the value of the shift s in eqn. (21), which was 0.5, are very accurately obtained. However the remaining eight eigenvalues do not correspond to the values given by the Companion Matrix Method. The reason for this is unclear though the occurrence of the complex conjugates in this problem suggests that a dominant solvent may not exist. However shifting the value of s to a complex value did not alter the result. Thus the reason for the failure of the factorization scheme in this case remains unclear.

In spite of the difficulties with the model problem it has served to illustrate the numerical methods. These methods have been applied to the more complex problem posed by the Orr-Sommerfeld equation and the boundary conditions corresponding to a non-isotropic compliant surface. Rather than detail the hydrodynamic properties of such a surface a special case will be considered. It was mentioned earlier that Bridges and Morris [5] had difficulty in applying matrix factorization techniques for the rigid wall boundary layer. The reason for this can be seen if the Companion Matrix Method is applied to the boundary value problem given in Section 2 for the case of a massive wall; that is as $C_M \to \infty$. In this case the compliant surface problem reduces to the rigid wall case. Figure 3 shows the resulting eigenvalue spectrum for $N = 24, R = 2240$ and $\omega = 0.05$. This case gives 96 finite eigenvalues. The eigenvalues shown in Fig. 3 contain both discrete eigenvalues and the Chebyshev approximation to the four branches of the continuous spectrum (see ref [16]). It can be seen that many of the eigenvalues are clustered around $\alpha = 0$. If no shift in the eigenvalue, as given by eqn. ($\cdot$), is used then the eigenvalues of the *minimal solvent* will not include the discrete eigenvalue close to $\alpha = 0.3$. The minimal solvent was sought in ref. [4] and the discrete eigenvalue could not be obtained. The same spectrum of eigenvalues is shown in the $c - plane$ in Fig. 4, where $c = \omega/\alpha$. The Tollmien-Schlichting instability is indicated. One branch of the continuous spectrum forms a semi-circle in the $c - plane$ between $c = 0$ and $c = 1$. The attempt by the finite Chebyshev series to approximate this branch is seen clearly in this figure. If the eigenvalue problem is shifted by $s = 0.3$ then Traub iteration gives the spectrum shown if Fig. 5. The dominant eigenvalues have been sought using the iteration scheme given in Section 3. All of the eigenvalues associated with the approximation to the continuous spectrum that gave values of c close to zero have been eliminated. The spectrum given in Fig. 5 was obtained with 5 first stage and 5 second stage iterations. No accurate computation times were obtained but the matrix factorization scheme was considerably faster than the Companion Matrix Method.

In this section several examples of the application of matrix factorization schemes have been given. It is clear that they offer a considerable advantage over other schemes in the solution of eigenvalue problems in which the eigenvalue appears nonlinearly. It is also clear that methods for factorizing matrix polynomials that have high rates of convergence are still needed.

5. REFERENCES.

[1] S. A. Orszag, "Accurate solution of the Orr-Sommerfeld stability equation," *J. Fluid Mechanics* **11** (1971), 689.

[2] D. J. Benney and S. A. Orszag, "Stability analysis for laminar flow control. Part I," *NASA CR*-2910 (1977).

[3] T. J. Bridges and P. J. Morris, "Differential eigenvalue problems in which the parameter appears nonlinearly," *J. Computational Physics* **55** (1984) 437.

[4] T. J. Bridges and P. J. Morris, "A note on boundary layer stability calculations," submitted to **AIAA Journal**, (1986).

[5] P. W. Carpenter and P. J. Morris, "The hydrodynamic stability of flows over non-isotropic compliant surfaces–numerical solution of the differential eigenvalue problem," **Numerical Methods in Laminar and Turbulent Flow** (ed. C. Taylor et. al.) (1985) 1613.

[6] M. O. Kramer, "Boundary layer stabilization by distributed damping," *J. Aero. Sci.* **24** (1957) 459.

[7] M. O. Kramer, "Boundary layer stabilization by distributed damping," *J. Aero/Space Sci.* **27** (1960) 69.

[8] M. Gaster and A. P. Daniel, "Practical investigation of boundary layer stabilization by compliant surfaces," **ONR Compliant Coating Workshop**, (1985) London.

[9] R. Grosskreutz, "An attempt to control boundary layer turbulence with nonisotropic compliant walls," *Univ. Sci. Journal (Dar es Salaam)* **1** (1975) 67.

[10] L. Fox and I. B. Parker, "**Chebyshev Polynomials in Numerical Analysis**," Oxford University Press, (1968).

[11] I. Gohberg, P. Lancaster, and L. Rodman, "**Matrix Polynomials,**" Academic Press, New York (1982).

[12] R. H. Bartels and G. W. Stewart, "A solution of the equation $AX + XB = C$," *Commun. ACM* **15** (1972) 820.

[13] G. H. Golub, S. Nash, and C. Van Loan, "A Hessenberg-Schur method for the problem $AX + XB = C$," *IEEE Trans. on Automatic Control*, **AC-24** (1979) 909.

[14] J. E. Dennis, J. F. Traub, and R. P. Weber, "Algorithms for solvents of matrix polynomials," *SIAM J. Numer. Anal.* **15** (1978) 523.

[15] J. F. Traub, "A class of globally convergent iteration functions for the solution of polynomial equations," *Math. Comput.* **20** (1966) 113.

[16] C. E. Grosch and H. Salwen, "The continuous spectrum of the Orr-Sommerfeld equation. Part 1. The spectrum and the eigenfunctions," *J. Fluid Mechanics* **87** (1978) 33.

α_{real}	α_{imag}
$0.4406718E+00$	$0.4510956E-14$
$-.2992489E+00$	$0.1022875E+01$
$-.2992489E+00$	$-.1022875E+01$
$0.3186915E+01$	$-.2216713E-08$
$-.3093725E+01$	$0.1475552E-08$
$0.6838718E+01$	$0.1359535E+01$
$0.6838718E+01$	$.1359535E+01$
$0.6186921E+01$	$0.2801617E-08$
$0.6912758E+01$	$0.3878241E+01$
$0.6912758E+01$	$-.3878241E+01$
$-.6061245E+01$	$0.1430764E-07$
$0.6813279E+01$	$0.7021262E+01$
$0.6813279E+01$	$-.7021262E+01$
$0.1049901E+02$	$0.4074714E-07$
$-.6727921E+01$	$0.1523704E+01$
$-.6727921E+01$	$-.1523704E+01$
$-.6934206E+01$	$0.4003945E+01$
$-.6934206E+01$	$-.4003945E+01$
$-.6876758E+01$	$0.7071108E+01$
$-.6876758E+01$	$-.7071108E+01$
$-.1046709E+02$	$0.5751595E-09$
$0.9339952E+01$	$0.1092799E+02$
$0.9339952E+01$	$-.1092799E+02$
$-.9412304E+01$	$0.1088636E+02$
$-.9412304E+01$	$-.1088636E+02$
$0.1471219E+18$	$0.0000000E+00$
$0.2924940E+16$	$0.0000000E+00$
$0.1196801E+16$	$0.0000000E+00$
$0.2813020E+16$	$0.0000000E+00$
$-.3321595E+16$	$0.0000000E+00$
$-.2542934E+17$	$0.0000000E+00$
$0.2955202E+16$	$0.0000000E+00$
$0.1349990E+17$	$0.0000000E+00$
$-.2681021E+16$	$0.0000000E+00$
$-.3430352E+16$	$0.0000000E+00$
$-.4477007E+16$	$0.0000000E+00$

Table I Eigenvalues of model problem
$\omega = \sqrt{3}/2$, $N = 12$, $\tan(\alpha) = 1/(\sqrt{3} + 2\alpha^2)$
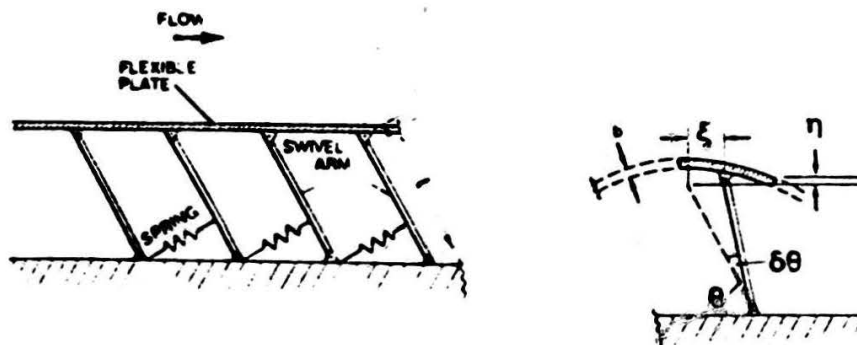
Fig. 1 Sketch of the compliant wall model.
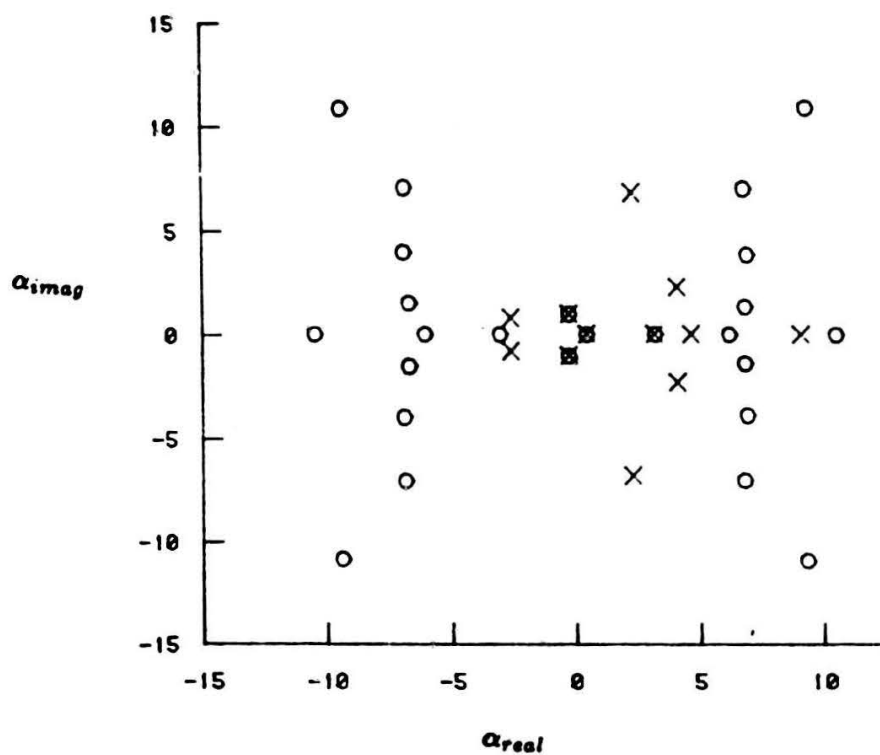
Fig. 2 Eigenvalues for the model problem. ◯ , Companion Matrix Method; × , Traub iteration.

64
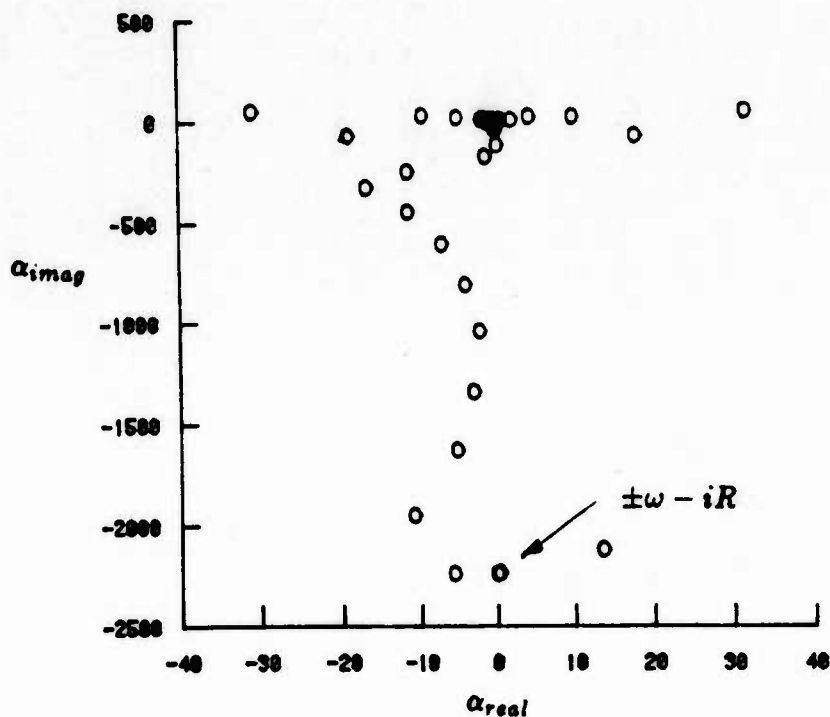
Fig. 3 Eigenvalue spectrum for the Orr-Sommerfeld problem: α-plane. $R = 2240$, $\omega = 0.05$, $N = 24$. Rigid wall case.
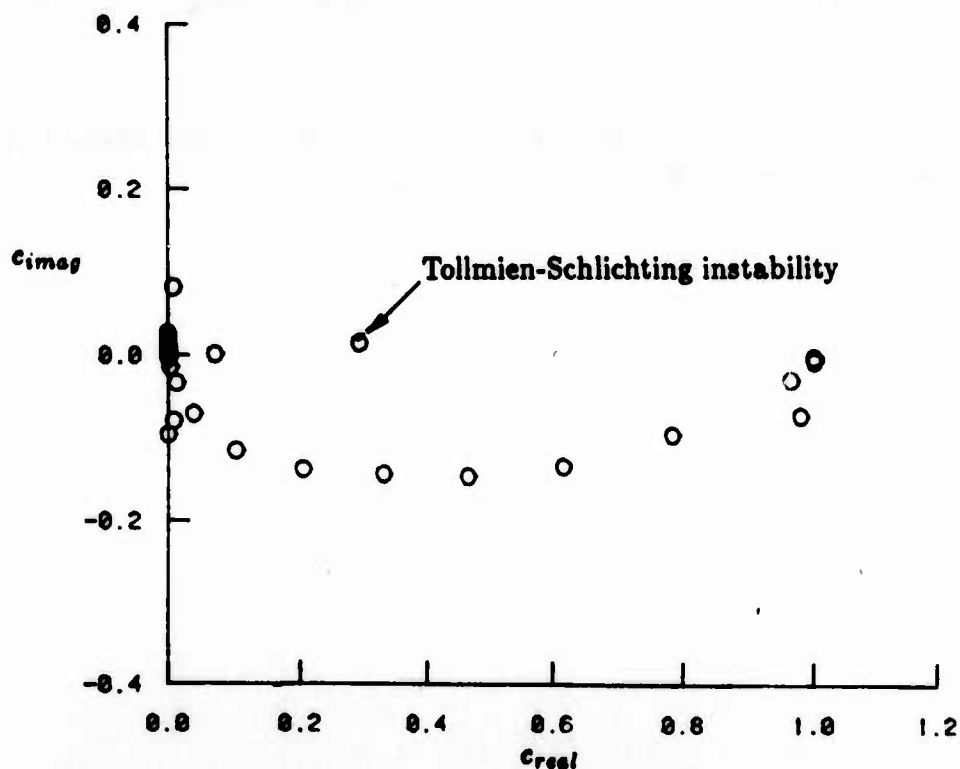
Fig. 4 Eigenvalue spectrum for the Orr-Sommerfeld problem: c-plane. Rigid wall case.
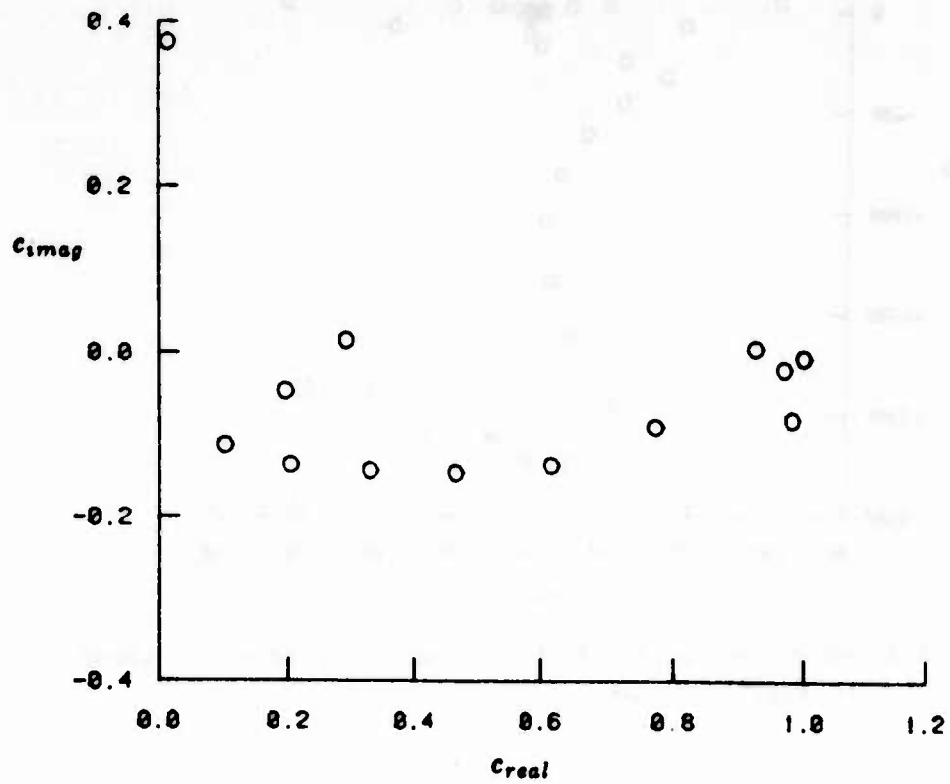
Fig. 5 Dominant eigenvalues for the Orr-Sommerfeld problem obtained by Traub iteration: c-plane. Rigid wall case.

A NUMERICAL STUDY OF THE EFFECT OF
CURVATURE ON DETONATION SPEED

B. Bukiet [1,2]

J. Jones [1,2]

Courant Institute of Mathematical Sciences
New York University
New York, N. Y. 10012

ABSTRACT

Two methods for computing the effect of curvature on the speed of finite reaction rate detonations are studied. One method involves fine grid computations using a method which gives a solution of high quality and is taken as exact. The other is based on recent work by one of the authors (J.J.) in which an asymptotic model for expanding detonation waves is presented and analyzed. Both methods assume cylindrical geometry.

The asymptotic model consists of a pair of quasi steady state ordinary differential equations for the flow velocity and a reaction progress variable. The equations are correct for large times and large radii. For each value of the shock radius, the speed of the weak detonation is well defined as the solution of a shooting problem between the shock and a critical point in the phase plane.

In this report we discuss the computational problems involved in applying these methods. We further show numerically that the model equations are accurate to first order in powers of the inverse radius. Finally, we discuss how this new theory may be used in conjunction with the method of front tracking to numerically solve detonation problems in which weak detonations develop due to the curvature of the geometry.

1. Introduction

We study the influence of the radius of curvature on the speed of a cylindrically expanding detonation wave with finite reaction rate. In doing so, we also study the transition from strong to weak detonations in an expanding geometry. Finally, we study the effect of curvature on the reaction zone. The central issue to be analyzed is the consequence of radially induced cooling on the chemical reaction. See [7] for a review of this topic and more generally of the theory of detonations in the presence of endothermic effects.

Two numerical methods are used to solve this problem. First, a one dimensional random choice computation with operator splitting for both the radial effects and the effects of the finite reaction rate is employed. Since this method resolves the reaction zone numerically (in contrast to [2]), it includes curvature effects on the detonation velocity. This method gives an accurate solution for grids fine enough to capture the dynamics within the reaction

67

zone. Next, we discuss the application of recent contributions by J. Jones [10], who derived a system of quasi steady state ordinary differential equations to describe expanding detonations. We solve these equations numerically to find the wave speed and to resolve the reaction zone. The solutions of Jones' equations are found to be correct to first order in powers of inverse radius thereby confirming and validating his analysis.

A motivation for this work was to enhance the front tracking algorithm (see [3]) to allow calculation of curvilinear detonation fronts in their transition from strong to weak detonations.

2. The Random Choice Computation

In this section, we discuss the solution to the equations of reactive gas dynamics with finite reaction rates in a symmetric geometry using the random choice method. The Zeld'ovich-Von Neumann-Doering (ZND) model of detonations (see [7]) is used. For this model, the equations of inviscid gas dynamics with cylindrical symmetry become

$$(2.1) \qquad w_t + f(w)_r = C - \alpha G ,$$

where

$$w = \begin{pmatrix} \rho \\ m \\ e \\ \lambda \end{pmatrix} \qquad f(w) = \begin{pmatrix} m \\ \dfrac{m^2}{\rho} + P \\ \dfrac{m}{\rho}(e + P) \\ m\dfrac{\lambda}{\rho} \end{pmatrix} ,$$

$$C = \begin{pmatrix} 0 \\ 0 \\ 0 \\ R(\lambda, T) \end{pmatrix} \qquad G = \begin{pmatrix} \dfrac{m}{r} \\ \dfrac{m^2}{\rho r} \\ \dfrac{m}{\rho r}(e + P) \\ 0 \end{pmatrix} ,$$

and

$$\alpha = \begin{cases} 0 & \text{for planar geometry} \\ 1 & \text{for cylindrical geometry} \\ 2 & \text{for spherical geometry} \end{cases}$$

C and αG are respectively, the source terms due to combustion and geometry. In these equations, ρ is the density of the gas, m is the momentum density, P is the pressure and λ is the mass fraction of burned gas ($0 \leq \lambda \leq 1$). The energy per unit volume, e, may be written as

$$e = \rho \epsilon + \frac{\rho u^2}{2}$$

where u is the velocity and ϵ is the specific internal energy. Assuming a polytropic equation of state,

68

$$\epsilon = \frac{P}{\rho(\gamma - 1)} + (1 - \lambda)q \ ,$$

with $\gamma > 1$. In order to simplify the formulas, the polytropic constant, γ, is assumed to have the same value in the unburned, burned and reacting gas. The heat released during combustion is q; T is the temperature ($T = P/\rho$) and $R(\lambda, T)$ is the reaction rate. We use Arrhenius kinetics, which yields an infinite reaction length. Thus,

$$R(\lambda, T) = \begin{cases} k(1 - \lambda) \exp\left(-\dfrac{E}{T}\right) & , \ T \geq T_c \\ 0 & , \ T < T_c \end{cases}$$

where k is the rate multiplier, and E is the activation energy. We introduce T_c, the critical temperature below which the reaction rate is taken to be identically zero, in order to allow for quenching and to eliminate the cold boundary effect. That is, if there were no critical temperature, the reaction rate would be positive even for cold gases. Then, the unburned gas would begin to burn before the shock wave encountered it.

To solve this system numerically, we employ operator splitting [13]. At the start of a time step, we solve the homogeneous system

$$(2.2) \qquad\qquad w_t + f(w)_r = 0$$

by the random choice method [9],[4]. The Newton's method of [2] is employed to solve the Riemann problems that arise in this computation. Next, we use the solution of eq. (2.2) as initial data for the system of ordinary differential equations for the geometrical source terms,

$$(2.3) \qquad\qquad w_t = -\alpha G \ ,$$

Finally, we use the solution of eq. (2.3) as initial data to solve

$$w_t = C ,$$

the equation for the source term due to chemistry. This sequential operator splitting calculation converges under mesh refinement. Colella, Majda and Roytburd [5] have used a three part splitting in their fractional step method computations for reacting gases.

A plot of pressure vs. distance for a stable planar reaction is shown in Fig. 2a at the start of a calculation, initialized with the steady state solution, and after several hundred time steps using the method described above. In addition, reactions which have parameters chosen to yield unstable detonations are modelled well by this method. In an example of an unstable detonation, our results agree with those of Erpenbeck [6], Mader [11] and Fickett and Wood [8] (see Fig. 2b).

We note that when performing these calculations one must take care not to introduce spurious effects due to the numerical modelling techniques. One should include enough grid points in the region of chemical activity in the reaction zone. Also, as the computation progresses, the region of chemical combustion grows for the Arrhenius model of kinetics. To deal with this problem, we eliminate the portion of the computational region more than a certain distance behind the initiating shock wave. In doing so, care must be taken to eliminate only regions in which there are very small variations in the gas states. Introducing even weak waves into the reacting gas in this elimination process is equivalent to releasing small amounts of energy on a slow time scale and can cause large errors. This phenomenon was studied by Bdzil [1].

3. The Asymptotic Method of Jones

In [10], J. Jones derived and analyzed a method of calculating the effect of curvature on the speed of an expanding cylindrical or spherical detonation wave to first order in powers of the inverse of radius of curvature. This theory also predicts the state of the gas through the reacting region.

The radius of curvature of the detonation wave is assumed to be much larger than the length of the reaction zone, where the reaction zone length is taken to be the distance from the initiating shock wave to the point at which 90% of the gas is burned. It is also assumed that the reaction has proceeded for many reaction zone lengths so that initial transients are eliminated. Thus the run has settled down to a quasi steady state. Further, the state of the unburned gas ahead of the shock is constant with zero velocity. Through the methods of perturbation theory, eliminating higher order terms, Jones derived the following system of ordinary differential equations from eq. (2.1):

$$(3.1) \qquad u_x = \frac{q(\gamma - 1)k(1 - \lambda)\exp\left(-\dfrac{E\gamma}{c^2}\right) - \dfrac{uc^2}{z}}{(\dot{z} - u)^2 - c^2}$$

$$\lambda_x = \frac{R(\lambda, T)}{\dot{z} - u}$$

$$c^2 = c_a^2 + \frac{\gamma - 1}{2}(\dot{z}^2 - (\dot{z} - u)^2) + q(\gamma - 1)\lambda$$

where c_a represents the sound speed in the unburned gas ahead of the shock, c is the speed of sound of the reacting gas, $c = \left\{\gamma P/\rho\right\}^{\frac{1}{2}}$, x is the distance behind the initiating shock wave, z is the radius of curvature of the shock, and $\dot{z}$ is the wave speed.

In the case of an undriven planar detonation, the reaction terminates at the Chapman-Jouguet (CJ) point on the Hugoniot curve. This is a sonic point. That is, a point at which the wave moves at sound speed with respect to the gas behind it. However, an expanding detonation is weakened by expansion induced rarefactions coming from behind the shock and the termination point for the reaction moves below the CJ point yielding a weak detonation. The flow is subsonic behind a shock but supersonic behind a weak detonation. Thus, a transition from subsonic to supersonic flow must occur in the reacting gas.

Since the denominator of the first of eqs. (3.1) vanishes at any sonic point, in order to have a smooth transition through a sonic point, the numerator must also vanish there. The transformation

$$y = \int \frac{dx'}{c(x')^2 - v(x')^2}$$

where $v = \dot{z} - u$, leads to the system

$$(3.2) \qquad v_y = q(\gamma - 1)k(1 - \lambda)\exp\left(-\frac{E\gamma}{c^2}\right) - \frac{\dot{z} - v}{z}c^2$$

$$\lambda_y = \frac{1}{v}k(1 - \lambda)\exp\left(-\frac{E\gamma}{c^2}\right)(c^2 - v^2)$$

70

where

$$c^2 = c_a^2 + \frac{\gamma - 1}{2}(\dot{z}^2 - v^2) + q(\gamma - 1)\lambda.$$

This transformation does not change the structure of the phase plane and the critical point conditions for this system are the same as the conditions for a smooth sonic transition mentioned above. These conditions are:

$$(3.3) \qquad q(\gamma - 1)k(1 - \lambda) \exp\left(-\frac{E\gamma}{c^2}\right) - \frac{\dot{z} - v}{z}c^2 = 0$$

$$\frac{1}{v}k(1 - \lambda) \exp\left(-\frac{E\gamma}{c^2}\right)(c^2 - v^2) = 0$$

$$c_a^2 + \frac{\gamma - 1}{2}(\dot{z}^2 - v^2) + q(\gamma - 1)\lambda = c^2.$$

From a computational point of view, eqs. (3.2) are easier to work with than eqs. (3.1), see also [10].

To solve for the wave speed and resolve the reaction zone we proceed as follows. We guess a value for v_0 (that is, v immediately behind the initiating shock) and solve for the critical point of the system of ordinary differential equations (3.2) by iterating on λ and v, using the equations

$$v = \left\{ \frac{2}{\gamma + 1} \left[c_a^2 + \frac{(\gamma - 1)\dot{z}^2}{2} + q(\gamma - 1)\lambda \right] \right\}^{\frac{1}{2}}$$

$$\lambda = 1 - \frac{(\dot{z} - v)c^2 \exp\left(\frac{E\gamma}{c^2}\right)}{qk(\gamma - 1)z}$$

which are derived from eqs. (3.3), using the fact that $v^2 = c^2$ at the sonic point. The square root takes the same sign as $\dot{z}$. We then integrate system (3.2) numerically with the initial conditions

$$v(0) = v_0$$
$$\lambda(0) = 0$$

to find the trajectory of the solution in the $v - \lambda$ plane for the given $\dot{z}$. We update the values of v_0 and $\dot{z}$ based on this trajectory by a bisection method until the trajectory passes within a specified tolerance of the critical point. The solution is continued through the critical point by finding the eigenvectors there. The equation, see [10],

$$p_y = -\rho v v_y$$

and the last of eqs. (3.2) are used to compute the pressure and density through the reacting region.

In Fig. 3a, we present the $v - \lambda$ phase plane portrait as well as the sonic locus for the value of $\dot{z}$ which yields a sonic transition for the given data. The curve passes through the critical point (S) after which the burning continues on the supersonic side of the sonic locus.

71

4. Results

In Fig. 4a, we compare a plot of the pressure, immediately behind the shock wave which initiates the detonation, vs. time for a planar CJ detonation, computed by the random choice method described in §2 ($\alpha = 0$), with a plot of the results from a cylindrical computation ($\alpha = 1$) and with the results of the method of Jones where the radius of curvature is assumed to be the same as for the cylindrical computation at all times. The vertical error bars give the highest and lowest values of pressure over each 100 time steps for the random choice computations. As expected, the pressures computed by the two cylindrical methods approach the planar CJ pressure just behind the shock as the radius of curvature increases. We note that the random choice computations are initialized with the planar steady state solution and it takes some time for the initial transients to disappear in the cylindrical run. When we initialized the random choice method with the results of Jones' method at a small radius the transients were much smaller but the results for larger radii were not significantly different from those of the planar initialization. A comparison of the pressures behind the shock wave using these initializations is seen in Fig. 4b. In these computations, we eliminated the regions more than 3 reaction zone lengths behind the initiating shock wave.

To exhibit the validity of Jones' method to leading order in inverse radius, we present, in Fig. 4c, a plot of pressure behind the initiating shock wave vs. inverse radius for the numerical methods described in §2 and §3. We also show the line predicted by the theory of Jones for the leading order corrections to pressure due to curvature, based on computations using Jones' equations with very large radius of curvature. The oscillations in the random choice computation are due to the numerical method and decrease with refinement of the grid. A similar plot is achieved for the corrections of detonation wave speed due to curvature (Fig. 4d).

Fig. 4e shows the states of a reacting gas for a planar CJ detonation, for an expanding cylindrical detonation using the method described in §2, and for the method of Jones at a fixed time. We have plotted pressure vs. specific volume along with the unburned and burned Hugoniot curves. It is plotted at a time when the radius of curvature is approximately 50 times the length of the reaction zone. The steady state planar wave is initiated by a shock and moves down the Rayleigh line from A to CJ (the CJ point) as the reaction progresses. This line, when extended, passes through the point representing the initial ahead state. The detonation waves for the two cylindrical methods are initiated by weaker shocks, corresponding to a lower pressure on the unburned Hugoniot (points F and D), and move down along the curves shown to a weak detonation. These curves do not terminate on the burned Hugoniot curve since it is computed from a planar theory. Wood and Kirkwood [14] have derived equations for the modifications of Hugoniot curves in a curved geometry. From Fig. 4e, we see the effect of curvature on the pressure just behind the shock and through the reaction zone. These computations show that a 12.5-17.5% smaller jump in pressure at the shock occurs in the curved geometry than in the corresponding planar calculation. Here the radius of curvature was approximately 50 times the reaction zone width.

5. Conclusions

We have shown that the derivation by J. Jones [10] of the corrections due to curvature of the speed of detonation waves and the pressure behind the shock wave are correct to first order in the inverse of radius of curvature. This validation was necessary since the passage from the original system of partial differential equations (2.1) to the ordinary differential equations (3.2) has not been shown rigorously.

The advantage of using the equations of Jones is considerable. Solving for the wave speed and the states of the reacting gas is usually accomplished in less than 10 CPU seconds on the ELXSI. We note that the rapid computation of wave speeds will enable two dimensional front tracking computations to be extended to include the effects of curvature on detonation waves in the near future. Although not directly comparable, the one dimensional computations with fully resolved chemical reactions, fine grids and Arrhenius kinetics,

72

exhibited in Figs. 4a-e, require approximately 40 hours of CPU time on the same machine. To avoid these slow computations, for practical computations one would normally use neither resolved chemical reactions, nor fine grids, nor Arrhenius kinetics.

The enhancement of the front tracking method ([3],[12]) using Jones' equations would involve modelling the reaction zone as being infinitely thin. Based on the divergence of the flow at each point on the detonation wave front, one would use the theory of Jones to find the speed of propagation of the detonation front at that point and the state of the gas behind the initiating shock wave as well as the state of the gas behind the completed chemical reaction. In this way, the transition from strong or CJ detonations to weak detonations can be modelled for two dimensional flows. This is not possible with the model employed in [2].

6. Acknowledgements

References

[1] J. B. BDZIL, "*Perturbation Methods Applied to Problems in Detonation Physics,*" Sixth International Symposium on Detonation, pp.352-370, 1977.

[2] B. BUKIET, "*Application of Front Tracking to Two Dimensional Curved Detonation Fronts,*" Submitted to SIAM J. Sci. Stat. Comp.

[3] I.L. CHERN, J. GLIMM, O. MCBRYAN, B. PLOHR, "*Front Tracking for Gas Dynamics,*" J. Comp. Phys., vol. 62, pp.83-110, 1986.

[4] A. J. CHORIN, "*Random Choice Solution of Hyperbolic Systems,*" J. Comp. Phys., vol. 22, pp.517-533, 1976.

[5] P. COLELLA, A. MAJDA AND V. ROYTBURD, "*Fractional Step Methods for Reacting Shock Waves,*" Lectures in Applied Mathematics, vol. 24, pp. 459-477, AMS, Providence, R.I. 1986.

[6] J. J. ERPENBECK, "*Stability of Idealized One-Reaction Detonations,*" Phys. Fluids, vol. 7, pp. 684-696, 1964.

[7] W. FICKETT AND W. C. DAVIS, "*Detonation,*" University of California Press, Berkeley, 1979.

[8] W. FICKETT AND W. W. WOOD, "*Flow Calculations for Pulsating One-Dimensional Detonations*", Phys. Fluids, vol. 9, pp. 903-916, 1966.

[9] J. GLIMM, "*Solutions in the Large for Nonlinear Hyperbolic Systems of Equations,*" Comm. Pure Appl. Math. vol. 18, pp.695-715, 1965.

[10] J. JONES, "*Asymptotic Analysis of an Expanding Detonation,*" In preparation, 1986.

[11] C. L. MADER, "*Numerical Modelling of Detonations,*" University of California Press, Berkeley, 1979.

[12] R. D. RICHTMYER AND K. W. MORTON, "*Difference Methods for Initial Value Problems,*" Interscience Publishers, New York, 1967.

[13] G. A SOD, "*A Numerical Study of a Converging Cylindrical Shock,*" J. Fluid Mech., vol. 83, pp.785-794, 1977.

[14] W. W. WOOD AND J. G. KIRKWOOD, "*Diameter Effect in Condensed Explosives. The relation between velocity and radius of curvature in the detonation wave,*" J. Chem. Phys., vol. 22, pp. 1920-4, 1954.

73

Pressure

600

500

400

300

0   10
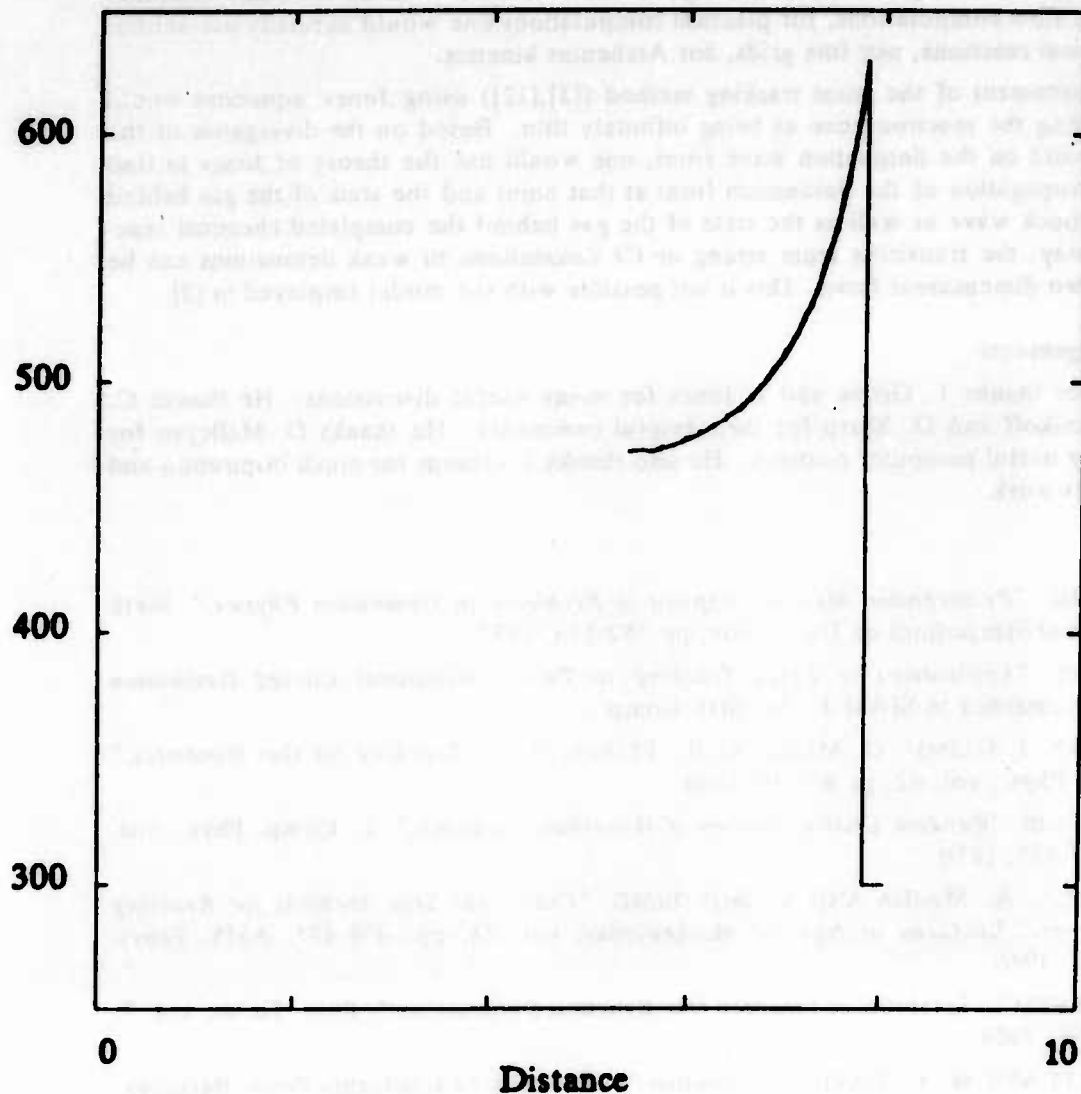
Distance

Fig. 2a. A 1D Stable Computation. *A plot of pressure vs. distance is shown for a planar detonation initialized with the steady state ZND solution. The initialized reaction and the reaction 1200 time steps later are superimposed so that both fronts are at the same location on the graph. The solution is clearly not deformed by the numerical method. The state ahead of the initiating shock (in units where the gas constant $R = 1$) has $P = 100$, $u = 0$, $\rho = 1.4$. The heat release ,q, is 300; $E = 100$; $T_c = 215$ and $\gamma = 1.1$. The distance from the initiating shock wave to the point where the gas is 90% burned is 0.75 and the grid spacing is 0.0125.*

74
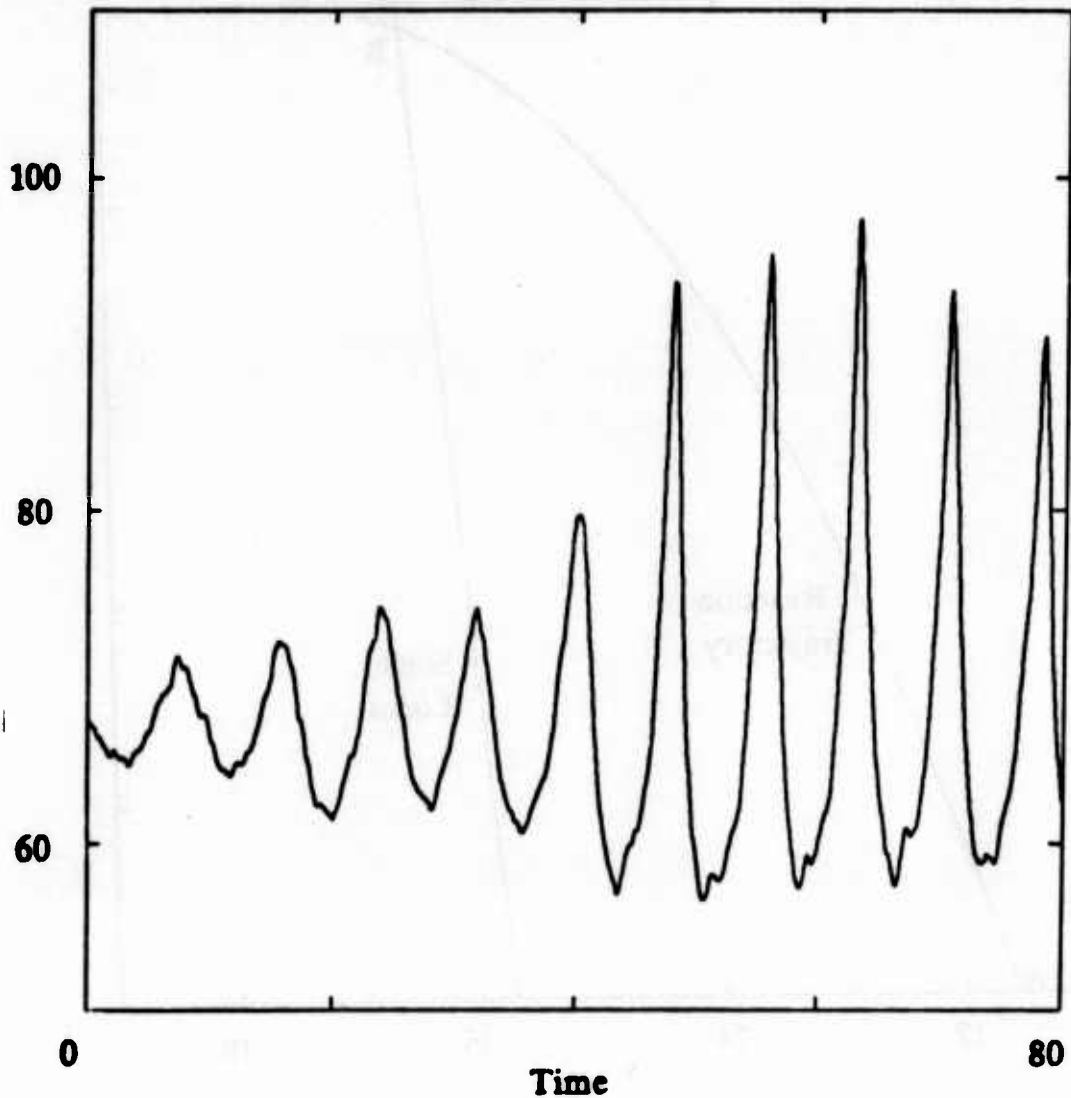
Pressure

Fig. 2b. A 1D Unstable Computation. *A plot of pressure behind the shock initiating the detonation vs. time for the data described in [11, pp. 18-19]. The reaction zone was initialized with length 1.75. The ahead state has $p = 1, u = 0, \rho = 1$. $q = 50, E = 50, \gamma = 1.2, k = 206$. The speed of the initialized wave is 1.265 times the CJ wave speed for the ahead state. Grid spacing is 0.05. The results are similar to those in [6] and [11].*
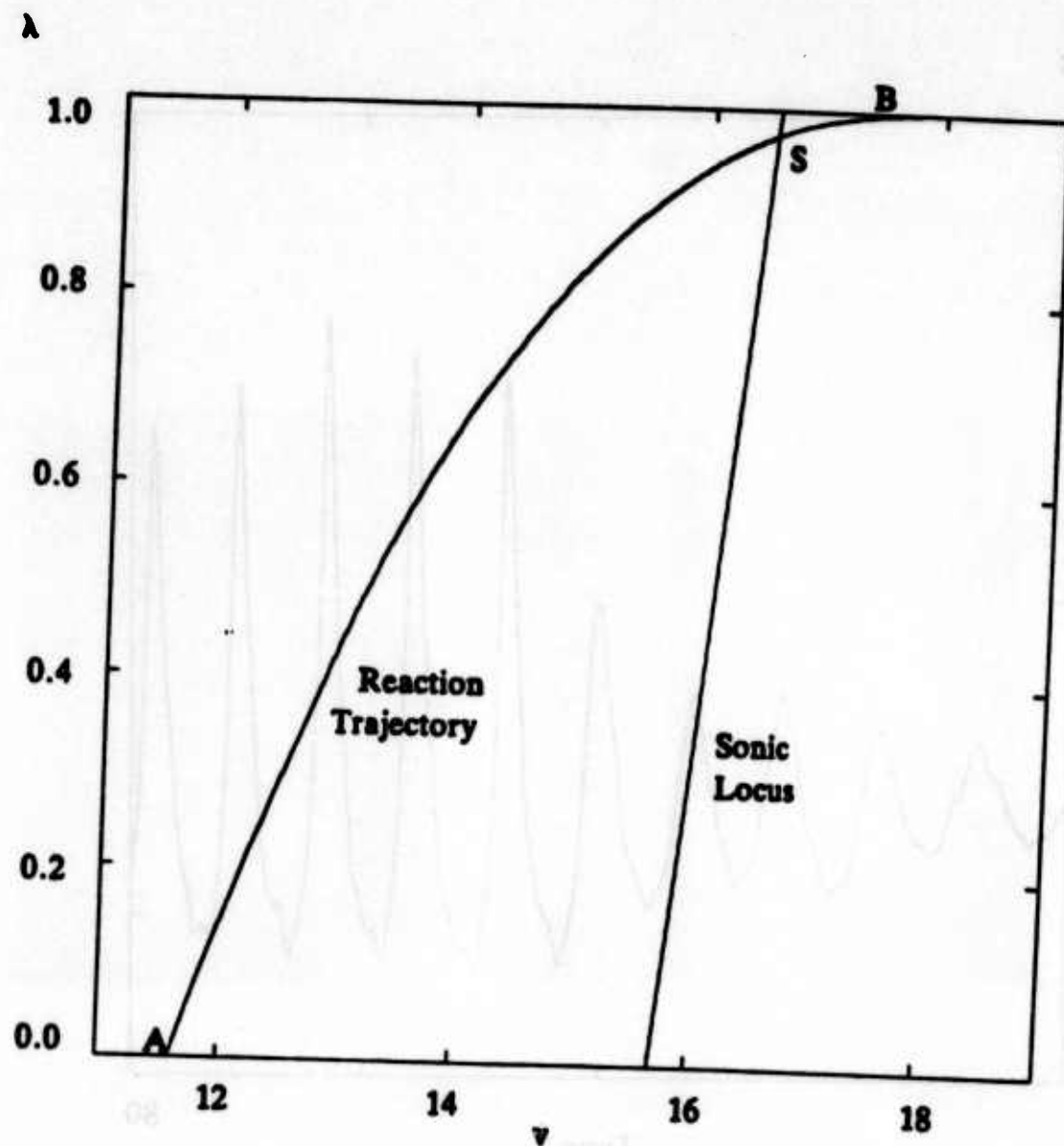
Fig. 3a. Phase Plane Portrait of the Reaction using Jones' Equations. *A plot of the trajectory through the sonic point (S) in the v − λ plane for the runs in the following figures when the radius of curvature is 50 times the reaction zone length. The sonic locus is also shown. A corresponds to the point behind the initiating shock wave while B represents the termination of the the reaction as a weak detonation.*

76
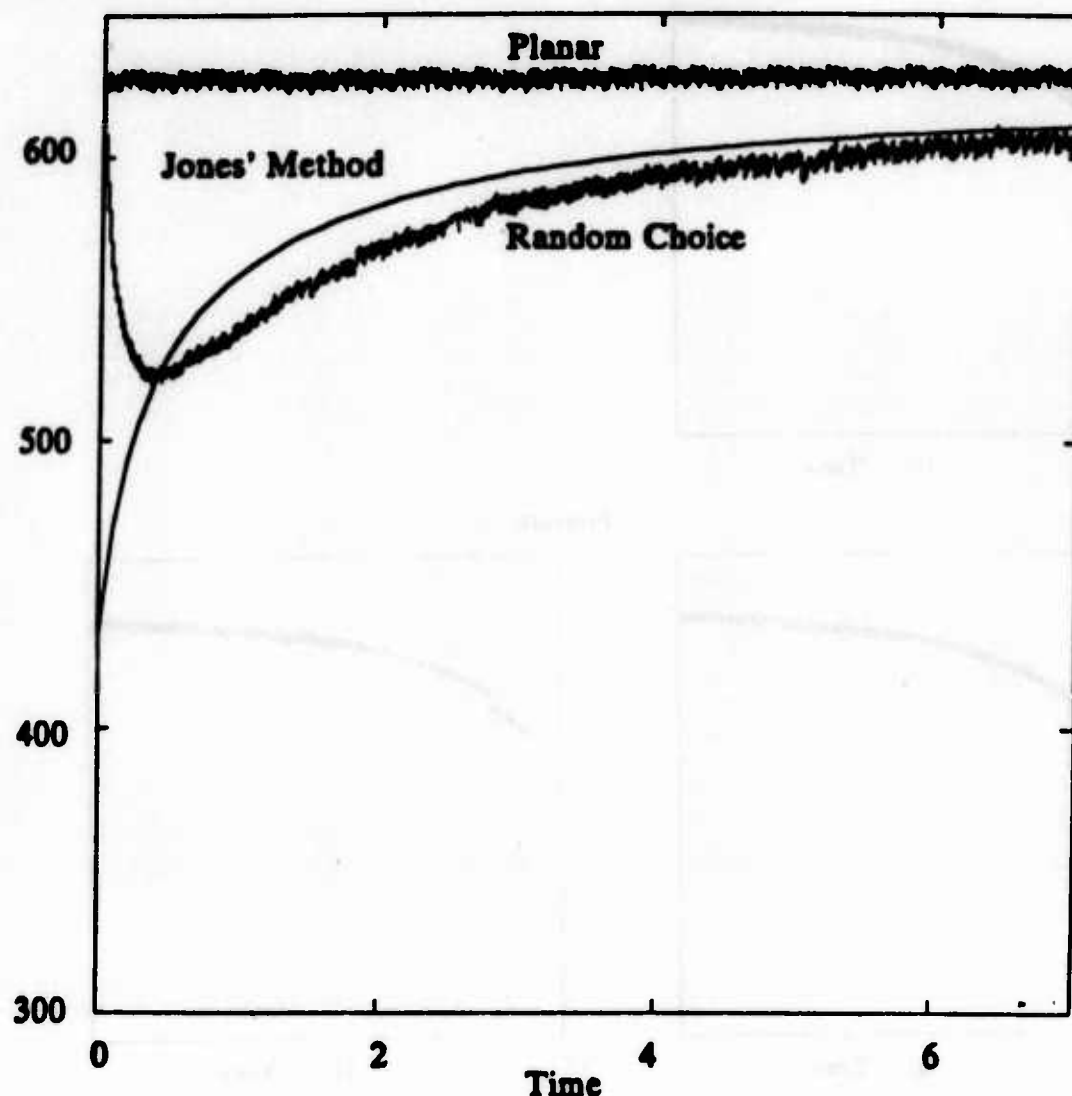
Fig. 4a. Effect of Curvature on Pressure behind the Initiating Shock. *A plot of pressure vs. time for planar and cylindrical computations using the random choice method of 52 and the solution to Jones' equations where the radius of curvature is assumed to be the same as for the cylindrical run by random choice. The error bars show the range of values over each 100 time steps for the random choice calculations. The ahead state has $p = 300$, $u = 0$, $\rho = 1.4$. The reaction zone has length 1, $q = 300$, $T_c = 215$, $\gamma = 1.1$, $E = 100$ and grid spacing 0.01.*
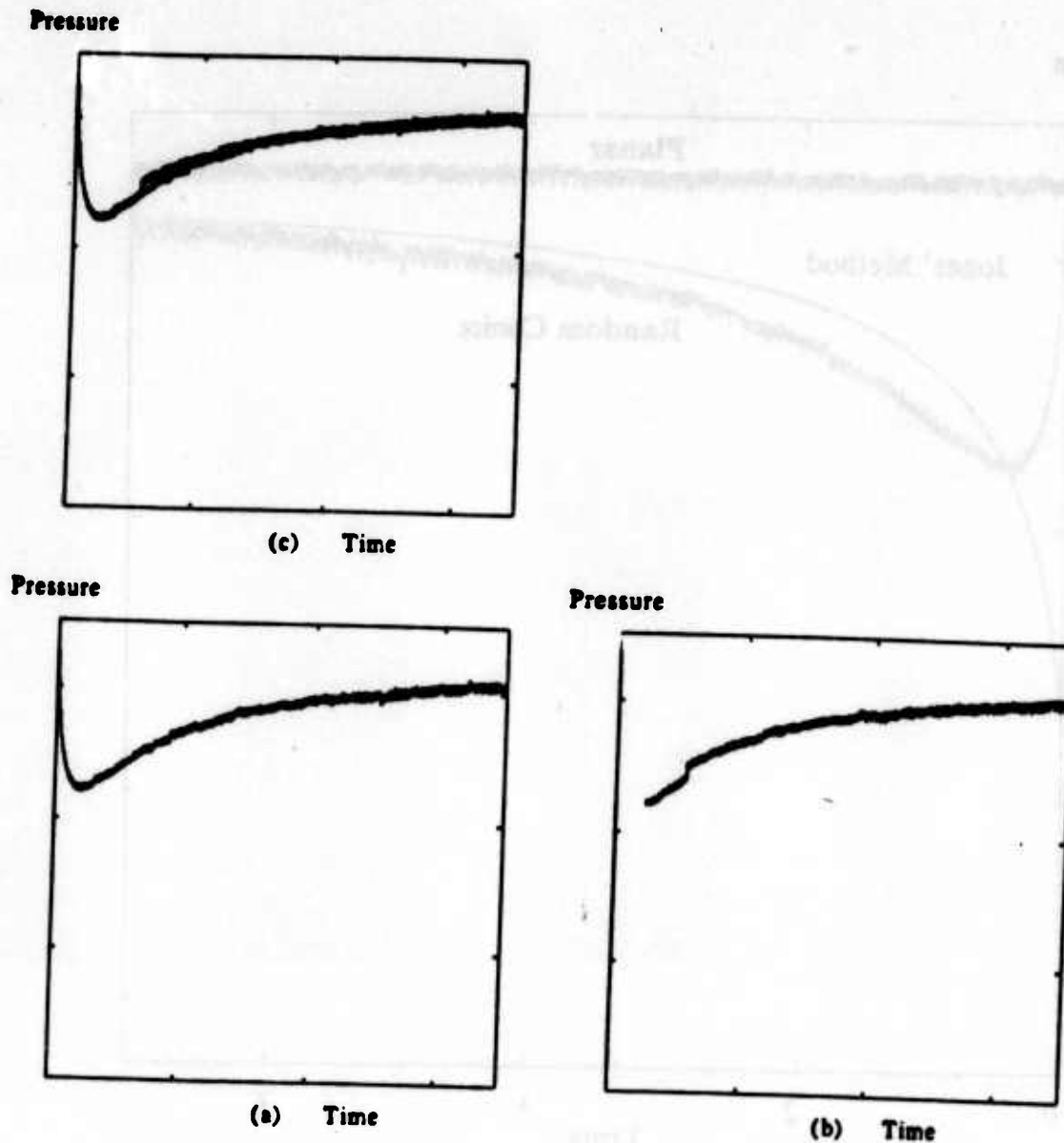
Fig. 4b. Comparison of Initializations. *The panels above show plots of pressure behind the initiating shock as a function of time for the same cylindrically expanding detonation problem as in Fig. 4a using the planar steady state initialization (a) and initialization by solution to Jones's method at a small radius (b). These two plots are superimposed in (c).*
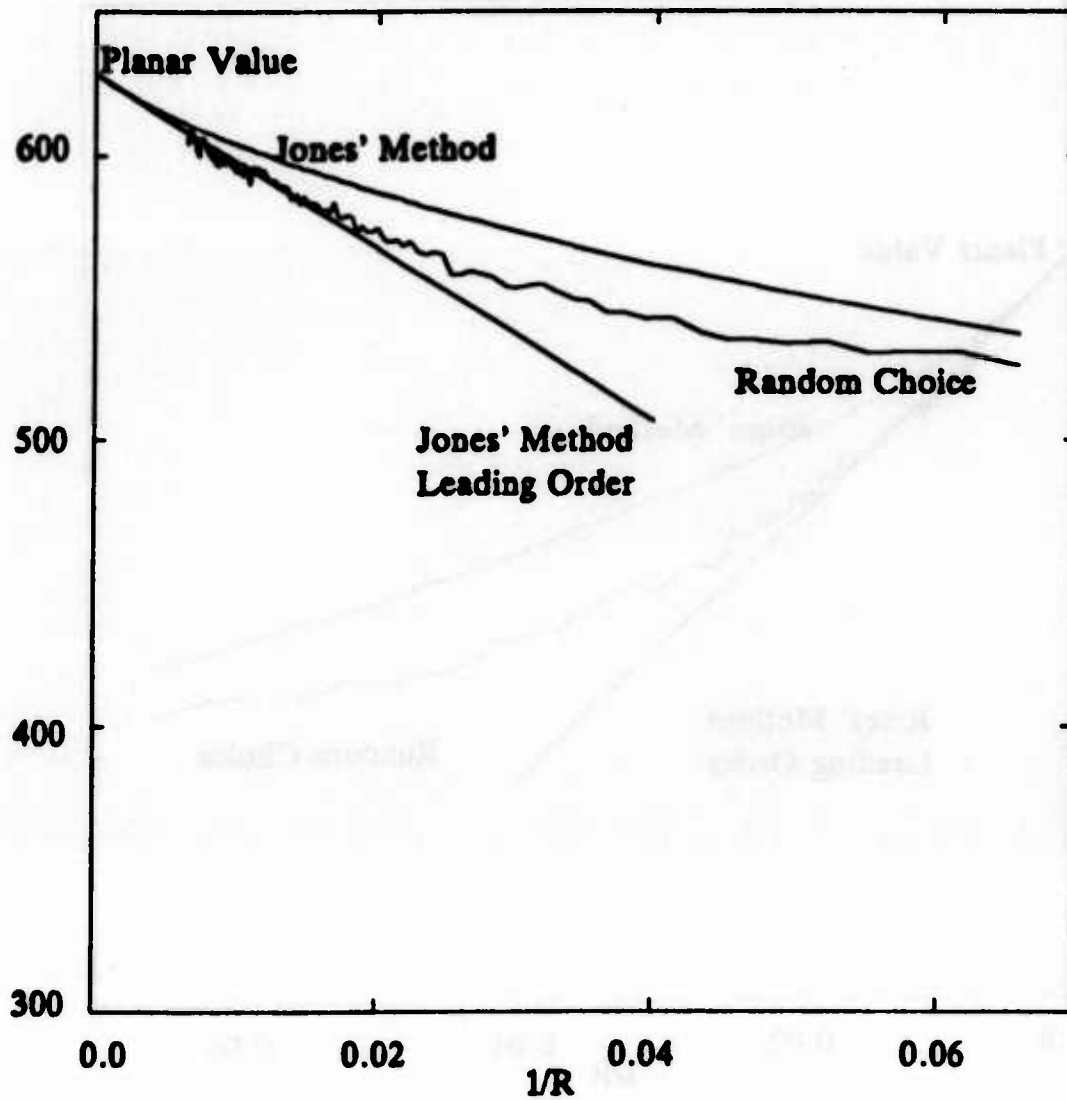
Fig. 4c. First Order Corrections to Pressure. *Pressure behind the initiating shock wave is plotted against inverse radius for the cylindrical computations of Fig. 4a. Also shown is the leading order correction predicted by solving Jones' equations for very large radii.*
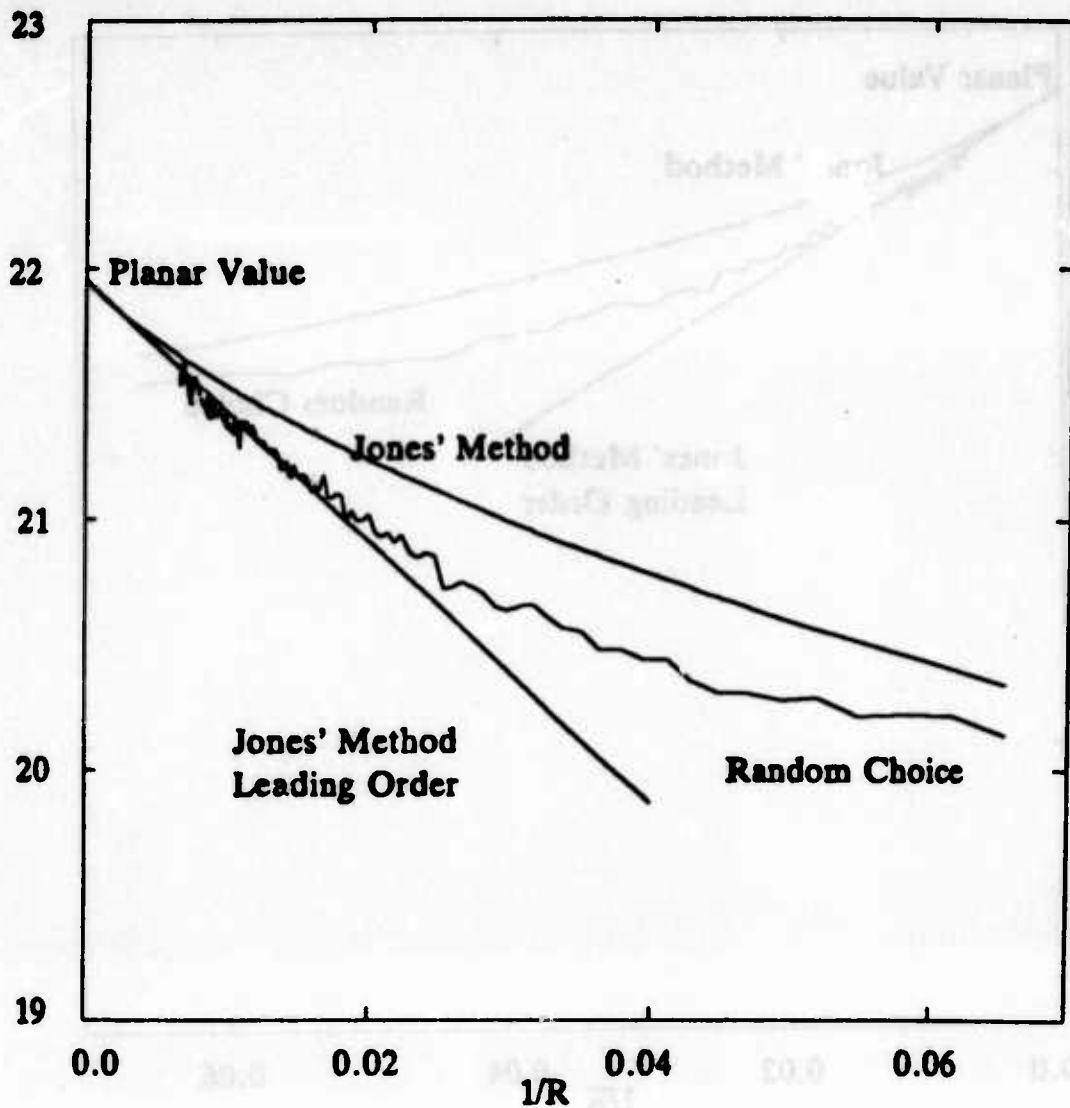
Fig. 4d. First Order Corrections to Wave Speed. *A plot of wave speed vs. inverse radius is shown corresponding to* Fig. 4c.

Pressure
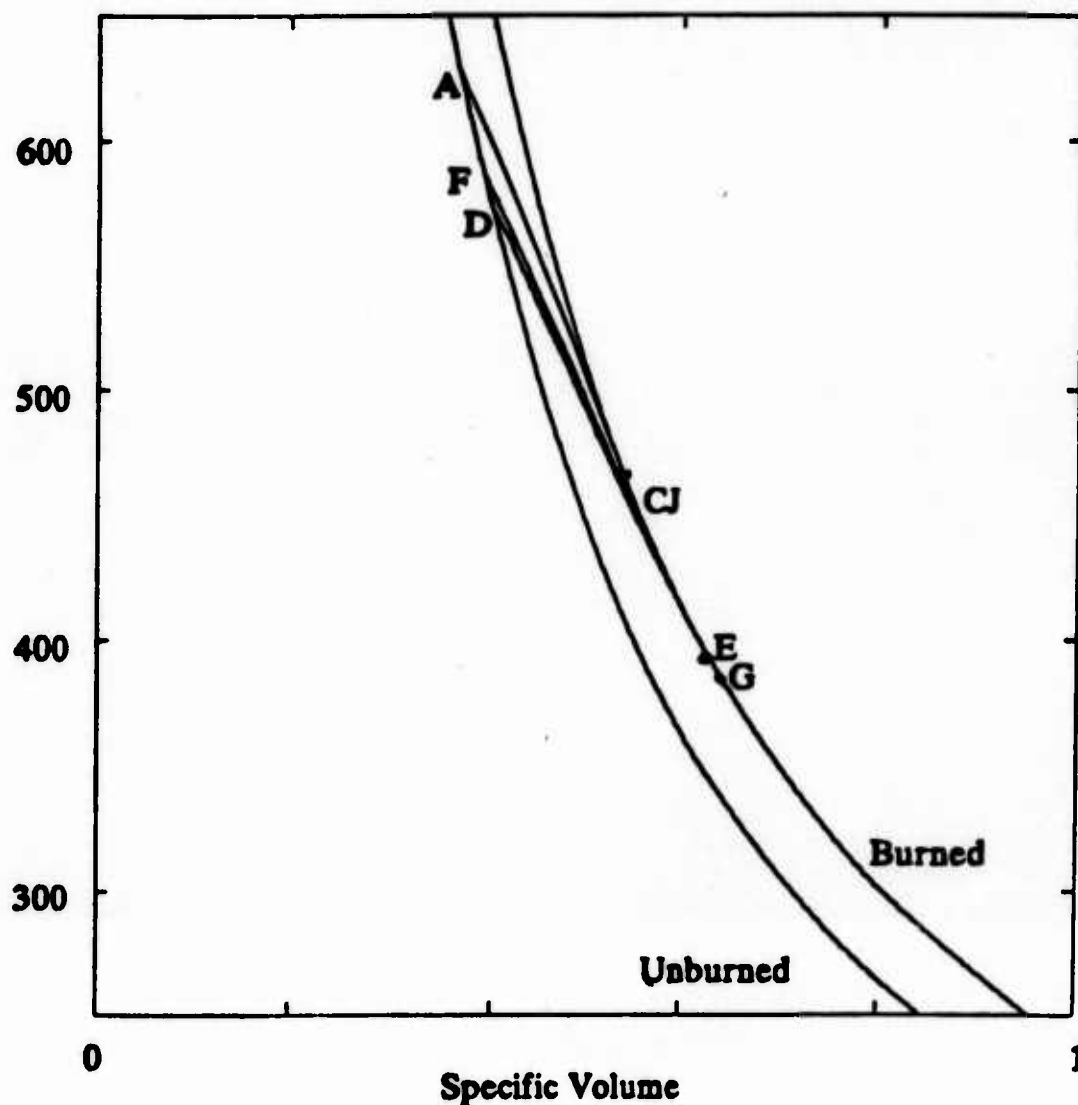
Fig. 4e. Effect of Curvature on the Hugoniot Diagram. *Pressure is plotted against specific volume for the calculations used in Figs. 4a-d. The unburned and burned Hugoniot curves are presented as well as the path through the reaction zone for a planar detonation (from A to CJ) and cylindrical detonations where the radius of curvature is approximately 59 times the reaction zone length by the random choice method (from D to E) and by Jones's method (from F to G). The pressure jump at the front is reduced by 12.5-17.5% by the curvature.*

81

PRESSURE TRANSIENTS IN A CAVITY DUE TO IMPULSIVE LOADS

C. Helleur

Defence Research Establishment, Valcartier, Canada

B. Tabarrok and R. G. Fenton

Department of Mechanical Engineering, University of Toronto

Abstract

The impact of a mass on a structural plate of a compartment causes plastic and elastic deformations which can give rise to pressure fluctuations of significant magnitude and duration due to the confined nature of the compartment and the low damping forces in the gas. This paper presents a procedure for calculating the pressure transients in an enclosed gas from an impact.

The procedure uses a formulation which gives the equations of motion in terms of a scalar momentum potential. This momentum potential is physically interpretable as a pressure impulse. With this formulation the transient pressure behavior of the gas is characterized by a single partial differential equation which is the wave equation in three dimensions. The spatial derivatives are treated by a finite element technique to obtain solutions for an arbitrary geometry of the enclosure.

The boundary conditions for the problem are that the normal velocity of the gas is compatible with the prescibed velocity of the enclosure walls. The rate of deformation of the wall resulting from the impact is approximated by modelling the region in the form of two concentric plastic hinges. Since the stress in the hinges must be at the yield value, it is possible to approximate the plastic deformation and hence the time history of the deformation.

1   Introduction

The subject of a projectile penetrating the wall of a cavity has received a great deal of attention because of possible military applications. A projectile which failed to perforate was considered of little interest. However there is increased interest in the transient response of an enclosed gas resulting from a mass impacting a cavity wall since it may cause sufficient deflection of the wall to give rise to pressure fluctuations of significant magnitude and duration.

This paper presents a procedure for calculating the pressure transients in an enclosed gas resulting from a deformation of the cavity wall. The suitability of this procedure to investigating pressure transients resulting from an projectile impact on the cavity wall is investigated.

2   Equations of Motion of a Gas

The force balance on an element of the gas is shown in figure 1. The components of displacement in the x_1, x_2, x_3 direction are denoted as u_1, u_2, u_3 respectively and the pressure is denoted with a 'p'.

Starting from the force balance in the x_1-direction we obtain

$$(p - (p + \frac{\partial p}{\partial x_1} dx_1))dx_2\, dx_3 = \rho dx_1\, dx_2\, dx_3 \frac{\partial^2 u_1}{\partial t^2} \tag{1}$$

which yields

$$-\frac{\partial p}{\partial x_1} = \rho \frac{\partial^2 u_1}{\partial t^2} \tag{2}$$

which can be generalized for the 'i' direction

$$-\frac{\partial p}{\partial x_i} = \rho \frac{\partial^2 u_i}{\partial t^2} \tag{3}$$

These represent the equilibrium equations of the gas. This formulation has the disadvantage that one can have infinite solutions for which u_1, u_2, u_3 are not zero whereas the volumetric strain is zero (i.e spurious solutions).

Impulse Formulation   In order to reduce the problem to a convenient form for solving the pressure impulse formulation of ref. [1] and [2] (i.e. $q = \int p dt$) is used.

The cause-effect relationship involving pressure impulse and the volumetric strain is shown here as the rate of change in pressure equal to minus the bulk modulus (K) times the volumetric strain (ϵ_{vol}).

$$\dot{q} = p = -K\,\epsilon_{vol} = -K(\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}) \tag{4}$$

This represents the constitutive equations of the gas in terms of the pressure impulse.
84

Differentiating eq. 3 with respect to x_1, x_2, x_3 respectively and substituting into eq. 4, we obtain eq. 5

$$\nabla^2 q = \frac{1}{C^2}\ddot{q} \tag{5}$$

$$\text{where } C^2 = \frac{K}{\rho}$$

On a fixed rigid wall, the normal displacement and hence also the normal velocity must vanish. For a surface which has a prescribed velocity V_n the normal component of the velocity of the gas and the wall must match.

$$V_n = \frac{du_n}{dt}$$

Therefore it follows from eq. 3 that

$$\frac{\partial q}{\partial n} = \rho V_n \tag{6}$$

This formulation has the advantage that it requires the solution of only one equation and that the boundary conditions are in terms of velocity. Analytical solutions to eq. 5 and eq. 6 can be found by classical means for cases where the physical problem has a simple geometry and boundary condition.

3 Finite Element Model

Since we wish to be able to model cavities of arbitrary geometry, it is necessary to solve the equation using finite elements. For this purpose it is useful to cast the essential equations of the problem into a variational statement.

Using the complementary energy principle, the increment of work done by violation of eq. 5 and eq. 6 results in eq. 7.

$$\int_V (\frac{-1}{\rho}\nabla^2 q + \frac{1}{K}\frac{\partial^2 q}{\partial t^2})\delta q \, dV + \int_s (\frac{1}{\rho}\frac{\partial q}{\partial n} + V_n)\delta q \, ds = 0 \tag{7}$$

Using Green's theorem, integrating with respect to time and integrating by parts we obtain the required variational statement

$$\int_{t_1}^{t_2}\int_V \left(\frac{\delta}{2\rho}\left((\frac{\partial q}{\partial x_1})^2 + (\frac{\partial q}{\partial x_2})^2 + (\frac{\partial q}{\partial x_3})^2 \right) - \frac{1}{2K}(\frac{\partial q}{\partial t})^2 \right) dV \, dt$$
$$+ \int_{t_1}^{t_2}\int_s V_n \delta q \, ds = 0 \tag{8}$$

The first integral is over the volume of the gas and the second integral is over that part of the surface where the normal velocity V_n is prescribed.

For use in cavities of arbitrary geometry an 8 noded isoparametric element (ref. [3]) as shown in figure 2 is used. The six faces of the element can be quadrilaterals of arbitrary shape, however, the 4 edges of the quadrilateral must be straight. The shape functions

$$
\begin{aligned}
N_1 &= 1/8(1 - \alpha_1)(1 - \alpha_2)(1 - \alpha_3) \\
N_2 &= 1/8(1 - \alpha_1)(1 + \alpha_2)(1 - \alpha_3) \\
N_3 &= 1/8(1 + \alpha_1)(1 + \alpha_2)(1 - \alpha_3) \\
N_4 &= 1/8(1 + \alpha_1)(1 - \alpha_2)(1 - \alpha_3) \\
N_5 &= 1/8(1 - \alpha_1)(1 - \alpha_2)(1 + \alpha_3) \\
N_6 &= 1/8(1 - \alpha_1)(1 + \alpha_2)(1 + \alpha_3) \\
N_7 &= 1/8(1 + \alpha_1)(1 + \alpha_2)(1 + \alpha_3) \\
N_8 &= 1/8(1 + \alpha_1)(1 - \alpha_2)(1 + \alpha_3)
\end{aligned}
\tag{9}
$$

transform the x_1, x_2, x_3 coordinates into the $\alpha_1, \alpha_2, \alpha_3$ coordinate system such that the 8 node element is transformed into a regular cube. Since the elements are isoparametric, the same shape function is used to interpolate the impulses (eq. 10) as is used to transform the coordinates.

$$
q(\alpha_1, \alpha_2, \alpha_3) = [N_1, N_2, \ldots N_8]
\begin{bmatrix}
q_1 \\
q_2 \\
\vdots \\
q_8
\end{bmatrix}
\tag{10}
$$

The impulse gradient can be expressed in terms of the nodal pressure impulses.

$$
\{\frac{\partial q}{\partial \alpha}\} = [B]\{q\}
\tag{11}
$$

The Jacobian matrix of transformation, represented here in symbolic form by the letter J, allows us to transform the impulse gradient from the x_1, x_2, x_3 coordinate system to the $\alpha_1, \alpha_2, \alpha_3$ coordinate system.

$$
\{\frac{\partial q}{\partial \alpha}\} = [J]\{\frac{\partial q}{\partial x}\}
\tag{12}
$$

The variational statement (eq. 8) has three terms. The first, shown in eq. 13, can be reduced and integrated, using eq. 11 and eq. 12, to produce the stiffness matrix $[K_e]$.

$$
\iiint \frac{\delta}{2\rho} \left[\{\frac{\partial q}{\partial x}\}^T \{\frac{\partial q}{\partial x}\} \right] dx_1 dx_2 dx_3 = \iiint \{q\}^T [B]^T [J^{-1}]^T [J^{-1}][B]\{q\}|J| d\alpha_1 d\alpha_2 d\alpha_3
$$

$$
= \frac{\delta}{2}\{q\}^T [K_e]\{q\}
\tag{13}
$$

It is important to bear in mind that in eq. 13 we have the discrete form of the kinetic energy and the dimensions of $[K_e]$ elements are velocity per unit momentum.

Consider the second term in the functional given in eq. 8. With a change in variables and integrating numerically using the Gauss-Legendre methods one obtains eq. 14.

$$\iiint \frac{\delta}{K} \left(\frac{dq}{dt}\right)^2 dx_1 dx_2 dx_3 = \delta\{\dot{q}\} \left[\iiint \frac{1}{2K}\{N\}^T\{N\}|J|d\alpha_1 d\alpha_2 d\alpha_3\right]\{\dot{q}\}$$

$$= \frac{\delta}{2}\{\dot{q}\}^T[M_e]\{\dot{q}\} \qquad (14)$$

$[M_e]$ is refered to as the element mass matrix but it should be noted that the dimensions of Me elements are displacements per unit force. In reality Me is the flexibility matrix.

Similarly the third term can be evaluated to produce the virtual work of the prescribed velocities in discrete form as follows

$$\int_s V_n \delta q ds = \iint \bar{V}_n \delta q \sqrt{1 + \left(\frac{\partial x_3}{\partial x_1}\right)^2 + \left(\frac{\partial x_3}{\partial x_2}\right)^2} \, dx_1 \, dx_2$$

$$= \iint \bar{V}_n \delta\{q\}^T\{N_s\}\sqrt{1 + \{x_3\}^T[B_s]^T[J_s^{-1}]^T[J_s^{-1}][B_s]\{x_3\}}|J|d\beta_1 \, d\beta_2$$

where $\{N_s\}$ = shape function on the face where V is prescribed
$\quad\quad\;\; [B_s]$ = gradient of shape function on the face where V is prescribed
$\quad\quad\;\; [J_s]$ = jacobian on the face where V is prescribed

Integrating using the Gauss Legendre method.

$$\int_s V_n \delta q ds = \delta\{q\}^T\{\bar{V}_e\} \qquad (15)$$

Having found the discrete forms of kinetic energy, complementary strain energy and the virtual work of prescibed velocities, we can write the discrete form of the Complementary Energy principle.

$$\int_{t_1}^{t_2} \left(\frac{\delta}{2}\{q\}^T[K_e]\{q\} - \frac{\delta}{2}\{\dot{q}\}^T[M_e]\{\dot{q}\} - \delta\{q\}^T\{V_e\}\right) dt$$

Expressing the functional in terms of the global matrices and carrying out the extremization we find the discrete equations of motion.

$$[K_g]\{q\} + [M_g]\{\ddot{q}\} = \{V_g\} \qquad (16)$$

To check that element matrices are correctly computed and assembled, natural frequencies of a cavity, modelled by different number of elements, were computed. The test cavity is a unit cube for which the natural frequencies can be determined analytically.

The first frequency is the zero frequency associated with a mode wherein the impulse q is constant and the mode associated with the first two non-zero frequency are standing waves for q, in the form of a cosine function.

The results of the investigation are shown in Fig. 3 as the convergence of the computed natural frequencies to the known natural frequencies as a function of the number of elements in the x_1-direction. These results verify the correctness of the element matricies as well as the connection process and give a good indication of the order of accuracy one can expect from eight noded elements.

Solution of Transient Equation   The matrix eq. 16 can be integrated using a moving polynomial solution

$$\{q\} = \{a_1\} + \{a_2\}t + \{a_3\}t^2 \tag{17}$$

Choosing equally spaced previous values of $\{q\}$, it is possible to solve for the $\{a\}$'s of eq. 17 and substituting into eq. 16 to result in eq. 18.

$$\left(\frac{1}{\Delta t^2}[M_g] + [K_g]\right)\{q_o\} = \{V_{q_{..}}\} + \frac{1}{\Delta t^2}[M_g]\left(2\{q_{-1}\{-\{q_{-2}\}\right)$$

or

$$[A]\{q_o\} = \{b\} \tag{18}$$

The response of the model is dependent on the number of elements used. To evaluate these effects we consider again the unit cube made up of two elements in the x_2 and x_3 direction and a variable number of elements in the x_1 direction. A velocity is imposed on the central node of the x_2, x_3 face of the cube which is initially at rest at time zero. The results are shown in Fig. 4 as the maximum pressure in the cavity as a function of time for different numbers of elements in the x_1 direction.

The response of the system will also vary with time step size(e.g. numerical damping). To illustrate this a unit cube is again subjected to a velocity at its center node. The results are shown in Fig. 5 as the maximum pressure in the cavity versus time for various time steps.

Due to the short time required for the cavity wall to reach its maximum velocity, it is essential that the model be capable of simulating the higher frequency components of the gas. This along with the fact that the mesh must be fine enough to adequately represent the localized applied velocity, makes it necessary to use a fine mesh and small time step. The large number of element will require a large amount of computer storage but matrix [A] is eq. 18 need only be inverted once if the integration time step is kept constant.

88

4 Application to Cavity Impact

The pressure pulse arises from the boundary condition that the normal velocity of the gas is compatible with the enclosed wall. Since this analysis was developed for the purpose of estimating pressure transients in a cavity resulting from an object impacting the cavity wall, a procedure for approximating the boundary conditions resulting from an impact has been included.

Fig. 6 shows a blunt object striking a cavity wall at normal incidence. The deformation process is divided into two phases. In the first phase the bulge, modelled by two concentric hinges, is accelerating until it attains the projectiles velocity. The second phase consists of the hinge and projectile decelerating together.

It is possible to approximate the velovity V_o and the times T_o and T_f using a procedure similar to that used in ref. [4]. The thrust on the target, F, is given by

$$F = [\sigma_{Y_c} + (V_p - V_b)^2 \rho] A_p$$

where A_p = cross-sectional area of the projectile

V_p = velocity of the projectile

V_b = velocity of the bulge

σ_{Y_c} = constrained uniaxial yield stress

This allows us to approximate the velocity history of the projectile and the cavity wall during the impact.

Fig. 7 shows the results for a 3 Kg projectile with a 60mm diameter and a velocity of 1Km/sec striking a hemisperical cavity with a 1M radius and a thickness of 30mm. The results are shown here as the pressure at the point of impact and at a point located at the center of the cavity versus time. The results show that severe pressures of very short duration will result.

5 Conclusion

In this paper a procedure for calculating the pressure transients in a cavity using a finite element method has been demonstrated. The results shown in this paper suggest that the procedure is suitable for determining the pressure transients in an armoured vehicle subjected to an impact subject to the conditions that the striking projectile does not perforate the cavity wall and that the velocity of the wall is below the speed of sound in the gas. This is generally the case when a velocity of a projectile is below the ballistic limit of a cavity wall.

References

[1] B. Tabarrok, Dual Formation for Acouto-Structura; Vibrations, *Int. J. num. Meth. Engng*, 13, 197-201 (1978).

[2] W. R. C. Underhill, Transient Pressure Variation of an Enclosed Gas by a Finite Element Method. Dissertation (M.A.Sc.) University of Toronto (1983).

[3] O.C. Zienkiewicz, An Introduction to The Finite Element Method, Mcgraw Hill, 1975, New York

[4] J. Liss, W. Goldsmith and J. M. Kelly, A Phenomenological Penetration Model of Plates, *Int. J. Impact Engng.* $\underline{1}$, 4, 321-341 (1983).
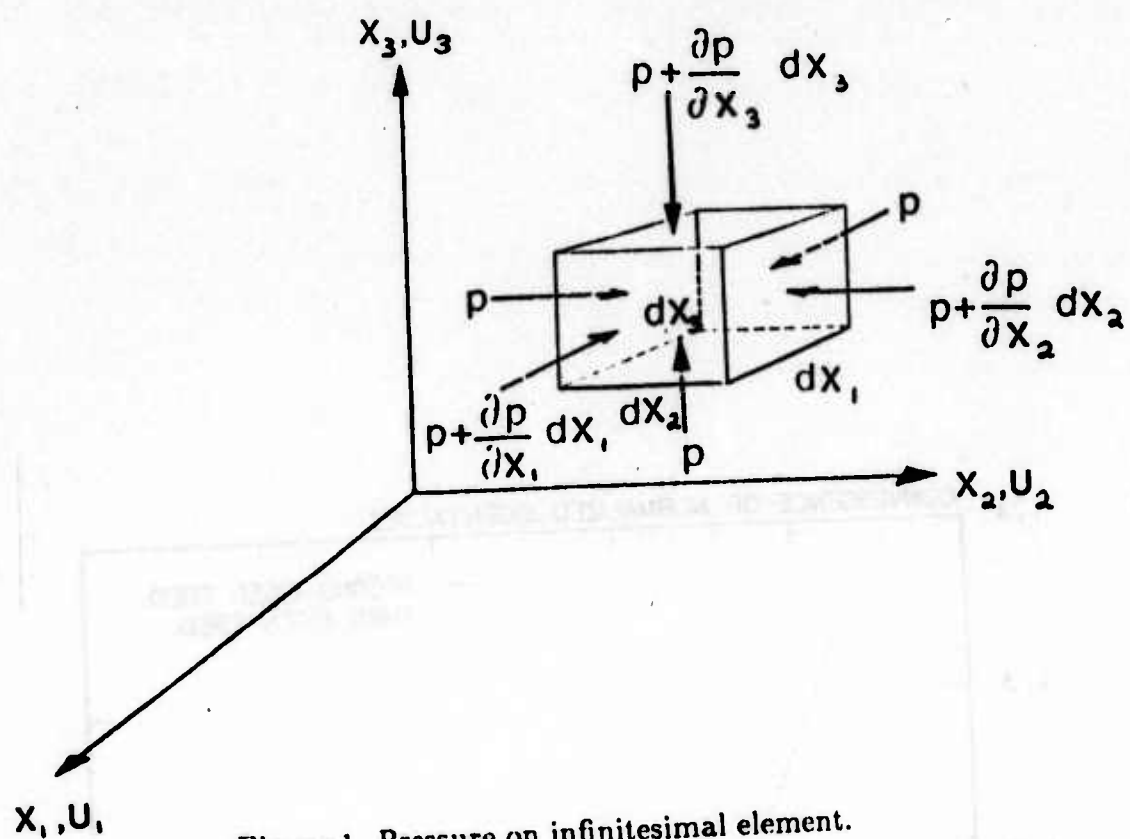
Figure 1. Pressure on infinitesimal element.
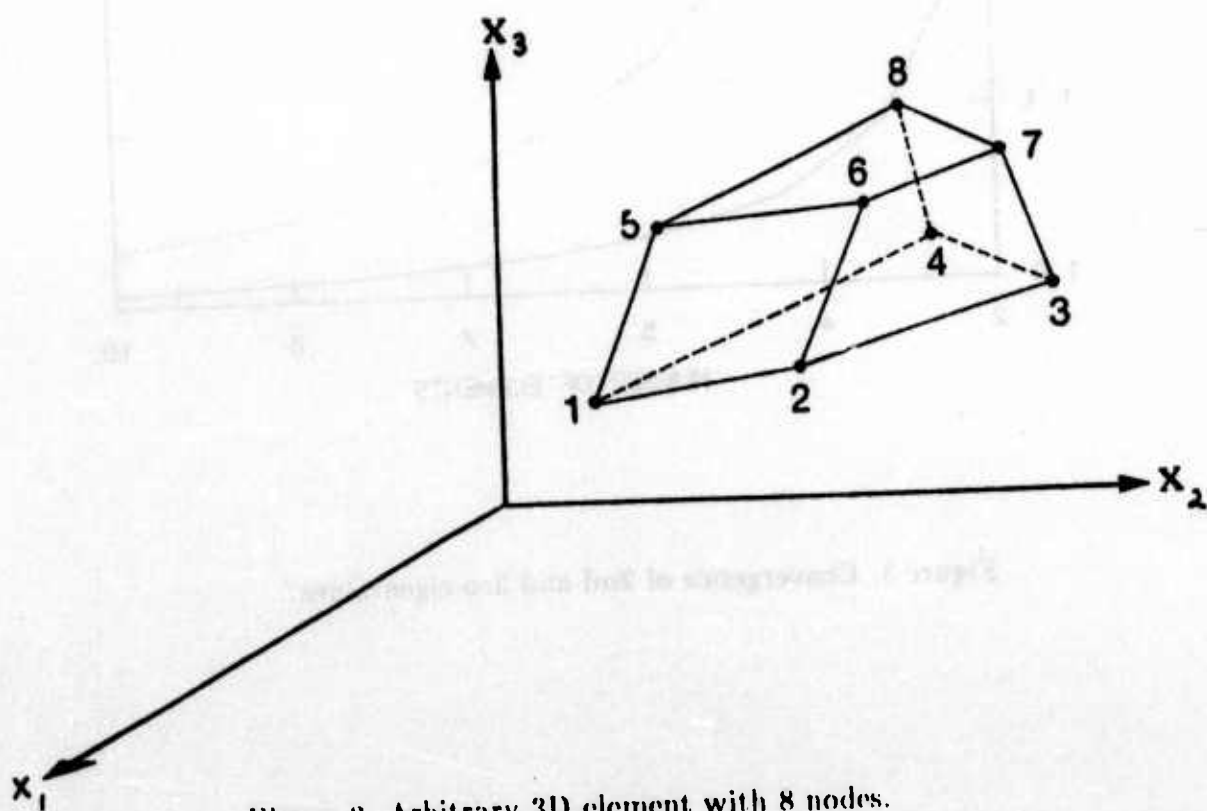
Figure 2. Arbitrary 3D element with 8 nodes.

91
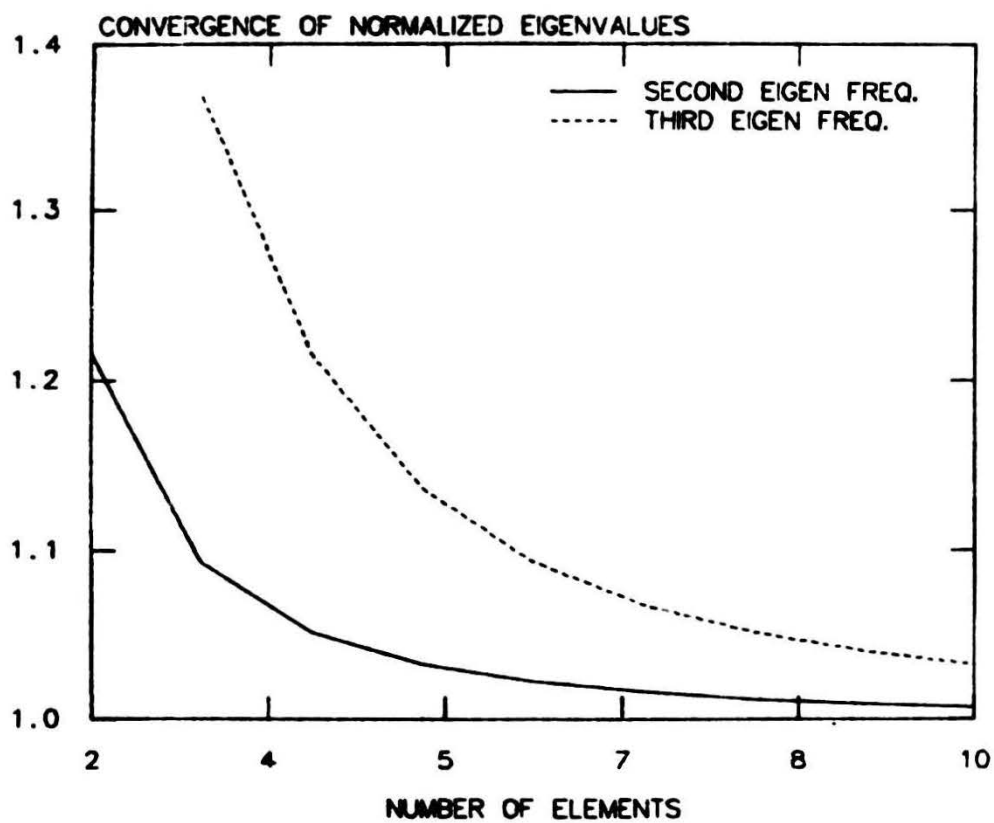
Figure 3. Convergence of 2nd and 3rd eigenvalues.

92
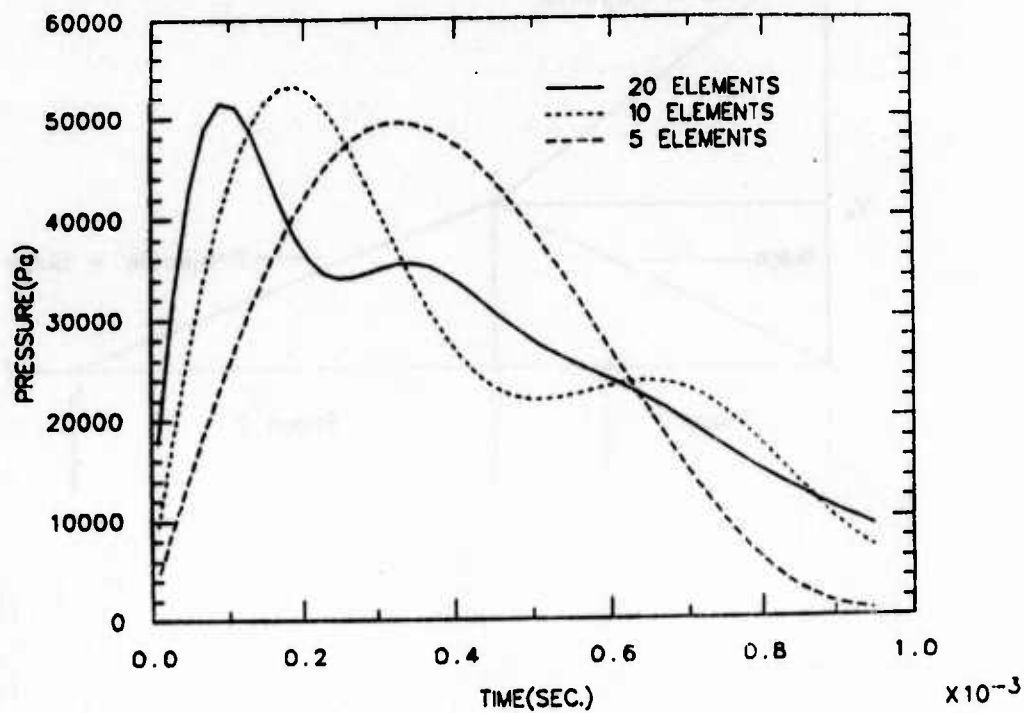
Figure 4. Effect of mesh size on response to unit step.
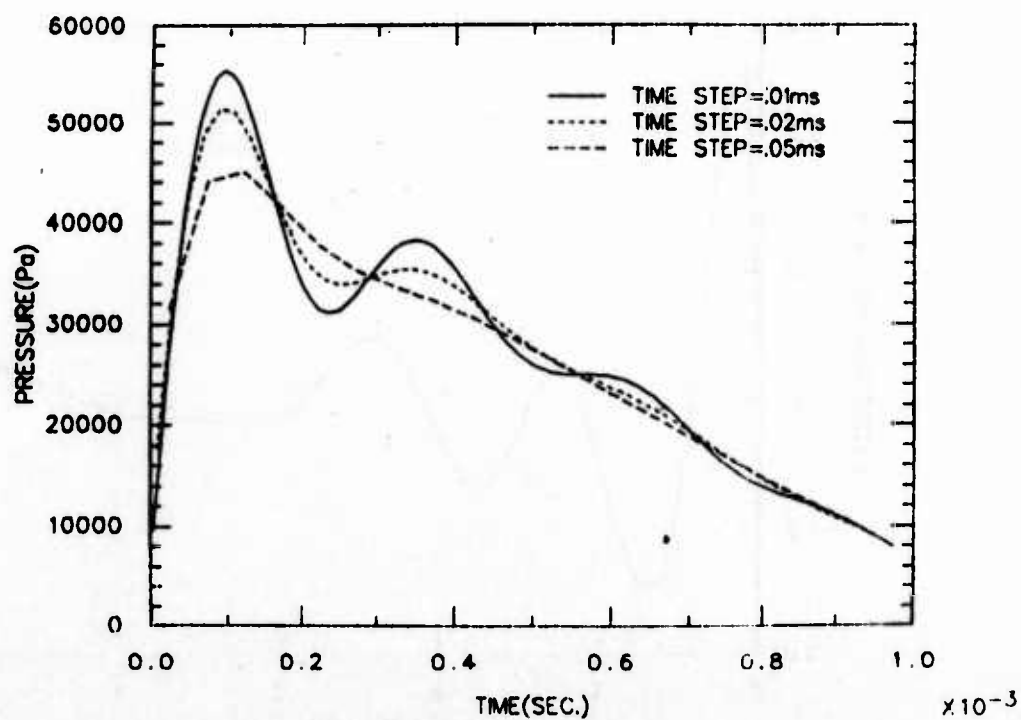
Figure 5. Effect of time step on response to unit step.

93
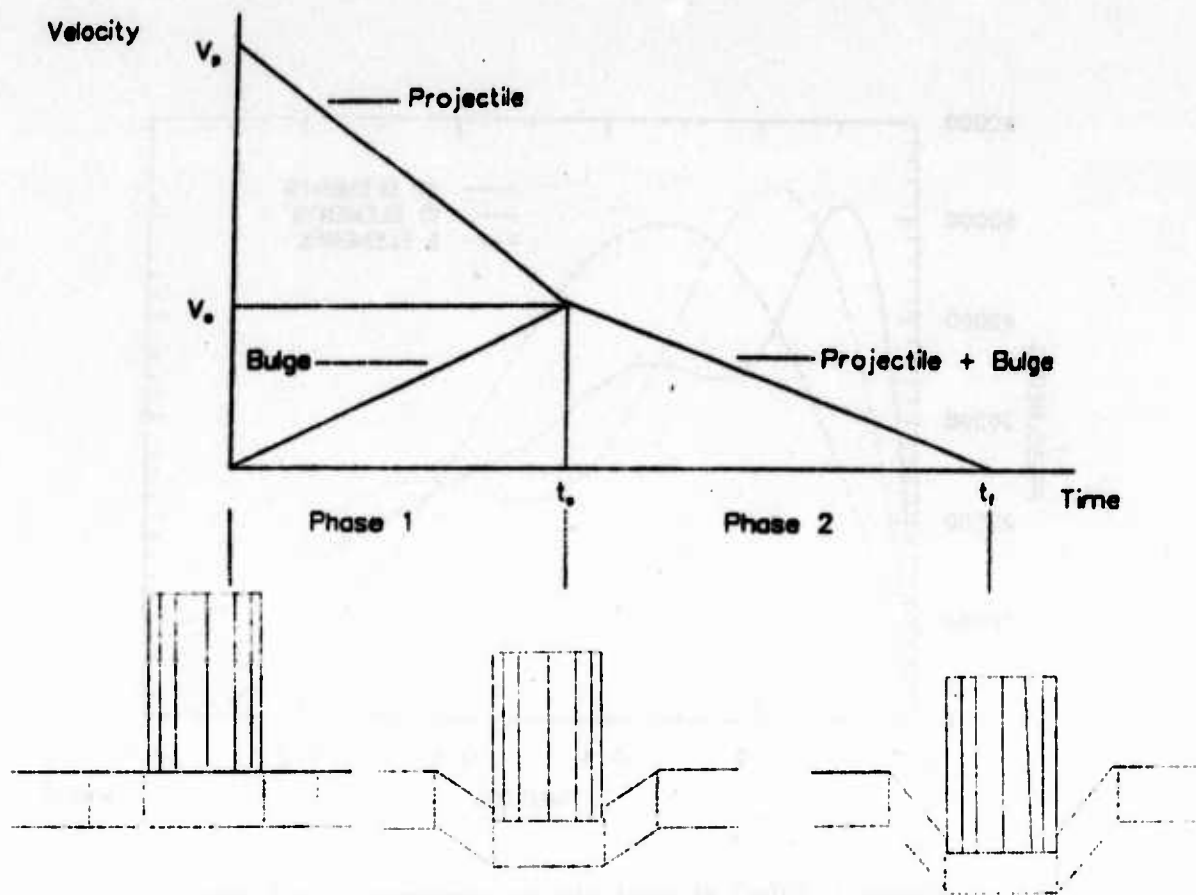
Figure 6. Idealization of impact
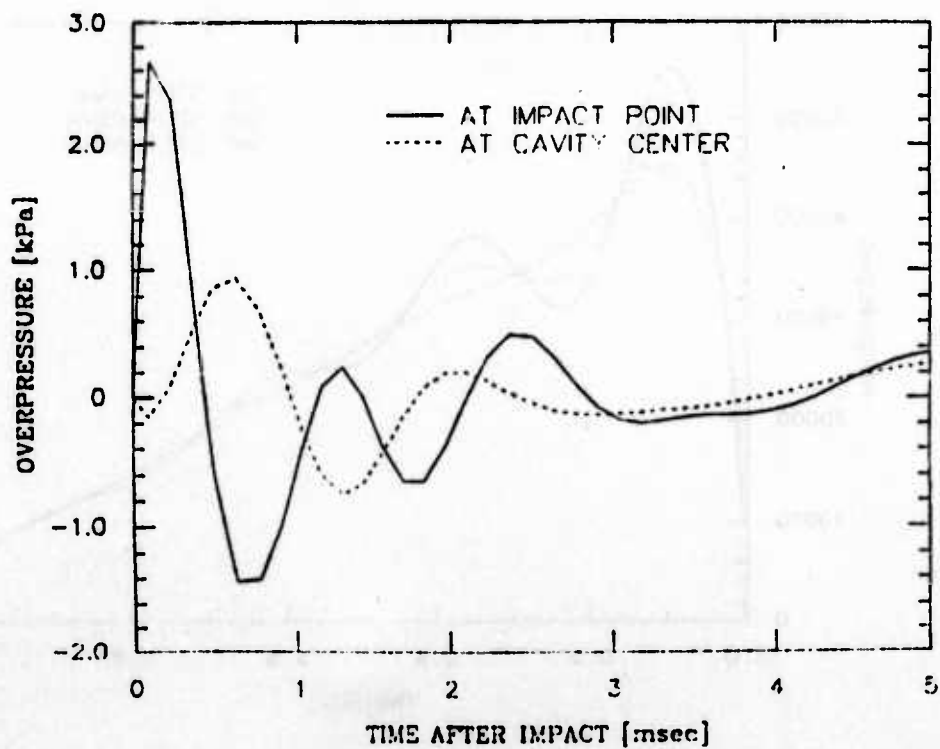
Figure 7. Response of hemispherical cavity to an impact.

94

Computation of Weight Functions in Two Dimensional Anisotropic Bodies †

T.-L. Sham
Department of Mechanical Engineering, Aeronautical Engineering & Mechanics
Rensselaer Polytechnic Institute, Troy, New York 12180-3590

Abstract

A finite element method introduced in [1] for computing Bueckner-Rice weight functions in finite bodies is described and the singular fields for a semi-infinite crack in an elastic and anisotropic body is given. These singular fields provided the required information for the computation of weight functions in anisotropic bodies under plane deformations.

1. Introduction

A finite element procedure, introduced in [1], has provided a unified approach in computing the Bueckner-Rice weight functions for all three fracture modes under either displacement, traction or mixed boundary conditions. This finite element procedure [1] is simple to implement and the results [1] obtained for two dimensional isotropic cracked solids are very accurate. In this study, this recently developed two and three dimensional finite element method is applied, as particular cases, to determining the weight functions in anisotropic bodies under plane deformations.

The synopsis of this paper is as follows. We first summarize in section 2 the finite element procedure introduced in [1] for determining the weight functions. This finite element method is valid for both two and three dimensional problems; however, in this paper, we shall concentrate on its two dimensional aspects. In section 3, we present the weight functions for a semi-infinite crack in an anisotropic full space which are required in the finite element procedure [1] for computing weight functions in anisotropic bodies.

2. Finite Element Method for Determining Weight Functions in Finite Bodies

Consider a two dimensional cracked body containing a single or a system of cracks. Let P be the specific crack tip at which we wish to determine the stress intensity factors. A crack tip cartesian coordinate system centered at P is employed with e_i being a set of unit base vectors. Roman subscripts have range 1 to 3 and summation convention is employed unless otherwise stated. We shall also use an in-plane polar coordinate system (r, θ) centered at the crack tip. Generalized plane deformation is assumed so that the stresses and strains are functions of the in-plane coordinates only. The stress intensity factors for an anisotropic solid can be defined by the traction vector acting on the plane directly ahead of the crack front as

$$K_i = \lim_{r \to 0} \sqrt{2\pi r}\ \sigma_{i2}(x_1, 0) \tag{1}$$

where K_1, K_2, K_3 are the mode II (in-plane shear mode), mode I (in-plane opening mode) and mode III (out-of-plane shear mode) stress intensity factors, respectively. The weight functions corresponding to the crack tip P will be denoted as $h_i(x; P)$ and they are vector-valued functions of position x. Under mixed boundary conditions and body force loading, the stress intensity factors K_i of P can be computed by [1,2]

$$K_i = \int_{S_T} \mathbf{T} \cdot \mathbf{h}_i\, dA + \int_{S_u} \mathbf{t}_i \cdot \mathbf{U}\, dA + \int_{\cdot} \mathbf{F} \cdot \mathbf{h}_i\, dV \tag{2}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the prescribed surface tractions and boundary displacements on the boundary S_T and S_u respectively; $\mathbf{F}$ is the prescribed body force field in the body with volume V; and t_i are the tractions generated by the weight functions h_i, each interpreted as displacement field, on the boundary S_u.

The weight functions h_i (Bueckner [3,4] and Rice [5]) are universal functions for given crack configuration, body geometry and material properties and are independent of loading systems. The stress field of h_i is in equilibrium with zero body force and it generates zero traction on all the crack faces and on the external boundary S_T. On the external boundary S_u where displacements are prescribed, the weight functions h_i are

---

zero. The weight functions yield elastic fields which give rise to unbounded energy for any finite region encompassing the crack tip P.

Let S_{int} be a suitably small bounding surface in the body which isolates the crack tip P from the rest of the body. The part of the body inside the surface S_{int} is referred to as region B and the remaining part of the body is denoted as region A. The unit outward normal to the bounding surface of region B is n. In region B, the weight functions h_i are decomposed into modified singular displacements $\hat{u}_i^s$ and modified regular displacements $\hat{u}_i^r$, viz

$$h_i = \hat{u}_i^s + \hat{u}_i^r \qquad (3)$$

The modified singular displacements $\hat{u}_i^s$ are constructed so that

(i)   they admit the same singularity as the weight functions h_i at the crack tip P; and

(ii)   they generate zero traction on the crack faces inside region B.

The modified regular displacements $\hat{u}_i^r$ are taken to be bounded at the crack tip P and they can be identified as the displacements in elastic crack analyses. The singular stress fields, σ_i^s, of $\hat{u}_i^s$ and the regular stress fields, σ_i^r, of $\hat{u}_i^r$ are self-equilibrating and they generate zero tractions on the crack faces inside region B.

It is noted that the modified singular displacements $\hat{u}_i^s$ defined as such do not lend themselves to a unique construction in general. Indeed, it is the non-uniqueness in their construction which allows us to make judicious choices of $\hat{u}_i^s$ - we can choose $\hat{u}_i^s$ to correspond to the simplest possible crack geometry.

2.1. Variational Principle for h_i and $\hat{u}_i^r$

The weight functions h_i in region A and the modified regular displacements $\hat{u}_i^r$ in region B of the cracked body under consideration can be determined by the finite element method introduced in [1]. This finite element method is based on the following minimum principle [1].

Define a functional H as

$$H[h_i, \hat{u}_i^r] = \int_A w(\varepsilon_i)\, dV + \int_B w(\varepsilon_i^r)\, dV - \int_{S_{int}} (-\sigma_i^s n) \cdot \hat{u}_i^r\, dA \qquad (no\ sum\ on\ i) \qquad (4)$$

where ε_i and ε_i^r are the strain fields corresponding to h_i in A and $\hat{u}_i^r$ in B respectively and w is the elastic strain energy density. The functional H is bounded and it is a functional of h_i in region A and $\hat{u}_i^r$ in region A. It has been proven in [1] that among all possible fields h_i in region A and $\hat{u}_i^r$ in region B which

(i)   satisfy the strain-displacement relations; and

(ii)   make h_i zero on the external boundary S_u and equal to the sum of $\hat{u}_i^s$ and $\hat{u}_i^r$ on the internal boundary S_{int}, where $\hat{u}_i^s$ are considered to be given;

the true fields $(h_i)^*$ in region A and $(\hat{u}_i^r)^*$ in region B minimize the functional H.

An implementation of this variational principle within the context of a displacement-based finite element method is given in [1] and this implementation can be incorporated into standard linear elastic finite element program with little programming efforts. This procedure is very similar to standard finite element methods and it involves prescribing

(i)   nodal forces corresponding to the tractions $-\sigma_i^s n$ on the internal boundary S_{int};

(ii)   nodal "effective" body forces [1] for the elements in region A which are adjacent to the internal boundnary S_{int}; and

(iii)   zero tractions and displacements on the external boundaries S_T and S_u respectively.

The nodes inside region B, including those on S_{int}, are interpreted as nodal unknowns for the modified regular displacements $\hat{u}_i^r$ and the remaining nodes represent the nodal values of the weight functions h_i.

Thus, in employing this procedure to determining weight functions in two dimensional anisotropic bodies, we first have to choose some modified singular displacements $\hat{u}_i^s$ which satisfy the conditions delineated above. The simplest candidates for this purpose are the weight functions of a semi-infinite crack in an anisotropic full space. They are given in the following section by ways of Rice's [5-7] crack front variation approach.

3. Weight Functions for a Semi-Infinite Crack in an Anisotropic Full Space

First, consider a crack in an elastic anisotropic solid of finite extent. The energy release rate G for the anisotropic solid may be expressed in terms of the stress intensity factors as [6-8]

$$G = K_i \Lambda_{ij} K_j \tag{5}$$

where Λ is symmetric and positive definite; it depends only on the elastic constants; and it can be related to the pre-logarithm energy factor of a straight dislocation line in an uncracked solid, lying parallel to the crack front in the cracked body. The energy release rate G can also be decomposed additively into three components, G_i, according to Irwin's concept of virtual crack extension [9] as

$$G_i = \lim_{\delta a \to 0} \frac{1}{2} \int_0^{\delta a} \sigma_{i2}(x_1,0) \ [\ u_i(x_1-\delta a,0_+) - u_i(x_1-\delta a,0_-) \ ] \ dx_1 \qquad (no \ sum \ on \ i)$$

where $\sigma_{i2}(x_1,0)$ is the traction acting on the e_2-plane ahead of the crack tip and $u_i(x_1-\delta a,0_+)$ and $u_i(x_1-\delta a,0_-)$ are the displacements on the upper and lower crack face respectively. It is customary to refer to G_1, G_2 and G_3 respectively as mode II, mode I and mode III energy release rate. Thus, for an anisotropic body, G_i can be related to the stress intensity factors via

$$G_i = \sum_{j=1}^{3} K_i \Lambda_{ij} K_j \qquad (no \ sum \ on \ i)$$

For an isotropic body, Λ is diagonal and eqn (5) reduces to the familiar Irwin relation

$$G = \frac{1-\nu^2}{E} (K_1{}^2 + K_2{}^2) + \frac{1+\nu}{E} K_3{}^2 \tag{6}$$

under plane strain conditions. Here, ν is the Poisson's ratio and E is the Young's modulus.

Adopting Rice's [5] crack front variation concept, let the anisotropic cracked solid be subjected under two linearly independent loading systems denoted by I and II respectively. Q^I and Q^{II} are the generalized loads and q^I and q^{II} are the corresponding work conjugate generalized displacements for the two loading systems. Suppose both loading systems are applied to the cracked body simultaneously. The change in strain energy due to virtual displacements δq^I and δq^{II}, and virtual crack front variation δa at fixed external forces is

$$\delta U = Q^I \delta q^I + Q^{II} \delta q^{II} - K_i \Lambda_{ij} K_j \delta a \tag{7}$$

where U is the strain energy and a is the crack length. From linearity, we may write

$$K_i = k_i^I(a) Q^I + k_i^{II}(a) Q^{II}$$
$$q^I = C_{I,I}(a) Q^I + C_{I,II}(a) Q^{II} \tag{8}$$
$$q^{II} = C_{II,I}(a) Q^I + C_{II,II}(a) Q^{II}$$

where $k_i^I(a)$ and $k_i^{II}(a)$ are the respective geometric dependent part of the stress intensity factors induced at the crack tip when loading systems I and II are applied individually to the cracked body. The C's are the compliances. A Legendre transformation of eqn (7) gives

$$\delta(U - Q^I q^I - Q^{II} q^{II}) = - q^I \delta Q^I - q^{II} \delta Q^{II} - G \delta a \tag{9}$$

The left hand side of eqn (9) is a perfect differential and this enables us to obtain the following reciprocal relation,

$$\left[\frac{\partial q^{II}(Q^I,Q^{II},a)}{\partial a} \right]_{Q^I,Q^{II}} = \left[\frac{\partial(K_i \Lambda_{ij} K_j)}{\partial Q^{II}} \right]_{Q^I,a} \tag{10}$$

Eqn (8) can be used to compute the derivative on the right hand side to obtain

$$\left[\frac{\partial q^{II}}{\partial a} \right]_{Q^I,Q^{II}} = 2 \Lambda_{ij} (k_i^I k_j^{II} Q^I + k_i^{II} k_j^{II} Q^{II}) \tag{11}$$

The interpretation of eqn (11) is as follows [5]. Suppose that we know the complete solution, in particular, $k_i^I(a)$ and $C_{II,I}$, when loading system I is applied. Setting $Q^{II} = 0$ in eqn (11), we obtain

97

$$\left[\frac{\partial q^{II}}{\partial a}\right]_{Q^I,\,Q^{II}=0} = 2\,K_i^I\,\Lambda_{ij}\,k_j^{II} \tag{12}$$

It is noted that by knowing the solution for loading system I, we can compute every term in eqn (12) except k_j^{II}. Thus, when eqn (12) is multiplied through by Q^{II}, we arrive at a useful relation

$$Q^{II}\left[\frac{\partial q^{II}}{\partial a}\right]_{Q^I,\,Q^{II}=0} = 2\,K_i^I\,\Lambda_{ij}\,K_j^{II} \tag{13}$$

for the stress intensities induced by non-zero loading system II alone.

In order to demonstrate how this relation is used, let the displacement fields u_1, u_2 and u_3 be three solutions to the elastic boundary value problem corresponding to the respective loading system Q_1^I, Q_2^I and Q_3^I which are linearly independent of each other. For our discussion, it suffices to assume that each loading, Q_i^I, induces *nonzero* stress intensity K_i^I alone at the crack tip. Employing these three solutions to relation (13), we obtain

$$Q^{II}\left[\frac{\partial q_i^{II}}{\partial a}\right]_{Q^I,\,Q^{II}=0} = S_{ij}\,\Lambda_{jm}\,K_m^{II} \tag{14}$$

where

$$S_{ij} = 2\,K_i^I\,\delta_{ij} \qquad (\textit{no sum on } i)$$

and δ_{ij} is the Kronecker delta. Since S_{ij} is diagonal and positive definite, it admits the inverse S_{ij}^{-1}, viz

$$S_{ij}^{-1} = \frac{1}{2\,K_i^I}\,\delta_{ij} \qquad (\textit{no sum on } i)$$

Thus eqn (14) can be inverted to obtain the important relation

$$K_i^{II} = Q^{II}\left[\Lambda_{ij}^{-1}\,S_{jm}^{-1}\left[\frac{\partial q_m^{II}}{\partial a}\right]_{Q^I,\,Q^{II}=0}\right] \tag{15}$$

Rice [5] has observed that the bracketed terms on the right do not depend on the nature of loading system I and hence they are universal functions for the given crack configuration, body geometry and elastic properties. In fact, these functions can be identified as Bueckner's [3] weight functions, h_i, namely

$$h_i = \Lambda_{ij}^{-1}\,S_{jm}^{-1}\,\frac{\partial u_m}{\partial a} \tag{16}$$

and the stress intensity factors can be computed by eqn (2) given above for a general loading system which consists of surface forces, boundary displacements and body forces. For the case of symmetrical mode I loading in an isotropic solid, eqn (16) reduces to the same expression as given by Rice [5],

$$h_2 = \frac{M}{2\,K_2(a)}\,\frac{\partial u_2(x,a)}{\partial a}$$

with M being related to the appropriate elastic constants. The singular stress fields, σ_i^I, of the weight functions h_i can also be composed from the stresses, σ_m, of u_m as

$$\sigma_i^I = \Lambda_{ij}^{-1}\,S_{jm}^{-1}\,\frac{\partial \sigma_m}{\partial a} \tag{17}$$

In order to employ eqns (16) and (17) to determine the weight functions and their corresponding stress fields for a semi-infinite crack in an anisotropic full space, we shall follow the plane anisotropic elasticity formulation, originally developed by Stroh [10,11], and subsequently used by Barnett and Asaro [8]; and Ting and coworkers [12-14], among others, to obtain three linearly independent solutions for the semi-infinite crack.

With respect to the coordinates introduced, the field equations for the anisotropic solid are

$$\varepsilon_{ij} = \frac{1}{2}\left[\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right]$$

$$\sigma_{ij} = C_{ijkm} \frac{\partial u_k}{\partial x_m} \tag{18}$$

$$\frac{\partial \sigma_{ij}}{\partial x_j} = 0$$

where $u = u(x_1, x_2)$; $\varepsilon = \varepsilon(x_1, x_2)$; and $\sigma = \sigma(x_1, x_2)$ are the displacements, strains and stresses respectively and plane strain conditions are assumed. A general solution to the equations in (18) is, [10,11]

$$u = A f(z) \tag{19}$$

with $z = x_1 + p x_2$, and A and p are complex. Substituting the general solution into eqns (18), the stresses can be expressed as

$$\sigma = \tau \frac{df}{dz}$$

$$\tau_{ij} = \left[C_{ijk1} + p \; C_{ijk2} \right] A_k$$

Equilibrium can be satisfied if

$$\left[C_{i1k1} + p \; (C_{i1k2} + C_{i2k1}) + p^2 \; C_{i2k2} \right] A_k = 0 \tag{20}$$

For non-trivial A, the determinant of the bracketed terms would have to be zero and this leads to a sextic characteristic equation for the roots p. Eshelby et al. [15] have shown that there are no real roots to this characteristic equation and thus the roots occur in complex conjugate pairs. Following Stroh [10,11], we shall introduce a complex vector L,

$$L_i = \tau_{i2} = (C_{i2k1} + p \; C_{i2k2}) A_k \tag{21a}$$

or

$$L_i = - (C_{i1k2} + p^{-1} \; C_{i1k1}) A_k \tag{21b}$$

Equation (20) can then be recasted as

$$- C_{i2j2}^{-1} \; C_{j2k1} A_k + C_{i2k2}^{-1} \; L_k = p \; A_i \tag{22a}$$

$$\left[C_{i1j2} \; C_{j2m2}^{-1} \; C_{m2k1} - C_{i1k1} \right] A_k - C_{i1j2} \; C_{j2k2}^{-1} \; L_k = p \; L_i \tag{22b}$$

These equations are in the form of standard eigenvalue problems with p being the eigenvalue and the vector { $A \mid L$ } being the eigenvector. Standard procedures (e.g. *EISPACK* [16]) can be employed to extract the eigenvalues and the eigenvectors efficiently. Since the eigenvalues are all complex, we shall order them such that p_α would have positive imaginary part and $\bar{p}_\alpha$ are their complex conjugates. Greek subscripts have range 1 to 3 but they do not conform to the summation convention. The corresponding eigenvectors will be denoted by A_α and L_α with complex conjugates $\bar{A}_\alpha$ and $\bar{L}_\alpha$. Assuming that all the eigenvalues p_α are distinct, the general solution u and σ can be expressed as a linear combination of all the eigenvectors as

$$u = 2 \; \text{Re} \sum_{\alpha=1}^{3} A_\alpha f_\alpha(z_\alpha)$$

$$\sigma = 2 \; \text{Re} \sum_{\alpha=1}^{3} \tau_\alpha \frac{df_\alpha}{dz_\alpha}$$

The cases of repeated roots for p_α can also be treated by using methods given by Ting and Chou [12]; and Ting [13] to construct the appropriate eigenvectors.

Introducing three vectors M_α (Stroh [10,11]) which are the reciprocal of L_α such that

$$M_\alpha \cdot L_\beta = \delta_{\alpha\beta}$$

The three linearly independent solutions for a semi-infinite crack, with each solution u_i corresponding to one stress intensity factor K_i being induced at the crack tip alone, can be written as [14]

$$u_i = \frac{\sqrt{2 \; r}}{\sqrt{\pi}} \; K_i \; \text{Re} \sum_{\alpha=1}^{3} (M_\alpha \cdot e_i) \; A_\alpha \; \xi_\alpha^{1/2} \qquad (\textit{no sum on } i)$$

99

and the respective stresses are

$$\sigma_i = \frac{K_i}{\sqrt{2\,\pi\,r}}\,\mathrm{Re}\sum_{\alpha=1}^{3}\,(\,\mathbf{M}_\alpha \cdot \mathbf{e}_i\,)\,\tau_\alpha\,\xi_\alpha^{-1/2} \qquad (\textit{no sum on } i\,)$$

where

$$\xi_\alpha = \cos\theta + p_\alpha \sin\theta$$

Eventhough the eigenvectors $\mathbf{A}_\alpha$, $\mathbf{L}_\alpha$ are determined to within arbitrary constants from eqns (22a,b), the products of the components of $\mathbf{A}_\alpha$, and $\mathbf{M}_\alpha$ are uniquely determined. The displacement and stress derivatives with respect to crack length can be derived by employing the following relations

$$\frac{\partial}{\partial a} = \frac{\partial r}{\partial a}\,\frac{\partial}{\partial r} + \frac{\partial\theta}{\partial a}\,\frac{\partial}{\partial\theta}$$

$$\frac{\partial r}{\partial a} = -\cos\theta$$

$$\frac{\partial\theta}{\partial a} = \frac{\sin\theta}{r}$$

to obtain

$$\frac{\partial u_i}{\partial a} = -\frac{K_i}{\sqrt{2\,\pi\,r}}\,\mathrm{Re}\sum_{\alpha=1}^{3}\,(\,\mathbf{M}_\alpha \cdot \mathbf{e}_i\,)\,\mathbf{A}_\alpha\,\xi_\alpha^{-1/2}\ (\textit{no sum on } i\,) \tag{23}$$

$$\frac{\partial\sigma_i}{\partial a} = \frac{K_i}{2\sqrt{2\,\pi}\,r^{3/2}}\,\mathrm{Re}\sum_{\alpha=1}^{3}\,(\,\mathbf{M}_\alpha \cdot \mathbf{e}_i\,)\,\tau_\alpha\,\xi_\alpha^{-3/2}\ (\textit{no sum on } i\,) \tag{24}$$

The matrix $\mathbf{A}$ which is related to the pre-logarithm energy factor of a straight dislocation lying parallel to crack front front can be expressed in terms of $\mathbf{A}_\alpha$ and $\mathbf{M}_\alpha$ as (Stroh [10,11])

$$\mathbf{A}^{-1} = 2\,\pi\,\mathbf{B}^{-1} \tag{25}$$

with

$$\mathbf{B} = \frac{\sqrt{-1}}{2}\sum_{\alpha=1}^{3}\,[\,\mathbf{A}_\alpha\,\mathbf{M}_\alpha - \overline{\mathbf{A}}_\alpha\,\overline{\mathbf{M}}_\alpha\,]$$

where the terms in the bracket are dyadic products of the respective vectors.

Thus, once the Stroh's eigenvalues p_α and the Stroh's eigenvectors $\mathbf{A}_\alpha$ and $\mathbf{L}_\alpha$ are computed for a given anisotropy of the body under consideration, the weight functions $\mathbf{h}_i$ and their associated singular stresses σ_i' for a semi-infinite crack can be obtained by using eqns (23-25) in eqns (16) and (17).

In the finite element implementation of the variational principle discussed in section 2, we would choose the modified singular displacements $\tilde{u}_i'$ and stresses σ_i' to be those corresponding to the semi-infinite crack geometry obtained above for all weight function computations in 2-D.

References

[1]   T.-L. Sham, to appear in *International Journal of Solids* and *Structure* (1986).

[2]   H.F. Bueckner, to appear in *International Journal of Solids* and *Structure*, (1986).

[3]   H.F. Bueckner, *Zeitschrift angew. Math. Mech.*, 50(1970) 529-533.

[4]   H.F. Bueckner, *Field Singularities and Related Integral Representations*, Mechanics of Fracture I: Methods of Analysis and Solution of Crack Problems, Noordhoff, Leyden (1973) 239-314.

[5]   J.R. Rice, *International Journal of Solids* and *Structure*, 8(1972) 751-758.

[6]   J.R. Rice, *Journal of Applied Mechanics*, 52(1985) 571-579.

[7]   J.R. Rice, *International Journal of Solids* and *Structure*, 21(1985) 781-791.

[8]   D.M. Barnett and R.J. Asaro, *Journal of the Mechanics & Physics of Solids*, 20(1972) 353-366.

[9]   G.R. Irwin, *Journal of Applied Mechanics*, 24(1957) 361.

[10] A.N. Stroh, *Philosophical Magazine*, 3(1958) 625-646.

[11] A.N. Stroh, *Journal of Mathematical Physics*, 41(1962) 77-103.

[12] T.C.T. Ting and S.C. Chou, *International Journal of Solids* and *Structure*, 17(1981) 1057-1068.

[13] T.C.T. Ting, *International Journal of Solids* and *Structure*, 18(1982) 139-152.

[14] T.C.T. Ting and P.H. Hoang, *International Journal of Solids* and *Structure*, 20(1984) 439-454.

[15] J.D. Eshelby, W.T. Read and W. Shockley, *Acta Metallurgica*, 1(1953) 1.

[16] *Matrix Eigensystem Routines - EISPACK Guide*, 2nd edition, B.T. Smith cl at., Springer-Verlag (1976).

MICROMECHANICS OF SHEAR BANDING IN HIGH STRENGTH STEEL

Dennis M. Tracey, Colin E. Freese, and Paul J. Perrone

Mechanics and Structures Division
U. S. Army Materials Technology Laboratory
Watertown, Massachusetts 02172-0001

ABSTRACT

The work is directed to the void softening mechanism of shear banding in ductile high strength steels. Elastic-plastic analyses of the field near a pair of interacting voids were conducted using a finite element formulation and large scale computational facilities. Results suggest dramatic intensification of strain between interacting voids. The nature of void interaction was found to be significantly different in the cases of nominal shear and uniaxial extension, consistent with experimental observations of void linking.

INTRODUCTION. Shear banding is a serious mode of degradation of high strength steel loaded into the plastic range and important design issues require an understanding of its causes. Controlled shear experiments have demonstrated that localization into narrow shear bands occurs at a material characteristic strain level. Banding occurs under both high-rate and quasi-static loadings. Once strain localizes, continued deformation to fracture occurs under decreasing applied stress. Hence, there is an underlying strain-softening mechanism associated with the banding event. In high rate loadings, thermal softening results from the heat of plastic deformation and near-adiabatic conditions. Thermal effects are insignificant in slow loading and thus other softening mechanism(s) must be involved.

Metallographic investigations[1] of sectioned shear specimens of high strength 4340 steel have found evidence that microscopic

103

voids play an important role in shear banding. Voids were observed at debonded grain refinement particles (0.1 micron size scale) and microcracks were found linking neighboring voids. It is speculated that the microcracks develop by coalescence of smaller scale voids which nucleate at strengthening particles as a result of the local strain intensification of the dominant pair.

In an attempt to elucidate the role of voids in the localization event, two-dimensional plane strain elastic-plastic analyses were conducted to establish the field solution between a pair of interacting voids under three different nominally uniform strain fields. The loadings considered were simple shear, uniaxial tension, and combined shear with extension. While the specified kinematic loadings represent nominally uniform strain fields, in the vicinity of the voids the stress and strain are found to be extremely complex with substantial interaction features.

NUMERICAL FORMULATION. The nonlinear elastic-plastic analysis was performed using an incremental finite element formulation.[2] Classical non-hardening Prandtl-Reuss constitutive theory was employed. The mesh used in the analysis is displayed in Figure 1. It consists mostly of quadrilaterals each subdivided into four linear strain triangles by its diagonals. As drawn, there are approximately 3000 degrees of freedom in the mesh. Symmetry allows a single quadrant to be analyzed in extensional loading, and in simple shear, one-half of the mesh is sufficient.

Loading was specified by displacement conditions on the outer boundary. In simple shear, the boundaries were constrained against motion in the y-direction, while displacement in the x-direction was specified as a linear function of y. In uniaxial extension, the boundaries were constrained against displacement in the x-direction, while y-displacement was specified as a linear function of y. The combined mode loading considered was a superposition of the simple shear and uniaxial

104

extension kinematic boundary conditions. Nominally, the simple shear problem involves no normal strain, while the uniaxial extension problem nominally involves no shear nor contractional components of normal strain.

The assumed displacement method of finite element analysis was used so that the problem at each load step was to find the vector of displacement increments Δu_i. If we consider components $i=1$ through $m-1$ as the specified non-zero boundary values and denote them as Δu_i^*, the stiffness equations for the degrees of freedom m through n take the form:

$$
\begin{bmatrix} K_{mm} & \cdots & K_{mn} \\ \vdots & \ddots & \vdots \\ K_{nm} & \cdots & K_{nn} \end{bmatrix} \begin{Bmatrix} \Delta u_m \\ \Delta u_n \end{Bmatrix} = \sum_{i=1}^{m-1} \Delta u_i^* \begin{Bmatrix} -K_{mi} \\ -K_{ni} \end{Bmatrix} \tag{1}
$$

The stiffness matrix $\underline{K}$ is assembled from element stiffness matrices $\underline{k}$ which have the form

$$
\underline{k} = \int_V \underline{B}^T \underline{D} \underline{B} \, dv \tag{2}
$$

where $\underline{B}$ follows from the displacement interpolation function and $\underline{D}$ is the incremental constitutive matrix relating stress and strain increments,

$$
\underline{\Delta \epsilon} = \underline{B} \, \underline{\Delta u} \tag{3}
$$
$$
\underline{\Delta \sigma} = \underline{D} \, \underline{\Delta \epsilon} \tag{4}
$$

It is the constitutive matrix $\underline{D}$ which is the source of the nonlinearity of the formulation. The rate form of the constitutive law has $\underline{D}$ dependent upon current stress state. If the current state is within the yield surface, then $\underline{D}$ represents linear elastic behavior. If the current stress state is on the yield surface, then $\underline{D}$ represents the ability of the material to plastically flow in the direction normal to the

105

yield surface. As load is increased, the stress distribution changes and the zone of plasticity expands. Thus, the appropriate $\underline{D}$ at a point continually changes.

In our numerical analysis, finite values are necessarily specified for the increments of the boundary displacements Δu_i^*. Thus, the appropriate $\underline{D}$ at each grid point will change within the load increment. An implicit approach is employed whereby a step average $\underline{D}$ is used that accounts for yielding, stress changes along the yield surface, and possibly unloading within the step. The solution for the increments Δu_i is gained using a successive approximation iterative procedure. At each iterate Δu_i^j, an average $\underline{D}$ is formed, if necessary, at each grid point, as follows:

$$\underline{D}^{av} = f_e \, \underline{D}^{el} + (1-f_e) \, (\underline{D}^{el} - 2G \, \underline{n} \, \underline{n}^T) \tag{5}$$

The weighting factor f_e and average yield surface unit normal $\underline{n}$ are defined in terms of Δu_i^j. $\underline{D}^{el}$ is the elastic constitutive matrix and G is the elastic shear modulus. The vector $\underline{n}$ is defined so that the stress state at the end of the increment satisfies the yield condition. The issue of stress scaling and associated load imbalance common to tangent approaches is not encountered in the implicit formulation.

Regardless of formulation, finite load steps involve a load path discretization error. This error was controlled by employing an adaptive load incrementation procedure. The size of the load increment (scale factor adjusting the magnitude of specified components Δu_i^*) is altered during the iteration to satisfy a condition on the stress solution. The condition dictates that the maximum deviatoric stress change (along the yield surface) accompanying plastic flow should equal a specified fraction of the uniaxial yield stress Y. In the results presented, this stress change fraction was specified equal to 0.05. Using a convergence test that has successive iterations with a relative difference in stress increments less

106

than 0.001, typically 4 iterations are found necessary at each load step. Typical solutions which trace plasticity from first yield to general yield conditions involve approximately 50 load steps. Hence, it is common to perform in the order of 200 solutions to the stiffness equation (1) in a particular load case analysis.

NUMERICAL SOLUTIONS. We discuss here details of the solution for the void pair shown in Figure 1 under three different modes of imposed strain. In each of the three analyses conducted, the solution history was incrementally traced from the point of first yield to the onset of general yielding throughout the band containing the voids.

First, we consider the solution for the case of imposed nominal shear. General yield in this case occurs at a shear strain slightly below the yield strain value $(Y/\sqrt{3})/G$. The plastic zone development is illustrated in Figure 2. Elements with a stress state satisfying the Mises yield condition are drawn at nominal strain levels of 60, 77, and 91% of the yield strain. First yield occurs at the surfaces of the the voids at positions roughly 45 degrees from the ligament. The imposed strain at first yield was found to equal 49% of the yield strain, suggesting an elastic concentration factor approximately equal to 2.0. As the plastic zones at the void surface grow, a separate distinct zone develops along the ligament. A mechanism for extensive local straining is possible once the zones link. Figure 3 illustrates the fact that on the ligament, the strain is fairly uniform over most of the load history. However, once general yielding conditions are achieved, maxima develop at a distance roughly 0.3D ahead of the voids. Figure 4 is a plot of local strain (maximum on the ligament) divided by nominal strain as a function of nominal strain. The plot demonstrates the surge in local straining that occurs once the plastic mechanism is established. The strain intensification rate reaches a value in excess of 30 once general yielding is achieved.

107

In the case of uniaxial extension, the nominal stress state is triaxial tension. For Poisson's ratio of 0.3, which was the value used in these analyses, yielding in the band without voids occurs at the uniaxial strain value of 1.3 Y/E. The plastic zone at a strain level of 1.27 Y/E is drawn in Figure 5a. The fully developed plasticity region is restricted to the void region at this nominal strain level. Figure 5b shows those elements which satisfy the near-isocloric condition which has the magnitude of the normal strain increments differing by less than 6%. At this strain level, the stress data along the ligament agrees very well with the logarithmic spiral slipline stress distribution, Figure 6. Consistent with the plastic zone development, the strain maximum is at the void surface in this problem throughout the loading history.

Under a state of combined extension and shear, with ϵ_{nom} equal to γ_{nom}, yielding occurs when the two strain components reach the value 0.98 Y/E. The plastic zone and near-isocloric region are drawn in Figure 8 for a strain level slightly exceeding this nominal yield value. In this problem, plastic zones grow from the void surfaces slightly skewed from the ligament. The separate zones from the voids merge by the development of a narrow plastic region between them oriented in the loading direction. The normal and shear stress distributions are provided in Figures 9 and 10. As general yield conditions are reached, the maximum normal stress is seen to shift to the center of the ligament. While the maximum shear stress is at or near the ligament center throughout loading, its value can be seen to at first increase and then decrease as general yield conditions are approached. An interesting aspect of the final shear stress distribution is the flat, near-zero valued regions near the voids, indicative of prevailing logarithmic spiral regions there. The strain distributions for this case show the maximum normal strain at the void surfaces at the beginning of loading. The maximum shifts to the center of the ligament once general yielding is achieved. The shear

strain is maximum at the ligament center throughout loading in this problem.

SUMMARY. The numerical solutions suggest that the interaction between voids in elastic-plastic deformation fields depends strongly upon the nominal straining mode. It can be readily inferred that the nature of the interaction is also strongly dependent upon the orientation of the voids with respect to the principal strain directions.

For the simple shear loading (with the void centerline aligned with the shear direction) the majority of the ligament experiences a reasonably similar strain history with magnitudes significantly in access of applied strain. The minimum strain region is at the void surface in this loading case. On the contrary, under uniaxial extension normal to the void centerline, the void surface experiences the maximum strain levels throughout the loading history. In combined loading, the strain distribution was shown to change character, first from void surface straining to strain maximum at the ligament center, once the plastic zones of the voids merge.

REFERENCES

1. M. Azrin, J. Cowie, and G. Olsen, private communication, Army Materials Technology Laboratory, 1986.

2. D. M. Tracey and C. E. Freese, "Adaptive Load Incrementation in Elastic-Plasic Finite Element Analysis," Computers and Structures, 13, pp. 45-53, 1981.
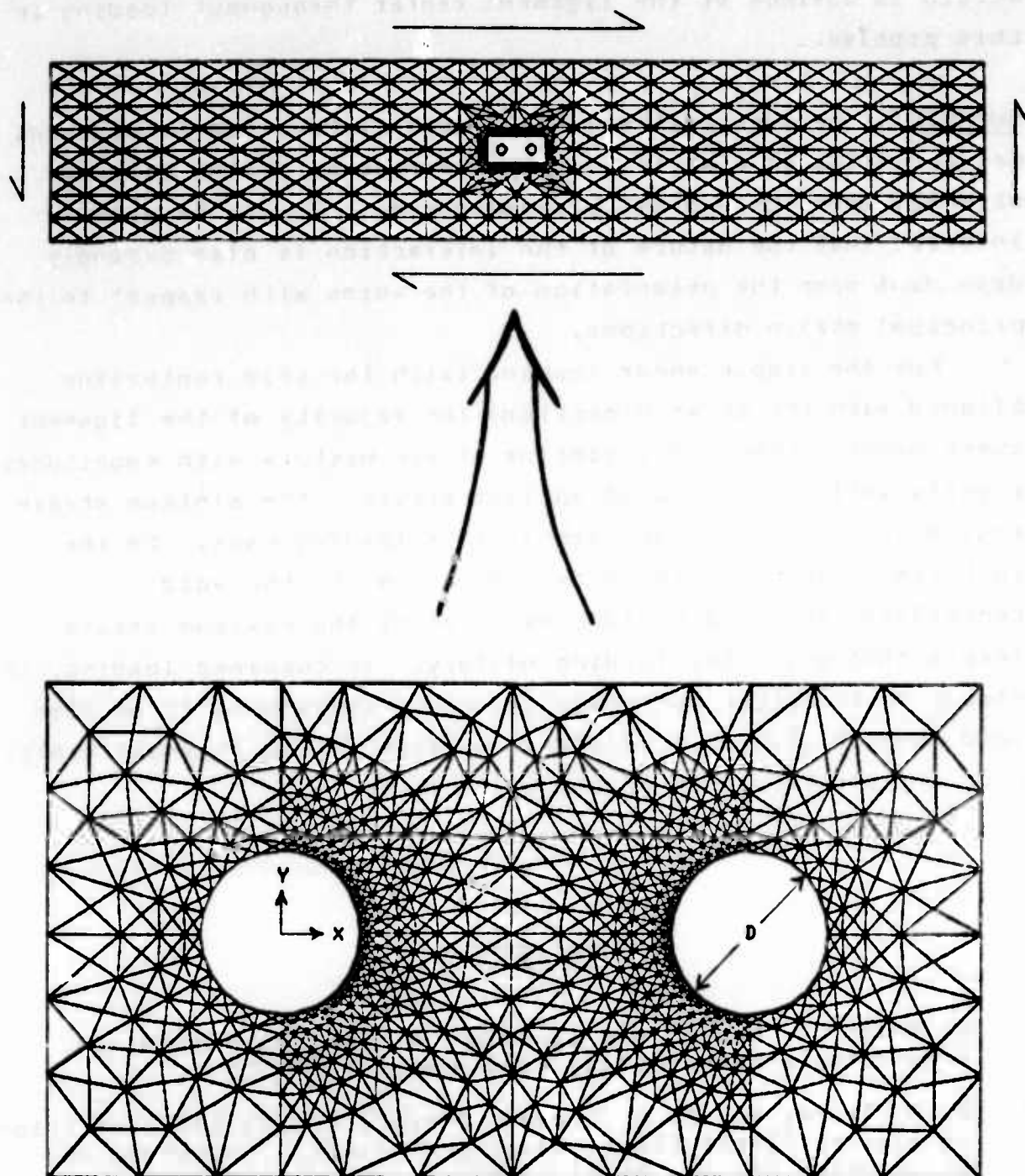
FIGURE 1 — Pair of voids in deformation band
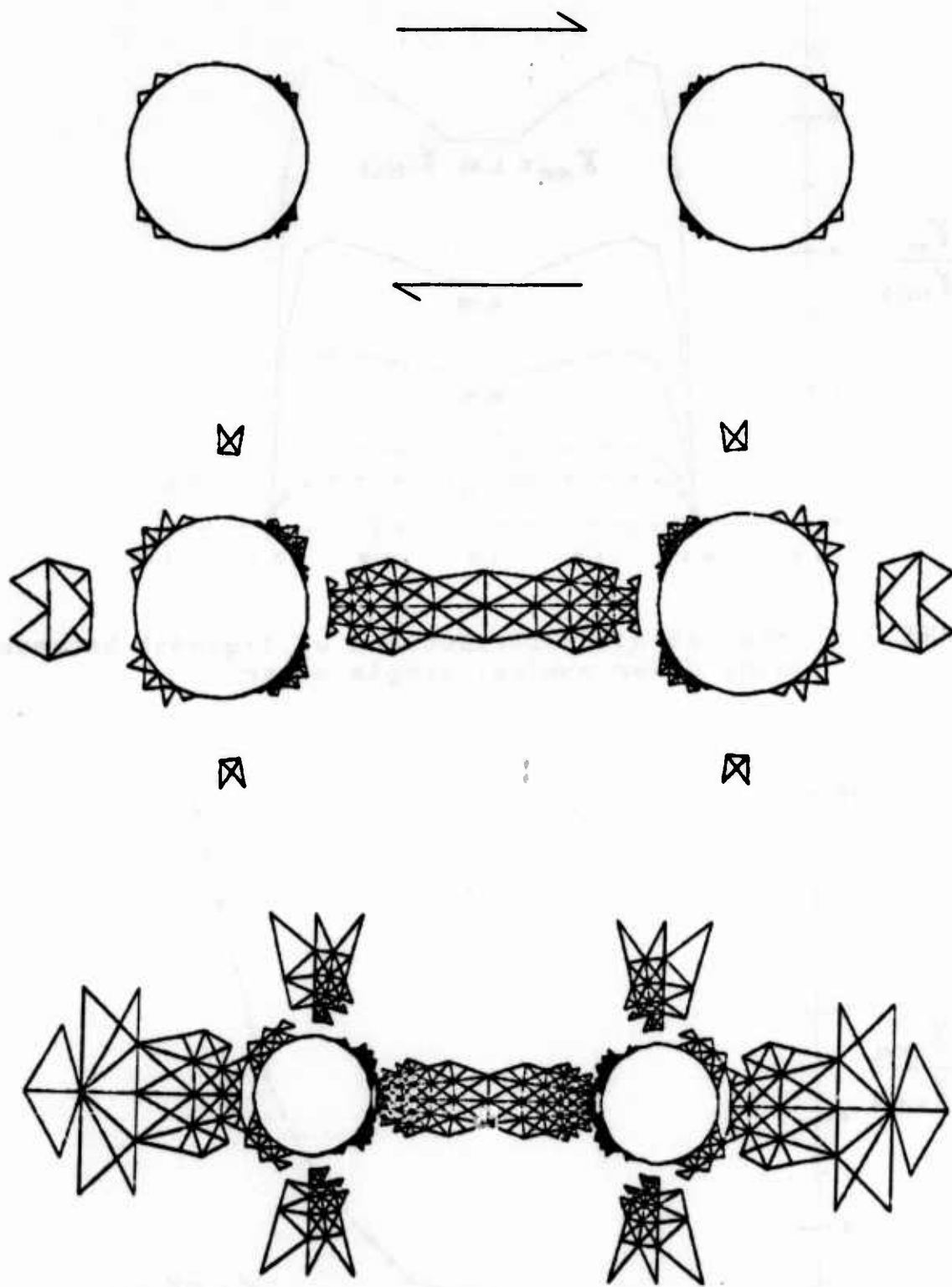
FIGURE 2 - Development of plastic instability between void
pair under simple shear. Plastic zones at nominal
shear strain levels 60, 77, and 91% of γ_{yield}

111
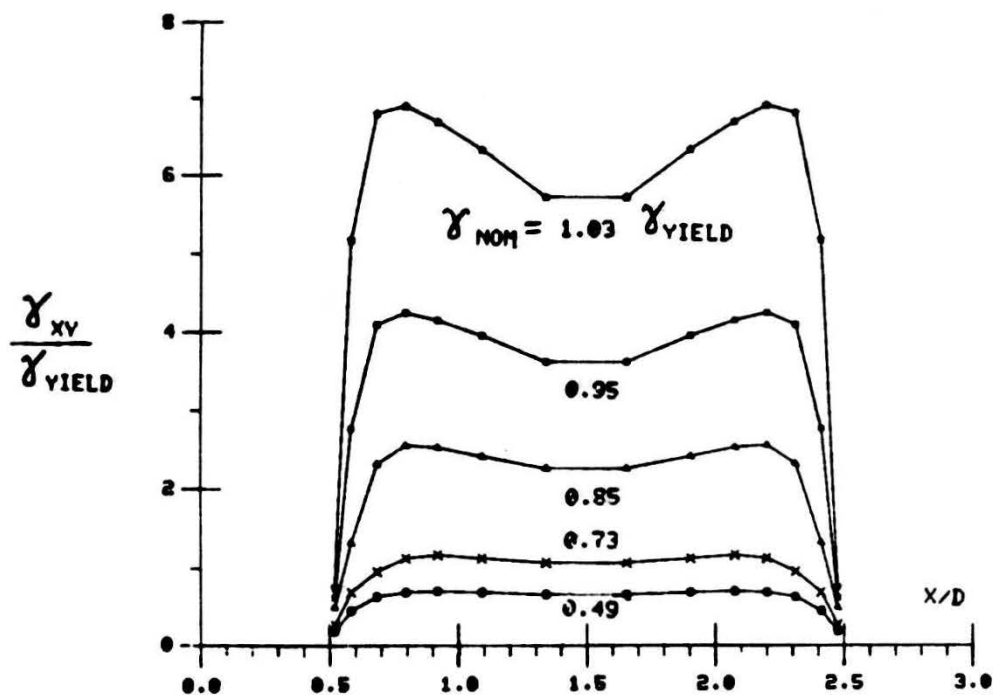
FIGURE 3 - Shear strain distributions on ligament between voids under nominal simple shear
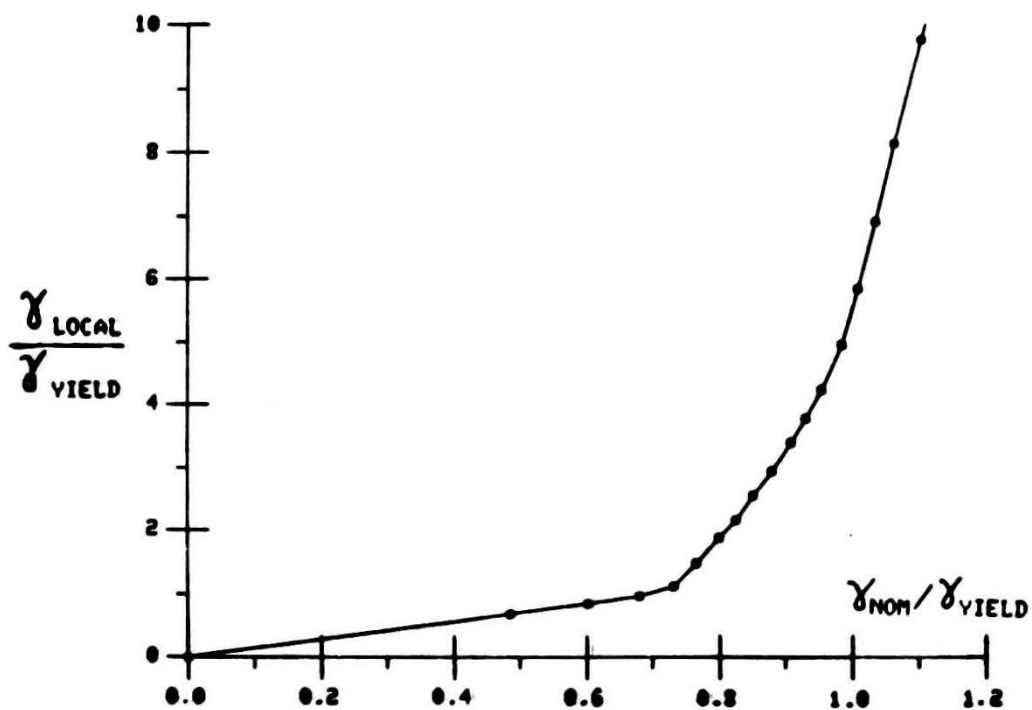
FIGURE 4 - Strain intensification between voids under nominal simple shear
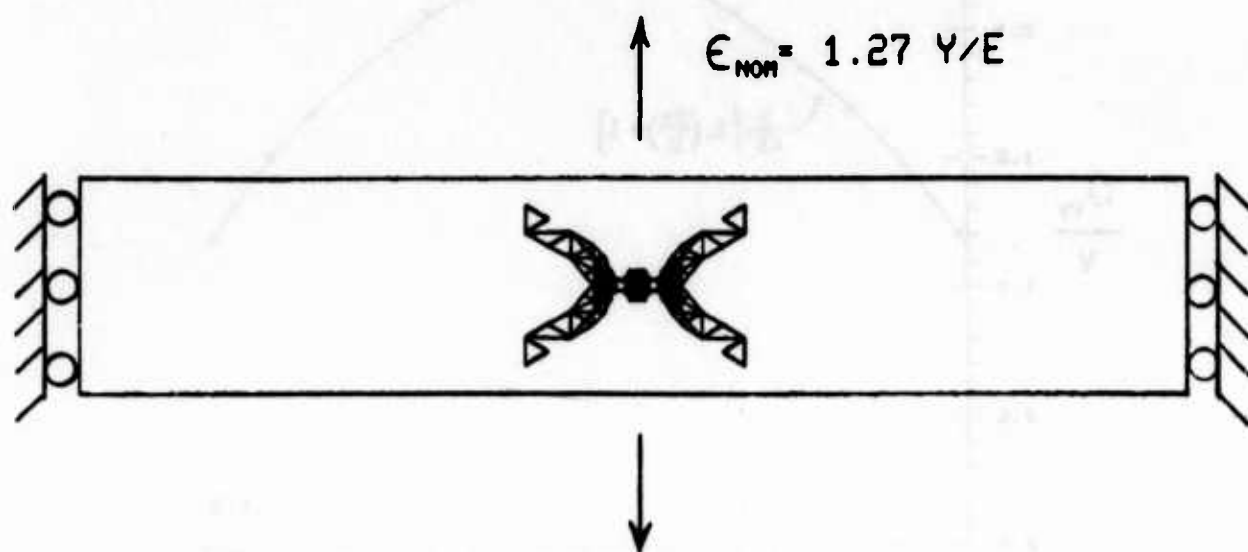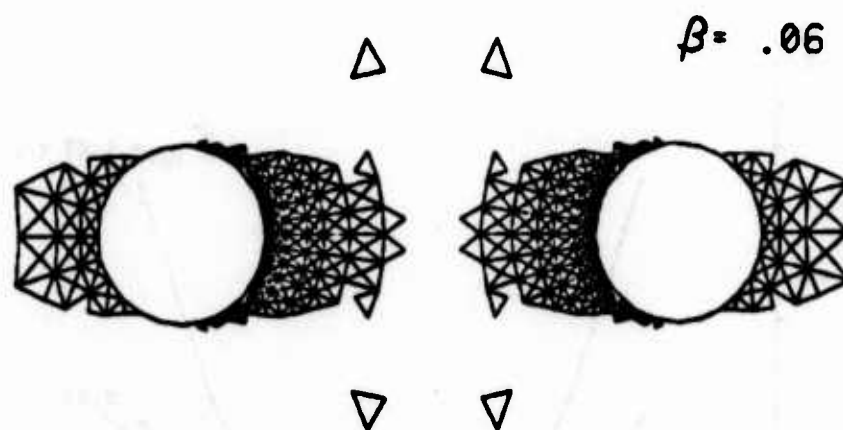
FIGURE 5 - (A) Plastic zone under nominal uniaxial strain
level 1.27 Y/E
(B) Fully developed near-isocloric portion of
plastic zone shown in (A)

113
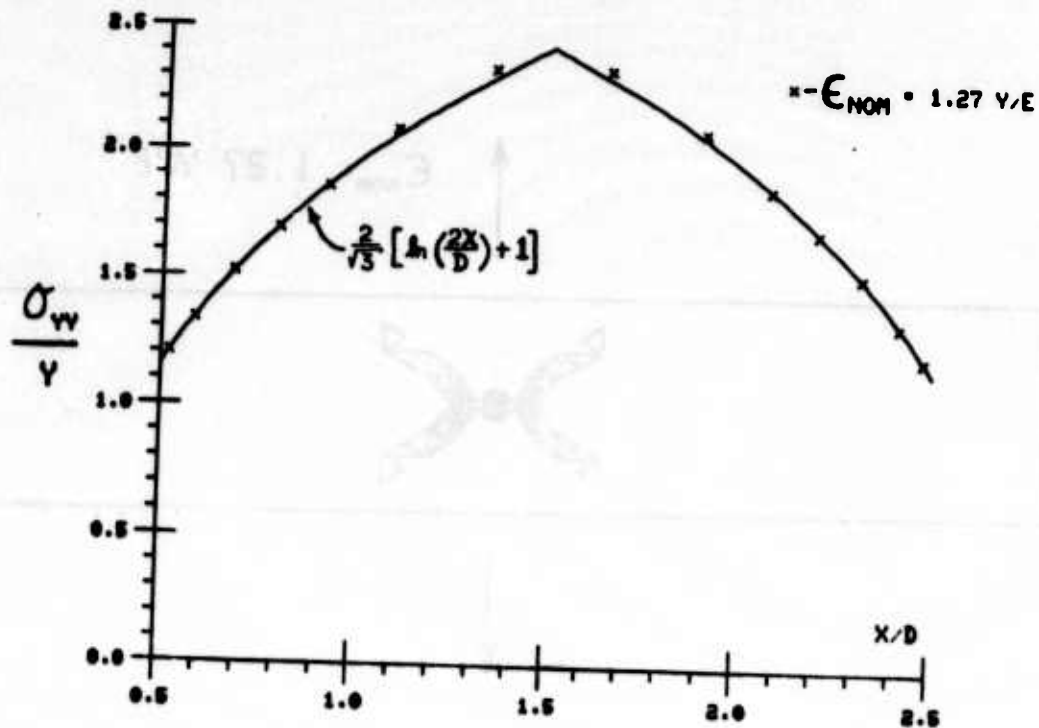
FIGURE 6 - Numerical data for normal stress on ligament at nominal uniaxial strain 1.27 Y/E compared with slipline theory logarithmic spiral distribution
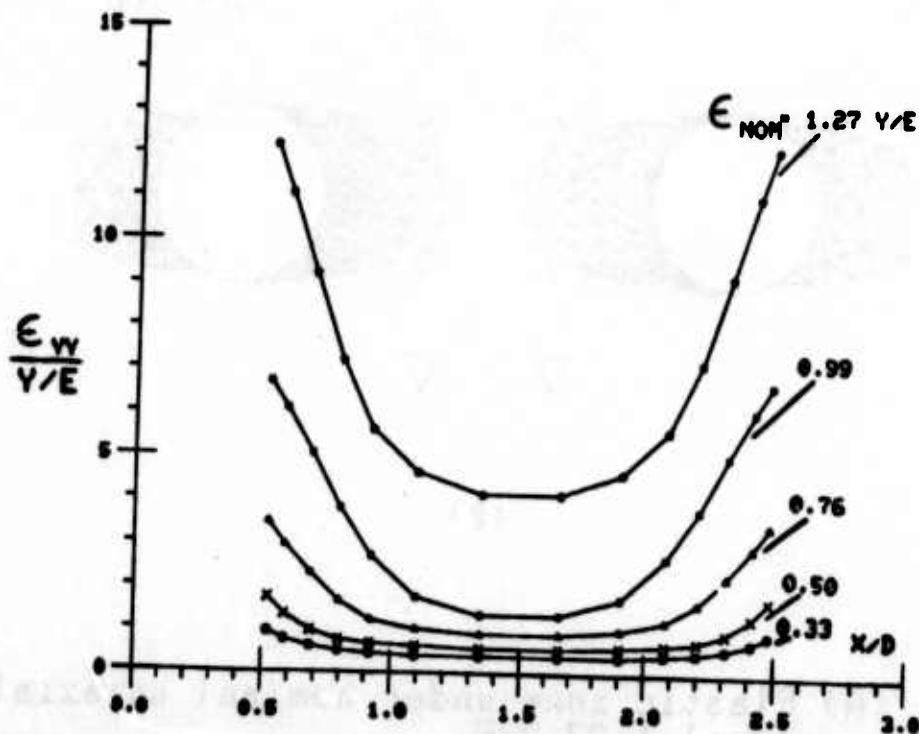
FIGURE 7 - Normal strain distribution on ligament between voids under nominal uniaxial strain

$\epsilon_{\text{NOM}} = \gamma_{\text{NOM}} = 0.99 \, Y/E$

(A)

$\beta = .06$
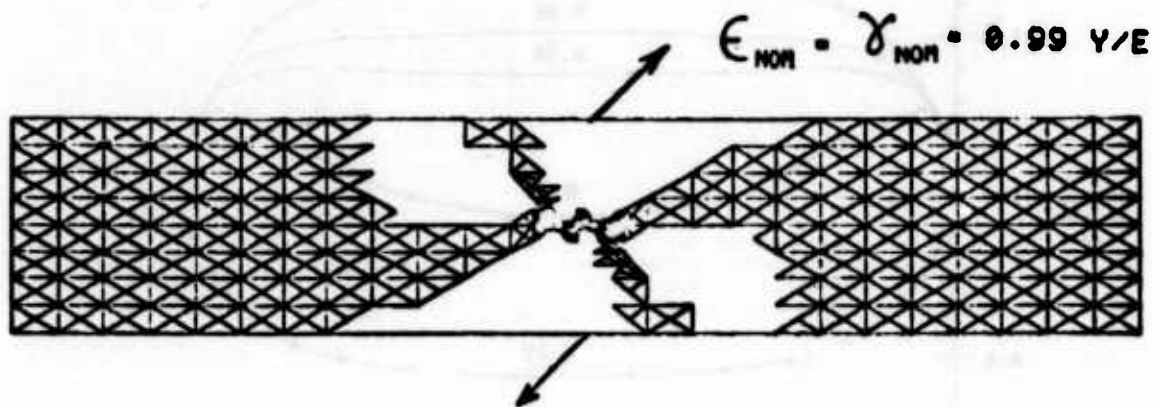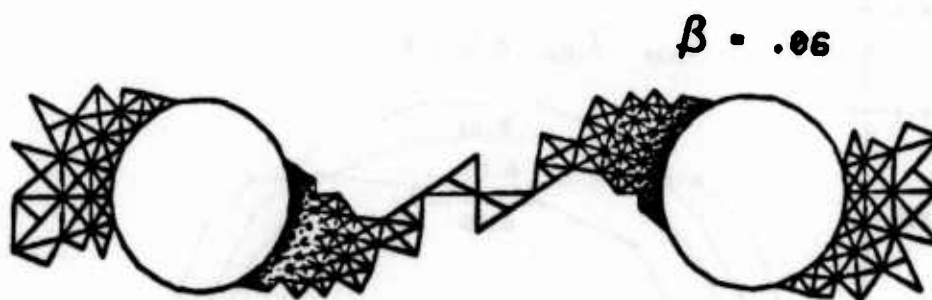
(B)

FIGURE 8 - (A) Plastic zone under nominal combined strain
level 0.99 Y/E
(B) Fully developed near-isocloric portion of
plastic zone shown in (A)

115
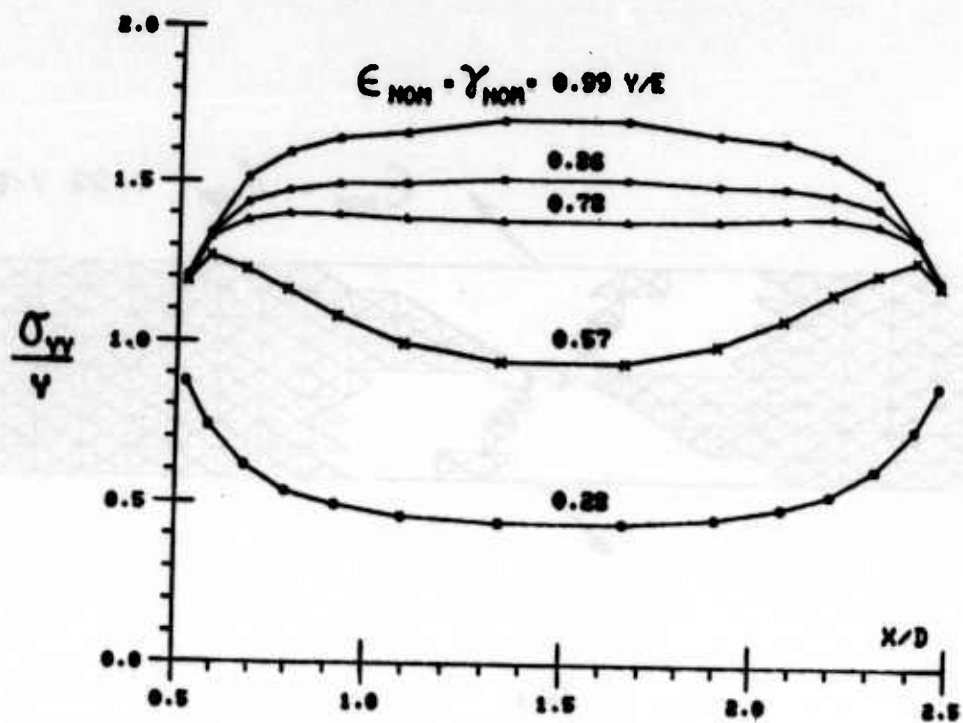
FIGURE 9 — Nominal stress on ligament at various levels of imposed combined strain
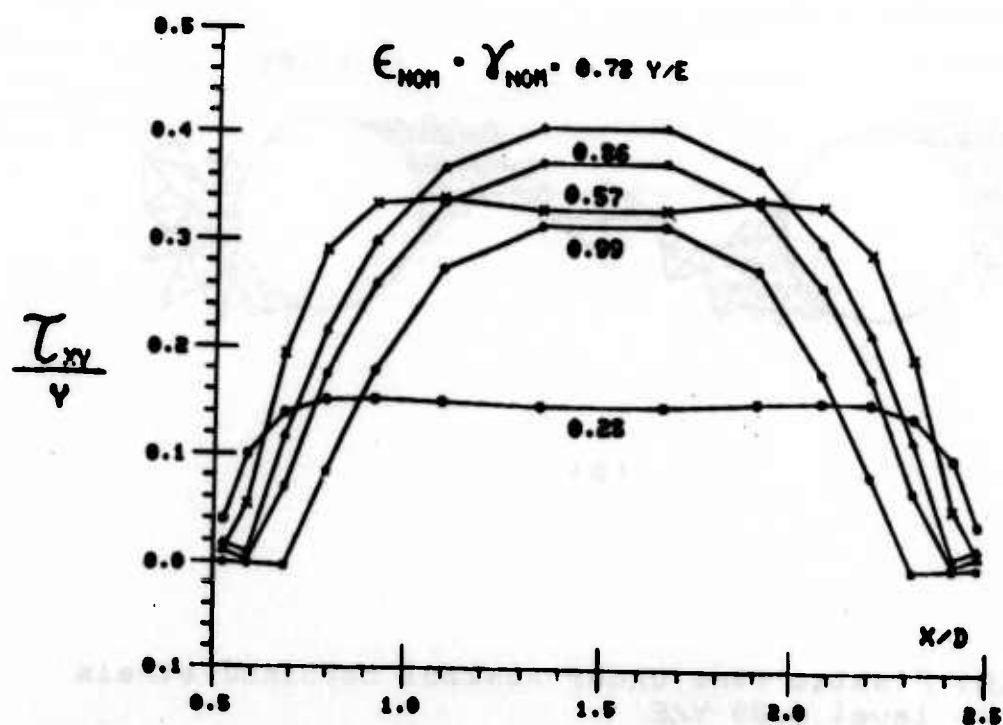
FIGURE 10 — Shear strain on ligament at various levels of imposed combined strain

116

FRACTALS, FRAGMENTATION, AND FAILURE

Donald L. Turcotte
Department of Geological Sciences
Cornell University
Ithaca, NY 14853

ABSTRACT. Many applied problems exhibit a fractal character; a necessary condition is that the fundamental phenomena be scale invariant over a reasonably wide range of scales. In many cases a renormalization group approach can give an applicable solution. A specific example is fragmentation. The number-size distribution for fragments often satisfies the power law fractal relation. A renormalization group approach can be used to obtain the fractal dimension associated with catastrophic fragmentation. The renormalization group approach can also be applied to the failure of a fractal network. A gridded network can be constructed that obeys fractal geometrical constraints. The elements of the network are given a statistical distribution of strengths. Stress transfer from failed elements to adjacent sound elements is an essential feature of the analysis. The renormalization group approach specifies a catastrophic failure criteria for the network. An example of an application is the failure of a stranded cable that has been constructed according to fractal constraints.

1. INTRODUCTION. It is recognized that there are a variety of scale invariant processes in nature; the concept of fractals provides a means of quantifying these processes. A fractal distribution can be defined by

$$N \sim r^{-D} \tag{1}$$

where N is the number (of objects) with a characteristic linear dimension greater than r, and D is the fractal dimension. The original definition of a fractal [1] related the length (perimeter) P of a coastline (or topographic contour) to the length of a yardstick r by

$$P \sim Nr \sim N^{1-D} \tag{2}$$

Typical coastlines or topographic contours have D = 1.2 - 1.3.

2. FRAGMENTATION. A material can be fragmented in a variety of ways. Rocks can be fragmented by joints and weathering. In this case the distribution of fragment sizes is likely to be determined by the preexisting planes of weakness in the rock. Fragments can also be produced by explosives. Again preexisting planes of weakness may determine the distribution of fragment sizes. Fragments can also be produced by impacts. Impacts are likely to have played a dominant role in determining the number-size relation for asteroids and meteorites.

A variety of statistical distributions have been used to correlate the number-size data on fragments. However, in many cases a power law relation was determined. Some of these results are given in Figure 1. Included are data

117

for broken coral [2], an underground nuclear explosion [3], and a basalt fragmented by an impacting projectile [4]. Other examples have been tabulated by Turcotte [5].

The applicability of a fractal distribution indicates that fragmentation is scale invariant over a wide range of scales. In order to model fragmentation we will use the renormalization group approach. For simplicity we will consider a cube of material with a linear dimension h as illustrated in Figure 2. This cube is referred to as a cell that is divided into eight cubic elements each with a dimension $h/2$. Attention is now focused on one of the cubic elements, and it becomes a cell of dimension $h/2$ at order 1. This cell is then divided into eight first-order elements each with a dimension $h/4$ as illustrated in Figure 2. The process is repeated at successively higher orders.
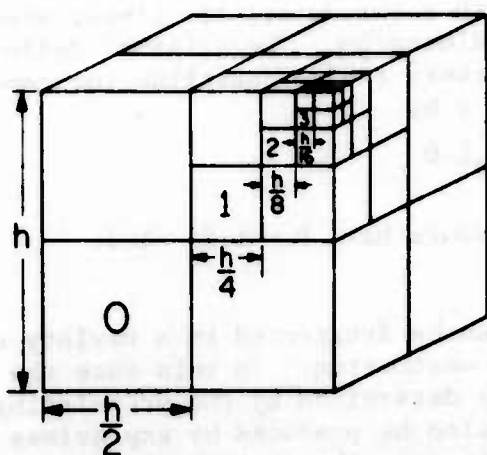
FIGURE 2. Illustration of the renormalization group approach to fragmentation. A zero-order cubic cell with dimension h is divided into eight cubic elements with dimension $h/2$. Each of these elements becomes a first-order cell and is divided into eight first-order elements with dimensions $h/4$. The process is repeated to higher orders.

The basic hypothesis of the renormalization group approach is the assumption that the probability that a cell will fragment into eight elements is the same at all orders. If initially there are N_0 cubes of dimension h, the number remaining after fragmentation is $N_0' = (1-p_c)N_0$. The number of fragments with dimension $h_1 = h/2$ is $N_1 = 8 \, p_c(1-p_c)N_0$, the number of fragments with dimension

$h_2 = h/2^2$ is $N_2 = (8p_c)^2 (1-p_c)N_0$, and generalizing the number of fragments with dimension $h_n = h/2^n$ is $N_n = (8p_c)^n (1-p_c)N_0$. The generalized form can be written

$$\ln \frac{h_n}{h} = - n \ln 2 \tag{3}$$

$$\ln \frac{N_n}{N_0'} = n \ln(8p_c) \tag{4}$$

And eliminating n gives

$$\frac{N_n}{N_0'} = \left(\frac{h_n}{h} \right)^{- \frac{\ln(8p_c)}{\ln 2}} \tag{5}$$

Comparison with (1) gives

$$D = \frac{\ln(8p_c)}{\ln 2} \tag{6}$$

Thus the fractal dimension D is directly related to probability that a fragment of a given size is broken into smaller elements. The probability p_c is dependent on the specific model chosen but the fractal dimension is independent of the model.

We next determine a specific value for D by specifying a fragmentation model. Following Allègre et al. [6] each element in a cell is hypothesized to be either fragile or sound. It is necessary to determine a condition for the probability that a cell is fragile p_n in terms of the probability that an element is fragile p_{n+1}. In each cell there can be zero to eight fragile elements; there are $2^8 = 256$ possible combinations. Excluding multiplicities, there are 22 topologically different configurations.

We hypothesize that the sides of a fragile element form planes of weakness. If the sides of fragile elements form an internal plane through the cell, the cell is assumed to be fragile. Examples of sound and fragile cells are given in Figure 3. In each case there are four fragile elements (shaded). In "a" no internal planes of weakness cut the cell, so it is sound; in "b" both a horizontal and a vertical plane of weakness cuts the cell and it is fragile.

a   b
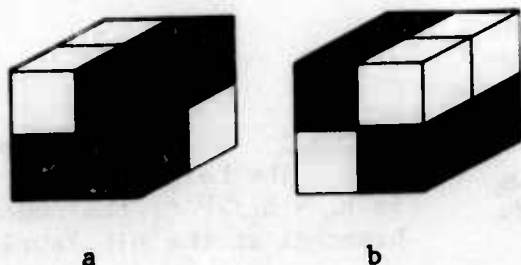
FIGURE 3. Each cubic cell contains four fragile elements (shaded) and four sound elements. In "a" the cell is sound, and in "b" the cell is fragile.

119

The probability P' that a cell is fragile is related to the probability P that an element is fragile by

$$p_n = 3 p_{n+1}^8 - 32 p_{n+1}^7 + 88 p_{n+1}^6 - 96 p_{n+1}^5 + 38 p_{n+1}^4 \qquad (7)$$

The dependence of p_n on p_{n+1} is given in Figure 4. The straight line $p_n = p_{n+1}$ is also included. The points 0 and 1 are stable fixed points of the system. The iterative relation crosses the straight line at $p_n = p_{n+1} = p_c = 0.490$. This is an unstable fixed point that separates the region of stable behavior from the region of unstable behavior. The critical probability p_c corresponds to catastrophic fragmentation of the object. Assuming that each fragmented cube breaks into eight pieces we find from (6) that $p_c = 0.490$ corresponds to D = 1.97.
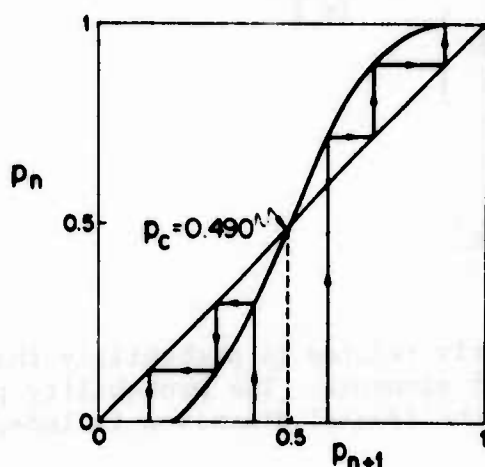
FIGURE 4. Probability of fragility at order n, p_n, as a function of the probability of fragility at order n+1, p_{n+1}, from (7). The critical probability p_c corresponding to catastrophic fragmentation is 0.490. The corresponding fractal dimension from (6) is 1.97.

3. NETWORK FAILURE. The renormalization group technique can also be applied to the failure of a fractal network [7,8]. The network is modeled as the fractal tree illustrated in Figure 5. Failure is associated with the application of a vertical load V to the tree. Each branch is given a random strength such that the probability of failure of the branch under load v is given by a Weibull distribution

$$P_1 = 1 - \exp[-(v/v_0)^2] \qquad (8)$$

where v_0 is a reference load. For each pair of branches the probability of failure of both is P_1^2, one is $2 P_1(1-P_1)$, and neither is $(1-P_1)^2$.
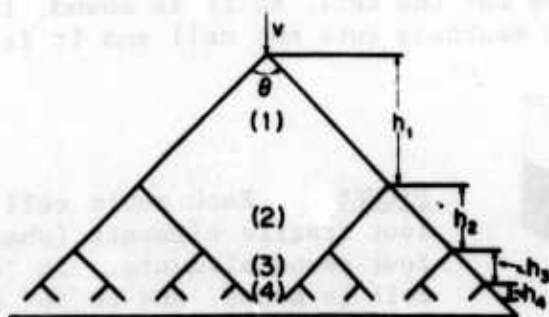
FIGURE 5. Illustration of a fractal tree. The height of the n+h level is $h_n = h_1/2^{n-1}$; the number of branches at the n+h level is 2^n.

However, we hypothesize that if one branch fails, the load is transferred to the adjacent branch in the pair. The second branch may suffer induced failure due to this transfer. We use the conditional probability $P_{2,1}$, that an unbroken branch already supporting a load v will fail when an additional load v is transferred to it from an adjacent broken branch. The probability that a pair will fail is given by

$$n^{-1}P_1 - {}^nP_1{}^2 + 2 \; {}^nP_1 \; (1-{}^nP_1) \; {}^nP_{2,1} \tag{9}$$

The conditional probability is given by

$${}^nP_{2,1} - \frac{{}^nP_2 - {}^nP_1}{1-{}^nP_1} \tag{10}$$

where ${}^nP_2$ is the probability of failure under load 2v. For the second order Weibull distribution given in (8) we have

$${}^nP_2 - 1-(1-{}^nP_1)^4 \tag{11}$$

Combining (9), (10), and (11) gives

$$n^{-1}P_1 - 2P_1 \; [1-(1-{}^nP_1)^4] - {}^nP_1{}^2 \tag{12}$$

This is a failure condition for a pair of branches.

It is implicit in the renormalization group approach that the failure condition (12) is applicable at all levels of the fractal tree. The dependence of $n^{+1}P_1$ on ${}^nP_1$ is given in Figure 6. Again the characteristic S-shaped curve is obtained, and the straight line $n^{-1}P_1 - {}^nP_1$ is included. Two stable points are 0 and 1 and the unstable fixed point is ${}^nP_1 - n^{-1}P_1 - P_1{}^* - 0.2063$. The corresponding value of the critical load from (8) is $V^* \sim 0.4806V_0$. It is interesting to note that this critical load is considerably less than the mean strength of a branch that is $\bar{v} - 0.8862v$ from (8).
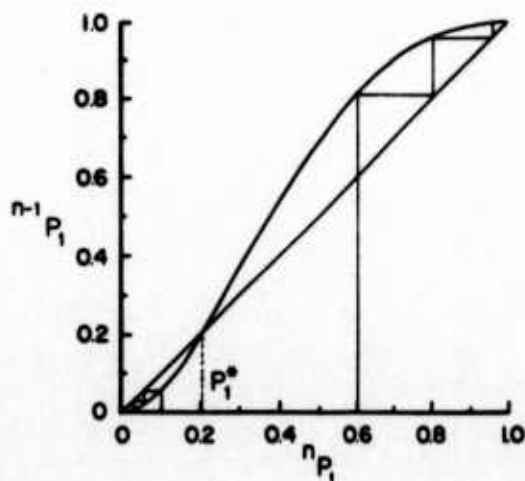
FIGURE 6. Probability of failure of a branch at the n-1 level $n^{-1}P_1$, as a function of the probability of failure at the n level, ${}^nP_1$, from (12). The critical probability P_* corresponding to catastrophic failure is 0.2063.

121

REFERENCES

1. B. Mandelbrot, How long is the coast of Britain? Statistical self-similarity and fractional dimension: Science 156, 636-638 (1967)

2. J.G. Bennett, Broken coal: J. Insti. Fuel 10, 22-39 (1936).

3. J.E. Schoutens, Empirical analysis of nuclear and high-explosive cratering and ejecta: Nuclear Geoplosics Sourcebook, Vol. LV, Part II, Section 4, Defense Nuclear Agency Report DNA 65 01H-4-2 (1979).

4. A. Fujiwara, G. Kamimoto, and A. Tsukamoto, Destruction of basaltic bodies by high-velocity impact: Icarus 31, 277-288 (1977).

5. D.L. Turcotte, Fractals and fragmentation: J. Geophys. Res. 91, 1921-1926 (1986).

6. C.J. Allègre, J.L. Le Mouel, and A. Provost, Scaling rules in rock fracture and possible implications for earthquake prediction: Nature 297, 47-49 (1982).

7. R.F. Smalley, D.L. Turcotte, and S.A. Solla, A renormalization group approach to the stick-slip behavior of faults: J. Geophys. Res. 90, 1894-1900 (1985).

8. D.L. Turcotte, R.F. Smalley, S.A. Solla, Collapse of loaded fractal trees: Nature 313, 671-672 (1985).

NUMERICAL SOLUTION TO A SYSTEM OF
RANDOM VOLTERRA INTEGRAL EQUATIONS*

N. Medhin[1], M. Sambandham[2] and C. K. Zoltani[3]

Abstract

In this article we present a brief summary of the numerical solution of a system of random Volterra integral equations. The methods we use are (i) Newton's method and (ii) successive approximation method. Based on the simulation, we discuss the mean and variance of the solution of a system of random Volterra integral equations.

[1]Department of Mathematics and Computer Science, Atlanta University, Atlanta, GA 30314.

[2]Center for Computational Sciences, Atlanta University, Atlanta, GA 30314; Department of Mathematics, Morehouse College, Atlanta, GA 30314.

[3]Ignition and Combustion Branch, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland 21005.

1. INTRODUCTION

The study of random Volterra integral equations and their applications, is an active area of research in probabilistic analysis. However, the numerical treatment of a system of random Volterra integral equations is yet to be explored. For a recent survey of approximate solution of random integral equations we refer to Bharucha-Reid and Christensen [4]. For the numerical treatment of random integral equations we refer to Bharucha-Reid [3,5], Becus [2], Christensen and Bharucha-Reid [6], Lax [8,9,10], Medhin and Sambandham [11,12], Sambandham [14,15], and Tsokes and Padgett [16]. Among other methods of successive approximation, stochastic approximation methods, method of moments, method of averaging, projection method, Newton's method, etc. are used to obtain the numerical solution of random integral equations. Most of the results in the literature are linear or one dimensinal equations. For the numerical treatment of deterministic integral equations we refer to Baker [1].

In this article we consider a system of random Volterra integral equations. By an application of successive approximation method and Newton's method, we examine the method of obtaining the numerical solution of a system of random Volterra integral equations. We organize our article as follows. By an application of Newton's method and successive approximation method, we develop useful numerical procedures respectively in Sections 2 and 3. In Section 4 we include a short discussion.

2. NEWTON'S METHOD

In this section we use Newton's method to obtain the numerical solution of a system of random Volterra integral equations.

Let (Ω, F, P) be a complete probability space and let X be a separable Hilbert space with *inner product* $(\cdot, \cdot)$. A mapping $T: \Omega \times X \to X$ is said to be *random* if and only if the function $<T(\omega)x,y>$ is a scalar valued random variable for every $x,y \in X$. In other words, $T(\omega)$ is a *random operator* if and only if $T(\omega)x$ is an X-valued random variable for every $x \in X$. A random operator is *linear* if $T(\omega)(\alpha x_1 + \beta x_2) = \alpha T(\omega)x_1 + \beta T(\omega)x_2$ a.s. for every $x_1, x_2 \in X$ and scalars α, β. $T(\omega)$ is said to be *bounded random operator* if there exists a non-negative real-valued random variable $M(\omega)$ such that for all $x_1, x_2 \in X$, $\|T(\omega)x_1 - T(\omega)x_2\| \leq M(\omega) \|x_1 - x_2\|$ a.s. If $T(\omega)$ is a bounded random operator on X, then $T^{-1}(\omega)$ is a random operator which maps $T(\omega)x$ into x a.s. $T(\omega)$ is said to be *invertible* if $T^{-1}(\omega)$ exists. If $T(\omega)$ is an invertible random

operator which is bounded, then $T^{-1}(\omega)$ is a bounded random operator also.

Now let us consider the random operator equation $T(\omega)x(\omega) = x(\omega)$, where $T(\omega): \Omega \times X \to X$ is a random nonlinear operator. For our purpose it is enough if $T(\omega)$ is defined on an open set U of X. Let $T(\omega)$ be continuously differentiable a.s. and let $x_0(\omega): \Omega \to U$ be an X-valued random variable such that $[I-T'(\omega)x_0]^{-1}: \Omega \times X \to X$ be defined and bounded, where $T'(\omega)x_0$ is random. It follows from the theorems of Hans [7] and Nashed-Salehi [13] that $[I-T'(\omega)x_0(\omega)]^{-1}$ is a random bounded linear operator .

Let $k(\omega): \Omega \to (0,1)$ be a real valued random variable; and let $T_0^{-1}(\omega)$ be a bounded linear random operator such that $\|T_0^{-1}(\omega)[I-T'(\omega)x_0(\omega)] - I\| \le k(\omega) < 1$. We denote by $B(x_0,r)$ the collection of all X-valued random variables $x(\omega)$ such that $x(\omega) \in U$ and $\|x(\omega)-x_0(\omega)\| \le r$. Then if

$$\|T_0^{-1}(\omega)[I-T(\omega)]x_0(\omega)\| \le r(1-k(\omega)),$$

there exists random variable $\hat{x}(\omega) \in B(x_0,r)$ such that $T(\omega)\hat{x}(\omega) = \hat{x}(\omega)$ a.s. (that is $\hat{x}(\omega)$ is a *fixed point* of $T(\omega)$), and the sequence of X-valued random variables defined by

$$x_{n+1}(\omega) = x_n(\omega) - T_0^{-1}(\omega)[I-T(\omega)]x_n(\omega), \qquad (2.1)$$

$n = 0,1,2,\ldots$, converges to $\hat{x}(\omega)$ a.s. For further discussion on this line we refer to [5,12].

As an application, we implement the above method to the following system. Consider

$$f_1(t,\omega) = g_1(t,\omega) + \int_0^t k_1(t,\tau,\omega)M_1(\tau,f_1(\tau,\omega),f_2(\tau,\omega),\nu_1(\omega))d\tau$$

(2.2)

$$f_2(t,\omega) = g_2(t,\omega) + \int_0^t k_2(t,\tau,\omega)M_2(\tau,f_1(\tau,\omega),f_2(\tau,\omega),\nu_2(\omega))d\tau,$$

where the solution vector is $(f_1(t,\omega),f_2(t,\omega))$ and the functions g_i, k_i, ν_i, $i = 1,2$ are assumed to be well defined so that $(f_1(t,\omega),f_2(t,\omega))$ exists with probability 1. Let

$$r_i^k(t,\omega) = f_i^k(t,\omega) - \int_0^t k_i(t,\tau,\omega)M_i(\tau,f_1^k(\tau,\omega),f_2^k(\tau,\omega),\nu_i(\omega))d\tau$$

$$- g_i(t,\omega), \quad i = 1,2, \quad k = 0,1,2,3,\ldots \quad (2.3)$$

$$\epsilon_1^k(t,\omega) - \int_0^t [k_1(t,\tau,\omega)\partial_{v_1}M_1(\tau,f_1^k(\tau,\omega),f_2^k(\tau,\omega),\nu_1(\omega))\epsilon_1^k(\tau,\omega)$$

$$+ k_1(t,\tau,\omega)\partial_{v_2}M_1(\tau,f_1^k(\tau,\omega),f_2^k(\tau,\omega),\nu_1(\omega))\epsilon_2^k(\tau,\omega)]d\tau$$

$$= r_1^k(t,\omega) \tag{2.4}$$

$$\epsilon_2^k(t,\omega) - \int_0^t [k_2(t,\tau,\omega)\partial_{v_1}M_2(\tau,f_1^k(\tau,\omega),f_2^k(\tau,\omega),\nu_2(\omega))\epsilon_1^k(\tau,\omega)$$

$$+ k_2(t,\tau,\omega)\partial_{v_2}M_2(\tau,f_1^k(\tau,\omega),f_2^k(\tau,\omega),\nu_2(\omega))\epsilon_2^k(\tau,\omega)]d\tau$$

$$= r_2^k(t,\omega) \tag{2.5}$$

where

$$k = 0,1,2,\ldots$$

$$\partial_{v_1}M_i(t,v_1,v_2,\nu) = \frac{\partial}{\partial v_1} M_i(t,v_1,v_2,\nu)$$

$$\partial_{v_2}M_i(t,v_1,v_2,\nu) = \frac{\partial}{\partial v_2} M_i(t,v_1,v_2,\nu)$$

$$i = 1,2.$$

Now we set

$$\begin{pmatrix} f_1^{k+1}(t,\omega) \\ f_2^{k+1}(t,\omega) \end{pmatrix} = \begin{pmatrix} f_1^k(t,\omega) \\ f_2^k(t,\omega) \end{pmatrix} - \begin{pmatrix} \epsilon_1^k(t,\omega) \\ \epsilon_2^k(t,\omega) \end{pmatrix}, \quad k = 0,1,2,\ldots \tag{2.6}$$

According to Newton's iteration $(f_1^k(t,\omega),f_2^k(t,\omega))$ converges to the solution of (2.2).

For numerical procedure, integrals in (2.3)-(2.5) can be evaluated by any suitable numerical procedure, for example, collocation or quadrature method.

Example 2.1: In the following example we illustrate the above procedure. Let

$$f_1(t) + \int_0^t t(1-\tau)[f_1^2(\tau)+f_2(\tau)+\nu(\omega)]d\tau = g_1(t,\omega),$$

$$\tag{2.7}$$

$$f_2(t) + \int_0^t (t-\tau)[f_1(\tau)+f_2^2(\tau)+\nu(\omega)]d\tau = g_2(t,\omega).$$

Then (2.3)-(2.5) reduce to

126

$$r_1^k(t,\omega) = f_1^k(t) + \int_0^t t(1-\tau)[f_1^{k^2}(\tau)+f_2^k+\nu(\omega)]d\tau - g_1(t,\omega)$$

$$(2.8)$$

$$r_2^k(t,\omega) = f_2^k(t) + \int_0^t (t-\tau)[f_1^k(\tau)+f_2^{k^2}(\tau)+\nu(\omega)]d\tau - g_2(t,\omega)$$

$$g_1(t,\omega) = t + t^2 - \frac{t^3}{2} + \frac{t^4}{3} + \frac{t^5}{4} + R(\omega)$$

$$g_2(t,\omega) = 1 + \frac{t^2}{2} + \frac{t^3}{6} + R(\omega)$$

$$\epsilon_1^k(t,\omega) - \int_0^t \{2t(1-\tau)f_1^k(\tau,\omega)\epsilon_1^k(t,\omega)+t(1-\tau)\epsilon_2^k(\tau,\omega)\}d\tau = r_1^k(t,\omega)$$

$$(2.9)$$

$$\epsilon_2^k(t,\omega) - \int_0^t \{(t-\tau)\epsilon_1^k(\tau,\omega)+2(t-\tau)f_2^k(\tau,\omega)\epsilon_2^k(\tau,\omega)\}d\tau = r_2^k(t,\omega)$$

$$(f_1^0(t,\omega),f_2^0(t,\omega)) = (R(\omega),1+R(\omega)) \qquad (2.10)$$

Using (2.6), (2.8)-(2.10) the numerical solution of (2.7) can be obtained.

Let $N(m,\sigma^2)$ denote normal distribution with mean m and variance σ^2 and $U(a,b)$ denote uniform distribution in the interval (a,b). In our numerical experiment we have taken as follows:

Example 2.1 (a): $\nu(\omega) \in N(0,.002^2)$, $R(\omega) \in U(-.001,.001)$

Example 2.1 (b): $\nu(\omega) \in N(0,.02^2)$, $R(\omega) \in U(-.01,.01)$

Example 2.1 (c): $\nu(\omega) \in N(0,.02^2)$, $R(\omega) \in U(-.1,.1)$

Example 2.1 (d): $\nu(\omega) \in N(0,.2^2)$, $R(\omega) \in U(-.1,.1)$.

We used trapezoidal rule in (2.8) and (2.9) for numerical integration. Our simulation results are presented in Tables 2.1 - 2.4 and Figures 2.1 - 2.4. These results are based on 30 samples and in each sample iteration was repeated until $|f_i^{k+1}(t,\omega)-f_i^k(t,\omega)| < .001$, i = 1,2. In Figures 2.1 - 2.4, ∘∘∘∘ and ···· denote respectively $E(f_1(t,\omega))$ and $E(f_2(t,\omega))$ and ···· and ——— denote repsectively $V(f_1(t,\omega))$ and $V(f_2(t,\omega))$.

Table 2.1.

	$\nu(\omega) \in N(0,0.002^2)$, $R(\omega) \in U(-0.001,0.001)$			
t	$E(f_1)$	$V(f_1)$	$E(f_2)$	$V(f_2)$
1	.009964	.000004	.999676	.000004
10	.099967	.000004	.999681	.000004
20	.199975	.000004	.999693	.000004
30	.299988	.000003	.999712	.000003
40	.400006	.000003	.999739	.000003
50	.500026	.000003	.999774	.000002
60	.600036	.000003	.999800	.000002
70	.700045	.000003	.999827	.000001
80	.800039	.000003	.999829	.000001
90	.900031	.000003	.999811	.000001
100	1.000028	.000003	.999764	.000001

Table 2.2.

	$\nu(\omega) \in N(0,0.02^2)$, $R(\omega) \in U(-0.01,0.01)$			
t	$E(f_1)$	$V(f_1)$	$E(f_2)$	$V(f_2)$
1	.009639	.000361	.996755	.000388
10	.099666	.000360	.996789	.000380
20	.199741	.000355	.996884	.000357
30	.299841	.000344	.997024	.000319
40	.399931	.000329	.997175	.000272
50	.499988	.000310	.997310	.000218
60	.599987	.000290	.997383	.000163
70	.699914	.000271	.997331	.000114
80	.799772	.000256	.997067	.000079
90	.899621	.000244	.996521	.000064
100	.999542	.000240	.995575	.000082

Table 2.3.

	$\nu(\omega) \in N(0,0.02)$, $R(\omega) \in U(-.1,.1)$			
t	$E(f_1)$	$V(f_1)$	$E(f_2)$	$V(f_2)$
1	.009639	.000361	.996755	.000388
10	.099663	.000360	.996785	.000380
20	.199736	.000355	.996876	.000357
30	.299834	.000344	.997012	.000319
40	.399922	.000328	.997160	.000272
50	.499974	.000310	.997288	.000218
60	.599971	.000290	.997355	.000164
70	.699886	.000271	.997282	.000115
80	.799745	.000255	.997014	.000079
90	.899590	.000244	.996455	.000065
100	.999513	.000239	.995503	.000083

Table 2.4

	$\nu(\omega) \in N(0,0.2)$, $R(\omega) \in U(-.1,.1)$			
t	$E(f_1)$	$V(f_1)$	$E(f_2)$	$V(f_2)$
1	.006387	.036141	.967551	.038808
10	.096353	.036049	.967691	.038019
20	.196225	.035588	.968063	.035683
30	.295926	.034612	.968529	.031979
40	.395261	.033163	.968807	.027200
50	.494169	.031373	.968641	.021754
60	.592422	.029403	.967498	.016252
70	.690021	.027502	.964827	.011429
80	.787226	.025913	.960106	.008192
90	.884353	.024770	.952363	.007708
100	.982290	.024262	.940524	.011188
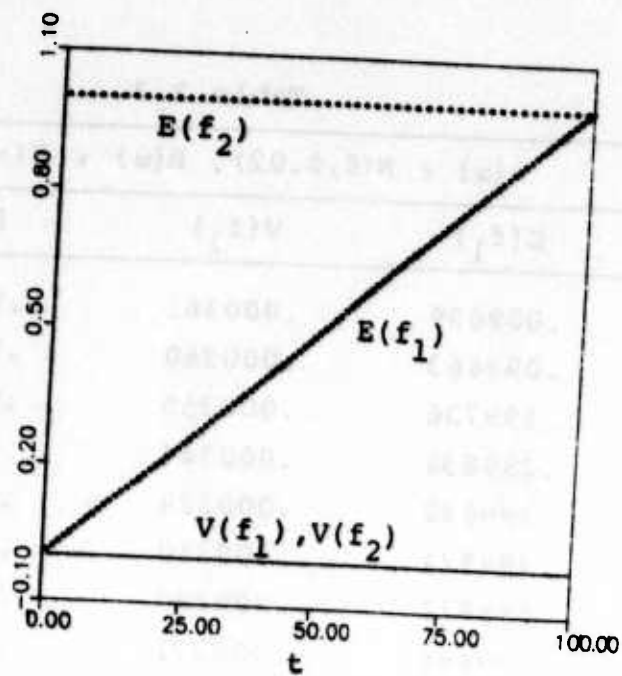
Fig. 2.1: Example 2.1 (a).

Fig. 2.2: Example 2.1 (b).

130
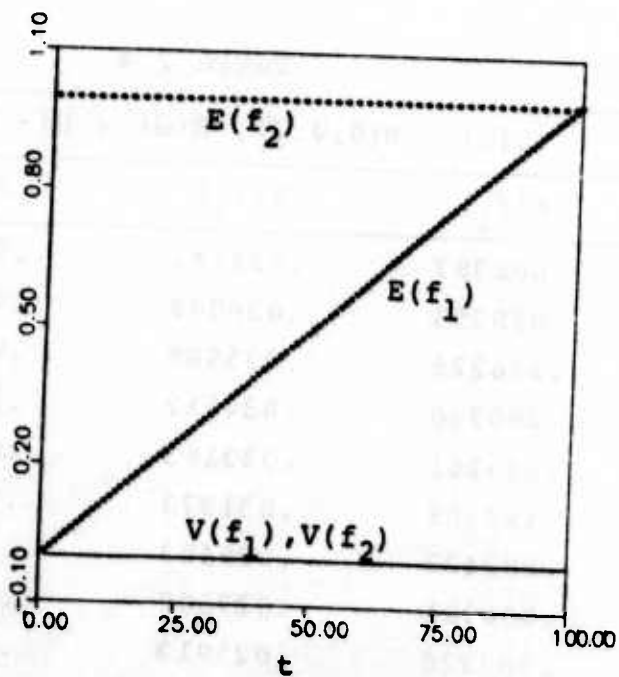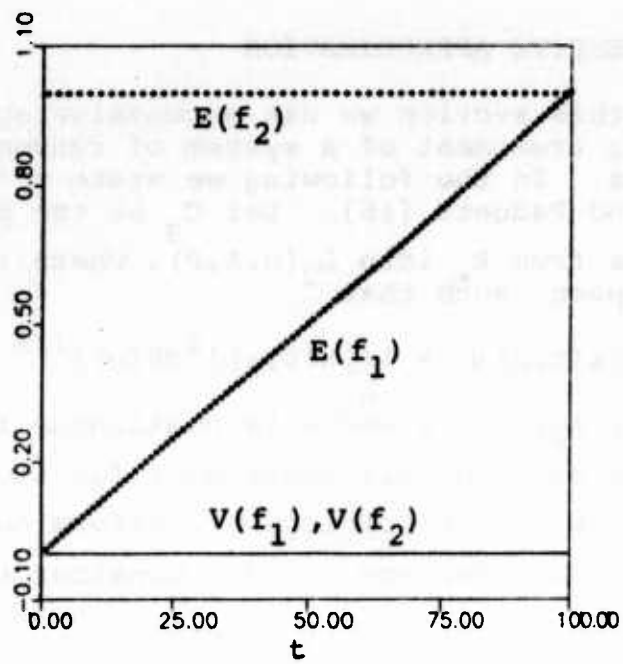
Fig. 2.3:   Example 2.1 (c).

Fig. 2.4:   Example 2.1 (d).

131
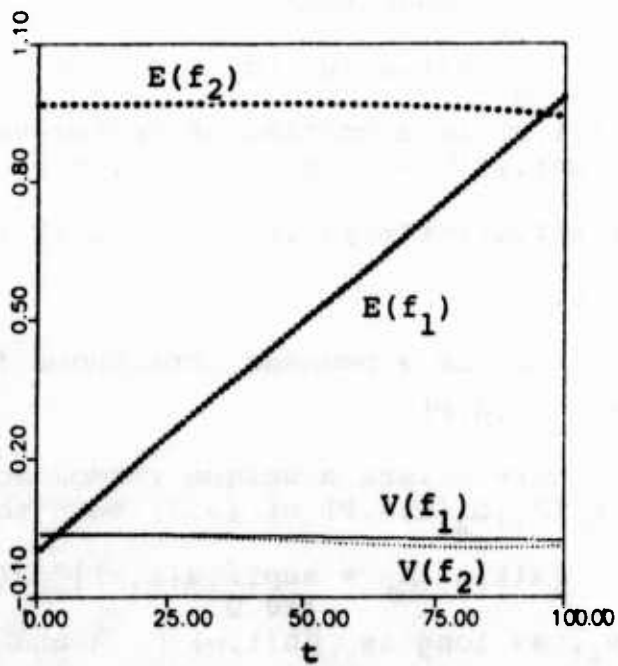
3. SUCCESSIVE APPROXIMATION

In this section we use successive approximation for the numerical treatment of a system of random Volterra integral equations. In the following we state a few useful results from Tsokes and Padgett [16]. Let C_g be the space of all continuous functions from R_+ into $L_2(\Omega, A, P)$, where (Ω, A, P) is a probability space, such that

$$\|x(t,\omega)\| = \{\int_\Omega |x(t,\omega)|^2 dP(\omega)\}^{1/2} \leq zg(t),$$

where $t \in R_+$, $z \geq 0$ and g is continuous function on R_+. Let C_c be the space of all continuous function from R_+ into $L_2(\Omega, A, P)$ with the topology of uniform convergence on the interval $[0, T]$ for any $T > 0$. Consider the random integral equation

$$x(t,\omega) = h(t,\omega) + \int_0^t k(t,\tau) f(\tau, x(\tau,\omega)) d\tau, \qquad (3.1)$$

where the following hypothesis hold.

(i) The function $(t,\tau) \to k(t,\tau)$ from the set $\Delta = \{(t,\tau), 0 \leq \tau \leq t < \infty\}$ onto R is continuous.

(ii) There exists a number $A > 0$ and a continuous function (on R_+) $g(t) > 0$ such that

$$\int_0^t |k(t,\omega)| g(\tau) d\tau \leq A, \qquad t \in R_+.$$

(iii) $f(t,x)$ is a continuous vector-valued function for $t \in R_+$, $\|x(t,\omega)\| \leq \rho$, such that $f(t,0) \in C_g$ and

$$\|f(t, x(t,\omega)) - f(t, y(t,\omega))\| \leq \lambda g(t) \|x(t,\omega) - y(t,\omega)\|$$

where $\lambda > 0$.

(iv) $h(t,\omega)$ is a bounded continuous function on R_+ with values in $L_2(\Omega, A, P)$.

Then there exists a unique random solution $x(t,\omega) \in C_c(R_+, L_2(\Omega, A, P))$ of (3.1) such that

$$\|x(t,\omega)\|_c = \sup_{t \geq 0} \{\int_\Omega |x(t,\omega)|^2 dP(\omega)\}^{1/2} \leq \rho \qquad (3.2)$$

for $t \in R_+$, as long as $\|h(t,\omega)\|$, λ and $\|f(t,\omega)\|_{cg}$ are small enough. For detail proof refer to [11, 16]. The method we use is Picard type successive approximation. We use a modified version of [16] to deal with systems and applied the results

132

to write a discrete version of the successive approximation.

We apply the method in the following example to obtain the numerical solution of a system of random Volterra integral equations.

Example 3.1: Let

$$x_1(t,\omega) = a_1(t,\omega) + \frac{1}{4} \int_0^t \tau e^{-\tau}[x_2(\tau,\omega) + \sin x_1(\tau,\omega)]d\tau,$$

$$x_2(t,\omega) = a_2(t,\omega) + \frac{1}{4} \int_0^t e^{\tau-t}[x_1(\tau,\omega) + \cos x_2(\tau,\omega)]d\tau,$$

(3.3)

where

$$a_1(t,\omega) = t - \frac{1}{8}(1+r_2(\omega))t^2 e^{-t} + \frac{1}{4}e^{-t}t\cos(t+r_1(\omega))$$

$$- \frac{1}{4}e^{-t}\sin(t+r_1(\omega)) + \frac{1}{4}e^{-t}\sin r_1(\omega) + r_1(\omega),$$

$$a_2(t,\omega) = \frac{5}{4} - \frac{1}{4}(t+r_1(\omega)) - \frac{1}{4}(1-r_1(\omega))e^{-t} - \frac{1}{4}\cos(1+r_2(\omega))$$

$$+ \frac{1}{4}\cos(1+r_2(\omega))e^{-t} + r_2(\omega).$$

By discretization (3.3) can be written as

$$x_1^{k+1}(\tau_n,\omega) = a_1(\tau_n,\omega) + \frac{1}{4}\sum_{i=0}^{n-1}\tau_i e^{-\tau_n}[x_2^k(\tau_i,\omega) + \sin x_1^k(\tau_i,\omega)] \times$$

$$(\tau_{i+1}-\tau_i)$$

(3.4)

$$x_2^{k+1}(\tau_n,\omega) = a_2(\tau_n,\omega) + \frac{1}{4}\sum_{i=0}^{n-1}e^{\tau_i-\tau_n}[x_1^k(\tau_i,\omega) + \cos x_2^k(\tau_i,\omega)] \times$$

$$(\tau_{i+1}-\tau_i).$$

$$k = 0,1,2,\ldots$$

Theoretical solution of (3.3) is $(t_1+r_1(\omega), 1+r_2(\omega))$. (3.4) can be simulated until $|x_i^{k+1}(t,\omega) - x_i^k(t,\omega)| < \epsilon$, $i = 1,2$.

For our numerical experiment we have assumed as follows:

Example 3.1 (a): $r_1(\omega), r_2(\omega) \in N(0,.002^2)$.

Example 3.1 (b): $r_1(\omega), r_2(\omega) \in N(0,.02^2)$.

Example 3.1 (c): $r_1(\omega), r_2(\omega) \in N(0,.2^2)$.

133

Example 3.1 (d): $r_1(\omega), r_2(\omega) \in U(-.001, .001)$

Example 3.1 (e): $r_1(\omega), r_2(\omega) \in U(-.01, .01)$

Example 3.1 (f): $r_1(\omega), r_2(\omega) \in U(-.1, .1)$.

We have presented our simulation results in Tables 3.1-3.6 and Figures 3.1-3.6. These results are based on 30 samples and each iteration was repeated until $|x_i^{k+1}(t, \omega) - x_i^k(t, \omega)| < .001$, $i = 1, 2$. In Figures 3.1 - 3.6, $\circ\circ\circ\circ$ and $\cdots\cdots$ denote respectively $E(x_1(t, \omega))$ and $E(x_2(t, \omega))$ and $\cdots\cdots$ and $\underline{\quad\quad}$ denote respectively $V(x_1(t, \omega))$ and $V(x_2(t, \omega))$.

4. DISCUSSION

The foregoing techniques demonstrate the usefulness and simplicity in applying Newton's and successive approximation methods to a system of random Volterra integral equations. Our numerical results show that mean of the sample solutions converge to the mean of the theoretical solution when variance decreases.

Other statistical parameters one can look at is risk functionals, confinement probability, skewness, kurtosis, correlation, etc. of random solutions. Also distribution of the numerical solution at different time units may be of interest.

134

Table 3.1

$$r_1(\omega), r_2(\omega) \in N(0, 0.002^2)$$

t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	.009867	.000030	1.009701	.000002
10	.097103	.000031	1.001172	.000004
20	.197879	.000032	1.001358	.000005
30	.300248	.000029	1.001692	.000005
40	.404133	.000029	1.001025	.000004
50	.508920	.000027	1.007050	.000004
60	.615807	.000033	1.007915	.000005
70	.724170	.000029	1.007578	.000004
80	.833203	.000037	1.007702	.000006
90	.943027	.000028	1.007992	.000005
100	1.052344	.000031	1.008623	.000004

Table 3.2

$$r_1(\omega), r_2(\omega) \in N(0, 0.02^2)$$

t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	.008556	.000357	1.018219	.000181
10	.098695	.000656	.999261	.000364
20	.197177	.000569	1.000663	.000513
30	.300128	.000444	1.003236	.000471
40	.404506	.000341	.995288	.000448
50	.509057	.000333	1.004797	.000441
60	.615514	.000486	1.014265	.000542
70	.725338	.000466	1.009957	.000383
80	.832995	.000515	1.008462	.000647
90	.942605	.000335	1.006821	.000461
100	1.044258	.000323	1.006896	.000408

Table 3.3

	$r_1(\omega), r_2(\omega) \in N(0, 0.2^2)$			
t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	-.004550	.046929	1.012795	.009134
10	.114598	.060710	.980120	.036398
20	.190134	.050747	.993800	.051145
30	.298866	.040816	1.018676	.046974
40	.408112	.030350	.937967	.044455
50	.510469	.032418	.981624	.043917
60	.612422	.043692	1.077264	.054399
70	.736662	.046105	1.032924	.038294
80	.830384	.043630	1.015674	.064831
90	.937779	.032423	.994332	.045982
100	.962477	.045704	.988586	.047095

Table 3.4

	$r_1(\omega), r_2(\omega) \in U(-0.002, 0.002)$			
t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	.010000	.000024	1.009160	.000000
10	.096964	.000023	1.001347	.000000
20	.197954	.000024	1.001442	.000000
30	.300235	.000024	1.001595	.000000
40	.404107	.000025	1.001491	.000000
50	.508935	.000025	1.007220	.000000
60	.615796	.000027	1.007402	.000000
70	.724064	.000026	1.007389	.000000
80	.833179	.000028	1.007598	.000001
90	.943050	.000027	1.008064	.000000
100	1.052986	.000027	1.008703	.000009

Table 3.5

	$r_1(\omega), r_2(\omega) \in U(-.02,.02)$			
t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	.009886	.000067	1.002753	.000018
10	.097296	.000070	1.001005	.000032
20	.197930	.000068	1.001496	.000040
30	.299994	.000065	1.002268	.000038
40	.404240	.000057	.999945	.000038
50	.509209	.000047	1.006492	.000041
60	.615406	.000063	1.009133	.000044
70	.724278	.000061	1.008070	.000035
80	.832759	.000075	1.007418	.000052
90	.942839	.000056	1.007543	.000039
100	1.050695	.000065	1.007700	.000030

Table 3.6

|	$r_1(\omega), r_2(\omega) \in U(-.2,.2)$			
t	$E(x_1)$	$V(x_1)$	$E(x_2)$	$V(x_2)$
1	.008749	.003090	.951889	.003010
10	.100586	.004457	.997168	.003305
20	.197686	.003644	1.002044	.004044
30	.297588	.003728	1.008994	.003835
40	.405573	.002929	.984492	.003804
50	.511933	.002624	.999228	.004051
60	.611481	.003274	1.026460	.004380
70	.726380	.004078	1.014866	.003524
80	.828506	.003568	1.005627	.005222
90	.940660	.003556	1.002309	.003895
100	1.027694	.003709	.997618	.004123
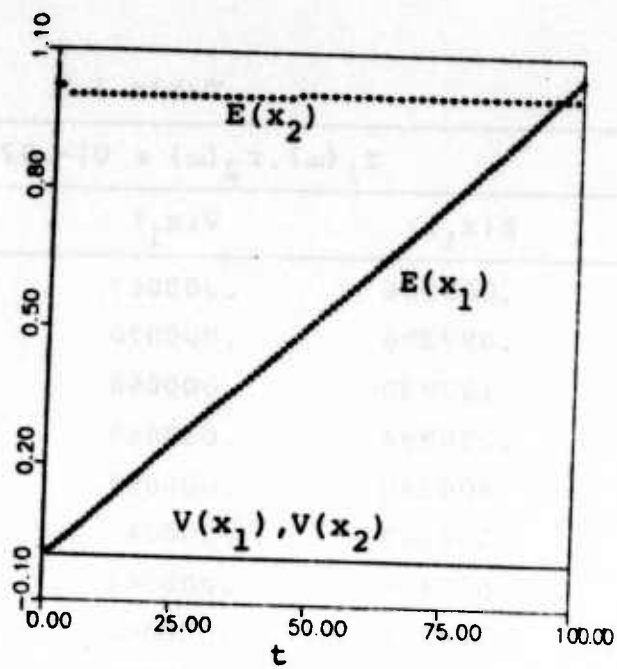
Fig. 3.1: Example 3.1 (a).

Fig. 3.2: Example 3.1 (b).

138
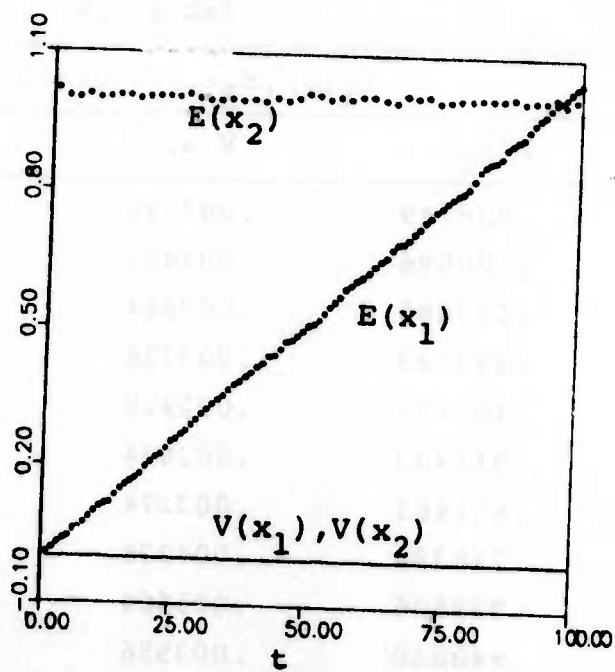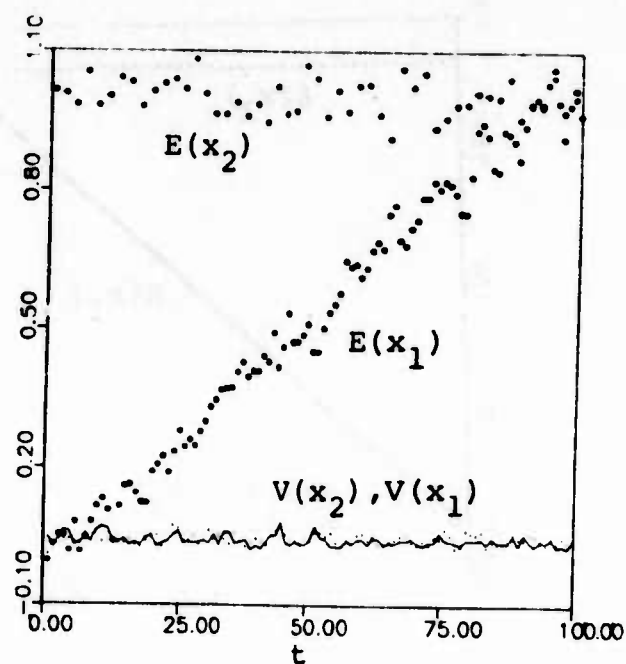
Fig. 3.3: Example 3.2 (c).

Fig. 3.4: Example 3.2 (d).

139
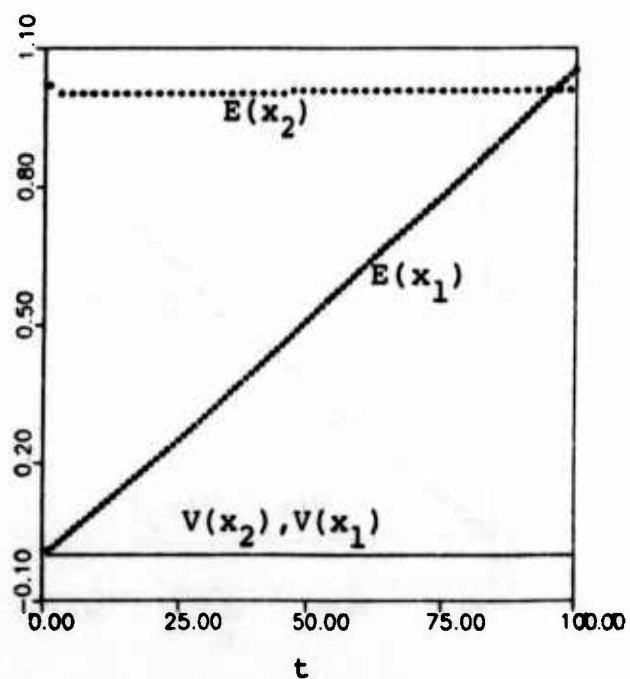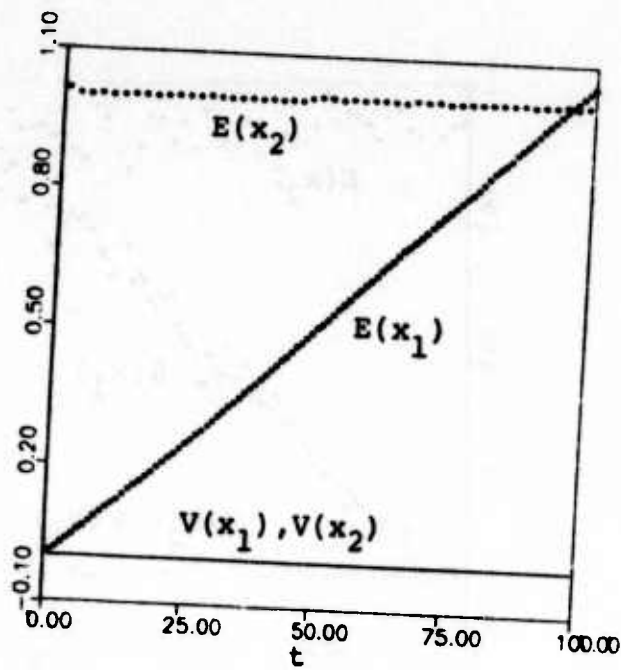
Fig. 3.5: Example 3.1 (e).

Fig. 3.6: Example 3.1 (f).

140
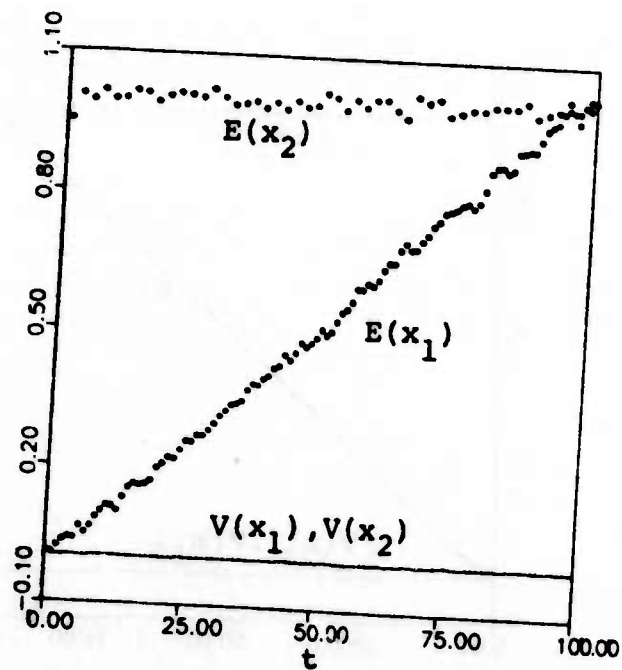
REFERENCES

1. Baker, C. T. H. The numerical treatment of integral equations. Clarendon Press, Oxford, 1977.

2. Becus, G. A. Successive approximations solution of a class of random equations. Approximate Solution of Random Equations (ed. A. T. Bharucha-Reid) North Holland, NY, 1979.

3. Bharucha-Reid, A. T. Random Integral Equations, Academic Press, NY, 1972.

4. Bharucha-Reid, A. T. and Christensen, M. J. Approximate solution of random integral equations: General methods. Math. Comp. Simulation 26 (1984) 321-328.

5. Bharucha-Reid, A. T. and Kannan, R. Newton's method for random operator equations, Nonlinear Anal. Theory Methods Appl. 4 (1980) 231-240.

6. Christensen, M. J. and Bharucha-Reid, A. T. Numerical solution of random integral equation I,II. J. Integ. Equation 3 (1981) 217-229, 333-344.

7. Hans, O. Random operator equations. Proc. Fourth Berkeley Symp. Math. Stat. Prob. 2 (1960) 185-202 (1961).

8. Lax, M. D. Method moments approximate solutions of random linear integral equations. J. Math. Analy. Appl. 58 (1977) 46-55.

9. Lax, M. D. Approximate solution of random differential and integral equations. Applied Stochastic Processes (ed. G. Adomian). Academic Press, NY 1980.

10. Lax, M. D. Solving random linear Volterra integral equations using the method of moments. J. Integral. Equation 3 (1981) 357-363.

11. Medhin, N. and Sambandham, M. Numerical solution of random Volterra integral equation I. Successive approximation method. To be published.

12. Medhin, N. and Sambandham, M. Numerical solution of random Volterra integral equation II. Newton's method. To be published.

13. Nashed, M. Z. and Salehi, H. Measurability of generalized inverses of random linear operators. SIAM J. Appl. Math. 25 (1973) 681-692.

141

14. Sambandham, M. Numerical solution of random linear Volterra integral equation. Trans. Third Army Conference on Appl. Math. Comp. ARO Report 86-1 (1986) 841-855.

15. Sambandham, M., Christensen, M. J. and Bharucha-Reid, A. T. Numerical solution of random integral equation III. Stoch. Analy. Appl. 3 (1985) 467-484.

16. Tsokes, C. P. and Padgett, W. J. Random integral equations with applications to life sciences and engineering. Academic Press, NY, 1974.

17. Wonk, A. A Course of Applied Functional Analysis. John Wiley and Sons, NY, 1979.

APPROXIMATE METHODS FOR
STRUCTURAL RELIABILITY

Mircea Grigoriu and Arnold Buss
Department of Structural Reliability
Cornell University
Ithaca, N.Y. 14850

ABSTRACT. Structural reliability can be defined as the probability that a random process $\underline{X}(t)$ remains in a domain of safe structural performance during a reference period. The process can model material properties, environmental loads, or outputs of mechanical systems subject to random inputs. In the time-independent case $\underline{X}(t) = \underline{X}$ and reliability can be approximated from a scalar quantity, the reliability index. This approximation is evaluated for Gaussian and non-Gaussian vectors and safe domains of various shapes. In the time-dependent case reliability is approximated by the mean rate at which $\underline{X}(t)$ crosses out of the safe domain. When $\underline{X}(t)$ is non-Gaussian it can be approximated by a memoryless transformation of a Gaussian process, called a translation process. Translation processes have identical marginal distributions and similar crossing properties as the original process $\underline{X}(t)$. The approximate method of reliability analysis based on translation processes is applied to several non-Gaussian processes and safe domains.

I. INTRODUCTION. Consider a structural component of strength X_2 subject to an uncertain axial load of X_1. The reliability of the component is equal to the probability $P_S = P\{X_1 \leq X_2\}$ and can be determined from the probability content of the safe domain $D = \{(x_1,x_2): x_1 - x_2 \leq 0\}$. The component fails with the probability $P_F = 1-P_S$. In general reliability problems the vector $\underline{X} = (X_1,X_2)$ of uncertain parameters is n-dimensional and can be time-dependent. The safe domain $D = \{\underline{x}: g(\underline{x}) \leq 0\}$ is a region in R^n and the boundary $\partial D = \{\underline{x}: g(\underline{x})=0\}$ of D is usually refered to as the limit state.

The objective of this paper is to examine and evaluate approximate methods for calculating the reliability P_S in general reliability problems involving time-invariant and time-dependent random vectors $\underline{X}$. It is assumed in the analysis of time-dependent problems that failure occurs at the first excursion out of the safe domain. Thus, failures due to changes in material characteristics under constant stress or damage accumulation caused by repeated loads are not investigated.

II. TIME INVARIANT PROBLEMS.

First let $\underline{X}$ be a Gaussian vector. Without loss of generality it is assumed that $\underline{X}$ has independent components of zero mean and unit variance. The reliability is

$$P_S = \int_D \varphi(\underline{x}) dx \qquad (1)$$

in which $\varphi(\underline{x}) = (2\pi)^{-n/2} \exp\{-\frac{1}{2}\sum_i x_i^2\}$. The determination of P_S in Eq. (1) usually requires numerical integration and, as a result, is impractical when $n \geq 3$. However, the reliability can be obtained in closed form in two cases corresponding to safe domains bounded by hyperplanes and hyperspheres. It is $\Phi(\beta)$ for linear limit states at a distance β from the origin and $F_{\chi_n}(\beta)$ for spherical limit states with radius β centered at the origin. Φ and F_{χ_n} denote, respectively, the standard Gaussian distribution and the chi distribution with n degrees of freedom.

The reliabilities corresponding to linear and spherical limit states can be employed to develop probability bounds for general domains. Consider, for example, the safe domain D in Figure 1 and let $\underline{x}_0$ be the point of ∂D closest to the origin, assumed to be unique. This point is usually refered to as the β-point and is at a distance $\beta = |\underline{x}_0|$ from the origin. If D is convex and the function $g(\underline{x})$ can be differentiated, the reliability P_S is bounded by

$$F_{\chi_n}(\beta) \leq P_S \leq \Phi(\beta) \qquad (2)$$

The bounds F_{χ_n} and $\Phi(\beta)$ on P_S are attractively simple. However, they become less informative as the dimension of the space n increases, as shown in Figure 2. Therefore, other methods are needed to approximate P_S.

The most accurate approximation available for P_S is based on an asymptotic evaluation of the integral in Eq. (1) as the distance β to the β-point increases indefinitely. It can be shown that P_F can be approximated asymptotically as $\beta \to \infty$ by [1]

$$P_{F,a} = \Phi(-\beta) \prod_{i=1}^{n-1} (1-k_i)^{-\frac{1}{2}} \qquad (3)$$

in which k_i are the principal curvatures of the β-point assumed

144

to satisfy the conditions $1>k_1 \geq k_2 \geq \ldots \geq k_{n-1}$ when $\beta = 1$. The same asymptotic result can be obtained for P_F if the limit state is approximated in a small neighborhood of the β-point by a quadratic tangent to the limit state at $\underline{x}_0$. Equation 3 can be generalized to the case where the limit state has finitely many β-points. In this case P_F is asymptotically equal to a sum of terms as in Eq. (3) corresponding to each β-point.

Figure 3 shows the asymptotic probabilities of failure in Eq. (3)., $P_{F,a}$, and the actual failure probability, P_F for various elliptical domains with limit states $(x_1/a)^2 + (x_2)^2 = \beta^2$. As expected, the ratio of these probabilities approaches unity as β increases [1]. It is approximately one for large values of a because the problem becomes one-dimensional, in which case $P_{F,a}$ is equal to P_F.

An alternative method for calculating P_F can be based on a simulation approach. Brute force simulation is impractical because P_F is usually smaller than 10^{-3}. However, an efficient simulation method can be developed to estimate P_F. Assume for simplicity that the function g specifying the limit state is the quadratic form

$$Z = g(\underline{X}) = \underline{X}^T \underline{a} \, \underline{X} + \underline{b}^T \underline{X} + \underline{c} \tag{4}$$

The Gaussian vector $\underline{X}$ can be expressed as $\underline{X} = \underline{\Lambda}R$ in which $\underline{\Lambda}$ is a random vector uniformly distributed on the unit sphere in R^n and R is a chi random variable with n degrees of freedom. Conditional on $\underline{\Lambda} = \underline{\lambda}$, the quadratic form is

$$Z|\underline{\Lambda} = \underline{\lambda}^T = (\underline{\lambda}^T \underline{a} \, \underline{\lambda}) R^2 + (\underline{b}^T \underline{\lambda}) R + c \tag{5}$$

The conditional probability of failure is

$$P_F(\underline{\lambda}) = P\{ Z > 0 \mid \underline{\Lambda} = \underline{\lambda} \} \tag{6}$$

and can be calculated by

$$P_F(\underline{\lambda}) = F_{\chi_n}(r_1(\underline{\lambda})) + 1 - F_{\chi_n}(r_2(\underline{\lambda})) \tag{7}$$

when, e.g., the roots $r_k(\underline{\lambda})$, k=1,2, of $(\underline{\lambda}^T \underline{a} \, \underline{\lambda})r^2 + (\underline{b}^T \underline{\lambda}) r + c = 0$ are positive and $r_1(\underline{\lambda}) \leq r_2(\underline{\lambda})$. An estimator of the probability of failure P_F is

145

$$\hat{P}_{F'} = \frac{1}{N} \sum_{j=1}^{N} P_F[\underline{\lambda}^{(j)}] \tag{8}$$

in which $\underline{\lambda}^{(j)}$ are samples of $\underline{\Lambda}$ and N denotes the number of these samples.

Tables 1 and 2 show values of $\hat{P}_F$ obtained in two samples of size $N=100$ for safe domains characterized by two quadratic forms:

$$Z = X_1^2 + X_2^2 - X_1 X_2 - (f_y/f_0)^2 \tag{9}$$

and

$$Z = \sum_{i=1}^{n} (X_i + \mu_i)^2 - \beta^2 \tag{10}$$

The first form corresponds to the von Mises strength criterion while the second one corresponds to a non-central chi-square variable with non-centrality parameter $\mu = \sum_{i=1}^{n} \mu_i^2$. From the tables, $\hat{P}_F$ satisfactorily approximates P_F in both cases.

When the components of $\underline{X}$ are independent but do not follow Gaussian distributions, the techniques discussed in this section can still be applied if the $\underline{x}$-space is mapped into a new space according to the transformation

$$U_i = \Phi^{-1} \circ F_i (X_i) \tag{11}$$

in which F_i is the distribution of X_i. The variables U_i are independent and follow standard Gaussian distributions. Figure 4, from reference [1] shows exact values of P_F and asymptotic approximations of the probability of a safe domain defined by the condition $\sum_{i=1}^{n} X_i - n - \alpha\sqrt{n}$ where the variables X_i are independent identically distributed exponential random variables with unit mean. The asymptotic approximation is also satisfactory in this case.

Techniques are also available to map vectors of dependent non-Gaussian variables into Gaussian vectors with independent components. Following such transformations one can directly apply any of the methods developed for Gaussian vectors.

146

III.TIME DEPENDENT PROBLEMS. Consider a process X(t) defined by

$$X(t) = \int_0^t h(t-\tau)g(\underline{Y}(\tau))d\tau \qquad (12)$$

in which $\underline{Y}(t)$ is a stationary Gaussian vector process in R^n with independent components. The process X(t) can be thought of as the response of a linear system with unit impulse response function h to the input $g(\underline{Y}(t))$. On the other hand, if h in Eq. (12) is δ (the Dirac delta function), then $X(t) = g(\underline{Y}(t))$ is a memoryless transformation of the Gaussian process $\underline{Y}(t)$. As in the time-invariant case, the safety condition requires that X(t) be smaller that an allowable threshold during a time period τ. The reliability $P_S(\tau)$ can be approximated by

$$P_S(\tau) = P\{X(0) \leq 0\} \exp\{-\tau \ \nu(0)\} \qquad (13)$$

in which $\nu(x)$ is the mean rate at which X(t) crosses a level x from below. Note that the reliability $P_S(\tau)$ depends only on the mean upcrossing rate $\nu(x)$ and the marginal distribution of X(t).

Three special cases in which $h = \delta$ are now considered. First, let $\underline{Y}(t)$ be a univariate Gaussian process and g be the identity function. Then X(t) coincides with $\underline{Y}(t)$ and, according to the Rice formula [6],

$$\nu(x) = [\dot{\sigma}_X/2\pi\sigma_X]\exp\{-[(x-m_X)/\sigma_X]^2/2\} \qquad (14)$$

in which m_X and σ_X^2 are the mean and variance of X(t) and $\dot{\sigma}_X^2$ is the variance of $\dot{X}(t)$.

Second, let $\underline{Y}(t)$ be a univariate Gaussian process with zero mean and unit variance and $g = F_X^{-1} \circ \Phi$, where F_X is any continuous distribution. The process X(t) in this case is called a translation process, and the marginal distribution of X(t) at any time t is F_X. If $\underline{Y}(t)$ is stationary and differentiable, so is X(t). The standard deviations of the derivatives of these processes are related by $\dot{\sigma}_Y = \eta\dot{\sigma}_X/\sigma_X$, in which $\eta = \{E[g'(Y(t))^2]\}^{-\frac{1}{2}}$. The constant η is generally close to unity [3]. The mean upcrossing rate of X(t) can be obtained from Eq. (14) and is

$$\nu(x) = [\dot\sigma_Y/2\pi]\exp\{-[g^{-1}(x)]^2/2\}$$

$$= [\eta\dot\sigma_X/2\pi\sigma_X]\exp\{-[g^{-1}(x)]^2/2\} . \tag{15}$$

In this case $\nu(x)$ depends only on F_X and $\dot\sigma_X$.

Consider any process that can be partially specified by its marginal distribution and covariance function. Such a process can be approximated by a translation process having the correct marginal distribution and covariance function. The mean upcrossing rate of this process can be approximated by Eq. (15). In many cases the translation approximation is conservative, meaning that it overestimates the actual value of $\nu(x)$.

A third type of memoryless transformation of $\underline{Y}(t)$ is

$$X(t) = \underline{Y}(t)^T \underline{a} \, \underline{Y}(t) + \underline{b}^T \underline{Y}(t) + c \tag{16}$$

Two special cases of this quadratic form are examined. Consider first the time-dependent form of the von Mises criterion that requires $X(t) = Y_1^2(t) + Y_2^2(t) - Y_1(t)Y_2(t)$ be smaller that a limit value x, where $\underline{Y}(t) = (Y_1(t), Y_2(t))^T$ is a bivariate Gaussian process with independent components having mean zero and unit variance. The mean upcrossing rate of $X(t)$ is [2]

$$\nu(x) = [\dot\sigma/\pi^{3/2}](2x/3)^{1/2} e^{-2x} \int_0^\pi (2 + \cos u)^{1/2} \exp\{x \cos u\} \, du \tag{17}$$

From the density of $X(t)$ a translation approximation may be obtained. Table 3 compares mean upcrossing rates ν_T for the translation approximation with the exact mean upcrossing rate in Eq. (17). The translation approximation is seen to be conservative for moderate to large thresholds. As another example, let $X(t) = \sum_{j=1}^{n} Y_j(t)^2$ be a chi-square process with n degrees of freedom in which $\underline{Y}(t) = (Y_1(t),...,Y_n(t))$ is a Gaussian vector whose components are independent identical univariate Gaussian processes with zero mean and unit variance. The mean upcrossing rate is

$$\nu(x) = \dot\sigma_X [x/2\pi n]^{1/2} f(x) \tag{18}$$

where $f(x) = (x/2)^{n/2-1} e^{-x/2}/2\Gamma(n/2)$. Table 4 gives ratios of ν to the mean upcrossing rate ν_T based on the translation

148

approximation. The translation approximation is conservative for large thresholds, and this observation is correct asymptotically as x→∞ [2].

In addition to the translation approximation, one can develop approximations analogous to the asymptotic approximations given in section II. These will not be pursued here, and the interested reader is refered to the relevant literature [1,5].

Now consider the more general case in Eq. (12) in which h is a function which vanishes for t<0. Since g is nonlinear, $g(\underline{Y}(t))$ is non-Gaussian and so is X(t). Direct determination of the probability law of X(t) from Eq. (12) is usually impractical. An approximation is preferred to estimate mean crossing rates of X(t). The approximation can be based on the simplified representation $\underline{m}+\underline{s}(t)\underline{Z}$ of $\underline{Y}(t)$ in which $\underline{Z}$ is a vector of independent standard normal random variables, $\underline{m}$ is a mean vector, and

$$
\underline{s}(t) = \begin{bmatrix} \underline{s}_1(t) & & 0 \\ & \ddots & \\ 0 & & \underline{s}_n(t) \end{bmatrix}
\tag{19}
$$

where $\underline{s}_j(t)$, j=1,...,n are vectors of sines and cosines with appropriate coefficients. In the special case in which $g(y) = y^2$ the process X(t) is the following quadratic form

$$
X(t) = \underline{Z}^T\underline{a}(t)\,\underline{Z} + \underline{b}^T(t)\,\underline{Z} + c(t)
\tag{20}
$$

In contrast to the form in Eq. (16), the coefficients of this quadratic form depend on time and they operate on the random vector $\underline{Z}$, while in Eq. (16) the coefficients are constant and operate on the random vector process $\underline{Y}(t)$. One can derive the characteristic functions of X(t) and $(X(t),\dot{X}(t))$, which can be used to find the marginal density and mean upcrossing rate of X(t) [2].

The representation in Eq. (20) can be applied to estimate the response of a structure to wind loads that are proportional to the square of the wind speed Y(t). It is assumed that the structure is modeled by a simple oscillator with response function $h(s) = \exp\{-\zeta\omega_0 s\}\sin(\omega_d s)/\omega_d$, $s\geq0$, where $\omega_d = \omega_0[1-\zeta^2]^{1/2}$. As a numerical example, Y(t) is taken to have mean 6.57 and discrete spectrum given in table 5. The system parameters are $\zeta = 0.0$, $\omega_0 = 6.28$. The marginal density of the response X(t)

149

is illustrated in figure 5. Mean upcrossing rates are found by the translation approximation and by the joint characteristic function method. A comparison of these in in table 6. As in the other cases considered, it is seen that the translation approximation is conservative with respect to the exact mean upcrossing rate.

IV.CONCLUSIONS. Approximate methods have been examined for the reliability analysis of time-independent and time-dependent problems. Probability bounds and asymptotic approximations have been developed for the estimation of the reliability of time-invariant structural problems. On the other hand, the reliability estimates for the time-dependent problems have been based on mean crossing rates out of a domain of safety and on translation approximations of these crossing rates. The translation approximations have been found to be conservative in the cases examined.

REFERENCES

[1] Breitung, K., Asymptotic Approximations for the Crossing Rates of Stationary Gaussian Vector Processes, Report LUTFD2, Department of Mathematical Statistics, Lund University, Sweden, January 1984.

[2] Buss, A. and Grigoriu, M., Extremes of Non-Gaussian Processes, submitted to publication.

[3] Grigoriu, M., Crossings of Non-Gaussian Translation Processes, Journal of Engineering Mechanics, ASCE, Vol. 110, No. EM4, April 1984, pp. 610-620.

[4] Grigoriu, M., Response of Linear Systems to Quadratic Excitations, Journal of Engineering Mechanics, Vol. 112, No. 6, June 1986, pp. 523-535.

[5] Lindgren, G., Extreme Values and Crossings for the χ^2-Process and Other Functions of Multidimensional Gaussian Processes, with Reliability Applications, Advances in Applied Probability, Vol. 12, 1980, pp. 746-774.

[6] Rice, S. O., Mathematical Analysis of Random Noise, Bell System Technical Journal, Vol. 24, 1945, pp.46-156.

[7] Rubinstein, R. Y., Simulation and the Monte Carlo Method, John Wiley, New York, 1981.

Table 1. Exact and Estimated Values of P_F in Eq. (9)

f_y/f_0	P_F	$\hat{P}_F$	
		Sample 1	Sample 2
1	5.75×10^{-1}	5.81×10^{-1}	5.68×10^{-1}
2	1.37×10^{-1}	1.41×10^{-1}	1.34×10^{-1}
3	1.83×10^{-2}	1.85×10^{-2}	1.77×10^{-2}
4	1.37×10^{-3}	134×10^{-3}	1.30×10^{-3}

Table 2 Exact and Estimated Values of P_F in Eq. (10)

$\sqrt{\mu/\beta}$	P_F	$\hat{P}_F$	
		Sample 1	Sample 2
0.0	1.61×10^{-5}	1.61×10^{-5}	1.61×10^{-5}
0.4	3.57×10^{-3}	7.02×10^{-3}	6.84×10^{-3}
0.8	2.19×10^{-1}	2.92×10^{-1}	2.64×10^{-1}
1.0	5.80×10^{-1}	6.21×10^{-1}	5.89×10^{-1}

Table 3. Exact and Approximate Mean Crossing Rates
for the von Mises Criterion

x	υ	υ/υ_T
1	2.49×10^{-1}	1.25
4	1.07×10^{-1}	0.96
9	1.98×10^{-2}	0.87
16	1.90×10^{-3}	0.83
25	9.42×10^{-5}	0.81

Table 4. Exact and Approximate Values of Mean Crossing Rates for the Chi Square Process

$\tilde{x}$	n = 1		n = 2		n = 5	
	ν	ν/ν_T	ν	ν/ν_T	ν	ν/ν_T
0	9.65×10^{-2}	1.08	7.34×10^{-2}	1.05	4.87×10^{-2}	1.03
3	1.16×10^{-2}	0.87	7.31×10^{-3}	0.88	3.56×10^{-3}	0.89
6	1.39×10^{-3}	0.84	4.81×10^{-4}	0.84	8.49×10^{-5}	0.84
9	1.66×10^{-4}	0.83	2.86×10^{-4}	0.81	1.46×10^{-6}	0.83

Table 5. Power Spectral Density of Wind Speed

ω_i	0.25	0.75	1.25	1.75	2.25	3.25	4.75	6.25	7.75	9.25
σ_i	1.26	0.34	0.20	0.13	0.11	0.14	0.09	0.07	0.05	0.05

Table 6. Exact and Approximate Values of Mean Crossing Rates for a Structure Subject to Wind Load

$\tilde{x}$	ν	ν/ν
0	6.15×10^{-1}	1.00
2	1.12×10^{-1}	0.81
4	7.47×10^{-3}	0.74
6	3.18×10^{-4}	0.70
8	1.02×10^{-5}	0.65
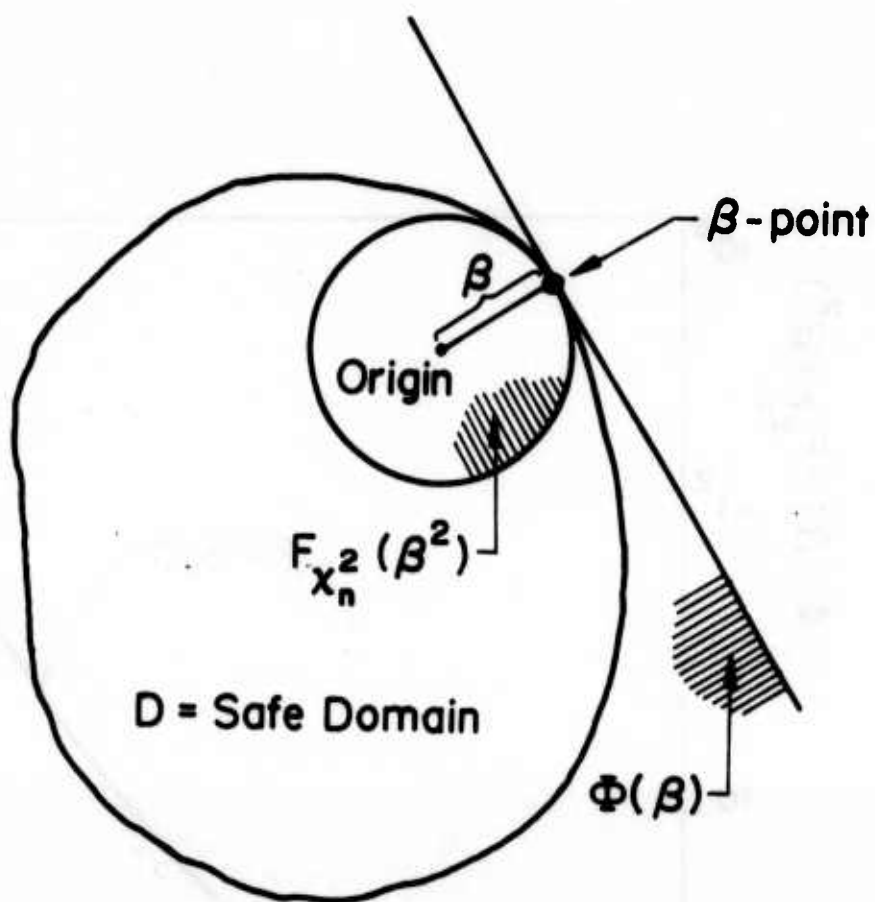
Figure 1. Bounds on Reliability.
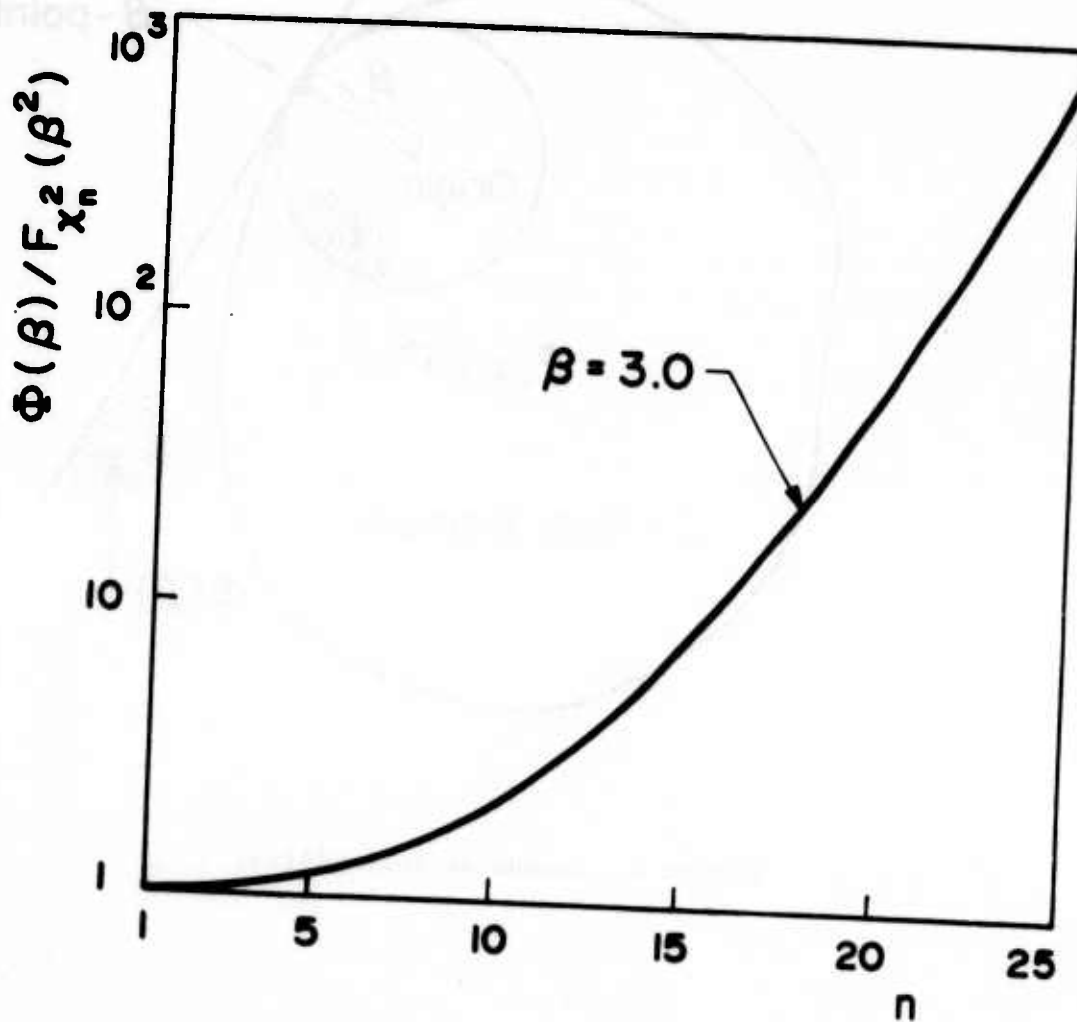
Figure 2. Ratio of Upper to Lower Bounds on Reliability.
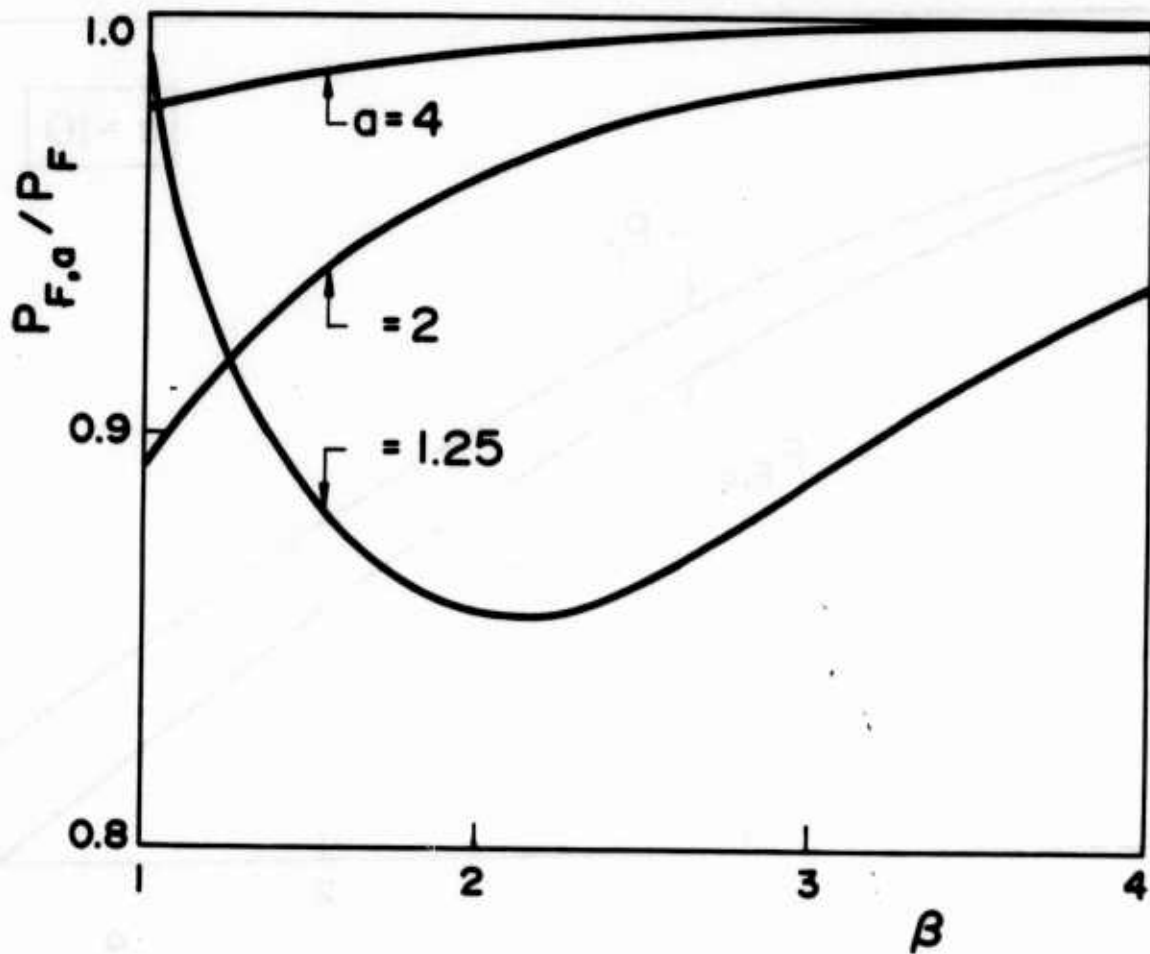
Figure 3. Ratios of Asymptotic to Exact Values
of the Probability of Failure for
Elliptical Safe Domains.

155
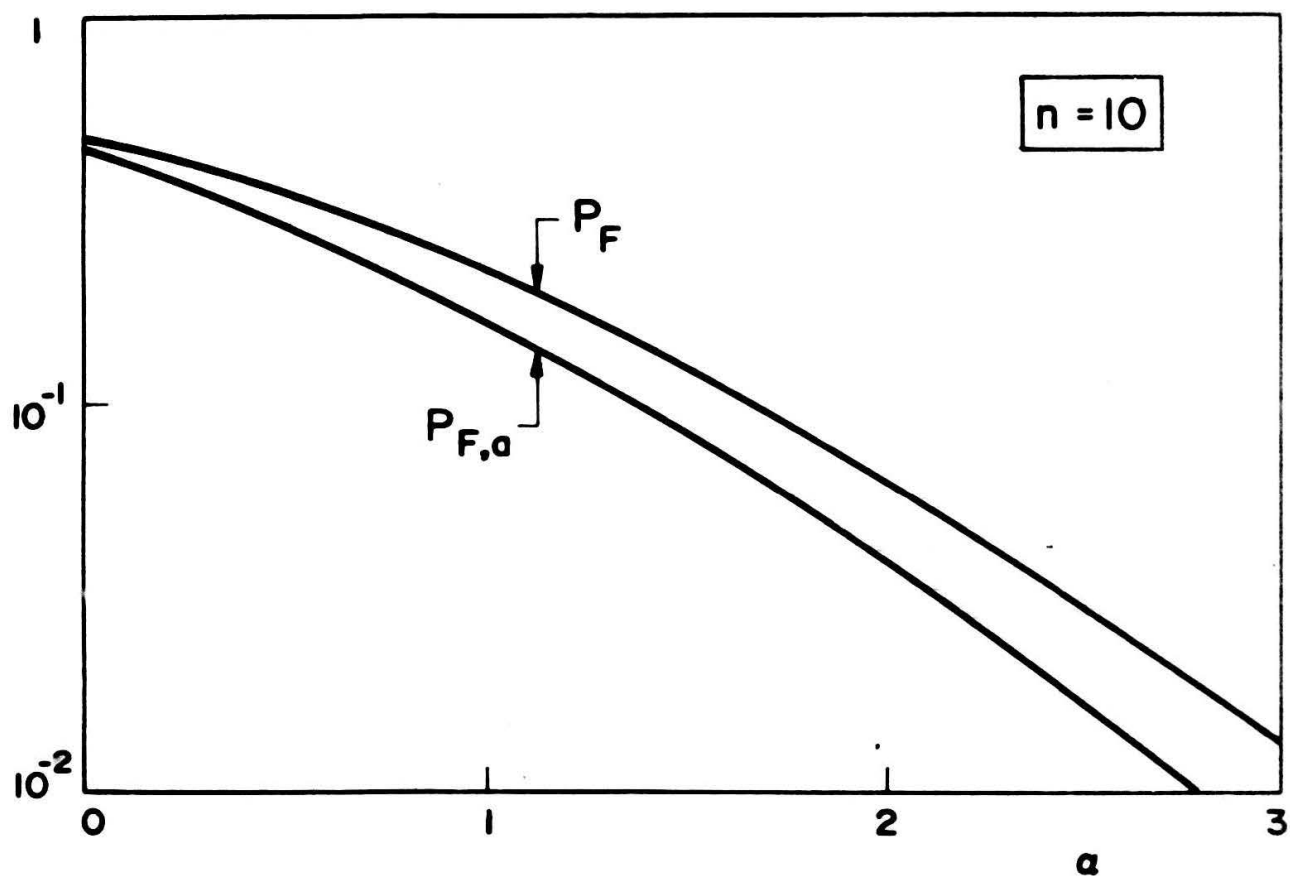
Figure 4. Exact and Asymptotic Values of the
Probability of Failure for Exponential
Random Variables.
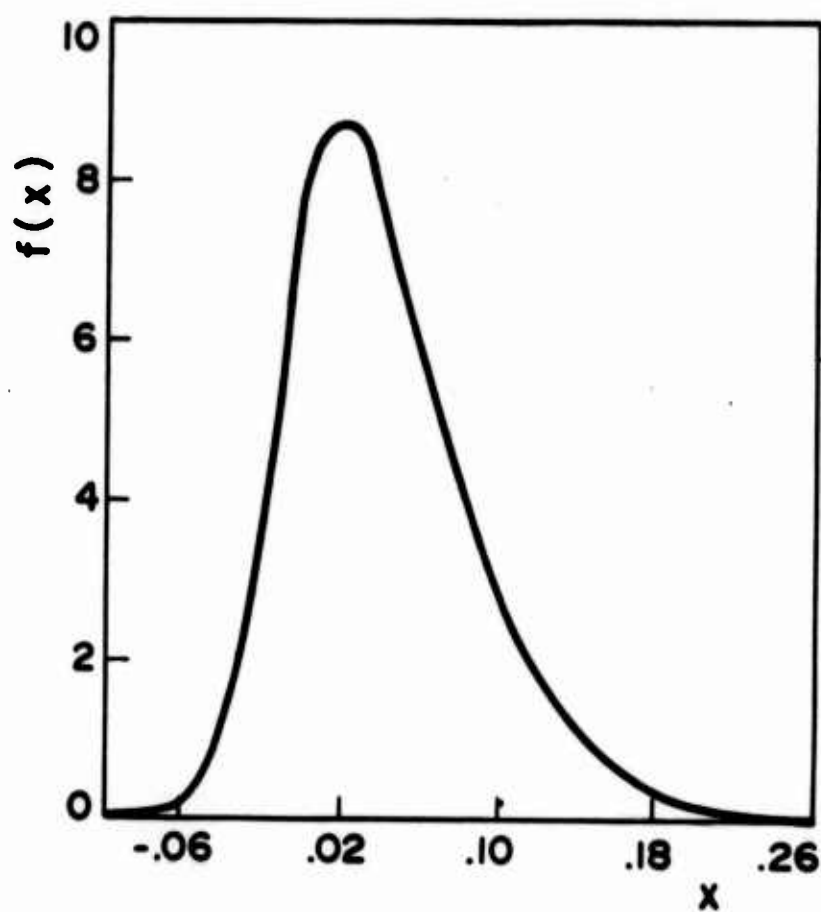
Figure 5. Marginal Density of the Response of a
Linear System to Wind Load.

Marc A. Bergen
Carnegie Mellon University
Pittsburg, Pennsylvania

§1. Introduction and Overview: An Operational Itô Calculus

In order to describe the method of random characteristics we begin by considering cases which lead to the classical second order Itô-theory. Let $\sigma: \mathbb{R}^m \to \mathbb{R}^m$ be of class C_b^1, and let $X = X_\sigma$ denote the vector field $X := \sigma \cdot \nabla$. As usual exp denotes the mapping $f(x) \mapsto f(\xi(1;x))$ where $\xi(\tau;x)$ satisfies $d\xi/d\tau = \sigma(\xi)$, $\xi(0) = x$. The Banach space on which $\exp(X)$ acts is $C(\mathbb{R}_*^m)$ with the supremum norm, $\mathbb{R}_*^m$ being the one point compactification of $\mathbb{R}^m$. It is clear that $\exp(tX)$, $t \in \mathbb{R}$, is the group generated by X. Observe that we are allowing time, t, to run backward and forward.

Let $\theta(t)$ be standard one-dimensional Brownian motion. The mgf of $\theta(t)$ is given by

$$\mathbb{E}\, e^{x\theta(t)} = e^{\frac{1}{2}x^2 t}; \quad x \in \mathbb{R}, \ t > 0. \tag{1}$$

We can ask about the validity of (1) if x were to be replaced by X above. Let us examine what each side of (1) would then mean. On the left-hand side $e^{X\theta(t)}$ would represent the random wave operator $f(x) \mapsto f(\xi(\theta(t);x))$, and so $\mathbb{E}\, e^{X\theta(t)}$ would be the operator

$$f(x) \mapsto \int_{\mathbb{R}} f(\xi(z;x))p(z,t)dz, \tag{2}$$

where p is the Gauss kernel $p(z,t) = 1/\sqrt{2\pi t}\, \exp(-z^2/2t)$. On the right-hand side we would have the semi-group operator $\exp(\frac{1}{2}tX^2)$, generated by $\frac{1}{2}X^2$. In fact with this interpretation (1) does remain valid when x is replaced with X.

159

The easiest way to see this is to use the fact that p is the fundamental solution of $\partial u/\partial t = \frac{1}{2} \partial^2 u/\partial z^2$. Thus, upon integrating by parts, one sees that

$$u(x,t) = \int_{\mathbb{R}} (e^{zX}f)(x)p(z,t)dz \qquad (3)$$

satisfies the evolution equation $du/dt = \frac{1}{2} X^2 u$, $u(0) = f$. It is interesting to interpret this representation (2) for the semi-group geometrically, in terms of the characteristics ζ. The domain of influence of a region $D \subset \mathbb{R}^m$, at any time $t > 0$, is that region spanned by the characteristic curves which pass through D at time zero. (Recall that "time" along the characteristics runs both backward and forward.)

Next let $\sigma_k: \mathbb{R}^m \to \mathbb{R}^m$, $1 \le k \le \ell$, be ℓ such maps of class C_b^1 and correspondingly let $X_k = X_{\sigma_k} = \sigma_k \cdot \nabla$. If $\theta(t) = (\theta_1(t),\ldots,\theta_\ell(t))$ is standard ℓ-dimensional Brownian motion then

$$\mathbb{E} \exp(<x,\theta(t)>) = \exp(\frac{1}{2} t \Sigma x_k^2); \quad x \in \mathbb{R}^\ell, \ t > 0. \qquad (4)$$

Now we can ask about replacing $x = (x_1,\ldots,x_\ell)$ with $X = (X_1,\ldots,X_\ell)$ in (4). At this point we already know what interpretation both sides of (4) would have. Here the random wave operator under consideration would be $\exp(<X,\theta(t)>)$, and the semi-group under consideration would be the one generated by $\frac{1}{2} \Sigma X_k^2$. This time the answer is YES (the replacement $x \to X$ in (4) is valid) if the X_k commute, but NO in

160

general.

Example I: $X_1 = x_2 \frac{\partial}{\partial x_1}$, $X_2 = \frac{\partial}{\partial x_2}$

$\mathbb{E} \exp(<X, \theta(t)>)$:

$$f(x) \mapsto \int_{\mathbb{R}^2} f(x_1 + z_1 x_2 + \frac{1}{2} z_1 z_2, \; x_2 + z_2) p(z_1, t) p(z_2, t) \, dz$$

$\exp(\frac{1}{2} t (X_1^2 + X_2^2))$:

$$f(x) \mapsto \mathbb{E} f(x_1 + x_2 \theta_1(t) + \int_0^t \theta_2(s) d\theta_1(s), \; x_2 + \theta_2(t)).$$

What is in fact true, however, is that for small t,
$\mathbb{E} \exp(<X, \theta(t)>)$ is very close to $\exp(\frac{1}{2} t \sum X_k^2)$, in the sense that

$$\prod_0^t \mathbb{E} \exp(<X, \theta(du)>) = \exp(\frac{1}{2} t \sum X_k^2). \tag{5}$$

If $T(t)$ denotes the operator $\mathbb{E} \exp(<X, \theta(t)>)$ then the product integral here indicates a Riemann-type strong limit

$$\prod_0^t T(du) = \text{st} - \lim_n \prod_i T(\Delta_{in}), \tag{6}$$

where $0 = t_{0n} < t_{1n} < \dots < t_{\nu_n n} = t$ forms a partition of $[0, t]$, $\Delta_{in} = t_{in} - t_{i-1n}$ and $\lim_n \max_i \Delta_{in} = 0$. We shall see that (6) follows from a version of Chernoff's product formula [8] which allows for variable step size. (See Pierre and Rihani [18].)

Actually for the case at hand where θ is Brownian motion, one can interchange the expectation and product integral in (5),

$$\prod_0^t \mathbb{E}\exp(<X,\theta(du)>) = \mathbb{E}\prod_0^t \exp(<X,d\theta(u)>), \qquad (7)$$

thereby unveiling sample path behavior. At this point to arrive at (7) we are exploiting the stationary independent increment property of Brownian motion. The product integral on the right in (7) is now a sample path limit, indicated by

$$\prod_0^t \exp(<X,d\theta(u)>) = st-\lim_n \prod_i \exp(<X,\Delta_{in}\theta>), \qquad (8)$$

where the partitions $0 = t_{0n} < t_{1n} < \ldots < t_{\nu_n n} = t$ are as above, and $\Delta_{in}\theta = \theta(t_{in}) - \theta(t_{i-1n})$. This resembles McKean's injection set-up for Lie groups [13, §4.8]. In fact if the vector fields X_k, $1 \le k \le \ell$, belong to a finite dimensional Lie algebra (e.g., $\sigma_k(x) = M_k x$ for $m \times m$ matrices M_k), then this fits precisely into McKean's set-up. The st-lim in (8) indicates a strong limit in the bounded linear operator sense, but we also need to specify the probabilistic mode of convergence. In what sense is the sequence of random variables

$$\left\| \prod_i \exp(<X,\Delta_{in}\theta>)f - \prod_0^t \exp(<X,d\theta(u)>)f \right\|$$

converging to zero, for each $f \in C(\mathbb{R}_*^m)$?

There is a nice concise way of describing the operator $\prod_i \exp(<X,\Delta_{in}\theta>)$, appearing on the right-hand side of (8). Let $\psi^{(n)}$ denote the piecewise-constant function $\psi^{(n)}(\tau) = \Delta_{in}\theta/\Delta_{in}$, $t_{i-1n} \le \tau \le t_{in}$; and let $\xi^{(n)}(\tau;x)$ denote the solution of $d\xi^{(n)}/d\tau = \Sigma\psi_k^{(n)}(t-\tau)\sigma_k(\xi^{(n)})$, $\xi^{(n)}(0) = x$. By a simple time scale one sees that

$$\prod_i \exp(\langle X, \Delta_{in}\theta\rangle): \quad f(x) \mapsto f(\xi^{(n)}(t;x)). \tag{9}$$

Observe that applying the operators in the order

$$\exp(\langle X, \Delta_{\nu_n n}\theta\rangle)\ldots\exp(\langle X, \Delta_{2n}\theta\rangle)\exp(\langle X, \Delta_{1n}\theta\rangle)$$

(application proceeds from right to left) leads to the time reversed evolution

$$f(x) \mapsto f(\xi_1(\Delta_{1n};\xi_2(\Delta_{2n};\ldots(\xi_{\nu_n}(\Delta_{\nu_n n};x))\ldots)))),$$

starting from ξ_{ν_n} and working back to ξ_1, where $\xi_i(\tau;x)$ denotes the solution of $d\xi_i/d\tau = \Sigma \Delta_{in}\theta_k/\Delta_{in} \cdot \sigma_k(\xi)$, $\xi_i(0) = x$. The time reversal here can be straightened out, though, by reversing time in the Brownian motion instead. Thus, if $\bar{\theta}(\tau) = \theta(t) - \theta(t-\tau)$ then $\psi^{(n)}(t-\tau) = \bar{L}_{in}\bar{\theta}/\bar{\Delta}_{in}$, $\bar{t}_{i-1n} < \tau \leq \bar{t}_{in}$, and $\bar{\Delta}$ refers to the reversed partition $\bar{t}_{in} = t - t_{\nu_n-in}$. Under the additional assumption that the first partial derivatives of σ are uniformly Lipschitz continuous, Stroock and Varadhan [21] have shown that for the sequence of dyadic partitions $\nu_n = [2^n t] + 1$, $t_{in} = i/2^n$, $0 \leq i \leq [2^n t]$, $t_{\nu_n n} = t$ one has the convergence in distribution $\xi^{(n)} \Rightarrow \xi$ where ξ is the solution of the Stratonovich stochastic differential equation

$$d\xi = \Sigma\sigma_k(\xi) \circ d\bar{\theta}_k, \quad \xi(0) = x. \tag{10}$$

(This is one of the advantages of using the vector field form $\frac{1}{2}\Sigma X_k^2$, rather than the form $\Sigma\Sigma a_{ij}\partial^2/\partial x_i\partial x_j + \Sigma b_i \partial/\partial x_i$, for the generator. Namely, the Itô stochastic differential

equation for its underlying diffusion corresponds simply to the above Stratonovich equation. On the other hand the vector field form is restrictive -- not every generator has this form, even if we allow an additional first order term X_0 as done below.) One should identify $\xi^{(n)}$ appearing in (9) as the solution of the equation obtained from (10) by replacing $\bar\theta$ with its piecewise linear interpolant passing through the interpolation points $(\bar{t}_{in}, \bar\theta(\bar{t}_{in}))$, $0 \leq i \leq \nu_n$. (Cf. Wong and Zakai [23].) Thus we discover the form that our limiting operator in (8) ought to have:

$$\prod_0^t \exp(<X, d\theta(u)>) : f(x) \mapsto f(\xi(t;x)), \tag{11}$$

where ξ is the solution of (10). It is this type of operator we refer to as a "random wave operator." When considered in terms of two parameters $\prod_s^t$, these operators form a random two parameter semi-group (with stationary independent increments, much like Brownian motion on a Lie group).

The representation (11) ought properly to be understood as "the fundamental theorem of calculus" for the product integrals of this form. It shows that the product integral obeys a certain differential equation (namely (10)), when considered as a function of its upper limit, and thus obviates the necessity of resort to partitions $0 = t_{0n} < t_{1n} < \ldots < t_{\nu_n n} = t$ for its evaluation. In fact, it relates the calculus of these product integrals to the technically rich and easily mastered second order, or Itô calculus.

The product formula (5) has implications concerning the support properties of $\exp(\frac{1}{2} t \Sigma X_k^2)$. For $z \in \mathbb{R}^\ell$ let us denote by $\xi(\tau;x,z)$ the solution of $d\xi/d\tau = \Sigma z_k \sigma_k(\xi)$, $\xi(0) = x$. Observe then that $T(\Delta_{in})$ from (6) can be expressed as

$$T(\Delta_{in}): f(x) \mapsto \int_{\mathbb{R}^\ell} f(\xi(1;x,z)) p(z,\Delta_{in}) dz, \tag{12}$$

where we have introduced the notation $p(z,t) = \Pi p(z_k,t)$. Thus in order that $x \in \text{supp } T(\Delta_{in}) f$ there must exist a set $C \subset \mathbb{R}^\ell$ of positive Lebesgue measure, for which $\xi(1;x,z) \in \text{supp } f$, $\forall z \in C$. We refer to C as a control set, through which ξ can be controlled to go from x (at time zero) into supp f (at time one). By a simple time scale we see that $\xi(1;x,z) = \xi(\Delta_{in};x, z/\Delta_{in})$, and we can use C/Δ_{in} to control ξ so as to enter supp f at time Δ_{in} (rather than time one). Extending this argument we can represent

$$\underset{i}{\Pi} T(\Delta_{in}): \tag{13}$$

$$f(x) \mapsto \int_{\mathbb{R}^\ell \times \ldots \times \mathbb{R}^\ell} f(\xi(t;x,\frac{z^{(\nu_n)}}{\Delta_{\nu_n n}},\ldots,\frac{z^{(1)}}{\Delta_{in}})) \underset{i}{\Pi} p(z^{(i)},\Delta_{in}) dz^{(i)}$$

where $\xi(\tau;x,z^{(1)},\ldots,z^{(\nu_n)}) = \xi(t;x,\psi^z)$ denotes the solution of $d\xi/d\tau = \Sigma \psi_k^z(\tau) \sigma_k(\xi)$, $\xi(0) = x$, and $\psi^z(\tau) = z^{(i)}$, $\bar{t}_{i-1n} \le \tau < \bar{t}_{in}$. Here we use the notation Z for $(z^{(1)},\ldots,z^{(\nu_n)})$. Thus for each choice $z^{(1)},\ldots,z^{(\nu_n)} \in \mathbb{R}^\ell$ we associate the piecewise constant function ψ^z which takes the value $z^{(i)}$ over the interval $[\bar{t}_{i-1n},\bar{t}_{in})$ obtained from the given

165

partition. (Recall that $\bar{t}_{in} = t - t_{\nu_n - in}$.) The controls, then, here are functions $\psi: [0,t) \to \mathbb{R}^\ell$ which are piecewise constant over the fixed partition intervals $[\bar{t}_{i-1n}, \bar{t}_{in})$. Each such function can be uniquely associated with some $Z \in \mathbb{R}^\ell \times \ldots \times \mathbb{R}^\ell$, and we use the notation ψ^Z to denote this correspondence. The action of the control ψ is given by the differential equation $d\xi/d\tau = \Sigma \psi_k(\tau) \sigma_k(\xi)$. It follows from (13) then that $x \in$ supp $\prod_i T(\Delta_{in})f$ only if there exists a set $C \subset \mathbb{R}^\ell \times \ldots \times \mathbb{R}^\ell$ of positive Lebesgue measure, for which $\xi(t; x, \psi^Z) \in$ supp f, $\forall Z \in C$. That is, we need to use ψ^Z to control the process ξ so as to go from x (at time zero) into supp f (at time t). Furthermore, if $f \geq 0$ then this control criterion is also a sufficient condition for $x \in$ supp $\prod_i T(\Delta_{in})f$.

Letting $n \to \infty$ in (6) effectively picks up all piecewise constant, or step functions ψ. (Especially if our techniques allow us to use arbitrary partitions. Otherwise we may be restricted, say, to dyadic step functions -- with dyadic step intervals.) The product formula (7) then implies that $x \in$ supp $\exp(\frac{1}{2} t \Sigma X_k^2)f$ only if there exist step function controls ψ, through which ξ can be controlled so as to go from x (at time zero) into supp f (at time t). It is clear from the equation $d\xi/d\tau = \Sigma \psi_k(\tau) \sigma_k(\xi)$ that this property is in fact independent of $t > 0$. We see from a time scale that if ξ can be controlled to get to supp f at time t, then it can be controlled to get to supp f at any other time t'.

However, we leave the "time t" requirement in the above
support-control condition, since it will be important when
we add on a first order vector field X_0, and describe
$\text{supp} \exp(t(\frac{1}{2} \Sigma X_k^2 + X_0))$. There the support-control condi-
tion will not be independent of t.

Using the representation (10), (11) we see that this
support-control condition above in fact amounts precisely
to the Stroock-Varadhan characterization for the support of
a diffusion [21]. Their result establishes that the support
of the diffusion ξ satisfying (10) is precisely the closure
of the set of $\eta \in C([0,\infty), \mathbb{R}^m)$ for which there exists a step
function $\psi : [0,\infty) \to \mathbb{R}^\ell$ such that $d\eta/d\tau = \Sigma \psi_k(\tau) \sigma_k(\eta)$,
$\eta(0) = x$.

<u>Example II</u>: $X_1 = 2x_2 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial x_2}$, $X_2 = x_3 \frac{\partial}{\partial x_3}$.

The propagation is confined to hyperbolic cylinders. If
supp f lies in one or both parts of the wedge
$a \le x_1^2 - 2x_2^2 \le b$ on one side (above/below) of the $x_1 x_2$-
plane, then so does $\text{supp} \exp(\frac{1}{2} t(X_1^2 + X_2^2)) f$.

Our next interest is in extending (7) to a product
integral representation for the general linear evolution
equation

$$\frac{\partial u}{\partial t} = [\frac{1}{2} \Sigma X_k^2(t) + X_0(t)]u + a(x,t)u + b(x,t), \qquad (14)$$

where the vector fields $X_k(t) = X_{\sigma_k}(t) = \sigma_k \cdot \nabla$ now come from
time dependent mappings $\sigma_k(x,t)$ from $\mathbb{R}^m \times [0,\infty) \to \mathbb{R}^m$,

$0 \leq k \leq \ell$. To begin with we need to identify $\exp(X + a(x))$, where $X = \sigma \cdot \nabla$ is a vector field, and $a: \mathbb{R}^m \to \mathbb{R}$ is a bounded measurable mapping. Since we want $\exp(t(X + a(x)))$ to be the group generated by $X + a(x)$ we define

$$\exp(X + a(x)): \quad f(x) \mapsto f(\xi(1;x)) \exp(\int_0^1 a(\xi(\tau;x)) d\tau) \qquad (15)$$

where, as above, $\xi(\tau;x)$ is the solution of $d\xi/d\tau = \sigma(\xi)$, $\xi(0) = x$. Consider now the product

$$\prod_i \exp(<X(t_{i-1n}), \Delta_{in}\theta> + [X_0(t_{i-1n}) + a(x, t_{i-1n})]\Delta_{in}),$$

involving the ℓ-tuple of time dependent vector fields $X(t) = (X_1(t), \ldots, X_\ell(t))$ along with $X_0(t)$ and the mapping $a(x,t)$. A careful "keeping track of things" reveals that this is the random wave operator

$$f(x) \mapsto f(\xi^{(n)}(t;x)) \exp(\int_0^t a^{(n)}(\xi^{(n)}(\tau;x), t-\tau) d\tau),$$

where $\xi^{(n)}(\tau;x)$ is the solution of $d\xi^{(n)}/d\tau = \Sigma \psi_k^{(n)}(t-\tau)\sigma_k^{(n)}(\xi^{(n)}, t-\tau) + \sigma_0^{(n)}(\xi^{(n)}, t-\tau)$, $\xi^{(n)}(0) = x$ and for $t_{i-1n} \leq \tau < t_{in}$

$$\psi^{(n)}(\tau) = \frac{\Delta_{in}\theta}{\Delta_{in}}, \qquad \sigma_k^{(n)}(\tau) = \sigma_k^{(n)}(x, t_{i-1n}),$$

$$a^{(n)}(x,\tau) = a(x, t_{i-1n}).$$

By reversing time it is easy to see that $\xi^{(n)}(\tau;x) = \bar{\xi}^{(n)}(t-\tau;x)$ where $\bar{\xi}^{(n)}$ satisfies $d\bar{\xi}^{(n)}/d\tau = -\Sigma \psi_k^{(n)}(\tau)\sigma_k^{(n)}(\bar{\xi}^{(n)}, \tau) - \sigma_0^{(n)}(\bar{\xi}^{(n)}, \tau)$, $\bar{\xi}^{(n)}(t) = x$. Thus one expects the product integral $\prod_s^t \exp(<X(u), d\theta(u)> + [X_0(u) + a(x,u)]du)$ to be the

random wave operator

$$f(x) \rightarrow f(\overline{\xi}(s;x))\exp(\int_s^t a(\overline{\xi}(\tau;x),\tau)d\tau),$$

where $\overline{\xi}(\tau;x)$ satisfies the backward Stratonovich differential equation

$$d\overline{\xi} = -\Sigma\sigma_k(\overline{\xi},\tau) \circ d\theta_k(\tau) - \sigma_0(\overline{\xi},\tau)d\tau, \quad \overline{\xi}(t) = x. \quad (16)$$

Thus, using the technique of variation of parameters for product integrals (Dollard and Friedman [9]), we see that the solution of (14) with initial condition $u(x,0) = f(x)$ is given by

$$u(x,t) = f(\overline{\xi}(s;x))\exp(\int_s^t a(\overline{\xi}(\tau;x),\tau)d\tau)$$

$$+ \int_s^t b(\overline{\xi}(r;x),r)\exp(\int_r^t a(\overline{\xi}(\tau;x),\tau)d\tau)dr. \quad (17)$$

§2. Extension to Higher Order Equations

In order to extend these ideas to higher order equations we need to introduce the fundamental solution $p_n(z,t)$ for the equation $\partial u/\partial t = (-1)^{(n/2)-1}/n! \, \partial^n u/\partial z^n$, n even. This function has the scale property $p_n(z,t) = t^{-1/n} p_n(t^{-1/n}z,1)$, and $p_n(z,1)$ is given by

$$p_n(z,1) = \frac{1}{\pi} \int_0^\infty \cos \lambda z \, \exp(-\frac{\lambda^n}{n!}) d\lambda. \qquad (1)$$

Associated with p_n is a generalized Brownian motion $\theta_{(n)}$ with transition densities $p_n(x,t)$ and infinitesimal·generator $(-1)^{(n/2)-1}/n! \, \partial^n/\partial x^n$. This process $\theta_{(n)}$ has been studied by several people ([10],[12], [14], [16]), and is not a genuine diffusion since its transition densities $p_n(x,t)$ are signed. In fact it does not even arise from a signed probability on path space, since such a measure would necessarily be of infinite total variation. Nevertheless if we are willing to work in a finitely additive setting it can be shown that $\theta_{(n)}$ generates an nth order analogue to Itô's stochastic calculus. In particular if $\theta_{(n)}(t) = (\theta_{(n)1}(t),\ldots,\theta_{(n)\ell}(t))$ is ℓ-dimensional generalized Brownian motion then

$$\mathbb{E} \exp(<x,\theta_{(n)}(t)>) = \exp(\frac{(-1)^{\frac{n}{2}-1}}{n!} t \, \Sigma x_k^n); \quad x \in \mathbb{R}^\ell, \, t > 0. \qquad (2)$$

This parallels (1.4) and we can again ask about replacing $x = (x_1,\ldots,x_\ell)$ with $X = (X_1,\ldots,X_\ell)$ for vector fields $X_k = \sigma_k \cdot \nabla$, $1 \leq k \leq \ell$. Can we expect

170

$$\exp((-1)^{\frac{n}{2}-1}/n! \; t\Sigma X_k^n) = \prod_0^t \mathbb{E}\exp(<X,\theta_{(n)}(du)>)$$

(3)

$$= \mathbb{E} \prod_0^t \exp(<X,d\theta_{(n)}(u)>),$$

where $\prod_0^t \exp(<X,d\theta_{(n)}(u)>)$ is to be interpreted as a random wave operator $f(x) \mapsto f(\xi(t;x))$ and ξ is the solution of the generalized Stratonovich differential equation

$$d\xi = \Sigma\sigma_k(\xi) \circ d\bar{\theta}_{(n)k}, \qquad \xi(0) = x? \qquad (4)$$

The use of the name Stratonovich here simply indicates that the generator of ξ is to have the invariant form $(-1)^{(n/2)-1}/n! \; \Sigma X_k^n$. (Cf. [20, §4].) The $\bar{\theta}$ again indicates a time reversal $\bar{\theta}(\tau) = \theta(t) - \theta(t-\tau)$, $0 \le \tau \le t$. The left equality in (3) can be understood without the need of setting up a generalized stochastic calculus. The operator $T(t) = \mathbb{E}\exp(<X,\theta_{(n)}(t)>)$ is given by

$$f(x) \mapsto \int_{\mathbb{R}^\ell} f(\xi(1;x,z))p_n(z,t)dz,$$

where ξ satisfies $d\xi/d\tau = \Sigma z_k\sigma_k(\xi)$, $\xi(0) = x$ and $p_n(z,t) = \Pi p_n(z_k,t)$. Thus the analogue of (1.13) holds here, with p replaced by p_n. In particular establishment of the first part of (3) would lead to the analogous support property for $(-1)^{(n/2)-1}/n! \; \Sigma X_k^n$.

The right equality in (3) comes in only as a handy calculational tool for the product integral. It shows that the calculus of these product integrals $\prod_0^t \exp(<X,d\theta_{(n)}(u)>)$ is

171

essentially an nth order analogue of the Itô calculus. This is that same "fundamental theorem of calculus" described above.

In order to allow for lower order terms in the generator we introduce the generalized Appell polynomials (see Bell [1]) $\phi_n: \mathbb{R}^{n-1} \to \mathbb{R}$ defined by

$$\phi_n(y_1,\ldots,y_{n-1}) = \frac{1}{n!} \frac{\partial^n}{\partial t^n} \exp(\Sigma y_j t^j)/_{t=0}. \qquad (5)$$

Let $\theta_{(n)}$ be a one-dimensional n^{th} order Brownian motion, and set $\theta_{(n)}^{(j)}(t) = \int_0^t d\theta_{(n)}^j$, $1 \le j \le n-1$. (See [10]; [16].) Then

$$\mathbb{E} \exp(\Sigma y_j \theta_{(n)}^{(j)}(t)) = \exp((-1)^{\frac{n}{2}-1} t\phi_n(y_1,\ldots,y_{n-1})). \qquad (6)$$

This leads us then to expect the following generalization of (3). Let $X_{jk} = \sigma_{jk} \cdot \nabla$, $1 \le j \le n-1$, $1 \le k \le \ell$, and $X_0 = \sigma_0 \cdot \nabla$ be smooth vector fields. Let $\theta_{(n)} = (\theta_{(n)1},\ldots,\theta_{(n)\ell})$ be ℓ-dimensional. Then

$$\exp(t(-1)^{\frac{n}{2}-1} \sum_k \phi_n(X_{1k},\ldots,X_{n-1k}) + X_0))$$

$$= \prod_0^t \mathbb{E} \exp(\Sigma_j \Sigma_k X_{jk} \theta_{(n)k}^j(du) + X_0 du) \qquad (7)$$

$$= \mathbb{E} \prod_0^t \exp(\Sigma_j \Sigma_k X_{jk} d\theta_{(n)k}^j(u) + X_0 du),$$

and $\prod_0^t \exp(\Sigma_j \Sigma_k X_{jk} d\theta_{(n)k}^j(u) + X_0 du)$ is to be interpreted as a random wave operator $f(x) \mapsto f(\xi(t;x))$, where $\xi(\tau;x)$ is the solution of the generalized Stratonovich differential equation

$$d\xi(\tau) = \sum_j \sum_k \sigma_{jk}(\xi) \circ d\bar{\theta}^{\,j}_{(n)k}(\tau) + \sigma_0(\xi)d\tau, \quad \xi(0) = x. \quad (8)$$

Again the use of the name Stratonovich here simply indicates that the generator of ξ is to have the invariant form $(-1)^{(n/2)-1} \sum_k \phi_n(X_{1k},\ldots,X_{n-1k}) + X_0$. We must explain here, though, what is meant by $\phi_n(X_1,\ldots,X_{n-1})$ for vector fields $X_1,\ldots,X_{n-1}$. Our convention is that all monomials are evaluated as symmetric products. For example if $a(y_1,y_2) = y_1y_2$ and $b(y_1,y_2) = y_1y_2^2$ then

$$a(X_1,X_2) = \frac{1}{2}(X_1X_2 + X_2X_1)$$

$$b(X_1,X_2) = \frac{1}{3}(X_1X_2^2 + X_2X_1X_2 + X_2^2X_1).$$

The operator $T(\Delta) = \mathbb{E}\exp(\sum_j \sum_k X_{jk}\theta^{\,j}_{(n)k}(\Delta) + X_0\Delta)$ is given by

$$f(x) \to \int_{\mathbb{R}^\ell} f(\xi(\Delta;x,z:\Delta))p_n(z,t)dz,$$

where ξ satisfies $d\xi/d\tau = \sum_j \sum_k z^{\,j}_k/\Delta \; \sigma_{jk}(\xi) + \sigma_0(\xi), \; \xi(0) = x$.

Another approach to allow for lower order terms, based on the Cameron-Martin formula, is that of Motoo [14] and Nishioka [16]. These authors consider what amounts here to product integrals

$$\prod_0^t \exp(<X,d\theta_{(n)}(u)> + \sum_j \sum_k \alpha_{jk}(x)d\theta^{\,j}_{(n)k}(u) + \alpha_0(x)du),$$

where $X = (X_1,\ldots,X_\ell)$ are vector fields $X_k = \sigma_k \cdot \nabla$, $1 \le k \le \ell$, the maps $\sigma_k: \mathbb{R}^m \to \mathbb{R}^m$ are C_b^{n-1} and the maps $\alpha_{jk}: \mathbb{R}^m \to \mathbb{R}$, $1 \le j \le n-1$, $1 \le k \le \ell$, and $\alpha_0: \mathbb{R}^m \to \mathbb{R}$ are C_b^1. Following our earlier steps one can see that this product integral ought

173

to be the random wave operator

$$f(x) \mapsto f(\xi(t;x))$$

$$\cdot \exp\{\int_0^t \sum_j \sum_k \alpha_{jk}(\xi(\tau;x)) \cdot d\bar{\theta}_{(n)k}^j(\tau) + \alpha_0(\xi(\tau;x))d\tau\}$$

(9)

where ξ is the solution of the generalized Stratonovich differential equation (4), and the stochastic integrals in (9) are generalized Stratonovich integrals. This means that

$$\int_0^t \alpha(\tau) \cdot d\theta^j(\tau) = \lim_n \sum \int_{\Delta_{i-1n}}^{\Delta_{in}} \alpha(\tau) \frac{(\Delta_{in}\theta)^j}{\Delta_{in}} d\tau.$$

The operator

$$T(\Delta) = \mathbb{E} \exp(<X, \theta_{(n)}(\Delta)>$$

$$+ \sum_j \sum_k \alpha_{jk}(x)\theta_{(n)k}^j(\Delta) + \alpha_0(x)\Delta)$$

(10)

is given by

$$f(x) \mapsto \int_{\mathbb{R}^\ell} f(\xi(\Delta;x, \frac{z}{\Delta}))$$

$$\cdot \exp\{\int_0^\Delta [\sum_j \sum_k \alpha_{jk}(\xi(\tau;x, \frac{z}{\Delta}))z_k^j/\Delta$$

(11)

$$+ \alpha_0(\xi(\tau;x, \frac{z}{\Delta}))]d\tau\}p_n(z,\Delta)dz,$$

where $\xi(\tau;x,z)$ satisfies $d\xi/d\tau = \sum z_k\sigma_k(\xi)$, $\xi(0) = x$. The analogue to (7) is then

174

$$\exp(t(-1)^{\frac{n}{2}-1} \sum_k \phi_n(X_k + \alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{n-1k}) + \alpha_0))$$

$$= \prod_0^t \mathbb{E} \exp(<X, \theta_{(n)}(du)> + \sum_j \sum_k \alpha_{jk}\theta^j_{(n)k}(du) + \alpha_0 du) \qquad (12)$$

$$= \mathbb{E} \prod_0^t \exp(<X, d\theta_{(n)}(u)> + \sum_j \sum_k \alpha_{jk}d\theta^j_{(n)k}(u) + \alpha_0 du),$$

where α_{jk} represents the operator which multiplies a function by $\alpha_{jk}(x)$. The rule here for evaluating $\phi_n(X + \alpha_1, \alpha_2, \ldots, \alpha_{n-1})$, where X is a vector field and $\alpha_i: \mathbb{R}^m \to \mathbb{R}$ are smooth functions, is exactly as above in (7): the monomials in ϕ_n are evaluated symmetrically. This representation (12) is analogous to what Simon [19, §15] calls the Feynman-Kac-Itô formula.

Actually the form of (9) appearing in Motoo [14] and Nishioka [16] is the non-symmetric form, involving generalized Itô rather than Stratonovich integrals. If one replaces the Stratonovich integrals in (9) with their Itô counterparts, then the operator $T(\Delta)$, representing the short time average, would have to be given by

$$f(x) \mapsto \int_{\mathbb{R}^\ell} f(\xi(\Delta; x, \frac{z}{\Delta}))$$
$$\cdot \exp\{\sum_j \sum_k \alpha_{jk}(x) z^j_k + \alpha_0(x)\Delta\} p_n(z, \Delta) dz, \qquad (13)$$

rather than (11). The counterpart to the left equality of (12) is

$$\exp(t(-1)^{\frac{n}{2}-1} \sum_k \phi_n(X_k + \alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{n-1k}) + \alpha_0))$$
$$= \prod_0^t T(du) \qquad (14)$$

175

but now $\phi_n(X + \alpha_1, \alpha_2, \ldots, \alpha_{n-1})$ has to be evaluated in its non-symmetric form, where all monomials involving X are evaluated by applying X first (i.e., on the right). In this case the coefficients α_{jk} need not be smooth, since none of their derivatives arise in the generator. The different generators corresponding to (11) and (13) reflect the difference in the stochastic calculus stemming from the Stratonovich and Itô integrals, respectively. For the special case $\xi(t;x) = x + \theta_{(n)}(t)$ in (9) (i.e., $\sigma_k \equiv 1$, $1 \leq k \leq \ell$), which is the case studied by Motoo and Nishioka, the conversion is simply based on

$$\int_0^t \alpha(x+\theta(\tau) \circ d\theta^j(\tau)$$

$$= \sum_{k=0}^{n-j} \int_0^t \frac{1}{(k+1)!} \alpha^{(k)}(x+\theta(\tau)) d\theta^{j+k}(\tau) \tag{15}$$

for θ one-dimensional. The advantage of the symmetric form (11) over the non-symmetric form (13) is that the former arises as a wave operator (namely (10)), whereas the latter does not. In general wave operators obtained through stochastic product integrals involve symmetric, or Stratonovich, stochastic integrals and stochastic differential equations, and correspondingly the generators take a symmetric form.

To prove the left equality of (12) we use a special case of the version of Chernoff's product formula appearing in Pierre and Rihani [18].

176

Theorem I: Let A be the generator of a linear contraction C_0 semi-group, and let $\{T(t): t > 0\}$ be a family of linear contractions satisfying

$$\lim_{t \downarrow 0} \frac{T(t)f - f}{t} = Af, \quad f \in D(A). \tag{16}$$

Then the strong product integral $\prod\limits_0^t T(du)$ exists and equals exp(tA).

We can extend this Theorem so as to allow A to be the generator of certain non-contractive C_0 semi-groups. The operators T(t) can be bounded linear operators satisfying (16), provided there exists a constant $\omega \geq 0$ such that

$$\| T(t) \| \leq e^{\omega t}, \quad t > 0. \tag{17}$$

One simply replaces T(t) with $e^{-\omega t}T(t)$ and A with A−ω, and then Theorem I applies.

177

References

1. Bell, E. T., Exponential polynomials, <u>Ann. Math.</u> 35 (1934), 258-277.

2. Berger, M. A. and Sloan, A. D., A method of generalized characteristics, <u>Mem. Amer. Math. Soc.</u> 266 (1982).

3. Berger, M. A. and Sloan, A. D., Explicit solutions of partial differential equations, in Non-Standard Analysis - Recent Developments, <u>Lecture Notes in Mathematics</u> 983, A. D. Hurd, ed., Springer-Verlag, N.Y., 1983.

4. Berger, M. A. and Sloan, A. D., Product formulas for semigroups with elliptic generators, <u>J. Func. Anal.</u> 57 (1984), 244-269.

5. Berger, M. A. and Sloan, A. D., Characteristic methods for multi-dimensional evolution equations, <u>J. Diff. Eqns.</u> 57 (1985), 15-43.

6. Berger, M. A. and Sloan, A. D., Product formulas for solutions of initial value partial differential equations I, <u>J. Diff. Eqns.</u> 57 (1985), 224-247.

7. Berger, M. A. and Sloan, A. D., Product formulas for solutions of initial value partial differential equations II, <u>J. Diff. Eqns.</u>, to appear.

8. Chernoff, P. R., Note on product formulas for operator semi-groups, <u>J. Func. Anal.</u> 2 (1968), 238-242.

9. Dollard, J. D. and Friedman, C. N., <u>Encyclopedia of Mathematics and its Applications</u>, Vol. 10: Product Integration, G. C. Rota, ed., Addison-Wesley, Reading, Mass., 1979.

10. Hochberg, K. J., A signed measure on path space related to Wiener measure, <u>Ann. Prob.</u> 6 (1978), 433-458.

11. Ibero, M., Intégrales stochastiques multicatives et construction de diffusions sur un groupe de Lie, <u>Bull. Sc. math.</u> (2e serie) 100 (1976), 175-191.

12. Krylov, V. Yu., Some properties of the distribution corresponding to the equation $\partial u/\partial t = (-1)^{q+1} \partial^{2q} u/\partial x^{2q}$, <u>Sov. Math. Dokl.</u> 1 (1960), 760-763.

13. McKean, H. P. Jr., <u>Stochastic Integrals</u>, Academic Press, New York, 1969.

178

14. Motoo, M., An analogue to the stochastic integral for $\partial/\partial t = -\Delta^2$, <u>Adv. in Prob.</u>, Vol. 7, M. Pinsky, ed., Marcel Dekker, New York, 1984, 323-338.

15. Nishioka, K., A complex measure related to the Schrödinger equation, <u>Adv. in Prob.</u>, Vol. 7, M. Pinsky, ed., Marcel Dekker, New York, 1984, 339-351.

16. Nishioka, K., Stochastic calculus for a class of evolution equations, <u>Jap. J. Math.</u> (New Series) 11 (1985), 59-102.

17. Nishioka, K., A stochastic solution of a higher order parabolic equation, preprint.

18. Pierre, M. and Rihani, M., About product formulas with variable step size, MRC Technical Summary Report #2783, January 1985.

19. Simon, R., <u>Functional Integration and Quantum Physics</u>, Academic Press, New York, 1979.

20. Stratonovich, R. L., A new representation for stochastic integrals and equations, <u>J. SIAM Control</u>, 4 (1986), 362-371.

21. Stroock, D. W. and Varadhan, S. R. S., On the support of diffusion processes with applications to the strong maximum principle, <u>Proc. Sixth Berkeley Symp. in Prob. and Stat.</u> III (1972), 333-359.

22. Stroock, D. W. and Varadhan, S. R. S., Limit theorems for random walks on Lie groups, <u>Sankhya</u> Ser. A (3) 35 (1973), 277-294.

23. Wong, E. and Zakai, M., Riemann-Stieltjes approximation of stochastic integrals, <u>Z. Wahr. und Verw. Gebiete</u> 12 (1969), 87-97.

Characteristic Functions of a Class of Probability Distributions

Siegfried H. Lehnigk, Huntsville, Alabama
Research Directorate
Research, Development, and Engineering Center
U. S. Army Missile Command
Redstone Arsenal, Alabama 35898-5248

Summary. The characteristic function of a class of continuous one-sided probability distributions is being considered. The distribution class contains three independent parameters; one of them represents scale, the other two determine initial and terminal shape of the associated probability density function. The analytical properties of the characteristic function depend heavily on the terminal shape parameter λ which may vary in the interval $(-\infty, 1)$. If $0 < \lambda < 1$, the characteristic function is many-valued with branch points at zero and infinity. Its principal branch is holomorphic and bounded upon analytic continuation (into the complex plane cut along the nonnegative real axis) from the primary element which is holomorphic in the open left-hand plane. If $\lambda = 0$, the primary element of the characteristic function is holomorphic in the half-plane left of the vertical line through the point $(b^{-1}, 0)$, b being the scale parameter. Upon continuation it becomes either a rational function (if the initial shape parameter is a nonpositive integer) with a pole at the point $(b^{-1}, 0)$ or a many-valued function with branch points at $(b^{-1}, 0)$ and infinity whose principal branch is holomorphic in the plane cut along the real axis from b^{-1} to infinity. If $\lambda < 0$, the characteristic function is an entire function of order greater than unity. It has no real zeros but an infinity of conjugate complex pairs of zeros even if the order is an even integer.

(To appear in Complex Variable: Theory and Application)

Poisson and Extreme Value Limit Theorems for Markov Random Fields*

Simeon M. Berman
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012

ABSTRACT

Let Z^m be the integer lattice in R^m, and let X_t, $t \in Z^m$, be a Markov random field. Let I_n be a rectangular box in Z^m with corner points having coordinates of the form $\pm n$. Define $M_n = \max(X_t: t \in I_n)$. The extreme value limit problem is as follows: Find conditions under which there exist a nondegenerate distribution function $G(x)$ and real sequences (a_n) and (b_n), with $a_n > 0$, such that the conditional probability

$$P(a_n^{-1}(M_n-b_n) < x \mid \underline{\text{Given }} X_s: s \in \underline{\text{boundary of }} I_n)$$

converges, for $n \to \infty$, to $G(x)$ at all points of continuity, and for all possible values of X_s on the boundary.

Here the extreme value limit problem is solved for a general class of Markov random fields. The conditions on the field are stated in terms of the system of nearest neighbor conditional distributions. These distributions are assumed to be invariant under translations in Z^m (homogeneity). Dobrusin's condition for regularity and mixing is also assumed to hold, so that there exists a unique stationary measure P.

In addition to these general conditions, the following more special conditions are also assumed:

1. For fixed t, the marginal distribution of X_t under the stationary measure belongs to the domain of attraction of an extreme value limiting distribution function $G(x)$ with normalizing sequences (a_n) and (b_n). This is equivalent to

$$\lim_{n \to \infty} P^n(X_t < a_n x + b_n) = G(x).$$

2. For all possible values of X_s for points s which are neighbors of O,

$$P(X_0 > u \mid X_s, s \text{ } \underline{a} \text{ } \underline{\text{neighbor}} \text{ } \underline{\text{of}} \text{ } 0) = O(P(X_0 > u)), \text{ for } u \to \infty.$$

This paper has been accepted for publication in <u>Advances in Applied Probability</u>.

A BOUND ON THE VARIATION BETWEEN TWO PROBABILITY MEASURES

IN TERMS OF THE INTENSITIES OF A DISCRETE POINT PROCESS

RELATIVE TO THESE PROBABILITIES.

G. R. Andersen

U.S. Army Ballistic Research Laboratory (BRL)

1. Introduction: In an application of discrete parameter point processes to a communication network problem at the BRL there was a need to measure the robustness of a point process intensity. That is, we wanted to know if small changes in the intensity of a point process implied small changes in the distribution of the point process. H. Rost proved such a result for continuous parameter point processes having absolutely continuous compensators, in the 1984 University of Strasbourg Seminar in Probability. If Rost had considered the case where the compensator was absolutely continuous with respect to an increasing process (instead of just Lebesgue measure), it would have been directly applicable to the discrete point process model. Rather than extend his result in this direction here, it was decided to see what would be required to prove an analogous result totally within the framework of discrete point processes. These processes are sequences of Bernoulli random variables (with no distributional assumptions or assumptions concerning independence) and so are of fundamental importance to probability theory.

To derive a discrete parameter analogue of Rost's result, we will require a sequence of four Lemmas. These Lemmas are known from the general theory (Jacod [1979], Itmi [1980], Bremaud [1981]) where they are proved in the case of continuous parameter marked point processes. Bremaud also treats in detail the case where the point process has an absolutely continuous compensator. The latter case does not apply to discrete point processes, but the form of the statements and the essential mechanics of the proofs for the discrete case can be inferred from Bremaud's presentation. The relationship of discrete point processes to the general marked point process of Jacod [1975] is given in Andersen [1986, Chapter 4]. The mathematical setting considered by Jacod is general enough to allow one to obtain the correct statements of the Lemmas for a discrete point process by the simple device of embedding such a process (and filtration) in a continuous parameter process (and filtration) which is constant between integer times.

In the case of discrete point processes, however, it is extremely easy and informative to derive these Lemmas directly from first principles and this is what we will do. The discrete analogue of Rost's Theorem simply does not follow from his result and so it is derived in Section 5. For a discussion on discrete point processes and their use in approximating continuous parameter point processes one can refer to Brown [1983].

Readers not already familiar with the relatively new martingale techniques might find

that the discrete form of stochastic calculus provides easy access to this area. These techniques have wide applicability to engineering, physics and statistics; for a small sample of applications to queuing, control, statistics, reliability and design of experiments see Bremaud [1981], Jacobsen [1982], Gill [1980]

2. **Notation and Preliminaries:** Let Z_+ be the set of non-negative integers. $(\Omega, F_\infty, (F_n), P)$ is called a **filtered probability space** if F_∞ is a σ-algebra of subsets of Ω and P is a probability measure on F_∞ with F_n a sub σ-algebra of F_∞, for each $n \epsilon Z_+$ and the sequence $n \rightarrow F_n$ is increasing (F_n is contained in F_{n+1}, for all $n \epsilon Z_+$). $X = (X_n, n \epsilon Z_+)$ is said to be a **stochastic process** if each X_n is a random variable on (Ω, F_∞). Let $\Delta X_k := X_k - X_{k-1}$ and define the process X_- by setting $(X_-)_n := X_{n-1}$ with $X_{-1} := 0$ for all $n \epsilon Z_+$. As always, the conditional expectation of a P-integrable random variable Z given the σ-algebra F_k is written $E(Z \mid F_k)$. In what follows, a constantly (and silently) used property of conditional expectation is that if g is a bounded F_k-measurable process, then $E(gZ \mid F_k) = gE(Z \mid F_k)$, a.s.P; the abbreviation "a.s.P" means "almost surely relative to the probability P". Its use with the last equation indicates that the random variables defined on either side of this equation are only equal on an event whose probability is one.

Let $X = (X_n)$ and $V = (V_n)$ be processes on (Ω, F_∞). Then the **transform of X by V**, denoted $V.X = ((V.X)_n)$, is the process defined by setting

$$(V.X)_n(w) := \sum_{k=0}^{n} V_k(w) \Delta X_k(w) ,$$

for all w in Ω. If X is a square integrable processes relative to P, then the **variance process** of X is denoted by $<X,X>$ and is defined by

$$<X,X>_n := \sum_{k=0}^{n} E((\Delta X_k)^2 \mid F_{k-1}),$$

a.s.P.

$X = (X_n)$ is said to be **adapted** to the filtration $F = (F_n)$ iff X_n is F_n-measurable for each n, while $V = (V_n)$ is said to be **previsible** relative to the filtration F iff V_n is F_{n-1}-adapted. If X is F-adapted, then X_- is F-previsible.

If M is a discrete parameter process on the filtered probability space $(\Omega, F_\infty, (F_n), P)$, then $M = (M_n, F_n)$ is an **(F,P)-martingale** iff

     (i)   M is adapted to F,

     (ii)   M has finite expectation

     (iii) $E(M_n \mid F_{n-1}) = M_{n-1}$ (a.s.P)

          for all $n \epsilon Z_+$

186

2.0.1. Remark. The following examples are immediate consequences of the definitions.

(1°) If M is an (F,P)-martingale and V is F-previsible. Then V.M is an (F,P)-martingale, if V.M is P-integrable.

(2°) If M is a square-integrable martingale, then $M^2 - <M,M>$ is a martingale.

(3°) If X, V and V.X are square integrable processes, and V is previsible, then $<V.X,V.X>_n = V^2.<X,X>_n$, a.s.P. and $E((V.X)^2) = E(V^2.<X,X>)$, if X is a martingale.

Additional notation used includes writing the "indicator function" of a set A as 1_A.

3. Discrete Point Processes on a Measure Space:
Let (Ω, F_∞) be a measurable space. Suppose that $(T_n, n\epsilon Z_+)$ is a strictly increasing sequence of $\overline{Z}_+ := Z_+ \bigcup \{\infty\}$ valued random variables relative to (Ω, F_∞). The statement that the sequence is "strictly increasing" means that for all $n\epsilon Z_+$,

$$T_n < T_{n+1} \quad \text{on} \quad [T_n < \infty] := \{w\epsilon\Omega : T_n(w) < \infty\}.$$

Thus defined, the sequence $(T_n, n\epsilon Z_+)$ is called a **discrete point process** (Dpp).

Given a discrete point process $(T_n, n\epsilon Z_+)$, it is customary to introduce the process, $N = (N_t, t \geq 0)$, corresponding to (T_n) by setting

$$N_t := \sum_{m \geq 1} 1_{[T_m \leq t]} \tag{1}$$

for $t \geq 0$. $N_t(w)$ counts the number of times that members of the sequence $(T_m(w), m \geq 1)$ fall in the interval $(0,t]$.

In the case treated here the "times" T_n take their values in $\overline{Z}_+$; they are "integer-valued". It follows then that

$$N_t = \sum_{m=1}^{\lfloor t \rfloor} 1_{[T_m \leq t]} \tag{2}$$

(This is because, while finite, the T_m are strictly increasing and integer-valued functions, so at most $\lfloor t \rfloor$ of them can occur before time t.) Note that $\lfloor t \rfloor$ represents the greatest integer less than or equal to t. There should be no confusion between this use of brackets and their use in specifying sets, as in $[T_m \leq t] := \{w : T_m(w) \leq t\}$.

For each $k\epsilon Z_+$, set

$$X_k(w) := \sum_{m=1}^{k} 1_{[T_m = k]}(w), \tag{3}$$

and $X_0(w) = 0$, for all $w\epsilon\Omega$. Then it is easy to see that

$$N_t = \sum_{k=0}^{\lfloor t \rfloor} X_k, \quad t \geq 0. \tag{4}$$

(Just insert the right side of (3) for X_k in (4) and interchange order of summation.) It is

sufficient for our purposes then to consider the counting process in (4) as just the stochastic sequence $N = (N_n, n \epsilon Z_+)$.

Starting with the discrete point process (T_n) we have defined the sequence $(X_n, n \epsilon Z_+)$ of Bernoulli 0,1-valued random variables on (Ω, F_∞). Each of the processes $N = (N_n)$ and $X = (X_n)$ are equivalent representations of the discrete point process (T_n). For example, if we are given a Bernoulli sequence $X = (X_n, n \epsilon Z_+)$ relative to (Ω, F_∞) we can define the sequence (T_n) by setting $T_0 :\equiv 0$ and

$$T_m = \inf\{k \epsilon \overline{Z}_+ : k > T_{m-1}, X_k = 1\} \tag{6}$$

for $m \geq 1$, when $\{ \cdots \} \neq \phi$ and equal to ∞ otherwise. The sequence $N = (N_n)$ is defined by

$$\Delta N_n = N_n - N_{n-1} = X_n \tag{7}$$

$N_{-1} = 0$, for $n \epsilon Z_+$, so that

$$N_n = \sum_0^n X_k. \tag{8}$$

4. Discrete Processes on a Filtered Probability Space

We begin with the filtered probability space $(\Omega, F_\infty, (F_n), P)$, a P-complete filtration (F_0 contains all P-null sets) and $F_\infty = \sigma(\bigcup_{n \geq 0} F_n)$.

Under this set-up an F-adapted $\{0,1\}$-Bernoulli process on (Ω, F_∞) with $X_0 = 0$ on Ω is said to be an **(F,P)-discrete point process** (Dpp).

The process N defined as in (8) is then an F-adapted process also and the sequence $(T_n, n \epsilon Z_+)$ defined by (6) is a sequence of F-stopping times. For the reasons noted earlier all three sequences are called **(F,P)- discrete point processes.**

For each $n \epsilon Z_+$, define the stochastic sequence $\lambda = (\lambda_n, n \epsilon Z_+)$ by

$$\lambda_n := E(X_n \mid F_{n-1}), \tag{9}$$

$n \geq 1$ and $\lambda_0 = 0$ on Ω. Then the process λ is said to be the **(F,P)-intensity** of the underlying (F,P) discrete point process.

When there is no ambiguity about which filtration or probability measure is being used we will sometimes drop one or both of the qualifiers F, P and just refer to the "intensity". On the other hand, when we must keep in mind that these processes depend on F and P we will write, for example, $\lambda = (\lambda_n, F_n, P)$ or just $\lambda = (\lambda_n, F_n)$. It will be understood that the index n is in Z_+. The following properties are immediate:

(a) $\lambda = (\lambda_n, F_n, P)$ is F-previsible and $0 \leq \lambda_n \leq 1$ a.s.P.

(b) $N = (N_n, F_n, P)$ is F-adapted with compensator $\Lambda_n = \sum_{k=0}^n \lambda_k$. That is,

$$m_n = N - A_n \qquad (10)$$

is an (F,P)-martingale.

4.0.1. Remark: As noted in the introduction, the following four Lemmas are known from the general theory (Jacod [1979], Bremaud [1981]) where they are proved in the case of continuous parameter marked point processes They are proved here totally within the framework of discrete point processes for the purpose of exposition and in the belief that Bernoulli variates are at the heart of most things probabilistic.

4.1. **Lemma:**

Suppose that $N = (N_n, F_n)$ *is a Dpp with F-intensity* $\lambda = (\lambda_n, F_n)$ *Let* $\mu = (\mu_k, F_k)$ *be a strictly positive F-previsible process and* $\psi_k := 1 + \lambda_k(\mu_k - 1)$. *Then* $\psi_k > 0$ *for all* $k \epsilon Z_+$ *and*

$$L_n = \prod_{k=0}^{n} \frac{\mu_k^{X_k}}{\psi_k} \qquad (11)$$

for $n \epsilon Z_+$ *defines a positive F-martingale,* $L = (L_n, F_n, P)$.

4.1.1. Remark: $L_0 = 1$ on Ω.

4.1.2. Remark: That ψ_k is positive for all k follows from (a) by treating the three cases $\lambda_k = 0$, $\lambda_k = 1$, and $0 < \lambda_k < 1$. To show that $L = (L_n, F_n)$ is a martingale just realize that since the X's take values in $\{0,1\}$,

$$\mu_k^{X_k} = \mu_k X_k + (1 - X_k).$$

Then, by the F-previsibility of μ and the definition of λ

$$E(\mu_k^{X_k} \mid F_{k-1}) = 1 + \lambda_k(\mu_k - 1) = \psi_k. \qquad (12)$$

Since ψ_k is F-previsible, it follows from (12) that

$$E(\frac{\mu_k^{X_k}}{\psi_k} \mid F_{k-1}) = 1,$$

Hence for $n \geq 1$, since $L_n = L_{n-1}\dfrac{\mu_n^{X_n}}{\psi_n}$ and L_{n-1} is F_{n-1}-measurable,

$$E(L_n \mid F_{n-1}) = L_{n-1}E(\frac{\mu_n^{X_n}}{\psi_n} \mid F_{n-1}) = L_{n-1}.$$

That is, L is an F-martingale and L is clearly positive.

4.1.3. Remark: Notice that since L is a martingale and $L_0 = 1$ on Ω, $EL_n = EL_0 = 1$ for all $n \geq 0$.

4.1.4. Remark: It is sometimes useful to write $\mu_k^{X_k} = 1 + (\mu_k-1)X_k$ in the definition of L_n. This is about all that is needed to prove the following discrete analogue of a result due to C. Doleans-Dade.

4.2. Lemma: $(\mu, \psi, \text{ and } X \text{ as in Lemma 4.1.})$
Set

$$g_k = (\mu_k - 1)/\psi_k \tag{13}$$

If

$$L_n = \prod_{k=0}^{n} \frac{\mu_k^{X_k}}{\psi_k}, \tag{14}$$

then

$$L_n = 1 + (gL_-.m)_n \tag{15}$$

and conversely.

4.2.1. Remark: Just observe that from (13) and (14) with $n \geq 1$,

$$L_n - L_{n-1} = L_{n-1}(1+(\mu_n - 1)X_n - \psi_n)/\psi_n = L_{n-1}g_n\Delta m_n,$$

where m is defined in (b) as $m = N - A$, so that $\Delta m_n = X_n - \lambda_n$. Summing both sides of the first equation gives

$$L_n - L_0 = \sum_{k=1}^{n} L_{k-1}g_k\Delta m_k = (L_-g.m)_n,$$

since $\Delta m_0 = X_0 - \lambda_0 = 0$. Because $L_0 = 1$ we have (15). The converse follows by reversing the argument.

4.2.2. Remark: We continue with $\lambda = (\lambda_n, F_n, P)$ as the (F,P)-intensity of a Dpp $X = (X_n, F_n, P)$.

4.3. Lemma:
Let L, μ and ψ be defined as in Lemma 4.1. Define a probability measure $\overline{P}$ on (Ω, F_∞) by setting

$$\overline{P}(A) = \int_A L_v(w)P(dw), \tag{16}$$

for all $A \epsilon F_\infty$. Then $X = (X_n, F_n, \overline{P})$ is a Dpp with $(F,\overline{P})$-intensity α, where

$$\alpha_k = \lambda_k\mu_k/\psi_k, \quad \text{a.s.P,} \tag{17}$$

$k=1,2,...,v$.

4.3.1. Remark: From (17), notice that $0 \leq \alpha_k = \mu_k\lambda_k/(1 - \lambda_k + \mu_k\lambda_k) \leq 1$, a.s.P, as it should.

190

4.3.2. Remark: Following Lemma 4.1, we noticed that the positive martingale of that Lemma had the property that $E(L_n) = 1$. It follows that the measure $\overline{P}$ defined above is a probability measure. If we let $E_Q Y$ denote the expectation of Y relative to the probability measure Q, then Bucy's Lemma (Bremaud [1981, p171]) allows us to write

$$E_{\overline{P}}(X_n \mid F_{n-1}) E_P(L_n \mid F_{n-1}) = E_P(L_n X_n \mid F_{n-1}).$$

Hence, writing E for E_P,

$$\alpha_n = E_{\overline{P}}(X_n \mid F_{n-1}) = \psi_n^{-1} E(\mu_n^{X_n} \mid F_{n-1}),$$

since

$$E(L_n \mid F_{n-1}) = L_{n-1} E(\mu_n^{X_n} \mid F_{n-1}),$$

$$E(L_n X_n \mid F_{n-1}) = L_{n-1} E(\mu_n^{X_n} \mid F_{n-1}) = L_{n-1} \psi_n,$$

and $L_{n-1} > 0$. The conclusion of the Lemma then follows by recalling that μ is F-previsible and noting that $\mu_n^{X_n} X_n = \mu_n X_n$, ($X_n$ takes only the values 0 and 1).

4.3.3. Remark: It follows immediately that $\alpha = 1$ iff $\lambda = 1$ and $\alpha = 0$ iff $\lambda = 0$, which is useful in attempting to solve (17) for μ in Section 5.

4.3.4. Remark: In the last Lemma we used the positive martingale L to define a probability measure $\overline{P}$ which was absolutely continuous relative to P. In the next Lemma we give a "converse" of that result. For this purpose we take $F_k = F_k^N := \sigma(N_1, \cdots, N_k)$, where N is a Dpp with (F^N, P) intensity λ.

4.4. Lemma:
Let $\overline{P}$ and P be a probability measures on (Ω, F_∞), $F_\infty = \sigma(\bigcup_k F_k^N)$, and suppose that $\overline{P}$ is absolutely continuous with respect to P,

$$\overline{P} \ll P.$$

Let $\overline{P}_n$ and P_n be the restrictions of $\overline{P}$ and P, respectively, to F_n^N, and define

$$L_n := \frac{d\overline{P}_n}{dP_n}, \tag{18}$$

the Radon-Nikodym derivative of $\overline{P}$ relative to P. Then there exists a positive, F^N– previsible process (μ_k) such that

$$L_n = \prod^n \frac{\mu_k^{X_k}}{\psi_k}, \tag{19}$$

where, as before, $\psi_k = 1 + \lambda_k(\mu_k - 1)$, and so N is an $(F^N, \overline{P})$ Dpp with $(F^N, \overline{P})$ intensity

$$\alpha_k = \lambda_k \mu_k / \psi_k, \tag{20}$$

for all $k \in Z_+$.

4.4.1. Remark: The general idea of the proof of this Lemma is to note that L, as defined in (18), is an (F^N, P) - martingale relative to the filtration generated by the discrete point process. This fact can be used to write L in the form of equation (15) in such a way that g in this equation is previsible. Lemma 4.2, with μ defined through (13) and g, then applies and L has the required product representation (19). The form of the new point process intensity then follows from Lemma 4.3.

5. Rost's Theorem for Discrete Point Processes:

Let F be the filtration used in the last Lemma, $F_n^N = \sigma(N_k, k \leq n)$.

Let $\overline{P}$ and P satisfy the assumptions of Lemma 4.4 and define $L = (L_n, F_n^N, P)$ as in equation (18). Then this Lemma together with Lemma 4.2 says that L satisfies the following transform ("integral") equation

$$L_n = 1 + (L_- . \xi)_n, \tag{21}$$

where $\xi_n = (g.m)_n$ is an (F_n^N, P) martingale by 1° of Section 2, since $m = N - A$ as defined in equation (10) is such a martingale and g given by (13) is F^N previsible.

The variation of the two probability measures $\overline{P}$ and P is connected to the process L through

$$P_v(A) - \overline{P}_v(A) = \int_A (1 - L_v) \, dP, \tag{22}$$

for all $A \epsilon F_v^N$, where v is some fixed positive integer. Since $\overline{P}_v$ and P_v are the restrictions of P to F_v^N and $F_v^N \subset F_\infty$ we can drop the subscripts on the left side of (22).

Roussas [1972] shows that the **total variation**, $Var_v(\overline{P}, P) := Var(\overline{P}_v, P_v)$, between $\overline{P}_v$ and P_v (or between $\overline{P}$ and P on F_v^N) defined by

$$Var_v(\overline{P}, P) := \sup\{ \, |\overline{P}(A) - P(A)| : A \epsilon F_v^N \, \}$$

satisfies

$$Var_v(\overline{P}, P) = E((1 - L_v)1_{[L_v \leq 1]}), \tag{23}$$

where the expectation on the right is with respect to the probability P.

Now we follow along the lines of Rost's proof to obtain a bound on $Var_v(\overline{P}, P)$. The form of his bound is of course different from the one that will be obtained here since his compensator is absolutely continuous relative to Lebesgue measure.

To obtain a bound on the left member of equation (23), we will decompose the right member into two parts in such a way that one part is small and the other is small only when the $(F^N, \overline{P})$-intensity α and the (F^N, P)-intensity λ are, in some sense, close.

Since $L_0 = 1$ and $L_v > 0$ on Ω, we decompose the event $[L_v \leq 1]$ into the union of two disjoint events: $[1 - \epsilon < L_v \leq 1]$ and $[0 < L_v \leq 1 - \epsilon]$. On the first event, $1 - L_v$ is between 0 and ϵ and on the second it is between ϵ and 1. It follows from (23) that

$$Var_v(\overline{P}, P) \leq \epsilon + P(\epsilon < 1 - L_v \leq 1). \tag{24}$$

Let

$$S = \inf\{k \in Z_+ : k \leq v \text{ and } |1 - L_k| > \epsilon \}, \tag{25}$$

if $\{ \cdots \} \neq \phi$ and equal v otherwise. Then

$$P(\epsilon < 1 - L_v \leq 1) \leq P(\max_{k \leq v} |1 - L_k| > \epsilon) \leq P(\epsilon < |1 - L_S|)$$

$$\leq \frac{1}{\epsilon^2}E((1 - L_S)^2) = \frac{1}{\epsilon^2}E((L_-\cdot\xi)_S^2), \tag{26}$$

the last equality being due to (21). Using (3°) of Section 2, we have

$$E((L_-\cdot\xi)_S^2) = E(L_-^2 \cdot <\xi,\xi>_S), \tag{27}$$

where

$$\Delta<\xi,\xi>_k = \Delta<g.m,g.m>_k = g_k^2\Delta<m,m>_k = g_k^2\lambda_k(1 - \lambda_k). \tag{28}$$

In order to obtain the last equality from $\Delta<m,m>_k$ refer to Section 2. Then

$$E((X_k - \lambda_k)^2 \,|\, F_{k-1}) = E(X_k^2 \,|\, F_{k-1}) - 2\lambda_k E(X_k \,|\, F_{k-1}) + \lambda_k^2$$

$$= \lambda_k(1 - \lambda_k),$$

since $X_k^2 = X_k$. From equations (25) through (28), we find that

$$P(\epsilon < 1 - L_v \leq 1) \leq \frac{1}{\epsilon^2}E\sum_1^S L_{k-1}^2 g_k^2\lambda_k(1 - \lambda_k)$$

$$\leq \left(\frac{1 + \epsilon}{\epsilon}\right)^2 E\sum_1^S g_k^2\lambda_k(1 - \lambda_k), \tag{29}$$

where we have used the definition of the stopping time S which provides $L_{k-1}(w) \leq 1+\epsilon$, for k going from 1 to S(w). Using Remark 4.3.3. and equations (13) and (17), one can show that

$$g_k^2\lambda_k(1 - \lambda_k) = \begin{cases} 0 & , \text{on } [\lambda_k = 0 \text{ or } \lambda_k = 1] \\ (\alpha_k - \lambda_k)^2/\lambda_k(1 - \lambda_k) & , \text{on } [0 < \lambda_k < 1] \end{cases} \tag{30}$$

Therefore, since $P(S \leq v) = 1$ and the quantities in (30) are non-negative, we can replace S in the expectation on the right of (29) by v to obtain

5.1. Theorem:
Let $\bar{P}$ and P be probability measures on the measure space (Ω, F_∞) with $\bar{P} \ll P$. Let ϵ be any real number such that $0 < \epsilon < 1$. If N is a discrete point process with an F^N-intensity α relative to $\bar{P}$ and an F^N-intensity λ relative to P, then

$$\text{Var}_v(\bar{P},P) \leq \epsilon + \left(\frac{1 + \epsilon}{\epsilon}\right)^2 E\sum_1^v g_k^2\lambda_k(1 - \lambda_k), \tag{31}$$

where the summands on the right satisfy (30).

193

5.1.1. Remark: Recall that the expectation,

$$C_v := E\sum_1^v g_k^2 \lambda_k (1 - \lambda_k),$$

on the right of (31) does not depend on ϵ and that there are no constraints on ϵ other than it is in the open interval (0,1). So, if the non-negative quantity C_v is less than 1, we can choose $\epsilon = C_v^{1/3}$. Then

$$\epsilon + \left(\frac{1+\epsilon}{\epsilon}\right)^2 C_v = C_v^{1/3} + (1 + C_v^{1/3})^2 C_v^{1/3} \leq 5C_v^{1/3}. \tag{32}$$

Therefore, (31) and (32) yield the following

5.2. **Theorem:**
Under the assumptions of Theorem 5.1,

$$\text{Var}_v(\overline{P}, P) \leq 5C_v^{1/3}. \tag{33}$$

5.2.1. Remark: Notice that since (23) holds and $L_v \geq 0$, a.s.P, we always have $\text{Var}_v(\overline{P}, P) \leq 1$. Hence, (33) is trivial when C_v is not less than 1.

5.2.2. Remark: Only the form of C_v differs from Rost's result.

5.2.3. Remark: The bound in (31) or (33) differs considerably from the usual L_1-bound discussed in Kabanov, Liptser and Shiryaev [1983] and Serfling [1978], Lemma 6.1.

5.2.4. Remark: The following example is due to Rost. It illustrates the use of (31) or (33). Suppose that $\alpha_k = \alpha_{k(v)}$, $1 \leq k \leq v$, $\lambda_k = \lambda$, a constant, and $|\alpha_k - \lambda| = O(v^{-\delta})$ where $1 > \delta > .5$. Then

$$0 \leq E\sum_1^v g_k^2 \lambda_k (1 - \lambda_k) \leq C (1/v^{2\delta-1}) \to 0,$$

as $v \to \infty$ and so $\text{Var}_v(\overline{P}, P) \to 0$.

By contrast, under these assumptions an L_1-bound on the total variation would be unbounded. This doesn't mean that either type of bound is better or worse than the other, just different.

194

REFERENCES

Aalen, O., Bredrup, E. (1985). Asymptotic Results for Estimators in Restricted Randomization Designs. *Scand J Statist* 12, 303-211.

Andersen, G. (1986). Survey and Introduction to Modern Martingale Theory. (Stopping Times, Semi-Martingales and Stochastic Integration). *Ballistic Research Laboratory Technical Report*, to be released.

Bremaud, P. (1981). *Point Processes and Queues Martingale Dynamics.* Springer-Verlag Series in Statistics.

Brown, T.C. (1983). Some Poisson Approximations Using Compensators. *The Annals of Probability* Vol. 11 No.3 726-744.

Doleans-Dade, C. (1970). Quelques applications de la formule de changement de variables pout les semi-martingales. *Z. Wahr. verw. Giebiete* 16 181-194.

Gill, R.D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Central Limit Theorems for Martingales. *Scand. J. Statistics* 9.

Itmi, M. (1980). Histoire Interne des processus ponctuels marqués stochastiques. Etude d'un probléme de filtrage. Thése de 3éme cycle U. de Rouen.

Jacobsen, M (1982). *Statistical Analysis of Counting Processes.* Lect. Notes in Statistics, No.12, Springer-Verlag.

Jacod, J. (1975). Multivartiate Point Processes: Predictable Projection Radon-Nikodym Derivatives Representation of Martingales. *Z. Wahr. verw. Giebiete* 32 235-253.

Jacod, J. (1979). *Calcul Stochastique et Problemes de Martingales.* Lect. Notes in Math. 714 Springer-Verlag.

Kabanov, Y.M. and Liptser, R.S. and Shiryaev, A.N. (1983). Weak and Strong Convergence of the Distributions of Counting Processes. *Theory of Probability and its Applications* Vol XXVIII 303-336.

Rost, H (1984). Diffusion De Spheres Dures Dans La Droite Reelle: Comportement Macroscopique Et Equilibre Local., *Sem. de Proba.* XVIII, Lect. Notes Math. #1059.

Roussas, G. (1972). *Contiguity of Probability Measures.* Cambridge University Press.

Serfling, R. (1978). Some Elementary Results on Poisson Approximation in a Sequence of Bernoulli Trials. *SIAM REVIEW*, Vol. 20, No. 3, 567-579.

SOME PROBLEMS OF ESTIMATION FROM
POISSON TYPE COUNTING PROCESSES

Michael J. Phelan
Operations Research & Industrial Engineering
Upson Hall
Cornell University
Ithaca, New York 14853

ABSTRACT. We survey some estimation problems in the area of life
history analysis. The problems we describe involve estimating an arbitrary
life-distribution and the transition probabilities of a Markov chain. In our
discussion we emphasize the role of the observation scheme, for example
survival testing versus renewal testing, and the role of the product-limit
estimator. In this connection we demonstrate the need for the family of
Poisson type counting processes in developing a unified methodology for
solving these problems.

1. INTRODUCTION. Many areas of science such as demography, medicine,
industrial reliability and epidemiology give rise to phenomena involving life
histories whose description characterize a family of stochastic processes.
Our interest lies, in particular, in problems where individual life histories
are viewed as realizations of a stochastic process moving among states in a
discrete state space (pure jump processes). For such processes the states
denote the status of an individual (insurance policy, technical component,
etc.) and transitions between states denote events of interest.

   To fix ideas consider a problem in epidemiology where one studies the
relationship between a particular exposure and the incidence of any disease
that may develop. For example, healthy individuals may be initially
classified with regard to cigarette exposure and followed forward in time to
determine which of heart disease or lung cancer develops. Thus "health,"
"heart disease" and "lung cancer" are three states in an individual's life
history and an event occurs when the individual moves from a healthy state to
one of the diseased states. This is the subject of cohort analysis where the
object of interest is the effect of exposure on rate of disease incidence
(see Breslow (1985)).

   The example above highlights a salient feature of problems in the area
of life history analysis. That is, individual life histories are influenced
by the presence of auxiliary processes, such as cigarette exposure, which are
seen to effect the rate at which events occur. A similar motivation lies

behind some random shock and wear models of system reliability where, in addition to system age, the hazard rate of time to failure depends on environmental stresses. In recent years statistical analysis for such dynamical phenomena have relied on counting processes and their compensators. A germinal paper in this regard is Aalen (1978) who first introduced the multiplicative intensity model to various life history problems such as survival analysis. Since then a great deal of activity in this area has taken place and an excellent exposition of the role of counting processes in life history analysis is to be found in Andersen and Borgan (1985).

Counting process models based on intensities generalize the Poisson process as a model for random events in time. These models assume the natural time parameter to be continuous. However, some phenomena such as consumer loan repayment behavior naturally occur in discrete-time. Moreover, longitudinal data sets in sociology often arise from panel designs which generate observed stochastic processes with discrete-time parameter. In our work we have found it useful to consider a family of counting processes which, in the terminology of Liptser and Shiryaev (1978), we call Poisson type counting processes. These counting processes are characterized in section 2 by their compensators whose pathwise Radon-Nikodym derivative relative to a fixed Borel measure is an observable predictable process. The model generalizes the multiplicative intensity and allows for a unified treatment of mixed discrete and continuous-time problems. We describe several examples including survival analysis with arbitrary distribution measure and Markov chains. In statistical applications each example gives rise to an estimation problem which we reduce to that of estimating the Borel measure mentioned above. In sections 3 and 4 we survey some recent results in this area which rely on the martingale dynamics over point processes as discussed, for example, by Liptser and Shiryaev (1978), Jacod (1975) and Boel et al. (1975).

2. POISSON TYPE COUNTING PROCESSES. We define the family of Poisson type counting processes and give a number of worked examples. Let $(\Omega, \mathcal{I}, P)$ denote a probability space and $F = \{\mathcal{I}_t, t \geq 0\}$ a given family of sub-σ-algebras of $\mathcal{I}$ where F is nondecreasing, right-continuous and complete relative to P. All of the standard terminology used below, such as adapted, predictable and compensator, are defined in, for example, Métevier (1982), and Liptser and Shiryaev (1978).

Let $N = \{N_t, \mathcal{I}_t, t \geq 0\}$ denote a counting process defined on $(\Omega, \mathcal{I}, P)$ so that the sample paths of N are right continuous step functions with jumps of size $+1$, $N_0 = 0$ and N_t is a $\mathcal{I}_t$-measurable random variable. Let B denote a fixed Borel measure over the Borel sets in $R_+ = [0, \infty)$ and let $Y = \{Y_t, \mathcal{I}_t, t \geq 0\}$ denote a nonnegative predictable process.

Definition 2.1. If N has compensator $A = \{A_t, \mathcal{I}_t, t \geq 0\}$ relative to F given by

$$A_t = \int_{(0,t]} Y_s B\{ds\},$$

198

then we say N is a <u>Poisson type counting process</u>.∎

By definition the process $M = \{N_t - A_t, \mathcal{F}_t, t \geq 0\}$ is a local square integrable martingale and, in the terminology of Jacod (1975), the kernel $A\{dt\}$ is the dual predictable projection to N and is unique to within stochastic equivalence. If $Y_t = \lambda > 0$ is constant for all $t \geq 0$ and B denotes Lebesgue measure, then $A_t = \lambda t$ and N is a simple Poisson point process. More generally, if B is absolutely continuous relative to Lebesgue measure μ, then definition 2.1 gives rise to the multiplicative intensity model where the intensity is given by $Y dB/d\mu$.

In life history analysis problems involving an event of a single type N_t denotes the number of occurrences of this event over $(0,t]$ and A_t denotes the cumulative conditional rate of event occurrence over $(0,t]$. The actual composition of the conditional rate depends explicitly on the underlying filtration F, often called a history, so that specification of the relative richness of F is important. In applications Y plays the role of auxiliary process which may be some measure of environmental exposure or censoring. For example, in epidemiology Y might denote a measure of cigarette exposure.

Consider the following examples of Poisson type counting processes.

<u>Example 2.1. Survival analysis with censored data</u>. For each $n \geq 1$ let X_i and U_i, $i = 1,\ldots,n$ denote $2n$ independent positive random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ with X_i or U_i almost surely finite for each i. X_i has distribution measure G and U_i has distribution measure H. The observable random variables $\tilde{X}_i$ and δ_i are given by $\tilde{X}_i = \min(X_i, U_i)$ and $\delta_i = 1(X_i \leq U_i)$, where $1(A)$ is the indicator function of event A. In applications X_i denotes the survival time and U_i denotes a censoring time so that this is a model for random right censorship.

A history $F = \{\mathcal{F}_t, t \geq 0\}$ will have to record the progress in the lifetimes of the individuals or components under test. In this case the natural history is given by

$$\mathcal{F}_t = \mathcal{F}_0 \vee \sigma-(\tilde{X}_i \leq s, \delta_i 1(\tilde{X}_i \leq t), s \leq t, i = 1,\ldots,n)$$

where $\mathcal{F}_0$ contains the P-null sets of $\mathcal{F}$ and their subsets.

In the counting process formulation of this model for each $i = 1,\ldots,n$ we define $N(i) = \{1(\tilde{X}_i \leq t, \delta_i = 1), t \geq 0\}$ and $Y(i) = \{1(\tilde{X}_i \geq t, t \geq 0)\}$ and let B denote the Borel measure generated by the function

$$B(t) = \int_{(0,t]} (1-G(s-))^{-1}G\{ds\} \quad , \quad t \geq 0$$

where $G(s-) = \lim_{u\uparrow s} G(s)$. Clearly the process $N(i)$ is a counting process equal to zero until the ith survival time elapses and has not been censored while $Y(i)$ is called a risk process and is equal to one as long as the ith unit remains alive or at risk to failure.

According to theorem 3.1.1, Gill (1980) $N(i)$ has compensator $A(i) = \{A_t(i), t \geq 0\}$ relative to F given by

$$A_t(i) = \int_0^t 1(\tilde{X}_i \geq s)B\{ds\}.$$

Since $Y(i)$ is a left-continuous process it is predictable (see for example Brémaud (1981)) so that $A(i)$ satisfies definition 2.1 and $N(i)$ is a Poisson type counting process. To prove that $A(i)$ is the compensator to $N(i)$ one can verify the martingale property directly using properties of conditional distributions.∎

In the example above the function $B(t)$, $t \geq 0$ may be interpreted as the cumulative age-specific mortality rate for an average or baseline individual. If the sampling population is homogeneous with respect to mortality, this is a reasonable model of failure rate. On the other hand, for heterogeneous sampling populations it is preferable to model failure rate as a function of an auxiliary random variable.

Example 2.2. Failure rate as a function of a random variable. Let $(\Omega, \mathcal{F}, P)$ denote a probability space on which two positive random variables X and Z are defined. The random variable X models survival time whereas Z denotes a measure of a characteristic of the unit of observation from a heterogeneous population or some environmental exposure. The effect of Z on failure rate is modeled as follows. Let $G(\cdot;Z)$ denote the conditional distribution of X given Z so that

$$G(t;Z) = P(X \leq t|Z) = 1 - \exp\{-ZB(t)\} \quad , \quad t \geq 0$$

where B is a continuous function and denotes the cumulative age-specific mortality rate for a baseline individual (i.e. $Z = 1$). If B admits a density h relative to Lebesgue measure, then given the event $\{X \geq t\}$ and Z the conditional failure rate at time t is given by

$$h(t;Z) = G'(t;Z)/(1 - G(t;Z)) = Zh(t) \quad , \quad t \geq 0.$$

The reader will recognize the conditional rate above as the landmark proportional hazards model introduced to incorporate heterogeneity in life-

testing problems by Cox (1972). The function h is the baseline hazard and
Z is called a covariate.

Let $N = \{1(X \leq t), t \geq 0\}$ denote the survival time counting process
and $F = \{\mathcal{I}_t, t \geq 0\}$ be the history defined by

$$\mathcal{I}_t = \sigma(Z) \vee \sigma(X \leq s, s \leq t).$$

Then according to the model above N has compensator $A = \{A_t, t \geq 0\}$
relative to F given by

$$A_t = \int_0^t Z1(X \geq s)B\{ds\}.$$

We identify $Y_s = Z1(X \geq s)$ in 2.1 and obtain that N is a Poisson type
counting process.

Let $\overline{G}$ denote the unconditional distribution measure of the random
variable X so that

$$\overline{G}(t) = P(X \leq t) = 1 - \exp\{-\overline{B}(t)\} \quad , \quad t \geq 0$$

where $\overline{B}$ is some nonnegative continuous function which uniquely determines
$\overline{G}$. We exploit the martingale approach to survival analysis to determine $\overline{G}$
by using the innovation theorem (see for example Aalen (1978)) to determine
$\overline{B}_X$ by a change of history. The problem may be reformulated as follows. Let
$F^X = \{\mathcal{I}_t^X, t \geq 0\}$ denote the internal history to N given by

$$\mathcal{I}_t^X = \sigma(X \leq s, s \leq t).$$

Note that for each $t \geq 0 \ \mathcal{I}_t^X \subset \mathcal{I}_t$ and consider the problem of determining
the compensator $\overline{A} = \{\overline{A}_t, t \geq 0\}$ to N relative the F^X. By theorem 18.3,
Liptser and Shiryaev (1978) the compensator $\overline{A}$ is given by

$$\overline{A}_t = \int_0^t \overline{z}(s)1(X \geq s)B\{ds\}$$

where $\overline{z}(s) = E(Z|X > s)$ so that $\overline{B}$ is given by $\int \overline{z}dB$. We identify
$Y_s = 1(X \geq s)$ and $\overline{B}$ with B in 2.1 to obtain that $\overline{A}$ is of the Poisson

type. Note that the change of history from F to F^X leaves N in the family of Poisson type counting processes, this is a general result.

In some models of heterogeneity the conditional expectation $\bar{z}$ is easy to calculate. For example suppose Y_2 is a Gaussian random variable with mean μ and variance σ^2. If $Z = Y^2$, then a direct argument using Bayes rule given in Yashin (1985) shows that the conditional density of Y given the event $\{X \geq t\}$ is a Gaussian density with mean and variance parameter $\mu[2\sigma^2 h(t) + 1]^{-1}$ and $\sigma^2[2\sigma^2 h(t) + 1]^{-1}$, respectively. From this fact $\bar{z}$ is easily calculated to be $\sigma^2[2\sigma^2 h(t) + 1]^{-1} + \mu^2[2\sigma^2 h(t) + 1]^{-2}$. ∎

Problems in survival analysis generate a counting process which counts a single event: transition from an initial state to the state "failure." In general, a univariate counting process, such as a renewal counting process, counts the repeated occurrences of a single event over time. Often problems in life history analysis involve multiple types of events occurring over time so that univariate counting processes are not sufficiently general for their study. For example, Markov chains are widely used in demography (e.g. Hoem (1971)), econometrics (e.g. Singer (1981)), and illness-death models (e.g. Mau (1986)) of broad appeal in insurance and medicine. The transitions among the states of a Markov chain may be viewed as events of different types; their being one event associated with each possible pairwise transition among states of the chain. This demands a multivariate counting process which is essentially a collection of univariate counting process where each member of the collection is associated with a particular pairwise interstate transition.

Example 2.3. Non-homogeneous Markov chains. We give a unified treatment to discrete and continuous-time Markov chains. Let $X = \{X_t, t \geq 0\}$ denote a Markov chain with finite state space E, defined on a probability space $(\Omega, \mathcal{I}, P)$. Assume that the sample paths of X are right-continuous with left-hand limits. If X is a discrete-time chain, then X is derived from a Markov chain $\{Y_n, n \geq 0\}$, say, by putting $X_t = Y_n$ for $n \leq t < n + 1$. Let $F^X = \{\mathcal{I}_t^X, t \geq 0\}$ denote the internal history of X given by

$$\mathcal{I}_t^X = \mathcal{I}_0 \vee \sigma(X_s, s \leq t)$$

where $\mathcal{I}_0$ contains the P-null set of $\mathcal{I}$ and their subsets. Observe that in the discrete-time case $\mathcal{I}_t^X = \mathcal{I}_{[t]}^X$, where $[t]$ denotes largest integer in t.

If X is a continuous-time Markov chain, then we assume that X admits the Q-matrix or intensity $Q(t) = (q_{ij}(t), i, j \in E)$ such that for all $t \geq 0$ and $i \neq j \in E$

$$q_{ij}(t) > 0, \; q_{ii}(t) < 0, \; \sum_j q_{ij}(t) = 0 \; \text{ and } \; \int_0^t q_{ij}(s)ds < \infty.$$

202

In this case the transition probabilities $P(s,t)$ are given by the product integral

$$P(s,t) = \prod_{(s,t]} (I + Q(u)\mu\{du\}) \quad 0 \leq s \leq t < \infty,$$

where μ denotes Lebesgue measure; see e.g. Aalen and Johansen (1978).

Alternatively, if X is a discrete-time Markov chain, then we assume that for each integer $n \geq 0$ X admits the transition probability matrix $P^n = (p_{ij}(n))$ such that for $i,j \in E$

$$p_{ij}(n) = P(X_{n+1} = j | X_n = i).$$

Using the same notation we define the discrete analog of the Q-matrix as follows. For each $n \geq 0$ and $i \neq j \in E$ let $q_{ij}(t) = p_{ij}(n)$ if $n \leq t < n + 1$, $q_{ii}(t) = -\Sigma_{j \neq i} q_{ij}(t)$ and $Q(t) = (q_{ij}(t), i,j \in E)$. Then it is easy to verify that the transition probabilities $P(s,t)$ for the discrete-time chain are given by

$$P(s,t) = \prod_{s < n \leq t} (I + Q(n)) = \prod_{s < u \leq t} (I + Q(u)\mu\{du\}) \quad 0 \leq s \leq t < \infty$$

where μ denotes counting measure with support $\{0,1,2,\ldots\}$. If the product is empty it is defined to be I.

Fix $i \neq j \in E$ and define $N(i,j) = \{N_t(i,j), t \geq 0\}$ to be a random process which counts the number of direct transition from i into j for the Markov chain X. Thus for $t \geq 0$

$$N_t(i,j) = \sum_{0 < s \leq t} 1(X_s = j, X_{s-} = i)$$

where $X_{s-} = \lim_{u \uparrow s} X_u$. Our object is to show that $N(i,j)$ is a Poisson type counting process with compensator $A(i,j) = \{A_t(i,j), t \geq 0\}$ relative to F^X given by

$$A_t(i,j) = \int_{(0,t]} 1(X_{s-} = i) q_{ij}(s)\mu\{ds\}$$

where μ is Lebesgue measure for continuous-time chains and counting measure for discrete-time chains.

By virtue of the Lévy formula (see e.g. Brémaud (1981)) it follows that for any $0 \leq s \leq t$

203

$$E(N_t(i,j) - N_s(i,j)|\mathcal{F}_s^X) = E\left[\sum_{s<u\leq t} 1(X_u = j, X_{u-} = i)|\sigma(X_s)\right]$$

$$= E\left[\int_s^t 1(X_{u-} = i)q_{ij}(u)\mu\{du\}|\sigma(X_s)\right]$$

$$= E(A_t(i,j) - A_s(i,j)|\mathcal{F}_s^X)$$

from which the martingale property for $N(i,j) - A(i,j)$ easily follows. Note that by the Markov property conditioning with respect to X_s is equivalent to conditioning with respect to $\mathcal{F}_s^X$. Thus if we identify $Y_s = 1(X_{s-} = i)$ and $B\{du\} = q_{ij}(u)\mu\{du\}$ in 2.1, then $N(i,j)$ is a Poisson type counting process. The proof of this without using the Lévy formula may be made directly in the discrete-time case, and Aalen and Johansen (1978) give an alternative proof for the continuous-time case.∎

Consider a life-testing situation in which at time zero a component is put on test and upon failure is immediately replaced with an identical component and so on. The lifetimes generated by this test procedure may be modeled as an ordinary renewal process and applications can be found in industrial life-testing and animal experimentation.

Example 2.4. Renewal testing. Let $S = \{S_n, n \geq 0\}$ denote a renewal process induced by the arbitrary distribution measure G. Define the renewal counting process $\pi = \{\pi_t, t \geq 0\}$ given by

$$\pi_t = \sum_{n=1}^{\infty} 1(S_n \leq t)$$

Define the history $F = \{\mathcal{F}_t, t \geq 0\}$ as follows. For $t \geq 0$

$$\mathcal{F}_t = \sigma(S_n \leq s, s \leq t, n \geq 1)$$

and consider the problem of finding the compensator $A' = \{A_t, t \geq 0\}$ to π relative to F.

For each $n \geq 1$ let $X_n = S_n - S_{n-1}$ ($S_0 = 0$) so that $X = \{X_n, n \geq 1\}$ is a sequence of independent random variables each with distribution G. Thus the conditional distribution of S_n given $\mathcal{F}_{S_{n-1}}$ (see Brémaud (1981) for a definition) is given by G^n where

$$G^n(t) = P(S_n \leq t | \mathcal{F}_{S_{n-1}}) = G(t - S_{n-1}) \quad , \quad t \geq S_{n-1}.$$

Hence according to proposition 3.1, Jacod (1975) the compensator A is given by

$$A_t = \int_0^t \frac{dG(L(s))}{1-G(L(s)-)} = \sum_{n=1}^{\pi_t} \int_0^{X_n} \frac{dG(s)}{1-G(s-)} + \int_0^{L(t)} \frac{dG(s)}{1-G(s-)} \quad , \quad t \geq 0$$

where $L(t) = t - S_{\pi_{t-}}$ is a left-continuous version of the backwards renewal time. Therefore because of the presence of the L-process with its well known saw toothed sample paths π is <u>not</u> in general a Poisson type counting process.

The regenerative structure of the compensator A suggests that the family of counting processes generated by the interrenewal times $\{X_n, n \geq 1\}$ are of the Poisson type. For each $n \geq 1$ define a counting process $N(n) = \{1(X_n \leq t), t \geq 0\}$ and the history $H = \{\mathcal{H}_t, t \geq 0\}$ by

$$\mathcal{H}_t = \sigma(X_n \leq s, s \leq t, n \geq 1)$$

Then by virtue of the independence of the X_n it is possible to show directly that $N(n)$ has compensator $A(n) = \{A_t(n), t \geq 0\}$ <u>relative to H</u> given by

$$A_t(n) = \int_0^t 1(X_n \geq s) \frac{G\{ds\}}{1-G(s-)}$$

which is clearly of the Poisson type. This may be proved either by an appeal to proposition 3.1, Jacod (1975), by virtue of example 2.1, or by directly verifying the martingale property. This representation is called the sojourn-time approach by Phelan (1986b) who applies an extension of it to problems of inference from Markov renewal processes.∎

Thus far only example 2.1 involved censoring. In practice censored processes or incomplete observations are the rule rather than the exception. Therefore one yardstick of the utility of a given probability model at analyzing life history data is its ability to incorporate general patterns of censoring. Poisson type counting processes meet this demand as is illustrated below.

205

Example 2.5. Censored processes. Let $N = \{N_t, \mathcal{F}_t, t \geq 0\}$ denote a Poisson type counting process with compensator $A = \{A_t, \mathcal{F}_t, t \geq 0\}$ where $F = \{\mathcal{F}_t, t \geq 0\}$ is a given history. Let $C = \{C_t, \mathcal{F}_t, t \geq 0\}$ be a $\{0,1\}$-valued predictable process used to model the censoring. Thus the counting process N is observable only on the set $\{t: C_t = 1\}$ otherwise we say the process is censored. This implies that the observable counting process $\tilde{N} = \{\tilde{N}_t, t \geq 0\}$ is given by the pathwise Stieltjes integral

$$\tilde{N}_t = \sum_{0 < s \leq t} C_s \Delta N_s = \int_0^t C_s dN_s$$

which is often called the censored counting process. Since C is bounded and predictable, by the theory of stochastic integration with respect to counting process martingales (see e.g. Liptser and Shiryaer (1978)) the process defined by the pathwise Stieltjes integral

$$\tilde{M}_t = \int_0^t C_s (dN_s - dA_s) \quad , \quad t \geq 0$$

is a $\mathcal{F}_t$-(local) martingale. Hence $\tilde{N}$ has compensator $\tilde{A} = \{\tilde{A}_t, t \geq 0\}$ relative to F given by

$$\tilde{A}_t = \int_0^t C_s dA_s = \int_0^t C_s Y_s B\{ds\}$$

where $Y = \{Y_t, t \geq 0\}$ and B are defined by 2.1. From this expression it is evident that $\tilde{N}$ is a Poisson type counting with auxiliary process $\{C_t Y_t, t \geq 0\}$ and Borel measure B which it inherits from N.∎

An example of a left-continuous (hence predictable) censoring process $\{C_t = 1(U_i \geq t)\}$ was given in example 2.1. Censoring of Markov chains is considered by Aalen and Johansen (1978) and Phelan (1986c) for models in continuous and discrete-time, respectively, and for the renewal process of example 2.4 by Phelan (1986a). A general discussion of censoring is found in Gill (1980), and in the context of nonparametric tests for comparison of counting processes in Andersen et al. (1982).

One can construct numerous other examples of Poisson type counting processes. For example, Brémaud (1981) constructs a G/M/1 Queue using Poisson counting processes. His departure process (see page 37) gives an

example of a censored homogeneous Poisson process and is therefore of the Poisson type. Although we have not done so it would be of interest to survey stochastic models of natural phenomena which generate Poisson type counting processes. Some examples that we are aware of include models for the mating behavior of fruit flies (Aalen (1978)), labor-force dynamics (Andersen (1985)) and screening carcinogenic chemicals in animal experiments (Mau (1986)).

In practical problems the measure B is unknown and requires estimation. The estimation of B is usually based on observations of the bivariate process (N,Y) over a period of time. A general solution to this problem involves an empirical process called the martingale estimator of B. This estimation procedure is presented next and is applied in section 4 to solve some estimation problems drawn from the models developed above.

3. ESTIMATION FROM POISSON TYPE COUNTING PROCESSES. Let $N =$ $\{N_t, \mathcal{F}_t, t \geq 0\}$ denote a Poisson type counting process with compensator $A = \{A_t, \mathcal{F}_t, t \geq 0\}$, auxiliary process $Y = \{Y_t, \mathcal{F}_t, t \geq 0\}$ and measure B. Definition 2.1 is extended in the following way. Let $J = \{t: \Delta B(t) > 0\}$, where $\Delta B(t) = B(t) - B(t-)$, be the countable set to which B assigns positive mass. If J is nonempty, then for each $t \in J$ we allow $\Delta N(t) > 1$ with positive probability. This extension is used below where the superposition of Poisson type counting processes has this property. If the process (N,Y) is observable over a period of time and B is unknown, then a statistical problem is to estimate B from (N,Y).

Define the predictable process $Y^+ = \{Y_t^+, t \geq 0\}$ given by

$$(3.1) \qquad Y_t^+ = (Y_t)^{-1} 1(Y_t > 0) \quad (0/0 = 0 \text{ by convention})$$

and the empirical process $\hat{B} = \{\hat{B}_t, t \geq 0\}$ given by the Stieltjes integral

$$(3.2) \qquad \hat{B}_t = \int_0^t Y_s^+ dN_s.$$

The process $\hat{B}$ is the proposed estimator of B and is called the martingale estimator by virtue of the observation that the process

$$M_t = \int_0^t Y_s^+ (dN_s - dA_s) = \hat{B}_t - \tilde{B}_t \ , \quad t \geq 0$$

is a (local) martingale where $\tilde{B}_t = \int_0^t 1(Y_s > 0) B\{ds\}$. This follows, for example, by an appeal to theorem 18.7, Liptser and Shiryaev (1978).

207

The statistical theory of the martingale estimator is based on asymptotics. Suppose we are given a sequence $\{N^n, Y^n, n \geq 1\}$ of Poisson type counting processes and their associated auxiliary processes. For each n define $\hat{B}^n = \{\hat{B}^n_t, t \geq 0\}$ from (N^n, Y^n) according to (3.2). We consider the asymptotic (i.e. as $n \to \infty$) properties of the sequence of estimators $\{\hat{B}^n, n \geq 1\}$. Suppose $\{a_n, n \geq 1\}$ is a sequence of positive numbers tending to infinity as n tends to infinity. If Y^n/a_n^2 converges uniformly to a function y in probability as $n \to \infty$, where y is bounded from zero on $[0,a]$, say, and (N^n, Y^n) is derived as the sum of independent Poisson type counting processes, then the following properties will typically hold:

a. <u>Consistency</u>. $\sup_{0 \leq t \leq a} |\hat{B}^n_t - B(t)| \to 0$ in probability as $n \to \infty$;

b. <u>Weak Convergence</u>. For $n \geq 1$ define $Y^n = \{a_n(\hat{B}^n_t - B(t)), t \geq 0\}$, then Y^n converges weakly to a Gaussian process Y^∞ of independent increments as $n \to \infty$. Weak convergence takes place in the space $D([0,a])$ endowed with the Skorohod topology (see Billingsley (1968)).

To prove these results one employs two fundamental tools: an inequality due to Lenglart (1977) and functional central limit theorems for semimartingales as developed in Jacod et al. (1982). To see why observe that for each $n \geq 1$ and $t \geq 0$

$$\hat{B}^n_t - B(t) = \hat{B}^n_t - \tilde{B}^n_t + \tilde{B}^n_t - B(t).$$

It has already been noted that $M^n = \{\hat{B}^n_t - \tilde{B}^n_t, t \geq 0\}$ is a (local) martingale, and $X^n = \{\tilde{B}^n_t - B(t), t \geq 0\}$, being the difference between two monotone processes, is a process of local bounded variation. Hence Y^n is a semimartingale (see Shiryaev (1981)). The conditions above may be used to show directly that $a_n X^n$ converges to zero in probability as $n \to \infty$. In this case the Lenglart inequality is applied to M^n to prove (a). Then martingale functional central limit theorems are applied to $a_n M^n$ to prove (b). We omit the details but note that in our work we have found it convenient to appeal to alternative criteria for tightness found in Jacod and Mémin (1980).

4. SURVEY OF ESTIMATION PROBLEMS . We give a survey of estimation problems and results in the areas of life-testing and Markov chain analysis. We begin with the problem of estimating an arbitrary life-distribution G in life-testing models and then consider the problem of estimating transition probabilities of a Markov chain. In our discussion we emphasize the importance of the observation scheme, for example survival testing versus renewal testing, and the role of product-limit estimators.

4.1 Estimating the life-distribution. Let G denote an arbitrary life-distribution and for $t \geq 0$ define $B(t) = \int_0^t (1 - G(s-))^{-1} G\{ds\}$. For the problem of estimating G we distinguish among three observation schemes.

a) **Survival testing.** For each $n \geq 1$ we observe n pairs $(\tilde{X}_i, \delta_i)$, $i = 1, \ldots, n$ of independent censored lifetimes $\tilde{X}_i$ and their associated censoring indicator $1 - \delta_i$. In the notation of example 2.1, define the aggregate processes $N^n = \{N_t^n, t \geq 0\}$ and $Y^n = \{Y_t^n, t \geq 0\}$ by $N_t^n = \Sigma_1^n N_t(i)$ and $Y_t^n = \Sigma_1^n 1(\tilde{X}_i \geq t)$, respectively. The statistic (N^n, Y^n) is used below to construct the estimator of G. ∎

b) **Renewal testing.** A single renewal process $S = \{S_n, n \geq 0\}$ is observed over an expanding time horizon $[0, T]$, $T > 0$. In the notation of example 2.4, define the aggregate processes $N^T = \{N_t^T, t \geq 0\}$ and $Y^T = \{Y_t^T, t \geq 0\}$ by

$$N_t = \sum_1^{\pi(T)} N_t(n) \quad \text{and} \quad Y_t = \sum_1^{\pi(T)} 1(X_n \geq t),$$

respectively. Here (N^T, Y^T) is the relevant statistic for estimating G. ∎

c) **Renewal testing with finite horizon and repetitions.** Fix $T > 0$. For each $n \geq 1$ we observe n independent renewal processes over $[0, T]$. In the notation of example 2.4, let $\pi(i)$ and $\{X_k(i), k \geq 1\}$ denote the renewal counting process and lifetimes, respectively, for the ith renewal process, $i = 1, \ldots, n$. Then define the aggregate processes $\bar{N}^n = \{\bar{N}_t^n, t \geq 0\}$ and $\bar{Y}^n = \{\bar{Y}_t^n, t \geq 0\}$ by

$$\bar{N}_t^n = \sum_{i=1}^{n} \sum_{\ell=1}^{\pi(i;T)} 1(X_\ell(i) \leq t) \quad \text{and} \quad \bar{Y}_t^n = \sum_{i=1}^{n} \sum_{\ell=0}^{\pi(i;T-t)} 1(X_{\ell+1}(i) \geq t),$$

respectively. The statistic $(\bar{N}^n, \bar{Y}^n)$, which is almost equivalent to aggregating (N^T, Y^T) over n independent realizations, is used to estimate G. ∎

For each observation scheme we define the empirical process $\hat{B}^n = \{\hat{B}_t^n, t \geq 0\}$, $\hat{B}^T = \{\hat{B}_t^T, t \geq 0\}$ and $\bar{B}_t^n, t \geq 0$ by

$$(4.0) \qquad \hat{B}_t^n = \int_0^t (Y_s^n)^+ dN_s^n$$

with $\hat{B}^T$ and $\bar{B}^n$ defined analogously from (N^T, Y^T) and $(\bar{N}^n, \bar{Y}^n)$, respectively. The processes $\hat{B}^n$, $\hat{B}^T$ and $\bar{B}^n$ are the proposed estimators of the measure B for observation scheme (a), (b) and (c), respectively, and

are used to construct product-limit estimators of G. Define the processes $\hat{G}^n = \{\hat{G}^n_t, t \geq 0\}$, $\hat{G}^T = \{\hat{G}^T_t, t \geq 0\}$ and $\bar{G}^n = \{\bar{G}^n_t, t \geq 0\}$ by

$$(4.1) \qquad \hat{G}^n_t = 1 - \prod_{0<s\leq t} (1 - \Delta\hat{B}^n_s) = 1 - \prod_{0<s\leq t} \left[1 - \frac{\Delta N^n_s}{Y^n_s} \right]$$

with $\hat{G}^T$ and $\bar{G}^n_T$ defined analogously from $\hat{B}^T$ and $\bar{B}^n$, respectively. The processes $\hat{G}^n$, $\hat{G}^T$ and $\bar{G}^n$ are the proposed product-limit estimators of G from the observation schemes (a), (b) and (c), respectively. The estimator $\hat{G}^n$ was first formally introduced to the statistical literature by Kaplan and Meier (1958), although its historical origins appear to date earlier (see Gill (1980)), whereas $\hat{G}^T$ and $\bar{G}^n$ are natural extensions of $\hat{G}^n$.

For each t such $G(t) < 1$, lemma 18.8, Liptser and Shiryaev (1978) implies that

$$(4.2) \qquad \frac{\hat{G}^n - G(t)}{1 - G(t)} = \int_0^t \frac{1 - \hat{G}^n_{s-}}{1 - G(s-)} f(\Delta B(s))(d\hat{B}^n_s - dB(s))$$

where $f(\Delta B(s)) = (1 - \Delta B(s))^{-1} 1(\Delta B(s) \leq_T 1)$. Of course it is possible to write analogous expressions involving $\hat{G}^T$ and $\bar{G}^n$. It turns out that these expressions are the key to proving the asymptotic properties of the product-limit estimators since they either define a martingale or can be well approximated by a martingale in probability.

The estimators $\hat{G}^n$, $\hat{G}^T$ and $\bar{G}^n$ are consistent and the normalized differences converge weakly to a Gaussian process of independent increments as n, T and n tend to infinity, respectively. Essentially these estimators inherit these properties from the estimators $\hat{B}^n$, $\hat{B}^T$ and $\bar{B}^n$ as may be proved by the methods of section 3. A detailed study of $\hat{G}^n$ is given by Gill (1980, 1983) although his proof of weak convergence relies on an elaborate construction in theorem 4.2.2, Gill (1980). An alternative proof of weak convergence for $\hat{G}^n$ is given by Phelan (1986a) which is based on the methods outline in section 3 and does not rely on any special constructions. The problem of consistency and weak convergence for the renewal testing estimator $\hat{G}^T$ is considered by Phelan (1986a). His model includes right censoring of the interrenewal times and his method is to show that the equivalent expression to (4.2) for $\hat{G}^T$ is well approximated by a martingale in probability for large T. Then the asymptotic (i.e. $T \uparrow \infty$) properties of $\hat{G}^T$ are established in a manner consistent with that used for $\hat{G}^n$. Finally, the estimator $\bar{G}^n$ is considered by Gill (1981) when G is restricted to being either purely discrete or continuous. He does not employ martingale techniques although we believe the approximation methods of Phelan (1986a) can be modified for this purpose. This would unify the asymptotic treatment of $\bar{G}^n$ with that of $\hat{G}^n$ and $\hat{G}^T$.

In closing this subsection we recall the model of example 2.2 for life-testing in heterogeneous populations or random environments. In the proportional hazards model the random variable Z depends on an unknown

parameter Θ, say, where inference for Θ is also of interest. This is a problem in the general theory of partial-likelihood (see Wong (1986)). In life history analysis this problem has been considered from the modern point of view using counting processes by Andersen and Gill (1982) (see also Prentice and Self (1983)). These authors, of course, generalize the problem to allow Z to be a time-dependent stochastic process depending on Θ.

4.2. <u>Estimating Markov transition probabilities</u>. For fixed $T > 0$ let $P = (P(s,t), 0 \leq t \leq T)$ denote the transition probabilities for a nonhomogeneous Markov chain. Consider the problem of estimating P under the following observation scheme. For each $n \geq 1$ we observe n independent Markov chains $X^i = \{X_t^i\, 0 \leq t \leq T\}$, $i = 1,\ldots,n$ each with finite state space E, transition probabilities P and arbitrary initial distribution.

Let μ denote either counting measure or Lebesque measure and suppose P admits a Q-matrix (cf. example 2.3)) relative to μ. For each $i,j \in E$ and $t \geq 0$ define $B_{ij}(t) = \int_0^t q_{ij}(s)\mu\{ds\}$ and let $B(t) = (B_{ij}(t), i,j \in E)$. We begin by estimating the matrix function $B = (B(t), t \geq 0)$. For $i \neq j \in E$ define $N^n(i,j) = \{N_t^n(i,j), t \geq 0\}$, $Y^n(i) = \{Y_t^n(i), t \geq 0\}$ and $\hat{B}^n(i,j) = \{\hat{B}_t^n(i,j), t \geq 0\}$ by

$$N_t^n(i,j) = \sum_{k=1}^{n} \sum_{0<s\leq t} 1(X_s^k = j, X_{s-}^k = i), \quad Y_t^n(i) = \sum_{k=1}^{n} 1(X_{t-}^k = i) \quad \text{and}$$

$$\hat{B}_t^n(i,j) = \int_0^t (Y_s^n(i))^+ dN_s^n(i,j)$$

and put $\hat{B}(i,i) = -\sum_{j\neq i}\hat{B}(i,j)$. The matrix valued process $\hat{B}^n = (\hat{B}^n(i,j), i,j \in E)$ is the martingale estimator of the cumulative rate matrix B and is used to construct a product-limit estimator of P. For $0 \leq s \leq t \leq T$ define the product-limit estimator $\hat{P}$ by

$$\hat{P}(s,t) = \prod_{s<u\leq t} (I + \Delta\hat{B}_u^n).$$

If the product is empty, then define $\hat{P}(s,t) = I$, the identity matrix. The estimator $\hat{P}$ is an empirical transition probability matrix which satisfies the Chapman-Kolmogorov equation and is the proposed estimator for discrete and continuous-time Markov chains.

For $i \neq j \in E$ define $\tilde{B}^n(i,j) = \{\tilde{B}_t^n(i,j), 0 \leq t \leq T\}$ by

$$\tilde{B}_t^n(i,j) = \int_0^t 1(Y_s^n(i) > 0)q_{ij}(s)\mu\{ds\}$$

211

and $\tilde{B}^n(i,i) = -\Sigma_{j\neq i}\tilde{B}^n(i,j)$. Let $\tilde{B}^n = (\tilde{B}^n_t(i,j), 0 \leq t \leq T, i,j \in E)$ and define the process $\tilde{P}^n$ by the product integral

$$\tilde{P}(s,t) = \prod_{s<u\leq t} (I + \frac{d\tilde{B}^n}{d\mu}(u)\mu\{du\}) \quad , \quad 0 \leq s \leq t \leq T.$$

According to theorem 3.1, Aalen and Johansen (1978) the following integral equation is valid

$$M^n_t = (\hat{P}(0,t)\tilde{P}^{-1}(0,t) - I) = \int_0^t \hat{P}(0,s-)(d\hat{B}^n_s - d\tilde{B}^n_s)\tilde{P}^{-1}(0,s) \quad , \quad 0 \leq t \leq T$$

where $M^n = \{M^n_t, 0 \leq t \leq T\}$ is a matrix-valued process whose ijth element is the sum of terms of the form

$$\int_0^t \hat{P}_{ik}(0,s-)(d\hat{B}^n_s(k,m) - d\tilde{B}^n_s(k,m)\tilde{P}^{mj}(0,s).$$

It turns out that M^n is a martingale and this fact is key to proving the asymptotic properties of $\hat{P}$ (cf. equation (4.2) for $\hat{G}^n$).

The estimator $\hat{P}$ is uniformly consistent over $[0,T]$ and the normalized difference $n^{1/2}(\hat{P} - P)$ converges weakly to a matrix-valued Gaussian process of independent increments as $n \to \infty$. This is proved by Aalen and Johansen (1978) and Phelan (1986b) in the continuous and discrete-time setting, respectively. Their treatment is general enough to allow for general patterns of censoring.

In closing this subsection we pose the problem of estimating the sojourn-time distribution G_i for each $i \in E$. This is a problem of estimating a family of life-distributions. In fact a product-limit estimator of G_i can be constructed from $\hat{B}^n(i,i)$ (see Aalen and Johansen (1978)) and may be studied according to the methods of section 4.1.

5. DISCUSSION. In this paper we have surveyed some estimation problems in life-testing and Markov chain analysis involving Poisson type counting processes. Our discussion underscores the importance of martingale theory and the product-limit estimator in providing for a unified theory and methodology.

Other inference problems, such as setting confidence bands, hypothesis testing and comparison of sub-populations, are covered by some of the references cited herein.

212

REFERENCES

[1] Aalen, O. (1978). Non-parametric inference for a family of counting processes. Ann. Statist. 6, 701-726.

[2] Aalen O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand. J. Statist. 5, 141-150.

[3] Andersen, P.K. (1985). Statistical models for longitudinal labour market data based on counting processes. In longitudinal analysis of labour market data (ed. J.J. Heckman and B. Singer) Cambridge University Press, New York.

[4] Andersen, P.K. and Borgan, Ø. (1985). Counting process models for life history data: A review. Scand. J. Statist. 12, 97-158.

[5] Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100-1120.

[6] Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1982). Linear non-parametric tests for comparison of counting processes, with applications to censored survival data (with discussion). Int. Statist. Rev. 50, 219-258. Correction 52, 225.

[7] Boel, R., Varaiya, P. and Wong, E. (1975). Martingales on jump processes. I: representation results. Siam. J. Control. 13, 999-1021.

[8] Billingsley, P. (1968). Convergence of Probability Measures. Wiley, New York.

[9] Brémaud, P. (1981). Point Processes and Queues: martingale dynamics. Springer-Verlag, New York.

[10] Breslow, N.E. (1985). Cohort analysis in epidemiology. In A Celebration of Statistics, ISI Centenary Volume, Springer-Verlag, New York.

[11] Cox, D.R. (1972). Regression models and life tables (with discussion) J.R. Statist. Soc. B. 34, 187-220.

[12] Gill, R.D. (1980). Censoring and Stochastic Integrals. Math. Centre Tracts 124, Mathematical Centre Amsterdam.

[13] Gill, R.D. (1981). Testing with replacement and the product-limit estimator. Ann. Statist. 9, 853-860.

[14] Gill, R.D. (1983). Large sample behavior of the product-limit estimator on the whole line. Ann. Statist. 11, 49-58.

[15] Hoem, J.M. (1971). Point estimation of forces of transition in demographic models. J.R. Statist. Soc. B. 33, 275-289.

[16] Jacod, J. (1975). Multivariate point processes: predictable projection, Radon-Nykodym derivatives, representations of martingales. Z. Wahrsch. verw. Geb. 31, 235-253.

[17] Jacod, J. and Mémin, J. (1980). Un nouveau critere de compacite relative pour des suites de processus. Sem. Prob. Rennes. 79, 1-27.

[18] Jacod, J., Klopotowski, A. and Mémin, J. (1982). Théoréme de la limite centrale et convergence fonctionelle vers un processus à accroissments independents: la méthode des martingales. Inst. Henri Poincaré, 18, 1-45.

[19] Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. J. Am. Statist. Ass. 53, 457-481.

[20] Lenglart, E. (1977). Relation de domination entre deux processus. Ann. Inst. Henri Poincaré 13, 171-179.

[21] Liptser, R.S. and Shiryaev, A.N. (1978). Statistics of Random Processes II, Springer-Verlag, New York.

[22] Mau, J. (1986). Nonparametric estimation of the integrated intensity of an unobservable transition in a Markov illness-death process. Stoch. Proc. Appls. 21, 275-290.

[23] Métevier, M. (1982). Semimartingales: a Course on Stochastic Processes. Walter de Gruyter, New York.

[24] Nelson, W. (1969). Hazard plotting for incomplete failure data. J. Qual. Technol. 1, 27-52.

[25] Phelan, M.J. (1986a). Life-testing and estimation with arbitrary distribution function. Technical Report 708, Cornell University.

[26] Phelan, M.J. (1986b). Nonparametric estimation from a censored Markov renewal process observed over a long period of time. Technical Report 706, Cornell University.

[27] Phelan, M.J. (1986c). Nonparametric estimation from discrete-time nonhomogeneous finite Markov chains with applications to inference from multiwave panel data. Technical Report 709, Cornell University.

[28] Prentice, R.L. and Self, S.G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. Ann. Statist. 11, 804-813.

[29] Singer, B. (1981). Estimation of nonstationary Markov chains from panel data. In K.F. Schuessler (Ed.), Sociological Methodology 1981. San Francisco: Jossey-Bass.

[30] Shiryaev, A.N. (1981). Martingales: recent developments, results and applications. Inst. Statist. Rev. 49, 199-234.

[31] Wong, W.H. (1986). Theory of partial likelihood. Ann. Statist. 14, 88-123.

[32] Yashin, A.I. (1985). Dynamics of survival analysis: conditional Gaussian property versus the Cameron-Martin formula. In Steklov Seminar, 1984: Statistics and Control of Stochastic Processes. N.V. Krylov, R.S. Lipster, and A.A. Novikov (eds.).

A HIERARCHICAL MULTISCALE PROCESSING OF IMAGES [*]

B. Gidas [**]
Division of Applied Mathematics
Brown University
Providence, RI 02912

ABSTRACT

We describe a new method for Digital Image Processing. It is based on a combination of Renormalization Group ideas and the Markov Random Field modeling of images. It provides a unifying procedure for performing a hierarchical, multiscale, coarse-to-fine analysis of image processing tasks such as restoration, texture classification, coding, motion analysis, etc. The method has been tested by a number of computer experiments. We report here two restoration experiments.

I. The Method.

Image processing problems (restoration, segmentation, texture classification, compression and coding, motion analysis, photomosaics, etc.), and Robotics Vision (automatic object recognition), deal with cooperative features that exist and interact on a large number of length scales - from the microscopic features of texture consisting of elementary "grains" to the macroscopic features characterizing large scale objects. Such multiscale, interdependent features appear in all situations of practical interest: Images obtained from aircrafts, various types of satellite data, thermal

---

images, robot vision fields, photon emission tomography and scans from nuclear magnetic resources, etc.

Our method [3] processes images in a <u>multiscale</u>, <u>coarse-to-fine</u>, <u>hierarchical</u> fashion. It is based on a probabilistic modeling of images (a Bayesian approach using Gibbs distributions [2]), and Renormalization Group ideas from Statistical Physics and Quantum Field Theory [6]. The method is highly parallel and efficiently implementable on parallel architectures. The procedure generates a (vertical) <u>cascade</u> of images from a given image. The top level of the cascade is the original image, while the bottom level of the cascade contains only the largest scale features of the original image. Each intermediate level represents features of length scale larger than the length scale of levels below it, and smaller length scale than the levels above it.

The method consists of two major stages, the <u>Renormalziation stage</u> and the <u>Processing stage</u>. The general formulation of the method with a number of computer experiments can be found in [3]. Here we present a simple form of the procedure (and two restoration experiments). We describe first the Renormalization stage: Given a $2^N \times 2^N$ image $L^{(0)}$ (to be, for example, restored, segmented, or coded), we construct a sequence of M≤N images $L^{(k)}$ of size $2^{N-1} \times 2^{N-k}$, $k = 1, ..., M$ (see Figure 1; here the cascade appears horizontal rather than vertical). The original image $L^{(0)}$

$L^{(0)}$ $L^{(1)}$ $L^{(2)}$
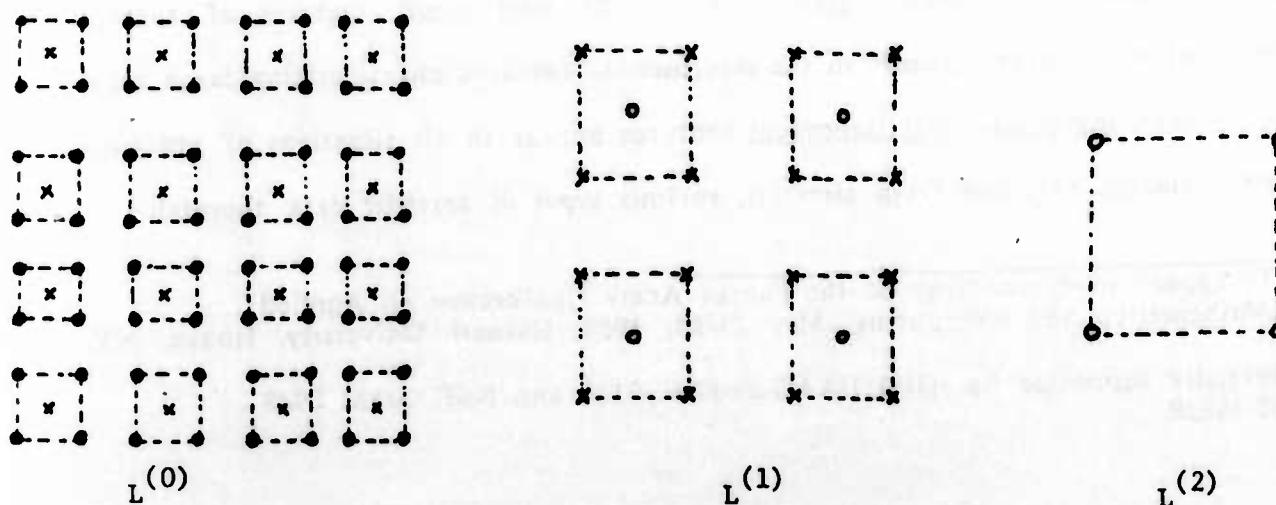
Figure 1: Cascade levels

216

has the _finest_ _grid_ (of lattice spacing, say, 1). Each lower image $L^{(k)}$, $k = 1$,..., M, has a _coarse_ _grid_ of lattice spacing 2^k. Each level $L^{(k)}$ is obtained from the previous level $L^{(k-1)}$ by dividing $L^{(k-1)}$ into 2×2 disjoint _cells_ and identifying each cell by its center. The set of these centers constitute the pixels of the level $L^{(k)}$. In Figure 1, the dotted squares of $L^{(0)}$ are centered at crosses which become the pixels of $L^{(1)}$. The dotted squares of $L^{(1)}$ are centered at circles which become the pixels of $L^{(2)}$, and so on. _One should think of each_ $L^{(k)}$, _k_ _= 1_ _,..., M, as being the_ _original image viewed from larger and larger distances._

Each image $L^{(k)}$, $k = 0$,..., M, is associated with a Gibbs distribution $P^{(k)}$. $P^{(0)}$ is the _prior_ or _posterior_ distrbution of $L^{(0)}$, depending on whether $L^{(0)}$ is undergraded or degraded. The distribution $P^{(0)}$ is estimated from the given data (and the degradation characteristics, if the data are degraded). The distribution $P^{(1)}$ is obtained from $P^{(0)}$ via a Renormalization Group transformation R. Similarly, $P^{(2)}$ is obtained from $P^{(1)}$ via R, and so on. At each level $k = 1$,..., M, the image $L^{(k)}$ together with the renormalization group transformation R _preserve all_ the information contained in the original image $L^{(0)}$.

The renormalziation group transformation R is specified in terms of certain _conditional probabilities_ Q as follows: Consider the ith-cell of level k-1 (see Figure 2). Let $x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$, $x_i^{(4)}$, be the gray levels at the four pixels of the ith-cell
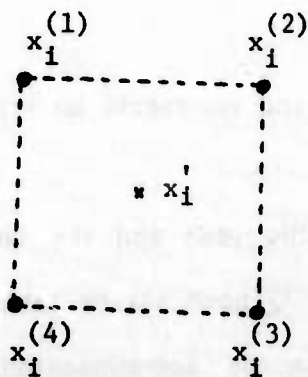
Figure 2: The ith-cell of (k-1)-level

of $L^{(k-1)}$ (if the original image is degraded, the x_i's are unknown). The center of the ith-cell (denoted by a cross in Figure 2), will become the ith pixel of the k-level. The gray level x_i' of this pixel in $L^{(k)}$ is chosen randomly from a conditional probability $Q(x_i' | x_i^{(1)}, ..., x_i^{(4)})$. An example of such a conditional probability is

$$Q(x_i' | x_i^{(1)}, ..., x_i^{(4)}) = \frac{\exp[\rho x_i'(x_i^{(1)} + ... + x_i^{(4)})]}{\sum\limits_{\{x_i^{(1)}, ..., x_i^{(4)}\}} \exp[\rho x_i'(x_i^{(1)} + ... + x_i^{(4)})]} \tag{1}$$

where ρ is an arbitrary parameter. If the gray levels are binary, i.e., $x_i = \pm 1$, then taking $\rho \to +\infty$ in (1), we obtain the "majority rule": If the majority of the $x_i^{(\alpha)}$'s is ± 1, then x_i' is ± 1, respectively. If there is tie among the $x_i^{(\alpha)}$'s, then x_i' is chosen +1 with probability ½. Let $x = \{x_i : i \in L^{(k-1)}\}$ and $x' = \{x_i' : i \in L^{(k)}\}$, be the gray level configurations of $L^{(k-1)}$ and $L^{(k)}$, respectively. Then the Renormalization Group transformation R is defined by

$$P^{(k)}(x') = R P^{k-1} = \sum_{\{x\}} R(x' | x) P^{(k-1)}(x) \tag{2}$$

where

$$R(x'/x) = \prod_{i \in L^{(k)}} Q(x_i' | x_i^{(1)}, ..., x_i^{(4)}) . \tag{3}$$

If the gray levels x_i are continuous, the sums in (1) and (2) should be replaced by integrals.

In general, there is [3] a freedom in choosing the cells and the conditional probabilities Q (i.e., the cells need not be squares, and Q need not be taken of the form (1)). Any a priori knowledge about $L^{(0)}$ can be accommodated in the

218

modeling of $L^{(0)}$, as well as in the choice of the cells and the conditional probabilities Q.

The above stage of constructing the <u>renormalized</u> Gibbs distributions $P^{(k)}$, $k = 1 ,..., M$, from $P^{(0)}$, is the <u>Renormalization stage</u> of our procedure (in this stage we start from the top level $L^{(0)}$ of the cascade and proceed towards the bottom level $L^{(M)}$ of the cascade). Next comes the <u>Processing stage</u> of our procedure. In imaging tasks such as restoration, texture classification, coding, etc., we start our processing from the bottom of the cascade. That is, we first process the coarsest-grid image $L^{(M)}$ (which contains large scale features only, and has very few degrees of freedom). Then we go upwards. We transmit the processed information from level M to level M-1, and process the <u>new</u> (smaller scale) features which appear in $L^{(M-1)}$ but <u>not</u> in $L^{(M)}$. We continue the process until we reach $L^{(0)}$, and thus process all the fine details of $L^{(0)}$.

During the kth step of the processing stage (i.e., in going from $L^{(k)}$ to level $L^{(k-1)}$) the number of <u>possible intensity images at the (k-1)-level constrained by the processed information at</u> the k-level is much smaller than the number of <u>all</u> intensity images at (k-1)-level <u>without</u> any constraint. This reduces drastically the number of computational steps needed to determine the (k-1)-level. This multiscale, coarse-to-fine processing of images, results to a rapid convergence and reduction of the computational cost.

The present approach to image processing problems is reminescent to the pyramid structures [1] and to the multi-grid method in partial differential equations [4,5]. However, our procedure is fundamentally different from these schemes, as are its most important properties.

In restoration problems, we often combine the above procedure with the annealing algorithm: The posterior distribution $P^{(0)}$ of $L^{(0)}$ depends on the

temperature T, i.e., $P^{(0)} = P_T^{(0)}$. Here the T-dependence enters through $\frac{1}{T} H$, where H is the underlying energy function. The T-dependence of the subsequent (renormalized) distributions $P_T^{(k)}$ is more complicated. The bottom level M of the cascade (which has very few degrees of freedom) can be restored by applying the annealing algorithm to $P_T^{(M)}$ (quite often, however, this level can be restored by a simple determination of the lowest energy). Having restored a level k (k = M, M-1 ,..., 1), we restore the (k-1)-level, by applying the annealing algorithm to the conditional probability

$$P_T(x^{(k-1)}|x^{(k)}) \equiv R(x^{(k)}|x^{(k-1)}) \frac{P_T^{(k-1)}(x^{(k-1)})}{P_T^{(k)}(x^{(k)})} . \tag{4}$$

where $x^{(k)}$ denotes the gray intensities of the k-level (already restored), and $x^{(k-1)}$ the gray intensities of the (k-1)-level (to be restored).

At each level of the restoration stage (i.e., in going from $L^{(k)}$ to $L^{(k-1)}$), we choose an <u>annealing schedule</u> of the form

$$T(t) = \frac{T_0}{1 + . \log t} , \quad t = 1, 2, \dots . \tag{5}$$

The initial temperature T_0 need not be the same at all cascade levels. In fact, choosing T_0 to increase as we move from the bottom of the cascase (coarse grid) toward the top of the cascade (fine grid), the algorithm is somewhat faster. In our experiments (Section II), we chose T_0 to be the same at all cascade levels. However, this T_0 is in general smaller than the T_0 needed for a direct annealing of the fine grid level $L^{(0)}$ only. There is a theoretical justification of this fact: the renormalization group "trajectories" [6] move towards the trivial zero-temperature "fixed point" as $T(t) \to 0$. Also, the overall convergence of the present procedure is

(in general) faster than the convergence of annealing applied directly to the fine grid distribution $P_{T(t)}^{(0)}$.

II. Experiments

The image processing method outlined in Section I, has been tested by a number of computer experiments [3]. Figures 3 and 4 show two restoration experiments.

Figure 3

The original signal (a) is a binary, "handrawn" signal with 1025 pixels. The degraded data (b) were obtained by adding a Gaussian noise of mean zero and variance $\sigma^2 = 1$. $(c_7 - c_0)$ represent eight restoration levels. Notice that the small pieces at the center and end of the original signal, do not appear until level (c_3). These pieces have a length scale smaller than the length scale of the "features" contained in $(c_7) - (c_4)$.

In this example, equation (2) can be solved exactly. The resulting algorithm is deterministic (i.e., no annealing or stochastic relaxation is needed), and extremely efficient.

Figure 4

The original image (a) is a 64×64 binary image. It was generated by the "spin-flip" algorithm. The degraded data (b) was generated by adding a Gaussian noise of mean zero and variance $\sigma^2 = 5$. $(c_2) - (c_0)$ represent three restoration levels: (c_2) 16×16, (c_1) 32×32, and (c_0) 64×64. At each cascade level we used the annealing algorithm (applied to (4)) with an initial temperature $T_0 = 1.5$, and performed five sweeps per level. For comparison, we restored the degraded image (b) by applying the annealing algorithm directly at the top level (64×64). With an initial temperature $T_0 = 3$ and 100 sweeps, the result of the annealing was not as good as the result of our procedure.

The final restoration (c_0) of the present procedure although satisfactory, contains some noise at the boundaries of the various regions. This noise could be eliminated by using (c_0) as the initial configuration of a deterministic descent algorithm. Since (c_0) is very near to the true "global minimum", any deterministic descent algorithm starting (c_0) would reach the true global minimum in a small number of iterations.

References

1. Burt, P.J.: "The Pyramid as a Structure for Efficient Computation", in Multiresolution Image Processing and Analysis, Springer-Verlag (1984), ed.: A. Rosenfeld.

2. Geman, S. and D. Geman: "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", IEEE Trans. PAMI-6 (1984), 721-741.

3. Gidas, B.: "A Renormalization Group Approach to Image Processing Problems", preprint, Brown University 1986 (submitted for publication).

4. Hackbush, W.: Multi-Grid Methods and Applications, Springer-Verlag, 1986.

5. Terzopoulos, D.: "Multilevel computational processes for visual surface reconstruction", Comp. Vision Graphics Image Process 24 (1983), 52-56.

6. Wilson, K. and J. Kogut: "The Renormalization Group and the E-expansion", Phys. Reports C12 (1974), 75-200.

Figure 3

ORIGINAL SIGNAL

(a)

NOISY DATA

(b)

RESTORATION 9 POINTS

(c_7)

RESTORATION 17 POINTS

(c_6)

RESTORATION 33 POINTS

(c_5)

RESTORATION 65 POINTS

(c_4)

223

Figure 3 (continued)

RESTORATION 129 POINTS

(c₃)

RESTORATION 257 POINTS

(c₂)

RESTORATION 513 POINTS

(c₁)

RESTORATION 1025 POINTS
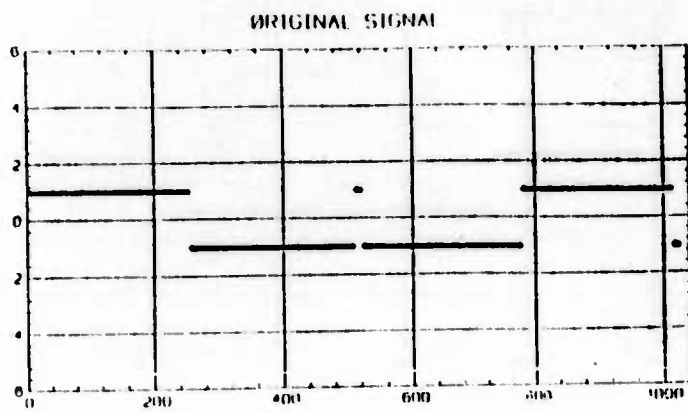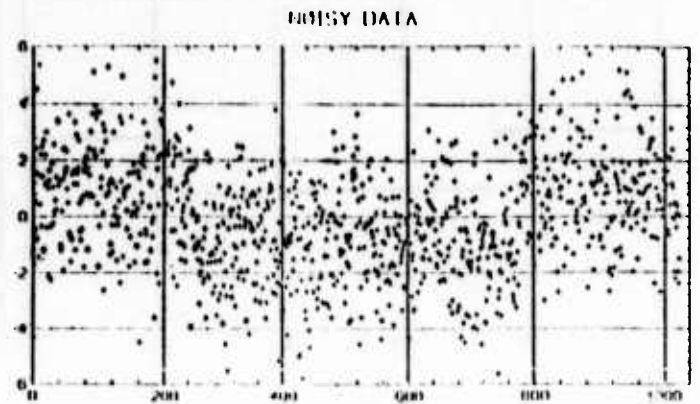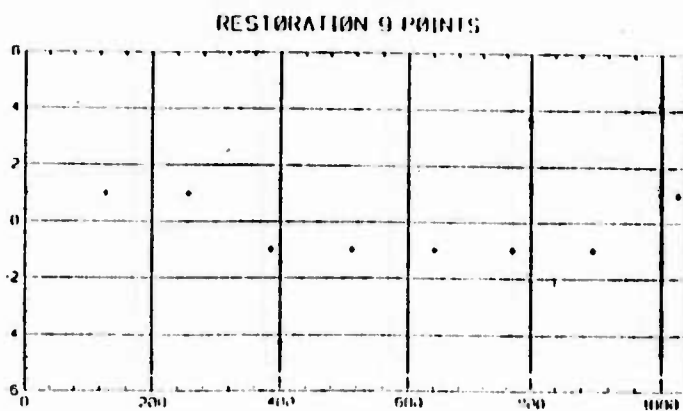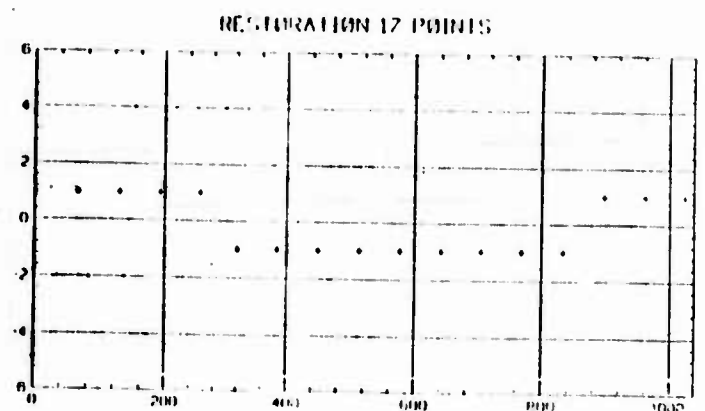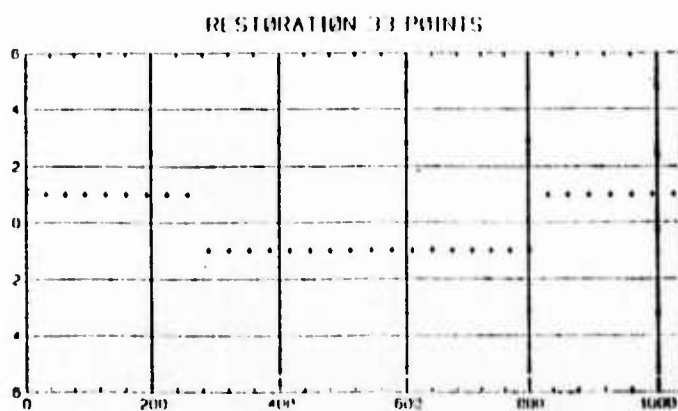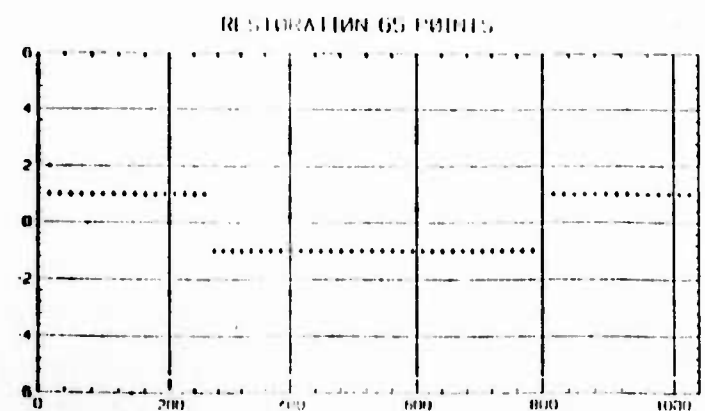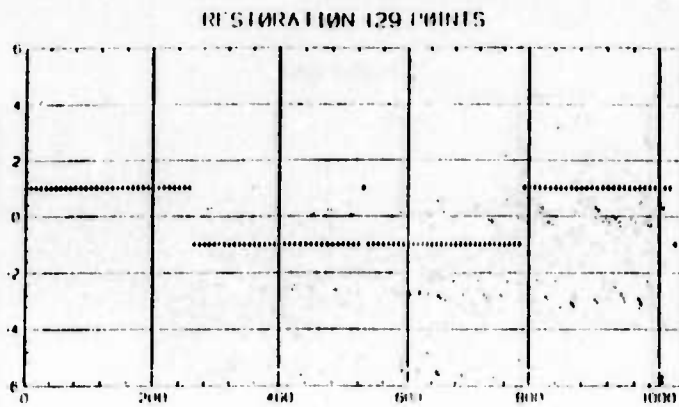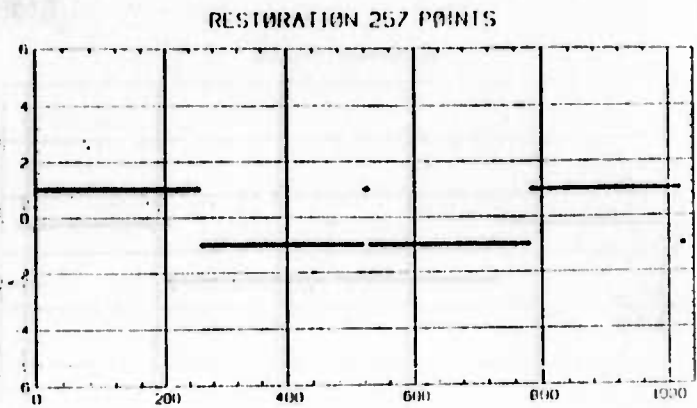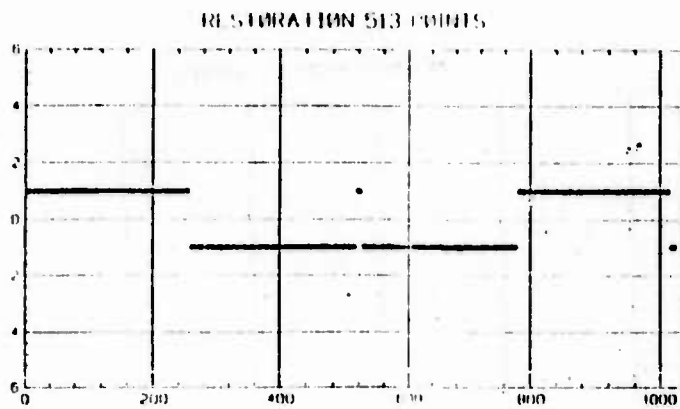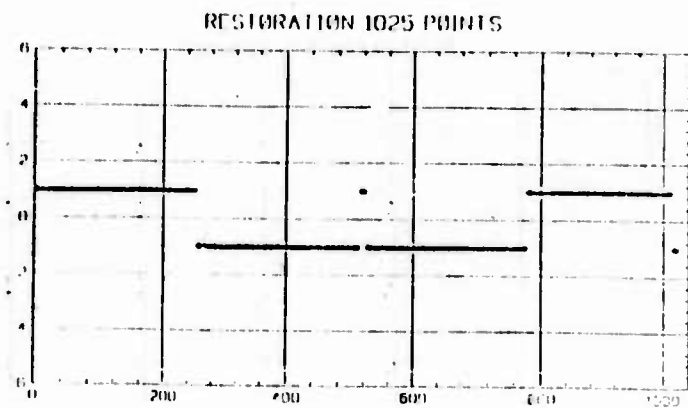
(c₀)
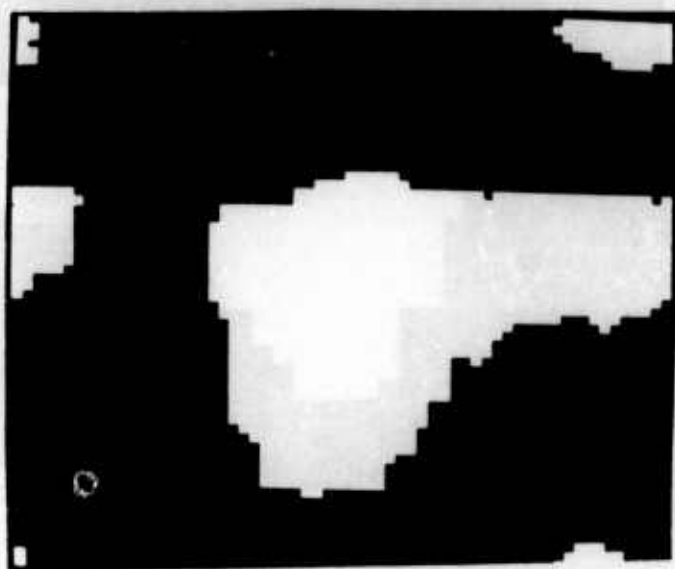
Figure 4

(a)

(b)
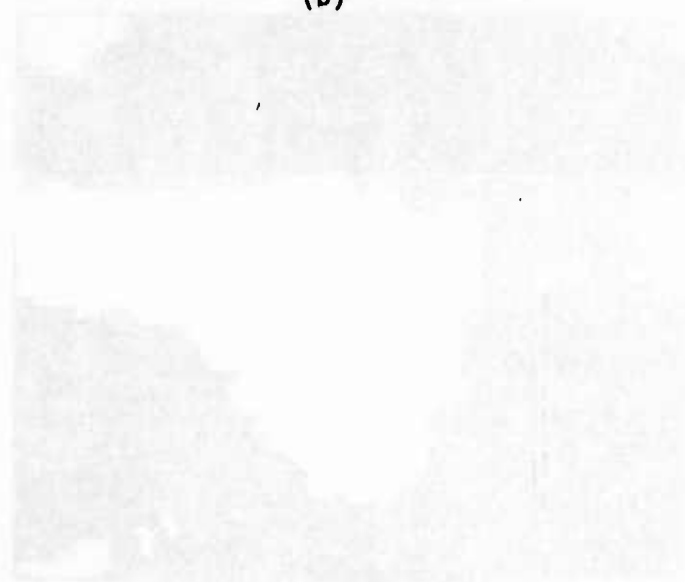
Figure 4 (Continued)

(c$_2$)

(c$_1$)

(c$_0$)

226
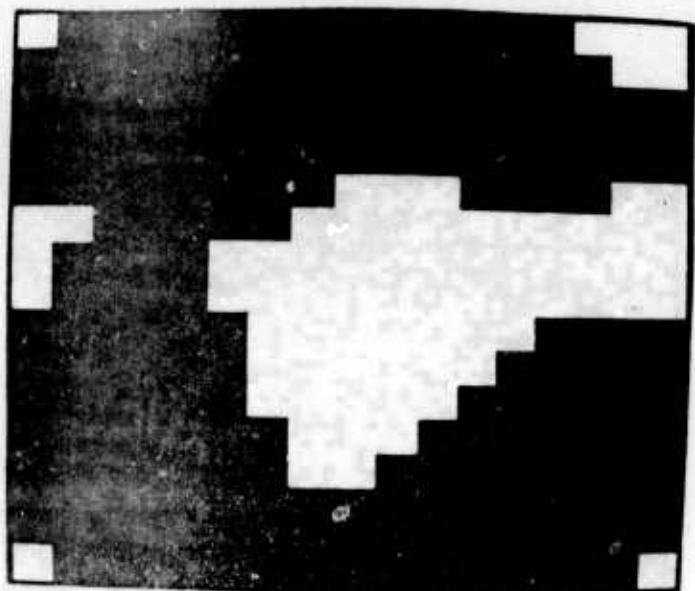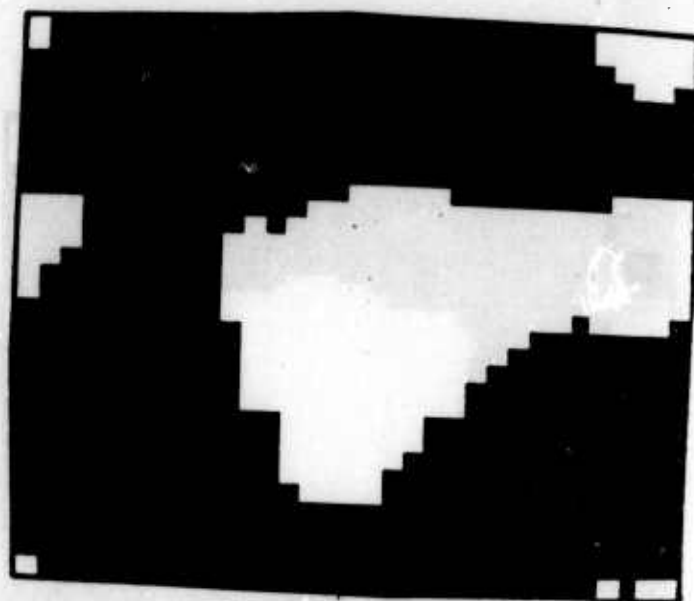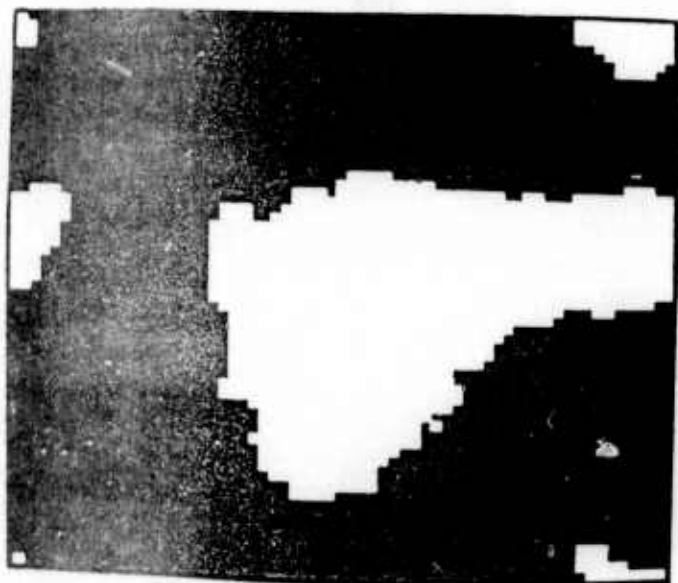
A Maximum Entropy Method for Expert System Construction

ALAN LIPPMAN
Division of Applied Mathematics
Brown University
Providence, R.I. U.S.A. 02912

Abstract   We consider the maximum entropy method of expert system construction. We show that the construction of the expert system is equivalent to the minimization of a convex function in as many dimensions as there were pieces of knowledge supplied the system. We show that in the case where the knowledge presented the system is self-contradictory, the minimization of this function creates an expert system for a set of constraints that is consistent and 'close' to the original inconsistent constraints. Monte Carlo methods for minimizing the function are discussed, and illustrated by computer experiment. One of the examples given suggests an approach to the problem of invariant optical character recognition.

Introduction   An expert system is designed to answer questions. We consider probabilistic expert systems — if the system is given an event, it should be able to calculate its probability. Such an expert system is actually a distribution on the set of all events we wish to consider. Typically the knowledge the system is based on will be insufficient to answer all questions. In many cases we wish to consider, the sheer size of the state space precludes such knowledge. A medical expert system could be asked for the probability of a disease given some combination of symptoms, yet the set of all possible combinations of symptoms is huge, and the knowledge base can not be expected to contain all the different probabilities. We desire our system to answer questions even in such cases, and to do so in a reasonable manner, much like a human expert would. For this purpose we consider 'the principle of maximum entropy'. Of all the distributions which satisfy the knowledge supplied the system, we will pick the one with maximum entropy to be our expert system. The entropy H of a distribution p is defined as

$$H(p) \quad = \quad -\sum_{\omega \in \Omega} p(\omega) \log p(\omega)$$

where ω is an event, Ω is the set of all events, and $p(\omega)$ is the probability of the event ω. Entropy has an information-theoretic meaning; the distribution with maximum entropy can be viewed as the one containing the least knowledge. By maximizing the entropy over all distributions that agree with the knowledge base, we are picking as our expert system the distribution that makes the fewest unnecessary 'assumptions'. For more information regarding the justification of the principle of maximum entropy, we refer the reader to [1].

1. Knowledge   The construction of a probabilistic expert system begins with knowledge. We classify as knowledge anything that answers probabilistic questions; we think of a probabilistic question as a function of probabilities and we consider an answer to be the value we say the function will take on. (This brings up a more general way to view knowledge; we could view an answer as specifying that the function, that defines the question, has a value in a certain range. We will not be using this type of answer.) Using these ideas, we see that the knowledge supplied the system can be broken up into distinct 'pieces' of knowledge, each of which corresponds to a distinct probabilistic question and its answer. Each piece of knowledge is a constraint that must be satisfied by the expert system; if the answer to the question we ask the system is included in the knowledge base, then the expert system's answer is constrained to duplicate it. We can write a constraint in its most general form as

$$B(p) \quad = \quad c$$

We consider a restriction of the above to the case where B is a linear function of p, our constraint can thus be written as

$$\sum_{\omega \in \Omega} b(\omega)p(\omega) = c$$

The above constraint is the same as specifying that the expected value of b is c. Since $c = c\sum_{\omega \in \Omega} p(\omega)$, we consider the function a, where $a(\omega) = b(\omega) - cp(\omega)$, and we re-write the above constraint as

$$\sum_{\omega \in \Omega} a(\omega)p(\omega) = 0$$

It may seem that this form is very restricted, but it is sufficient for several important types of constraints ([3],[4]). It is capable of representing any piece of knowledge that can be put in terms of the expected value of a function; it can thus represent knowledge about marginal, joint and conditional probabilities. To illustrate this consider the following example:

$$p(\omega \in S_1 | \omega \in S_2) = .5$$

Using Bayes's rule, we can re-write the above as

$$\frac{p(\omega \in S_1 \cap S_2)}{p(\omega \in S_2)} = .5$$

This can be written as

$$p(\omega \in S_1 \cap S_2) - .5p(\omega \in S_2) = 0$$

which is the same as

$$\sum_{\omega \in \Omega} \left(\chi_{S_1 \cap S_2}(\omega) - .5\chi_{S_2}(\omega) \right) p(\omega) = 0,$$

where χ_S denotes the indicator function on the set of events in S; when $\omega \in S$ we will have $\chi_S(\omega) = 1$, when $\omega \notin S$ we will have $\chi_S(\omega) = 0$.

1. **Lagrange Multipliers**   Recall our goal. We wish to find a distribution that satisfies a set of constraints, and has higher entropy than any other such distribution. Using the form for knowledge that we introduced in the previous section, we can state the problem as follows:

$$\max \left(- \sum_{\omega \in \Omega} p(\omega) \log p(\omega) \right)$$

over all p satisfying   (1)

$$\sum_{\omega \in \Omega} a_i(\omega)p(\omega) = 0 \qquad i = 1, \ldots m$$

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

$$p(\omega) \geq 0 \qquad \omega \in \Omega$$

With suitable care, we can use Lagrange multipliers to reduce the above, constrained, problem to an unconstrained problem. In order to apply the theory of Lagrange multipliers

228

we must add some assumptions. For the complete details, we refer the reader to [2]; here we will note that the required assumptions are that

$$\{a_i(\cdot)\} \quad \text{are linearly independent vectors.}$$

$$\exists p(\omega) \quad \text{such that} \quad \sum_{\omega \in \Omega} a_i(\omega) p(\omega) = 0 \forall i \quad \text{and} \quad p(\omega) > 0 \forall \omega \in \Omega \qquad (2)$$

The Lagrangian is

$$\sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad + \quad \sum_{i=1}^{m} \lambda_i \left(\sum_{\omega \in \Omega} a_i(\omega) p(\omega) \right) \quad + \quad \delta \left(\sum_{\omega \in \Omega} p(\omega) - 1 \right)$$

We know from Lagrange multiplier theory that there exist $\bar{\lambda} = \{\bar{\lambda}_1, \ldots, \bar{\lambda}_m\}$ and $\bar{\delta}$ such that the derivative of the Lagrangian (with respect to λ_i, δ and $p(\omega)$) at $\bar{\lambda}, \bar{\delta}$ is zero, and that such $\bar{\lambda}, \bar{\delta}$ define local extrema. Performing some algebraic manipulations, we arrive at the following equations

$$p(\omega) \quad = \quad \exp\left(-\sum_{i=1}^{m} \bar{\lambda}_i a_i(\omega) \right) \Big/ \sum_{\hat{\omega} \in \Omega} \exp\left(-\sum_{i=1}^{m} \bar{\lambda}_i a_i(\hat{\omega}) \right) \qquad \forall \, \omega \in \Omega$$

$$\frac{\partial}{\partial \lambda_k} \sum_{\omega \in \Omega} \exp\left(-\sum_{i=1}^{m} \lambda_i a_i(\omega) \right) \Big|_{\lambda = \bar{\lambda}} \quad = \quad 0 \qquad k = 1, \ldots, m \qquad (3)$$

The function $\sum_{\omega \in \Omega} \exp(-\sum_{i=1}^{m} \lambda_i a_i(\omega))$ will be denoted $Z(\lambda)$, where $\lambda = \{\lambda_1, \ldots, \lambda_m\}$. The function Z is sometimes called the partition function.

Notice that Z is a convex function. Hence there is at most one $\bar{\lambda}$, corresponding to the global minimum of Z, at which Z has an extremal point (*i.e.*, $\frac{\partial}{\partial \lambda_i} Z(\lambda)|_{\lambda = \bar{\lambda}} = 0 \forall i$).

Under the assumptions (2) we know that such a $\bar{\lambda}$ must exist, hence the maximum entropy distribution exists and is unique. A interesting property of Z is that when the assumptions (2) do not hold, Z has no extremal point (see [2] for the details). Hence, if we try to minimize Z and succeed, we have found the maximum entropy distribution (since the maximum entropy distribution is defined, through (3), by the $\bar{\lambda}$ at which the minimum occurs). Our computational goal (section 4) will thus be to minimize Z. We note that there have been many ideas and methods for the computation of the maximum entropy distribution, some involving Lagrange multipliers, others not; some examples are [1],[5]-[7]. The method we use is based on work by Geman[3] and Geman[4].

3. Contradictions   Let us consider the case where the assumptions (2) do not hold. We will still assume that the a_i are linearly independent. a_i will usually be a simple function of ω (for example a_i is often a combination of indicator functions), in such cases independence is relatively easy to verify. If the constraints are dependent, some can be removed so as to provide independence. Hence, the restriction that the a_i be independent is often easy to satisfy.

More hazardous is the assumption that

$$\exists p(\omega) \quad \text{such that} \quad \sum_{\omega \in \Omega} a_i(\omega) p(\omega) = 0 \forall i \quad \text{and} \quad p(\omega) > 0 \forall \omega \in \Omega$$

This assumption can fail in two fundamentally different ways. The first occurs when there exists distributions p that satisfy the constraints (so $\sum_{\omega \in \Omega} a_i(\omega) p(\omega) = 0$), but all such

distributions assign probability zero to some events. The other way this assumption can fail is when there exists no p that satisfies the constraints. This is the case when the constraints are self-contradictory. In both the above situations we can show that trying to minimize the partition function (i.e., driving the gradient of Z close to zero) and using the form for the probabilities found by the use of Lagrange multipliers (3), does something useful.

The original constraints we supplied the system with were

$$\sum_{\omega \in \Omega} a_i(\omega)p(\omega) = 0 \qquad i = 1, \dots, m \qquad (4)$$

Consider the system of constraints

$$\sum_{\omega \in \Omega} a_i(\omega)p(\omega) = \epsilon_i \qquad i = 1, \dots, m \qquad (5)$$

When $\|\epsilon\| = (\epsilon_1^2 + \dots \epsilon_m^2)^{\frac{1}{2}}$ is small enough, we would expect the two systems of constraints to be interchangeable. Now, let p_λ be defined as follows

$$p_\lambda(\omega) = \exp\left(-\sum_{i=1}^m \lambda_i a_i(\omega)\right) \Big/ Z(\lambda)$$

where Z is defined with respect to the constraints (4). If Z has an extrema at $\bar{\lambda}$ then $p_{\bar{\lambda}}$ is the maximum entropy distribution for the constraints (4). For any λ, we can show (see [2]) that p_λ is the maximum entropy distribution for the system of constraints

$$\sum_{\omega \in \Omega} a_i(\omega)p(\omega) = \sum_{\omega \in \Omega} a_i(\omega) \exp\left(-\sum_{i=1}^m \lambda_i a_i(\omega)\right) \Big/ Z(\lambda) = -\nabla_i Z(\lambda) \Big/ Z(\lambda)$$

where ∇_i is the ith component of the gradient. So, if, for a given λ, $\|\nabla Z(\lambda)/Z(\lambda)\|$ is small (corresponding to $\|\epsilon\|$ being small in (5)), then p_λ is the maximum entropy distribution for a system of constraints that is close to the original system of constraints.

Hence, our desire is to find a λ such that $\|\nabla Z(\lambda)/Z(\lambda)\|$ is small. In light of this, let us examine the cases where the assumptions (2) do not hold. When a system of constraints has as its only solutions distributions p that assign probability zero to some events, we can show (see [2]) that $Z(\lambda)$ is bounded below by 1. Hence, all we need to do is make the gradient of Z arbitrarily small, and we will. have found a λ that defines a maximum entropy distribution which satisfies constraints arbitrarily close to those originally supplied. When the constraints are contradictory, we can show (see [2]) that when ∇Z goes to zero, Z will also. But, we can also show (see [2]) that using a continuous gradient descent method (define $\lambda(t)$ by the O.D.E. $d/dt\lambda_i(t) = -\nabla_i Z(\lambda(t))/\|\nabla Z(\lambda(t))\|$, with the initial condition $\lambda_i(0) = 0 \forall i$) to minimize Z yields a path $\lambda(t)$ such that $\|\nabla Z(\lambda(t))/Z(\lambda(t))\|$ decreases as t increases. In this sense, we get a maximum entropy distribution for a consistent set of constraints that approximates the inconsistent set.

4. Minimizing the Partition Function   In this section we consider the computational side of finding a maximum entropy distribution. Recall that finding the maximum entropy distribution is equivalent to minimizing a convex function, the partition function $Z(\lambda)$, as we showed in section 2. Recall also that $Z(\lambda)$ and $\nabla Z(\lambda)$ are defined by sums over all elements in Ω. When Ω has a small number of elements, computation is simple. The gradient of Z

can be calculated exactly, and Z minimized by gradient descent. However, even in a small letter recognition problem, for example, we may have our letters described by ten features, each feature being able to take on thirty different values. This yields a state space with 30^{10} elements, and a sum of 30^{10} terms (each of which involves the exponential of a sum of m terms), as would be necessary to explicitly compute ∇Z, is beyond the practical limits of computation. This difficulty can be overcome by estimating the direction of ∇Z, instead of calculating it exactly. Before we delve too deeply (for more details see [2]), let us first outline the general idea. The crucial observation is that we can find a distribution p_λ, such that $\nabla Z(\lambda)/Z(\lambda)$ is just an expected value (with respect to the distribution p_λ) of some simple function. Notice that $\nabla Z(\lambda)/Z(\lambda)$ supplies us with both the direction of the gradient (so we can minimize Z by gradient-descent type methods) and also tells us how close we are to satisfying our constraints (see section 3). Since by using Monte Carlo type methods we can simulate such a distribution p_λ, and since the sample mean from a simulation is close to the real expected value, we can actually approximate $\nabla Z(\lambda)/Z(\lambda)$ without doing a size of Ω number of calculations. The idea of using sampling to find the direction of the gradient of Z, was first proposed by Geman [3].

Consider a distribution p_λ on the space Ω where the probability of the event ω, $p_\lambda(\omega)$ is defined, as before, as

$$p_\lambda(\omega) = \frac{\exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)}{\sum_{\omega \in \Omega} \exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)}$$

The expected value of the function $f(\omega)$ with respect to the distribution p_λ is

$$E_\lambda(f) = \sum_{\omega \in \Omega} f(\omega)p_\lambda(\omega) = \frac{\sum_{\omega \in \Omega} f(\omega)\exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)}{\sum_{\omega \in \Omega} \exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)}$$

and for the function $-a_i(\omega)$ we have

$$E_\lambda(-a_i) = \frac{-\sum_{\omega \in \Omega} a_i(\omega)\exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)}{\sum_{\omega \in \Omega} \exp(-\sum_{i=1}^{m} a_i(\omega)\lambda_i)} = \frac{\nabla_i Z(\lambda)}{Z(\lambda)}$$

Since all a gradient descent method needs is the direction of the gradient, we can use the above. Also note that a measure of how close we are, at a certain λ, to satisfying the original constraints is just $\|E_\lambda(a_i)\| = \|\nabla Z(\lambda)/Z(\lambda)\|$ (see section 3).

Now that we have $\nabla Z/Z$ in terms of an expected value we come to the problem of simulation. The goal is to find an ergodic sequence ω^i with marginal distribution p_λ. In this way the sample expected value of $f(\omega)$ using S samples is

$$\frac{1}{S}\sum_{j=1}^{S} f(\omega^j) \tag{6}$$

and for S large this should be close to the true value of $E_\lambda(f)$.

The method we use to find an ergodic sequence requires that Ω have some sort of neighborhood structure, we will thus revise our view of the state space Ω. For the sake of clarity we will consider each event in Ω as the state of a 1-dimensional lattice with N_1 elements. An element ω in Ω will thus be of the form $\omega = \{\omega_1, \ldots, \omega_{N_1}\}$. Furthermore, each component of ω, ω_k will be restricted to N_2^k different values (Ω will thus have $\prod_{k=1}^{N_1} N_2^k$ elements). We note that with a lattice structure Ω can get very large, without much effort.

231

Our ergodic sequence will start at a random point ω^1 in Ω. We will then pick ω^{l+1} given ω^l as follows; we will fix most of the components of ω^{l+1} to be the same as the value for ω^l. The values we don't fix will allow ω^{l+1} to be any element in some subset of Ω, call it T, containing r elements, $\{t_1, \ldots, t_r\}$. We will then randomly pick an element from T, according to the probabilities $p_\lambda(t_i)$, to be ω^{l+1}. This is done as follows: we first calculate $Z(\lambda) \cdot p_\lambda(t_i)$ for every t_i in T, then we randomly (according to a uniform distribution) pick a number between 0 and $Z(\lambda) \cdot \sum_{i=1}^{r} p_\lambda(t_i)$. This randomly chosen number will be between $Z(\lambda) \cdot p_\lambda(t_j)$ and $Z(\lambda) \cdot p_\lambda(t_{j+1})$ for some t_j in T, and we will let $\omega^{l+1} = t_j$. Note that $Z(\lambda) \cdot p_\lambda(t_i)$ equals $\exp \sum_{i=1}^{m} (-a_i(t_i)\lambda_i)$, which just requires an order of m operations to calculate. Since we usually have m less than several hundred, we are in good computational shape. Of course, one has to be carefull when picking T at each step l (i.e., decide which components of ω^l to fix) in order to avoid creating numerical artifacts. This is not too difficult; one approach is to fix components in a random order and with equal likelihood. This method of finding an ergodic sequence is known as Stochastic Relaxation [3] and is closely related to the Metropolis Algorithm [8].

Now let us say a word about the minimization of Z, given that we have estimates for the gradient. We note that finding the gradient is still a computationally difficult task, and hence we desire to use a method that requires the direction of the gradient at as few a number of points as possible. A discrete analog of the continuous gradient descent scheme, suggested in section 3 for handling contradictions, would prove too costly; we will therefore assume that the constraints are not self-contradictory, so any method that drives ∇Z to zero will be acceptable.

We implement a modification of the standard gradient descent method. Typically, gradient descent refers to constant small steps in the direction opposite to the gradient. Instead, to minimize the number of times we need to compute the gradient, we employ a slight modification. We will still move in the direction opposite to the gradient, but the size of the step we take will not be constant. When we begin we will pick a value for our step-size δ (positive). We will always start at $\lambda^0 = 0$, since this corresponds to the uniform distribution on Ω, a logical starting point. At the point λ^i we will find a λ^{i+1} such that $\lambda^{i+1} = \lambda^i - \hat{\delta} \nabla Z(\lambda^i)/Z(\lambda^i)$ where $\hat{\delta}$ is picked as follows. Since $Z(\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i))$ is a convex function of δ its derivative (with respect to δ) can only be zero for at most one δ, which we shall call $\bar{\delta}$. If $\bar{\delta}$ does not exist, then $Z(\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i))$ is a decreasing function of δ, and since $Z(\lambda)$ is bounded below by zero, we would have $\nabla Z(\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i))$ going to zero; so for δ large enough $\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i)$ would define an adequate solution (section 3). When $\bar{\delta}$ does exist, we see that the derivative of $Z(\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i))$ (with respect to δ) is negative for all δ greater than zero and less than $\bar{\delta}$. Hence, picking $\hat{\delta}$ between 0 and $\bar{\delta}$ would yield $Z(\lambda^{i+1})$ less than $Z(\lambda^i)$. However, the closer $\hat{\delta}$ is to $\bar{\delta}$ the smaller $Z(\lambda^{i+1})$ will be. We will pick $\hat{\delta}$ between $\bar{\delta}$ and $\bar{\delta}/\epsilon$ (ϵ around 2) by doing a binary search: if the dot product $\nabla Z(\lambda^i) \cdot \nabla Z(\lambda^i - \hat{\delta} \nabla Z(\lambda^i)/Z(\lambda^i))$ (remember that Z is positive, so the sign of this term is computable even without normalizing) is negative we try $\hat{\delta} = \hat{\delta}/\epsilon$, if positive we try $\hat{\delta} = \hat{\delta}\epsilon$. When $\nabla Z(\lambda^i) \cdot \nabla Z(\lambda^i - \hat{\delta} \nabla Z(\lambda^i)/Z(\lambda^i))$ switches sign from the last $\hat{\delta}$ to the current $\hat{\delta}$, we will have completed our search in the direction $\nabla Z(\lambda^i)$; we will define λ^{i+1} using the $\hat{\delta}$ (choosing from either the current or the last) for which the sign was positive. In this manner we will be sure that $Z(\lambda^{i+1}) < Z(\lambda^i)$. Making ϵ smaller (close to, but above, one) yields higher accuracy, but since our gradients are not exact, and since we would need to find many more gradients, a computationally expensive task, it is not worth it. We save the value of $\hat{\delta}$ that we used last, for the next step, since it is usually of the correct magnitude.

A useful feature of the above method is that it provides a means to test our sampling method. As we increase δ, the dot product $\nabla Z(\lambda^i) \cdot \nabla Z(\lambda^i - \delta \nabla Z(\lambda^i)/Z(\lambda^i))$ should be a

decreasing function. Likewise as we decrease δ it should be an increasing function (keeping δ positive). If this is true for the sampled values of the dot product, we can have more confidence in the sampling method. If it is not, we know that our estimate of the gradient is wrong, in which case we can take appropriate action. We can increase the number of samples we are averaging over (S in (6)), we can start at multiple begining points ($\omega^1, \bar{\omega}^1, \ldots$) and then average over the different trials ($\{\omega^1, \ldots, \omega^l, \bar{\omega}^1, \ldots, \bar{\omega}^l, \ldots\}$), we can discard the first n elements of the series to get rid of the effects of the random starting point, etc... There are many things that can, at the expense of increased computation, be done to improve the accuracy of the sampling method.

The next section is composed of two examples. The first is a test of our simulation methods. We construct a distribution and extract statistics. We then find the maximum entropy distribution. We then calculate $\nabla Z/Z$ exactly, and see that the simulation was successful (since the values for $\nabla Z/Z$ are quite small). The state space for this example is of size 2^{24}, so the exact calculation of ∇Z was quite lengthy.

The second example we consider is the problem of letter recognition. Sample letters were presented and features extracted from them. Statistics of the features conditioned on the letter served as our knowledge. The maximum entropy distribution was found and used to identify the sample letters. Considering the primitiveness of the features the results are encouraging.

5. Results

Example 1: A test of our method

In this section we conduct a test of our simulation methods. We will consider a distribution on a large state space and use the statistics generated by the distribution to form constraints. We use sampling methods to conduct the gradient descent (section 4), and find a point $\bar{\lambda}$ that will serve as our guess for the extremal point of the partition function. We then compute $\nabla Z(\bar{\lambda})/Z(\bar{\lambda})$ exactly. This will serve to tell us how the statistics generated by the distribution generated by $\bar{\lambda}$ differ from the statistics of the original distribution. We will present (on the following pages) the statistics of the original distribution, the estimated value of $\nabla Z(\bar{\lambda})/Z(\bar{\lambda})$ and the true value of $\nabla Z(\bar{\lambda})/Z(\bar{\lambda})$.

We will have as our state space, Ω, the set of all strings of length 24 composed of 1's and -1's, so Ω has 2^{24} elements. We picked this state space so that the exact calculation of ∇Z and Z is possible, although quite lengthy. The distribution $\hat{p}$ we use to generate statistics is

$$\hat{p}(\omega) = \frac{\exp(-\sum_{i=1}^{24} \sum_{k=i}^{24} \omega_i W(i,k)\omega_k - \sum_{i=1}^{24} T(i)\omega_i)}{\sum_{\omega \in \Omega} \exp(-\sum_{i=1}^{24} \sum_{k=i}^{24} \omega_i W(i,k)\omega_k - \sum_{i=1}^{24} T(i)\omega_i)}$$

Where $W(i,j)$ was picked randomly to be either $+c$ or $-c$, and $T(i)$ was picked randomly to be either $+d$ or $-d$. The values of d and c were chosen so that the distribution $\hat{p}$ is neither too flat nor too sharp. We used $c = 1/5$ and $d = 1/2$. Our constraints are the expected values of ω_i and $\omega_i \omega_j$ with respect to $\hat{p}$. Our constraints are thus

$$E(\omega_i) - \sum_{\omega \in \Omega} \omega_i p(\omega) = 0 \quad \text{for all } i$$

$$E(\omega_i \omega_j) - \sum_{\omega \in \Omega} \omega_i \omega_j p(\omega) = 0 \quad \text{for all } i, j \text{ with } i \geq j$$

where $E(\omega_i)$ and $E(\omega_i \omega_j)$, the expected values with respect to $\hat{p}$, were computed exactly.

233

We can write the partition function, $Z(\lambda)$ (where $\lambda = \{\lambda_1, \ldots \lambda_{300}\}$) as

$$Z(\lambda) = \sum_{\omega \in \Omega} \exp\left(\sum_{i=1}^{24} \sum_{k=i}^{24} \lambda_{(i-1)\cdot(25-(i/2))+k-i+1}(\omega_i \omega_k - E(\omega_i \omega_k)) + \sum_{i=1}^{24} \lambda_{276+i}(\omega_i - E(\omega_i))\right)$$

On the computational side of things we used 20 starting points in the sampling. Each sample involved 80 steps, $\{\omega^1, \ldots, \omega^{80}\}$, the last 50 being kept to form the expected value. Each step was composed of randomly dividing the 24 components of the string into six groups (four in each). We then chose a group and, holding the other groups fixed, picked a value for it according to the distribution p_λ (see section 4). We repeated this procedure until each of the six groups had been allowed to vary once.

We note that the computational time taken to conduct all the steps of the gradient descent (involving the estimation of $\nabla Z/Z$ several hundred times) was less that that needed to do one exact computation of $\nabla Z/Z$.

Recalling that $\bar{\lambda}$ was our estimate, we have (see section 3 and 4)

$$E_{\bar{\lambda}}(\omega_i \omega_k) = E(\omega_i \omega_k) - \nabla_{(i-1)\cdot(25-(i/2))+k-i+1} Z(\lambda)/Z(\lambda)$$
$$E_{\bar{\lambda}}(\omega_i) = E(\omega_i) - \nabla_{276+i} Z(\lambda)/Z(\lambda)$$

$E_{\bar{\lambda}}$ being the expected value under the distribution generated by $\bar{\lambda}$. The percent error in $E_{\bar{\lambda}}$ is

$$\frac{\text{the true value of } \nabla_i Z/Z}{\text{the value of the associated statistic in the original system}}$$

One measure of the "fit" of the maximum entropy distribution generated by $\bar{\lambda}$ is the median value of the percent error, which was, for the $\bar{\lambda}$ we found, .07. So, compared to the original statistics, the errors in the statistics for the maximum entropy distribution generated by $\bar{\lambda}$ were typically small.

The results on the following pages contain more detailed information about the behavior of the maximum entropy distribution generated by $\bar{\lambda}$. They are the statistics of the original distribution, the estimated value of $\nabla Z(\bar{\lambda})/Z(\bar{\lambda})$ and the true value of $\nabla Z(\bar{\lambda})/Z(\bar{\lambda})$. The results are presented in ten row, thirty column tables. (On the first row we will have $\{\nabla_1 Z/Z, \ldots, \nabla_{10} Z/Z\}$, on the second $\{\nabla_{11} Z/Z, \ldots, \nabla_{20} Z/Z\}$, etc.) They are presented in such a way that the statistics in the first table have the same position in their table as the gradient associated with that statistic has in its own table. The statistics concerning the $E(\omega_i \omega_j)$ are thus on the top of the table.

234

0.0806	0.4255	-0.5389	0.4194	-0.0658	0.1038	-0.2769	-0.1695	-0.0471	-0.4299
-0.2675	0.1587	-0.4326	-0.1210	-0.0087	-0.5435	-0.2733	-0.1378	0.0675	-0.2034
-0.1561	0.1139	-0.2435	0.1101	-0.2267	0.0143	-0.1123	-0.1405	-0.0727	0.0058
0.2384	-0.1810	0.1295	0.1596	-0.0861	-0.2118	-0.1079	-0.2927	-0.1603	-0.0798
0.1153	-0.2382	0.0820	0.0935	0.0209	-0.3710	0.1078	-0.0310	0.0578	-0.2163
-0.2566	-0.0483	-0.3900	0.0760	-0.0324	-0.1739	-0.1297	0.1580	-0.5501	-0.3170
-0.0611	0.1214	-0.3010	0.1056	0.2089	-0.2999	-0.4132	-0.0192	-0.2100	0.2303
-0.0626	0.0139	0.3400	0.2605	-0.3386	0.5548	0.3427	-0.1447	0.4944	0.2766
0.1199	0.1702	0.4366	0.1741	-0.0355	0.1152	-0.1655	0.0974	-0.0307	0.1474
-0.0506	-0.1471	-0.3767	0.1875	-0.4381	-0.2993	-0.1297	-0.2945	-0.0289	-0.1637
0.0578	-0.1158	-0.2955	0.1512	-0.1849	0.3361	0.2909	-0.0898	0.1631	0.0670
0.0747	0.2564	-0.0816	-0.1574	-0.0171	0.2316	0.2813	0.0894	-0.2684	-0.1118
0.1283	-0.3382	0.2184	0.1454	0.0056	0.0848	0.1245	-0.1346	0.2802	-0.1257
-0.1996	0.1433	0.0238	-0.0056	0.0272	-0.3076	-0.1960	-0.1132	-0.2609	0.1727
0.1411	0.0855	0.1670	-0.1307	-0.0344	0.0115	-0.0492	-0.1137	0.3927	0.4310
-0.1274	-0.1325	0.1253	0.0540	-0.3286	0.0894	-0.1906	0.2345	-0.1451	0.1649
-0.1359	-0.1806	0.1306	0.2210	-0.0150	0.0971	-0.2461	-0.0658	-0.1536	-0.0483
-0.0332	0.0306	0.1210	0.1267	-0.0879	-0.0752	-0.2189	-0.0902	0.1518	-0.1577
0.1225	-0.1729	-0.0798	0.0621	0.1784	0.1587	0.0587	0.2528	0.0247	0.0028
0.4491	0.1652	0.0265	-0.1744	0.0841	0.0775	-0.2494	0.1086	-0.0153	0.3228
0.0290	0.1046	0.0138	-0.0861	0.2085	0.1336	-0.0243	0.2994	0.1015	-0.0400
-0.3274	-0.3316	0.1087	-0.0558	0.0223	0.1356	-0.3317	-0.1941	0.0738	-0.1794
0.0815	0.3093	-0.1086	0.3438	0.0098	0.1848	0.1888	0.3935	0.1908	0.0734
0.1954	0.1346	0.1088	-0.0214	-0.0820	0.0429	0.2873	0.1646	-0.1003	-0.0602
-0.0993	-0.1701	-0.0309	-0.2673	-0.2077	0.1756	-0.2132	-0.1913	0.3897	0.1054
-0.1502	0.3540	-0.0165	-0.2966	0.3280	-0.1534	0.0292	0.0823	0.0419	-0.2966
0.0808	-0.1662	0.1384	0.2146	0.1215	-0.0259	0.1721	-0.1384	0.3438	-0.0841
0.0553	0.0776	0.1831	-0.1699	-0.1746	-0.0239	-0.6172	-0.3503	-0.5228	0.6238
-0.4085	0.1994	0.0478	0.3981	0.1448	-0.0911	0.4876	0.1390	-0.2291	0.4696
0.2816	0.0204	0.7344	0.3949	-0.0182	-0.0226	0.3911	0.0047	-0.3163	0.3157

Table of $E(\omega_i \omega_j)$, $E(\omega_i)$

0.0181	0.0044	0.0103	-0.0576	0.0205	0.0128	0.0354	-0.0573	-0.0112	-0.0207
0.0526	-0.0433	0.0646	0.0483	-0.0115	0.0238	0.0049	-0.0582	0.0352	0.0075
0.0459	-0.1033	-0.0199	0.0230	-0.0314	-0.0270	-0.0332	0.0172	0.0277	0.0107
0.0089	-0.0006	0.0140	0.0496	-0.0075	0.0112	0.0152	-0.0301	-0.0083	-0.0053
-0.0019	-0.0347	0.0754	-0.0593	-0.0005	-0.0170	-0.0432	-0.0054	-0.0363	-0.0052
-0.0027	0.0156	0.0104	0.0496	-0.0028	0.0264	0.0278	0.0160	-0.0186	-0.0580
-0.0189	0.0001	-0.0323	0.0506	-0.0181	-0.0392	0.0497	-0.0578	-0.0270	-0.0151
0.0076	-0.0114	0.0145	-0.0614	-0.0072	-0.0147	0.0072	-0.0130	-0.0220	-0.0437
0.0466	0.0200	0.0033	-0.0422	0.0862	-0.0520	0.0346	0.0025	0.0689	0.0136
0.0134	0.0287	0.0575	-0.0071	0.0232	0.0174	-0.021?	0.0441	0.0726	-0.0476
0.0137	0.0614	0.0371	-0.0227	0.0176	0.0547	-0.0176	0.0389	-0.0030	0.0127
0.0183	0.0545	-0.0850	-0.0454	0.0182	0.0142	0.0230	0.0347	-0.0603	-0.0647
-0.0331	0.0112	0.0113	0.0776	-0.0295	0.0342	0.0191	-0.0370	0.0079	0.0057
-0.0248	0.0096	0.0523	0.0355	-0.0403	-0.0448	-0.0213	-0.0560	-0.0749	0.0524
0.0834	-0.0371	0.0368	-0.0412	0.0673	-0.0717	-0.0453	0.0335	-0.0058	-0.0162
0.0240	-0.0730	-0.0221	-0.0160	0.0146	0.0747	0.0260	0.0130	-0.0213	-0.0192
-0.0263	-0.0442	-0.0489	0.0198	0.0603	-0.0337	0.0176	-0.0155	-0.0457	0.0396
0.0524	-0.0090	-0.0063	0.0133	-0.0652	-0.0188	-0.0057	-0.0195	0.0259	0.0063
0.0252	-0.0466	-0.0233	0.0099	-0.0020	-0.0233	0.0006	-0.0170	-0.0228	-0.0263
0.0516	0.0211	-0.0190	-0.0321	-0.0085	-0.0110	0.0063	0.0075	0.0340	-0.0543
-0.0406	-0.0124	-0.0615	-0.0315	0.0615	0.0441	-0.0648	-0.0273	0.0446	-0.1162
0.0178	-0.0483	-0.0144	0.0403	0.1209	-0.0200	-0.0304	0.0106	0.0494	-0.0770
0.0250	0.0564	0.0125	-0.0317	-0.0714	0.0483	0.0162	0.0110	-0.0118	0.0310
-0.0513	-0.0110	-0.0483	-0.1200	0.0366	0.0231	0.0208	-0.0155	0.0306	0.0119
0.0016	0.0281	-0.0513	0.0288	-0.0188	-0.0227	-0.0221	0.0133	-0.0107	0.0304
-0.0363	-0.0077	-0.0792	0.0160	0.0388	0.0365	-0.0204	-0.0463	-0.0299	0.0240
0.0323	0.0200	-0.0552	0.0275	0.0142	-0.0724	0.0058	-0.0108	-0.0077	-0.0415
-0.0188	0.0411	0.0186	0.0623	0.0610	-0.0424	0.0195	-0.0346	-0.0599	-0.0191
0.0643	0.0026	0.0519	0.0355	0.0610	-0.0524	0.0213	-0.0826	0.0023	-0.0123
-0.?108	0.0230	0.0034	-0.0136	-0.0389	0.0168	0.0367	-0.1384	0.0327	0.0507

Table of estimated $\nabla Z(\lambda)/Z(\lambda)$

-0.0233	0.0004	0.0116	-0.0173	0.0168	0.0138	-0.0258	-0.0143	-0.0304	-0.0050
0.0088	0.0029	0.0204	0.0477	0.0017	0.0095	-0.0259	0.0077	0.0056	0.0386
0.0042	-0.0317	-0.0117	-0.0152	-0.0069	-0.0288	-0.0116	-0.0147	0.0091 .	0.0114
0.0049	-0.0140	0.0130	0.0087	0.0282	-0.0004	-0.0071	-0.0014	-0.0171	0.0035
0.0081	-0.0035	0.0215	-0.0188	0.0214	-0.0027	-0.0297	0.0244	0.0249	0.0135
-0.0135	-0.0196	-0.0130	0.0241	-0.0123	0.0262	0.0044	0.0198	0.0049	-0.0208
0.0026	-0.0019	0.0015	0.0301	-0.0010	-0.0112	0.0188	-0.0195	-0.0145	0.0258
-0.0129	0.0259	0.0057	-0.0064	0.0000	-0.0185	-0.0125	-0.0276	-0.0139	0.0189
0.0029	0.0281	-0.0028	-0.0001	0.0279	-0.0257	-0.0197	0.0059	0.0068	0.0244
-0.0388	0.0301	0.0052	0.0022	0.0205	0.0093	0.0014	0.0183	0.0120	-0.0009
0.0000	0.0401	0.0010	0.0006	-0.0094	0.0058	-0.0060	0.0031	0.0023	-0.0132
0.0213	0.0045	-0.0235	-0.0115	0.0045	0.0065	-0.0008	0.0152	-0.0188	-0.0323
0.0065	-0.0001	-0.0046	0.0057	-0.0079	0.0034	0.0077	-0.0069	0.0025	0.0225
-0.0036	0.0243	-0.0186	-0.0180	0.0083	-0.0144	-0.0173	-0.0136	-0.0128	0.0019
0.0324	-0.0094	-0.0030	-0.0055	-0.0238	-0.0071	-0.0105	0.0030	0.0184	0.0057
-0.0095	-0.0159	0.0169	0.0132	-0.0046	-0.0230	-0.0011	0.0057	-0.0165	0.0074
-0.0256	-0.0285	-0.0130	0.0238	0.0146	-0.0065	0.0043	-0.0159	-0.0240	0.0129
-0.0079	0.0203	0.0315	0.0142	0.0012	0.0306	0.0129	-0.0136	-0.0041	0.0124
-0.0097	-0.0131	0.0050	0.0218	0.0098	0.0016	0.0128	0.0009	-0.0316	-0.0288
0.0036	0.0049	-0.0353	-0.0014	-0.0258	-0.0010	-0.0069	-0.0120	0.0307	-0.0157
-0.0292	0.0070	-0.0269	-0.0089	-0.0022	0.0189	-0.0253	-0.0034	0.0195	-0.0142
-0.0053	-0.0039	0.0180	0.0210	0.0176	0.0094	-0.0033	0.0144	0.0152	-0.0099
0.0133	-0.0107	-0.0375	-0.0175	-0.0173	-0.0007	0.0311	0.0001	0.0005	0.0259
0.0120	0.0049	-0.0465	-0.0200	-0.0194	-0.0037	-0.0021	0.0017	-0.0100	-0.0199
-0.0240	-0.0005	-0.0344	0.0008	-0.0130	0.0004	-0.0170	-0.0019	0.0167	-0.0035
-0.0038	-0.0087	-0.0272	0.0035	0.0050	-0.0045	0.0021	-0.0245	0.0099	-0.0006
-0.0143	0.0111	0.0005	0.0045	0.0213	-0.0234	0.0275	-0.0046	0.0039	-0.0108
0.0041	0.0229	0.0031	0.0017	-0.0148	0.0223	-0.0016	-0.0240	-0.0092	-0.0002
0.0066	0.0200	0.0274	0.0393	0.0226	-0.0146	0.0205	0.0019	-0.0198	-0.0030
-0.0054	0.0204	0.0006	0.0308	-0.0525	0.0347	0.0002	-0.0478	-0.0175	0.0187

Table of true $\nabla Z(\lambda)/Z(\lambda)$

Example 2: Letter recognition

Let us consider the problem of invariant letter recognition. We will be presented with a picture of a letter of unknown size, orientation and font, and we wish to find out which letter it is. For the sake of simplicity we will use simple images with just two grey levels (black and white), and we will just consider the capital letters $A, \ldots, G$.

There are many ways to approach this problem, the one we will consider is based on feature extraction. We will deal with the invariance of the problem by extracting features (scalars) that are independent of the orientation or size of the letter. Our expert system will be a distribution on the space of features and labels, where the latter identify the letter. This distribution will be used to find the probabilities of the labels conditioned on the observed features (*e.g.*, $P($ 'the letter is an A' | feature$_1 = 5$, feature$_2 = 6) = .3$)

Choosing the features is a crucial task, and should be given as much consideration as the construction of the expert system that uses them. Features can be roughly separated into two groups, local and global. Global features deal with the whole picture and are what we used in the results presented in the following pages. Local features deal with the local behavior of the picture elements. Hence, local features are ideal for occluded pictures. Local features seem more powerful and, it is our belief, will be essential for a true solution to the invariant character recognition problem.

The features we used in our example were non-standard. They were picked because they seemed reasonable and not too difficult to compute. They mostly deal with holes and indentations. A hole being a white (non-letter) region completely surrounded by the letter (typically A has a hole, C does not), and an indentation being a white region that is connected, is in the convex hull (the convex hull of the set S is the smallest convex set containing S) of the letter, and yet not a hole. Some thought will show that this is exactly what we mean by an indentation (typically O has no indentations, T has two). Below are listed twelve of the features we use.

1	The size of the largest hole	/ The size of the convex hull of the letter
2	The size of the second largest hole	/ The size of the convex hull of the letter
3	The size of the third largest hole	/ The size of the convex hull of the letter
4	The size of the largest indentation	/ The size of the convex hull of the letter
5	The size of the second largest indentation	/ The size of the convex hull of the letter
6	The size of the third largest indentation	/ The size of the convex hull of the letter
7	The ratio of longest to shortest axis of the largest hole	
8	The ratio of longest to shortest axis of the largest indentation	
9	The ratio of longest to shortest axis of the second largest indentation	
10	The ratio of longest to shortest axis of the third largest indentation	
11	The total area of the indentations in the largest hole	/ The size of the letter
12	The total area of the indentations in the largest indentation	/ The size of the letter

We also have several other features that deal with the points that span the convex hull. We construct these features as follows. Let our original set of points be the smallest set that spans the convex hull of the letter. At every step remove one point from our set of points, picked to maximize the area spanned by the remaining points. Continue doing this until no points are left. Our final features shall be

13	The number of points that span the convex hull of the letter	
14	The area spanned by six remaining points	/ The area of the convex hull
15	The area spanned by five remaining points	/ The area of the convex hull

237

16   The area spanned by four remaining points   / The area of the convex hull

17   The area spanned by three remaining points / The area of the convex hull

These features are useful since they tell us how curved the letter is. For example the letter 'E' is square so the area left after removing all but four points should be large, but when only three points remain the number should be much smaller.

The knowledge we used to form our expert system deals with the expected values of the features, the features squared, and the products of selected features, all conditioned on the letter (eg. $E(\text{feature}_1 |$ the letter is an $A) = C_1$, $E((\text{feature}_1)^2 |$ the letter is an $A) = C_2$, $E(\text{feature}_1 \cdot \text{feature}_2 |$ the letter is an $A) = C_3$). We also give our system the very important piece of knowledge that the letters are of equal probability (each of probability 1/7). It would be nice to use all the products of features as constraints, but with seven letters and 17 features we would have several thousand constraints and this is computationally difficult. In the experiment that yielded the results on the following pages we used the conditional probabilities of only fifteen different products. The total number of constraints was thus 350.

On the computational side we use several techniques to improve the behavior of the gradient descent. The first technique deals with the constraints themselves. In general a constraint is of the form $E(G_i) = C_i$, where G_i is some function and C_i is a constant. Typically we find C_i by taking a test sample, and using the sample mean of G_i. However this does have problems. Since our gradient descent is not exact we typically end up with $\nabla_i Z/Z$ small for all i and typically of the same order of magnitude, but not exactly zero. Thus if we have as a constraint $E(100 \cdot G_j) = 100 \cdot C_j$ for some j, then $E(G_j) = C_j$ will come much closer to being true in the resulting maximum entropy distribution than if we had $E(G_j) = C_j$ as the constraint. Also, there is a problem caused by wanting our expert system to recognize things not in the sample that formed our constraints. In the following results we trained our system with 5 samples of each letter. Now, what will happen if we try to recognize a letter that was not in the population we used to train the system? We would like it to be recognized, especially if it is similar to the original population. This does not always happen. One particular case of this problem is caused by boundry effects. If a feature has range 0 to 1, and all the sample letter C's had value 0 for this feature, then the only way to satisfy the constraint $E(\text{feature} \mid$ letter is a $C) = $ sample mean $= 0$, is to have $P(\text{feature} = 0 \mid$ letter is a $C) = 1$. If we present a C which has value .001 for this feature, it will not be recognized. While this problem could be cured by having a large sample (and should be), it and the previous problem can both be dealt with by scaling and slightly modifying the constraint functions. For the full details we refer the reader to [2].

Now let us consider the sampling method itself. In this problem the constraints have a rather odd form, almost all of them are conditioned on the letter. This can lead to difficulties in the sampling method. When the letter is an A, for example, the features tend to have certain values, as the constraints specify. At every step in the sampling we go through the feature vector, holding most of the features fixed and then picking those that are not fixed according to a distribution. However when the label is 'A', the features tend to stay within a certain range. When it comes time to fix the features and vary the label, the distribution that we use to pick a label, being generated by features that correspond to an A, will emphasize the label 'A'. This is to be expected, since we can think of the label 'A' as corresponding to some region in the state space, and forming a sort of 'well' in the energy landscape (a region of very likely events, corresponding to 'A's, surrounded by a region of low probability that corresponds to feature values not associated with any letter). Once such a 'well' is entered it can be difficult to get out of. So, if the label 'A' is turned on it tends to stay on, and our sample will quite possibly over-emphasize one particular letter at the expense of the rest.

Since all the constraints involving the features are conditional, we can use the following

change in the sampling method to cure the problem mentioned above. We will first fix label at 'A'. We will then conduct gradient descent on Z until the values of $\nabla_i Z/Z$ are small for all the i's corresponding to constraints conditioned on the label being 'A'. Then we fix the label 'B' and continue. After we have gone through all the letters (in our case A,...,G) we start sampling normally (letting the label vary). Since the values of $\nabla_i Z/Z$ are small for all i's corresponding to conditional constraints, we need only conduct gradient descent until the constraint that all letters be equally likely is (close to being) satisfied.

Now let us present the results. The letters we wished to identify are on the following pages. They are the same letters that were used to find the sample means in the constraints. The probabilities of the labels conditioned on the observed features, given by the maximum entropy expert system, is provided underneath the letters. Only the top three probabilities are listed for each letter, in the interest of saving space. The energies are also listed, where the energy is $\sum_{i=1}^{m} -a_i(\omega)\bar{\lambda}_i$ (where ω is the element of Ω corresponding to the fee' ire vector plus the hypothesized label, and $\bar{\lambda}$ is the result of our minimization of Z). The energies are given to provide some comparison between different letters ("this E looks more like an E than that E").

1

2

3

4

5
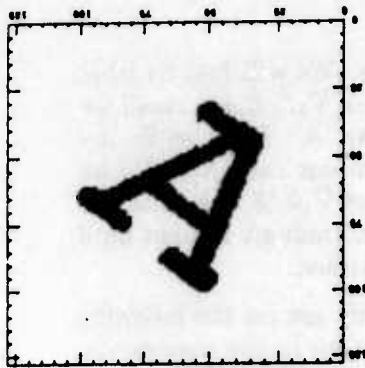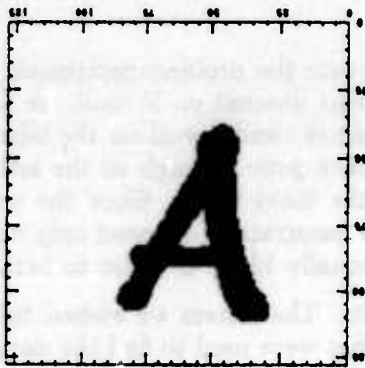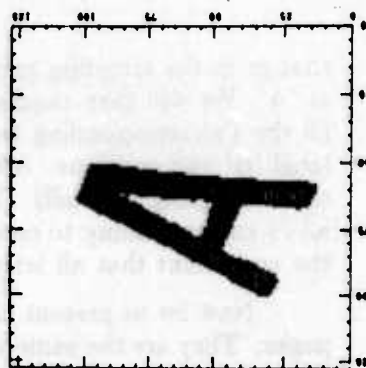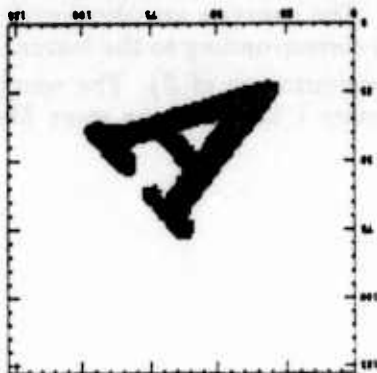
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
A1	it is an A	0.6719273	3.8487754
A1	it is an E	0.0002583	11.7124262
A1	it is an F	0.3277985	4.5665264
A2	it is an A	0.9809950	2.1446524
A2	it is an E	0.0012927	8.7764645
A2	it is an F	0.0177039	6.1594334
A3	it is an A	1.0000000	0.2035229
A3	it is an B	0.0000000	18.8109589
A3	it is an E	0.0000000	23.3654041
A4	it is an A	0.9920666	1.6209198
A4	it is an E	0.0000050	13.8176394
A4	it is an F	0.0079282	6.4502902
A5	it is an A	0.9974482	5.4865880
A5	it is an E	0.0001584	14.2343798
A5	it is an F	0.0023898	11.5205803

240

1

2

3

4

5
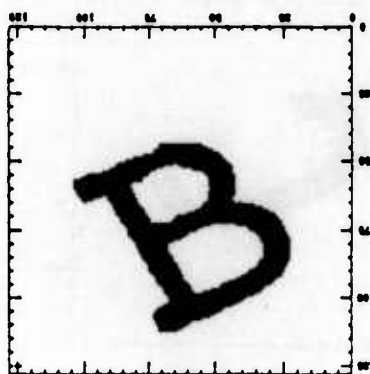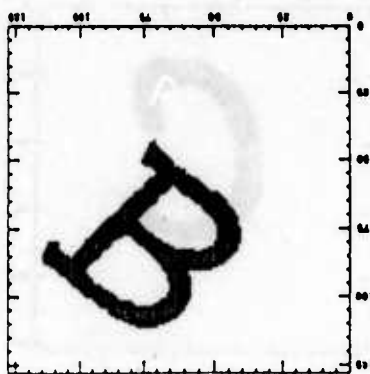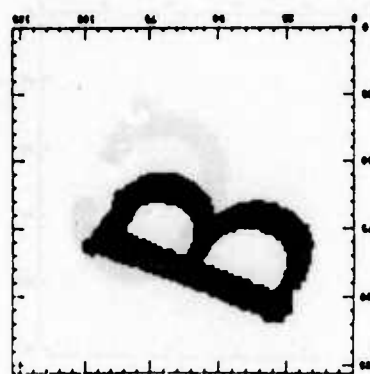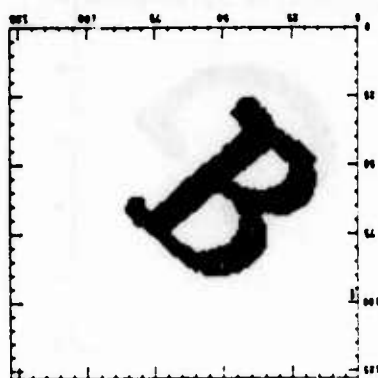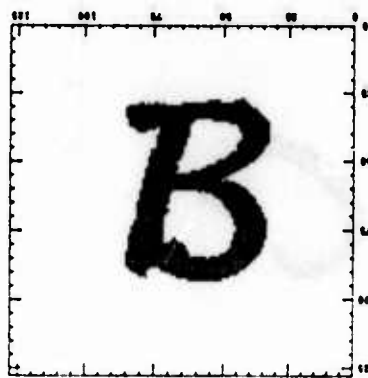
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
B1	it is an B	1.0000000	5.7593293
B1	it is an C	0.0000000	26.1249256
B1	it is an D	0.0000000	26.0998402
B2	it is an B	1.0000000	6.2038503
B2	it is an C	0.0000000	26.3224182
B2	it is an E	0.0000000	26.6758900
B3	it is an A	0.0000000	29.8165340
B3	it is an B	1.0000000	9.1290588
B3	it is an E	0.0000000	30.7371597
B4	it is an B	0.9999986	5.9010286
B4	it is an E	0.0000006	20.2056236
B4	it is an F	0.0000003	20.8549023
B5	it is an B	0.9999995	5.5978689
B5	it is an E	0.0000000	23.7445774
B5	it is an F	0.0000004	20.2729683

241

1

2

3

4

5
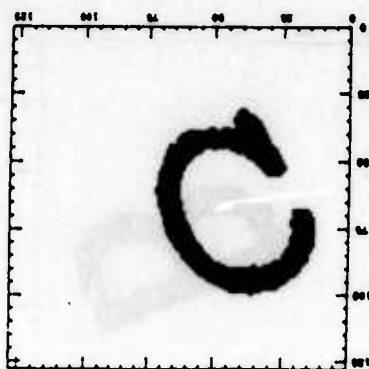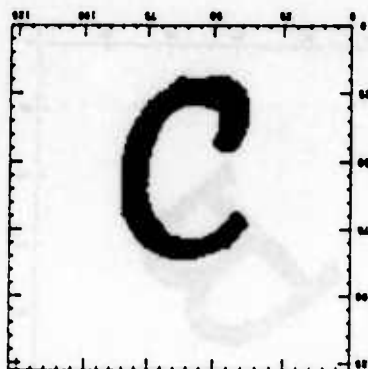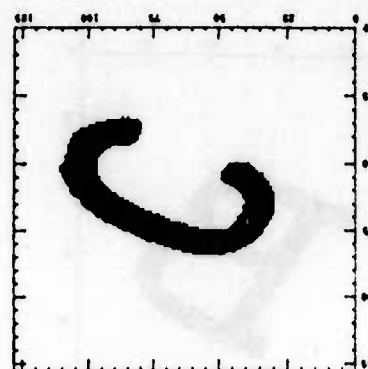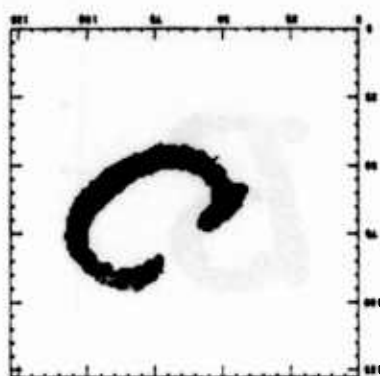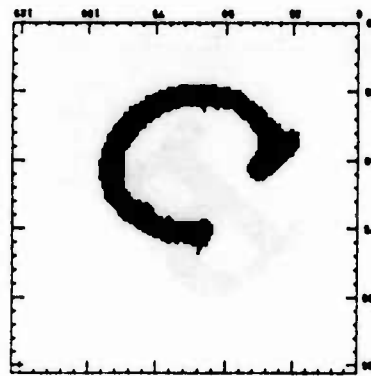
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
C1	it is an C	0.6604182	-1.9814487
C1	it is an E	0.0184070	1.5986940
C1	it is an G	0.3191914	-1.2543660
C2	it is an C	0.6313014	-1.3120894
C2	it is an E	0.1306149	0.2634406
C2	it is an G	0.2368994	-0.3319414
C3	it is an C	0.5955555	-1.6064481
C3	it is an E	0.0437944	1.0035404
C3	it is an G	0.3588811	-1.0999445
C4	it is an C	0.6606517	-0.5129181
C4	it is an E	0.0676684	1.7656897
C4	it is an G	0.2695387	0.3835969
C5	it is an C	0.6386604	-1.4192295
C5	it is an E	0.0491251	1.1457740
C5	it is an G	0.3097548	-0.6956378

242

1

2

3

4

5
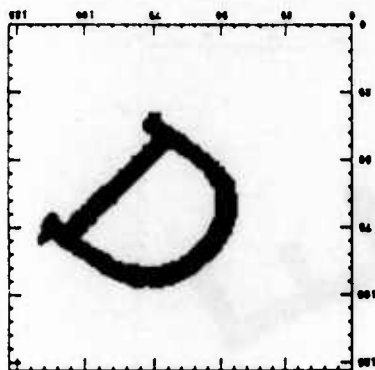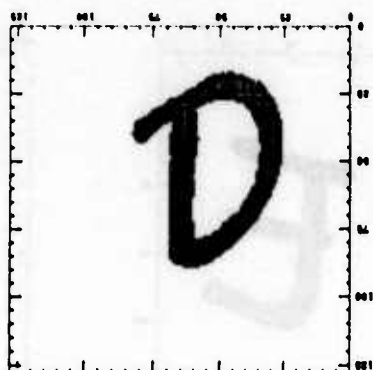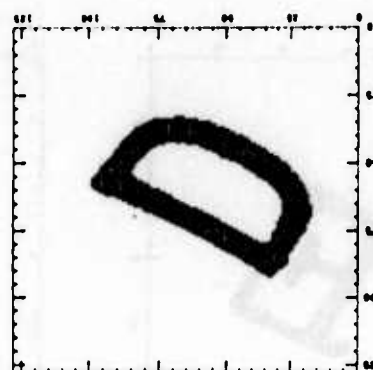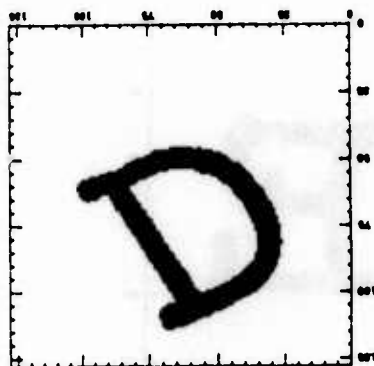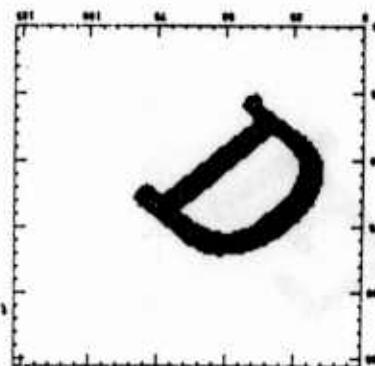
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
D1	it is an A	0.0000051	19.0304546
D1	it is an B	0.0011396	13.6119480
D1	it is an D	0.9988480	6.8360224
D2	it is an B	0.0003877	12.2251072
D2	it is an D	0.9994236	4.3704495
D2	it is an G	0.0000976	13.6045828
D3	it is an B	0.0000000	89.0601807
D3	it is an D	1.0000000	-12.7670460
D3	it is an F	0.0000000	82.0110168
D4	it is an B	0.0005200	12.7396383
D4	it is an D	0.9994300	5.1784315
D4	it is an G	0.0000154	16.2575455
D5	it is an A	0.0000039	17.1760368
D5	it is an B	0.0003292	12.7477312
D5	it is an D	0.9996585	4.7292614

1

2

3

4

5
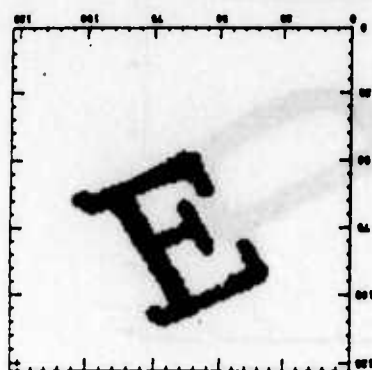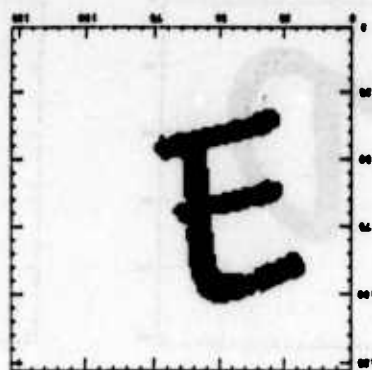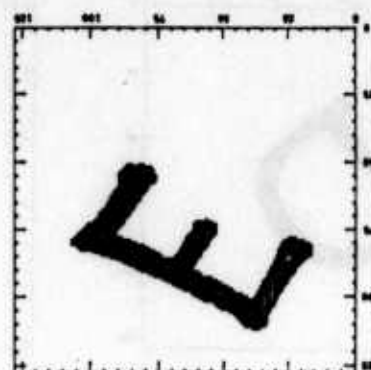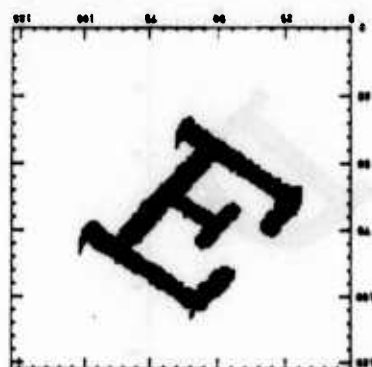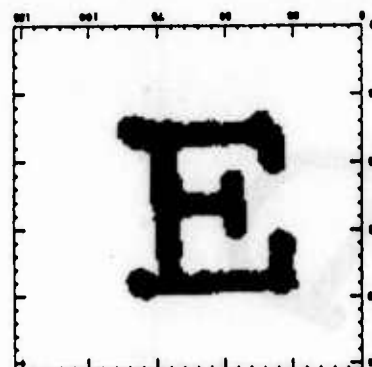
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
E1	it is an A	0.0000728	7.3868618
E1	it is an E	0.9531993	-2.0928898
E1	it is an F	0.0467241	0.9226741
E2	it is an C	0.0000507	7.1914301
E2	it is an E	0.9832692	-2.6817935
E2	it is an F	0.0166214	1.3983997
E3	it is an A	0.0003051	8.8423738
E3	it is an C	0.0000295	11.1771812
E3	it is an E	0.9996634	0.7478167
E4	it is an A	0.0003749	6.4038081
E4	it is an E	0.9340211	-1.4169102
E4	it is an F	0.0655886	1.2391865
E5	it is an A	0.0000976	6.7060304
E5	it is an E	0.9618846	-2.4899280
E5	it is an F	0.0380000	0.7413796

1

2

3

4

5
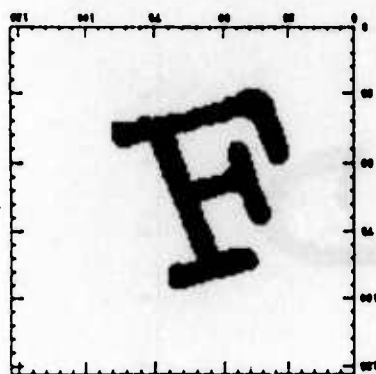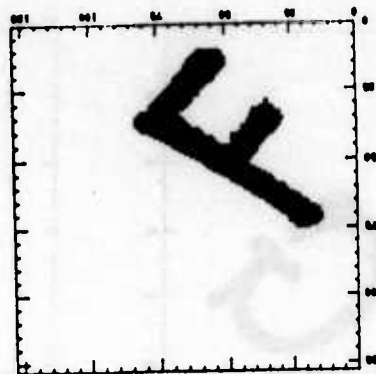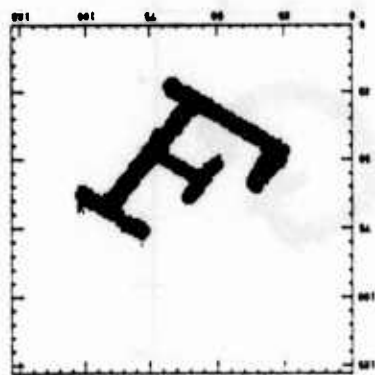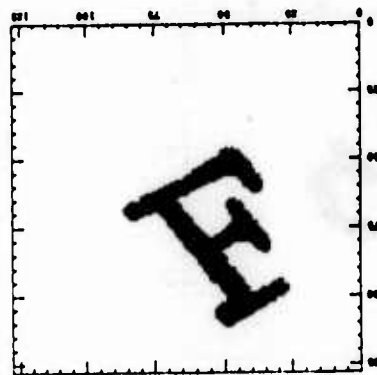
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
F1	it is an A	0.0009967	5.9281058
F1	it is an E	0.0000012	12.6835003
F1	it is an F	0.9990017	-0.9819657
F2	it is an A	0.0136954	3.9498260
F2	it is an E	0.0032687	5.3824973
F2	it is an F	0.9830301	-0.3237556
F3	it is an A	0.0021422	5.9868760
F3	it is an E	0.0016594	6.2422500
F3	it is an F	0.9961985	-0.1552548
F4	it is an E	0.8462385	-0.0166720
F4	it is an F	0.1515211	1.7034043
F4	it is an G	0.0020051	6.0284510
F5	it is an A	0.0003445	8.3701258
F5	it is an E	0.8101026	0.6073851
F5	it is an F	0.1895521	2.0598819

1

2

3

4

5
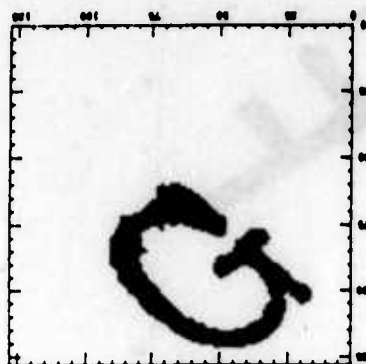
Picture	Hypothesis	Probability of Hypothesis	Energy of Hypothesis
G1	it is an C	0.1939279	1.1415343
G1	it is an E	0.2315701	0.9641383
G1	it is an G	0.5542700	0.0913689
G2	it is an C	0.2378499	0.9435763
G2	it is an E	0.1392609	1.4788672
G2	it is an G	0.6008883	0.0168073
G3	it is an C	0.3128236	0.1166506
G3	it is an E	0.0234408	2.7078128
G3	it is an G	0.6587726	-0.6280884
G4	it is an C	0.2702022	0.0306356
G4	it is an E	0.0150385	2.9191945
G4	it is an G	0.7069070	-0.9310930
G5	it is an C	0.3253015	-0.3527871
G5	it is an E	0.1510101	0.4146184
G5	it is an G	0.5172358	-0.8165337

246

Bibliography

[1] Edwin T. Jaynes, 'On The Rational of Maximum-Entropy Methods,' *Proc. of the IEEE* **70** (1982), 939–952.

[2] Alan F. Lippman, 'A Maximum Entropy Method for Expert System Construction,' *Ph.D. Thesis, Brown University* (1986).

[3] Stuart Geman, 'Stochastic Relaxation Methods for Image Restoration and Expert Systems,' To appear in: *Automated Image Analysis: Theory and Experiments, D.B. Cooper, R.L. Launer, and D.E. McClure, Eds.* New York: Academic Press.

[4] Donald Geman, Personal Communication (1985).

[5] Censor,Y., Elfving,T., Herman,G.T., Kuo,Y.H., and Lent,A., 'On The Relationship Between "MART" and Bregman's Algorithm for Entropy Maximization Over Linear Inequalities,' **In preparation**.

[6] Peter Cheeseman, 'A Method of Computing Maximum Entropy Probability Values for Expert Systems,' **Preprint** SRI International, Menlo Park, California.

[7] Gull,S.F. and Skilling,J., 'The Entropy of an Image', *Proceedings of the Second Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics,'* **To appear**

[8] Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H., and Teller,E., 'Equations of State Calculations by Fast Computing Machines,' *J. Chem. Phys.* **21** (1953), 1087-1091.

PROBABILISTIC FINITE ELEMENTS AND POTENTIAL APPLICATIONS TO FRACTURE*

Wing Kam Liu, Ted Belytschko,
Glen Besterfield, and A. Mani

Northwestern University
Department of Mechanical and Nuclear Engineering
Evanston, Illinois, 60201, U.S.A.

ABSTRACT. Methodologies for treating random field problems by finite elements are described. The methods are based on second-moment analysis procedures, but they are remarkably robust and are able to deal with substantial nonlinearities. Both static and dynamic problems have been considered. Some applications to linear and elastic-plastic structures are described along with potential applications to fracture which are now being considered.

I. INTRODUCTION. The probabilistic analysis of engineering problems by finite element methods is currently a dynamic area of research. The most widespread statistical approach for analyzing probabilistic systems is by simulation, the direct Monte Carlo Simulation [1-3] being the most frequently used. Since the accuracy of the statistical results is dependent on the number of samples, the analysis can be prohibitively expensive for large systems. Although simulation techniques can be applied to linear and nonlinear systems, they are in general quite inefficient. Thus, there is considerable interest in non-statistical approaches, such as second-moment analysis and Probabilistic Finite Element Methods (PFEM). For linear systems, second-moment analysis techniques [3,4] have proven to be effective in structural mechanics. But, the extension of second-moment analysis to nonlinear structural dynamics is not currently feasible. Consequently, recent developments in the statistical analysis of linear and nonlinear structural dynamics have been advanced with PFEM.

Although the development of PFEM is a relatively new area of research, the amount of literature is quite broad [5-13]. The authors' research has encompassed both static and dynamic linear PFEM as well as recent advances in nonlinear PFEM. The development of PFEM for static linear analysis with material randomness is discussed in Ref. [9]. In the application of PFEM for linear dynamics, secular terms arise in the statistical distributions causing erroneous results [8,11,13]. In Refs. [8,9], the PFEM is extended to static and dynamic nonlinear analysis with material and geometric nonlinearities. Extensive research has been done in the application of PFEM for elastic/ plastic materials [9,10]. Recently, the PFEM has been developed using a potential energy variational principle [11]. In this manner, problems with random materials, shapes, body forces, and boundary conditions can be easily incorporated into the PFEM. In Ref. [12], the probabilistic potential energy

variational principle is extended to a three-field Hu-Washizu variational principle. The PFEM has proved to be a very efficient means of non-statistical analysis for linear and nonlinear continuum in statics and dynamics. Currently, the majority of the research in PFEM is directed toward improved nonlinear analysis.

It has been observed [8] that the secular terms arise in nonlinear transient analysis as well. Elimination of these secular terms is not as straightforward as in linear transient analysis, and is a current topic of research. The nonlinear probabilistic analysis herein, is, therefore, restricted to statics.

In the next section, the linear transient PFEM equations and the scheme for eliminating secularities are outlined. In Section III, the PFEM equations for nonlinear statics are derived. In Section IV, the effectiveness of PFEM and the scheme for eliminating secularities are demonstrated. In Section V, the conclusions and potential applications to fracture are discussed.

II. PFEM FOR TRANSIENT ANALYSIS. As a consequence of applying PFEM for transient analysis, secular terms arise in the higher order equations and hence, all statistical results [8]. Many theoretical methods have been developed for eliminating secularities and the literature is quite extensive. Secular terms erroneously result from the perturbation process causing the higher order equations to increase indefinitely with time or until damped away. Thus, secularities cause all statistical results such as the expectation and variance of displacement to be unbounded for long times. The characteristics of secularities and a method for their removal have been developed for a single degree-of-freedom random oscillator [13], but to the authors' knowledge no methods have been developed for PFEM. Consequently, there is a considerable need to develop means for eliminating secular terms in PFEM.

Initially, consider a structural dynamic system governed by the following linear system of equations which are developed from a finite element discretiztion:

$$\underset{\sim}{M}\,\underset{\sim}{\ddot{d}} + \underset{\sim}{C}\,\underset{\sim}{\dot{d}} + \underset{\sim}{K}\,\underset{\sim}{d} = \underset{\sim}{F} \tag{2.1}$$

where $\underset{\sim}{M}$, $\underset{\sim}{C}$, and $\underset{\sim}{K}$ are the mass, damping and stiffness matrices, respectively; $\underset{\sim}{F}$ is the external force vector; $\underset{\sim}{d}$ is the displacement vector; and a superscript dot represents time differentiation. The mass is assumed to be deterministic whereas the stiffness and damping are assumed to be functions of a generalized variance vector Var(b) where $b(\underset{\sim}{x})$ is a random field. The basic idea in applying second-moment analysis to develop PFEM involves expanding all random functions about the mean value of the random field $b(\underset{\sim}{x})$, denoted by $\overline{b}(\underset{\sim}{x})$. That is, for a small parameter ε, the random function $\underset{\sim}{d}(\underset{\sim}{b},t)$ is expanded about $\overline{b}(\underset{\sim}{x})$ via a second-order perturbation at a given $\underset{\sim}{x}$ and the random field is discretized along with the random functions as follows:

$$\underset{\sim}{d} = \underset{\sim}{d}_0 + \epsilon \sum_{i=1}^{q} \underset{\sim}{d}_{b_i} \Delta b_i + \frac{1}{2} \epsilon^2 \sum_{i,j=1}^{q} \underset{\sim}{d}_{b_i b_j} \Delta b_i \Delta b_j \tag{2.2}$$

where $\underset{\sim}{d}_0$, $\underset{\sim}{d}_{b_i}$, and $\underset{\sim}{d}_{b_i b_j}$ represent the mean, the first order variation about $\underset{\sim}{\bar{b}}$, and the second order variation about $\underset{\sim}{\bar{b}}$ of the displacement; Δb_i represents the first order variation of b_i about $\bar{b}_i$; and q is the number of random variables. Complete details of this procedure can be found in Refs. [7,8]. Similar expansions are done for $\underset{\sim}{F}$, $\underset{\sim}{K}$, and $\underset{\sim}{C}$. Substitution of these expansions for $\underset{\sim}{d}$, $\underset{\sim}{F}$, $\underset{\sim}{K}$, and $\underset{\sim}{C}$ into Eq. (2.1) yields the following three equations for $\underset{\sim}{d}_0$, $\underset{\sim}{d}_{b_i}$, and $\underset{\sim}{d}_2$:

Zeroth Order Equation

$$\underset{\sim}{M} \ddot{\underset{\sim}{d}}_0 + \underset{\sim}{C}_0 \dot{\underset{\sim}{d}}_0 + \underset{\sim}{K}_0 \underset{\sim}{d}_0 = \underset{\sim}{F}_0 \tag{2.3}$$

First Order Equation

$$\underset{\sim}{M} \ddot{\underset{\sim}{d}}_{b_i} + \underset{\sim}{C}_0 \dot{\underset{\sim}{d}}_{b_i} + \underset{\sim}{K}_0 \underset{\sim}{d}_{b_i} = \underset{\sim}{F}_{1b_i} \quad , \quad i = 1, \ldots, q \tag{2.4}$$

where

$$\underset{\sim}{F}_{1b_i} = \underset{\sim}{F}_{b_i} - (\underset{\sim}{C}_{b_i} \dot{\underset{\sim}{d}}_0 + \underset{\sim}{K}_{b_i} \underset{\sim}{d}_0) \quad , \quad i = 1, \ldots, q \tag{2.5}$$

Second Order Equation

$$\underset{\sim}{M} \ddot{\underset{\sim}{d}}_2 + \underset{\sim}{C}_0 \dot{\underset{\sim}{d}}_2 + \underset{\sim}{K}_0 \underset{\sim}{d}_2 = \underset{\sim}{F}_2 \tag{2.6}$$

where

$$\underset{\sim}{F}_2 = \sum_{i=1}^{q} \left\{ \frac{1}{2} \underset{\sim}{F}_{b_i b_i} - \frac{1}{2} \underset{\sim}{C}_{b_i b_i} \dot{\underset{\sim}{d}}_0 - \frac{1}{2} \underset{\sim}{K}_{b_i b_i} \underset{\sim}{d}_0 - \underset{\sim}{C}_{b_i} \dot{\underset{\sim}{d}}_{b_i} \right.$$

$$\left. - \underset{\sim}{K}_{b_i} \underset{\sim}{d}_{b_i} \right\} \mathrm{Var}(b_i) \tag{2.7}$$

and

$$\underset{\sim}{d}_2 = \frac{1}{2} \sum_{i=1}^{q} \underset{\sim}{d}_{b_i b_i} \mathrm{Var}(b_i) \tag{2.8}$$

The solution of Eqs. (2.3) and (2.6) yields $\underset{\sim}{d}_0$ and $\underset{\sim}{d}_2$, respectively, whereas the solution of Eqs. (2.4) requires q solutions to obtain $\underset{\sim}{d}_{b_i}$. In Eqs. (2.4) through (2.8) it has been assumed that b_i and b_j are uncorrelated for $i \neq j$,

251

thereby enabling the full covariance matrix $Cov(b_i, b_j)$ to be expressed as a diagonal variance matrix $Var(b_i)$ for $i = j$ and zero for $i \neq j$. It is noted that in order to reduce the computations further, a transformed random variable can be introduced [9]. After the zeroth order equation is solved for $\underset{\sim}{d}_0$, the q first order forcing functions given by Eq. (2.5) can be evaluated. Since the first order forcing function is a function of the zeroth order solution, part of its effect will be resonant causing secularities in the first order solution [11,13]. The second order forcing function is a function of the first order solution in addition to the second order solution, thus secularities also result in the second order solution. When damping is present in the system, the effect of secularities is present until it is damped away for long durations. The secular terms in the first and second order solutions erroneously result from the perturbation process. Therefore, the method presented in this paper for removal of secularities in PFEM involves removing the resonant part from the first and second order forcing functions.

The mean and variance of displacement are defined by

$$E[\underset{\sim}{d}] = \int_{-\infty}^{\infty} \underset{\sim}{d} \, f(\underset{\sim}{b}) d\underset{\sim}{b} \qquad (2.9)$$

and

$$Var(\underset{\sim}{d}) = \int_{-\infty}^{\infty} (\underset{\sim}{d} - \underset{\sim}{d}_0)^2 f(\underset{\sim}{b}) d\underset{\sim}{b} \qquad (2.10)$$

respectively, where $f(\underset{\sim}{b})$ is the probability density function. Once Eqs. (2.3), (2.4), and (2.6) are solved for $\underset{\sim}{d}_0$, $\underset{\sim}{d}_{b_i}$, and $\underset{\sim}{d}_2$, respectively, substitution of the expansion for $\underset{\sim}{d}$ given by Eq. (2.2) into Eqs. (2.9) and (2.10) yields the second order accurate expectation and first order accurate variance of displacement given by

$$E[\underset{\sim}{d}] = \underset{\sim}{d}_0 + \underset{\sim}{d}_2 \qquad (2.11)$$

and

$$Var(\underset{\sim}{d}) = \sum_{i=1}^{q} (\underset{\sim}{d}_{b_i})^2 Var(b_i) \qquad (2.12)$$

252

respectively. Since d_2 in Eq. (2.11) has secular terms present, the expectation of displacement will increase indefinitely with time. Similarly, the variance of displacement will also increase indefinitely with time due to secularities in d_{b_i}. Similar expressions to Eqs. (2.11) and (2.12) can be developed for strain and stress. The statistical results for strain and stress will also be invalid for long times. Thus, there is a considerable need to develop methods to eliminate secularities in PFEM so all statistical results are bounded.

There is a vast amount of literature available dealing with the analytical removal of secularities but no methods have been developed for the numerical elimination of secularities in PFEM. The method presented herein for numerical elimination of secular terms involves using Fourier Analysis to separate the resonant and non-resonant parts from the first and second order forcing functions. By performing Fourier Analysis on the time series for $F_{1b_i}(d_0)$ and $F_2(d_0, d_{b_i})$ with a Fast Fourier Transform (FFT), the time series can be separated as follows

$$F_{1b_i}(d_0) = F_{1b_i}^R + F_{1b_i}^{NR} \qquad (2.13)$$

and

$$F_2(d_0, d_{b_i}) = F_2^R + F_2^{NR} \qquad (2.14)$$

where the superscripts R and NR represent the resonant and non-resonant parts, respectively. Once the forcing functions are separated, only the non-resonant parts of F_{1b_i} and F_2 are evaluated in the first and second order equations given by Eqs. (2.4) and (2.6) yielding solutions which are devoid of secularities. In order to remove the resonant part of the forcing functions, the frequency spectra of the system must be known. To aid in this part of the analysis, a highly eficient eigenvalue routine using Lanczos coordinates is incorporated to obtain a reduced system tridiagonal eigenproblem [14]. The resonant part is then removed by weighting all coefficients in the Fourier series which fall within a designated range of the system natural frequencies [13]. That is, coefficients which are very close to the natural frequencies are almost entirely eliminated whereas coefficients which are separated from the natural frequencies are unaffected. Applicable frequency weighting windows include cosine and $(cosine)^2$. This procedure provides an effective and efficient procedure for eliminating secularities from PFEM so all statistical results are bounded. Another advantage to using a Lanczos coordinate reduced basis is the solution of a reduced system of equations [15].

III. PFEM FOR NONLINEAR STATICS. The PFEM equations for nonlinear statics of a continuum, incorporating material nonlinearities, can be derived

using the approach followed in Section II. The discretized equilibrium equations governing the nonlinear statics are:

$$\underset{\sim}{f}(\underset{\sim}{d},\ \underset{\sim}{b}) = \underset{\sim}{F}(\underset{\sim}{b}) \tag{3.1}$$

where $\underset{\sim}{f}$, $\underset{\sim}{F}$ and $\underset{\sim}{d}$ are the internal force, external force and displacement vectors respectively and $\underset{\sim}{b}$ is the discretized random vector of size q, [9].

The randomness could arise from loading and/or material properties. The zeroth, first and second-order equations corresponding to Eq. (3.1) are:

Zeroth Order Equation

$$\underset{\sim}{\bar{f}} = \underset{\sim}{\bar{F}} \tag{3.2}$$

First-Order Equation

$$\underset{\sim}{\bar{K}}\ \underset{\sim}{\bar{d}}_{b_i} = \underset{\sim}{\bar{F}}_{i+2} \qquad\qquad i = 1, \ldots, q \tag{3.3a}$$

and

$$\underset{\sim}{\bar{F}}_{i+2} = \underset{\sim}{\bar{F}}_{b_i} - \underset{\sim}{\bar{f}}_{b_i} \qquad\qquad i = 1, \ldots, q \tag{3.3b}$$

where $\underset{\sim}{\bar{K}}$ is the tangent stiffness matrix.

Second-Order Equation

$$\underset{\sim}{\bar{K}}\ \underset{\sim}{\bar{d}}_2 = \underset{\sim}{\bar{F}}_2 \tag{3.4a}$$

where

$$\underset{\sim}{\bar{d}}_2 = \frac{1}{2} \sum_{i,j=1}^{q} \underset{\sim}{\bar{d}}_{b_i b_j}\, \mathrm{Cov}(b_i, b_j) \tag{3.4b}$$

and

$$\underset{\sim}{\bar{F}}_2 = \sum_{i,j=1}^{q} \left\{ \frac{1}{2} \underset{\sim}{\bar{F}}_{b_i b_j} - \frac{1}{2} \underset{\sim}{\bar{f}}_{b_i b_j} - \underset{\sim}{\bar{K}}_{b_i} \underset{\sim}{\bar{d}}_{b_j} \right\} \mathrm{Cov}(b_i, b_j) \ . \tag{3.4c}$$

The computational effort in solving Eqs. (3.3) through (3.4) can be reduced significantly by transforming the full covariance matrix, $\mathrm{Cov}(b_i, b_j)$, to a diagonal variance matrix, $\mathrm{Var}(c_i)$ [9]. Usually, only n (n<q) <u>highest</u> values of $\mathrm{Var}(c_i)$ are necessary [9,10]. Using the random vector $\underset{\sim}{c}$, the first

254

and second order equations are simplified to:

First Order Equations

$$\underset{\sim}{\overline{K}}\, \overline{\underset{\sim}{d}}_{c_i} = \overline{\underset{\sim}{F}}_{i+2} \qquad\qquad i = 1, \dots, q \qquad\qquad (3.5a)$$

where

$$\overline{\underset{\sim}{F}}_{i+2} = \overline{\underset{\sim}{F}}_{c_i} - \overline{\underset{\sim}{f}}_{c_i} \qquad\qquad i = 1, \dots, q \qquad\qquad (3.5b)$$

and

$$\overline{\underset{\sim}{f}}_{c_i} = \int_{\Omega} \underset{\sim}{B}^T \, \overline{\underset{\sim}{\sigma}}_{c_i} \Big|_{\underset{\sim}{d} = \overline{\underset{\sim}{d}}} \, d\Omega \quad . \qquad\qquad (3.5c)$$

Second-Order Equations

$$\underset{\sim}{\overline{K}}\, \overline{\underset{\sim}{d}}_2 = \overline{\underset{\sim}{F}}_2 \qquad\qquad (3.6a)$$

where

$$\overline{\underset{\sim}{F}}_2 = \sum_{i=1}^{q} \{ \tfrac{1}{2} \overline{\underset{\sim}{F}}_{c_i c_j} - \tfrac{1}{2} \overline{\underset{\sim}{f}}_{c_i c_j} - \overline{\underset{\sim}{K}}_{c_i} \overline{\underset{\sim}{d}}_{c_i} \} \mathrm{Var}(c_i) \qquad\qquad (3.6b)$$

$$\overline{\underset{\sim}{K}}_{c_i} = \int_{\Omega} \underset{\sim}{B}^T \overline{\underset{\sim}{C}}_{Tc_i} \underset{\sim}{B} \, d\Omega \qquad\qquad (3.6c)$$

and

$$\overline{\underset{\sim}{f}}_{c_i c_j} = \int_{\Omega} \underset{\sim}{B}^T \, \overline{\underset{\sim}{\sigma}}_{c_i c_j} \Big|_{\underset{\sim}{d} = \overline{\underset{\sim}{d}}} \, d\Omega \quad . \qquad\qquad (3.6d)$$

Once $\overline{\underset{\sim}{d}}$, $\overline{\underset{\sim}{d}}_{c_i}$ and $\overline{\underset{\sim}{d}}_2$ are obtained, the mean and autocovariance matrices of the displacement can be computed from:

$$E[\underset{\sim}{d}] = \overline{\underset{\sim}{d}} + \overline{\underset{\sim}{d}}_2 \qquad\qquad (3.7a)$$

and

$$\mathrm{Cov}(d^i, d^j) = \{ \sum_{r=1}^{q} \overline{d}_{c_r}^{\,i} \, \overline{d}_{c_r}^{\,j} \, \mathrm{Var}(c_r) \} \quad . \qquad\qquad (3.7b)$$

Next, the mean and autocovariance matrices of the stress can be similarly computed. At any point (usually an integration point) in the domain Ω, $\bar{\sigma}$ is computed from Eq. (3.2) and:

$$[\bar{\sigma}]_{c_i} = \bar{\sigma}_{c_i}\Big|_{\underset{\sim}{d}=\bar{\underset{\sim}{d}}} + \bar{C}_T \, B \, \bar{\underset{\sim}{d}}_{c_i} \tag{3.8a}$$

$$[\bar{\sigma}]_{c_i c_i} = \bar{\sigma}_{c_i c_i}\Big|_{\underset{\sim}{d}=\bar{\underset{\sim}{d}}} + \bar{C}_{Tc_i} \, B \, \bar{\underset{\sim}{d}}_{c_i} + \bar{C}_T \, B \, \bar{\underset{\sim}{d}}_{c_i c_j} \tag{3.8b}$$

where [ ] denotes total derivative and $\underset{\sim}{C}_T$ represents the tangent constitutive matrix. Thus,

$$E[\underset{\sim}{\sigma}] = \bar{\underset{\sim}{\sigma}} + \bar{\underset{\sim}{\sigma}}_2 \tag{3.9a}$$

where

$$\bar{\underset{\sim}{\sigma}}_2 = \frac{1}{2} \sum_{i=1}^{q} [\bar{\underset{\sim}{\sigma}}]_{c_i c_i} \, \text{Var}(c_i) \tag{3.9b}$$

and

$$\text{Cov}(\sigma^i, \sigma^j) = \left\{ \sum_{r=1}^{q} [\vec{\sigma^i}]_{c_r} [\vec{\sigma^j}]_{c_r} \, \text{Var}(c_r) \right\} . \tag{3.9c}$$

Evaluation of Internal Force/Stress Derivatives

It is seen that, in all the first and second-order equations derived in Eqs. (3.5a) and (3.6b), the derivatives of the internal force and stress are required. Direct evaluation of these derivatives are not possible, clearly, as the internal force and stress are implicit functions of the random vector $\underset{\sim}{c}$. Usually in such cases, these derivatives are replaced by their finite-difference counterparts [16,17]. Employing central-difference approximations,

$$\bar{\underset{\sim}{\sigma}}_{c_i}\Big|_{\underset{\sim}{d}=\bar{\underset{\sim}{d}}} \approx \frac{1}{2\Delta c_i} (\bar{\underset{\sim}{\sigma}}^+ - \bar{\underset{\sim}{\sigma}}^-)\Big|_{d=\bar{d}} \tag{3.10a}$$

and

$$\bar{\underset{\sim}{\sigma}}_{c_i c_i}\Big|_{\underset{\sim}{d}=\bar{\underset{\sim}{d}}} \approx \frac{1}{\Delta c_i \Delta c_i} (\bar{\underset{\sim}{\sigma}}^+ - 2\bar{\underset{\sim}{\sigma}} + \bar{\underset{\sim}{\sigma}}^-)\Big|_{\underset{\sim}{d}=\bar{\underset{\sim}{d}}} \tag{3.10b}$$

256

where

$$\bar{\sigma}^+ = \sigma(\bar{c} + \Delta c_i) \tag{3.10c}$$

$$\bar{\sigma}^- = \sigma(\bar{c} - \Delta c_i) \tag{3.10d}$$

and Δc_i are defined as

$$\Delta c_i = (0, \ldots, 0, \Delta c_i, 0, \ldots, 0)^T . \tag{3.10e}$$

The first and second-order derivatives of the internal force can then be obtained from Eqs. (3.5c) and (3.6c), respectively. The derivatives of the tangent constitutive matrix, in Eqs. (3.6c) and (3.8b), can also be approximated similarly.

IV. NUMERICAL EXAMPLES. The method presented in Section II for the elimination of secularities in transient PFEM is demonstrated by application to a multiple degree-of-freedom transmission tower. The effectiveness of the method for removing secularities from PFEM (NOS), is compared to the standard PFEM solution with secularities (SEC), and a Monte Carlo Simulation (MCS) with 400 samples. The random material properties are incorporated into the system by choosing Young's Modulus for elements 1-4 and 6-9 as uncorrelated normal random variables with a coefficient of variation of 5%. Rayleigh stiffness proportional damping is added to the system enabling the model to incorporate random stiffness and random damping. The performance of the method is presented in Figs. 1 and 2 for sinusoidal excitation.

The problem statement is presented in Fig. 1 for a 15 node/32 bar transmission tower with 26 degrees-of-freedom. The system has a first mode natural frequency of 8.7 cps and Rayleigh stiffness proportional damping with damping ratio equivalent to 0.1% of first mode. The expectation and variance of the x-displacement of node 2 are shown in Figs. 2a and 2b for a $(\text{cosine})^2$ weighting window, respectively. Since the second order solution is negligible compared to the zeroth order solution, secularities are only slightly evident in the expectation. In Fig. 2b, the variance of displacement exhibits secularities in the SEC which die out after 6 secs. due to damping. Initially all three methods are in agreement but the SEC begins to deviate from the MCS due to secularities until they are damped away. The method presented in this paper (NOS) removes the secularities from the SEC bringing it into agreement with the MCS. Initially, the NOS removes too much from SEC which is probably due to the solution being heavily dominated by the transient part. The method presented is valid for coefficients of variation up to 20% as in the PFEM.

In the next application, the PFEM procedure for nonlinear statics is demonstrated. The problem analyzed is an elastic-plastic plate with a circular hole and subjected to uniform, compressive loading (Fig. 3). The load is assumed to be random with a coefficient of variation of 10% and a correlation length (λ) of 3L (Fig. 3). The response statistics viz., mean and variance with respect to incremental loading and the spatial correlation of the response are studied. The mean and variance of the displacement, at Node 400, are plotted in Figs. 4a and 4b. These results show good agreement with those obtained by Monte Carlo Simulation (MCS) [4,10] of 400 realizations. The maximum coefficient of variation of the displacement is found to be ~10%.

257

The mean and variance of the compressive stress, in Element 15, are plotted in Figs. 4c and 4d. The mean stress is in good agreement with the simulation results, whereas the variance of stress shows some disagreement, particularly at larger loads. As the load is increased, the variance of stress increases and, after a certain load, starts to decrease. Since the material is assumed to be elastic-plastic, with a ratio of elastic modulus to plastic modulus as 100, it is nearly perfectly plastic for large strains. Therefore, once the material starts yielding at a point, the stress is nearly bounded above by the yield-stress. This causes the variance of stress to fall to a near-zero level, with increasing loading (Fig. 4d). The maximum coefficient of variation of the stress is also found to be ~10%.

The displacement correlation (w.r.t. Node 400) along the y-axis and the stress correlation (w.r.t. Element 7) along the x-axis are plotted in Figs. 5a and 5b. The displacement shows almost complete correlation (i.e., 1.0). However, while the stress shows complete correlation near Element 7, it drops drastically to very low correlations near the ends. The elements near the circular hole are in a plastic state and the stresses in these elements are near the yield stress. At the same time the elements far from the hole are elastic and the stresses vary appreciably with changes in load. The correlation between the elastic stress and the plastic stress, which changes very little with load , is very low and this explains the low stress correlation near the hole. The low stress correlation near the far end of the x-axis (Fig. 5b) seems due to the low variance of stress there. The mean stress and the variance of stress along the x-axis are plotted in Figs. 5c and 5d, respectively, for a particular load. The stress variance is low at both ends and in between, near the hole, it peaks. The stress in this region is in the transition state from elastic to plastic and so the stress variance is high.

V. CONCLUSIONS. The validity of PFEM, for uncertainties as large as 10% (i.e., coefficient of variation is 10%) and under substantial material nonlinearity, has been demonstrated in the previous section. Also, the effectiveness of the scheme in removing secularities from the transient statistics is brought out. Based on this scheme, extension can be made to remove secularities from nonlinear transient statistics as well. Also, the PFEM can be extended to handle geometric randomness. Efforts are being made to achieve these two goals.

The PFEM and related procedures [5-13] have been applied in the past to study the effect of randomness in structural dynamics, linear and nonlinear response of continua, and buckling and collapse analysis. While such wide applications of PFEM in structural mechanics have been achieved, from the point of view of reliability and failure analysis the statistical aspects of fracture mechanics assume importance. Numerical methods, such as PFEM, for studying these aspects are very scarce. The fracture related quantities such as fracture toughness, initial and ultimate yield stress, the number, size and orientation of the cracks, voids and inclusions are usually hard to determine exactly. These and other quantities, which govern the crack growth, rate of crack growth, the direction of crack propagation and the eventual failure of the structure, can be modelled as random material or geometric quantities. Fracture studies, incorporating such randomness in PFEM, could give an insight on the fracture statistics. Based on the experience obtained so far, such studies using PFEM, seems promising.

REFERENCES

1. Lin, Y. K., _Probabilistic Theory of Structural Dynamics_, McGraw-Hill, New York, 1967.

2. Vanmarcke, E., _Random Fields, Analysis and Synthesis_, MIT Press, Second Printing, 1984.

3. Ang, A. H. S. and Tang, W. H., _Probability Concepts in Engineering Planning and Design_, Vol. I: _Basic Principles_, John Wiley and Sons, 1975.

4. Ma, F., "Extension of Second Moment Analyis to Vector-Valued and Matrix-Valued Functions," submitted to _International Journal of Nonlinear Mechanics_.

5. Nakagini, S., Hisada, T. and Toshimitsu, K., "Stochastic Time-History Analysis of Structural Vibration with Uncertain Damping," ASME, PVP-Vol. 93, pp. 109-120, 1984.

6. Lawrence, M. A., "A Basis Random Variable Approach to Stochastic Finite Elements," to be presented at First World Congress on Computational Mechanics, September 1986, Austin, Texas.

7. Liu, W. K., Belytschko, T. and Mani, A., "A Computational Method for the Determination of the Probabilistic Distribution of the Dynamic Response of Structures," _Computer-Aided Engineering_, ASME PVP-Vol. 98 (5), pp. 243-248, 1985.

8. Liu, W. K., Belytschko, T. and Mani, A., "Probabilistic Finite Element Methods for Nonlinear Structural Dynam cs," _Computer Methods in Applied Mechanics and Engineering_, 56, pp. 61-81, 1986.

9. Liu, W. K., Belytschko, T. and Mani, A., "Random Field Finite Elements," to appear in _International Journal of Numerical Methods in Engineering_.

10. Liu, W. K., Belytschko, T. and Mani, A., "Applications of Probabilistic Finite Element Methods in Elastic/Plastic Dynamics," to appear in _Journal of Engineering for Industry_.

11. Liu, W. K., Belytschko, T., Mani, A. and Besterfield, G., "A Variational Principle for Probabilistic Mechanics," to appear in the text _Finite Element Methods for Plate and Shell Structures_, Vol. 2: _Formulations and Algorithms_, eds. T. J. R. Hughes and E. Hinton, Pineridge Press, Swansea, U.K.

12. Liu, W. K., Besterfield, G. and Belytschko, T., "A Probabilistic Hu-Washizu Variational Principle," to be presented at 28th AIAA Structural Dynamics and Materials Conference, April 1987, Monterey, California.

13. Liu, W. K., Belytschko, T. and Besterfield, G., "Transient Probabilistic Systems," to be presented at First World Congress on Computational Mechanics, September 1986, Austin, Texas.

14. Ojalvo, I. U., "Proper Use of Lanczos Vectors for Large Eigenvalue Problems," <u>Computers and Structures</u>, Vol. 20, pp. 115-120, 1985.

15. Nour-Omid, B. and Clough, R. W., "Dynamic Analysis of Structures Using Lanczos Co-ordinates," <u>Earthquake Engineering and Structural Dynamics</u>, Vol. 12, pp. 565-577, 1984.

16. Arora, J. S. and Hang, E. J., "Efficient Methods of Optimal Structural Design," <u>Journal of Engineering Mechanics Division</u>, ASCE, 104 (EM3), pp. 663-680, 1978.

17. Haftka, R. T., "Techniques for Thermal Sensitivity Analysis," <u>International Journal for Numerical Methods in Engineering</u>, Vol. 17, pp. 71-80, 1981.

Fig. 1   Problem Statement 1: Transmission Tower with 15 Nodes/32 Bars.
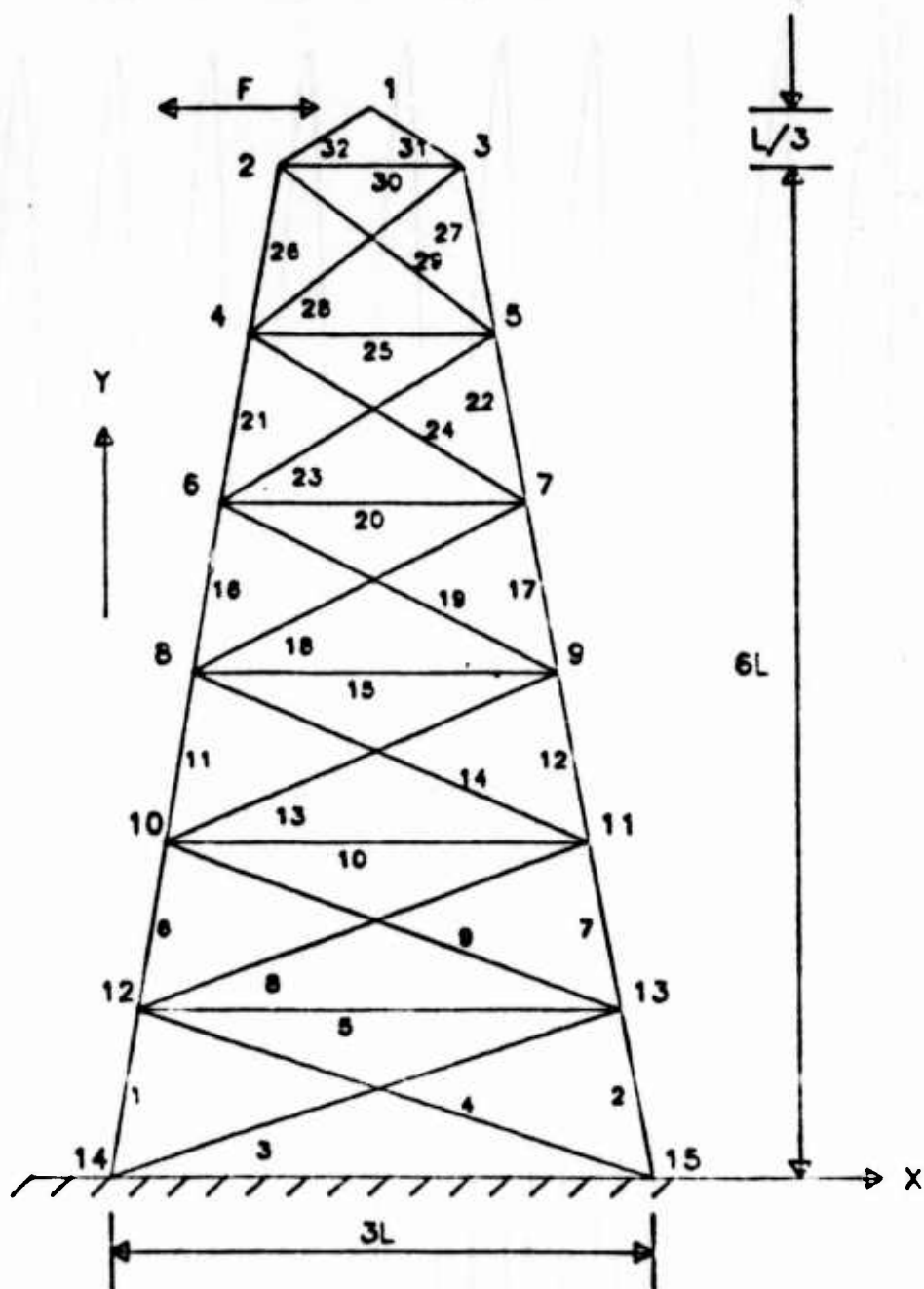
DENSITY (ρ) = 0.00776
AREA (A) = 6
YOUNGS MODULUS (E) = 30000000
LENGTH (L) = 60
FIRST NAT. FREQ. (ω_1) = 8.7 CPS
SINUSOIDAL LOADING (F) = 2 CPS
COEFF. OF VARIATION = 5 %
MCS SAMPLES = 400
RANDOM VARIABLE = E
ELEMENTS 1−4,6−9

Comparison of the Expectation of Node 2 x-Displacement, between SEC, MCS and NOS.

EXPECTATION OF NODE 2 X-DISPLACEMENT

Fig. 2a
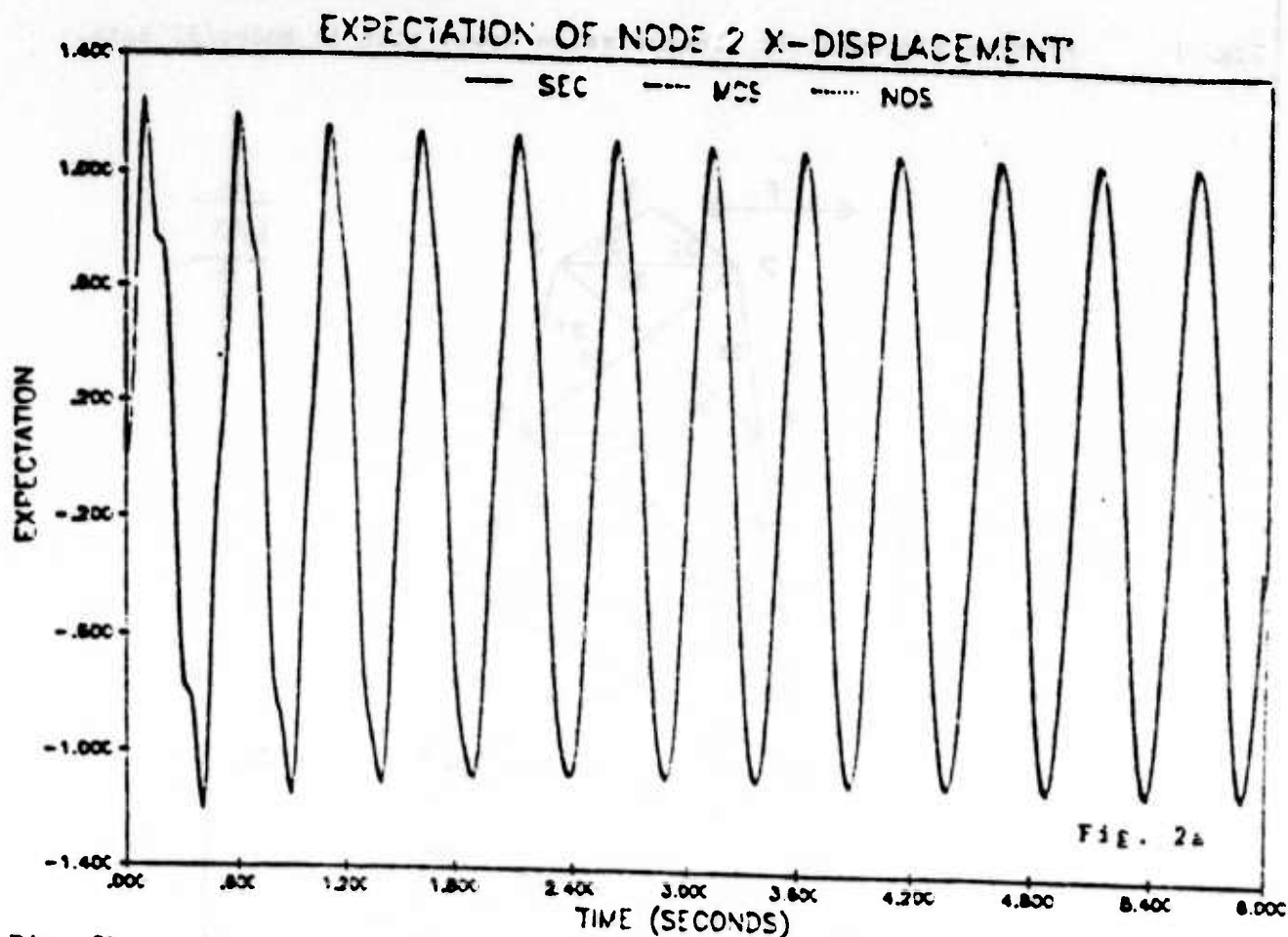
Fig. 2b Comparison of the Variance of Node 2 x-Displacement, between SEC, MCS, and NOS.

VARIANCE OF NODE 2 X-DISPLACEMENT

Fig. 2b

262
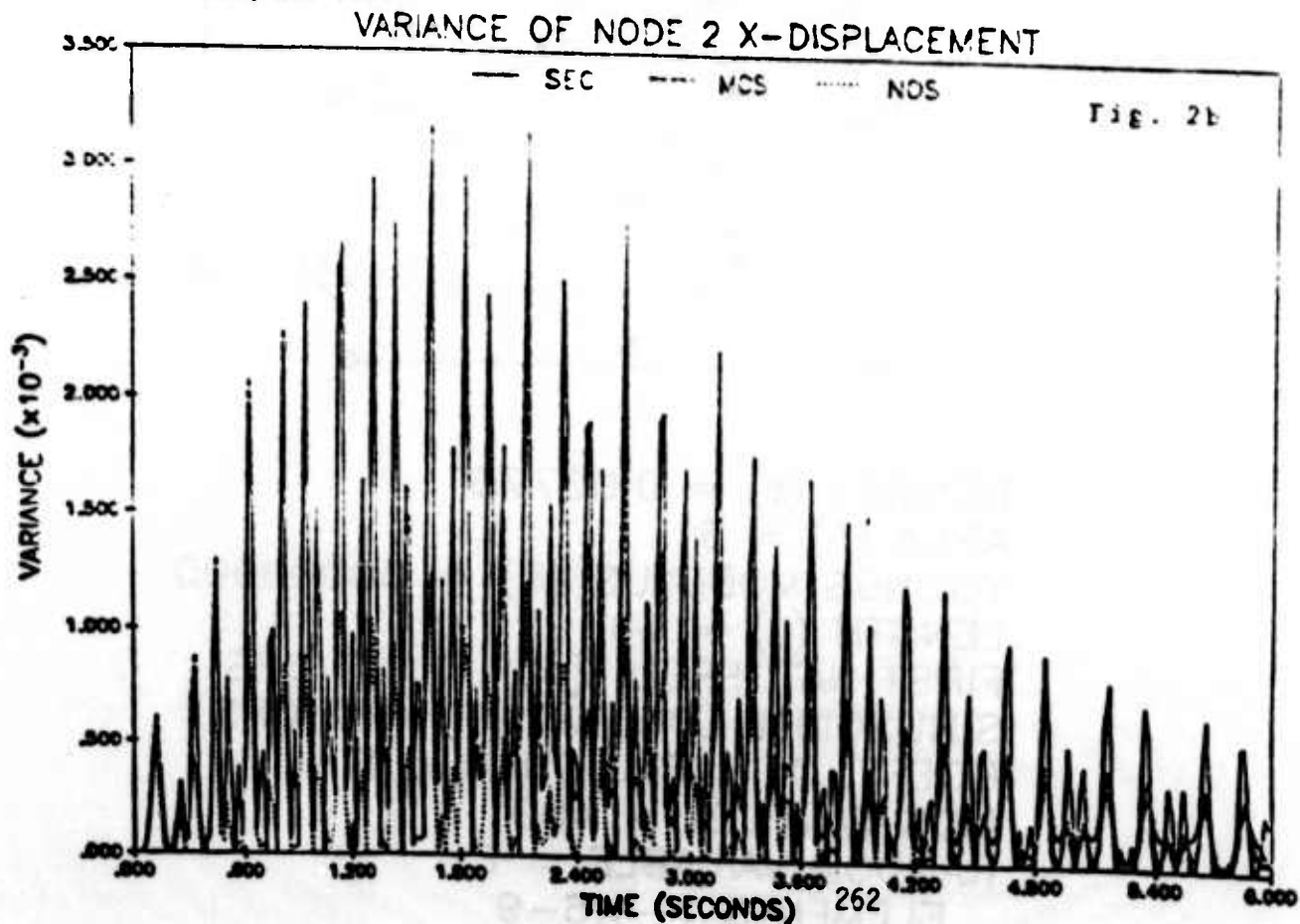
Fig. 3   Problem Statement 2: Elastic-Plate with a Circular Hole.

Elastic Plastic Plate with a Hole
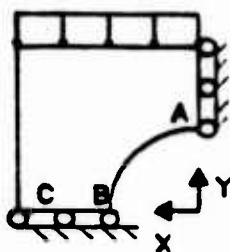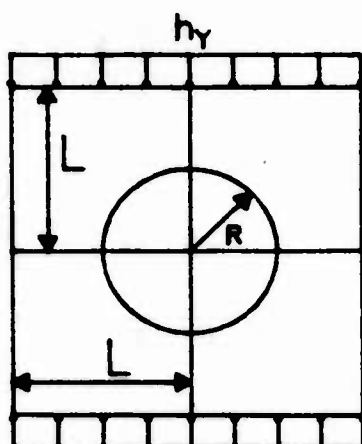
E $= 30. \times 10^6$
E_T $= 30. \times 10^4$
σ_Y $= 25000.0$
(Isotropic Hardening)
ν $= 0.3$
L $= 6.0$, R $= 3.0$

4 Node 2D Cont. Element
    (Plane Stress)
400 Nodes, 360 Elements
Node 400   Point A
Element 15 Point B
Element 7 Point C

Random Load Characteristics

Size of Random Load Vector (q)= 12
Coefficient of Variation= 0.10

Load Steps	0001	0002	0003	0004	0005	0006	0007	0008
Mean Load (h_Y)	2000	4000	4100	4200	4300	4400	4500	4600

Spatial Correlation of Random Load

$R(x_i, x_j) = \exp(-ABS(x_i - x_j)/\lambda)$
$\lambda = 3L$

263

Fig. 4a Comparison of the Mean Displacement at Node 400, between PFEM and MCS.

Fig. 4b Comparison of the Variance of Displacement at Node 400, between PFEM and MCS.
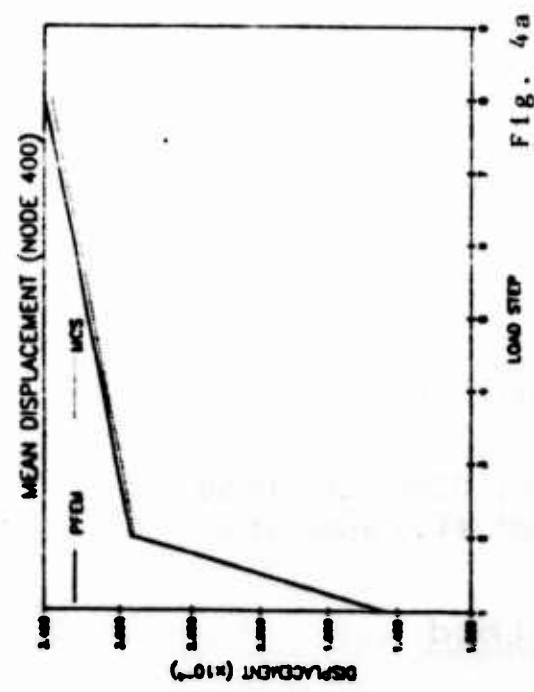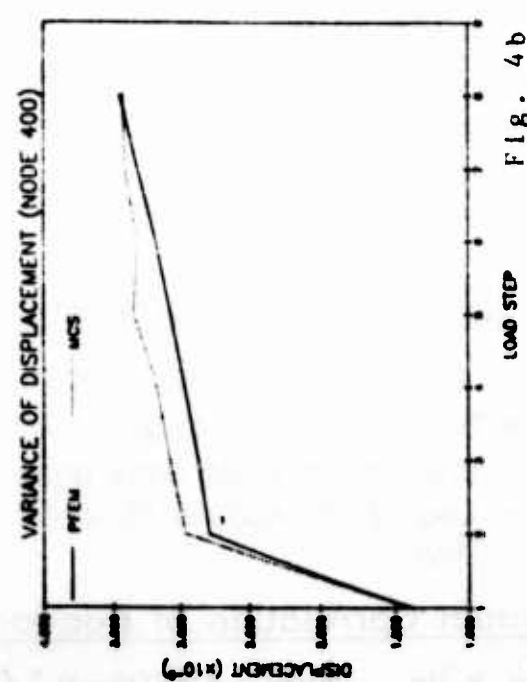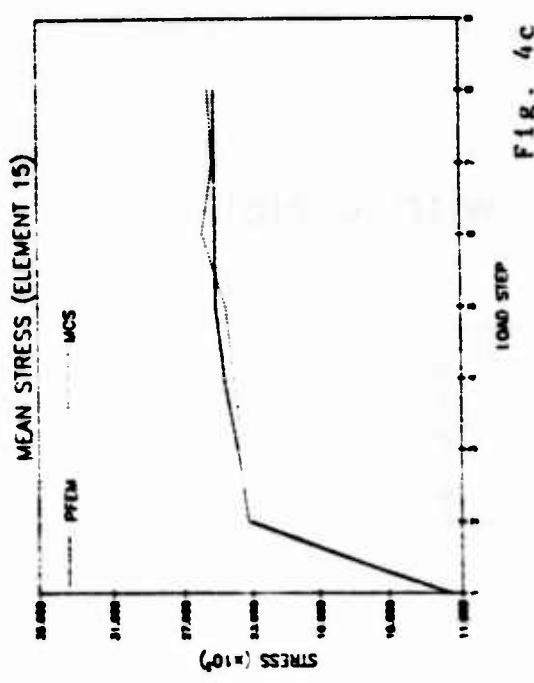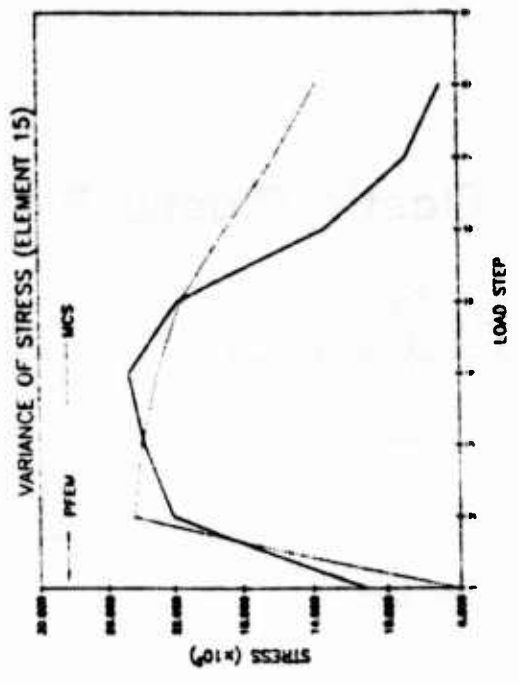
Fig. 4c Comparison of the Mean Stress in Element 15, between PFEM and MCS.

Fig. 4d Comparison of the Variance of Stress in Element 15, between PFEM and MCS.

Fig. 5a   Spatial Correlation of Displacement along y-axis, w.r.t. the
           Displacement at Node 400, by PFEM and MCS.

Fig. 5b   Spatial Correlation of Stress along x-axis, w.r.t. the Stress in
           Element 7, by PFEM and MCS.
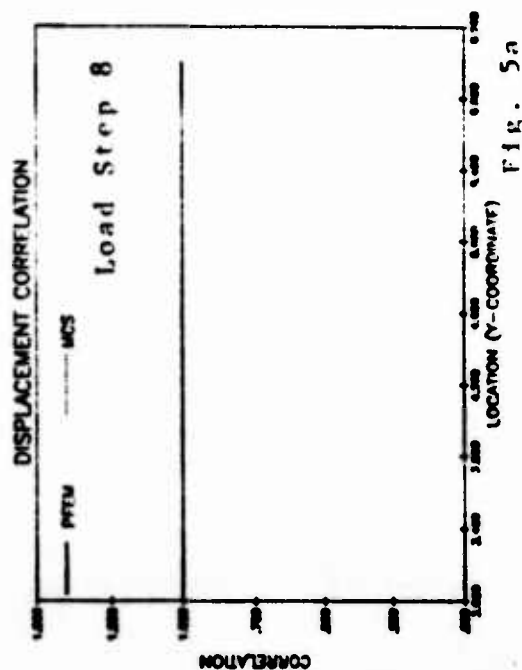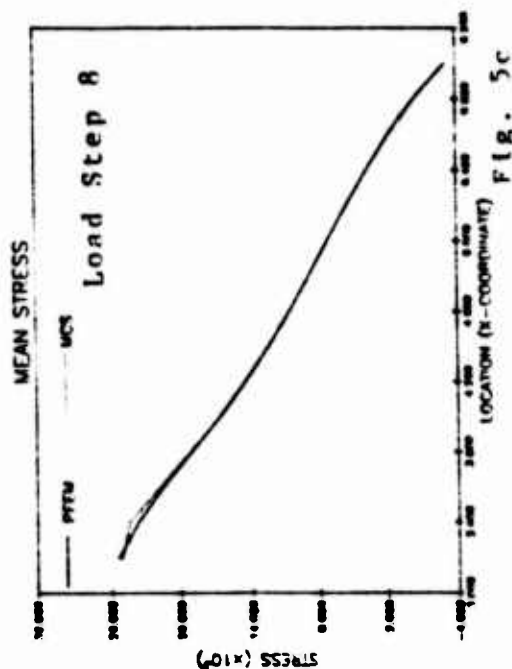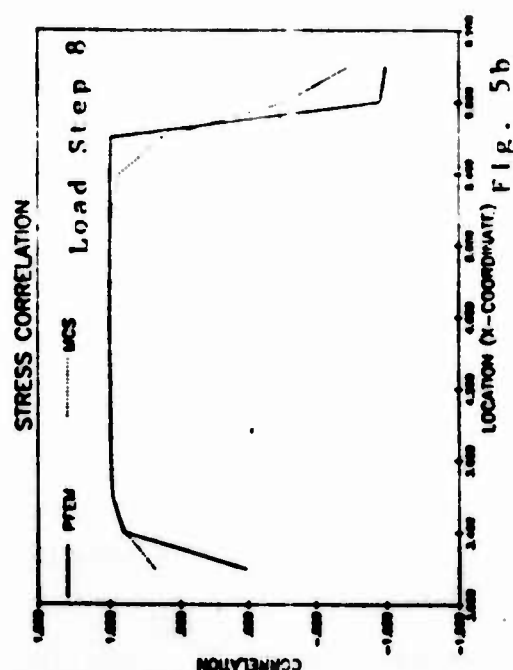
Fig. 5c   Comparison of the Mean Stress along x-axis at load step 8, between
           PFEM and MCS.
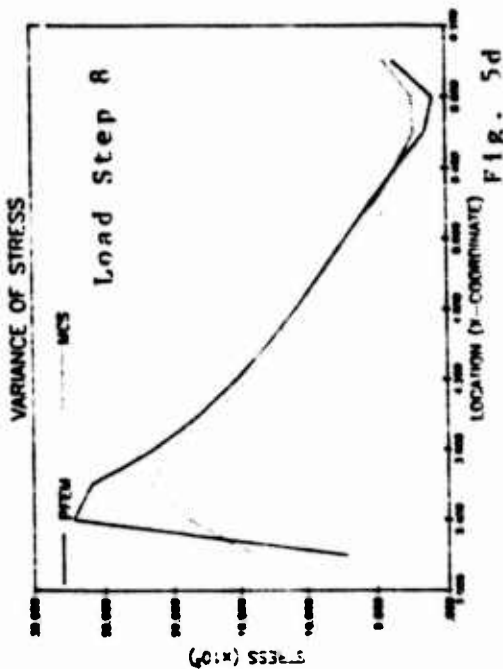
Fig. 5d   Comparison of the Variance of Stress along x-axis at load step 8,
           between PFEM and MCS.

265

LIMIT THEOREMS FOR THE SIZE EFFECT
IN THE LIFETIME DISTRIBUTION
OF A FIBROUS COMPOSITE

S. Leigh Phoenix and Chia-Chyuan Kuo[*]
Sibley School of Mechanical and Aerospace Engineering
Cornell University
Ithaca, New York 14853

ABSTRACT. A composite material is a parallel arrangement of stiff brittle fibers in a flexible matrix. Under load fibers fail, and the loads of failed fibers are locally redistributed onto nearby survivors through the matrix. In this paper we develop a new technique for computing the probability of failure under a previously studied model of the failure process. In this model, known as the chain-of-bundles model, failure occurs when all fibers fail in at least one bundle. A recursion and limit theorem are obtained which apply separately to static strength and fatigue lifetime depending on the composite loading and the probability model for the failure of individual fibers under their own loads. The limit theorem yields an approximation for the distribution function for composite lifetime which is of the form $1 - [1 - W(t)]^{mn}$ where $W(t)$ is a characteristic distribution function and mn is the composite volume, reflecting a size effect. A similar result holds also for static strength. In both cases such a result was conjectured several years ago. This limit theorem is obtained from the recursion upon applying a key theorem in the theory of the renewal equation. In the proofs three technical conditions arise which must be verified in specific applications. In the case of static strength these conditions are quite easy to verify, but in the case of fatigue lifetime the verification is generally difficult, and entails considerable numerical computation.

[*] Present Address: Kendall Company, 95 West St., Walpole, MA 02081.

I. INTRODUCTION. In this paper we present a new recursive technique and a limit theorem for an earlier, idealized model of the failure process in a fibrous composite material. For previous work on static strength see Harlow and Phoenix (1978, 1981, 1982), Harlow (1985), and Smith (1980, 1982, 1983), and in the case of time dependent fatigue see Tierney (1982) and Phoenix and Tierney (1983). The present paper is an abbreviated version of a forthcoming paper by Kuo and Phoenix (1987).

To review the model, we consider a simple composite which is an arrangement of n parallel filaments along a line to form a planar tape, or in a circle to form a tube. The loading, which is a specified function of time, is simple tension in the fiber direction. The actual failure process to be modeled begins when fibers fail randomly in both time and position, and locally their original loads are transferred to adjacent fibers which then become overloaded. In time some of these overloaded fibers fail too, and clusters of several contiguous breaks appear. Eventually one of these clusters grows to an unstable size, turns into a catastrophic crack and fails the composite.

To model this failure process the composite is partitioned into a series of m short sections called bundles, each containing n fibers elements of length δ, the effective load transfer length. The failure process is localized within the bundles, and the composite is treated as a weakest-link arrangement of its m bundles, each carrying the externally applied load. The mn fiber elements are treated as statistically independent entities under an identical prescribed load history on each (though their failure times within a bundle will be dependent because of the load transfer process which will cause the individual fiber load histories to differ); thus the bundles are statistically independent. Throughout we speak of load on a 'force-per-fiber' basis; that is, the load is the total external force on the composite divided by n. Henceforth our modeling will be in terms of the fiber elements, and for brevity in the notation we refer to these as the 'fibers'.

Load-sharing rule for fibers. If the bundle load is ℓ, a surviving fiber carries load $K_r \ell$ where K_r is called a load concentration factor, and r is the number of consecutive failed fibers immediately adjacent to this survivor (counting on both sides). Also set $K_0 = 1$.

As in most previous analyses, the first load-sharing rule we consider is

(1.1) $K_r = 1 + r/2$, $r = 0,1,2,... $,

wherein the load of a failed fiber is redistributed in equal portions onto its two nearest surviving neighbors, one on each side. In linear bundles which may have fibers failed at the bundle edge, this rule has a slight deficiency since there are no exterior fibers to carry some of the shifted load. To avoid these difficulties we will also consider circular bundles which have no such edges. Here we need to take care in the situation where only one last fiber remains since one would expect that fiber to carry the total load $n\ell$, whereas $K_{n-1} \ell = (n+1)\ell/2$. Thus we will consider instead $K^*_{n-1} = n$.

An alternate rule is based on elastic calculations in a planar lattice by Gotlib, El'yashevich and Svetlov (1973), namely $K_r = (1+r)^{1/2}$, $r = 0,1,2,...$. This rule more accurately models the fiber loads once r becomes large and reflects results from fracture mechanics where the stresses at the crack tip grow as the square root of the crack length.

An important feature exploited in our later analysis is that none of the load of a failed fiber is redistributed beyond the two flanking nearest survivors. The mechanical analysis of Hedgepeth (1961) shows this assumption to be somewhat oversimplified (as it would be for the alternate rule), but the results of Pitt and Phoenix (1983) suggest that this shortcoming is minor, provided that most of the redistributed load appears on the nearest survivors.

Load histories. We let $\ell(t)$, $t \geq 0$ be the load history which we apply to the composite. In general $\ell(t)$ can be any positive function of $t \geq 0$. However, in the setting of static strength we work with the linear load $\ell(t) = t$, since in this case the failure time and the load at failure will be identical. In fatigue lifetime the simplest model is $\ell(t) = L$, $t \geq 0$. Note that the actual fiber load histories will differ from $\ell(t)$ as neighboring fibers fail.

Distribution functions for lifetime. We let $H_{m,n}(t;\ell)$ be the distribution function for the failure time of the composite under load $\ell(t)$, $t \geq 0$. Also let $G_n(t;\ell)$ be the distribution function for the

failure time of a single bundle. Since the individual fibers, and thus the bundles, will be statistically independent entities we have the simple connection

(1.2) $H_{m,n}(t;\ell) = 1 - [1 - G_n(t;\ell)]^m$, $\quad t \geq 0$.

The main task is thus to calculate $G_n(t;\ell)$.

Model for the failure of fibers. We let $F(t;\lambda)$, $t \geq 0$ be the distribution function for the failure time of a single fiber under its load history $\lambda(t)$, $t \geq 0$. To model fiber failure it is convenient to use the concept of a _standard representative fiber_ as introduced by Tierney (1982). First associate with the fiber a random variable Z which follows the unit exponential distribution

(1.3) $\tilde{F}(z) = 1 - \exp\{-z\}$, $\quad z \geq 0$.

Then given the load history $\lambda(t)$, $t \geq 0$ on the fiber, let $\Theta(t;\lambda)$ be the cumulative hazard function (CHF) for failure, and assume it to be a non-anticipating _functional_ of λ. Also we assume $\Theta(t;\lambda)$ is increasing and right-continuous in $t \geq 0$ for fixed λ, and $\Theta(t;\lambda)$ is monotone in λ; that is, if $\lambda_1(s) \geq \lambda_2(s)$ for all $0 \leq s \leq t$ then $\Theta(t;\lambda_1) \geq \Theta(t;\lambda_2)$. Then under λ, the failure time T of the fiber is the smallest value of $T \geq 0$ for which

(1.4) $\Theta(T;\lambda) \geq Z$.

By this construction we have

(1.5) $F(t;\lambda) = 1 - \exp\{-\Theta(t;\lambda)\}$, $\quad t \geq 0$.

Under a common fiber load history λ the lifetimes of the individual fibers are assumed to be statistically independent; that is, the Z's are independent from fiber to fiber. However, in a bundle the individual fiber load histories will begin to differ as neighboring fibers fail, and the fiber lifetimes will become dependent. This is where the main complication arises.

In the setting of _static strength_, a fiber has random strength X which is independent of both its load history, and the strength of

270

other fibers. We assume X has distribution function $F_0(x)$, $x \geq 0$, which we write as

$$(1.6) \qquad F_0(x) = 1 - \exp\{-\Theta_0(x)\}, \quad x \geq 0$$

for some suitable increasing function $\Theta_0(x)$, $x \geq 0$; of course, $F_0(x)$ is common to all fibers. Then in this case the CHF becomes

$$(1.7) \qquad \Theta(t;\lambda) = \sup_{0 \leq s \leq t} \Theta_0 (\lambda(s)) .$$

Some specific cases of this model are as follows: The first and simplest case is known as the 'pure flaw' model which has been considered by Harlow (1985). In this case a fiber is assumed to have zero strength with probability ϕ and unit strength with probability $1-\phi$. Thus

$$(1.8) \qquad \Theta_0(x) = \begin{cases} 0 & , & x < 0 , \\ -\ln(1-\phi) & , & 0 \leq x < 1 , \\ \infty & , & 1 \leq x . \end{cases}$$

A second case is where fibers follow a Weibull distribution for strength:

$$(1.9) \qquad F_0(x) = 1 - \exp(-x^\gamma) , \quad x \geq 0$$

where $\gamma > 0$ is a constant. This is the model studied by Harlow and Phoenix (1978, 1981, 1982) and Smith (1980, 1982, 1983) among others. Then $\Theta_0(x) = x^\gamma$, $x \geq 0$. In both these cases the lifetime T and the strength X are identical under $\lambda(t) = t$, and the same will be true for the composite.

In the case of <u>time dependent fatigue</u> we may consider the CHF of the form

$$(1.10) \qquad \Theta(t;\lambda) = (\int_0^t \lambda(s)^\rho \, ds)^\beta , \quad t \geq 0 ,$$

where $\beta > 0$ and $\rho > 0$ are constants. Various versions of this model have been studied by Tierney (1982) and Kuo (1983).

<u>Failure process in a bundle</u>. For a bundle, we assign the independent random variables $Z_1, \ldots, Z_n$, one to each fiber, and assume these also follow (1.3). Then given particular realizations of the Z_i's, the bundle load history $\mathbf{L}$, and the load-sharing K_r's we can solve explicitly for the fiber failure times denoted $T_{(1)}, \ldots, T_{(n)}$ and for the bundle failure time $T_n = \max\{T_{(1)}, \ldots, T_{(n)}\}$. (See Phoenix and Tierney (1983) for example.) Note that while the bundle load history is $\mathbf{L}(t)$, $t \geq 0$, the individual fiber load histories $\lambda_1(t), \ldots, \lambda_n(t)$ may involve some $K_r \mathbf{L}(t)$ as neighbors fail, and these must be used in the analysis.

<u>Outline of the paper</u>. The key quantity to consider is $Q_n(t) \equiv 1 - G_n(t)$, $t \geq 0$, the probability a bundle survives to time t. (Henceforth we generally suppress $\mathbf{L}(t)$ in the notation unless germane.) In Section 2 we develop a recursion formula (Theorem 1) for $Q_n(t)$ for both planar and circular bundles. In Section 3 we obtain the main limit theorem, Theorem 2. In Section 4 we recast this theorem into a key approximation for $G_n(t)$ and $H_{m,n}(t)$ which involves two functions: a characteristic distribution function W(t) and a boundary function $\pi(t)$. The form of the approximation is thus $H_{m,n}(t) \cong 1 - [1 - W(t)]^{mn} \pi(t)^m$. Also $\pi(t) = 1$ for circular bundles, and typically deviates negligibly from one for planar bundles. This theorem and resulting approximation essentially confirm a conjecture first posed by Harlow and Phoenix(1978) in the static case and Tierney (1982) in the time dependent case. Harlow (1985) first confirmed the conjecture in the simplest case, the pure flaw model.

To use Theorem 2 in specific applications, three technical conditions, (3.3), (3.4) and (3.5) must be verified. Roughly speaking, these conditions involve showing that at time t the probability of having a lone survivor in a bundle of n fibers divided by the survival probability $[1 - W(t)]^n$ is small compared to one, and furthermore diminishes very rapidly as $n \rightarrow \infty$. In Section 5 we verify these conditions for the static cases, giving ranges for the model parameters under which the results hold. However, generally we cannot give justification in the time dependent model without introducing additional conditions which are physically justifiable, but seem to be irrelevant from numerical calculations.

272

II. RECURSION ANALYSIS. For the most part the dependence of all quantities on time t will be suppressed in the notation. We consider only values of t for which $F(t;\mathcal{L}) < 1$ since otherwise the problem is trivial.

Linear bundles. We first consider linear bundles wherein the fibers are arranged along a line from left to right. For a given fiber we let the symbols 'X' and 'O' denote failure and survival, respectively. A bundle of n fibers clearly has 2^n configurations of failed and surviving fibers. For example, for n = 6 a possible configuration is OXXOXO. Failure is defined as the configuration XXX...X and we let A_n be the set of all $2^n - 1$ remaining survival configurations, that is, all configurations of n fibers with at least one 'O'. Thus we formally define $G_n = Pr\{XXX...X\}$ and $Q_n \equiv Pr\{A_n\} = 1 - G_n$.

Next we let E_i be the configuration of i fibers which has all 'X's except for an 'O' at the very left, that is, $E_1 = \{O\}$, $E_2 = \{OX\}$, $E_3 = \{OXX\}$, and so on. In the analysis to follow we decompose A_n into the disjoint subsets $A_{n,1}$, $A_{n,2}$, ..., $A_{n,n}$ where $A_{n,i}$ contains all elements of A_n whose right-most i fibers are in the configuration E_i. Thus

(2.1) $$A_n = \bigcup_{i=1}^{n} A_{n,i}$$

and defining $Q_{n,i} = Pr\{A_{n,i}\}$ we have

(2.2) $$Q_n = \sum_{i=1}^{n} Q_{n,i}, \quad n \geq 1.$$

For two sets of survival configurations A and B we define the new set $A * B$ through the operation '*' as the set of all configurations generated by attaching a configuration from B to the right end of one from A. By inspection, we have the general recursive relationships

(2.3) $$A_{n+1,1} = A_n * \{O\}, \quad n \geq 1$$

and

(2.4) $$A_{n+1,i} = A_{n,i-1} * \{X\}, \quad 2 \leq i \leq n+1,$$

starting with $A_{1,1} = A_1 = \{O\}$. We will also use the operation '*' to join two configurations, rather than sets.

Decomposable configurations for linear bundles. A configuration Y in A_n is said to be decomposable if there exist two adjacent survivors or 'O's in Y. For example, $Y = \{OXOOX\}$ is decomposable whereas $Y' = \{OXOX\}$ is not. If Y is decomposable then clearly $Y = Y_1 * Y_2$ for some Y_1 in A_r and Y_2 in A_{n-r} where $1 \leq r \leq n-1$ and where Y_1 has a survivor at its right end while Y_2 has a survivor at its left end. The importance of this concept is that $Pr\{Y\} = Pr\{Y_1\}$ $Pr\{Y_2\}$. (To see this one must use the concept of standard representative fibers as described in Section 1.) In other words, the probability of a decomposable configuration occurring at time t is the product of the probabilities for the component configurations viewed as smaller distinct bundles.

For certain configurations which cannot be decomposed, we will later need bounding probabilities written in terms of probabilities for smaller configurations, as in the following lemma.

Lemma 1. Let $Y_1 \in A_n$ and $Y_2 \in A_m$ such that either Y_1 has an 'O' at its right end, or Y_2 has an 'O' at its left end. Then for $Y = Y_1 * Y_2 \in A_{n+m}$ we have

(2.5) $Pr\{Y\} \leq Pr\{Y_1\}\, Pr\{Y_2\}$.

Proof: To prove this lemma think in terms of the standard representative fibers of Section 1 where $Z_1, ..., Z_n$ and $Z_{n+1}, ...,Z_{n+m}$ are associated with the fibers which may yield the respective configurations Y_1 and Y_2, and altogether the configuration Y. First, if $Z_1, ...,Z_{n+m}$ have values such that $Y \in A_{n+m}$ results at time t where fiber n is surviving while fiber n+1 is failed, then these same values will automatically produce $Y_1 \in A_n$ and $Y_2 \in A_m$. The reverse, however, is not true in that some survival configuration other than Y may result for the bundle of size m+n. Hence $Pr\{Y\} \leq Pr\{Y_1\}\, Pr\{Y_2\}$, proving the lemma.

Special non-decomposable configurations. Two sets are crucial to the analysis: The first is $F_{n,i}$ which is the subset of $A_{n,i}$ whose elements have an 'O' on the left and are not decomposable.

The second is $R_{n,i}$ which is the subset of $A_{n,i}$ whose elements have an 'X' on the left and are also <u>not</u> decomposable. Define $f_{n,i} \equiv \Pr\{F_{n,i}\}$ and $r_{n,i} \equiv \Pr\{R_{n,i}\}$ and also set $r_{0,1} \equiv 1$, $r_{n,n} \equiv 0$ and otherwise $r_{j,i} \equiv f_{j,i} \equiv 0$ for all $j < i$.

By inspection we see that

$$(2.6) \qquad \overset{n}{\underset{i=2}{\cup}} F_{n,i} = \overleftarrow{R}_{n,1}, \quad n \geq 2,$$

where the symbol "$\leftarrow$" above a set means that each of its configurations has its entries written down in reverse order. For example, if $B = \{XO, XOXXO\}$ then $\overleftarrow{B} = \{OX, OXXOX\}$. (We also apply "$\leftarrow$" to a single configuration with the same meaning.) Furthermore, by studying survival configurations for small n we see the structure

$$F_{1,1} = \{O\}$$

$$(2.7) \qquad F_{2,1} = \{\emptyset\}$$

$$F_{n,1} = \overset{n-1}{\underset{i=2}{\cup}} F_{n-i,1} \cdot \overleftarrow{E}_i, \quad n \geq 3$$

where we assume $\emptyset \cdot Y = Y \cdot \emptyset = \emptyset$ for any Y in A_n. It is also true that

$$(2.8) \qquad F_{n+1,1} = \{O\} \cdot R_{n,1}$$

and

$$(2.9) \qquad R_{n+1,1} = (\overset{n-1}{\underset{i=2}{\cup}} R_{n,i} \cup \{X X \dots X\}) \cdot \{O\}.$$

We now obtain some key recursions.

<u>Lemma 2</u>. For linear bundles we have the recursion

$$(2.10) \qquad Q_{n,i} = \sum_{j=1}^{n} Q_{n-j,1} f_{j,i} + r_{n,i}, \quad 1 \leq i \leq n,$$

starting with $Q_{0,1} \equiv 1$ and $Q_{0,i} \equiv 0$ for $i \geq 2$.

<u>Proof</u>: It suffices to prove the case i = 1. If n = 1 the result is obvious. Next take n ≥ 2, and suppose $Y \in A_{n,1}$ but Y is not decomposable. Then either $Y \in F_{n,1}$ or $Y \in R_{n,1}$, accounting for j = n in the sum. On the other hand, if $Y \in A_{n,1} - (F_{n,1} \cup R_{n,1})$, then Y is decomposable and there exists some j such that $Y \in A_{n-j,1} * F_{j,1}$ and $1 \le j \le n-1$. The sum follows from disjointedness of the various configurations.

<u>Circular bundles</u>. In the case of circular bundles, the previous analysis must be modified. Again we label the fibers consecutively from 1 to n starting arbitrarily, but fibers 1 and n are now adjacent. We still write out a configuration in a linear fashion, though with this adjacency of the first and last fibers understood. This latter aspect forces a modification of the earlier concept of a decomposable configuration. To be decomposable a configuration Y for a circular bundle must now have either three adjacent survivors, or, two or more pairs of adjacent survivors.

We define A_n and Q_n as before but for n ≥ 2 let $A_{n,0}$ be the subset of A_n whose elements have two or more adjacent survivors. Furthermore, for n ≥ 3 we partition $A_{n,0}$ into two subsets $A_{n,01}$ and $A_{n,02}$ where $A_{n,01}$ contains exactly those elements of $A_{n,0}$ which are decomposable, and $A_{n,02}$ contains the rest. Lastly we let $Q_{1,0} = Pr\{O\}$, $Q_{2,0} = Pr\{OO\}$ and $Q_{n,0} = Pr\{A_{n,0}\}$, n ≥ 3.

<u>Lemma 3</u>. For circular bundles we have the recursion

$$(2.11) \qquad Q_{n,0} = n f_{n,1} + \sum_{j=1}^{n} Q_{n-j,0} f_{j,1} \, , \quad n \ge 1 \, ,$$

where $Q_{0,0} \equiv 0$ and $f_{n,1}$ is as defined for linear bundles.

<u>Proof</u>: See Kuo and Phoenix (1987).

<u>Theorem 1</u>. For both linear and circular bundles we have the recursion

$$(2.12) \qquad Q_n = \sum_{j=1}^{n} Q_{n-j} f_{j,1} + \xi_n \, , \quad n \ge 1 \, ,$$

starting with $Q_0 \equiv 1$, where for linear bundles

$$(2.13) \qquad \xi_n = \sum_{m=0}^{n} r_{n-m,1} r_{m,1} + \sum_{j=2}^{n} r_{n,j} - \sum_{m=1}^{n} \sum_{j=2}^{n} f_{n-m,1} r_{m,j}$$

and for circular bundles

276

(2.14) $\xi_n = n\, f_{n,1} + r_{n,1}^{(0)} - \displaystyle\sum_{j=1}^{n} r_{n-j,1}^{(0)}\, f_{j,1}$.

where $r_{n,1}^{(0)}$ is analogous to $r_{n,1}$ for linear bundles (and must include all unique rotations). The proof of this theorem is long and will appear in Kuo and Phoenix (1987).

III. BEHAVIOR OF Q_n AS n GROWS LARGE. We show here that for both linear and circular bundles Q_n has the structure $Q_n = \chi^{-n}(\pi + o_n)$ for suitable functions χ, π and o_n where $o_n \to 0$ as $n \to \infty$. Also, χ is the same for both bundles, and π, which apparently reflects edge effects, is identically one in the circular case. We begin with some key lemmas, definitions and assumptions.

Lemma 4. For linear bundles

(3.1) $\displaystyle\sum_{n=0}^{\infty} f_{n,1} \le 1$.

Proof: Recall (2.20) and take i = 1. Since $r_{0,1} = 1$ we have $\tilde{R}_1(s) > 1$ whence $\tilde{F}_1(s) < 1$. In view of (2.19) Abel's lemma (Karlin and Taylor (1975)) gives us (3.1).

Next let χ be the solution to

(3.2) $\displaystyle\sum_{n=1}^{\infty} f_{n,1}\, \chi^n = 1$

and note that $\chi \ge 1$ by Lemma 1, since $f_{0,1} = 0$.

Technical conditions. We now make some technical assumptions needed later. For a linear bundle of n fibers we recall E_n was the configuration {OXX ... X}. Let $e_n \equiv \Pr\{E_n\}$ and assume two conditions are satisfied, namely

(3.3) $\displaystyle\sum_{n=2}^{\infty} e_n\, \chi^n < 1$

and

277

(3.4) $$\sum_{n=1}^{\infty} n\, e_n \chi^n < \infty.$$

We also let h_n be the probability that only one fiber survives in a linear bundle of n fibers, and assume

(3.5) $$\sum_{n=1}^{\infty} h_n \chi^n < \infty.$$

(By the principle of monotone convergence these sums will indeed converge.) Lastly we assume t is such that χ is finite. This is guaranteed if $F(t;\ell) < 1$.

We now give several lemmas whose proofs appear in Kuo and Phoenix (1987).

Lemma 5. For linear bundles

(3.6) $$\sum_{n=1}^{\infty} n\, f_{n,1}\, \chi^n < \infty.$$

Lemma 6. For linear bundles

(3.7) $$\tilde{R}_1(\chi) < \infty.$$

Lemma 7. For linear bundles

(3.8) $$\tilde{R}(\chi) < \infty.$$

Lemma 8. For circular bundles

(3.9) $$\tilde{R}_1^{(0)}(\chi) < \infty.$$

Lemma 9. For both circular and linear bundles the sequence $\{Q_n \chi^n\}_{n=0}^{\infty}$ is bounded.

Lemma 10. For both linear and circular bundles

(3.10) $$\sum_{n=1}^{\infty} |\xi_n|\, \chi^n < \infty.$$

Thus we may state the key result.

<u>Theorem 2</u>. For both linear and circular bundles

(3.11) $\lim_{n \to \infty} Q_n \chi^n = \pi$

where

(3.12) $\pi = \begin{cases} \tilde{R}_1(\chi)^2 / \sum_{n=1}^{\infty} n f_{n,1} \chi^n & \text{for linear bundles ,} \\ 1 & \text{for circular bundles .} \end{cases}$

<u>Proof</u>: From Theorem 1 we may write the renewal equation

(3.13) $Q_n \chi^n = \sum_{j=1}^{n} Q_{n-j} \chi^{n-j} f_{j,1} \chi^j + \xi_n \chi^n$

where ξ_n is given respectively by (2.13) and (2.14) for linear and circular bundles. To (3.13) we may apply a key theorem in the theory of the renewal equation as given in Karlin and Taylor (1975). The key conditions for this theorem are (3.2), Lemma 5, Lemma 9, Lemma 10, and

(3.14) $\gcd\{n \mid f_{n,1} > 0\} = 1$,

which is obvious. The theorem yields (3.11) where

(3.15) $\pi = \sum_{n=0}^{\infty} \xi_n \chi^n / \sum_{n=1}^{\infty} n f_{n,1} \chi^n .$

In the linear case $\xi_0 = r_{0,1}$ and (2.13) yields

(3.16) $\sum_{n=0}^{\infty} \xi_n \chi^n = \tilde{R}_1(\chi)^2 + \sum_{j=2}^{\infty} \tilde{R}_j(\chi) - \tilde{F}_1(\chi) \sum_{j=2}^{\infty} \tilde{R}_j(\chi) = \tilde{R}_1(\chi)^2$

in view of (3.2). In the circular case $\xi_0 = r_{0,1}^{(0)} = 1$, and (2.14) similarly yields

(3.17) $\sum_{n=0}^{\infty} \xi_n \chi^n = \sum_{n=1}^{\infty} n f_{n,1} \chi^n .$

Thus (3.12) follows from (3.15) to (3.17), proving the theorem.

IV. BEHAVIOR OF $H_{m,n}(t)$ AS m AND n GROW LARGE. We now recast the results of Theorem 2 into a more useful form from the point of view of applications. Let

(4.1) $W(t) = 1 - 1/\chi(t)$, $t \geq 0$.

Then since $G_n(t) = 1 - Q_n(t)$ we may recast Theorem 2 as

(4.2) $G_n(t) = 1 - [1 - W(t)]^n [\pi(t) + o_n(t)]$, $t \geq 0$,

where $o_n(t) \to 0$ and $n \to \infty$ for each $t \geq 0$. From (1.2) the distribution function for the failure time of the composite is

(4.3) $H_{m,n}(t) = 1 - [1 - W(t)]^{mn} [\pi(t) + o_n(t)]^m$, $t \geq 0$.

Shortly we show that $W(t)$ is typically a proper distribution function in $t \geq 0$. Also $\pi(t)$, which is identically one for circular bundles (Theorem 2) is typically very close to one for linear bundles and usually $\pi(0) = 1$. It appears that $\pi(t)$ plays the role of a bundle edge term, and may be neglected for larger n. Thus when m and n are both large and of the same order, the resulting approximation is

(4.4) $H_{m,n}(t) \approx 1 - [1 - W(t)]^{mn}$, $t \geq 0$.

Of course, the accuracy of this approximation depends on the speed with which $o_n(t) \to 0$. Limited numerical studies show that once n reaches a moderate size, $o_n(t)$ decreases by orders of magnitude with each unit increase in n, so that the convergence is extremely fast.

Because of its importance we call $W(t)$, $t \geq 0$ the characteristic distribution function for failure. To see that it is indeed a distribution function we note that for circular bundles

$W(t) = 1 - [Q_n(t)]^{1/n} [1 + o_n(t)]^{-1/n}$.

Since $Q_n(t)$ is nondecreasing in t and $o_n(t) \to 0$ as $n \to \infty$ for each $t \geq 0$, it is easy to argue that $W(t)$ must be nondecreasing. We recall $\chi(t) \geq 1$ and from the definition (3.2) of $\chi(t)$ we have $f_{1,1}(t)\chi(t) \leq 1$. Since $f_{1,1}(t) = Pr\{0\} = 1 - F(t;\mathcal{L})$ we use (4.1) to obtain

(4.5) $0 \leq W(t) \leq F(t;\mathcal{L})$, $t \geq 0$.

Now as $t \to \infty$ we have $F(t;\ell) \to 1$ but it is more difficult to argue that $W(t) \to 1$ since this requires $\chi(t) \to \infty$ and this is not easily seen from the definition (3.2) of $\chi(t)$. However, a lower bound $W^*(t)$ on $W(t)$ can be obtained in many cases which satisfies $W^*(t) \to 1$ as $t \to \infty$. Multiplying (3.7) by χ^n and summing on n leads to

$$(4.6) \qquad \sum_{j=1} e_j \chi^j \geq 1 .$$

Unfortunately, simple expressions for $e_j(t)$ are not usually possible, but in applications one can usually show that

$$(4.7) \qquad e_j(t) \leq A \, B(t)^j , \quad t \geq 0 ,$$

where A is a positive constant and B(t) is some positive function satisfying $B(t) \to 0$ as $t \to \infty$. Then $\chi^*(t)$ which solves

$$(4.8) \qquad A \sum_{j=1} (\chi^*(t) \, B(t))^j = 1$$

will be a lower bound on $\chi(t)$. Since $\chi^*(t)B(t)$ must be a constant in (4.8) we will have $\chi^*(t) \to \infty$ and $W(t) \to 1$ as $t \to \infty$. Loosely speaking condition (4.8) will tend to be satisfied when the survival probability for a single fiber diminishes sharply to zero with increasing K_j. (Recall $e_j = Pr\{OXX...X\}$.) This will depend on both the upper tail behavior of $F(t;\lambda)$ and how fast K_r grows in r.

Numerical Calculation of W(t). The exact calculation of W requires the calculation of χ using (3.2). We let $\underline{B}_r$ be the r x r matrix

$$(4.9) \qquad \underline{B}_r = \begin{bmatrix} 0 & 1 & & & \\ 0 & 0 & 1 & & \\ & & ... & & \\ 0 & 0 & 0 & ... & 1 \\ f_{r,1} & f_{r-1,1} & ... & f_{2,1} & f_{1,1} \end{bmatrix}$$

where we recall

$$f_{1,1} = Pr\{O\}$$
$$f_{2,1} = 0$$
$$f_{3,1} = Pr\{OXO\}$$

281

(4.10) $f_{4,1} = \Pr\{OXXO\}$
  $f_{5,1} = \Pr\{OXOXO, OXXXO\}$
  $f_{6,1} = \Pr\{OXXXXO, OXXOXO, OXOXXO\}$
  •
  •
  • ,

these being dependent on t. Then $1 - W = 1/\chi$ is the spectral radius of the infinite matrix

(4.11) $\underline{B}_\infty = \lim_{r \to \infty} \underline{B}_r$.

To calculate $1/\chi$ numerically, first calculate in succession $1/\chi_1$, $1/\chi_2$, ... as the largest eigenvalues of the respective matrices $\underline{B}_1$, $\underline{B}_2$, ... This requires being able to calculate probabilities for the configurations in (4.10), and this is usually possible for configurations up to length 12 or so. Since $\chi_r \to \chi$ as $r \to \infty$ choose χ_r where r is large enough for the convergence to be essentially complete. In this regard note that $H_{m,n}(t) \approx mnW(t)$ according to (4.4), where mn is typically very large (say 10^9). Thus "essentially complete" means that changes in $mn(\chi_r-1)/\chi_r$ must be small compared to one. In any case $\chi_r \geq \chi$ so that $W_r \equiv 1 - 1/\chi_r$ will be an upper bound on W. In applications, r of the order of 10 often suffices.

Behavior of $\pi(t)$. As mentioned, $\pi(t)$ appears to play the role of a boundary or edge term, and is identically one for circular bundles. For linear bundles it may be shown that $\pi(0) = 1$ when $\mathcal{L}(0) = 0$ even when $F(0;\mathcal{L}) = \phi > 0$ as in the pure flaw model. This is shown in Kuo and Phoenix (1987).

V. APPLICATIONS AND VERIFICATION OF TECHNICAL ASSUMPTIONS.

To apply the previous results in specific cases, we must verify the key technical assumptions (3.3) to (3.5). Here we do this for the 'pure flaw' model for fibers to illustrate some useful procedures and difficulties.

'Pure flaw' model for fibers. We recall the simple model (1.8) where a fiber has unit strength with probability $1-\phi$ or has zero strength with probability ϕ. The composite loading we recall is $\mathcal{L}(t)$

= t, t ≥ 0. Before beginning, we mention that Harlow (1985) used a very different recursive approach to study the planar case of the model. He arrived at essentially the same structure (4.3) for $H_{m,n}(t)$ though it is difficult to demonstrate that all his quantities are equivalent to ours.

Considering t = 0 first, we are able to evaluate all the major quantities. First we show W(0) = 0 even though $F(0,\mathcal{L}) = \phi > 0$. When t = 0, it may be shown that (4.6) is

(5.1) $$\sum_{j=1}^{\infty} e_j \chi^j = 1, \quad (t = 0).$$

Since $e_j = Pr\{E_j\} = \phi^{j-1}(1-\phi)$ we may evaluate the sum in (5.1) to obtain

(5.2) $$(1-\phi)\phi\chi/(\phi(1-\phi\chi)) = 1.$$

This yields $\chi = 1$ so that W(0) = 0. At the end of Section 4 we pointed out that $\pi(0) = 1$ for both circular and linear bundles. Turning to the three conditions (3.3) to (3.5) we first note that $h_n = ne_n$ so that the third is equivalent to the second. Since $e_n = \phi^{n-1}(1-\phi)$ and $\chi = 1$ we have

$$\sum_{n=2}^{\infty} e_n \chi^n = \phi < 1 \quad \text{and} \quad \sum_{n=1}^{\infty} n e_n \chi^n = 1/(1-\phi) < \infty.$$

Finally, it is easy to see that $G_n(0) = \phi^n$ so from (4.2) the residue term is $o_n(0) = -\phi^n$.

Next we consider t such that 0 < t < 1, and we choose k such that $K_{k-1}t < 1 \le K_k t$. The interpretation of k is that under the composite load t an intact fiber will fail once it develops k failed neighbors (counting on both sides). To verify the three conditions (3.3) to (3.5) it is easiest to use a simple upper bound on $\chi(t)$, namely $1/(1-F(t;\mathcal{L})) = 1/(1-\phi)$. Also $e_n = \phi^{n-1}(1-\phi)$ for $1 \le n \le k$ and is zero otherwise. Thus for the first condition (3.3),

(5.3) $$\sum_{n=2}^{\infty} e_n \chi^n \le \sum_{n=1}^{k-1} \phi^n (1-\phi)(1-\phi)^{-(n+1)} \le \phi/(1-2\phi),$$

283

which is less than one provided $\phi < 1/3$. For the second and third conditions (again $h_n = ne_n$)

$$(5.4) \qquad \sum_{n=1}^{\infty} ne_n \chi^n \leq \frac{(1-\phi)}{\phi} \sum_{n=1}^{\infty} n[\phi/(1-\phi)]^n = (1-\phi)^2/(1-2\phi)^2 ,$$

which is finite for $\phi < 1/2$. Thus all conditions are met independently of t for $\phi < 1/3$. The case $t = 1$ is trivial since $G_n(1) = 1$.

Turning to the calculation of $W(t)$, the simplest situation is when $1/K_1 \leq t < 1$. Studying (4.10) we get $f_{1,1} = (1-\phi)$ and $f_{n,1} = 0$ for $n \geq 2$. Thus (3.2) yields $\chi = 1/(1-\phi)$ so $W(t) = \phi$. The next simplest situation is when $1/K_2 \leq t < 1/K_1$ so that (4.10) yields $f_{1,1} = (1-\phi)$, $f_{2,1} = 0$, $f_{3,1} = \phi(1-\phi)^2$ and $f_{n,1} = 0$ for $n \geq 4$. Thus (3.2) yields

$$(5.5) \qquad (1-\phi)\chi + \phi(1-\phi)^2 \chi^3 = 1 .$$

While we could solve for χ explicitly, we are usually interested in small values of ϕ in applications. We find $\chi = 1/(1-2\phi^2) + O(\phi^3)$ whence $W(t) = 2\phi^2 + O(\phi^3)$. The next easiest case is $1/K_3 \leq t < 1/K_2$. Studying (4.10) we find the new $f_{n,1}$'s are $f_{4,1} = \phi^2(1-\phi)^2$ but otherwise $f_{n,1} = 0$ for n even and $f_{n,1} = \phi^{(n-1)/2}(1-\phi)^{(n+1)/2}$ for n odd. The series (3.2) may be evaluated to yield

$$(5.6) \qquad (1-\phi)\chi + \phi(1-\phi)\chi^2 + \phi^2(1-\phi)^2\chi^4 - \phi^3(1-\phi)^3 \chi^6 = 1 ,$$

where in the process we find $\chi < [\phi(1-\phi)]^{-1/2}$. Then χ is the real solution to (5.6), and must be determined numerically. For small ϕ we find $\chi = 1/(1 - 3\phi^3) + O(\phi^4)$ whence $W(t) = 3\phi^3 + O(\phi^4)$.

For smaller t and k the $f_{n,1}$ in (4.10) become more complicated. However, it appears to be generally true that

$$(5.7) \qquad W(t) = k\phi^k + O(\phi^{k+1}) , \qquad 1/K_k \leq t < 1/K_{k-1}$$

for $k = 1,2,\ldots$. Thus we see that $W(0) = 0$ and $W(t)$ increases in steps at the time points $t_k = 1/K_k$, $k = 1,2,\ldots$ where the number of steps becomes infinite as $t \downarrow 0$. The above results agree with those of Harlow (1985), who points out that $W(t) \leq k\phi^k$ at least for $1 \leq k \leq 5$.

Turning to $\pi(t)$ for the case of linear bundles, we find from (3.12) that $\pi(t) = 1$ for $1/K_1 \leq t < 1$ and $\pi(t) = 1 + 3\phi^2 + O(\phi^3)$ for $1/K_2 \leq t < 1/K_1$. In general it appears that

$$(5.8) \quad \pi(t) = 1 + (k-1)(k+1)\phi^k + O(\phi^{k+1}) , \quad 1/K_k \leq t < 1/K_{k-1} .$$

Lastly we note that Harlow (1985) numerically calculated the maximum deviation

$$(5.9) \quad \varepsilon_{m,n} = \sup_{0 \leq t \leq 1} | H_{m,n}(t) - \{1 - [1 - W(t)]^{mn}\} |$$

for linear bundles and various combinations of m, n and ϕ in order to study the error in the approximation (4.4). First his results suggest that (5.7) is an extremely accurate approximation for W(t) for $\phi \leq 0.1$ and k up to 12, which is as far as his results go. Second, almost all the deviation he observed in (5.9) can to be accounted for by using the approximation (5.8) for $\pi(t)$ instead of putting $\pi(t) = 1$. In other words, the boundary effects in the planar composites, though small, seem to dominate the residue $o_n(t)$.

For fibers with Weibull strength (see (1.9)), the calculation of W(t) and $\pi(t)$ must be done numerically and will not be considered here. Insight into their behavior can be obtained from Harlow and Phoenix (1981, 1982) where a different recursive approach was used to study the first occurrence of k adjacent breaks. In fact, the results here essentially verify a conjecture which arose there.

For the time-dependent fatigue model (1.10),verification of conditions (3.3) to (3.5) has proven to be elusive except for $\rho = \beta = 1$.

A practical solution is to restrict the load on a fiber to ℓ_{max}, that is, to take $F(t;\ell) = 1$ as soon as $\ell(t)$ on a fiber exceeds ℓ_{max}. In practice ℓ_{max} would be the theoretical atomic bond strength for the material. With this limitation, if k is chosen such that $K_{k-1}L < \ell_{max} \leq K_k L$ then the sums in conditions (3.3) to (3.5) need only be considered for n up to k. Numerical calculations can be carried out for k up to about 10, and for $K_k = 1 + k/2$ this means for $L > \ell_{max}/6$. This happens to be sufficient for many applications. Numerical results suggest that the conditions hold for $\beta\rho > 3$, and in fact, ℓ_{max}, if sufficiently large, seems to have little to do with the convergence.

References:

Coleman, B. D. (1958) Statistics and time dependence of mechanical breakdown in fibers. J. Appl. Phys. 29, 968-983.

Gotlib, Yu Ya, El'yashevich, A. M. and Svetlov, Yu E. (1973) Effect of microcracks on the local stress distribution in polymers and their deformation properties. Network model. Soviet Physics - Solid State 14, 2672-2677.

Harlow, D. G. and Phoenix, S. L. (1978) The chain of bundles probability model for the strength of fibrous materials II: A numerical study of convergence. J. Composite Materials 12, 314-334.

Harlow, D. G. and Phoenix, S. L. (1981) Probability distribution for the strength of composite materials II: a convergent sequence of tight bounds. Internat. J. Fracture 17, 601-630.

Harlow, D. G. and Phoenix, S. L. (1982) Probability distributions for the strength of fibrous materials I: Two-level failure and edge effects. Adv. Appl. Prob. 14, 68-94.

Harlow, D. G. (1985) The pure flaw model for chopped fibre composites. Proc. R. Soc. London A 397, 211-232.

Hedgepeth, J. M. (1961) Stress Concentrations in Filamentary Structures, NASA Technical Note D-882.

Karlin, S. and Taylor, H. M. (1984) An Introduction to Stochastic Modeling. Academic Press, New York.

Kuo, C. C. (1983) Recursion Formulas and Limit Theorems for the Lifetime Distribution of a Model Fibrous Composite, Ph.D. Thesis, Cornell University.

Kuo, C. C. and Phoenix, S. L. (1987) Recursions and limit theorems for the strength and lifetime distributions of a fibrous composite. J. Applied Probability (to appear).

Phoenix, S. L. and Tierney, L-J. (1983) 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.

Pitt, R. E. and Phoenix, S. L. (1983) Probability distributions for the strength of composite materials IV: localized load-sharing with tapering. Internat. J. Fracture 22, 243-276.

Smith, R. L. (1980) A probability model for fibrous materials with local load sharing. Proc. R. Soc. London A 372, 539-553.

Smith, R. L. (1982) A note on a probability model for fibrous composites. Proc. R. Soc. London A 382, 179-182.

Smith, R. L. (1983) Limit theorems and approximations for the reliability of load-sharing systems. Adv. Appl. Prob 15, 304-330.

Taylor, H. M. and Karlin, S. (1975) A First Course in Stochastic Processes. Academic Press (2nd Ed.), New York.

Tierney, L. (1982) Asymptotic bounds on the time to fatigue failure of bundles of fibers under local load sharing, Adv. Appl. Prob. 14, 95-121.

PHASE SPACE METHODS AND PATH INTEGRATION:

A MICROSCOPIC APPROACH TO DIRECT AND INVERSE WAVE PROPAGATION

Louis Fishman
Department of Civil Engineering
The Catholic University of America
Washington, D.C. 20064 USA

ABSTRACT. This project focuses on the development of new, multidimensional algorithms for direct acoustic propagation and generalized acoustic tomography at the level of the scalar Helmholtz equation. The general aim is the continued detailed development of the ideas originally outlined several years ago. Phase space, or "microscopic," methods and path (functional) integral representations provide the appropriate framework to extend homogeneous Fourier methods to inhomogeneous environments. The path integrals furnish the principal representation of the Helmholtz propagator and, subsequently, through direct computation, the basis for the direct numerical algorithms. There are two complementary approaches to the analysis and computation of the n-dimensional Helmholtz propagator. The first is essentially a factorization/parabolic-based (one-way) phase space path integration/invariant imbedding approach. This results in a marching algorithm which generalizes the Tappert/Hardin split-step FFT algorithm for one-way wave propagation, a nonperturbative incorporation of backscatter effects which generalizes Kennett's algorithm in reflection seismology for two-way wave propagation, and the basis for the formulation and solution of corresponding arbitrary-dimensional nonlinear inverse problems. The numerical algorithms based on these modern, "microscopic" methods directly compute pseudo-differential and Fourier integral operators, incorporate phase space filtering, and are ideally suited for computers which provide either a vector or a parallel pipe type of operation. Extensive testing has, so far, been very promising. While the first approach starts from a transversely inhomogeneous formulation and, subsequently, builds in backscatter effects, the second approach constructs elliptic-based (two-way) path integral representations of the propagator for general range-dependent environments from the outset. A particular approximate path integral construction (Feynman/Garrod) results in a true path functional, suggesting the underlying stochastic foundations of the Helmholtz equation. It appears to be a viable computational approximation for a useful range of propagation experiments and can be numerically evaluated by standard Monte Carlo (statistical) methods. A more detailed examination and approximate construction of the underlying stochastic process would provide for both more accurate and widely applicable path integral representations and direct numerical simulation techniques.

I. INTRODUCTION. Direct wave propagation modeling plays a significant role in such fields as underwater communication, radio transmission through the atmosphere, laser propagation, and earthquake prediction. Likewise, the corresponding inverse problems are at the heart of such areas as submarine detection, CAT scan technology, soft-tissue diffraction tomography, the mapping of the interior earth, and oil

exploration. In all of these and many other examples, relatively fast and accurate numerical algorithms are necessary.

The analysis and fast, accurate numerical computation of the wave equations of classical physics are often quite difficult for rapidly changing, multidimensional environments extending over many wavelengths. For the most part, classical, "macroscopic" methods have resulted in direct wave field approximations (perturbation theory, ray-theory asymptotics, modal analysis, hybrid ray-mode methods), derivations of approximate wave equations (scaling analysis, field splitting techniques, formal operator expansions), and discrete numerical approximations (finite differences, finite elements, spectral methods). In the last several decades, however, mathematicians studying linear partial differential equations have developed, in the language of physicists, a sophisticated, "microscopic" phase space analysis. In conjunction with the global functional integral techniques pioneered by Wiener (Brownian motion) and Feynman (quantum mechanics), and so successfully applied today in quantum field theory and statistical physics, the n-dimensional classical physics propagators can be both represented explicitly and computed directly. The phase space, or "microscopic," methods and path (functional) integral representations provide the appropriate framework to extend homogeneous Fourier methods to inhomogeneous environments, in addition to suggesting the basis for the formulation and solution of corresponding arbitrary-dimensional nonlinear inverse problems. Moreover, it is in phase space, rather than in configuration space, that, from a mathematical perspective, the interesting geometry takes place.

II. PHASE SPACE AND PATH INTEGRAL CONSTRUCTIONS. For the n-dimensional scalar Helmholtz equation, there are two complementary approaches to this analysis and computation, as illustrated in Figure 1. The first is essentially a factorization/path integration/invariant imbedding approach. For transversely inhomogeneous environments, implying medium homogeneity with respect to a single distinguished direction, the n-dimensional Helmholtz equation can be exactly factored into separate, physical forward and backward, one-way wave equations, following from spectral analysis [1-5]. The forward evolution (one-way) equation

$$(i/\bar{k})\partial_x \phi^+(x,\underline{x}_t) + (K^2(\underline{x}_t) + (1/\bar{k}^2 \nabla_t^2)^{1/2} \phi^+(x,\underline{x}_t) = 0 , \tag{1}$$

where $K(x)$ is the refractive index field and $\bar{k}$ is a reference wave number, is the formally exact wave equation for propagation in a transversely inhomogeneous half-space supplemented with appropriate outgoing wave radiation and initial-value conditions. While functions of a finite set of commuting self-adjoint operators can be defined through spectral theory, functions of noncommuting operators are represented by pseudo-differential operators [2,5]. The formal wave equation (1) is now written explicitly as a Weyl pseudo-differential equation in the form

$$(i/\bar{k})\partial_x \phi^+(x,\underline{x}_t) + (\bar{k}/2\pi)^{n-1} \int_{R^{2n-2}} \frac{d\underline{x}_t' d\underline{p}_t}{}$$

$$\cdot \Omega_B(\underline{p}_t,(\underline{x}_t + \underline{x}_t')/2) \exp(i\bar{k}\underline{p}_t\cdot(\underline{x}_t - \underline{x}_t'))\phi^+(x,\underline{x}_t') = 0. \tag{2}$$
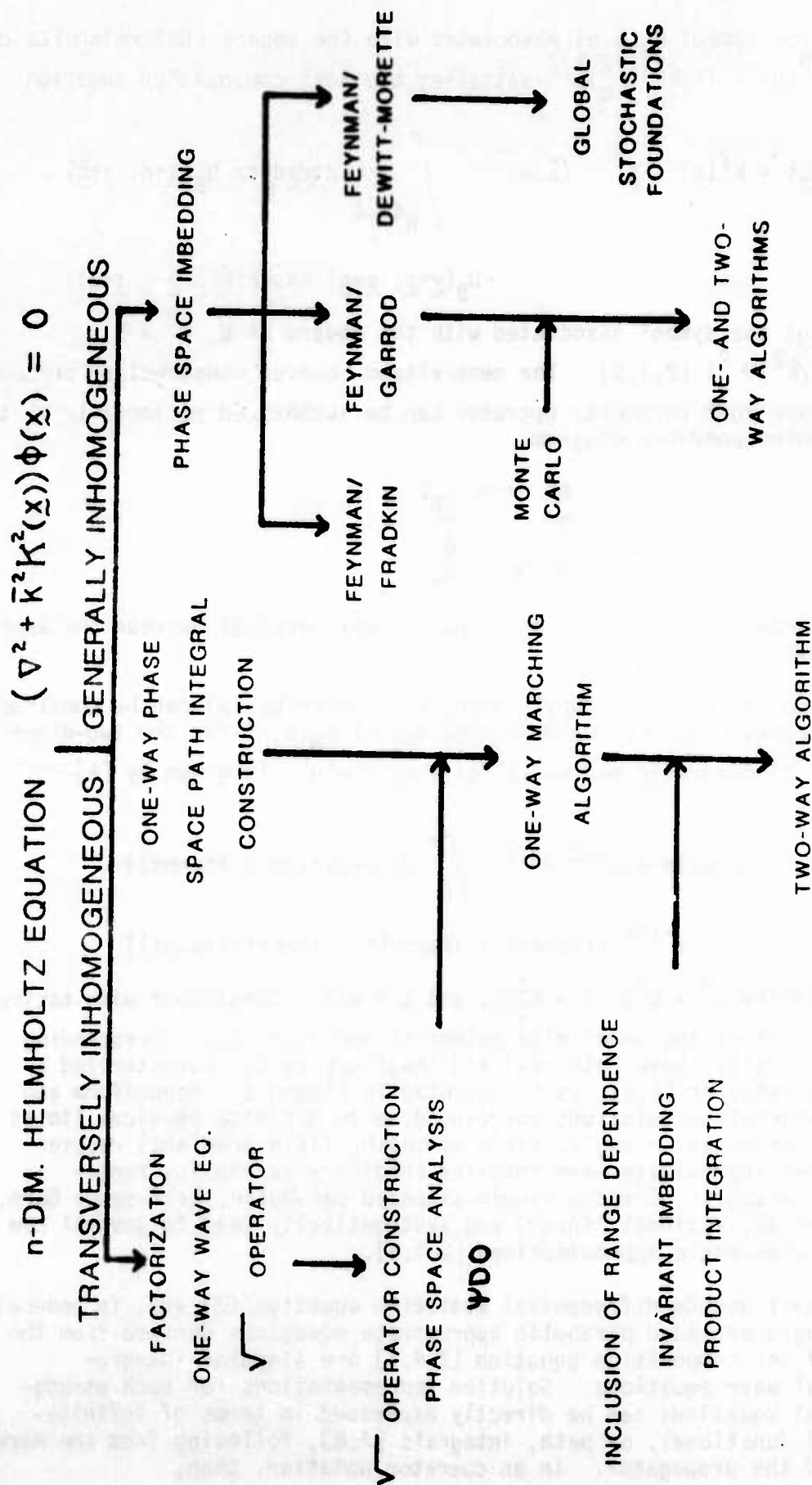
Fig. 1. Two complementary approaches to the analysis and computation of the n-dimensional scalar Helmholtz equation.

In Eq.(2), the symbol $\Omega_B(\underline{p},\underline{q})$ associated with the square root Helmholtz operator $B = (K^2(\underline{q}) + (1/\bar{k}^2)\nabla_{\underline{q}}^2)^{1/2}$ satisfies the Weyl composition equation

$$\Omega_B^2(\underline{p},\underline{q}) = K^2(\underline{q}) - \underline{p}^2 = (\bar{k}/\pi)^{2n-2} \int_{R^{4n-4}} d\underline{t}\,d\underline{x}\,d\underline{y}\,d\underline{z}\; \Omega_B(\underline{t}+\underline{p},\; \underline{x}+\underline{q})$$

$$\cdot \Omega_B(\underline{y}+\underline{p},\; \underline{z}+\underline{q})\, \exp(2i\bar{k}(\underline{x}\cdot\underline{y} - \underline{t}\cdot\underline{z})) \qquad (3)$$

with $\Omega_B^2(\underline{p},\underline{q})$ the symbol associated with the square of B, $B^2 = (K^2(\underline{q}) + (1/\bar{k}^2)\nabla_{\underline{q}}^2)$ [2,3,5]. The generalized Fourier construction procedure for the square root Helmholtz operator can be summarized pictorially by the following correspondence diagram

$$\begin{array}{ccc} B^2 & \Longleftrightarrow & \Omega_B^2 \\ \uparrow & & \Updownarrow \\ B & \Longleftrightarrow & \Omega_B \end{array}$$

where the arrows symbolize the one- and two-way mappings between the appropriate quantities.

Exact solutions of the Weyl composition equation (3) can be constructed in several cases [6]. For example, the symbol $\Omega_B(p,q)$ for the two-dimensional (n = 2) quadratic medium, $K^2(q) = K_0^2 + w^2q^2$, is given by [6]

$$\Omega_B(p,q) = -(\exp(i\pi/4)\varepsilon^{1/2}/\pi^{1/2}) \int_0^\infty dt\, \exp(i(Yt + X\tanh t))$$

$$\cdot t^{-1/2}\, (iY\,\mathrm{sech}\,t + iX\,\mathrm{sech}^3 t - (\mathrm{sech}\,t)(\tanh t)) \qquad (4)$$

with $X = (1/\varepsilon)(w^2q^2 - p^2)$, $Y = K_0^2/\varepsilon$, and $\varepsilon = w/\bar{k}$. Consistent with taking the square root of the indefinite Helmholtz operator, the corresponding symbols, generally, have both real and imaginary parts characterized by oscillatory behavior [4,6], as illustrated in Figure 2. Nonuniform and uniform perturbation solutions corresponding to definite physical limits (frequency, propagation angle, field strength, field gradient) recover several known approximate wave theories (ordinary parabolic, range-refraction parabolic, Grandvuillemin-extended parabolic, half-space Born, Thomson-Chapman, rational linear) and systematically lead to several new full-wave, wide-angle approximations [2-4,6].

The exact pseudo-differential evolution equation (2) and, in general, the wide-angle extended parabolic approximate equations derived from the analysis of the composition equation [2-4,6] are singular integro-differential wave equations. Solution representations for such pseudo-differential equations can be directly expressed in terms of infinite-dimensional functional, or path, integrals [7,8], following from the Markov property of the propagator. In an operator notation, then,

$$\exp(i\bar{k}Bx) = \lim_{N \longrightarrow \infty} \prod_{j=1}^{N} \exp(i\bar{k}B\Delta x_j) \tag{5}$$

where $\Delta x_j = x/N$, symbolically representing the propagator in terms of the infinitesimal propagator. As the operator symbol is not simply quadratic in $\underline{p}$, the configuration space Feynman path integral formulation is not appropriate, necessitating the more general phase space construction [4,7]. This results in a parabolic-based (one-way) Hamiltonian phase space path integral representation of the propagator in the form [3,7]

$$G^{+}(x,\underline{x}_t|0,\underline{x}_t') = \lim_{N \longrightarrow \infty} \int_{R^{(n-1)(2N-1)}} \prod_{j=1}^{N-1} d\underline{x}_{jt} \prod_{j=1}^{N} (\bar{k}/2\pi)^{n-1} d\underline{p}_{jt}$$

$$\cdot \exp(i\bar{k} \sum_{j=1}^{N} (\underline{p}_{jt} \cdot (\underline{x}_{jt} - \underline{x}_{j-1t}) + (x/N) H(\underline{p}_{jt},\underline{x}_{jt},\underline{x}_{j-1t}))) \tag{6}$$

where
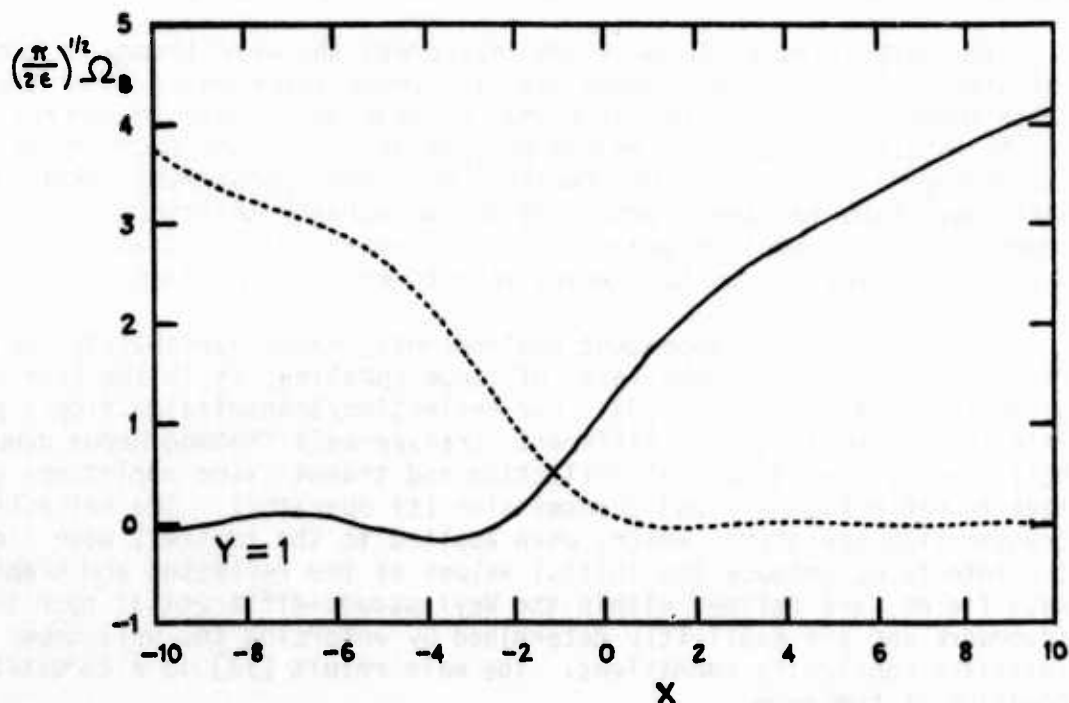
Fig. 2.   The real (———) and imaginary (----) parts of the $n = 2$ quadratic medium symbol as a function of X for Y = 1.

293

$$H(\underline{p}, \underline{q}'', \underline{q}') = (\bar{k}/2\pi)^{n-1} \int_{R^{2n-2}} d\underline{s} d\underline{t} \, F(\underline{q}' - \underline{q}'', \underline{s})$$

$$\cdot h_B(\underline{p}, ((\underline{q}'' + \underline{q}')/2) - \underline{t}) \exp(i\bar{k}\underline{s} \cdot \underline{t}). \tag{7}$$

In Eq.(7), $F(\underline{u}, \underline{v})$ and $h_B(\underline{p}, \underline{q})$ are related to the operator symbol $\Omega_B(\underline{p}, \underline{q})$ by

$$\hat{\Omega}_B(\underline{u}, \underline{v}) = F(\underline{u}, \underline{v}) \, \hat{h}_B(\underline{u}, \underline{v}) \tag{8}$$

where $\hat{\Omega}_B$ and $\hat{h}_B$ are the corresponding Fourier transforms [2,3,7].

The nonuniqueness of the lattice-approximation path integral representation is readily understood in terms of different discretizations, or quadratures, of the symbolic functional integral and corresponds to the representation of a given (fixed) operator by different operator-ordering, or pseudo-differential operator, schemes [2,3,7,8]. More fundamentally, in analogy with the Schrödinger equation for particle motion on a Riemannian space and the thermodynamic (Fokker-Planck) equation for particle diffusion, the algorithmic Helmholtz path integral construction reflects the stochastic ..ature of the integration [4,9]. Further, both the macroscopic and microscopic (infinitesimal) half-space propagators can be·formally expressed as Fourier integral operators with complex phase [4]. The phase space path integral, thus, represents the macroscopic Fourier integral operator in terms of the N-fold application of the microscopic, or infinitesimal, Fourier integral operator in a manner which can be related to the global geometrical-optics construction of the macroscopic operator [4,5].

The path integral formulation interprets the wave theory in terms of an infinitesimal propagator summed over all phase space paths. For the Helmholtz theory, the exact infinitesimal propagator is not, in general, given by the locally homogeneous medium propagator, as in the ordinary parabolic (Schrödinger) propagator construction [8]. The approximate extended parabolic wave theories then correspond to approximate infinitesimal propagators summed over the complete phase space. In retaining the "sum over all paths," diffraction, or full-wave, effects are incorporated.

For weakly range-dependent environments, range variability can be, at first, accommodated at the level of range updating, as in the case of the parabolic path integral [1,8]. For reflection/transmission from a planar interface separating two (different) transversely inhomogeneous acoustic half-spaces, the concept of reflection and transmission amplitudes generalizes to reflection (r) and transmission (t) operators. The reflection and transmission operators, which, when applied to the incident wave field at the interface, produce the initial values of the reflected and transmitted wave fields, are defined within the Weyl pseudo-differential operator framework and are explicitly determined by enforcing the well-known interface continuity conditions. The main result [10] is a composition equation of the form

$$\Omega_{BL}(\underline{p}, \underline{q}) - \Omega_{BR}(\underline{p}, \underline{q}) = (\bar{k}/\pi)^{2n-2} \int_{R^{4n-4}} d\underline{t} d\underline{x} d\underline{y} d\underline{z} \, (\Omega_{BL}(\underline{t} + \underline{p}, \underline{x} + \underline{q}) +$$

294

$$\Omega_{BR}(\underline{t}+\underline{p}, \underline{x}+\underline{q})) \, \Omega_r(\underline{y}+\underline{p}, \underline{z}+\underline{q}) \, \exp(2i\bar{k}(\underline{x}\cdot\underline{y} - \underline{t}\cdot\underline{z})) \qquad (9)$$

for the reflection operator symbol $\Omega_r(\underline{p},\underline{q})$ and an analogous equation for the transmission operator symbol $\Omega_t(\underline{p},\underline{q})$. The inclusion of a planar transition region of arbitrary length and inhomogeneity can be accomplished by factorization methods in conjunction with invariant imbedding [4,11]. Invariant imbedding constructs the initial-value system for the reflection and transmission operators associated with the transition region, transforming the Helmholtz boundary-value problem into an initial-value problem. A discretized formulation [11] provides the extension of Kennett's method [4,11] in reflection seismology. The resultant forward and backward wave fields propagating in the transversely inhomogeneous half-spaces are represented by the one-way path integrals, while, within the transition region, a formal path integral representation of the propagator can be expressed as a product integral [8]. This takes the form [4]

$$G = \int_a^x \exp(i\bar{k}\underline{\underline{H}}(s)ds) = \lim_{N\longrightarrow\infty} \prod_{j=1}^N \exp(i\bar{k}\underline{\underline{H}}(s_j)\Delta s_j) \qquad (10)$$

where $s_j = a + (j-1/2)\Delta s_j$, $\Delta s_j = (x-a)/N$, a denotes the transition region boundary, $\underline{\underline{H}}$ is the appropriate first-order Helmholtz equation matrix operator [2,4], and with the product of exponential factors ordered from right (lower j) to left (higher j) reflecting the noncommutativity of the matrix operator $\underline{\underline{H}}$ at different x. While product integration-based path integral constructions have been applied to the problems of nonrelativistic electron spin and the Dirac equation, such infinite products of matrices are, generally, only tractable in simple limiting cases [4,8].

Rather than starting from a transversely inhomogeneous formulation and, subsequently, building in backscatter effects, the generalization of Fourier methods to arbitrary inhomogeneous environments and the construction of a dynamical basis for the Helmholtz equation can proceed, in the second approach, from the construction of truly global configuration space path integrals, which attempt to generalize, for example, the homogeneous half-space result [3,7]

$$G^+(x,\underline{x}_t|0,\underline{x}_t') = \lim_{N\longrightarrow\infty} \int_{R^{(n-1)(N-1)}} \prod_{j=1}^{N-1} dx_{jt} \, (i\pi xN^{(n-1)N/2})$$

$$\cdot (\bar{k}K_0/2\pi\delta_{(n-1)N+1})^{((n-1)N+1)/2} H_{((n-1)N+1)/2}^{(1)}(\bar{k}K_0\delta_{(n-1)N+1})) \qquad (11)$$

where

295

$$\delta_{(n-1)N+1} = (N \sum_{j=1}^{N} (\underline{x}_{jt} - \underline{x}_{j-1t})^2 + x^2)^{1/2} \qquad (12)$$

and $H_\nu^{(1)}(\xi)$ is the Hankel function. These elliptic-based (two-way) constructions, originating from the Fourier transform relationship between the Helmholtz and Schrödinger (parabolic) propagators, result in the approximate Feynman/Garrod path integral [3,7]

$$G(\underline{x}|\underline{x}') \simeq (-1/2\bar{k}^2) \lim_{N \longrightarrow \infty} \int_{R^{n(2N-1)}} \prod_{j=1}^{N-1} d\underline{x}_j \prod_{j=1}^{N} (\bar{k}/2\pi)^n d\underline{p}_j \frac{\exp(i\bar{k}S_N)}{(1/2 - \Sigma)} \qquad (13)$$

where

$$S_N = \sum_{j=1}^{N} \underline{p}_j \cdot (\underline{x}_j - \underline{x}_{j-1}) \qquad (14)$$

corresponds to an appropriate discretized action and

$$\Sigma = (1/N) \sum_{j=1}^{N} (\underline{p}_j^2/2 + V(\underline{x}_j)) \qquad (15)$$

plays a role analogous to an average energy with the identification $V(\underline{x}) = (-1/2)(K^2(x) - 1)$. For a transversely inhomogeneous half-space, partial integration of Eq.(13) in conjunction with the reflection principle (or method of images) results in [3,7]

$$G^+(x,\underline{x}_t|0,\underline{x}_t') \simeq \lim_{N \longrightarrow \infty} \int_{R^{(n-1)(2N-1)}} \prod_{j=1}^{N-1} d\underline{x}_{jt} \prod_{j=1}^{N} (\bar{k}/2\pi)^{n-1} d\underline{p}_{jt}$$

$$\cdot \exp(i\bar{k}(S_N + 2^{1/2}x(1/2 - \Sigma)^{1/2})) \qquad (16)$$

with S_N and Σ taking on their appropriate forms in one-lower dimension.

Formally reducing both the full- and transversely inhomogeneous half-space phase space Feynman/Garrod path integrals to configuration space path integrals [7] establishes the path functional character of the representation. Moreover, the approximate Feynman/Garrod path integral is exact in the homogeneous medium limit, incorporates significant backscatter information, and contains both the geometrical (ray) acoustic and ordinary parabolic approximations. This configuration space formulation for the two-way problem, initially based on a variational principle and phase space constructions, seeks to express the propagator in terms of a phase

296

functional evaluated over an appropriate path space, as symbolically expressed in the Feynman/DeWitt-Morette representation [3,7,9]. This takes the form

$$G(\underline{x}|\underline{x}') = (-1/2\bar{k}^2) \int_E D(\underline{z}) \exp(i\bar{k}\bar{w}(\underline{z})) \tag{17}$$

where

$$\bar{w} = \int_{\underline{x}'}^{\underline{x}} \|d\underline{z}\| (1 - 2V(\underline{z}))^{1/2} \tag{18}$$

is the analog of the action associated with a "free particle" on a space with the metric

$$d\underline{l}^2 = (1 - 2V(\underline{z})) \| d\underline{z} \|^2 \tag{19}$$

and where E represents the space of paths from $\underline{x}'$ to $\underline{x}$ such that

$$1/2 = (1/\tau) \int_0^{\tau} dt ((1/2)\| d\underline{z}(t)/dt \|^2 + V(\underline{z}(t))) \tag{20}$$

with the constraints

$$\underline{z}(0) = \underline{x}' ,$$
$$\underline{z}(\tau) = \underline{x} . \tag{21}$$

The dynamical basis of the Helmholtz equation can, thus, be viewed in terms of a stochastic process embodying fixed "average energy" paths, or, alternatively, in terms of "free particle" motion [3,7,9].

   III. COMPUTATIONAL ALGORITHMS. Direct integration of the one-way phase space path integral provides the computational basis for the pseudo-differential wave equation (2). Choosing the standard ordering, $F(\underline{u},v) = \exp(-i\bar{k}\underline{u}\cdot v/2)$, in Eqs. (6), (7), and (8) results in a numerically more efficient post-point marching algorithm in the form

$$\phi^+(x+\Delta x, \underline{x}_t) \simeq \int_{R^{n-1}} d\underline{p}_t \exp(i\bar{k}\underline{p}_t\cdot\underline{x}_t) (\exp(i\bar{k}\Delta x h_B(\underline{p}_t,\underline{x}_t)) \hat{\phi}^+(x,\underline{p}_t)) \tag{22}$$

where $\hat{\phi}^+$ is the Fourier-transformed wave field and

$$h_B(\underline{p}_t,\underline{x}_t) = (\bar{k}/\pi)^{n-1} \int_{R^{2n-2}} d\underline{s}d\underline{t} \; \Omega_B(\underline{s},\underline{t}) \exp(-2i\bar{k}(\underline{x}_t - \underline{t})\cdot(\underline{p}_t - \underline{s})). \tag{23}$$

297

This marching algorithm provides the generalization of the Tappert/Hardin split-step FFT algorithm [1] to the full one-way (factored Helmholtz) wave equation. For a two-dimensional model ocean/bottom propagation environment with a perfectly reflecting ocean surface, the Fourier transform of the wave field in Eq.(22) is replaced by a discrete fast sine transform and the inverse transform is evaluated by a rectangular rule integration, enabling the propagated wave field to be expressed in the matrix form

$$\phi^+(x+\Delta x, z_n) = \sum_m A_{nm} \hat{\phi}^+(x, p_m) \tag{24}$$

for each depth point z_n. In Eq.(24), ϕ^+ and $\hat{\phi}^+$ are column vectors and the matrix $\underline{\underline{A}}$ is defined by its matrix elements

$$A_{nm} = \eta \sin(\bar{k}p_m z_n + \bar{k}\Delta x h_B^0(p_m, z_n)) \exp(i\bar{k}\Delta x h_B^e(p_m, z_n)) \tag{25}$$

where h_B^e and h_B^0 are the even and odd parts with respect to p of $h_B(p,z)$ in Eq.(23) and η is an appropriate transform normalization constant [1,4,12].

The principal idea underlying the practical implementation of the phase space marching algorithm is the construction of a small number of approximate operator symbols, which, when taken together, allow for wave field computations over a very wide range of model environments and propagation parameters. In conjunction with a study of exactly soluble cases of the Weyl composition equation [6], high-frequency, real Weyl high-frequency, uniform high-frequency, and low-frequency approximate symbols have been constructed [2-4,6]. Of particular significance is the fact that the manner of marching the radiation field is independent of the medium and any approximation to the square root Helmholtz operator, resulting in a modular code architecture and highly versatile propagation program. Moreover, the propagation models constructed and computed through the code correspond to singular integro-differential equation as well as partial differential equation approximations to the one-way wave equation. Indeed, this numerical algorithm represents one of the very few attempts to compute directly with pseudo-differential and Fourier integral operators. For the two-dimensional case, the range-incrementing procedure is just a sequence of matrix multiplications, and, thus, ideally suited for computers which provide either a vector or a parallel pipe type of operation. Phase space filtering reduces both the size of the matrix multiplication and the number of matrix elements initially computed, in particular, reducing the total range-incrementing computational time by almost an order of magnitude for typical model calculations [4].

Numerical results of transmission loss (dB re 1 m) as a function of range (km) for a number of model ocean/bottom propagation experiments demonstrate the computational viability of the factorization-/path integration-based phase space marching algorithm [4,12]. Several propagation experiments are summarized in Figures 3, 5, and 7, with the corresponding transmission loss curves compared with a reference Fast Field Program (FFP) algorithm [4,12] in Figures 4, 6, and 8.
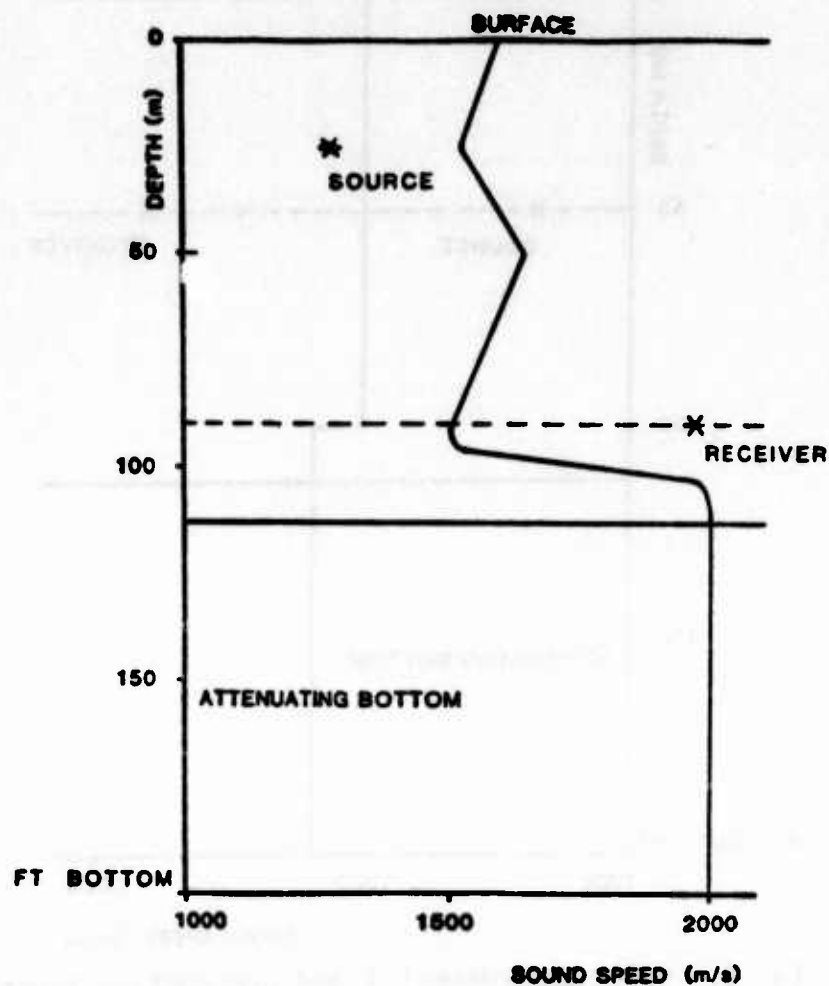
Fig. 3. Model environment 1 and propagation experiment.
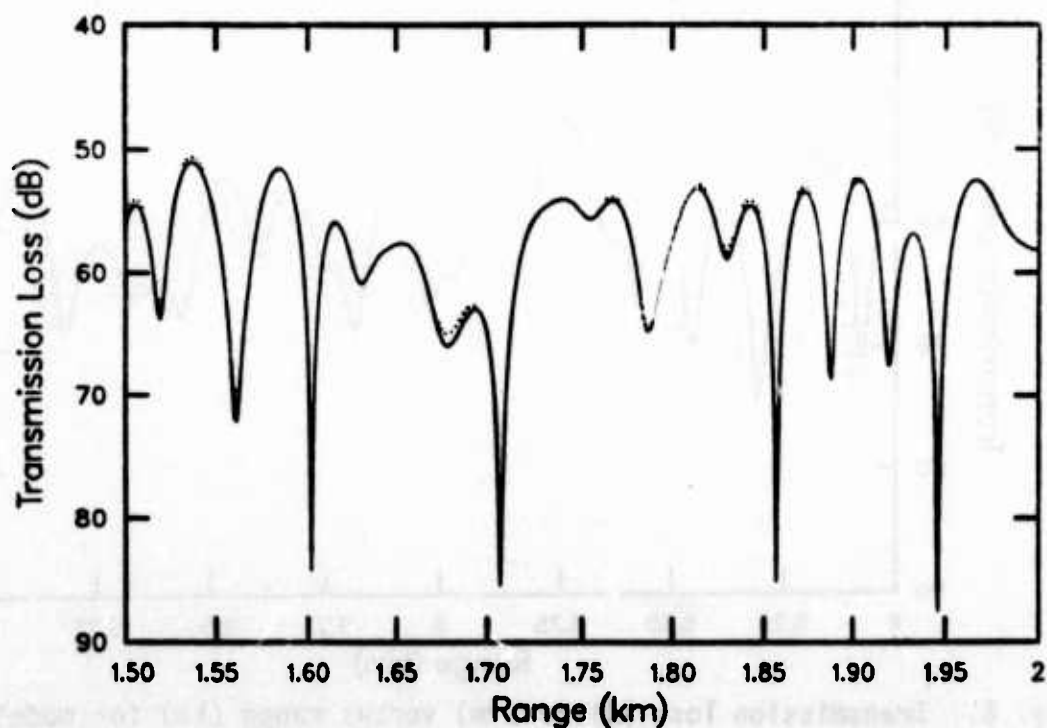
Fig. 4. Transmission loss (dB re 1 m) versus range (km) for model environment 1 at 400 Hz. (———) High Frequency (80 degree filter) (....) FFP

299
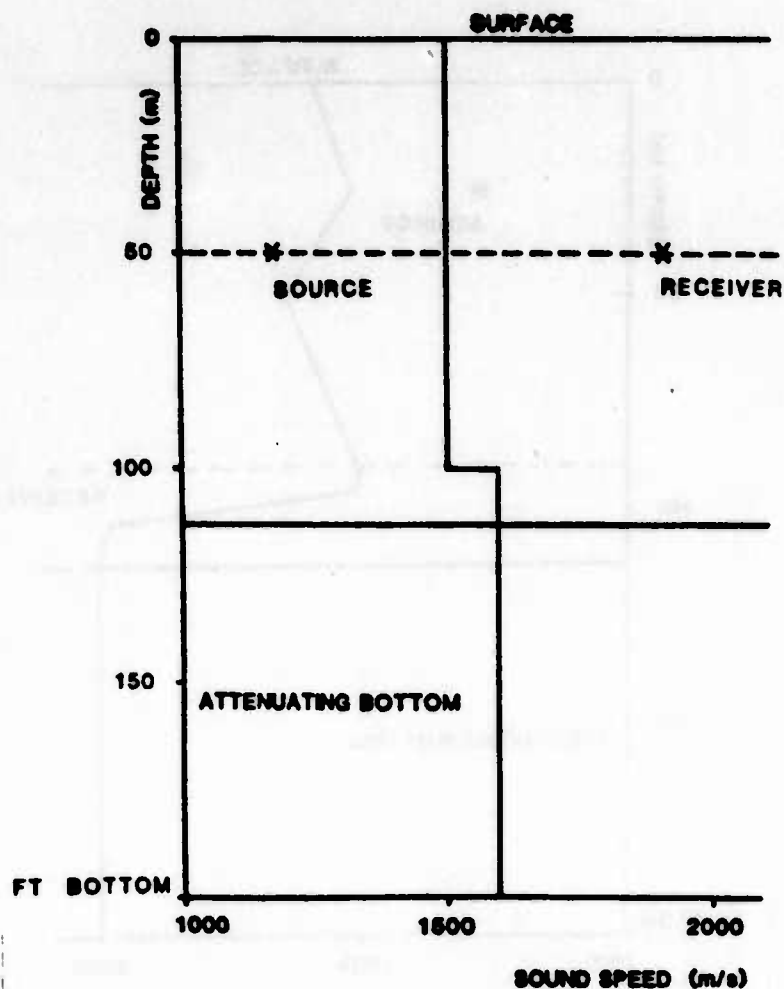
Fig. 5. Model environment 2 and propagation experiment.
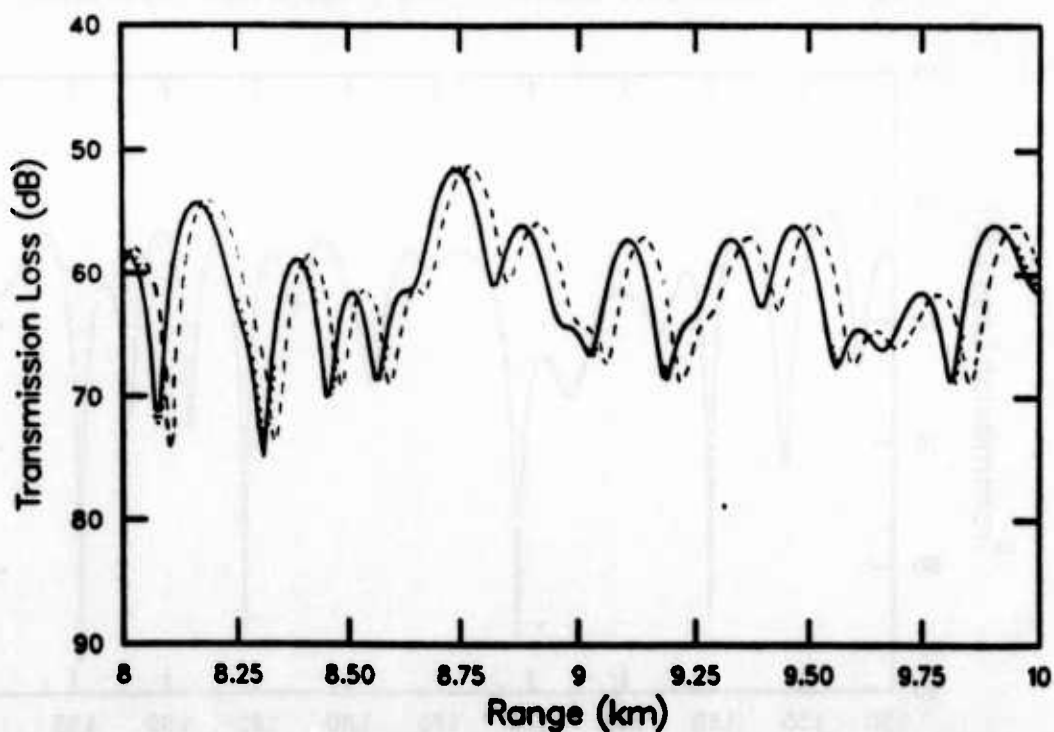
Fig. 6. Transmission loss (dB re 1 m) versus range (km) for model
environment 2 at 250 Hz. (———) High Frequency (60 degree filter)
(----) High Frequency (....) FFP
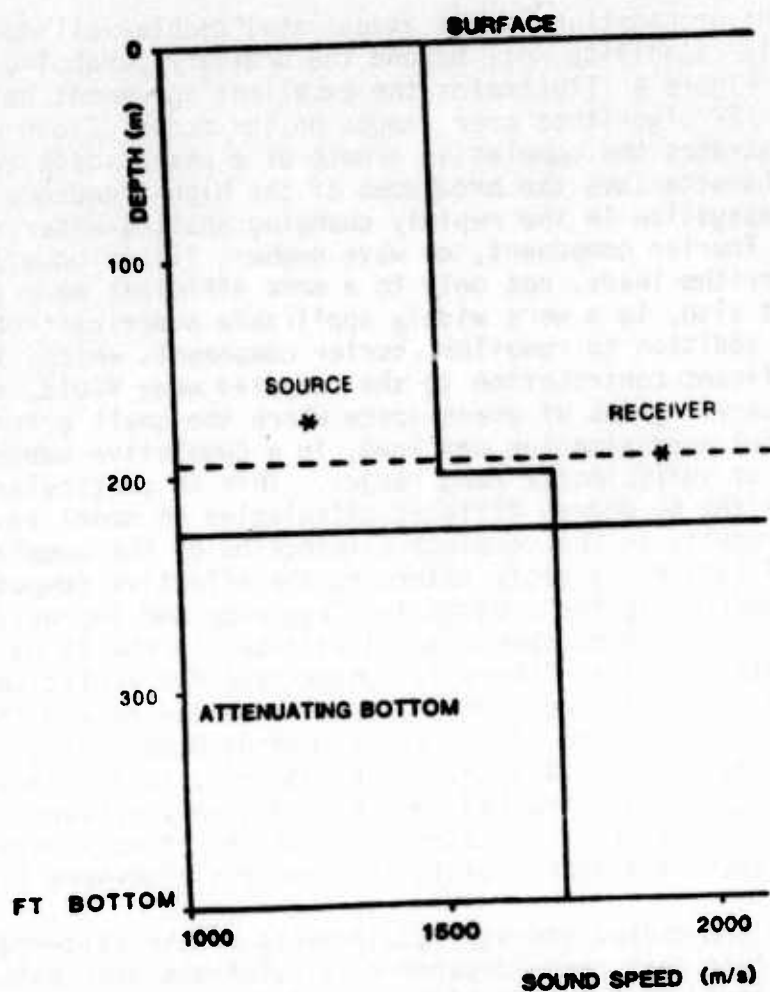
Fig. 7. Model environment 3 and propagation experiment.
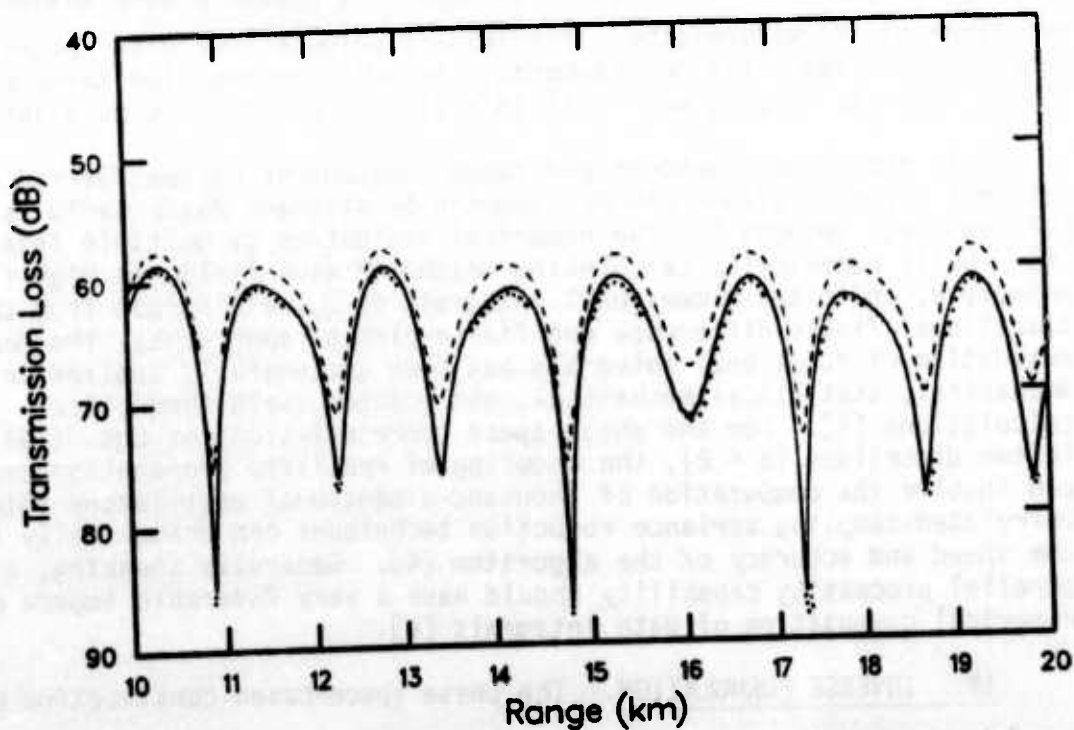
Fig. 8. Transmission loss (dB re 1 m) versus range (km) for model environment 3 at 25 Hz. (———) Real Weyl High Frequency (----) High Frequency (....) FFP

301

For 400 Hz propagation in the exaggerated double-well model of Figure 3, a wide-angle capability well beyond the ordinary parabolic approximation is required. Figure 4 illustrates the excellent agreement between the high-frequency and FFP algorithms over ranges on the order of 500 wavelengths. Figure 6 illustrates the cumulative growth of a phase shift error at long range which characterizes the breakdown of the high-frequency algorithm in the 250 Hz propagation in the rapidly changing shallow-water model of Figure 5. Combining Fourier component, or wave number, filtering with the high-frequency algorithm leads, not only to a more efficient and, thus, faster algorithm, but also, to a more widely applicable numerical scheme. The filtering, in addition to removing Fourier components which, in principle, make no significant contribution to the computed wave field, eliminates those unnecessary regions of phase space where the small error in the high-frequency symbol approximation can lead, in a cumulative manner, to serious discrepancies at sufficiently long ranges. This is particularly well illustrated in the 60 degree filtered calculation on model environment 2 at 250 Hz which results in the complete elimination of the cumulative phase shift error (Figure 6), greatly extending the effective computational range. Sufficiently decreasing the propagation frequency and increasing the jump discontinuity in the sound speed, as illustrated in the 25 Hz propagation in the shallow-water model of Figure 7, demonstrate the violation of energy conservation inherent in the high-frequency wave theory and the now-rapid decay with increasing range of the corresponding numerical algorithm. This growth in the wave field, illustrated in Figure 8, is eliminated by the real Weyl high-frequency algorithm [4], which effectively restores energy conservation, as is also illustrated in Figure 8. A more detailed discussion of these and other points is presented elsewhere [1,4,12].

The speed and modest storage requirements of the filtered one-way algorithm indicate that range-dependent calculations over extended environments should be feasible with current supercomputer technology. Both range-updating and the numerical calculation of the reflected and transmitted fields from an interface should be possible over distances on the order of 10^4 wavelengths. Preliminary computations with range-dependent Munk-profile deep ocean environments, including propagation through extended shadow regions, compare well with adiabatic normal-mode calculations.

Both the range-dependent and range-independent Feynman/Garrod path integral representations can be computed by standard Monte Carlo (statistical sampling) methods for the numerical evaluation of multiple integrals [4]. While numerically calculating Helmholtz wave fields as high (in principle, infinite)-dimensional integrals is quite distinct from the more traditional finite-difference and finite-element approaches, the Monte Carlo evaluation of functional integrals has been successfully applied in quantum mechanical, statistical mechanical, and quantum field theoretical calculations [4]. For the phase space representations of Eqs. (13) and (16) in two dimensions (n = 2), the modeling of realistic propagation experiments can involve the computation of thousand-dimensional oscillatory integrals. Correlated-sampling variance reduction techniques can dramatically improve the speed and accuracy of the algorithm [4]. Generally speaking, a large parallel processing capability should have a very favorable impact on the numerical computation of path integrals [4].

IV. INVERSE FORMULATION. The phase space-based construction of the

square root Helmholtz operator provides the basis for a formulation of the inverse algorithms mentioned in the Introduction. Mathematically, the refractive index field (or its square) is reconstructed from the full-space Helmholtz Green's function G through the relationship

$$B(\underline{x}_t,\underline{x}_t') = (2i/\bar{k}) \lim_{x \to 0} (\partial_x^2 G(x,\underline{x}_t|0,\underline{x}_t')).$$

(26)

The symbol $\Omega_B(\underline{p},\underline{q})$ is then constructed through an inverse Fourier transform of the kernel function $B(\underline{x}_t,\underline{x}_t')$ and subsequently yields the refractive index field upon a direct application of the Weyl composition equation (3) for $|\underline{p}| = 0$. In the homogeneous medium limit, the direct evaluation of the composite symbol reduces to the square of the symbol, $\Omega_B 2(\underline{p},\underline{q}) = \Omega_B^2(\underline{p},\underline{q})$.

The inverse algorithm proceeds around the correspondence diagram (pictorial summary) in a counterclockwise fashion. The direct propagation algorithm requires the inversion of Eq.(3) while the inverse propagation algorithm only requires a direct computation of Eq.(3). Thus the direct propagation problem has been transformed into an "inverse" problem while the wave field inversion problem has been reformulated, in an appropriate sense, as a direct calculation.

The factorization algorithm exactly inverts the inherently nonlinear relationship between the wave field data and the refractive index field as reflected in the Lippmann-Schwinger equation for the propagator [3]. Most importantly, it is a multidimensional formulation. For the "physical experiment," a point source is introduced into the medium defining the initial-value (x = 0) plane. The second derivative with respect to the range of the wave field is then determined as a function of the point source and receiver positions. Collecting the data on the initial-value plane would most probably limit the application of the algorithm to specific types of bore-hole experiments. Moreover, mathematically, the inversion requires the evaluation of singular integrals (generalized functions). Collecting data on a downfield plane (x > 0) leads to a transmission experiment similar to the oceanic sound speed profile inversion method of DeSanto [3]. The downfield wave field provides for an appropriate analytic continuation in the factorization algorithm and connects the analysis with the inverse diffraction problem [3].

The transmission, or propagation, formulation is analogous to tomography. The reference wave number in the factorization analysis corresponds to 2π/(Planck's constant) as opposed to its square playing the role of an energy. The source generation and data collection over parallel planes then naturally correspond to the multidirectional insonifying plane waves and subsequent angular data collection of fixed-energy (frequency) diffraction tomography [3]. For range-dependent environments, the inclusion of backscatter effects, even in an approximate manner, would then provide the basis for a generalized acoustic tomography, extending the diffraction algorithms based on the Born, Rytov, or distorted-wave Born approximations [3]. The nonlinear factorization and subsequent weak-backscatter perturbation theory would extend the linearized weak-scattering treatments into the nonlinear regime. This can be attempted in two ways. Formal field splitting analysis provides the basis for a weak-backscatter perturbation theory within the framework of invariant imbedding [2-4]. The arbitrary-

303

dimensional nature of the factorization analysis in conjunction with mathematical imbedding concepts provides the basis for a spatial-dimensional perturbation theory [2-4]. This essentially involves treating the spatial dimension of both the Helmholtz operator, in general, and the refractive index field, in particular, as a variable and subsequently studying the structure of the resulting family of systems indexed in this manner. For the case of two (different) transversely inhomogeneous half-spaces separated by a planar interface, an inverse algorithm can be initially based on the composition equation (9).

For a transversely inhomogeneous environment, the factorization inversion model invites comparison with "effective one-dimensional" stratified environmental models such as that of Stickler and Deift [4]. In both models, the location of the field source (finite) and the data measurements is within the scattering region. Most importantly, the factorization method is a direct inversion of an arbitrary-dimensional propagation equation which requires less symmetry than those models (i.e., Stickler-Deift) reducible to the standard one-dimensional formulation of Deift-Trubowitz [4] or Gelfand-Levitan [4]. Thus for example, in a general n-dimensional Cartesian formulation, the refractive index field can be a function of as many as (n-1) coordinates in the factorization model, while a function of only one coordinate in an "effective one-dimensional" model. The experiment envisioned and the distinguished direction differ in the two models. In the transversely inhomogeneous environment, the direction in which there is medium homogeneity is distinguished, while in the "effective one-dimensional" model, the one direction in which there is medium inhomogeneity is, in effect, distinguished. Data, in both cases, is collected perpendicular to the distinguished direction. The Stickler-Deift model is essentially a one-dimensional scattering experiment with the surface data, in effect, reflection coefficient data. Thus unlike the transmission experiment, which extensively samples the region of inhomogeneity, in the factorization model, the Stickler-Deift analysis does not account for the presence of "trapped modes" [4]. The formal inclusion of a specific pressure-release surface within the pseudo-differential operator framework would allow for a stratified environmental model and the subsequent quantitative comparison with the Stickler-Deift model.

For applied inverse problems, approximate inversions may prove adequate. Approximate inversion algorithms follow readily from the perturbative treatments of the Weyl composition equation. $K^2(\underline{q})$ is related to $\Omega_B(0,\underline{q})$ in a quadratic fashion and through a linear integral relationship, respectively, in the high-frequency ($\bar{k} \longrightarrow \infty$) and weak-inhomogeneity (Born) limits. In particular, the high-frequency algorithm is based upon choosing, in practice, a $|\underline{p}|$ such that the symbol approaches its asymptotic form, $\Omega_B(\underline{p},\underline{q}) \sim (K^2(\underline{q}) - \underline{p}^2)^{1/2}$. The approach to the asymptotic regime in phase space is governed both by the magnitude of $K^2(\underline{q}) - \underline{p}^2$ (large) and the variation of the refractive index field on the wavelength scale (small). Figure 9 illustrates the high-frequency inversion for the case of a quadratic medium. Applying the full composition equation for the inversion would result in a linear function in X for the real part and an imaginary part which is identically zero. Finally, weighted Hilbert space methods for incorporating prior estimates appear to be applicable to the Fourier-based factorization approach [4].

ACKNOWLEDGMENTS

REFERENCES

1. J.A. Davis, D. White, and R.C. Cavanagh, "NORDA Parabolic Equation Workshop," NORDA tech. note 143, Naval Ocean Research and Development Activity, NSTL Station (1982).
2. L. Fishman and J.J. McCoy, Derivation and application of extended parabolic wave theories. Part I. The factorized Helmholtz equation, J. Math. Phys., 25 (2): 285 (1984).
3. L. Fishman and J.J. McCoy, Factorization, path integral representations, and the construction of direct and inverse wave propagation theories, IEEE Trans. Geosc. Rem. Sens., GE-22 (6): 682 (1984).
4. L. Fishman, J.J. McCoy, and S.C. Wales, Factorization and path integration of the Helmholtz equation: numerical algorithms, J. Acoust. Soc. Am., submitted for publication (1986).
5. M.E. Taylor, "Pseudodifferential Operators," Princeton University Press, Princeton (1981).
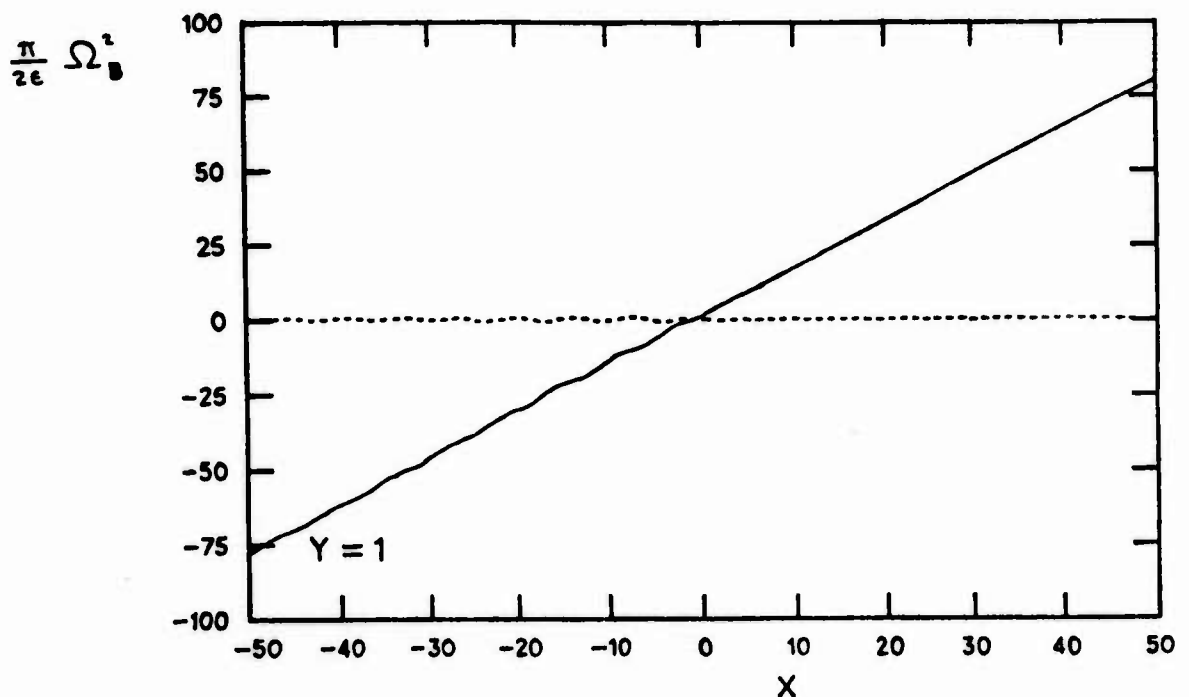
Fig. 9. The real (——) and imaginary (----) parts in the high-frequency approximate inversion for the n = 2 quadratic medium.

6. L. Fishman and A. Whitman, Exact and uniform perturbation solutions of the Helmholtz composition equation, $\underline{J}$. $\underline{Math}$. $\underline{Phys}$., submitted for publication (1986).

7. L. Fishman and J.J. McCoy, Derivation and application of extended parabolic wave theories. Part II. Path integral representations, $\underline{J}$. Math. $\underline{Phys}$., 25 (2): 297 (1984).

8. L.S. Schulman, "Techniques and Applications of Path Integration," Wiley, New York (1981).

9. C. DeWitt-Morette, A. Maheshwari, and B. Nelson, Path integration in nonrelativistic quantum mechanics, $\underline{Phys}$. Rep., 50 (5): March (1979).

10. J.J. McCoy, L. Fishman, and L.N. Frazer, Reflection and transmission at an interface separating transversely inhomogeneous acoustic half-spaces, $\underline{Geophys}$. J. R. Astr. $\underline{Soc}$., to appear (1986).

11. J.J. McCoy and L.N. Frazer, Propagation modelling based on wave field factorization and invariant imbedding, $\underline{Geophys}$. $\underline{J}$. $\underline{R}$. $\underline{Astr}$. $\underline{Soc}$., to appear (1986).

12. L. Fishman and J.J. McCoy, A new class of propagation models based on a factorization of the Helmholtz equation, $\underline{Geophys}$. $\underline{J}$. $\underline{R}$. $\underline{Astr}$. $\underline{Soc}$., 80: 439 (1985).

SCALE INVARIANT EQUATIONS FOR RELATIVISTIC WAVES

Richard A. Weiss
Environmental Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. The basic trace equation of relativistic thermodynamics is decoupled into two Callan-Symanzik type renormalization group equations that connect the matter fields with the thermodynamic gauge parameters. These equations determine two characteristic curves along which the solution to the trace equation assumes a simple form. The differential equation describing the variation of the Grüneisen parameter with pressure is derived. A perturbation procedure is applied to the potential form of the renormalization group equations in order to develop the corresponding potential form of the renormalization equations for waves in a relativistic medium. A method for calculating the Debye temperature for the excited states of solids and quantum liquids is developed. The amplitude and spectrum of waves in thermodynamic media are calculated. A simple equation is derived that scales the wave amplitudes for different material densities (pressures). The results of this paper will have applications to nuclear blast loadings, the interaction of directed energy beams with matter, and to various high density geophysical and astrophysical phenomena.

1. INTRODUCTION. The renormalization group was originally developed for problems in quantum field theory.[1,2] But over the years it has become an important technique in many areas of physics including, phase transitions, critical phenomena, hydrodynamics, and statistical mechanics.[3-5] The renormalization group consists of a set of continuous transformations that establish a correspondence between sets of parameters that define physically different states. In particular, the renormalization group gives a correspondence between systems having different correlation lengths. The correlation length of a physical system is the distance over which local particle densities are correlated. Ordinarily the correlation length is approximately equal to the range of interaction between two component particles, however, near the critical point of a fluid the correlation length is much greater than the range of pair interactions.[6] The renormalization group is commonly described by a set of differential equations for the physical state parameters.[3]

A set of renormalization group equations in potential form has been developed for the ground state parameters of a relativistic thermodynamic system.[7] In this case the correspondence between sets of parameters refers to a change of the local scale (gauge), and this change of scale is equivalent to a change in the correlation length. The potential form of the renormalization group equations consists of a set of differential equations for the two gauge parameters of relativistic thermodynamics.[7] These equations are obtained by requiring the basic trace equation of relativistic

thermodynamics to be invariant under a local scale transformation that corresponds to a change in the correlation length of the system.

The trace equation of relativistic thermodynamics is written as[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 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. The trace equation (1) can be rewritten as[7]

$$\left(1 - b + T\frac{\partial}{\partial T} - bV\frac{\partial}{\partial V}\right)E - 3\left(1 + \gamma + V\frac{\partial}{\partial V} - \gamma T\frac{\partial}{\partial T}\right)P \tag{2}$$

$$= \left(1 - b^a + T\frac{\partial}{\partial T} - b^aV\frac{\partial}{\partial V}\right)E^a$$

where E = relativistic energy density = U/V, E^a = nonrelativistic energy density, and where[7]

$$\gamma = \frac{V}{C_V}\left(\frac{\partial P}{\partial T}\right)_V \tag{3}$$

$$b = \frac{T(\partial P/\partial T)_V}{(P - K_T)} \tag{4}$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \tag{5}$$

where γ = Grüneisen parameter, C_V = heat capacity at constant volume, and

$$K_T = -V\left(\frac{\partial P}{\partial V}\right)_T \tag{6}$$

$$K_T^a = -V\left(\frac{\partial P^a}{\partial V}\right)_T \tag{7}$$

308

are the relativistic and nonrelativistic values of the bulk modulus respectively. The parameters b and γ are the gauge parameters of relativistic thermodynamics.

For a solid or low temperature quantum system the nonrelativistic state equation of the ground state is assumed to have the following form[8],[9]

$$E^a = E^a_o + E^a_j T^j + \cdots \tag{8}$$

$$P^a = P^a_o + P^a_j T^j + \cdots \tag{9}$$

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. Typical examples of systems that are described by equations (8) and (9) are[8]

> j = 1   high temperature solid
> j = 2   low temperature Fermi gas
> j = 5/2   low temperature molecular Bose gas
> j = 4   low temperature solid

A commonly used descriptor of the thermal state equations given by equations (8) and (9) is the nonrelativistic zero-temperature value of the Grüneisen parameter that is defined by[8],[9]

$$\gamma^a_o = \frac{P^a_j}{E^a_j} = \frac{1}{(j-1)} \frac{1}{E^a_j} \frac{d}{dV}(VE^a_j) \tag{10}$$

except for j = 1. Here γ^a_o = nonrelativistic zero-temperature value of the Grüneisen parameter, and V = volume of the material system. When j = 1 , γ^a_o = 2/3 . The zero temperature value of the nonrelativistic bulk modulus is given by $K^a_o = ndP^a_o/dn$, 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[8],[9]

$$E = E_o + E_j T^j + \cdots \tag{11}$$

$$P = P_o + P_j T^j + \cdots \tag{12}$$

$$\gamma_o = \frac{P_j}{E_j} = \frac{1}{(j-1)} \frac{1}{E_j} \frac{d}{dV}(VE_j) \tag{13}$$

except for $j = 1$, when $\gamma_o = 2/3$, 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 Grünesien 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 (8) through (13) yields the following ground state equations[8]

$$E_o - 3[(1 + \gamma_o)P_o - K_o] = E_o^a \tag{14}$$

$$E_j \left(1 + j + \frac{j\gamma_o P_o}{P_o - K_o} + 3n \frac{d\gamma_o}{dn} \right) = E_j^a \left(1 + j + \frac{j\gamma_o^a P_o^a}{P_o^a - K_o^a} \right) \tag{15}$$

The potential forms of the ground state renormalization group equations for the gauge parameters b and γ are determined from the requirement of local scale invariance of equation (2).[7] A form of the renormalization group equations that is commonly used in quantum field theory and the theory of critical phenomena are the Callan-Symanzik equations.[1-3] In this paper the Callan-Symanzik form of the renormalization group equations for the relativistic ground state of thermal media will be obtained directly from equation (2).

Two forms of the renormalization group equations for radiation in matter are obtained in this paper. The potential form of these equations is obtained by a perturbation procedure that is applied to the potential form of the renormalization group equations for the ground state. The Callan-Symanzik form of the renormalization group equations for radiation will be obtained by a perturbation procedure applied directly to equation (2).

An important task of modern physics is the determination of the effects of gauge (scale) invariance on the ground state and excitations of matter. These effects have been treated for the ground state of a thermodynamic system.[8] An approximate treatment has been developed for waves in solids and quantum liquids by assuming that the diffuse radiation factor and the radiation Grüneisen parameter are equal.[7] This paper presents a completely general procedure for calculating the relativistic amplitude and spectrum of elastic waves in solids and quantum liquids. A set of coupled second order radiation equations is developed that determines the relativistic energy density and Grüneisen parameter for radiation in solids and Fermi and Bose liquids.

2. RENORMALIZATION GROUP EQUATIONS FOR THE GROUND STATE.

The ground state of a relativistic thermodynamic medium is described by equation (2) where γ and b are gauge parameters.[7] Equation (2) can be decoupled into two independent equations by noting that E and P are related by the Gibbs-Helmholtz relation as follows

$$\left(\frac{\partial U}{\partial V}\right)_T = E + V\left(\frac{\partial E}{\partial V}\right)_T = T\left(\frac{\partial P}{\partial T}\right)_V - P \tag{16}$$

With the introduction of a Langrange undetermined multiplier η , equation (16) can be rewritten as

$$\eta\left(1 + V \frac{\partial}{\partial V}\right)E + \eta\left(1 - T \frac{\partial}{\partial T}\right)P = 0 \tag{17}$$

Combining equations (2) and (17) yields the following decoupled equations

$$[T \frac{\partial}{\partial T} + (\eta-b)V \frac{\partial}{\partial V} + \eta + 1 - b]E = (T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a)E^a \tag{18}$$

$$[V \frac{\partial}{\partial V} - (\gamma - \frac{\eta}{3})T \frac{\partial}{\partial T} - \frac{\eta}{3} + \gamma + 1]P = 0 \tag{19}$$

The undetermined multiplier η is in general a function of V and T. From equation (19) it follows that

$$\frac{\eta}{3} = \frac{V \frac{\partial P}{\partial V} - \gamma T \frac{\partial P}{\partial T} + (\gamma + 1)P}{P - T \frac{\partial P}{\partial T}} \tag{20}$$

except when the denominator is zero (as in the case of an ideal gas for which $P = nRT$). For $T = 0$, equations (18) through (20) become

$$(\eta_o V \frac{\partial}{\partial V} + \eta_o + 1)E_o = E_o^a \tag{21}$$

$$(V \frac{\partial}{\partial V} - \frac{\eta_o}{3} + \gamma_o + 1)P_o = 0 \tag{22}$$

$$\frac{\eta_o}{3} = \frac{V}{P_o} \frac{dP_o}{dV} + \gamma_o + 1 \tag{23}$$

311

where η_0 = the Lagrange undetermined multiplier for $T = 0$. It is easy to show that combining equation (21) through (23) by eliminating the Lagrange multiplier and using the $T = 0$ form of equation (16) which is

$$P_o = -V \frac{dE_o}{dV} - E_o \tag{24}$$

gives the $T = 0$ ground state equation (14).

Equations (18) and (19) can be rewritten as

$$(T \frac{\partial}{\partial T} + f \frac{\partial}{\partial v} + M)E = \psi^a \tag{25}$$

$$(T \frac{\partial}{\partial T} + h \frac{\partial}{\partial v} + N)P = 0 \tag{26}$$

where

$$V = e^v \tag{27}$$

$$f = \eta - b \tag{28}$$

$$h = \frac{1}{\eta/3 - \gamma} \tag{29}$$

$$M = f + 1 \tag{30}$$

$$N = h - 1 \tag{31}$$

$$\psi^a = (T \frac{\partial}{\partial T} - b^a \frac{\partial}{\partial v} + 1 - b^a)E^a \tag{32}$$

Equations (25) and (26) are immediately recognized to be similar in form to the Callan-Symanzik equations that describe the renormalization group.[1,3] The physical meaning of equations (25) and (26) is that T and V can be considered to be arbitrary parameters and that the trace equation of relativistic thermodynamics is form invariant for any choices of values for T and V, i.e., arbitrary values of temperature and volume are acceptable in equation (2). Equations (25) and (26) connect the matter fields E and P to the gauge fields γ and b. Using equations (25) and (26) allows f and h to be written as

$$f = \frac{T \frac{\partial E}{\partial T} + E - \psi^a}{P - T \frac{\partial P}{\partial T}} \tag{33}$$

$$h = \frac{P - T \frac{\partial P}{\partial T}}{P - K_T} \tag{34}$$

which for the case T = 0 become

$$f_o = \frac{E_o - E_o^a}{P_o} \tag{35}$$

$$h_o = \frac{P_o}{P_o - K_o} \tag{36}$$

The functions f and h are the thermodynamic analogs of the Gell-Mann-Low functions.[1]

In analogy to equation (27) the introduction of

$$T = e^t \tag{37}$$

allows equations (25) and (26) to be rewritten in simpler form by evaluating the derivatives along two characteristic curves as follows

$$\frac{dE}{dt} + ME = \psi^a \tag{38}$$

$$\frac{dP}{dt} + NP = 0 \tag{39}$$

The two characteristic curves are defined by

$$f(V,T) = \frac{dv}{dt} = \frac{T}{V} \frac{dV}{dT} \tag{40}$$

$$h(V,T) = \frac{dv}{dt} = \frac{T}{V} \frac{dV}{dT} \tag{41}$$

313

The characteristic equations (38) and (39) can be solved formally as

$$E = Ce^{-\int Mdt} + e^{-\int Mdt} \int \psi_a e^{\int Mdt} \, dt \tag{42}$$

$$P = De^{-\int Ndt}$$

where C and D are constants. Therefore, along the characteristic curves the solutions to equations (25) and (26) assume a simple form.

The requirement of local scale (gauge) invariance demands that equation (2) be invariant under transformations of the form $P \rightarrow P' = Pe^{-\phi}$ and $E \rightarrow E' = Ee^{-\psi}$ where ϕ and ψ are functions of V and T.[7] These transformations describe a correspondence between physical states having different correlation lengths. The scale invariance condition is taken to be analogous to the condition of local gauge invariance, and yields the following differential equations for the gauge parameters using the $e^{-\phi}$ and $e^{-\psi}$ transformations[7]

$$\left(\frac{d\gamma}{dP}\right)^- = \frac{\gamma \dfrac{T}{\phi} \dfrac{\partial \phi}{\partial T} - \dfrac{V}{\phi} \dfrac{\partial \phi}{\partial V}}{P - T \dfrac{\partial P}{\partial T} + PT \dfrac{\partial \phi}{\partial T}} \tag{44}$$

$$\left(\frac{db}{dE}\right)^- = \frac{\dfrac{T}{\psi} \dfrac{\partial \psi}{\partial T} - b \dfrac{V}{\psi} \dfrac{\partial \psi}{\partial V}}{E + v \dfrac{\partial E}{\partial V} - Ev \dfrac{\partial \psi}{\partial V}} \tag{45}$$

These are the potential forms of the renormalization group equations for the relativistic ground state of a thermodynamic system; they correspond to scale transformations with negative signs in front of the potential functions ϕ and ψ.

From equations (44) and (45) it is clear that the denominators of these equations are not symmetrical under $\phi \rightarrow -\phi$ and $\psi \rightarrow -\psi$. In fact, the requirement of scale invariance for equation (2) under the transformations $P \rightarrow P' = Pe^{+\phi}$ and $E \rightarrow E' = Ee^{+\psi}$ yields the following result

$$\left(\frac{d\gamma}{dP}\right)^+ = \frac{\gamma \dfrac{T}{\phi} \dfrac{\partial \phi}{\partial T} - \dfrac{V}{\phi} \dfrac{\partial \phi}{\partial V}}{P - T \dfrac{\partial P}{\partial T} - PT \dfrac{\partial \phi}{\partial T}} \tag{46}$$

314

$$\left(\frac{db}{dE}\right)^+ = \frac{\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}}{E + V\frac{\partial E}{\partial V} + EV\frac{\partial\psi}{\partial V}} \tag{47}$$

However, the physical derivatives of the gauge parameters must be independent of the signs of the potential functions that appear in the scale transformations, and the simplest way to accomplish this is to assume that the physical derivatives are given by the following symmetric forms

$$\frac{d\gamma}{dP} = \frac{1}{2}\left[\left(\frac{d\gamma}{dP}\right)^- + \left(\frac{d\gamma}{dP}\right)^+\right] \tag{48}$$

$$\frac{db}{dE} = \frac{1}{2}\left[\left(\frac{db}{dE}\right)^- + \left(\frac{db}{dE}\right)^+\right] \tag{49}$$

Equations (48) and (49) can be rewritten as

$$\frac{d\gamma}{dP} = \frac{\left(P - T\frac{\partial P}{\partial T}\right)\left(\gamma\frac{T}{\phi}\frac{\partial\phi}{\partial T} - \frac{V}{\phi}\frac{\partial\phi}{\partial V}\right)}{\left(P - T\frac{\partial P}{\partial T} + PT\frac{\partial\phi}{\partial T}\right)\left(P - T\frac{\partial P}{\partial T} - PT\frac{\partial\phi}{\partial T}\right)} \tag{50}$$

$$\frac{db}{dE} = \frac{\left(E + V\frac{\partial E}{\partial V}\right)\left(\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}\right)}{\left(E + V\frac{\partial E}{\partial V} - EV\frac{\partial\psi}{\partial V}\right)\left(E + V\frac{\partial E}{\partial V} + EV\frac{\partial\psi}{\partial V}\right)} \tag{51}$$

Equations (50) and (51) can be rewritten as

$$\frac{d\gamma}{dP} = \frac{\left(P - T\frac{\partial P}{\partial T}\right)\left(\gamma\frac{T}{\phi}\frac{\partial\phi}{\partial T} - \frac{V}{\phi}\frac{\partial\phi}{\partial V}\right)}{\left(P - T\frac{\partial P}{\partial T}\right)^2 - P^2\left(T\frac{\partial\phi}{\partial T}\right)^2} \tag{50a}$$

$$\frac{db}{dE} = \frac{\left(E + V\frac{\partial E}{\partial V}\right)\left(\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}\right)}{\left(E + V\frac{\partial E}{\partial V}\right)^2 - E^2\left(V\frac{\partial\psi}{\partial V}\right)^2} \tag{51a}$$

Equation (50) has the obvious property that for an ideal gas $d\gamma/dP = 0$ because in this case $P = nRT$, and this agrees with the fact that the Grüneisen parameter for an ideal gas is given by $\gamma = 2/3$. Equations (50) and (51) are the potential forms of the renormalization group equations for the relativistic thermodynamic ground state.

The potential function ϕ is related to the Debye temperature θ_D for high temperatures $(T > \theta_D)$ by $\phi = \theta_D/T$, and for low temperatures $(T < \theta_D)$ by $\phi = T/\theta_D$. For high temperatures, the use of $\phi = \theta_D/T$ in equation (50) gives

$$\frac{d\gamma}{dP} = \left(T\frac{\partial P}{\partial T} - P\right)\left[\gamma\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) + \frac{V}{\theta_D}\frac{\partial\theta_D}{\partial V}\right] / (A_H B_H) \tag{52}$$

where

$$A_H = T\frac{\partial P}{\partial T} - P + P\frac{\theta_D}{T}\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) \tag{53}$$

$$B_H = T\frac{\partial P}{\partial T} - P - P\frac{\theta_D}{T}\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) \tag{54}$$

Equation (52) can be rewritten as

$$\frac{d\gamma}{dP} = \frac{\left(T\frac{\partial P}{\partial T} - P\right)\left[\gamma\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) + \frac{V}{\theta_D}\frac{\partial\theta_D}{\partial V}\right]}{\left(T\frac{\partial P}{\partial T} - P\right)^2 - P^2\left(\frac{\theta_D}{T}\right)^2\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right)^2} \tag{52a}$$

Since in general the following conditions hold for ordinary materials at high temperature[11]

$$T\frac{\partial P}{\partial T} > P \tag{55}$$

$$\left|\frac{V}{\theta_D}\frac{\partial\theta_D}{\partial V}\right| \gtrsim \gamma\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) \tag{56}$$

it follows from equation (52) that in general $d\gamma/dP < 0$. The slow varying Grüneisen parameter condition $(d\gamma/dP \sim 0)$ in equation (52) yields[7]

$$\gamma = \frac{-\frac{V}{\theta_D}\left(\frac{\partial\theta_D}{\partial V}\right)_T}{\left[1 - \frac{T}{\theta_D}\left(\frac{\partial\theta_D}{\partial T}\right)_V\right]} \tag{57}$$

316

which is a standard equation of high pressure physics.[7] If θ_D does not depend appreciably on temperature it follows from equation (57) that

$$\gamma \approx - \frac{V}{\theta_D} \frac{\partial \theta_D}{\partial V} \tag{58}$$

For low temperatures the use of $\phi = T/\theta_D$ in equation (50) yields

$$\frac{d\gamma}{dP} = \left(P - T \frac{\partial P}{\partial T}\right)\left[\gamma\left(1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}\right) + \frac{V}{\theta_D} \frac{\partial \theta_D}{\partial V}\right] / (A_L B_L) \tag{59}$$

where

$$A_L = P - T \frac{\partial P}{\partial T} + P \frac{T}{\theta_D}\left(1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}\right) \tag{60}$$

$$B_L = P - T \frac{\partial P}{\partial T} - P \frac{T}{\theta_D}\left(1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}\right) \tag{61}$$

Equation (59) can be rewritten as

$$\frac{d\gamma}{dP} = \frac{\left(P - T \frac{\partial P}{\partial T}\right)\left[\gamma\left(1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}\right) + \frac{V}{\theta_D} \frac{\partial \theta_D}{\partial V}\right]}{\left(P - T \frac{\partial P}{\partial T}\right)^2 - P^2\left(\frac{T}{\theta_D}\right)^2\left(1 - \frac{T}{\theta_D} \frac{\partial \theta_D}{\partial T}\right)^2} \tag{59a}$$

The slow varying Grüneisen parameter condition ($d\gamma/dP \sim 0$) for equation (59) also yields the result in equation (57). For the case $T = 0$, equation (59) becomes

$$\frac{d\gamma_o}{dP_o} = \left(\gamma_o + \frac{V}{\theta_D^o} \frac{\partial \theta_D^o}{\partial V}\right) / P_o \tag{62}$$

where $\theta_D^o =$ the $T = 0$ value of the Debye temperature. The slow varying Grüneisen parameter approximation applied to equation (62) yields

$$\gamma_o \approx - \frac{V}{\theta_D^o} \frac{\partial \theta_D^o}{\partial V} \tag{63}$$

317

3. LOCAL GAUGE INVARIANCE FOR PHASE ROTATIONS.

The sign of the zero temperature derivative $d\gamma_0/dP_0$ that appears in equation (62) is not obtained from the previous analysis. An indication of what this sign is can be obtained by introducing complex local gauge transformations for the pressure, energy density, and the gauge parameters γ and b. This corresponds to phase rotations of the pressure and energy density relative to the gauge parameters as follows, $\tilde{P}' = \tilde{P}e^{\pm i\phi}$ and $\tilde{E}' = \tilde{E}e^{\pm i\psi}$, and correspondingly $\tilde{\gamma}' = \tilde{\gamma}e^{\pm iS}$ and $\tilde{b}' = \tilde{b}e^{\pm iW}$, where $\tilde{P}$, $\tilde{E}$, $\tilde{\gamma}$ and $\tilde{b}$ must now be taken as complex numbers whose magnitudes are respectively P, E, γ and b, and where $S = S(\phi)$ and $W = W(\psi)$ are the phase angles of the γ and b gauge parameters respectively. Thus whereas the real exponentials $e^{\pm\phi}$ and $e^{\pm\psi}$ correspond to changes in the pressure and energy density, the complex exponentials correspond to phase rotations where the magnitudes P, E, γ, and b are held fixed.

The phase angles $S(\phi)$ and $W(\psi)$ are determined from the condition of local gauge invariance on the fundamental trace equation (2). It is first noted that for P, E, γ, and b held fixed it follows that

$$\frac{d\tilde{\gamma}}{d\tilde{P}} = \frac{\gamma}{P}\frac{dS}{d\phi} \qquad\qquad \frac{d\tilde{b}}{d\tilde{E}} = \frac{b}{E}\frac{dW}{d\psi} \tag{63a}$$

Then it follows that the local gauge invariance conditions and the symmetrization equations (48) and (49) yield the following renormalization group equations for phase rotations in analogy to equations (50a) and (51a)

$$\frac{\gamma}{P}\frac{dS}{d\phi} = \frac{\left(P - T\frac{\partial P}{\partial T}\right)\left(\gamma\frac{T}{\phi}\frac{\partial\phi}{\partial T} - \frac{V}{\phi}\frac{\partial\phi}{\partial V}\right)}{\left(P - T\frac{\partial P}{\partial T}\right)^2 + P^2\left(T\frac{\partial\phi}{\partial T}\right)^2} \tag{63b}$$

$$\frac{b}{E}\frac{dW}{d\psi} = \frac{\left(E + V\frac{\partial E}{\partial V}\right)\left(\frac{T}{\psi}\frac{\partial\psi}{\partial T} - b\frac{V}{\psi}\frac{\partial\psi}{\partial V}\right)}{\left(E + V\frac{\partial E}{\partial V}\right)^2 + E^2\left(V\frac{\partial\psi}{\partial V}\right)^2} \tag{63c}$$

Note the positive sign in the denominators.

For high temperatures, $\phi = \theta_D/T$ and $S = S_D/T$, where S_D = characteristic temperature of the Grüneisen parameter phase angle. In this case equation (63b) becomes

$$\frac{\gamma}{P}\frac{dS}{d\phi} = \frac{\left(T\frac{\partial P}{\partial T} - P\right)\left[\gamma\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right) + \frac{V}{\theta_D}\frac{\partial\theta_D}{\partial V}\right]}{\left(T\frac{\partial P}{\partial T} - P\right)^2 + P^2\left(\frac{\theta_D}{T}\right)^2\left(1 - \frac{T}{\theta_D}\frac{\partial\theta_D}{\partial T}\right)^2} \tag{63d}$$

where

$$\frac{dS}{d\phi} = \frac{\frac{\partial S_D}{\partial V}\,dV - \frac{S_D}{T}\left(1 - \frac{T}{S_D}\frac{\partial S_D}{\partial T}\right)dT}{\frac{\partial \theta_D}{\partial V}\,dV - \frac{\theta_D}{T}\left(1 - \frac{T}{\theta_D}\frac{\partial \theta_D}{\partial T}\right)dT} \tag{63e}$$

A comparison of equations (52) and (63d) shows that at high temperatures $d\gamma/dP < 0$ and $dS/d\phi < 0$.

For low temperatures let $\phi = T/\theta_D$ and $S = T/S_D$ and get from equation (63b)

$$\frac{\gamma}{P}\frac{dS}{d\phi} = \frac{\left(P - T\frac{\partial P}{\partial T}\right)\left[\gamma\left(1 - \frac{T}{\theta_D}\frac{\partial \theta_D}{\partial T}\right) + \frac{V}{\theta_D}\frac{\partial \theta_D}{\partial V}\right]}{\left(P - T\frac{\partial P}{\partial T}\right)^2 + P^2\left(\frac{T}{\theta_D}\right)^2\left(1 - \frac{T}{\theta_D}\frac{\partial \theta_D}{\partial T}\right)^2} \tag{63f}$$

where the derivative $dS/d\phi$ is evaluated as follows

$$\frac{dS}{d\phi} = \frac{\theta_D}{S_D}\frac{\left(1 - \frac{T}{S_D}\frac{\partial S_D}{\partial T}\right)dT - \frac{T}{S_D}\frac{\partial S_D}{\partial V}\,dV}{\left(1 - \frac{T}{\theta_D}\frac{\partial \theta_D}{\partial T}\right)dT - \frac{T}{\theta_D}\frac{\partial \theta_D}{\partial V}\,dV} \tag{63g}$$

Thus at low temperatures $dS/d\phi > 0$ and from equations (59) and (63f) it follows that $d\gamma/dP > 0$ at low temperatures. In particular the $T = 0$ case of equation (63g) is

$$\left(\frac{dS}{d\phi}\right)_o = \frac{\theta_D^o}{S_D^o} \tag{63h}$$

where θ_D^o and S_D^o are the $T = 0$ values of θ_D and S_L respectively. Therefore for $T = 0$, equation (63f) becomes

$$\frac{\gamma_o}{P_o}\frac{\theta_D^o}{S_D^o} = \left(\gamma_o + \frac{V}{\theta_D^o}\frac{\partial \theta_D^o}{\partial V}\right)/P_o \tag{63i}$$

Note that $P_o < 0$ for bound systems such as solids at standard pressure.

319

A comparison of equation (62) and (63i) then shows that

$$\frac{P_o}{\gamma_o} \frac{d\gamma_o}{dP_o} = \frac{\theta_D^o}{S_D^o}$$

(63j)

and therefore it follows that if $\theta_D^o > 0$ and $S_D^o > 0$ then $d\gamma_o/dP_o > 0$, and from equation (62) one concludes that

$$\gamma_o > \left| \frac{V}{\theta_D^o} \frac{\partial \theta_D^o}{\partial V} \right|$$

(63k)

Since in general $d\gamma/dP < 0$ at high temperatures, while equation (63j) shows that $d\gamma/dP > 0$ at $T = 0$, it follows that as the temperature of a solid or quantum liquid is lowered a value of the temperature is finally reached at which point the sign of $d\gamma/dP$ changes from a negative to a positive value.

It should be pointed out that a similar analysis is not possible for the angles ψ and W (associated with E and b respectively) that appear in equation (63c) because the gauge parameter $b \to 0$ when $T \to 0$. This condition combined with equation (63c) suggests that ψ has the following low temperature form

$$\psi(V,T) = A(V) \exp \left(\int G(V,T) dT \right)$$

(63ℓ)

where $G(V,T) =$ some polynomial function of T .

4. RENORMALIZATION GROUP EQUATIONS FOR EXCITATIONS.

When electromagnetic or mechanical waves of small amplitude are present in a thermodynamic medium, the renormalization group equations for the excitations can be obtained either directly as a perturbation on the ground state equation (2) for the energy density, or as a perturbation on the potential forms of the ground state renormalization group equations given by equations (44) through (49). When excitations are present the pressure, energy density, bulk modulus, and heat capacity become $P + P_r$, $E + E_r$, $K_T + K_{Tr}$, and $C_V + C_{Vr}$ respectively, where $P_r =$ radiation pressure, $E_r =$ radiation energy density, and where

$$K_{Tr} = n \left(\frac{\partial P_r}{\partial n} \right)_T = \text{radiation bulk modulus}$$

$$C_{Vr} = v \left(\frac{\partial E_r}{\partial T} \right)_V = \text{radiation heat capacity}$$

320

The Grüneisen parameter for a thermodynamic medium with excitations is obtained from equation (3) as follows

$$\gamma + \delta_r = \frac{V}{C_V + C_{Vr}} \frac{\partial}{\partial T}(P + P_r) \tag{64}$$

where δ_r = change in the system Grüneisen parameter due to the presence of radiation. Expanding equation (64), subtracting equation (3), and keeping only first order terms gives

$$\delta_r = \frac{V}{C_V} \frac{\partial E_r}{\partial T} (\gamma_r - \gamma) = \frac{C_{Vr}}{C_V} (\gamma_r - \gamma) \tag{65}$$

where γ_r = Grüneisen parameter for the radiation field itself and is given by

$$\gamma_r = \frac{\partial P_r}{\partial T} \bigg/ \frac{\partial E_r}{\partial T} \tag{66}$$

The gauge parameter b for an excited thermodynamic medium is obtained from equation (4) to be

$$b + \beta_r = \frac{T\left(\frac{\partial P}{\partial T} + \frac{\partial P_r}{\partial T}\right)}{P - K_T + P_r - K_{Tr}} \tag{67}$$

where β_r = change in b parameter due to the presence of radiation in the system. Expanding equation (67), keeping only first order terms, and subtracting equation (4) gives

$$\beta_r = \frac{T \frac{\partial P_r}{\partial T}}{P - K_T} - \frac{T \frac{\partial P}{\partial T} (P_r - K_{Tr})}{(P - K_T)^2} \tag{68}$$

$$= b_r \frac{P_r - K_{Tr}}{P - K_T} - \frac{T \frac{\partial P}{\partial T} (P_r - K_{Tr})}{(P - K_T)^2}$$

where b_r = radiation gauge parameter given by

321

$$b_r = \frac{T \frac{\partial P_r}{\partial T}}{P_r - K_{Tr}} \tag{68a}$$

The parameters γ_r and b_r are the two radiation gauge parameters of the thermal medium.

The renormalization group equations for radiation can be put into a Callan-Symanzik form by first combining equation (2) with equations (64) and (67) as follows

$$\left[1 - (b + \beta_r) + T \frac{\partial}{\partial T} - (b + \beta_r)V \frac{\partial}{\partial V} \right] (E + E_r) \tag{69}$$

$$- 3\left[1 + \gamma + \delta_r + V \frac{\partial}{\partial V} - (\gamma + \delta_r)T \frac{\partial}{\partial T} \right] (P + P_r)$$

$$= \left[1 - (b^a + \beta_r^a) + T \frac{\partial}{\partial T} - (b^a + \beta_r^a)V \frac{\partial}{\partial V} \right] (E^a + E_r^a)$$

where β_r^a is given by the nonrelativistic analog of equation (68). Subtracting equation (2) from equation (69), and keeping only first order radiation terms yields the following radiation equation

$$\left(1 - b + T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V} \right) E_r - \beta_r\left(T \frac{\partial P}{\partial T} - P \right) \tag{70}$$

$$- 3\left[\left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T} \right) P_r - \delta_r\left(T \frac{\partial P}{\partial T} - P \right) \right]$$

$$= \left(1 - b^a + T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} \right) E_r^a - \beta_r^a\left(T \frac{\partial P^a}{\partial T} - P^a \right)$$

where the following standard thermodynamic relation was used

$$\frac{\partial U}{\partial V} = E + V \frac{\partial E}{\partial V} = T \frac{\partial P}{\partial T} - P \tag{71}$$

Equation (70) is a first order radiation equation that can be applied to any thermodynamic system such as gases, solids, and quantum liquids.

Equation (70) 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 (71), which for radiation becomes

$$\frac{\partial U_r}{\partial V} = E_r + V \frac{\partial E_r}{\partial V} = T \frac{\partial P_r}{\partial T} - P_r \tag{72}$$

Introducing a radiation Lagrange multiplier η_r as follows

$$\eta_r \left[E_r + V \frac{\partial E_r}{\partial V} + P_r - T \frac{\partial P_r}{\partial T} \right] = 0 \tag{73}$$

allows equation (70) to be separated as follows

$$\left(T \frac{\partial}{\partial T} + f_r \frac{\partial}{\partial v} + M_r \right) E_r - \beta_r \left(T \frac{\partial P}{\partial T} - P \right) = \psi_r^a \tag{74}$$

$$\left(T \frac{\partial}{\partial T} + h_r \frac{\partial}{\partial v} + N_r \right) P_r - h_r \delta_r \left(T \frac{\partial P}{\partial T} - P \right) = 0 \tag{75}$$

where v is defined in equation (27), and where

$$f_r = \eta_r - b \tag{76}$$

$$h_r = (\eta_r/3 - \gamma)^{-1} \tag{77}$$

$$M_r = f_r + 1 \tag{78}$$

$$N_r = h_r - 1 \tag{79}$$

$$\psi_r^a = \left(T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V} + 1 - b^a \right) E_r^a - \beta_r^a \left(T \frac{\partial P^a}{\partial T} - P^a \right) \tag{80}$$

Equations (72), (74), and (75) are coupled radiation equations that give E_r, P_r, and η_r. These equations simplify somewhat for solids and quantum liquids.[7]

The potential form of the renormalization group equations for radiation will now be obtained by a perturbation procedure that is applied to equations (44) through (49). When excitations are present the renormalization group equations (44) through (47) become

$$\frac{\left(\frac{d\gamma}{dP}\right)^- + \left(\frac{d\delta_r}{dP}\right)^-}{\left(1 + \frac{dP_r}{dP}\right)} = \frac{(\gamma + \delta_r)\frac{T}{(\phi + \phi_r)}\left(\frac{\partial\phi}{\partial T} + \frac{\partial\phi_r}{\partial T}\right) - \frac{V}{(\phi + \phi_r)}\left(\frac{\partial\phi}{\partial V} + \frac{\partial\phi_r}{\partial V}\right)}{P + P_r - T\left(\frac{\partial P}{\partial T} + \frac{\partial P_r}{\partial T}\right) + (P + P_r)T\left(\frac{\partial\phi}{\partial T} + \frac{\partial\phi_r}{\partial T}\right)} \tag{81}$$

$$\frac{\left(\frac{d\gamma}{dP}\right)^+ + \left(\frac{d\delta_r}{dP}\right)^+}{\left(1 + \frac{dP_r}{dP}\right)} = \frac{(\gamma + \delta_r)\frac{T}{(\phi + \phi_r)}\left(\frac{\partial\phi}{\partial T} + \frac{\partial\phi_r}{\partial T}\right) - \frac{V}{(\phi + \phi_r)}\left(\frac{\partial\phi}{\partial V} + \frac{\partial\phi_r}{\partial V}\right)}{P + P_r - T\left(\frac{\partial P}{\partial T} + \frac{\partial P_r}{\partial T}\right) - (P + P_r)T\left(\frac{\partial\phi}{\partial T} + \frac{\partial\phi_r}{\partial T}\right)} \tag{82}$$

$$\frac{\left(\frac{db}{dE}\right)^- + \left(\frac{d\beta_r}{dE}\right)^-}{\left(1 + \frac{dE_r}{dE}\right)} = \frac{\frac{T}{(\psi + \psi_r)}\left(\frac{\partial\psi}{\partial T} + \frac{\partial\psi_r}{\partial T}\right) \quad (b + \beta_r)\frac{V}{(\psi + \psi_r)}\left(\frac{\partial\psi}{\partial V} + \frac{\partial\psi_r}{\partial V}\right)}{E + E_r + V\left(\frac{\partial E}{\partial V} + \frac{\partial E_r}{\partial V}\right) - (E + E_r)V\left(\frac{\partial\psi}{\partial V} + \frac{\partial\psi_r}{\partial V}\right)} \tag{83}$$

$$\frac{\left(\frac{db}{dE}\right)^+ + \left(\frac{d\beta_r}{dE}\right)^+}{\left(1 + \frac{dE_r}{dE}\right)} = \frac{\frac{T}{(\psi + \psi_r)}\left(\frac{\partial\psi}{\partial T} + \frac{\partial\psi_r}{\partial T}\right) - (b + \beta_r)\frac{V}{(\psi + \psi_r)}\left(\frac{\partial\psi}{\partial V} + \frac{\partial\psi_r}{\partial V}\right)}{E + E_r + V\left(\frac{\partial E}{\partial V} + \frac{\partial E_r}{\partial V}\right) + (E + E_r)V\left(\frac{\partial\psi}{\partial V} + \frac{\partial\psi_r}{\partial V}\right)} \tag{84}$$

Expanding equations (81) through (84), keeping only first order terms, and subtracting equations (44) through (47) gives

$$\left(\frac{d\delta_r}{dP}\right)^- = \left(\frac{d\gamma}{dP}\right)^- \frac{dP_r}{dP} + \frac{A_r - \left(\frac{d\gamma}{dP}\right)^- B_r^-}{P - T\frac{\partial P}{\partial T} + PT\frac{\partial\phi}{\partial T}} \tag{85}$$

$$\left(\frac{d\delta_r}{dP}\right)^+ = \left(\frac{d\gamma}{dP}\right)^+ \frac{dP_r}{dP} + \frac{A_r - \left(\frac{d\gamma}{dP}\right)^+ B_r^+}{P - T\frac{\partial P}{\partial T} - PT\frac{\partial\phi}{\partial T}} \tag{86}$$

$$\frac{dP_r}{dP} = \frac{K_{Tr} + \frac{\partial P_r}{\partial T}n\frac{dT}{dn}}{K_T + \frac{\partial P}{\partial T}n\frac{dT}{dn}} \tag{86a}$$

324

$$A_r = \frac{T}{\phi}\left(\delta_r \frac{\partial \phi}{\partial T} + \gamma \frac{\partial \phi_r}{\partial T}\right) + \frac{\phi_r}{\phi}\left(\frac{V}{\phi}\frac{\partial \phi}{\partial V} - \gamma \frac{T}{\phi}\frac{\partial \phi}{\partial T}\right) - \frac{V}{\phi}\frac{\partial \phi_r}{\partial V} \tag{87}$$

$$B_r^- = P_r - T \frac{\partial P_r}{\partial T} + P_r T \frac{\partial \phi}{\partial T} + PT \frac{\partial \phi_r}{\partial T} \tag{88}$$

$$B_r^+ = P_r - T \frac{\partial P_r}{\partial T} - P_r T \frac{\partial \phi}{\partial T} - PT \frac{\partial \phi_r}{\partial T} \tag{89}$$

and

$$\left(\frac{d\beta_r}{dE}\right)^- = \left(\frac{db}{dE}\right)^- \frac{dE_r}{dE} + \frac{C_r - \left(\frac{db}{dE}\right)^- D_r^-}{E + V \frac{\partial E}{\partial V} - EV \frac{\partial \psi}{\partial V}} \tag{90}$$

$$\left(\frac{d\beta_r}{dE}\right)^+ = \left(\frac{db}{dE}\right)^+ \frac{dE_r}{dE} + \frac{C_r - \left(\frac{db}{dE}\right)^+ D_r^+}{E + V \frac{\partial E}{\partial V} + EV \frac{\partial \psi}{\partial V}} \tag{91}$$

$$\frac{dE_r}{dE} = \frac{E_r + P_r - T \frac{\partial P_r}{\partial T} + \frac{\partial E_r}{\partial T} n \frac{dT}{dn}}{E + P - T \frac{\partial P}{\partial T} + \frac{\partial E}{\partial T} n \frac{dT}{dn}} \tag{91a}$$

$$C_r = \frac{T}{\psi}\frac{\partial \psi_r}{\partial T} + \frac{\psi_r}{\psi}\left(b \frac{V}{\psi}\frac{\partial \psi}{\partial V} - \frac{T}{\psi}\frac{\partial \psi}{\partial T}\right) - \frac{V}{\psi}\left(b \frac{\partial \psi_r}{\partial V} + \beta_r \frac{\partial \psi}{\partial V}\right) \tag{92}$$

$$D_r^- = E_r + V \frac{\partial E_r}{\partial V} - E_r V \frac{\partial \psi}{\partial V} - EV \frac{\partial \psi_r}{\partial V} \tag{93}$$

$$D_r^+ = E_r + V \frac{\partial E_r}{\partial V} + E_r V \frac{\partial \psi}{\partial V} + EV \frac{\partial \psi_r}{\partial V} \tag{94}$$

where $(d\gamma/dP)^-$ and $(d\gamma/dP)^+$ are given by equations (44) and (46), while $(db/dE)^-$ and $(db/dE)^+$ are given by equations (45) and (47).

The symmetric equations that describe the variation of the radiation gauge parameters are then given by

$$\frac{d\delta_r}{dP} = \frac{1}{2}\left[\left(\frac{d\delta_r}{dP}\right)^- + \left(\frac{d\delta_r}{dP}\right)^+\right] \tag{95}$$

$$\frac{d\beta_r}{dE} = \frac{1}{2}\left[\left(\frac{d\beta_r}{dE}\right)^- + \left(\frac{d\beta_r}{dE}\right)^+\right] \tag{96}$$

Equations (95) and (96) are the potential forms of the renormalization group equations for a thermodynamic system that contains radiation. These equations determine the radiation potentials ϕ_r and ψ_r. The potential function ϕ_r is related to the radiative change in the Debye temperature θ_{Dr} by $\phi_r = \theta_{Dr}/T$ for high temperatures and by $\phi_r = T/\theta_{Dr}$ for low temperatures. The derivatives on the left side of equations (95) and (96) can be evaluated using equations (65) and (68) and the following simple relationships

$$\frac{d\delta_r}{dP} = \frac{n\dfrac{\partial\delta_r}{\partial n} + \dfrac{\partial\delta_r}{\partial T}\,n\,\dfrac{dT}{dn}}{K_T + \dfrac{\partial P}{\partial T}\,n\,\dfrac{dT}{dn}} \tag{97}$$

$$\frac{d\beta_r}{dE} = \frac{n\dfrac{\partial\beta_r}{\partial n} + \dfrac{\partial\beta_r}{\partial T}\,n\,\dfrac{dT}{dn}}{E + P - T\dfrac{\partial P}{\partial T} + \dfrac{\partial E}{\partial T}\,n\,\dfrac{dT}{dn}} \tag{98}$$

5. WAVES IN SOLIDS AND QUANTUM LIQUIDS. Excitations in relativistic solids and quantum liquids have already been considered using some simplifying assumptions.[7] A general procedure for calculating the amplitude and spectrum of relativistic waves in solids and quantum liquids will be outlined here. The energy density and pressure for radiation in these systems is written as

$$E_r^a = E_{or}^a + E_{jr}^a T^j + \cdots \tag{99}$$

$$P_r^a = P_{or}^a + P_{jr}^a T^j + \cdots \tag{100}$$

and

$$E_r = E_{or} + E_{jr} T^j + \cdots \tag{101}$$

$$P_r = P_{or} + P_{jr} T^j + \cdots \tag{102}$$

326

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 (66) and (99) through (102) to be

$$\gamma_{or}^a = \frac{P_{jr}^a}{E_{jr}^a} \qquad \gamma_{or} = \frac{P_{jr}}{E_{jr}} \tag{103}$$

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.

The basic relativistic equations describing excitations in solids and quantum liquids are written as (equations 76 and 77 of Reference 7)

$$E_{or} - 3[(1 + \gamma_o)P_{or} - K_{or}] - 3\frac{E_{jr}}{E_j}P_o(\gamma_{or} - \gamma_o) = E_{or}^a \tag{104}$$

$$jE_j(\alpha K_{or} - \beta P_{or}) + E_{jr}S_{jr} = jE_j^a\left(\alpha^a K_{or}^a - \beta^a P_{or}^a\right) + E_{jr}^a T_{jr}^a \tag{105}$$

where

$$S_{jr} = 1 + j + \frac{jP_o\gamma_{or}}{P_o - K_o} + 3n\frac{d\gamma_{or}}{dn} - 3(j-1)(\gamma_{or} - \gamma_o)^2 \tag{106}$$

$$T_{jr}^a = 1 + j + \frac{jP_o^a\gamma_{or}^a}{P_o^a - K_o^a} \tag{107}$$

327

and where

$$\alpha = \frac{\gamma_o P_o}{(P_o - K_o)^2} \qquad\qquad \alpha^a = \frac{\gamma_o^a P_o^a}{(P_o^a - K_o^a)^2} \qquad\qquad (108)$$

$$\beta = \frac{\gamma_o K_o}{(P_o - K_o)^2} \qquad\qquad \beta^a = \frac{\gamma_o^a K_o^a}{(P_o^a - K_o^a)^2} \qquad\qquad (109)$$

Equations (104) and (105) can be deduced directly from equation (70) by using equations (11) and (101). For example, the expression for δ_r that appears in equation (65) can be evaluated for the zero temperature case of solids and quantum liquids to be

$$\delta_{or} = \frac{E_{1r}}{E_j} (\gamma_{or} - \gamma_o) \qquad\qquad (110)$$

Using the following basic relationships

$$P_{or} = n \frac{dE_{or}}{dn} - E_{or} \qquad\qquad (111)$$

$$K_{or} = n \frac{dP_{or}}{dn} = n^2 \frac{d^2 E_{or}}{dn^2} \qquad\qquad (112)$$

allows equations (104) and (105) to be written as

$$3n^2 \frac{d^2 E_{or}}{dn^2} - 3(1 + \gamma_o) n \frac{dE_{or}}{dn} + (3\gamma_o + 4) E_{or} + 3 \frac{E_{1r}}{E_j} P_o (\gamma_o - \gamma_{or}) = E_{or}^a \qquad\qquad (113)$$

$$\alpha n^2 \frac{d^2 E_{or}}{dn^2} - \beta n \frac{dE_{or}}{dn} + \beta E_{or} + \frac{E_{1r} S_{1r}}{j E_j}$$

$$= \frac{E_j^a}{E_j} \left[\alpha^a n^2 \frac{d^2 E_{or}^a}{dn^2} - \beta^a \left(n \frac{dE_{or}^a}{dn} - E_{or}^a \right) \right] + \frac{E_{1r}^a T_{1r}^a}{j E_j} \qquad\qquad (114)$$

328

The quantities S_{jr} and T_{jr}^a are functions of γ_{or} and γ_{or}^a, while the ratios E_{jr}/E_j, E_j^a/E_j, and E_{jr}^a/E_j are functions of γ_{or}, γ_{or}^a, γ_o and γ_o^a.[7] Equations (113) and (114) are simultaneous second order differential equations that determine E_{or} and γ_{or} in terms of E_{or}^a, γ_{or}^a, and in terms of the parameters of the ground state.

Equations (113) and (114) must be solved simultaneously to obtain the relativistic wave amplitude and the relativistic wave number in terms of the corresponding nonrelativistic values. However, in order to obtain an approximate value for the relativistic wave amplitude and wave number, only equation (113) will be used. The zero temperature values of the nonrelativistic and relativistic radiation energy densities are respectively written as[10]

$$E_{or}^a = \frac{1}{4} K_o^a k_a^2 A_a^2 \tag{115}$$

$$E_{or} = \frac{1}{4} K_o k^2 A^2 \tag{116}$$

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 equations (115) and (116) into equation (113) gives

$$\frac{3}{4} K_o n^2 \frac{d^2}{dn^2}(k^2 A^2) + \frac{3}{4}\left[2n \frac{dK_o}{dn} - (1 + \gamma_o)K_o\right] n \frac{d}{dn}(k^2 A^2) \tag{117}$$

$$+ \frac{1}{4} k^2 A^2 \left[3n^2 \frac{d^2 K_o}{dn^2} - 3(1 + \gamma_o)n \frac{dK_o}{dn} + (3\gamma_o + 4)K_o\right] + g = \frac{1}{4} k_a^2 A_a^2 K_o^a$$

where

$$g = 3 \frac{E_{jr}}{E_j} P_o(\gamma_o - \gamma_{or}) \tag{118}$$

As a crude approximation take $\gamma_{or} = 1/3$, $E_{jr}/E_j = E_{or}/E_o$, and $P_o \sim n^{\sigma_o}$ where σ_o = adiabatic index, and assume kA is not explicitly density dependent, and get

$$g \sim 3E_{or}(\sigma_o - 1)\left(\gamma_o - \frac{1}{3}\right) \tag{119}$$

and

$$k^2 A^2 \left[3n^2 \frac{d^2 K_o}{dn^2} - 3(1 + \gamma_o)n \frac{dK_o}{dn} + (3\sigma_o\gamma_o - \sigma_o + 5)K_o\right] = k_a^2 A_a^2 K_o^a \tag{120}$$

329

Using $K_o \sim n^{\sigma_o}$ in equation (120) gives

$$k^a A^2 = \frac{k_a^2 A_a^2 K_o^a / K_o}{3\sigma_o^2 - 7\sigma_o + 5} \tag{121}$$

$$\approx k_a^2 A_a^2 \left(\frac{3\sigma_o^2 - 3\sigma_o(2 + \gamma_o) + 3\gamma_o + 4}{3\sigma_o^2 - 7\sigma_o + 5} \right)$$

The relative values of kA and $k_a A_a$ at low and high densities depend on the values of σ_o and γ_o at these densities. For a low density Fermi gas where $\sigma_o = 5/3$ and $\gamma_o = 2/3$ one has $kA = 0.77 \, k_a A_a$, for a solid where $\sigma_o \sim 8$ and $\gamma_o \sim 3.83$ the result is $kA = 0.69 \, k_a A_a$. The high density limit of equation (121) is somewhat more delicate. If the high density limit is associated with asymptotic freedom, then $\sigma_o = 4/3$ and $\gamma_o = 1/3$ and $kA = k_a A_a$. On the other hand, if at high densities the interactions increase without limit and $\sigma_o \to \infty$, but with $\gamma_o = $ constant, then $kA = k_a A_a$. However, γ_o is probably a function of σ_o and may be written as $\gamma_o = \sigma_o - 4/3$.[8] In this case equation (121) goes as $\sigma_o/(3\sigma_o^2)$ as $\sigma_o \to \infty$ so that $kA/(k_a A_a) \to 0$. This behaviour contrasts with the results of the first order radiation differential equation approximation of Reference 7, where $kA/(k_a A_a) \to 3\sigma_o^2/\sigma_o$ and is large for $\sigma_o \to \infty$ with $\gamma_o = $ constant while $kA/(k_a A_a) \to \sigma_o/\sigma_o = 1$ for $\sigma_o \to \infty$ with $\gamma_o = \sigma_o - 4/3$. Finally for the case of asymptotic freedom with $\sigma_o = 4/3$ and $\gamma_o = 1/3$ the first order differential equation approximation gives $kA = 0.65 \, k_a A_a$. Thus the effect of relativistic thermodynamics on wave amplitudes is system and model dependent.

6. RELATIVISTIC PHASE VELOCITY. A general procedure is given for determining the relativistic phase velocity for waves in solids and quantum liquids. The procedure will be first to determine E_{or}^a and γ_{or}^a from the values of the nonrelativistic sound speed, then to solve for the relativistic quantities E_{or} and γ_{or} using quations (113) and (114), and then finally working backward to determine the relativistic sound speed from the relativistic energy density and Grüneisen parameter for the radiation.

The nonrelativistic expression for the phase velocity of mechanical waves is given by[8]

$$\left(\frac{W^a}{c} \right)^2 = \frac{K_T^a + \gamma^a T \frac{\partial P^a}{\partial T}}{\Sigma^a + \gamma^a \theta^a} \tag{122}$$

where

$$\Sigma^a = E^a + P^a + K_T^a - T \frac{\partial P^a}{\partial T} \tag{123}$$

330

$$\Theta^a = T \frac{\partial E^a}{\partial T} + T \frac{\partial P^a}{\partial T} \tag{124}$$

where W^a = nonrelativistic sound speed, and c = light speed. The zero temperature limit of equation (122) is[7],[8]

$$\left(\frac{W^a_o}{c}\right)^2 = \frac{K^a_o}{E^a_o + P^a_o + K^a_o} \tag{125}$$

where W^a_o = sound speed in a T = 0 solid or quantum liquid.

The nonrelativistic diffuse radiation factor is defined to be

$$\Gamma^a_r = \frac{P^a_r}{E^a_r} \tag{126}$$

For isotropic radiation the diffuse radiation factor can be expressed as follows[10]

$$\Gamma^a_r = \frac{1}{3} + \frac{n}{W^a} \frac{dW^a}{dn} \tag{127}$$

The phase velocity that appears in equation (122) through (124) can be expanded in powers of the temperature, so that the diffuse radiation factor can be written in the following general form

$$\Gamma^a_r = \Gamma^a_{or} + \Gamma^a_{jr} T^j + \cdots \tag{128}$$

Therefore the coefficients of the diffuse radiation are expressed, through the phase velocity W^a, in terms of the ground state functions E^a_o, P^a_o, E^a_j, and P^a_j.

Using equations (99) and (100) to represent the radiation pressure and energy density that appear in the defining equation (126) for the diffuse radiation factor, expanding the denominator, and keeping only first order terms yields the following results

$$\Gamma_{or}^a = \frac{P_{or}^a}{E_{or}^a} \tag{129}$$

$$\Gamma_{jr}^a = \frac{E_{jr}^a}{E_{or}^a} \left(\gamma_{or}^a - \Gamma_{or}^a \right) \tag{130}$$

where the left hand side of these equations are obtained from the sound speed.

Equations (129) and (130) can be used to determine E_{or}^a, E_{jr}^a, and γ_{or}^a. For instance, placing equation (111) into equation (129) gives

$$\Gamma_{or}^a = \frac{n}{E_{or}^a} \frac{dE_{or}^a}{dn} - 1 \tag{131}$$

which is a differential equation that can be solved for E_{or}^a, since Γ_{or}^a is known from the ground state parameters through equations (122) and (127). In fact, the solution of equation (131) is

$$E_{or}^a = nD_{or}^a \exp \left(\int^n \Gamma_{or}^a \frac{dn}{n} \right) \tag{132}$$

where D_{or}^a = constant. The determination of E_{jr}^a and γ_{or}^a from equation (130) goes as follows. It is easily shown that[7]

$$E_{jr}^a = nD_{jr}^a \exp \left[- (j-1) \int^n \gamma_{or}^a \frac{dn}{n} \right] \tag{133}$$

where D_{jr}^a = constant. Placing equation (133) into equation (130) gives an integral equation which can be solved for γ_{or}^a. In this way the nonrelativistic radiation energy density and Grüneisen parameter, E_{or}^a and γ_{or}^a respectively, can be determined from the phase velocity.

The corresponding relativistic values of the radiation energy density E_{or} and Grüneisen parameter γ_{or} are obtained by the solution of the simultaneous equations (113) and (114). The relativistic thermal energy density coefficient is then determined by[7]

332

$$E_{jr} = nD_{jr} \exp\left[- (j-1) \int^{n} \gamma_{or} \frac{dn}{n} \right] \qquad (134)$$

where D_{jr} = constant. The relativistic diffuse radiation factor coefficients are then calculated by

$$\Gamma_{or} = \frac{P_{or}}{E_{or}} = \frac{n}{E_{or}} \frac{dE_{or}}{dn} - 1 \qquad (135)$$

$$\Gamma_{jr} = \frac{E_{jr}}{E_{or}} (\gamma_{or} - \Gamma_{or}) \qquad (136)$$

The relativistic diffuse radiation factor is then written as

$$\Gamma_r(n,T) = \Gamma_{or} + \Gamma_{jr} T^j + \cdots \qquad (137)$$

$$= \frac{P_r}{E_r}$$

Finally the relativistic phase velocity is obtained as a solution to the following equation

$$\Gamma_r = \frac{1}{3} + \frac{n}{W} \frac{dW}{dn} \qquad (138)$$

which can be written as

$$\frac{W}{c} = \exp\left[- \int_{n}^{\infty} (\Gamma_r - \frac{1}{3}) \frac{dn}{n} \right] \qquad (139)$$

If it is assumed that the diffuse radiation factor is independent of temperature it follows from equations (136) and (137) that

$$\Gamma_r = \Gamma_{or} = \gamma_{or} \qquad (140)$$

In this case it follows that $P_{or} = \gamma_{or} E_{or}$, and the wave equations (104) and (105) reduce to a set of coupled first order differential equations instead of the second order differential equations that appear in equations (113) and (114).[7]

7. **SCALING THE WAVE AMPLITUDE.** This section obtains a simple expression for the wave amplitude in a T = 0 system. Combining equation (135) with the T = 0 form of equation (138) yields

$$\frac{n}{E_{or}} \frac{dE_{or}}{dn} = \frac{4}{3} + \frac{n}{W_o} \frac{dW_o}{dn} \tag{141}$$

where W_o = zero temperature value of the relativistic phase velocity. An equation analogous to (141) holds for the corresponding nonrelativistic quantities. The solution of equation (141) is easily obtained to be

$$E_{or} = G_r W_o n^{4/3} \tag{142}$$

where G_r = constant independent of n. The relationship between wave number and phase velocity is $W_o = \omega/k$, where ω = angular frequency. Using this in equation (116) gives the following expression for the radiation energy density

$$E_{or} = \frac{1}{4} \frac{\omega^2 A^2 K_o}{W_o^2} \tag{143}$$

Combining equation (142) and (143) gives the wave amplitude as

$$A^2 = \frac{4G_r W_o^3 n^{4/3}}{\omega^2 K_o} \tag{144}$$

Let n and n_1 be two particle number densities, then it follows from equation (144) that

$$\left[\frac{A(n)}{A(n_1)}\right]^2 = \left[\frac{W_o(n)}{W_o(n_1)}\right]^3 \left(\frac{n}{n_1}\right)^{4/3} \frac{K_o(n_1)}{K_o(n)} \tag{145}$$

334

which is the scaling equation for wave amplitudes under a change in density in a T = 0 system. Equation (145) is expected to be a good approximation for finite temperature systems if the corresponding finite temperature parameters are used.

8. ELECTROMAGNETIC WAVES IN MATTER. The relativistic calculation of the energy density and phase velocity of electromagnetic waves in matter proceeds in a manner analogous to the case of mechanical waves. The relativistic and nonrelativistic electromagnetic energy densities at zero temperature are given by

$$\bar{E}_{or} = \frac{1}{2} \left(\epsilon_o E^2 + \mu_o H^2 \right) \tag{146}$$

$$E_{or}^a = \frac{1}{2} \left(\epsilon_o^a E_a^2 + \mu_o^a H_a^2 \right) \tag{147}$$

where E and H = relativistic electric and magnetic radiation fields respectively; E_a and H_a = nonrelativistic electric and magnetic radiation fields respectively; ϵ_o, μ_o and ϵ_o^a, μ_o^a = zero temperature values of the relativistic and nonrelativistic permittivities and permeabilities respectively. The thermal part of the radiation energy density and pressure is written in the form of equations (101) through (103), and therefore the determination of $\bar{E}_{or}$ and γ_{or} is necessary for a relativistic description of electromagnetic waves in matter. The crude approximation $\gamma_{or} = 1/3$ is made in this section, and the problem is to determine ϵ_o, μ_o, E, and H .

It will be assumed that the nonrelativistic values of the zero temperature values of the permittivity and permeability can be represented by some theoretical expressions in terms of the density, pressure and Grüneisen parameters as follows

$$\epsilon_o^a = X[n, P_o^a(n), \gamma_o^a(n)] \tag{148}$$

$$\mu_o^a = Y[n, P_o^a(n), \gamma_o^a(n)] \tag{149}$$

Then the relativistic values ϵ_o and μ_o are determined using the same functional relationships but now evaluated for the relativistic values of the pressure and Grüneisen parameter as follows

$$\epsilon_o = X[n, P_o(n), \gamma_o(n)] \tag{150}$$

$$\mu_o = Y[n, P_o(n), \gamma_o(n)] \tag{151}$$

The relativistic values of P_o and γ_o are obtained from the solution of the simultaneous equations (14) and (15). Thus the relativistic values of ϵ_o and μ_o are obtained indirectly from the ground state solution of the relativistic trace equation (1).

A complete relativistic thermodynamic description of electromagnetic waves in matter requires the determination of E_{or} and γ_{or} by the simultaneous solution of equations (113) and (114). But for an approximate solution only equation (113) can be used with $\gamma_{or} = 1/3$. Placing equation (146) and (147) into equation (113) gives the following differential equations for E and H

$$\frac{3}{2} \epsilon_o n^2 \frac{d^2}{dn^2}(E^2) + \frac{3}{2}\left[2n \frac{d\epsilon_o}{dn} - (1 + \gamma_o)\epsilon_o\right]n \frac{d}{dn}(E^2) \tag{152}$$

$$+ \frac{1}{2} E^2\left[3n^2 \frac{d^2\epsilon_o}{dn^2} - 3(1 + \gamma_o)n \frac{d\epsilon_o}{dn} + (3\gamma_o + 4)\epsilon_o\right] + g_E = \frac{1}{2} \epsilon_o^a E_a^2$$

$$\frac{3}{2} \mu_o n^2 \frac{d^2}{dn^2}(H^2) + \frac{3}{2}\left[2n \frac{d\mu_o}{dn} - (1 + \gamma_o)\mu_o\right]n \frac{d}{dn}(H^2) \tag{153}$$

$$+ \frac{1}{2} H^2\left[3n^2 \frac{d^2\mu_o}{dn^2} - 3(1 + \gamma_o)n \frac{d\mu_o}{dn} + (3\gamma_o + 4)\mu_o\right] + g_H = \frac{1}{2} \mu_o^a H_a^2$$

where g_E and g_H are obtained from equations (118) and (119) to be given approximately as

$$g_E \sim \frac{3}{2} \epsilon_o E^2(\sigma_o - 1)(\gamma_o - \frac{1}{3}) \tag{154}$$

$$g_H \sim \frac{3}{2} \mu_o H^2(\sigma_o - 1)(\gamma_o - \frac{1}{3}) \tag{155}$$

Combining equations (152) through (155) yields the following equations for the relativistic values of the electric and magnetic fields in matter assuming that E and H are not explicitly density dependent

$$E^2\left[3n^2 \frac{d^2\epsilon_o}{dn^2} - 3(1 + \gamma_o)n \frac{d\epsilon_o}{dn} + (3\sigma_o\gamma_o - \sigma_o + 5)\epsilon_o\right] = \epsilon_o^a E_a^2 \tag{156}$$

$$H^2\left[3n^2 \frac{d^2\mu_o}{dn^2} - 3(1 + \gamma_o)n \frac{d\mu_o}{dn} + (3\sigma_o\gamma_o - \sigma_o + 5)\mu_o\right] = \mu_o^a H_a^2 \tag{157}$$

336

Assuming $\varepsilon_o \sim n^{\rho_o}$, $\mu_o \sim n^{\nu_o}$, and $P_o \sim n^{\sigma_o}$ in equations (156) and (157) gives the following approximate equations

$$E^2 = \frac{E_a^2 \varepsilon_o^a / \varepsilon_o}{3\rho_o^2 - 6\rho_o - \sigma_o + 3(\sigma_o - \rho_o)\gamma_o + 5} \tag{158}$$

$$H^2 = \frac{H_a^2 \mu_o^a / \mu_o}{3\nu_o^2 - 6\nu_o - \sigma_o + 3(\sigma_o - \nu_o)\gamma_o + 5} \tag{159}$$

where

$$\sigma_o = \frac{n}{P_o}\frac{dP_o}{dn} = \frac{K_o}{P_o} \tag{160}$$

$$\rho_o = \frac{n}{\varepsilon_o}\frac{d\varepsilon_o}{dn} \tag{161}$$

$$\nu_o = \frac{n}{\mu_o}\frac{d\mu_o}{dn} \tag{162}$$

The determination of the relativistic phase velocity for electromagnetic waves in matter proceeds in a manner similar to that for the case of mechanical waves that was treated in equations (122) through (139) except that the temperature dependent nonrelativistic phase velocity is given by

$$\left(\frac{W^a}{c}\right)^2 = (\varepsilon^a \mu^a)^{-1} \tag{163}$$

where ε^a and μ^a = nonrelativistic permittivity and permeability respectively. The zero temperature limit of equation (163) is written as

$$\left(\frac{W_o^a}{c}\right)^2 = (\varepsilon_o^a \mu_o^a)^{-1} \tag{164}$$

337

From the phase velocities given in equations (163) and (164) one can calculate the nonrelativistic radiation energy density and Grüneisen parameters, E_{or}^a and γ_{or}^a respectively, by the procedure outlined in equations (126) through (133). Then the solution of equations (113) and (114) yields the corresponding relativistic radiation energy density and Grüneisen parameter, E_{or} and γ_{or} respectively. From E_{or} and γ_{or} one obtains an estimate of the relativistic diffuse radiation factor Γ_r by the procedure outlined in equations (134) through (138). Finally the relativistic phase velocity for electromagnetic waves in matter is given by

$$\frac{W}{c} = \exp\left[-\int_0^n (\frac{1}{3} - \Gamma_r) \frac{dn}{n}\right] \tag{165}$$

If it is assumed that the diffuse radiation factor is independent of temperature the substitution $\Gamma_r = \gamma_{or}$ can be made in equation (165).

9. CONCLUSION. The trace equation of the relativistic thermodynamic ground state is reduced to two Callan-Symanzik type renormalization group equations that connect the matter fields E and P with the thermodynamic gauge fields γ and b. The gauge parameters are necessary to insure that the trace equation is invariant under a local scale transformation. The assumption of local scale invariance under changes of the correlation length of the system leads in a natural way to a set of differential equations for the gauge parameters. These are the potential forms of the renormalization group equations for the ground state. The renormalization group equations for radiation in matter can be written in terms of radiation potentials or in the form of radiative Callan-Symanzik equations. The radiation equations for a general thermodynamic system are applied to waves in solids and quantum liquids, and a set of coupled second order differential equations are developed that determine the relativistic radiation energy density and Grüneisen parameter. Finally, a simple scaling relation is developed for the amplitude of waves propagating through materials of different density.

No mass or energy scale occurs in the equations of relativistic thermodynamics, but the temperature and volume scales that appear in these equations is similar to the mass cutoff parameter that appears in the Callan-Symanzik equations of quantum field theory.[1] Therefore in analogy to the dimensional transmutation of Coleman and Weinberg there may appear a mass associated with the gauge bosons that correspond with the gauge parameters γ_r and b_r.[12] On the other hand, the ground state of a relativistic thermodynamic system may exhibit a broken symmetry in which case the gauge bosons can become massive by the Higgs mechanism.[1] In either case massive thermal gauge bosons should exist that are associated with the thermodynamic gauge parameters γ_r and b_r. The gauge boson associated with the Grüneisen parameter should exist even for $T = 0$ solids or quantum liquids. Therefore new physical phenomena are expected to occur in bulk matter that is subjected to high pressures. In addition, the results of this paper should have engineering and geophysics applications.

REFERENCES

1. Huang, K., <u>Quarks Leptons and Gauge Fields</u>, World Scientific, Singapore, 1982.

2. Moriyasu, K., <u>An Elementary Primer for Gauge Theory</u>, Heyden & Son, Philadelphia, 1983.

3. Pfeuty, P. and Toulouse, G., <u>Introduction to the Renormalization Group and to Critical Phenomena</u>, John Wiley, New York, 1977.

4. Wilson, K. G. and Kogut, J., "The Renormalization Group and the ϵ Expansion", Phys. Rep. 12C, p. 75, 1974.

5. Wilson, K. G., "The Renormalization Group and Critical Phenomena", Rev. Mod. Phys. Vol. 55, No. 3, July 1983.

6. Sengers, A. L., Hocken, R., and Sengers, J. V., "Critical Point Universality and Fluids", Physics Today, pp. 42, Dec. 1977.

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., <u>Relativistic Thermodynamics</u>, Vols 1 and 2, Exposition Press, New York, 1976.

9. Zharkov, V. N. and Kalinin, V. A., <u>Equations of State for Solids at High Pressures and Temperatures</u>, Consultants Bureau, New York, 1971.

10. Brillouin, L., <u>Tensors in Mechanics and Elasticity</u>, Academic Press, New York, 1964.

11. Boehler, R., "Adiabats of Quartz, Coesite, Olivine, and Magnesium Oxide to 50 KBAR and 1000 K, and the Adiabatic Gradient in the Earth's Mantle", Journal of Geophysical Research, Vol 87, No. B7, 5501-5506, July 10, 1982.

12. Coleman, S. and Weinberg, E., "Radiative Corrections as the Origin of Spontaneous Symmetry Breaking", Phys. Rev., D7, p. 1888, 1974.

RELATIVISTIC WAVE EQUATIONS FOR REAL GASES

Richard A. Weiss
Environmental Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. The relativistic wave equation for a generalized thermo-
dynamic system is developed. The solution of this equation is obtained
for the real gases whose pressure is described by a virial expansion. A
procedure is given for calculating the relativistic amplitude and phase
velocity for mechanical waves propagating in real gases. The relativistic
wave amplitude is calculated by a virial expansion whose coefficients are
determined from the wave equation. The relativistic effects on wave propa-
gation in gases manifest themselves only through the third virial coeffi-
cient, and therefore these effects are expected to be observed only at high
pressures such as found in atmospheric nuclear explosions, the interaction
of directed energy beams with the atmosphere, stellar atmospheres, and in
high-pressure-physics laboratory experiments. The effects of curvature
waves in spacetime on the pressure of real gases are also considered, and
applications to the detection of gravitational radiation are suggested.

1. INTRODUCTION. Local gauge (scale) invariance plays a fundamental
role in the description of diverse physical phenomena.[1-3] The requirement
of local scale invariance suggests that relativistic thermodynamics can be
formulated on the basis of a relativistic trace equation that relates the
pressure and internal energy fields to a set of gauge parameters.[4,5] The
trace equation for a thermodynamic system can be written as a partial dif-
ferential equation involving the energy density, pressure, and two gauge
parameters.[5] The scale transformations refer to changes in the correlation
length of the system, and the scale invariance establishes a correspondence
between different physical states of a relativistic thermodynamic system.
This correspondence is encompassed by the renormalization group differential
equations that describe the variation of the gauge parameters with the
magnitude of ambient matter fields such as pressure and energy density.

For the case where the thermodynamic system has a well defined zero
temperature state, such as is the case for solids and quantum liquids, the
trace equation leads to a set of coupled second order differential equations
for the simultaneous determination of the zero temperature values of the
pressure and Grüneisen parameter.[4] For real gases whose pressure is described
by a virial expansion, the trace equation yields a relativistic expression
for the third virial coefficient.[4] This paper derives the relativistic
equation for radiation in a generalized thermodynamic system, and then de-
rives the equations that are necessary to calculate the wave amplitudes
and phase velocity for waves in real gases. This is done by a perturbation
procedure that is applied to the basic trace equation that describes the

ground state of a relativistic thermodynamic system.

The trace equation of relativistic thermodynamics is written as[4]

$$U + T\left(\frac{dU}{dT}\right)_{PV} - 3V \frac{d}{dV}(PV)_U = U^a + T\left(\frac{dU^a}{dT}\right)_{P^aV} \tag{1}$$

where U = relativistic 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. It is easy to show that equation (1) can be written as follows[4]

$$\frac{\partial}{\partial T}(TU) - bV \frac{\partial U}{\partial V} - 3V\left[\frac{\partial}{\partial V}(PV) - \gamma \frac{\partial U}{\partial V}\right] \tag{2}$$

$$= \frac{\partial}{\partial T}(TU^a) - b^a V \frac{\partial U^a}{\partial V}$$

where

$$\gamma = \frac{V}{C_V}\left(\frac{\partial P}{\partial T}\right)_V \tag{3}$$

$$b = \frac{T(\partial P/\partial T)_V}{(P - K_T)} \tag{4}$$

$$b^a = \frac{T(\partial P^a/\partial T)_V}{(P^a - K_T^a)} \tag{5}$$

and where γ = relativistic Grüneisen parameter, C_V = relativistic heat capacity at constant volume, and where

$$K_T = - V\left(\frac{\partial P}{\partial V}\right)_T \tag{6}$$

$$K_T^a = -V\left(\frac{\partial P^a}{\partial V}\right)_T \tag{7}$$

are the relativistic and nonrelativistic values of the bulk modulus respectively. The nonrelativistic Grüneisen parameter is defined as follows

$$\gamma^a = \frac{V}{C_V^a}\left(\frac{\partial P^a}{\partial T}\right)_V \tag{8}$$

where C_V^a = nonrelativistic heat capacity and constant volume. Equation (2) can be rewritten in terms of the energy density as follows[5]

$$\left(1 - b + T\frac{\partial}{\partial T} - bV\frac{\partial}{\partial V}\right)E - 3\left(1 + \gamma + V\frac{\partial}{\partial V} - \gamma T\frac{\partial}{\partial T}\right)P \tag{9}$$

$$= \left(1 - b^a + T\frac{\partial}{\partial T} - b^a V\frac{\partial}{\partial V}\right)E^a$$

where $\bar{E} = U/V$ = relativistic energy density, and $E^a = U^a/V$ = nonrelativistic energy density. The parameters γ and b are the two gauge parameters of relativistic thermodynamics.[5]

Wave motion in relativistic gases can be of two types. The first corresponds to mechanical vibrations of the gas which results in pressure changes in time and space. This type of wave motion is described by a relativistic wave equation for real gases. Such an equation can be developed by first considering relativistic waves in a completely general thermodynamic medium and then specializing to the case of real gases. It is required to find both the relativistic amplitude and sound speed for waves in real gases. In order to do this the nonrelativistic wave amplitude and phase velocity must first be determined. The relativistic effects appear only in the third and higher virial coefficients of the real gas state equation.[4] Therefore it is necessary to solve the relativistic radiation equation for real gases to determine the relativistic value of the third virial coefficient for radiation in the real gas system. The relativistic diffuse radiation factor and the relativistic sound speed in real gases are then determined from the values of the relativistic third virial coefficient for radiation. These effects should be important only for real gases at high pressures where the third virial coefficient contributes significantly to the equation of state.

A second kind of wave motion in gases can be induced by the coupling of the wave motions in spacetime with some characteristic parameter of real gases. The wave motions in spacetime are gravitational waves. The attempts

at detecting gravitational radiation over the past twenty years by various methods including the use of solid body resonance detectors have not been successful.[6-8] This may be due to the lack of adequate sensitivity of present day detectors because the cosmic sources of gravity waves are thought to be very weak.[9-11] On the other hand the lack of positive experimental results using solid body detectors may be due to a basic insensitivity of this type of detector, and it has been suggested that real gases and liquids may be better suited for a detector material because the third virial coefficient is expected to be sensitive to changes in the metric of spacetime.[12] This paper calculates the adiabatic changes in gas pressure that are expected to occur in a detector that is subjected to the tidal effects of gravity waves.

The procedure followed in this paper is to: a) review the theory of the relativistic ground state of real gases, b) determine the equations that describe relativistic waves in a generalized thermodynamic medium, c) develop a simple nonrelativistic calculation of the amplitude of waves in real gases, d) determine the solution of the wave equation for real gases by performing a perturbation calculation on the relativistic ground state equation for real gases, e) determine the relativistic values of the wave amplitude and phase velocity, f) determine the adiabatic changes of pressure, volume, and temperature for a real gas that is interacting with gravitational radiation.

2. RELATIVISTIC GROUND STATE OF REAL GASES. The form of the solution of the trace equation (1) depends on the type of physical system being considered. For real gases the nonrelativistic and relativistic pressure, energy density, bulk modulus, and molar heat capacity (specific heat) are written in virial form respectively as[13,14]

$$P^a = nR^aT[1 + nB^a(T) + n^2C^a(T) + \cdots] \tag{10}$$

$$E^a = nR^aT\left[\frac{3}{2} - nT\frac{\partial B^a}{\partial T} - \frac{1}{2}n^2T\frac{\partial C^a}{\partial T} - \cdots\right] \tag{11}$$

$$K_T^a = nR^aT[1 + 2nB^a(T) + 3n^2C^a(T) + \cdots] \tag{12}$$

$$\tilde{C}_V^a = R^a\left[\frac{3}{2} - n\left(T^2\frac{\partial^2 B^a}{\partial T^2} + 2T\frac{\partial B^a}{\partial T}\right) - \frac{1}{2}n^2\left(T^2\frac{\partial^2 C^a}{\partial T^2} + 2T\frac{\partial C^a}{\partial T}\right) - \cdots\right] \tag{13}$$

and

$$P = nRT[1 + nB(T) + n^2C(T) + \cdots] \tag{14}$$

344

$$E = nRT\left[\frac{3}{2} - nT\frac{\partial B}{\partial T} - \frac{1}{2}n^2T\frac{\partial C}{\partial T} - \cdots\right] \qquad (15)$$

$$K_T = nRT\left[1 + 2nB(T) + 3n^2C(T) + \cdots\right] \qquad (16)$$

$$\tilde{C}_V = R\left[\frac{3}{2} - n\left(T^2\frac{\partial^2 B}{\partial T^2} + 2T\frac{\partial B}{\partial T}\right) - \frac{1}{2}n^2\left(T^2\frac{\partial^2 C}{\partial T^2} + 2T\frac{\partial C}{\partial T}\right) - \cdots\right] \qquad (17)$$

where

$$n = N/V = 1/\tilde{V} \qquad (18)$$

where N = number of moles, $\tilde{V}$ = molar volume; R^a, $B^a(T)$ and $C^a(T)$ = nonrelativistic values of the gas constant, second virial coefficient, and third virial coefficient respectively; R, B(T), and C(T) = relativistic values of the gas constant, second virial coefficient, and third virial coefficient respectively; and $\tilde{C}_V^a$ and $\tilde{C}_V$ = nonrelativistic and relativistic values of the molar heat capacity (specific heat) respectively. The relationship between the extensive, intensive, and molar quantities that are used in this paper is as follows

$$C_V^a = N\tilde{C}_V^a = \left(\frac{\partial U^a}{\partial T}\right)_V \qquad (19)$$

$$C_V = N\tilde{C}_V = \left(\frac{\partial U}{\partial T}\right)_V \qquad (20)$$

$$E^a = n\tilde{U}^a = U^a/V \qquad (21)$$

$$E = n\tilde{U} = U/V \qquad (22)$$

where $\tilde{U}^a$ and $\tilde{U}$ = nonrelativistic and relativistic internal energy per mole.

The relationship between the relativistic and the nonrelativistic functions that appear in equations (10) through (17) are given by[4]

$$R = R^a \tag{23}$$

$$B(T) = B^a(T) \tag{24}$$

$$C(T) = C^a(T) - 3[B^a(T)]^2 \, \ell n \, \psi^a \tag{25}$$

where

$$\psi^a = \frac{T}{T_R} \left| \frac{B^a(T)}{B^a(T_R)} \right|^{2/3} = \frac{T}{T_{CR}} \left| \frac{B^a(T)}{B^a(T_{CR})} \right|^{2/3} \tag{26}$$

and where T_R = relativity temperature constant, and T_{CR} = conjugate relativity temperature constant. The relationship between T_R and T_{CR} is shown in Figure 1. Thus the relativistic effects enter the real gas state equation only through the third and higher virial coefficients; the ideal gas term and the second virial coefficient are unaffected.

The relativity temperature T_R and the conjugate relativity temperature T_{CR} are related to the critical temperature of a real gas. The conditions for the critical point can be expressed in terms of the second and third virial coefficients as follows[15]

$$B(T_{crit}) = - \tilde{V}(T_{crit}) \tag{27}$$

$$3C(T_{crit}) = \tilde{V}^2(T_{crit}) \tag{28}$$

or equivalently

$$3C(T_{crit}) = B^2(T_{crit}) \tag{29}$$

where T_{crit} = critical temperature. Equations (24), (25), and (29) give the critical point condition as[16]

$$C^a = \frac{1}{3} [B^a]^2 (1 + 9 \, \ell n \, \psi^a) \tag{30}$$

and gives the relationship between T_{crit} and T_R (or T_{CR}) that is shown in Figure 2. Figure 3 gives the dependence of the critical molar volume on the relativity temperature.

The Grüneisen function can be evaluated for real gases using equation (14) which gives

$$\left(\frac{\partial P}{\partial T}\right)_n = nR[1 + nf_1(T) + n^2 f_2(T) + \cdots] \tag{31}$$

where

$$f_1(T) = T \frac{\partial B}{\partial T} + B \tag{32}$$

$$f_2(T) = T \frac{\partial C}{\partial T} + C \tag{33}$$

and equation (17) which gives

$$\frac{1}{\tilde{C}_V} = \frac{2}{3R} [1 + ng_1(T) + n^2 g_2(T) + \cdots] \tag{34}$$

where

$$g_1(T) = \frac{2}{3}\left(T^2 \frac{\partial^2 B}{\partial T^2} + 2T \frac{\partial B}{\partial T}\right) \tag{35}$$

$$g_2(T) = \frac{1}{3}\left(T^2 \frac{\partial^2 C}{\partial T^2} + 2T \frac{\partial C}{\partial T}\right) \tag{36}$$

$$+ \frac{4}{9}\left(T^2 \frac{\partial^2 B}{\partial T^2} + 2T \frac{\partial B}{\partial T}\right)^2$$

Then equations (3), (31), and (34) give the relativistic Grüneisen parameter as

347

$$\gamma = \frac{2}{3} \left[1 + n\gamma_1(T) + n^2 \gamma_2(T) + \cdots \right] \tag{37}$$

where

$$\gamma_1(T) = f_1(T) + g_1(T) \tag{38}$$

$$\gamma_2(T) = f_2(T) + g_2(T) + f_1(T)g_1(T) \tag{39}$$

Expressions analogous to equations (31) through (39) hold for the nonrelativistic Grüneisen parameter.

3. EXCITATIONS IN THERMODYNAMIC SYSTEMS. This section considers mechanical radiation in thermal media. Only small amplitude vibrations are considered. When radiation is present in a thermal system the relativistic energy density, pressure, bulk modulus, and heat capacity are written as, $E + E_r$, $P + P_r$, $K_T + K_{Tr}$, and $C_V + C_{Vr}$ respectively, while the corresponding nonrelativistic quantities become $E^a + E_r^a$, $P^a + P_r^a$, $K_T^a + K_{Tr}^a$, and $C_V^a + C_{Vr}^a$ respectively, where

E_r^a and E_r = nonrelativistic and relativistic radiation energy density respectively

P_r^a and P_r = nonrelativistic and relativistic radiation pressure respectively

$$K_{Tr} = n\left(\frac{\partial P_r}{\partial n}\right)_T = \text{relativistic bulk modulus of the radiation}$$

$$K_{Tr}^a = n\left(\frac{\partial P_r^a}{\partial n}\right)_T = \text{nonrelativistic bulk modulus of the radiation}$$

$$C_{Vr} = V\left(\frac{\partial E_r}{\partial T}\right)_V = \text{relativistic heat capacity of radiation}$$

$$C_{Vr}^a = V\left(\frac{\partial E_r^a}{\partial T}\right)_V = \text{nonrelativistic heat capacity of radiation}$$

The radiation terms are assumed to be much smaller than the ground state terms, i.e., $E_r \ll E$ and $P_r \ll P$.

348

The Grüneisen parameter γ and the gauge parameter b become $\gamma + \delta_r$ and $b + \beta_r$, where δ_r and β_r = incremental changes in the parameters γ and b when radiation is present in the system. The increment in the Grüneisen parameter of the system due to the presence of small amplitude radiation is obtained from the defining equation (3) by noting that

$$\gamma + \delta_r = \frac{V}{C_V + C_{Vr}} \frac{\partial}{\partial T} (P + P_r) = \frac{V}{C_V \left(1 + \frac{C_{Vr}}{C_V}\right)} \frac{\partial}{\partial T} (P + P_r) \tag{40}$$

Expanding the denominator in equation (40), keeping only first order terms, and finally substracting equation (3) gives

$$\delta_r = \frac{V}{C_V} \frac{\partial E_r}{\partial T} (\gamma_r - \gamma) \tag{41}$$

$$= \frac{V}{C_V} \left[E_r \frac{\partial \Gamma_r}{\partial T} + (\Gamma_r - \gamma) \frac{\partial E_r}{\partial T} \right] \tag{42}$$

where γ_r = relativistic Grüneisen parameter of the radiation itself, and is defined as

$$\gamma_r = \frac{V}{C_{Vr}} \frac{\partial P_r}{\partial T} = \frac{\partial P_r}{\partial T} \bigg/ \frac{\partial E_r}{\partial T} \tag{43}$$

and where Γ_r = relativistic diffuse radiation factor which is defined by

$$\Gamma_r = \frac{P_r}{E_r} \tag{44}$$

Note that a comparison of equations (41) and (42) shows that if Γ_r is independent of temperature, then $\Gamma_r = \gamma_r$.

Similarly, the increment in the gauge parameter b due to the presence of radiation in the medium is obtained from equation (4) by observing that

$$b + \beta_r = \frac{T\left(\frac{\partial P}{\partial T} + \frac{\partial P_r}{\partial T}\right)}{P - K_T + P_r - K_{Tr}} = \frac{T\left(\frac{\partial P}{\partial T} + \frac{\partial P_r}{\partial T}\right)}{\left(P - K_T\right)\left(1 + \frac{P_r - K_{Tr}}{P - K_T}\right)}$$ (45)

Expanding the denominator in equation (45), keeping only first order terms, and subtracting equation (4) gives

$$\beta_r = \frac{T\frac{\partial P_r}{\partial T}}{P - K_T} - \frac{T\frac{\partial P}{\partial T}\left(P_r - K_{Tr}\right)}{\left(P - K_T\right)^2}$$ (46)

$$= \frac{\gamma_r T\frac{\partial E_r}{\partial T}}{P - K_T} - \frac{\gamma T\frac{\partial E}{\partial T}\left(P_r - K_{Tr}\right)}{\left(P - K_T\right)^2}$$

$$= \frac{T}{P - K_T}\frac{\partial}{\partial T}\left(\Gamma_r E_r\right) - \frac{T\frac{\partial P}{\partial T}}{\left(P - K_T\right)^2}\left[\Gamma_r E_r + V\frac{\partial}{\partial V}\left(\Gamma_r E_r\right)\right]$$

Note that equation (46) can be rewritten as

$$\beta_r = b_r\frac{P_r - K_{Tr}}{P - K_T} - \frac{T\frac{\partial P}{\partial T}\left(P_r - K_{Tr}\right)}{\left(P - K_T\right)^2}$$ (47)

where b_r = radiation gauge parameter given by

$$b_r = \frac{T\frac{\partial P_r}{\partial T}}{P_r - K_{Tr}}$$ (48)

The parameters γ_r and b_r are the two gauge parameters for radiation in a thermal medium.

350

The corresponding nonrelativistic values of the δ_r and β_r are given by

$$\delta_r^a = \frac{V}{C_V^a} \frac{\partial E_r^a}{\partial T} (\gamma_r^a - \gamma^a) \tag{49}$$

$$\beta_r^a = \frac{T}{P^a - K_T^a} \frac{\partial P_r^a}{\partial T} - \frac{T \frac{\partial P^a}{\partial T}}{\left(P^a - K_T^a\right)^2} \left[\Gamma_r^a E_r^a + V \frac{\partial}{\partial V} \left(\Gamma_r^a E_r^a\right)\right] \tag{50}$$

where γ_r^a is given by the nonrelativistic analog of equation (43), and Γ_r^a is given by the nonrelativistic analog of equation (44). Note also that δ_r and β_r are small quantities because E_r is assumed to be small compared to E. But the radiation gauge parameters γ_r and b_r, and the diffuse radiation factor Γ_r, are not small quantities since they are defined as the ratio of two small numbers.

When radiation is present in a general thermodynamic system, equation (9) can be written as

$$\left[1 - \left(b + \beta_r\right) + T \frac{\partial}{\partial T} - \left(b + \beta_r\right)V \frac{\partial}{\partial V}\right]\left(E + E_r\right) \tag{51}$$

$$- 3\left[1 + \gamma + \delta_r + V \frac{\partial}{\partial V} - \left(\gamma + \delta_r\right)T \frac{\partial}{\partial T}\right]\left(P + P_r\right)$$

$$= \left[1 - \left(b^a + \beta_r^a\right) + T \frac{\partial}{\partial T} - \left(b^a + \beta_r^a\right) V \frac{\partial}{\partial V}\right]\left(E^a + E_r^a\right)$$

Subtracting equation (9) from equation (51) and keeping only first order terms yields the following first order radiation equation

$$\left(1 - b + T \frac{\partial}{\partial T} - bV \frac{\partial}{\partial V}\right) E_r - \beta_r\left(T \frac{\partial P}{\partial T} - P\right) \tag{52}$$

$$- 3\left[\left(1 + \gamma + V \frac{\partial}{\partial V} - \gamma T \frac{\partial}{\partial T}\right) P_r - \delta_r\left(T \frac{\partial P}{\partial T} - P\right)\right]$$

$$= \left(1 - b^a + T \frac{\partial}{\partial T} - b^a V \frac{\partial}{\partial V}\right) E_r^a - \beta_r^a\left(T \frac{\partial P^a}{\partial T} - P^a\right)$$

351

where the following standard thermodynamic relationship was used

$$\frac{\partial U}{\partial V} = E + V \frac{\partial E}{\partial V} = T \frac{\partial P}{\partial T} - P \tag{53}$$

Equation (52) can also be written as

$$\frac{\partial}{\partial T} \left(TU_r \right) - bV \frac{\partial U_r}{\partial V} - \beta_r V \frac{\partial U}{\partial V} \tag{54}$$

$$- 3V \left[\frac{\partial}{\partial V} \left(VP_r \right) - \gamma \frac{\partial U_r}{\partial V} - \delta_r \frac{\partial U}{\partial V} \right]$$

$$= \frac{\partial}{\partial T} \left(TU_r^a \right) - b^a V \frac{\partial U_r^a}{\partial V} - \beta_r^a V \frac{\partial U^a}{\partial V} .$$

where $U_r = VE_r$ = relativistic radiation internal energy, and $U_r^a = VE_r^a$ = non-relativistic radiation internal energy. Equation (52) or equation (54) can serve as the basic first order relativistic thermodynamic equation governing radiation in a thermal medium. Equations (52) and (54) are completely general and can be used to derive the radiation equations for real gases. To do this it is first necessary to develop a nonrelativistic theory of mechanical radiation in real gases so that the terms on the right hand side of equations (52) or (54) can be evaluated.

4. NONRELATIVISTIC THEORY OF WAVES IN REAL GASES. A simple nonlinear nonrelativistic calculation of the radiation energy density and pressure for waves in real gases is presented that will allow the calculation of the nonrelativistic amplitude of the waves. The nonrelativistic radiation pressure and energy density are written in a virial form analogous to the ground state equations (10) and (11) as follows

$$P_r^a = \Gamma_{ro}^a nR_r^a T + n^2 R^a TB_r^a(T) + n^3 R^a TC_r^a(T) + \cdots \tag{55}$$

$$E_r^a = nR_r^a T - n^2 R^a T^2 \frac{\partial B_r^a}{\partial T} - \frac{1}{2} n^3 R^a T^2 \frac{\partial C_r^a}{\partial T} - \cdots \tag{56}$$

where $\Gamma_{ro}^a = 1/3$ = diffuse radiation factor for an ideal gas, and where the nonrelativistic radiation coefficients R_r^a , $B_r^a(T)$, and $C_r^a(T)$ are to be determined from a simple model that describes the vibrations in a real gas.

352

The form of the energy density in equation (56) follows from equation (55) by the requirement that

$$\frac{\partial U_r^a}{\partial V} = T \frac{\partial P_r^a}{\partial T} - P_r^a = E_r^a - n \frac{\partial E_r^a}{\partial n} \qquad (57)$$

The functions δ_r^a, β_r^a, γ_r^a, and Γ_r^a that appear in the right hand side of the wave equation (52) can be calculated in terms of the nonrelativistic radiation virial coefficients from equations (55) and (56) by using equations (49) and (50) and the nonrelativistic analogs of equations (43) and (44).

The nonrelativistic vibrations in a mechanical medium have an energy density given by[17]

$$E_r^a = \frac{1}{4} k_a^2 A_a^2 K_T^a \qquad (58)$$

where k_a = nonrelativistic wave number, and A_a = nonrelativistic wave amplitude. The wave number and wave amplitude that appear in equation (58) are also expected to have a virial expansion of the form

$$k_a^2 A_a^2 = k_o^2 A_o^2 \left[1 + n\alpha_1^a(T) + n^2\alpha_2^a(T) + \cdots \right] \qquad (59)$$

where α_1^a and α_2^a are unknown functions of temperature that are to be determined, and where k_o and A_o = known wave number and wave amplitude respectively associated with waves in an ideal gas. Combining equations (12), (58) and (59) gives

$$E_r^a = \frac{1}{4} k_o^2 A_o^2 nR^a T \left[1 + n\left(2B^a + \alpha_1^a\right) + n^2\left(3C^a + 2\alpha_1^a B^a + \alpha_2^a\right) + \cdots \right] \qquad (60)$$

Comparing equations (56) and (60) gives

$$R_r^a = \frac{1}{4} k_o^2 A_o^2 R^a \qquad (61)$$

$$-T \frac{\partial B_r^a}{\partial T} = \frac{1}{4} k_o^2 A_o^2 \left(2B^a + \alpha_1^a\right) \qquad (62)$$

353

$$-T \frac{\partial C_r^a}{\partial T} = \frac{1}{2} k_o^2 A_o^2 \left(3C^a + 2\alpha_1^a B^a + \alpha_2^a \right) \tag{63}$$

Equation (61) immediately determines the value of the radiation coefficient R_r^a, but further equations in addition to equations (62) and (63) are needed to determine the radiation virial coefficients B_r^a and C_r^a. This is so because the functions α_1^a and α_2^a are also unknown and need to be determined.

The additional equations needed in conjunction with equations (62) and (63) are those involving the nonrelativistic diffuse radiation factor Γ_r^a defined by

$$P_r^a = \Gamma_r^a E_r^a \tag{64}$$

Combining equations (55), (56) and (64) gives the following expression for Γ_r^a

$$\Gamma_r^a = \Gamma_{ro}^a + n\Gamma_{r1}^a + n^2 \Gamma_{r2}^a + \cdots \tag{65}$$

where as before $\Gamma_{ro}^a = 1/3$ and

$$\Gamma_{r1}^a = \frac{R^a}{R_r^a} \left[B_r^a + \Gamma_{ro}^a T \frac{\partial B_r^a}{\partial T} \right] \tag{66}$$

$$\Gamma_{r2}^a = \frac{R^a}{R_r^a} \left[C_r^a + \frac{1}{2} \Gamma_{ro}^a T \frac{\partial C_r^a}{\partial T} + \Gamma_{r1}^a T \frac{\partial B_r^a}{\partial T} \right] \tag{67}$$

where R^a/R_r^a is given by equation (61). But it is well known that the general expression for the diffuse radiation factor is[17]

$$\Gamma_r^a = \frac{1}{3} + \frac{n}{W^a} \frac{dW^a}{dn} \tag{68}$$

where W^a = phase velocity of mechanical waves in a thermodynamic medium. The phase velocity of waves in a general thermodynamic medium is given by[4]

354

$$\left(\frac{W^a}{c}\right)^2 = \frac{K_T^a + \gamma^a T \frac{\partial P^a}{\partial T}}{\Sigma^a + \gamma^a \theta^a} \tag{69}$$

where

$$\Sigma^a = E^a + P^a + K_T^a - T \frac{\partial P^a}{\partial T} \tag{70}$$

$$\theta^a = T \frac{\partial E^a}{\partial T} + T \frac{\partial P^a}{\partial T} \tag{71}$$

Thus in general $W^a = W^a(n,T)$ and substitution of equation (69) into equation (68) and expanding in terms of the ground state virial coefficients automatically determines the expansion coefficients for the diffuse radiation factor given in equation (65). Therefore it will be assumed that Γ_{ro}^a, Γ_{r1}^a, Γ_{r2}^a, and so on, can be obtained from the sound speed and are known functions of temperature through the ground state virial expansion coefficients. Then equations (66) and (67) can be integrated to obtain the nonrelativistic radiation virial coefficients $B_r^a(T)$ and $C_r^a(T)$. Finally equations (62) and (63) can be used to calculate α_1^a and α_2^a as follows

$$\alpha_1^a(T) = -\frac{4}{k_o^2 A_o^2} T \frac{\partial B_r^a}{\partial T} - 2B^a \tag{72}$$

$$\alpha_2^a(T) = -\frac{2}{k_o^a A_o^2} T \frac{\partial C_r^a}{\partial T} - 3C^a - 2\alpha_1^a B^a \tag{73}$$

Then $k_a A_a$ can be determined from equation (59). It will be assumed that $k_a = k_o$ and therefore equation (59) gives the nonrelativistic wave amplitude.

5. SOLUTION OF THE WAVE EQUATION FOR REAL GASES. The solution of the radiation equation (52) for the real gases can be most easily obtained from equations (14), (15) and (23) through (26) that describe the relativistic ground state solution of equation (1) for the real gases. The relativistic expressions for the radiation pressure and energy density are written in a form analogous to equations (55) and (56) as follows

355

$$P_r = \Gamma_{ro} nR_r T + n^2 RTB_r(T) + n^3 RTC_r(T) + \cdots \tag{74}$$

$$E_r = nR_r T - n^2 RT^2 \frac{\partial B_r}{\partial T} - \frac{1}{2} n^3 RT^2 \frac{\partial C_r}{\partial T} - \cdots \tag{75}$$

where the relativistic radiation parameters Γ_{ro} and R_r, and the relativistic radiation virial coefficients B_r and C_r, are to be determined from the solution of the radiation equation (52). The functions δ_r, β_r, γ_r, and Γ_r that appear in equation (52) can be calculated from equations (74) and (75) by using equations (42), (43), (44), and (46).

The solution of the radiation equation (52) for the real gases can be immediately obtained from the ground state solution of equation (1), as given by equations (23) through (26) for the real gases, by a simple perturbation method applied to this solution. Thus when mechanical radiation is present in a real gas, equations (23) through (26) become

$$R + R_r = R^a + R_r^a \tag{76}$$

$$B(T) + B_r(T) = B^a(T) + B_r^a(T) \tag{77}$$

$$C(T) + C_r(T) = C^a(T) + C_r^a(T) - 3\left[B^a(T) + B_r^a(T)\right]^2 \ln\left(\psi^a + \psi_r^a\right) \tag{78}$$

Subtracting equations (23) through (25) from equations (76) through (78) respectively and keeping only first order terms yields

$$R_r = R_r^a \tag{79}$$

$$B_r = B_r^a \tag{80}$$

$$C_r = C_r^a - 3\left[2B^a B_r^a + (B_r^a)^2\right] \ln \psi^a - 3(B^a + B_r^a)^2 \ln\left(1 + \frac{\psi_r^a}{\psi^a}\right) \tag{81}$$

$$C_r \approx C_r^a - 6B^a B_r^a \ln \psi^a - 3(B^a)^2 \psi_r^a / \psi^a \tag{81a}$$

where the following first order approximation has been used

356

$$\ln\left(\psi^a + \psi_r^a\right) = \ln\,\psi^a + \ln\left(1 + \frac{\psi_r^a}{\psi^a}\right) \approx \ln\,\psi^a + \frac{\psi_r^a}{\psi^a} \tag{82}$$

to obtain the result in equation (81a). The small radiation term ψ_r^a that occurs in equation (81a) is obtained from the defining equation (26) as follows

$$\psi^a + \psi_r^a = \frac{T}{T_R}\left|\,\frac{B^a(T) + B_r^a(T)}{B^a(T_R) + B_r^a(T_R)}\,\right|^{2/3} \tag{83}$$

Expanding the denominator in equation (83), and subtracting equation (26), and finally dividing by ψ^a yields the following first order approximation

$$\frac{\psi_r^a}{\psi^a} \approx \frac{2}{3}\left[\frac{B_r^a(T)}{B^a(T)} - \frac{B_r^a(T_R)}{B^a(T_R)}\right] \tag{84}$$

$$\approx \frac{2}{3}\left[\frac{B_r^a(T)}{B^a(T)} - \frac{B_r^a(T_{CR})}{B^a(T_{CR})}\right]$$

Equations (79) through (81) give the relativistic radiation virial coefficients in terms of the corresponding nonrelativistic radiation virial coefficients and in terms of the second order ground state virial coefficient $B^a(T)$. Note that at the Boyle temperature T_B , at which $B^a(T_B) = 0$,

it follows from equation (81a) that $C_r(T_B) = C_r^a(T_B)$. Also note that at the relativity temperature T_R (or at the conjugate relativity temperature T_{CR}) it follows from equation (26) that $\psi^a = 1$, and from equation (84) that $\psi_r^a = 0$, so that $C_r(T_R) = C_r^a(T_R)$ and $C_r(T_{CR}) = C_r^a(T_{CR})$. Therefore any experimental test that is conducted to determine the difference between $C_r(T)$ and $C_r^a(T)$ should exclude the temperature regions around T_B , T_R and T_{CR} . A similar result is already known for the ground state third virial coefficient.[4]

6. RELATIVISTIC WAVE AMPLITUDE AND PHASE VELOCITY. Relativistic effects on waves in real gases will manifest themselves in the amplitude and dispersive properties of the waves. Therefore it is important to be able to calculate the relativistic amplitude and phase velocity of waves in real gases and to compare them with their corresponding nonrelativistic values. The relativistic energy density for mechanical waves in a real gas is written in analogy to equation (56) as[17]

357

$$E_r = \frac{1}{4} k^2 A^2 K_T \tag{85}$$

where k = relativistic wave number, A = relativistic amplitude, and where K_T is given by equation (16). In a form similar to that of equation (59), the product $k^2 A^2$ is written as

$$k^2 A^2 = k_o^2 A_o^2 \left[1 + n\alpha_1(T) + n^2\alpha_2(T) + \cdots \right] \tag{86}$$

where the relativistic functions $\alpha_1(T)$ and $\alpha_2(T)$ need to be determined. Combining equations (16), (75), (85), and (86) gives

$$R_r = \frac{1}{4} k_o^2 A_o^2 R = R_r^a \tag{87}$$

$$-T \frac{\partial B_r}{\partial T} = \frac{1}{4} k_o^2 A_o^2 (2B + \alpha_1) \tag{88}$$

$$-T \frac{\partial C_r}{\partial T} = \frac{1}{2} k_o^2 A_o^2 (3C + 2\alpha_1 B + \alpha_2) \tag{89}$$

Because $B(T) = B^a(T)$ and $B_r(T) = B_r^a(T)$ it follows from equations (62) and (88) that $\alpha_1 = \alpha_1^a$. The value of α_2 is obtained from equation (89) to be

$$\alpha_2 = - \frac{2}{k_o^2 A_o^2} T \frac{\partial C_r}{\partial T} - 3C - 2\alpha_1^a B^a \tag{90}$$

where C_r is given by equation (81) and C is given by equation (25). In this way the relativistic expression for $k^2 A^2$ given by equation (86) can be calculated in terms of nonrelativistic quantities. Since $\alpha_1 = \alpha_1^a$, it is clear that relativistic effects affect only the second order and higher terms in equation (86). Essentially this is due to the fact that only the third and higher virial coefficients of the ground state are affected by relativity as shown in equations (24) and (25).

The relativistic phase velocity can be obtained by first noting that the relativistic diffuse radiation factor is obtained from equations (44), (74), and (75) to be

358

$$\Gamma_r = \Gamma_{ro} + n\Gamma_{r1} + n^2\Gamma_{r2} + \cdots \tag{91}$$

where

$$\Gamma_{ro} = \Gamma_{ro}^a = 1/3 \tag{92}$$

$$\Gamma_{r1} = \Gamma_{r1}^a \tag{93}$$

$$\Gamma_{r2} = \frac{R}{R_r}\left[C_r + \frac{1}{2}\Gamma_{ro}T\frac{\partial C_r}{\partial T} + \Gamma_{r1}T\frac{\partial B_r}{\partial T}\right] \tag{94}$$

where C_r is given by equation (81). Therefore $\Gamma_r(n,T)$ can be evaluated in terms of nonrelativistic quantities. The relativistic sound speed can then be calculated by solving the following equation

$$\Gamma_r(n,T) = \frac{1}{3} + \frac{n}{W}\frac{dW}{dn} \tag{95}$$

or

$$\frac{W}{c} = \exp\left[- \int_n^\infty \left(\Gamma_r - \frac{1}{3}\right)\frac{dn}{n}\right] \tag{96}$$

where c = light speed.

7. GRAVITATIONAL WAVES IN REAL GASES. It has been suggested that real gases can possibly be used in a gravity wave detector.[12] This is possible because the relativity temperature parameter T_R that occurs in the state equation of relativistic real gases is a measure of the interaction of the real gases with the vacuum state, and gravity waves are oscillations of the vacuum, i.e., waves of curvature in spacetime. Gravity waves are shear-like in nature and are not expected to directly change the volume, pressure, or temperature of a gas, liquid or solid. Thus in the case of the Weber bar design for a gravity wave detector, only the surface shear strain is attempted to be measured, but no success has been reported.[6-11]

The interactions of real gases are of dipole-dipole, dipole-quadrupole, and quadrupole-quadrupole types.[13] These interactions depend on the separation and shape of the molecules through their dipole, quadrupole, and

higher moments.[13] The values of T_R and T_{crit} depend on these multipole moments as both temperatures are species dependent.[4] The tidal nature of gravity waves will alter the multipole moments across a volume of gas, and will produce a gradient of T_R across the volume of gas in a detector. Gravity wave detector calculations must be done in conjunction with the relativistic state equations of the materials used in a detector. Real gases and liquids exhibit a critical point, and the critical temperature is related to the relativity temperature by equation (30). Solids, on the other hand, do not have a parameter akin to T_R in their relativistic state equations.[12] Gases and liquids are expected to be sensitive to gravity waves while solids are not expected to show any response.

The values of the relativity temperature T_R and the critical temperature T_{it} are expected to vary across one volume of a gaseous gravitational wave detector due to the tidal effects of gravity waves. Heat exchange in the detector gas will tend to produce a uniform change in temperature. The tidal effects of gravitation can be described by the difference between the metric $g_{\mu\nu}$ for gravitational waves and the Minkowski metric $g^o_{\mu\nu} = (1, 1, 1, -1)$ which is written as[18]

$$h_{\mu\nu} = g_{\mu\nu} - g^o_{\mu\nu} \equiv h \qquad (97)$$

where the values of the small dimensionless number h give a measure of the strength of gravitational radiation at the detector.

The gravitational potential that is associated with this weak gravitational field is $\chi = hc^2$. In the presence of a gravitational field the energy of a body is altered by the following quasi-static factor[19]

$$\left(1 + \frac{2\chi}{c^2}\right)^{1/2} \qquad (98)$$

so that the effects of a gravitational wave on the relativity temperature is to give it the value

$$T_{RG} = \left(1 + \frac{2\chi}{c^2}\right)^{1/2} T_R \qquad (99)$$

$$= (1 + 2h)^{1/2} T_R$$

$$\approx (1 + h) T_R$$

where T_{RG} = value of relativity temperature in the presence of gravity waves. The change in the value of T_R due to ambient gravity waves is therefore[12]

$$\delta T_R = h T_R \tag{100}$$

A similar analysis holds for the conjugate relativity temperature T_{CR}. The order of magnitude change in the value of the relativity temperature depends on the value of h at the detector.

Many studies have been done on the relative strengths of possible astronomical sources of gravity waves.[9-11] These sources include pulsars $10^{-27} < h < 10^{-24}$, supernovae $10^{-22} < h < 10^{-19}$, and binary stars $h < 10^{-21}$. It is possible that the galactic center radiates gravity waves with $h < 10^{-16}$. Taking $T_R \sim 100°K$ gives $10^{-25} < \delta T_R < 10^{-14}$ as a likely range for the change in the relativity temperature of a gas due to astronomical sources of gravity waves. The corresponding changes in pressure, temperature, and volume in a gaseous gravitational wave detector will now be calculated.

8. GENERALIZED FORCE ASSOCIATED WITH RELATIVITY TEMPERATURE. The generalized work done during a change of volume and a change of the relativity temperature of the system is given by

$$dW = PdV + S_R dT_R \tag{101}$$

$$= -\frac{PN}{n^2} dn + S_R dT_R$$

where S_R = generalized force associated with T_R. Clearly S_R has the dimensions of an entropy. The generalized force associated with dV is clearly the system pressure P. The generalized forces can be calculated using the Gibbs-Helmholtz equation which states that if a generalized work is written as E dq, where E = generalized force associated with a physical variable q, then[20]

$$\left(\frac{\partial U}{\partial q}\right)_{T,V} = T\left(\frac{\partial E}{\partial T}\right)_{q,V} - E \tag{102}$$

For instance E might be an electric field and q an electric charge. The Gibbs-Helmholtz equations associated with the situation in equation (101) are

$$\left(\frac{\partial U}{\partial V}\right)_{T,T_R} = T\left(\frac{\partial P}{\partial T}\right)_{V,T_R} - P \tag{103}$$

$$\left(\frac{\partial U}{\partial T_R}\right)_{T,V} = T\left(\frac{\partial S_R}{\partial T}\right)_{T_R,V} - S_R \tag{104}$$

Equation (104) can be used to determine the function S_R .

An expression for S_R can easily be obtained from equation (104) by making the substitution $S_R = Ts_R$ because then equation (104) becomes

$$\left(\frac{\partial U}{\partial T_R}\right)_{T,V} = T^2\left(\frac{\partial s_R}{\partial T}\right)_{T_R,V} \tag{105}$$

Combining equation (15) and (18) with equation (105) gives

$$-\frac{1}{2} NRT^2 n^2 \frac{\partial^2 C}{\partial T \partial T_R} = T^2\left(\frac{\partial s_R}{\partial T}\right)_{T_R,V} \tag{106}$$

which reduces immediately to

$$s_R = -\frac{1}{2} NRn^2 \left(\frac{\partial C}{\partial T_R}\right)_T \tag{107}$$

Finally $S_R = Ts_R$ gives

$$S_R = -\frac{1}{2} NRTn^2 \left(\frac{\partial C}{\partial T_R}\right)_T \tag{108}$$

which can be written per unit volume as

$$S_R/V = -\frac{1}{2} RTn^3 \left(\frac{\partial C}{\partial T_R}\right)_T \tag{108A}$$

362

or in molar quantities as

$$\tilde{S}_R = - \frac{1}{2} RTn^2 \left(\frac{\partial C}{\partial T_R} \right)_T \tag{108B}$$

where C = relativistic third virial coefficient. This generalized force (entropy) will be used subsequently to calculate the changes in volume, temperature, and pressure in a gas due to gravity waves.

The derivative in equation (108) can be evaluated by using equations (25) and (26) which give

$$T_R \frac{\partial C}{\partial T_R} = - 3 [B^a(T)]^2 \frac{T_R}{\psi^a} \frac{\partial \psi^a}{\partial T_R} \tag{109}$$

and

$$\frac{T_R}{\psi^a} \frac{\partial \psi^a}{\partial T_R} = - \left[1 + \frac{2}{3} \frac{T_R}{B^a(T_R)} \frac{\partial B^a(T_R)}{\partial T_R} \right] \tag{110}$$

For reference it is noted also that

$$\frac{T}{\psi^a} \frac{\partial \psi^a}{\partial T} = 1 + \frac{2}{3} \frac{T}{B^a(T)} \frac{\partial B^a(T)}{\partial T} \tag{111}$$

Combining equations (108B), (109), and (110) gives the final result as

$$\tilde{S}_R(n,T,T_R) = - \frac{3}{2} R \frac{T}{T_R} n^2 [B^a(T)]^2 \left[1 + \frac{2}{3} \frac{T_R}{B^a(T_R)} \frac{\partial B^a(T_R)}{\partial T_R} \right] \tag{112}$$

The entropy $\tilde{S}_R$ is a purely relativistic quantity that is associated with the variation of T_R and is related to the interaction of the vacuum state with the molecules of a real gas. Equations (23) through (26) and equation (112) represent a relativistic thermodynamic analog of the Casimir effect of quantum electrodynamics.[21]

363

9. ADIABATIC CHANGES OF TEMPERATURE, VOLUME, AND PRESSURE. The first law of thermodynamics for the relativistic real gas can be written as follows

$$dU = dQ - PdV - S_R dT_R \tag{113}$$

$$= dQ + \frac{PN}{n^2} dn - S_R dT_R$$

where $dQ = TdS =$ increment of heat associated with the absorption of gravity waves by a real gas, and $dS =$ corresponding increase in entropy. Because the internal energy is a state function it has a perfect differential which can be written as

$$dU = \left(\frac{\partial U}{\partial T}\right)_{V,T_R} dT + \left(\frac{\partial U}{\partial V}\right)_{T,T_R} dV + \left(\frac{\partial U}{\partial T_R}\right)_{V,T} dT_R \tag{114}$$

Using the Gibbs-Helmholtz equations (103) and (104) brings equation (114) into the following form

$$dU = \left(\frac{\partial U}{\partial T}\right)_{V,T_R} dT + \left[T\left(\frac{\partial P}{\partial T}\right)_{V,T_R} - P\right] dV + \left[T\left(\frac{\partial S_R}{\partial T}\right)_{V,T_R} - S_R\right] dT_R \tag{115}$$

Placing equation (115) into equation (113) gives the following expression for the heat increment

$$dQ = \left(\frac{\partial U}{\partial T}\right)_{V,T_R} dT + T\left(\frac{\partial P}{\partial T}\right)_{V,T_R} dV + T\left(\frac{\partial S_R}{\partial T}\right)_{V,T_R} dT_R \tag{116}$$

The condition for adiabatic processes is given by $dQ = 0$ or

$$C_V dT + T\left(\frac{\partial P}{\partial T}\right)_{V,T_R} dV + T\left(\frac{\partial S_R}{\partial T}\right)_{V,T_R} dT_R = 0 \tag{117}$$

where C_V is given by equations (17) and (20). It will be assumed that gravity wave interactions with the real gases are sufficiently rapid that they

364

can be described as adiabatic processes. The general expression for the change of pressure in a gas due to the passage of a gravity wave will be written as

$$dP = \left(\frac{\partial P}{\partial T}\right)_{V,T_R} dT + \left(\frac{\partial P}{\partial V}\right)_{T,T_R} dV + \left(\frac{\partial P}{\partial T_R}\right)_{T,V} dT_R \tag{118}$$

Combining equations (14), (16), and (118) gives

$$dP = nR[1 + nf_1(T) + n^2 f_2(T) + \cdots] dT \tag{119}$$

$$+ K_T \frac{dn}{n} + RTn^3 \frac{\partial C}{\partial T_R} dT_R$$

where f_1 and f_2 are given by equations (32) and (33) respectively. Several special cases will now be examined.

Using equation (117) allows several interesting adiabatic situations to be considered.

<u>Case a.</u> Adiabatic Change of Temperature at Constant Volume.

For this case equation (117) gives

$$dT\big|_{S,V} = -\frac{T}{\tilde{C}_V}\left(\frac{\partial \tilde{S}_R}{\partial T}\right)_{V,T_R} dT_R \tag{120}$$

Combining equations (108) and (120) gives

$$dT\big|_{S,V} = \frac{Rn^2 CJT}{2\tilde{C}_V T_R} dT_R \tag{121}$$

where the dimensionless quantity J is given by

$$J = \frac{T_R}{C} \frac{\partial C}{\partial T_R} + \frac{TT_R}{C} \frac{\partial^2 C}{\partial T \partial T_R} \tag{122}$$

365

The second derivative that occurs in equation (122) is calculated using equation (109) as follows

$$TT_R \frac{\partial^2 C}{\partial T \partial T_R} = -6B^a(T)T \frac{\partial B^a}{\partial T} \frac{T_R}{\psi^a} \frac{\partial \psi^a}{\partial T_R} \tag{123}$$

The result in equation (121) can be rewritten using equation (34) as follows

$$dT|_{S,V} = \frac{CJn^2 T}{3T_R} \left(1 + g_1 n + g_2 n^2 + \cdots \right) dT_R \tag{124}$$

where g_1 and g_2 are given by equations (35) and (36) respectively. The sign of the temperature change given by equation (124) depends on the sign of the product CJ which is temperature dependent and can be positive or negative according to the value of temperature being considered.

Case b. Adiabatic Change of Volume at Constant Temperature.

By using the definition of the Grüneisen function given in equation (3) it follows from equations (117) and (120) that

$$d\tilde{V}|_{S,T} = -\frac{\tilde{V}}{\gamma \tilde{C}_V} \left(\frac{\partial \tilde{S}_R}{\partial T}\right)_{V,T_R} dT_R \tag{125}$$

$$= \frac{\tilde{V}}{\gamma T} dT|_{S,V}$$

$$= -\frac{dn}{n^2}\bigg|_{S,T}$$

Combining equations (34), (37), (121), and (125) gives

$$d\tilde{V}|_{S,T} = -\frac{dn}{n^2}\bigg|_{S,T} = \frac{RnCJ}{2\gamma \tilde{C}_V T_R} dT_R \tag{126}$$

$$= \frac{nCJ}{2T_R} [1 - f_1 n + (f_1^2 - f_2)n^2 - \cdots] dT_R \tag{127}$$

366

where f_1 and f_2 are given by equations (32) and (33) respectively.

<u>Case c.</u> Adiabatic Change in Pressure at Constant Volume.

Placing equation (124) into equation (119) with dn = 0 gives

$$dP\big|_{S,V} = \frac{RTCn^3}{3T_R} (F_o + F_1 n + F_2 n^2 + \cdots) \, dT_R \tag{128}$$

where

$$F_o = J + \frac{3T_R}{C} \frac{\partial C}{\partial T_R} \tag{129}$$

$$F_1 = \gamma_1 J \tag{130}$$

$$F_2 = \gamma_2 J \tag{131}$$

and where γ_1 and γ_2 are defined in equation (38) and (39) respectively. An equivalent expression for dP can also be written in terms of the Grüneisen parameter as follows

$$dP\big|_{S,V} = \frac{RTCn^3}{T_R} \left[\frac{\gamma J}{2} + \frac{T_R}{C} \frac{\partial C}{\partial T_R} \right] dT_R \tag{132}$$

Substituting the power series expansion for γ given by equation (37) into equation (132) yields the result given in equation (128). Thus $dP \sim n^3$ for low densities.

<u>Case d.</u> Adiabatic Change in Pressure at Constant Temperature.

Combining equation (126) with equation (119) for dT = 0 yields

$$dP\big|_{S,T} = \frac{RTCn^3}{T_R} \left[\frac{T_R}{C} \frac{\partial C}{\partial T_R} - \frac{JK_T}{2n\gamma \check{C}_V T} \right] dT_R \tag{133}$$

Using equations (16), (34), and (35) allows equation (133) to be rewritten as

367

$$dP|_{S,T} = \frac{RTCn^3}{2T_R} (G_0 + G_1 n + G_2 n^2 + \cdots) dT_R \qquad (134)$$

where

$$G_0 = 2 \frac{T_R}{C} \frac{\partial C}{\partial T_R} - J \qquad (135)$$

$$G_1 = J(f_1 - 2B) \qquad (136)$$

$$G_2 = J(f_2 - f_1^2 + 2f_1 B - 3C) \qquad (137)$$

where J is given by equation (122) and f_1 and f_2 are given by equations (32) and (33) respectively. Therefore at low densities $dP \sim n^3$. Because in general $dP/P \sim n^2$ for low densities, the efficiency of a gaseous gravitational wave detector can be improved by increasing the density of the gas in the detector.

Consider now the case of a constant pressure system. From equation (118) it follows that the constant pressure condition is written as

$$\left(\frac{\partial P}{\partial V}\right)_{T,T_R} dV + \left(\frac{\partial P}{\partial T}\right)_{V,T_R} dT + \left(\frac{\partial P}{\partial T_R}\right)_{V,T} dT_R = 0 \qquad (138)$$

Two cases of the constant pressure system are of interest.

Case e. Change in volume at Constant Pressure and Temperature.

From equation (138) and equations (6), (14), and (23) through (26) it follows that

$$d\tilde{V}|_{P,T} = \frac{\tilde{V}}{K_T} \left(\frac{\partial P}{\partial T_R}\right)_{T,V} dT_R \qquad (139)$$

$$= \frac{RTn^2}{K_T} \frac{\partial C}{\partial T_R} dT_R$$

$$= -\frac{2\tilde{S}_R}{K_T} dT_R$$

where $\tilde{S}_R$ is given by equation (108B). Equation (139) can be rewritten using equation (16) as follows

$$d\tilde{V}|_{P,T} = n[1 - 2nB + n^2(4B^2 - 3C) - \cdots] \frac{\partial C}{\partial T_R} dT_R \qquad (140)$$

Case f. Change in Temperature at Constant Pressure and Volume.

From equation (138) and equations (3), (14), and (23) through (26) it follows that

$$dT|_{P,V} = - \frac{\tilde{V}}{\gamma \tilde{C}_V} \left(\frac{\partial P}{\partial T_R} \right)_{T,V} dT_R \qquad (141)$$

$$= - \frac{RTn^2}{\gamma \tilde{C}_V} \frac{\partial C}{\partial T_R} dT_R$$

Using equations (34) and (37) allows equation (141) to be rewritten as

$$dT|_{P,V} = - Tn^2[1 - f_1 n + (f_1^2 - f_2)n^2 - \cdots] \frac{\partial C}{\partial T_R} dT_R \qquad (142)$$

where f_1 and f_2 are given by equations (32) and (33) respectively.

10. CONCLUSION. The description of relativistic wave motion in real gases must include the coupling of the matter and radiation fields with the thermodynamic gauge parameters for matter and radiation. This means that the quantities P, γ, b and P_r, γ_r, b_r are coupled as shown in equation (52). This is true for wave motion in any relativistic physical system. The form of the relativistic third virial coefficient of the ground state of a real gas is affected by the ground state gauge parameters. When mechanical radiation is present in real gases, the third virial coefficient of the radiation itself is correspondingly affected by both the ground state and radiation gauge parameters. Because only the third and higher virial coefficients are affected by the gauge parameters, measurable relativistic effects should be observed only at high pressures such as can occur in nuclear explosions in the atmosphere, during the interaction of directed energy beams with the atmosphere, or in high pressure laboratory experiments. The tidal effects of gravitational radiation are expected to appear in the third and higher virial coefficients of the real gases, and therefore these gases under high pressure can serve as suitable materials for a gravitational wave detector.

369

REFERENCES

1. Quigg, C., <u>Gauge Theories of the Strong, Weak, and Electromagnetic Interactions</u>, Addison-Wesley, New York, 1983.

2. Moriyasu, K., <u>An Elementary Primer for Gauge Theory</u>, World Scientific, Singapore, 1983.

3. Cheng, R. P. and Li, L. F., <u>Gauge Theory of Elementary Particle Physics</u>, Oxford, New York, 1984.

4. Weiss, R. A., <u>Relativistic Thermodynamics</u>, Vols. 1 and 2, Exposition Press, New York, 1976.

5. 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 Report 86-1, May 13-16 1985, p. 717.

6. Weber, J., "Gravitational Radiation", Phys. Rev. Lett., 18, 498, 1967.

7. Papini, G., "Gravitational Radiation and its Detection", Can. J. Phys., 52, 880, 1973.

8. Braginsky, V. B. and Manukin, A. B., <u>Measurement of Weak Forces in Physics Experiments</u>, Chicago Univ. Press, p. 98, 1977.

9. Press, W. H. and Thorne, K. S., Annual Reviews of Astronomy and Astrophysics, 10, 335 (1972).

10. Ostriker, J. P., article in <u>Sources of Gravitational Radiation</u>, edited by L. Smarr (Cambridge Univ. Press, 1979), p. 461.

11. Thorne, K. S., "Gravitational Wave Research: Current Status and Future Prospects", Revs. Mod. Phys., Vol 52, No. 2, Part 1, April 1980.

12. Weiss, R. A., "A Gaseous Gravitational Wave Detector", article in <u>After Einstein</u>, edited by P. Barker and C. G. Shugart, Memphis State University Press, 1981, pp. 103.

13. Hirschfelder, J. O., Curtiss, C. F. and Bird, R. B., <u>Molecular Theory of Gases and Liquids</u>, John Wiley, New York, 1954.

14. Beattie, J. A., "Thermodynamic Properties of Real Gases and Mixtures of Real Gases", article in <u>Thermodynamics and Physics of Matter</u>, edited by F. D. Rossini, Princeton University Press, 1955, pp. 240.

15. Rice, O. K., "Critical Phenomena", article in <u>Thermodynamics and Physics of Matter</u>, edited by F. D. Rossini, Princeton University Press, 1955, pp. 438.

16. Weiss, R. A., "Relativistic Effects on the Critical Point of Gases and Liquids", Bulletin Amer, Phys. Soc., NF 17, 1980.

17. Brillouin, L., <u>Tensors in Mechanics and Elasticity</u>, Academic Press, New York, 1964.

18. Weinberg, S., <u>Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity</u>, John Wiley, New York, 1972.

19. Møller, C., <u>The Theory of Relativity</u>, Oxford, London, 1955.

20. Epstein, P., <u>Textbook of Thermodynamics</u>, John Wiley, New York, 1937.

21. Brevik, I. and Kolbenstvedt, H., "Attractive Casimir Stress on a Thin Spherical Shell", Can. J. Phys., Vol. 63, 1985.

(1)

$T_B^* = 3.42$

$T^* = \frac{k}{\varepsilon} T$

k = Boltzmann constant

ε = Potential parameter of Lennard-Jones potential

(2)

$\tilde{V}_{crit}^* = \tilde{V}(T_{crit})/b_0$

$b_0 = \frac{2}{3} \pi N_A \sigma^3$

N_A = Avogadro's number

σ = Range parameter of Lennard-Jones potential
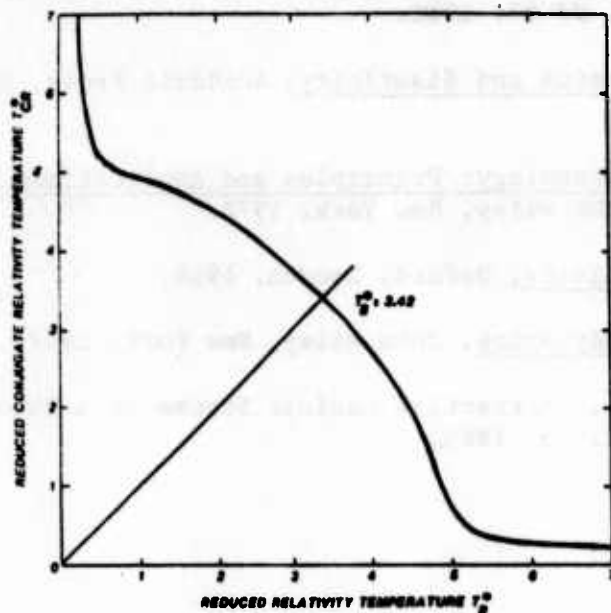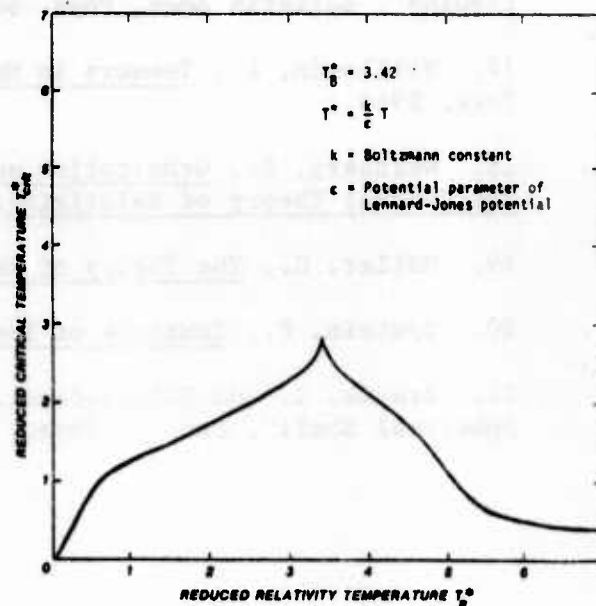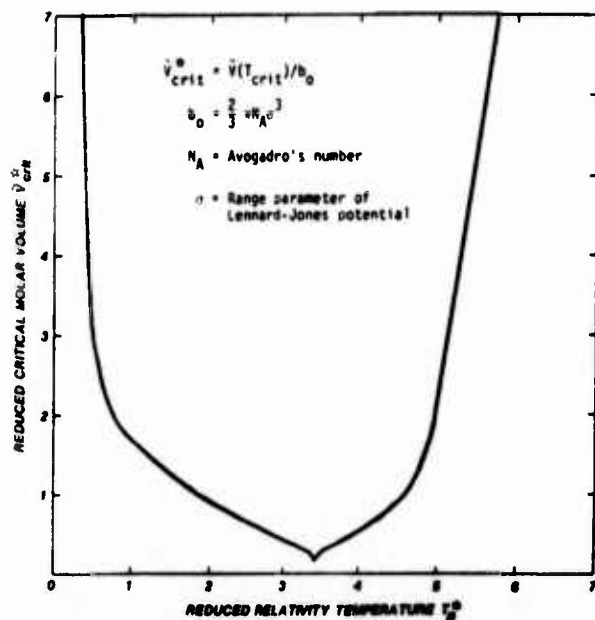
(3)

Figure 1. Relationship between the relativity temperature and the conjugate relativity temperature.

Figure 2. Dependence of reduced critical temperature on the reduced relativity temperature. Note that $T_{crit}^* < T_B^*$ and $B^a(T_{crit}) < 0$.

Figure 3. Dependence of the reduced critical molar volume on the reduced relativity temperature. Note that $\tilde{V}_{crit} = -B(T_{crit})$.

372

HAMILTONIAN DEFORMATIONS OF INTEGRABLE, NONLINEAR FIELD EQUATIONS (WITH APPLICATIONS TO OPTICAL FIBERS)

C. R. Menyuk,[a,b] P. K. A. Wai,[b] H. H. Chen,[b] and Y. C. Lee[b,c]

ABSTRACT. In integrable, nonlinear systems an arbitrarily shaped initial pulse is known to break up into a series of solitons and a dispersive wave component. It has been shown both analytically and numerically that this behavior persists when substantial Hamiltonian deformations, which destroy the system's integrability are present. By contrast, this behavior is usually destroyed by non-Hamiltonian deformations even when they are quite small. Hence, it is usually sufficient to know a deformation's character to immediately determine its effect on solitons. Application of this result to optical fiber communication is discussed.

I. INTEGRABLE EQUATIONS. It may seem odd at first that anything which sounds as esoteric as Hamiltonian deformations could have something useful to tell us about optical fibers. We believe, however, that our results are a nice example of how a physical/mathematical principle when properly understood can lead to important insights into the operation of real-world devices.

Many, if not most, physical systems exhibit turbulent or chaotic behavior in at least some regimes. Such systems are appropriately modeled by equations like the Navier-Stokes equation which has turbulent solutions at high Reynolds numbers and is used to study fluids. In many important cases, however, the physical systems exhibit nice, coherent behavior over a wide range of parameters. That is particularly the case in devices which are useful for something, as opposed to systems which are handed to us by nature, since one usually wants the device to behave in a nice, predictable manner.

Nonlinear field equations which always exhibit coherent behavior include the

[a] Science Applications International Corp., 1710 Goodridge Drive, McLean, VA 22102

[b] Dept. of Physics and Astronomy, Univ. of Maryland, College Park, MD 20742

[c] Center for Nonlinear Studies, Los Alamos Scientific Laboratory, MS-258, Los Alamos, NM 87545

sine-Gordon equation

$$u_{xt} = \sin u,$$

(1)

which has been used to model self-induced transparency [1], the Korteweg-de Vries equation

$$u_t - 6uu_x + u_{xxx} = 0,$$

(2)

which has been used to model water waves in shallow channels [2] and ion-acoustic waves in plasmas [3], and the nonlinear Schrödinger equation

$$iu_t + \frac{1}{2}u_{xx} + |u|^2 u = 0,$$

(3)

which has been used to model Langmuir waves in plasmas [4] and light pulses in optical fibers [5]. We will be discussing this last application in far more detail at a later point.

These equations are often referred to as "integrable." That is to say, they have a number of special properties which the vast majority of field equations do not have. Among the most important of these properties is a spectral transformation which can be considered to be a nonlinear Fourier transform. The spectral transform can be used to solve these special equations just like the usual Fourier transform can be used to solve linear field equations. The transformation procedure is shown schematically in Fig. 1 for the nonlinear Schrödinger equation, assuming that the initial data $u(x, t = 0)$ falls off sufficiently rapidly as $x \to \pm\infty$ [6]. The spectral transformation yields $[r(\xi, 0), \varsigma_j(0), C_j(0)]$. The quantity $r(\xi, 0)$ depends continuously on the variable ξ and is directly analogous to the usual Fourier transform, although it is not identical. Physically, it corresponds to a dispersive wave whose amplitude vanishes as $t \to \infty$. In addition, the spectral transform yields a number $N \geq 0$ of discrete pairs (ς_j, C_j) which have no analogy in the usual Fourier transform. These pairs correspond to solitons, nonlinear wave packets which propagate without dispersing.

Knowing the solution at $t = 0$, it is possible to immediately write down the solution at $t = \tau$. It is [6]

$$r(\xi, \tau) = r(\xi, 0) \exp(2i\xi^2\tau),$$
$$\varsigma_j(\tau) = \varsigma_j(0),$$
$$C_j(\tau) = C_j(0) \exp(2i\varsigma_j^2\tau).$$

(4)

One can use the inverse spectral transform to determine $u(\xi, \tau)$. The significance of the spectral transform is that it allows us to determine $u(\xi, \tau)$ in three steps, shown as solid lines in Fig. 1, no matter what the size of τ. If one were to use the direct route shown as a dashed line in Fig. 1, one would in general need to cut the time

374

$$u_t = \frac{i}{2} u_{xx} + i|u|^2 u$$
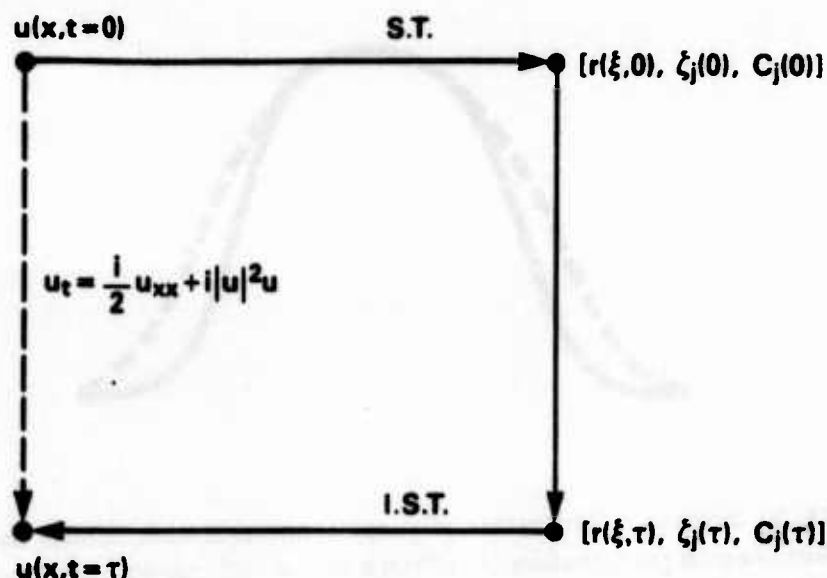
FIGURE 1. Schematic illustration of the way in which the spectral transform and its inverse can be used to solve the nonlinear Schrödinger equation.

axis into a number of pieces proportional to τ and determine the solution iteratively. Thus, there exists some time τ beyond which the indirect approach always wins. While this indirect approach has not been used much to date, there are a number of problems where it would be useful.

The issue that we are concerned with in this presentation is that there are many systems which behave integrably. That is to say, initial data breaks up into a dispersive wave component and a number of solitons (or, more precisely, solitary waves). It is natural to suppose that these systems can be well-modelled by one of the integrable equations. If the actual system were to be perturbed away from the integrable system by an amount of order ϵ, then one might expect that the integrable behavior would only appear for a time of order ϵ^{-1}. On a longer time scale, solitons would be destroyed. This expectation is borne out in practice when the perturbations are dissipative or have an explicit space or time dependence; however, when the perturbations are Hamiltonian and independent of space and time, that is no longer the case. Indeed, the systems appear to act integrably on arbitrarily long time scales. Moreover, they continue to act integrably when the Hamiltonian deviations are so large that they can no longer be referred to as perturbations, but must be considered substantial deformations. Why are systems so rugged under the influence of Hamiltonian deformations, and what are the implications for practical devices like optical fibers? We will be addressing these issues in the following sections.

While this sort of behavior can be seen in a large number of real physical

375
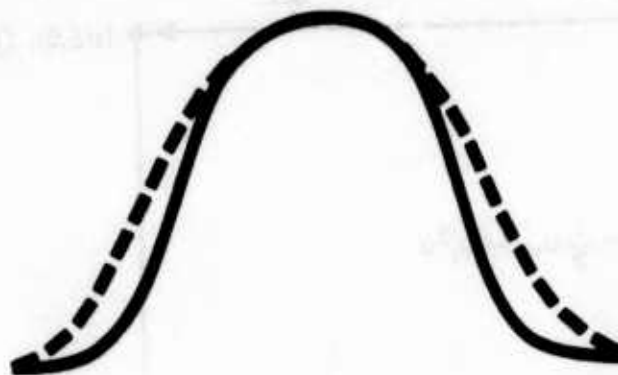
FIGURE 2. Schematic illustration of shape renormalization. Under the influence of a Hamiltonian perturbation, a soliton's shape will change from that shown as a solid line to that shown as a dashed line.

systems, we mention here two numerical examples closely related to the nonlinear Schrödinger equation

$$iu_t + \frac{1}{2}u_{xx} + |u|^2 u = -\left\{1 - |u|^2 - \exp(-|u|^2)\right\}u, \tag{5}$$

which has been used to model Langmuir waves in plasmas [7] and

$$iu_t + \frac{1}{2}u_{xx} + |u|^2 u = i\beta u_{xxx}, \tag{6}$$

which has been used to study light pulses in optical fibers near the zero-dispersion point [8,9]. The deformations in both Eqs. (5) and (6) are Hamiltonian. We emphasize numerical results because in simulated systems the effect of dissipation can be completely eliminated which can never be the case in real systems. In numerical solutions of Eq. (5), initial data are seen to break into a soliton and dispersive waves when $|u|$ is as large as 2, so that the term on the right is making a large contribution. Similar results are found when Eq. (6) is solved with β arbitrarily large. Clearly, then, the right-hand side can be a large deformation indeed! It should be noted that while solitons continue to exist, their shapes and frequency shifts are observed to change from what is predicted by the nonlinear Schrödinger equation, as shown schematically in Fig. 2, something which must be explained theoretically.

II. HAMILTONIAN SYSTEMS. In order to demonstrate that Eqs. (5) and (6) are Hamiltonian, it is sufficient to show that they can be derived from a

Hamiltonian functional. In the case of Eq. (6), this functional is

$$H = -\frac{i}{2} \int_{-\infty}^{\infty} dx \left[u_x u_x^* - |u|^4 + i\beta(u_x u_{xx}^* - u_{xx} u_x^*) \right].$$ (7)

Letting $q \equiv u$ and $p \equiv u^*$, one can show

$$\dot{q} = \frac{\delta H}{\delta p}, \qquad \dot{p} = -\frac{\delta H}{\delta q},$$ (8)

as is appropriate for Hamiltonian systems, where the derivatives $\delta x/\delta y$ are functional derivatives. A similar result can be obtained for Eq. (5). Such systems are often referred to as infinite-dimensional Hamiltonian systems because each point in x can be considered a separate degree-of-freedom. When one states that a Hamiltonian system is integrable, one generally means that a canonical transformation exists which yields a Hamiltonian independent of the new coordinates, depending only on the new momenta. This point of view seems different from that of the previous section where we said that the nonlinear Schrödinger equation could be solved by making a spectral transformation; in fact, these two points of view are equivalent. The spectral transformation turns out to be a canonical transformation which yields a Hamiltonian only depending on the momenta.

Before demonstrating this point explicitly, it is useful to turn to a simpler example to explain how these canonical transformations work. They are important because when integrable field equations with Hamiltonian perturbations are considered, it is possible to find an infinite series of canonical transformations which eliminates order-by-order the dependence on the coordinates. This result explains qualitatively why integrable behavior is rugged under the influence of Hamiltonian deformations. (At least when the deformations are small!)

The example we will consider is a simple, finite-dimensional system

$$H = \sum_i \frac{\omega_i}{2}(p_i^2 + q_i^2).$$ (9)

The canonical transformation $(p_i, q_i) \rightarrow (P_i, Q_i)$, where

$$p_i = (2P_i)^{1/2} \cos Q_i, \qquad q_i = (2P_i)^{1/2} \sin Q_i,$$ (10)

reduces the Hamiltonian to the desired form

$$H = \sum_i \omega_i P_i,$$ (11)

377

which depends only on the momenta. As a consequence, the momenta are constant in time, while the coordinates vary linearly. Writing the equations of motion,

$$\dot{P}_i = 0, \qquad \dot{Q}_i = \omega_i, \tag{12}$$

we obtain,

$$P_i = P_{i,0}, \qquad Q_i = Q_{i,0} + \omega_i t, \tag{13}$$

where $P_{i,0}$ and $Q_{i,0}$ are constants of integration. In similar fashion, if we make the transformation $u \to [P(\xi), Q(\xi), P_j, Q_j]$, where, in terms of the spectral data,

$$P(\xi) = \frac{i}{\pi} \ln[1 + |r(\xi)|^2], \qquad Q(\xi) = \arg r(\xi),$$
$$P_j = 2i\varsigma_j, \qquad Q_j = -\ln C_j, \tag{14}$$

we find that the transformed Hamiltonian becomes [6]

$$H = \int_{-\infty}^{\infty} d\xi \left[2\xi^2 P(\xi)\right] + \frac{i}{6} \sum_j P_j^3, \tag{15}$$

which only depends on the momenta. Hence, just as in the previous case, the momenta are constant in time while the coordinates vary linearly.

Suppose now that we perturb the finite-dimensional system by adding cubic terms to the Hamiltonian,

$$H = \sum_i \frac{\omega_i}{2}(p_i^2 + q_i^2) + ap_i^3 + bp_1^2 q_1 + \cdots \tag{16}$$

In the limit where p_i and q_i are small, this perturbation only makes a small contribution to the Hamiltonian. As long as all the ω_i are incommensurable, it is possible to find a canonical transformation, using the Lie transform method or the Poincaré-von Zeipel method, which eliminates the cubic terms at the expense of introducing fourth and higher order terms [10], *i.e.* there exists a transformation $[p_i, q_i] \to [\tilde{p}_i, \tilde{q}_i]$, such that our Hamiltonian becomes

$$H = \sum_i \frac{\omega_i}{2}(\tilde{p}_i^2 + \tilde{q}_i^2) + \tilde{a}\tilde{p}_1^4 + \cdots \tag{17}$$

We can then eliminate the fourth order terms by making another, analogous transformation and continue in this fashion order-by-order. Physically, this series of transformations is possible because when q_i and p_i are sufficiently small, the effect of the cubic perturbations, for the vast majority of initial conditions, is to deform the orbit of the pair (p_i, q_i) without destroying its neutral stability. By contrast, a
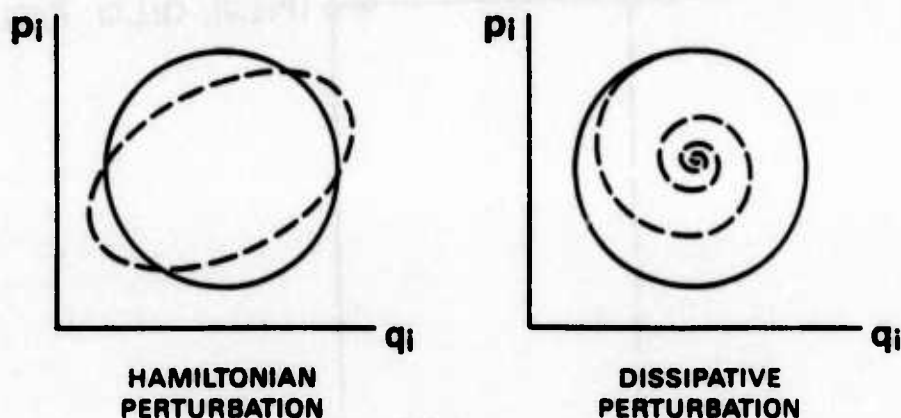
FIGURE 3. Effect of Hamiltonian and non-Hamiltonian perturbations. A Hamiltonian perturbation slightly deforms the trajectory, but it remains neutrally stable. A dissipative perturbation, no matter how small, leads to a spiral trajectory which ultimately falls into the origin.

dissipative perturbation, no matter how small, will lead to a fundamental change in orbit topology as shown qualitatively in Fig. 3.

A similar series of transformations exists for Hamiltonian perturbations of integrable field equations. We recall that the original transformation $u \rightarrow [P(\xi), Q(\xi), P_i, Q_i]$ yields quantities which evolve linearly in time when u_t is given by the nonlinear Schrödinger equation. That is no longer the case once the equations are perturbed. However, at any given order, the canonical transformations yield a new set of quantities $[\tilde{P}(\xi), \tilde{Q}(\xi), \tilde{P}_i, \tilde{Q}_i]$ which evolve linearly in time *through the order to which we are working*, i.e.

$$\tilde{P}(\xi) = \tilde{P}_0(\xi), \qquad \tilde{Q}(\xi) = \tilde{Q}_0(\xi) + \Omega(\xi)t,$$
$$\tilde{P}_j = \tilde{P}_{j,0}, \qquad \tilde{Q}_j = \tilde{Q}_{j,0} + \Omega_j t, \qquad (18)$$

where $\tilde{P}_0(\xi)$, $\tilde{Q}_0(\xi)$, $\tilde{P}_{j,0}$, $\tilde{Q}_{j,0}$, $\Omega(\xi)$, and Ω_j are all constant in time. Hence, just as in the integrable case, it is possible to integrate the equation in a fixed number of steps, as shown schematically in Fig. 4, independent of the length of time τ over which one wishes to determine the solution. Why then are these perturbed equations not also considered integrable? The reason is that in general this series of transformations is only convergent for special choices of the initial conditions; otherwise, the series is merely asymptotic, and only a finite number of transformations can be usefully made.

At every order of the transformation, one finds that the topology of the solution is unchanged; it still consists of a number of solitons which do not change in

u(x,t=0)   S.T.   C.T. $[\widetilde{P}(\xi,0),\ \widetilde{Q}(\xi,0),\ \widetilde{P}_j(0),\ \widetilde{Q}_j(0)]$

I.S.T.   I.C.T. $[\widetilde{P}(\xi,\tau),\ \widetilde{Q}(\xi,\tau),\ \widetilde{P}_j(\tau),\ \widetilde{Q}_j(\tau)]$

u(x,t = τ)
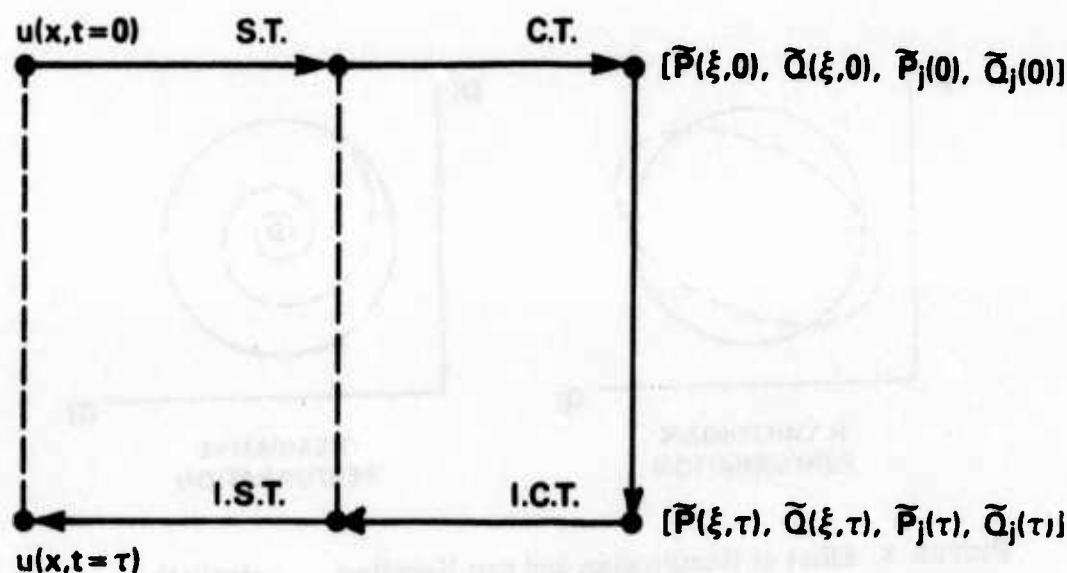
FIGURE 4. Schematic illustration of the integration procedure for the perturbed nonlinear Schrödinger equation when the perturbations are Hamiltonian. One first makes a spectral transformation followed by a series of canonical transformations to arrive at variables which evolve linearly in time. Having calculated the new variables at the time τ, one reverses the original sequence of transformations to determine u.

time when well-separated and a dispersive wave component [11,12]. Hamiltonian perturbations lead to no fundamental changes in the structure of the solution, in contrast to dissipative perturbations.

It should be noted that the results just described have only been demonstrated in detail when the underlying, integrable system is the Korteweg-de Vries equation [11,12], although the nature of the derivation makes it seem clear that similar results will hold when the underlying system is the nonlinear Schrödinger equation or any of a set of similar field equations. We are presently studying these systems.

III. OPTICAL FIBERS. Optical fibers consist of a glass core surrounded by a glass cladding; the index of refraction in the core is slightly higher than in the cladding, implying that waves will propagate. Essentially, they are trapped by total internal reflection [13].

If the core is sufficiently small, $\simeq 8\mu$m in diameter or less, then only a single mode, the HE_{11} mode, propagates, eliminating intermodal dispersion. Nonetheless,
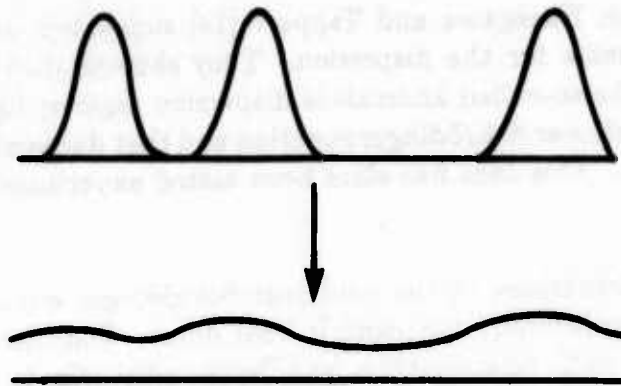
FIGURE 5. The effect of dispersion is illustrated schematically. As pulses propagate along the fiber, they broaden, eventually overlapping.

single mode dispersion remains a serious problem limiting the (bit rate) × (propagation length) values which can be attained in modern-day systems. The way in which dispersion limits the bit rate is shown in Fig. 5. A train of pulses is launched in the fiber. When a pulse is present in a given time slot, it is counted as a 1-bit, and, when it is absent, it is counted as a 0-bit. As the pulses propagate along the fiber, the dispersion causes spreading; after a long length, it is impossible for the detection system to tell whether there is a 1-bit or a 0-bit in any given slot.

For any given length, there is an optimum pulse size which yields the maximum bit rate possible. If the pulse is too narrow initially, then it has a large bandwidth and spreads very quickly due to the dispersion. If the pulse is too large initially, then it stays too large. It is conventional to measure pulse widths in units of time. If we write the initial pulse width as τ_0, then the final pulse width after going through a fiber of length L is

$$\tau = \tau_0 + \frac{\lambda^3}{c^2} n \left| \frac{d^2 n}{d\lambda^2} \right| \frac{L}{\tau_0}, \tag{19}$$

where λ is the light's wavelength, n is the index of refraction in the fiber, and c is the speed of light in a vacuum. The minimum fiber loss rate is 0.2 db/km when $\lambda = 1.55 \mu$m, from which we infer a 20 km propagation length before the signal loss becomes severe [5,13]. From Eq. (19), we then infer a maximum bit rate of 5 Gbit/sec. While this figure is quite large, the bit rates in communication systems have been rising roughly exponentially as a function of time over the last two centuries, with a break to a faster rise after 1950, as shown in Fig. 6. Unless this curve magically bends over in the near future, it is clear that this bit rate will

381

soon be achieved.

Some years ago, Hasegawa and Tappert [14] suggested using the Kerr non-linearity to compensate for the dispersion. They showed that in the wavelength range $\lambda > 1.3\mu$m, the so-called anomalous dispersion regime, light pulses are well-described by the nonlinear Schrödinger equation and that dispersionless propagation is therefore possible. This idea has since been tested experimentally and found to be feasible [15,16].

Significant deformations of the nonlinear Schrödinger equation, both Hamiltonian and non-Hamiltonian, can exist in real fibers. Hamiltonian deformations include cubic dispersion, birefringence, and finite radial effects. Non-Hamiltonian deformations include attenuation and Raman or Brillouin scattering. From the results of the previous sections, we may infer the following: Hamiltonian deformations, even large deformations, will have no adverse effect on the solitons; their shapes may be slightly different from what the nonlinear Schrödinger equation predicts, but they will still exist and propagate. By contrast, non-Hamiltonian deformations are very destructive and must be dealt with in some way. The power of this result is that it is not necessary to do any detailed analysis; it is only necessary to determine the nature of the deformation, and one immediately knows whether it is likely to cause trouble.

In order to verify these theoretical considerations and to determine the maximum deformations which will still allow solitons to propagate in real fibers, our group has in the past year mounted a systematic numerical investigation of all the deformations which can play a major role in optical fibers. We have begun by examining the behavior of pulses which are injected at the zero dispersion point, $\lambda \simeq 1.3\mu$m. At this point, the usual quadratic dispersion goes to zero, unveiling the effect of the cubic dispersion. Using appropriately normalized variables, one then finds

$$iu_s - iu_{rrr} + |u|^2 u = 0, \tag{20}$$

where s represents the length along the fiber and r the time variation in the group velocity frame. Note that we have reversed the roles of space and time from the "standard" roles seen in Eqs. (1-3); we do so because the pulses in fibers are initially specified for all time at a given point in space, rather than the reverse. We can obtain Eq. (20) from Eq. (6) by letting $s = t, r = x/\beta^{1/3}$, and by letting $\beta \to \infty$.

It is of great practical interest to operate as close to the zero dispersion point as possible. Since the dispersion is minimal at this point, the power needed to generate a soliton is also minimal. Indeed, it may be possible to reduce the power requirement to the point where a single laser diode can generate the pulses—a very desirable result indeed! Previous workers had assumed that pulses launched at

382
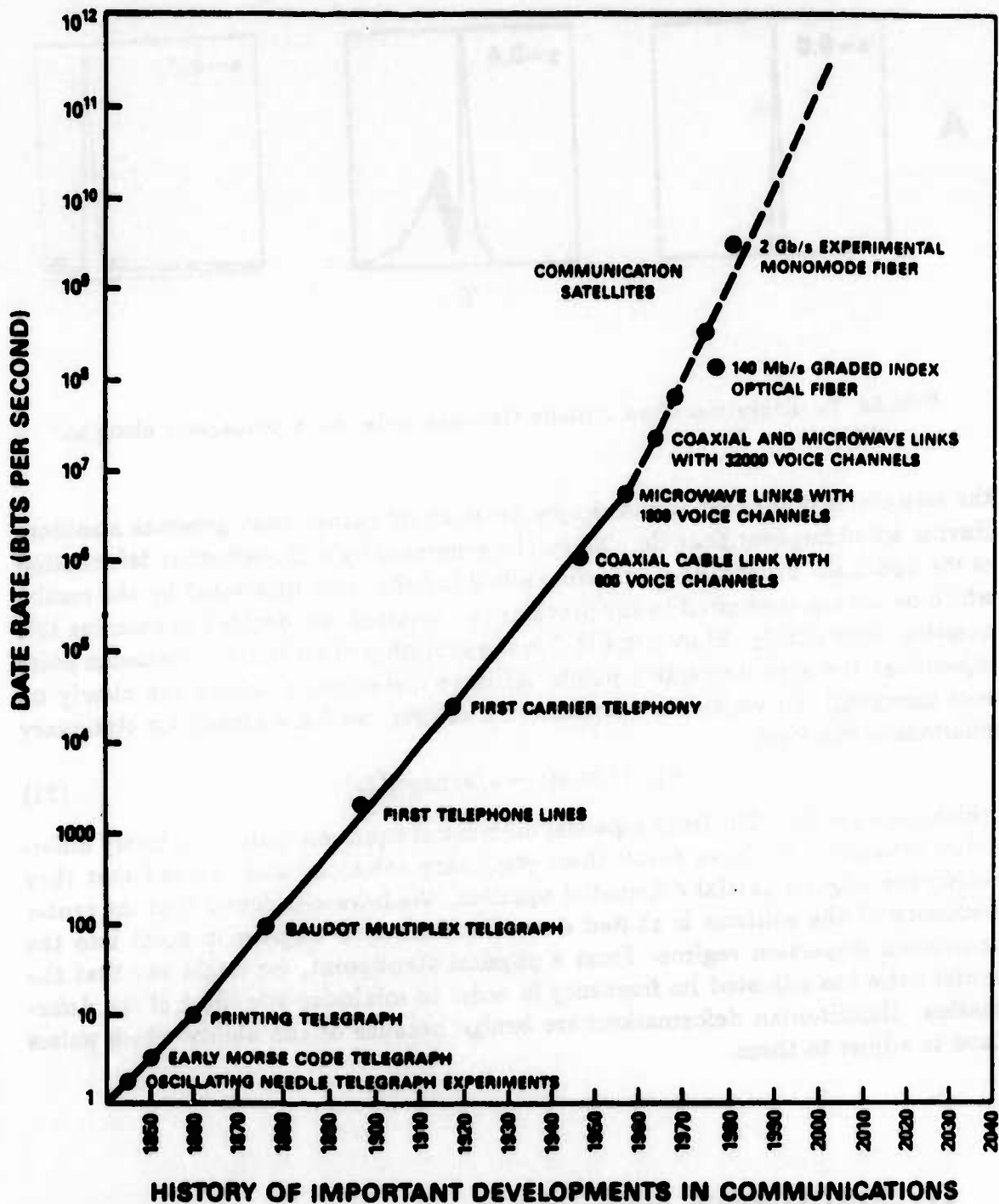
FIGURE 6. Bit rates in communication systems over the last two centuries. Note the break in the curve which occured about thirty years ago, roughly coincident with the invention of the laser.
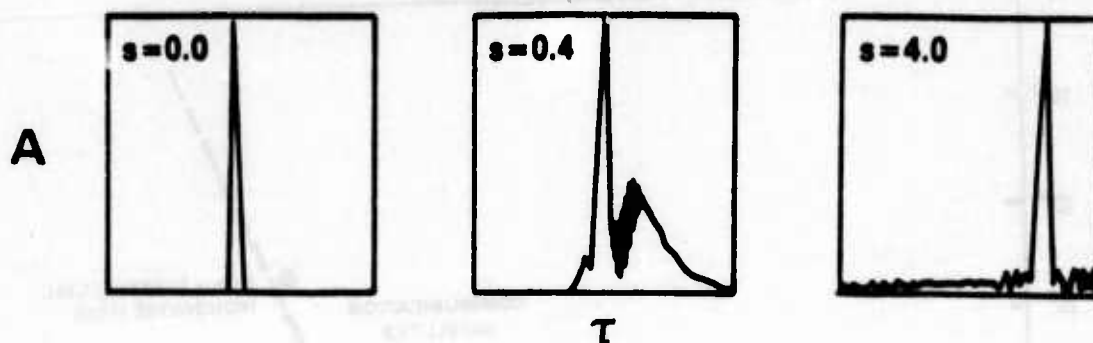
FIGURE 7. Evolution of an initially Gaussian pulse as it propagates along an optical fiber.

the zero dispersion point would simply break apart rather than generate a soliton. Having noted however that Eq. (20) can be generated by a Hamiltonian deformation of the nonlinear Schrödinger equation, albeit infinite, and motivated by the results which have been presented in the previous two sections, we decided to examine this question more closely. Shown in Fig. 7 is the evolution of an initially Gaussian pulse injected at the zero dispersion point. At large distances, a soliton can clearly be seen emerging! To verify the existence of a soliton, we have looked for stationary solutions of the form

$$u(s,\tau) = u(\tau - s/v)\exp(i\Omega s), \tag{21}$$

which converts Eq. (20) from a partial differential equation into an ordinary differential equation. We have found these stationary solutions and checked that they satisfy the original partial differential equation. We have also found that the center frequency of the solitons is shifted down from the zero dispersion point into the anomalous dispersion regime. From a physical standpoint, we might say that the initial pulse has adjusted its frequency in order to minimize the effect of the deformation. Hamiltonian deformations are benign because of the ability which pulses have to adjust to them.

IV. CONCLUSION. Solitons persist in the face of large Hamiltonian deformations while non-Hamiltonian deformations usually destroy them. This result has important technical implications for light propagation in optical fibers. By simply determining whether a deformation is Hamiltonian or non-Hamiltonian, we can immediately tell whether or not it is likely to cause trouble. Since this result is quite general, it is likely to be of importance not only in fibers, but in many other physical

systems as well.

REFERENCES

[1] S. L. McCall and E. L. Hahn, "Self-induced transparency by pulsed coherent light," Phys. Rev. Lett. **18**, 908–911 (1967).

[2] D. J. Korteweg and G. De Vries, "On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves," Philos. Mag. Ser. 5 **39**, 422–443 (1895).

[3] H. Washimi and T. Taniuti, "Propagation of ion acoustic solitary waves of small amplitude," Phys. Rev. Lett. **17**, 996–998 (1966).

[4] V. E. Zakharov, "Collapse of Langmuir waves," *Sov. Phys. JETP* **35**, 908–914 (1972).

[5] A. Hasegawa and Y. Kodama, "Signal transmission by optical solitons in a monomode fiber," Proc. IEEE **69**, 1145–1150 (1981).

[6] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform* (SIAM, Philadelphia, 1981). See Chapter 1.

[7] M. D'Evelyn and G. J. Morales, "Properties of large amplitude Langmuir solitons," **21**, 1997–2008 (1978).

[8] P. K. A. Wai, C. R. Menyuk, Y. C. Lee, and H. H. Chen, "Nonlinear pulse propagation in the neighborhood of the zero dispersion wavelength of monomode optical fibers," Optics Lett. **11**, 464–468 (1986).

[9] P. K. A. Wai, C. R. Menyuk, Y. C. Lee, and H. H. Chen, "Solitons at the zero-dispersion wavelength of single-mode fibers," *Technical Digest of the XIV I.Q.E.C. Conference, June 9–13, San Francisco, CA* (Optical Society of America, Washington, 1986), pp. 126–128.

[10] A. J. Lichtenberg and M. A. Lieberman, *Regular and Stochastic Motion* (Springer-Verlag, New York, 1983). See Chapter 2.

[11] C. R. Menyuk, "Lie perturbation theory for an infinite-dimensional Hamiltonian system" in *Proceedings of the 15th International Colloquium on Group*

Theoretical Methods in Physics, edited by W. W. Zachary (World Scientific Publ., Singapore, 1984).

[12] C. R. Menyuk, "Origin of solitons in the 'real' world," Phys. Rev. A **33**, 4367–4374 (1986).

[13] G. K. Keiser, *Optical Fiber Communication* (McGraw-Hill, New York, 1983). See Chapter 3.

[14] A. Hasegawa and F. Tappert, "Transmission of stationary nonlinear optical pulses with birefringent fibers," Appl. Phys. Lett. **23**, 142–144 (1973).

[15] L. F. Mollenauer, R. H. Stolen, and J. P. Gordon, "Experimental observation of picosecond pulse narrowing and solitons in optical fibers," Phys. Rev. Lett. **45**, 1045–1048 (1980).

[16] L. F. Mollenauer and R. H. Stolen, "Solitons in optical fibers," Laser Focus **18**(4), 193–198 (1982).

The Effects of Boundary Conditions on Electromagnetic Pulses

K. C. Heaton

Defence Research Establishment Valcartier

Abstract

As is well known, extremely energetic explosions are capable of generating intense electric and magnetic fields. When these explosions occur near a good electrical conductor, such as the Earth, the behaviour of the electric and magnetic fields is different to that obtained from isolated explosions. In particular, the continuity of the tangential components of the quasi-static electric field between the air and the ground requires the vanishing of the electric field along the surface of the Earth. In previous studies, this boundary condition has been held to imply that only electric fields whose angular dependence is given by odd spherical harmonics contribute to the total field above the ground.

In this work, it is shown that solutions exist to Maxwell's equations which satisfy the boundary conditions at the Earth's surface for the even spherical harmonics and which are not zero throughout all space.

Maxwell's equations for the fields are solved numerically, and results presented which indicate that the contribution of these fields to the total electric field may be significant at certain angles.

1 Introduction

It is well known that extremely energetic chemical explosions can produce electric and magnetic signals of appreciable magnitude at considerable distances from the location of the explosion (Glasstone and Dolan 1977). These fields seem to be generated by two distinct mechanisms: 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).

However, as one might have expected from the larger energies involved in nuclear explosions, the electromagnetic fields produced in these cases are of proportionately greater magnitudes. These are generated by electric currents caused by Compton scattering of electrons by X- and γ-rays from the nuclear explosion. The fields caused by nuclear explosions are generally known as electromagnetic pulses (EMP) (e. g. Longmire and Gilbert 1980, Longmire 1978). In the case of the nuclear explosions, the fields which are not generated by Compton scattering can be significant only at very late times.

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. In the case of nuclear explosions, the fields generated can possess 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). Another effect of interest associated with nuclear explosions is the presence of lightning flashes at times of up to 1 second after the explosion (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.

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 in 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, it is shown that the condition that the quasi-static electric field vanish along the surface of the Earth does not imply that the even multipoles of the field can not exist, and further, that these fields can be of appreciable magnitudes. At the surface of the Earth, these even multipole fields can induce a surface charge density which counteracts the radial field there and hence satisfy the boundary conditions. At locations other than the surface of the Earth, this cancellation is not complete, leaving a finite field composed of the sums of the original even multipole field and the fields produced by the surface charge induced at the Earth's surface. Sample calculations for estimates of typical field strengths are presented.

2 Maxwell's Equations for the Quasi-static Phase of EMP

The two relevant time dependent Maxwell equations are

$$\frac{\partial \vec{B}}{\partial t} = -\vec{\nabla} \times \vec{E}, \tag{1}$$

$$\epsilon \frac{\partial \vec{E}}{\partial t} = -\vec{j} + \frac{1}{\mu} \vec{\nabla} \times \vec{B}, \tag{2}$$

where $\vec{B}$ is the magnetic intensity in webers/m^2, $\vec{E}$ the electric field in volts/m, $\vec{J}$ the current density in amps/m^2, 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. (1) and (2) become

$$\vec{\nabla} \times \vec{E} = 0, \tag{3}$$

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{J}, \tag{4}$$

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

$$\vec{J} = \vec{J}_s + \sigma \vec{E} \tag{5}$$

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 1978). However, up to fields of strength ~ 100 kV/m, this dependence is small and can be neglected.

After substituting eq. (5) into eq. (4) and taking the divergence, one obtains

$$- \vec{\nabla} \cdot (\sigma \vec{E}) = \vec{\nabla} \cdot \vec{J}_s. \tag{6}$$

In order to satisfy eq. (3), the electric field must be derivable from a potential, thusly:

$$\vec{E} = -\vec{\nabla}\Phi \tag{7}$$

where

$$\Phi = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} Z_n^m(r)\, S_n^m(\theta, \phi). \tag{8}$$

In eq. (8), Z_n^m is the function containing the radial dependence of the potential associated with each surface spherical harmonic, S_n^m. The surface spherical harmonics of angular order m and rank n are defined by

$$S_n^m(\theta, \phi) = P_n^m(\cos\theta)e^{im\phi} \tag{9}$$

where the associated Legendre functions, P_n^m, are given by

$$P_n^m(\cos\theta) = (-1)^m \sin^m\theta \frac{d^m P_n(\cos\theta)}{d(\cos\theta)^m} \tag{10}$$

389

and the Legendre polynomials, P_n , by

$$P_n(\cos\theta) = \frac{(-1)^n}{2^n n!} \frac{d^n(\sin^{2n}\theta)}{d(\cos\theta)^n}. \tag{11}$$

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.

The substitution of eqs. (7) - (11) into eq. (6) yields, after considerable simplification,

$$\vec{\nabla} \cdot \vec{J}_s = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} \left(\sigma \frac{d^2 Z_n^m}{dr^2} + \left(\frac{2\sigma}{r} + \frac{d\sigma}{dr}\right) \frac{dZ_n^m}{dr} - n(n+1)\frac{\sigma}{r^2} Z_n^m \right) S_n^m(\theta, \phi). \tag{12}$$

It is assumed that the divergence of $\vec{J}_s$, the source current density, can be expressed by

$$\vec{\nabla} \cdot \vec{J}_s = \sum_{m=-\infty}^{\infty} \sum_{n=0}^{\infty} F_n^m(r) S_n^m(\theta, \phi) \tag{13}$$

where F_n^m is the function containing the radial dependence of the divergence of the source current density associated with each surface spherical harmonic, S_n^m ; 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).

Multiplying eq. (12) by the spherical harmonic $S_{n'}^{m'}$, and using

$$\int_0^{2\pi}\int_0^\pi S_n^m(\theta,\phi) S_{n'}^{m'}(\theta,\phi) \sin\theta \, d\theta \, d\phi = (-1)^m \frac{4\pi}{(2n+1)} \delta_{n'}^n \delta_{-m'}^m, \tag{14}$$

one can separate the radial functions associated with each spherical harmonic, and obtain a 2nd order differential equation for Z_n^m , the radial dependence of the electric potential, thusly:

$$\sigma \frac{d^2 Z_n^m}{dr^2} + \left(\frac{2\sigma}{r} + \frac{d\sigma}{dr}\right) \frac{dZ_n^m}{dr} - n(n+1)\frac{\sigma}{r^2} Z_n^m = F_n^m. \tag{15}$$

Grover (1980) and others (e. g. Hodgdon 1984) have derived similar equations, with the important difference that the summations in eqs. (12) - (13) were taken over only the odd values of n. 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 F_n^m is 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 at the surface of the Earth, assuming it to be an infinite plane located at $\theta = 90°$, are:

$$\hat{n} \times (\vec{E}_2 - \vec{E}_1) = 0, \tag{16}$$

$$\hat{n} \cdot (\vec{D}_2 - \vec{D}_1) = \varpi, \tag{17}$$

$$\hat{n} \cdot (\vec{B}_2 - \vec{B}_1) = 0, \tag{18}$$

$$\hat{n} \times (\vec{H}_2 - \vec{H}_1) = \vec{K}, \tag{19}$$

(Jackson 1962, Stratton 1941). In eqs. (16) - (19), the variables with subscript 1 refer to the Earth, and those with with subscript 2 refer to the air. As before, $\vec{E}$ is the electric field, and $\vec{B}$ the magnetic induction. As well, $\vec{D}$ is the electric displacement, and $\vec{H}$ the magnetic field. ϖ is the surface charge density on the Earth, $\vec{K}$ the surface current density, and $\hat{n}$ the outward unit normal to the surface of the Earth. In this paper, we shall make use only of eqs. (16) - (17), but eqs. (18) - (19) are included for the sake of completeness.

At this stage, it will be assumed that all physical processes involved in the explosion and the field are symmetrical with respect to 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. (8) - (15) is always 0, and hence that one is left only with a summation over the rank n.

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. Bearing in mind, then, that these boundary conditions can only be said to apply to that part of the Earth in which the conductivity is at least an order of magnitude greater than that in the air, eqs. (16) - (17) become:

$$\hat{n} \times \vec{E}_2 = 0, \tag{20}$$

$$\hat{n} \cdot \vec{D}_2 = \varpi. \tag{21}$$

For simplicity, it will nonetheless be assumed that the boundary conditions, eqs. (20) - (21), apply on the whole surface of the Earth. In the numerical calculations, this simply means that one must confine oneself to region in which this applies. 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. (7) - (8) into eqs. (20) - (21), one obtains

$$\sum_{n=0}^{\infty} \left(\sin\theta \frac{dZ_n^0}{dr} P_n(\cos\theta) + \cos\theta \frac{Z_n^0}{r} \frac{dP_n}{d\theta}(\cos\theta) \right)\bigg|_{\theta=90°} = 0 \tag{22}$$

and

$$\sum_{n=0}^{\infty} \left(\cos\theta \frac{dZ_n^0}{dr} P_n(\cos\theta) - \sin\theta \frac{Z_n^0}{r} \frac{dP_n}{d\theta}(\cos\theta) \right)\bigg|_{\theta=90°} = -\frac{\varpi}{\epsilon_0} \tag{23}$$

as boundary conditions. The summations over the odd Legendre polynomials P_n vanish, leaving only the summations over the even polynomials to be satisfied. The usual practice (Hodgdon 1984, Grover 1980) has been to satisfy the boundary condition by insisting $\dfrac{dZ_n^0}{dr} = Z_n^0 = 0$ 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 a 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.

Let the total potential, Φ_T , throughout all space be given by

$$\Phi_T = \Phi + \Phi_1 \tag{24}$$

where Φ is the potential as given by eqs. (8) and (12), caused by the source and conduction currents, and Φ_1 is the potential induced by the surface charge density ϖ to counteract Φ at the surface of the Earth. Substituting eq. (24) into eqs. (16) - (17), one obtains

$$\left. \frac{\partial \Phi}{\partial r} \right|_{\theta=90^\circ} = - \left. \frac{\partial \Phi_1}{\partial r} \right|_{\theta=90^\circ} , \tag{25}$$

$$\frac{1}{r} \left(\frac{\partial \Phi_1}{\partial \theta} + \frac{\partial \Phi}{\partial \theta} \right) \bigg|_{\theta=90^\circ} = \frac{\varpi}{\epsilon_0}. \tag{26}$$

Since eq. (25) is true over the whole $\theta = 90^\circ$ plane , one can integrate eq. (25) over the surface of the Earth to obtain

$$\Phi_1(\rho) = -\Phi(\rho) + C \tag{27}$$

where $\Phi_1(\rho)$ is the potential along the Earth and C is a constant of integration. Since at great distances from the initial explosion, the potential caused by it must drop to 0, $C = 0$.

Evidently,

$$\Phi_1(\rho, z) = \int_0^\infty f(k) e^{-kz} J_0(k\rho) \, dk \tag{28}$$

where $\Phi_1(\rho, z)$ is now the potential throughout all space due to the surface charge induced on the Earth, as a function of the cylindrical co-ordinates ρ and z centred at $r = 0$, $J_0(k\rho)$ is the 0th order Bessel function and $f(k)$ is an unknown function to be determined from the boundary conditions (Jackson 1962). Multiplying both sides of eq. (28) by $\rho J_0(k'\rho)$ and integrating with respect to ρ from 0 to ∞ , one obtains

$$\int_0^\infty \rho \Phi_1(\rho, z) J_0(k'\rho) \, d\rho = \int_0^\infty f(k) \frac{e^{-kz}}{k} \delta(k - k') \, dk. \tag{29}$$

Setting $z = 0$ in eq (29), one can evaluate the right hand integral to obtain an expression for $f(k)$:

$$f(k) = \int_0^\infty k\rho' \Phi_1(\rho') J_0(k\rho') \, d\rho'. \tag{30}$$

Substituting eq. (30) into eq. (28) one obtains finally that

$$\Phi_1(\rho, z) = \int_0^\infty \int_0^\infty k e^{-kz} J_0(k\rho) \rho' \Phi_1(\rho') J_0(k\rho') \, d\rho' \, dk. \tag{31}$$

Since the potential $\Phi_1(\rho)$ on the surface of the Earth is known from eq. (27), in principle, eq. (31) provides a means of calculating the potential resulting from the surface charge induced on the Earth to cancel the field there. Using the series expansion for the 0th order Bessel function,

$$J_0(k\rho') = \sum_{l=0}^\infty (-1)^l \frac{1}{2^{2l} \, l! \, \Gamma(l+1)} k^{2l} \rho'^{2l}, \tag{32}$$

eq. (31) becomes

$$\Phi_1(\rho, z)$$
$$= \sum_{l=0}^\infty (-1)^l \frac{1}{2^{2l} \, l! \, \Gamma(l+1)} \int_0^\infty \Phi_1(\rho') \rho'^{2l+1} \, d\rho' \int_0^\infty k^{2l+1} J_0(k\rho) e^{-kz} \, dk \tag{33}$$

where Γ is the gamma function. From Gradshteyn and Ryzhik (1971), pg 711,

$$\int_0^\infty x^{\mu-1} J_\nu(\beta x) e^{-\alpha x} \, dx$$

$$= (\alpha^2 + \beta^2)^{-\frac{1}{2}\mu} \, \Gamma(\nu + \mu) \, P_{\mu-1}^{-\nu} \left(\frac{\alpha}{(\alpha^2 + \beta^2)^{\frac{1}{2}}} \right), \tag{34}$$

$$\alpha > 0, \, \beta > 0, \, Re(\nu + \mu) > 0$$

where J_ν is the νth order Bessel function. As can be seen, eq. (33) may be evaluated by the use of eq. (34), with $\alpha = z$, $\beta = \rho$, $\mu = 2l + 2$, $\nu = 0$. Hence,

$$\int_0^\infty k^{2l+1} J_0(k\rho) e^{-kz} \, dk = \frac{1}{(z^2 + \rho^2)^{l+1}} \Gamma(2l + 2) P_{2l+1} \left(\frac{z}{(z^2 + \rho^2)^{\frac{1}{2}}} \right), \tag{35}$$

$$z > 0, \, \rho > 0.$$

By definition,

$$r = (z^2 + \rho^2)^{\frac{1}{2}},$$

$$\cos \theta = \frac{z}{(z^2 + \rho^2)^{\frac{1}{2}}}, \tag{36}$$

$$\Gamma(n + 1) = n!, \, n \in I,$$

393

and so

$$\Phi_1(r,\theta) = \frac{1}{r^2} \sum_{l=0}^{\infty} (-1)^l \frac{(2l+1)!}{2^{2l}(l!)^2} \frac{P_{2l+1}(\cos\theta)}{r^{2l}} A_l \tag{37}$$

where

$$A_l = \int_0^{\infty} \Phi_1(\rho') \rho'^{2l+1} \, d\rho' \tag{38}$$

It should be noted that eq. (37) is not valid for $z = 0$ (i. e. $\theta = 90°$). However, the potential and radial field are known on that plane since they must exactly counteract those produced by the source and conduction currents. In principle, then, the fields caused by the surface charge density on the Earth are known.

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 depend 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 even the multipole fields can contribute to the total field.

3   Numerical Methods

Before one attempts numerical solutions of the field equations, eq. (15), 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 pairs of ion-electron pairs. At early times, the electronic conductivity dominates; at late times, the ionic dominates. The expression for the total conductivity is hence

$$\sigma = \sigma_e + \sigma_I \tag{39}$$

where σ_e, the electronic conductivity, is given by

$$\sigma_e = e \mu_e \frac{S}{\alpha_e} \tag{40}$$

and σ_I, the ionic conductivity, is given by

$$\sigma_I = 2e \mu_I \left(\frac{S}{\gamma_I} \right)^{\frac{1}{2}} \tag{41}$$

(Downey 1983, Wyatt 1980, Grover 1980). In eqs. (40) and (41) 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

$$S = S_0 \frac{\exp(-r/\lambda)}{r^2} \tag{42}$$

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

$$\begin{aligned}
F_0(t) &= -3.9 \times 10^{-22} Y_0 N_a \exp(-8.33 \times 10^2 t), \\
G_0(t) &= 8.2 \times 10^{-22} Y_0 N_a \exp(-8.33 \times 10^2 t), \\
H_0(t) &= -2.8 \times 10^{-23} Y_0 N_a \exp(-16.7t), \\
F(r,t) &= \frac{F_0(t)}{r^2} \left[\exp(-2.65 \times 10^{-5} \rho_0 r) - \exp(-1.04 \times 10^{-4} \rho_0 r) \right], \\
G(r,t) &= \frac{G_0(t)}{r^2} \exp(-4.61 \times 10^{-5} \rho_0 r), \\
H(r,t) &= \frac{H_0(t)}{r^2} \left[\exp(-2.20 \times 10^{-5} \rho_0 r) - \exp(-4.78 \times 10^{-5} \rho_0 r) \right], \\
X(r,t) &= F(r,t) + H(r,t), \\
Y(r,t) &= 16F(r,t) + 1.3H(r,t), \\
U(r,t) &= G(r,t), \\
V(r,t) &= -G(r,t).
\end{aligned} \tag{43}$$

In terms of the functions defined in eq. (43), the source current densities are given by

$$\begin{aligned}
J_r &= F(r,t)(1 + 16\cos\theta), \\
J_\theta &= G(r,t)(1 - \cos\theta),
\end{aligned} \tag{44}$$

for ground capture sources, and

$$\begin{aligned}
J_r &= H(r,t)(1 + 1.3\cos\theta), \\
J_\theta &= 0,
\end{aligned} \tag{45}$$

395

for air capture sources (Downey 1983). In eqs. (43) - (45), Y_0 is the total yield in kilotons, N_a is the number of neutrons/ kiloton, ρ_0 is the air density in mg/cm³, r is the radial distance from the blast in metres, θ is the polar angle, t the retarded time in seconds, J_r the radial current density in abamps/cm² , and J_θ the polar current density in abamps/cm². The total current density at any retarded time t must be the vector sum of eqs. (44) - (45) . Hence, the components of of the source current density are

$$J_r = X(r,t) + Y(r,t) \cos \theta, \tag{46}$$

$$J_\theta = U(r,t) + V(r,t) \cos \theta. \tag{47}$$

Therefore,

$$\vec{\nabla} \cdot \vec{J_s} = \left(\frac{\partial X}{\partial r} + 2 \frac{X}{r} \right) + \left(\frac{\partial Y}{\partial r} + 2 \frac{Y}{r} \right) \cos \theta + \frac{U}{r} \frac{\cos \theta}{\sin \theta} + \frac{V}{r} \frac{\cos^2 \theta - \sin^2 \theta}{\sin \theta}. \tag{48}$$

Then, the substitution of eq. (48) into eq. (13), along with the use of eq. (14), yields

$$F_0^0 = \frac{\partial X}{\partial r} + 2 \frac{X}{r},$$

$$F_1^0 = \frac{\partial Y}{\partial r} + 2 \frac{Y}{r} + \frac{3\pi}{4} \frac{U}{r},$$

$$F_n^0 = \frac{(2n+1)\pi}{8} \frac{V}{r} \sum_{k=0}^{n/2} (-1)^k \frac{(2n-2k)!(n-2k+2)(n-2k)}{2^{2n-2k} k!(n-k)! \left((n/2-k+1)! \right)^2} \tag{49}$$

for n even , $n \geq 2$,

$$F_n^0 = (2n+1)\pi \frac{U}{r} \sum_{k=0}^{(n-1)/2} (-1)^k \frac{(2n-2k)!}{2^{2n-2k} k!(n-k)! \left(((n-1)/2-k)! \right)^2 (n-2k+1)}$$

for n odd , $n \geq 3$.

By substituting eq. (49) back into eq. (15), it is possible to solve numerically for the radial part of the potential and the field. It should, however, be noted that eq. (49) must be converted into amps/m³ 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

$$\frac{dZ_n^0}{dr} = y_1 \frac{ML}{QT^2},$$

$$Z_n^0 = y_2 \frac{ML^2}{QT^2},$$

$$r = r^* L,$$

$$t = t^* T, \tag{50}$$

$$\sigma = \sigma^* \frac{TQ^2}{ML^3},$$

$$F_n^0 = \left(F_n^0 \right)^* \frac{Q}{TL^3},$$

396

where

$$Q = L^3 T \left(F_n^0(r_0) \right),$$

(51)

$$M = L^3 T \left[\left(F_n^0 \right)^2 / \left(\frac{d\sigma}{dr} \right) \right]_{r=r_0},$$

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 equation, eq. (15) becomes

$$\frac{dy_1}{dr^*} = \frac{-1}{\sigma^*} \left(\frac{2\sigma^*}{r^*} + \frac{d\sigma^*}{dr^*} \right) y_1 + n(n+1) \frac{y_2}{r^{*2}} + \frac{\left(F_n^0 \right)^*}{\sigma^*},$$

(52)

$$\frac{dy_2}{dr^*} = y_1.$$

Equations (52) were solved using a four-point Runge-Kutta algorithm with automatic error controls. It was necessary to find initial solutions to begin the integration. Unfortunately, the field equations, eq. (52), the expression for the conductivity σ, eq. (39), and for the excitation function F_n^0, eq. (49), all possess the unfortunate property of singularity at the origin. This implies that the expressions used for σ, $\vec{J}_s$, and F_n^0 cease to be applicable close to the blast site and others must be used. The derivation of these, however, presents considerable problems. Instead of attempting to determine initial values for the field and potential near the blast site, it was decided to use the fact that at very large distances from the blast, both field and potential must be zero. Hence, if one starts the integration at a sufficiently great distance from the blast site and integrates inwards to $r = 0$, the initial values of y_1 and y_2 can both be set to zero. In all of the cases examined in this work, it was found that the excitation function F_n^0 and the source current $\vec{J}_s$ were negligible $\left(\left(F_n^0 \right)^* \ll 10^{-10} \right)$ at distances of $r = 28$ kilometres. Accordingly, the integration was begun at that point with $y_1 = y_2 = 0$ and stopped at $r = .2$ kilometres. At that point, the expressions for the current density and the conductivity, eqs. (40) and (41) will certainly cease to apply (Downey 1983, Wyatt 1980). In point of fact, $r = .4$ kilometres is probably the limit to which eqs. (40) and (41) are even approximately accurate, but the integration was carried out to $r = .2$ kilometres simply for completeness, although the values obtained for y_1 and y_2 for $r < .4$ km. are of questionable accuracy.

Two other numerical problems arose in the evaluation of eqs. (37) and (38) for the even multipole fields, both partially caused by the inapplicability of the conductivity model near the explosion site. At some point within a 4 km. radius around the explosion site, the conductivity will become very large both with respect to that of the Earth and absolutely. Within that radius, the potential in the air

397

must be constant in the quasi-static approximation in order that the field fall to zero there. This implies that the potential $\Phi_1(\rho)$ will not be given by the negative of $\Phi(\rho)$, where $\Phi(\rho)$ is a solution to eqs. (52) with $\sigma^\cdot$ and $(F_n^0)^\cdot$ as given by eqs. (39) - (51). Therefore, the true value of $\Phi_1(\rho)$ in eq. (38) is effectively unknown at values of $r < .4$ km. in the absence of a model for σ and $\vec{J}_s$ near the explosion site, although the continuity requirements on the fields and potential place limits on the magnitude of the error. The problem was addressed by halting the integration of eq. (38) at $r = .4$ km. At worst, the vanishing of the field and the continuity conditions on the potential imply that the error could be no worse than that obtained by holding Φ_1 in eq. (38) constant at its value at $\rho_0 = .4$ km. and integrating. That is, the error term for A_l , ϵ_l, resulting from the termination of the integration at $\rho = .4$ km. , should obey the condition

$$ |\epsilon_l| \leq \left. \frac{|\Phi(\rho_0)|\,\rho_0^{2l+2}}{2l+2} \right|_{\rho_0=.4} . \tag{53} $$

Equation (53) was evaluated for values of Φ arising from the original fields for $n = 2$ and $n = 0$ in eqs. (52) and found to be at least 3 orders of magnitude smaller than A_l for each value of l . A more serious problem concerns the nature of the potential $\Phi_1(\rho)$ in eq. (38). It is evident that A_l will be finite as $r \to \infty$ only if $\Phi_1(\rho)$ decays exponentially as a function of ρ . Since $\Phi_1(\rho)$ was obtained from the numerical solution of the field equations, eq. (52), its error terms can propagate through the integral A_l , coming to dominate over the true value, especially at high values of ρ . A related problem concerns the series, eq. (37). Since it represents a physical quantity, the potential, it must converge everywhere. However, at values of r near the origin, it will diverge because the values of $\vec{J}_s$ and σ are unbounded in that neighbourhood. For high values of r , the series will diverge because of the accumulation of numerical errors.

In order to limit the influence of these types of errors, the following procedure was adopted· Instead of evaluating A_l for each value of l, the quantity

$$ B_l' = \int_0^\infty \Phi_1(\rho') \left(\frac{\rho'}{r}\right)^{2l+1} d\rho' \tag{54} $$

was evaluated for each r and l of interest, with the factor $r^{-(2l+1)}$ effectively acting as an integrating factor. The ratio

$$ R = \left| \frac{C_{l+1}'}{C_l'} \right| \tag{55} $$

where

$$ C_l' = \frac{(2l+1)!}{2^{2l}(l!)^2} P_{2l+1}(\cos\theta) B_l' \tag{56} $$

was computed at each value of r and θ until $R > 1$. Since numerical trials had demonstrated that if $R > 1$ were true at $l = l_0$, it would also be true for all $l > l_0$,

l_0 , the value of l at which $R > 1$, was taken to be the point at which numerical errors were beginning to force the divergence of the series. The series was then truncated at the last value of l for which $R < 1$. Since the true value of B'_{l+1} was assumed to be less than B'_l in order to prevent divergence, this implies that the value calculated for Φ_1 could differ from the true value by no more than B'_{l_0-1} . It should be noted that this procedure does not ensure convergence of the series eq. (37) ; it merely discards those series that are felt to be demonstrably divergent. Unfortunately, since at certain values of r and θ , $R > 1$ for the first two elements of the series, it is not possible even to estimate the magnitudes of the fields and potential there from eq (37). The calculation of the B'_l was done with Simpson's second rule and a base point spacing of .025 km for the fields generated in reaction to the $n = 2$ component of the original field and a base point spacing of .05 km for those generated in reaction to the $n = 0$ component of the original field. The values of B'_l were checked against numerical error by halving the size of the base point spacings.

4   Numerical Results and Analysis

Figures 2-5 show the potential, total electric field, radial electric field and tangential field for the dipole electric field (i. e. for $n = 1$ in eq. (15)) 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. (43)). 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. (42) 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.

Table 1: Comparisons of Calculated EMP Electric Fields

Radius	Total Field	Total Field $\theta = 0$	Total Field $\theta = 0$	Total Field $n = 1, \theta = 0$
	(Wyatt 1980)	(Grover 1980)	(Downey 1983)	(Present Work)
(m)	(kV/m)	(kV/m)	(kV/m)	(kV/m)
500	390	45	23	304
900	164	21	19	128
1300	114	15	13	62

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. 2-5, the total field reaches a maximum of ∼ 300 kV/m at .5 km and falls to less than 1 kV/m at 3.5 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. 2-5 should constitute the greater part of the total electric field. It is encouraging that the magnitudes calculated are in excess of those needed to produce nuclear lightning over much of the range in which they were observed (900 - 1400 m from the blast) at time scales of 1 msec (Wyatt 1980). The values shown for the fields in Figs 2-5 are of the same order of magnitude as those obtained for the same conditions by Wyatt (1980), using two separate conductivity models. Wyatt's values for the fields are listed in Table 1, and compared with the ones obtained here, as well as with 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 well, it should be emphasized that the values included from Figs. 2-5 of this paper are only the dipolar component of the total field. Nonetheless, it is evident that there is a significant difference among the results obtained in the four works cited. In Downey's and Grover's cases, the results are likely attributable to the different conductivity models used. Downey (1983) 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. From this, it seems likely that the true form of the air conductivity would be quite important in any calculation of the of the fields.

Figures 6-9 show the sextopole fields and potential (i. e. for $n = 3$ in eq. (15)). As can be seen, the fields are considerably smaller than for the dipolar field, but

still significant.

Figures 10-17 show the fields and potential for the quadrapolar fields (i. e. $n = 2$ in eq. (15)). Figures 10-12 show the fields and potentials obtained from the solution of eq. (15): that is, they show the fields and potentials generated by the source currents without the addition of the fields and potentials due to the charge density on the surface of the Earth. Figures 13-14 show the fields generated by the surface charge density created by the $n = 2$ fields for various angles of interest.

Three features are of special interest in these figures. The first is that the magnitudes of these fields are frequently of the same order of magnitude as the fields which induced the surface charge density. This implies that the whole complex of fields generated by the even multipoles of the source current can contribute significantly to the total fields, acting at some angles to increase the magnitude of the fields and at others to decrease them.

The other two features of interest concern the two types of anomalous behaviour of the curves at low values of r. Examples of one type are the jumps exhibited near 1100 metres and, less noticeably, near 1600 metres, in the graph of the tangential field at $\theta = 89°$ in Fig. 14. These arise from the accumulation of numerical errors in the calculation of the series coefficients B_l^i, as discussed above. The jumps occur at values of r at which one is able to truncate the series, eq. (37), at a higher value of l than at the preceding value of r, and hence are a representation of the truncation error. Because the convergence of the series is most difficult to assure at low values of r, this error is most severe there. The second type of anomalous behaviour in the curves is the relatively abrupt change in the fields exhibited by the radial fields for $\theta = 0$ and $\theta = 30°$ for $r < 1500$ metres in Fig. 13, and by the tangential field for $\theta = 60°$ in Fig. 14. These features do not occur at places where the number of terms in the series has been increased and so do not seem to be due to truncation errors. They may, however, be an artefact of the models chosen for σ and $\vec{J_s}$. As discussed above, the expressions for the source current density and the conductivity become increasingly inaccurate near the site of the explosion, and the peculiar behaviour of the curves may be a reflection of that.

Figures 15-17 display the effects of the potential and fields produced by the surface charge density on the ones produced by the source currents. As noted above, the effects of the surface charge are significant. Figure 17 is of special interest, since it displays the potential near the Earth's surface. Since the boundary conditions require that the total potential be zero at the surface, one would expect the two potentials to counteract each other to some degree near the Earth's surface as they in fact do. The cancellation would be expected to be complete only at the Earth's surface, where, of course, the expression for the field induced by the surface charges, eq. (37), is not valid.

Figures 18-24 show the fields and potential obtained for the monopole portion of the fields(i. e. $n = 0$ in eq. (15)). This component of the field is of interest both because of its relatively large magnitude as well as for its peculiar structure. Since the potential generated by the source current density has no θ dependence, the

401

only tangential field present is due entirely to the surface charge density. Figures 18-19 show the radial field and potential due to the current density, and Figs. 20-21 the fields created by the surface charges. Again, many of the same curious features observed in the quadrapolar fields and potential are present here, and for the same reasons. It should also be remarked that the convergence of the series and integrals connected with eqs. (37) and (38) is much less satisfactory here than in the quadrapolar case. When the base point spacing in the numerical calculation of B_l' was halved for $n = 2$, the difference between the two calculations of the field, using the two values for each B_l' in eq. (37), was on the order of 1 volt/metre. When the same procedure was followed for the $n = 0$ case, the difference between the two calculations of the field could be as high as 60 volts/metre, although this lessened to 2-3 volts/metre at $r = 5$ km. This probably reflects the difficulty of obtaining a fit for the θ independent parts of the conductivity and source currents. Finally, Figs. 22-24 show the resulting fields and potentials when those due to the source currents and the surface charges are combined.

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. In that event, the even multipole fields induce a surface charge density which cancels the radial field at the surface of the conductor, but which leaves a non-zero even multipole field elsewhere in space.

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 fields generated by a typical explosion. It was found that the dipole field dominated, but that the contribution of the other multipole fields to the total field was significant. In particular, the calculated values of the field were sufficient to produce the lightning which has been observed to accompany nuclear explosions. This result is in agreement with the calculations performed by Wyatt(1980) but contradicts those done by Downey (1983) and Grover (1980). The difference is probably attributable to a different set of initial conditions and atmospheric conductivity model. In passing, it should also be noted that the computational algorithm developed in this work does not require a knowledge of the initial conditions at the blast, but only of those at large distances from the explosion.

Efforts are currently being made to extend this work by incorporating the effects of the induced magnetic fields and more accurate, self-consistent models for the conductivity and source currents.

References

[1] Bacon, D.P. and Cherin, D.P. 1984, *Dust Induced Electro-Magnetic Noise (DIEMN)*, Science Applications Intl. Corp, Maclean, Virginia

[2] Downey, J.R. 1983, M. Sc. Thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

[3] Glasstone, S. and Dolan P. J. 1977, *The Effects of Nuclear Weapons*, United States Department of Defence, Washington, D.C.

[4] Gradshteyn, I.S. and Ryzhik, I.M. 1965, *Tables of Integrals, Series and Products*, Academic Press, New York,

[5] Hodgdon, K.M. 1984, M. Sc. Thesis, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

[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] Jackson, J. D. 1962, *Classical Electrodynamics*, John Wiley and Sons Inc., New York

[8] Lee, K.S.H. 1980, *EMP Interaction: Principles, Techniques, and Reference Data*, Dikewood Industries, Albuquerque, New Mexico

[9] Longmire C.L. 1978, *IEEE Transactions on Antennna and Propagation*, AP-26, no. 1, 3

[10] Longmire, C.L. and Hobbs, W.E. 1979, *Fireball Effects in Late Time EMP from Surface Bursts*, Mission Research Corporation, Santa Barbara, California

[11] Longmire, C.L. and Gilbert, J. L. 1980, *Theory of EMP Coupling in the Source Region*, Mission Research Corporation, Santa Barbara, California

[12] Stratton, J. A. 1941, *Electromagnetic Theory*, McGraw-Hill Book Co. Inc., New York

[13] Uman, M.A., Seacord, D.F., Price, G.H., and Pierce, E.T. 1972, *Journal of Geophysical Research*, 77, 1591

[14] Wilhelm, H.E. 1983, *Appl. Phys. B*, 31, 107

[15] _____ 1984, *J. Appl. Phys.*, 56, no. 5, 1285

[16] 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

Figure 2

POTENTIAL AT t = 1 msec

$\theta = 0°$
$\theta = 30°$
$\theta = 45°$
$\theta = 60°$
$\theta = 89°$

RADIUS (M)

POTENTIAL (kV)

$\times 10^5$

Figure 3

TOTAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TOTAL FIELD (kV/M)

Figure 4

RADIAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

RADIAL FIELD (kV/M)

TANGENTIAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TANGENTIAL FIELD (kV/M)

408
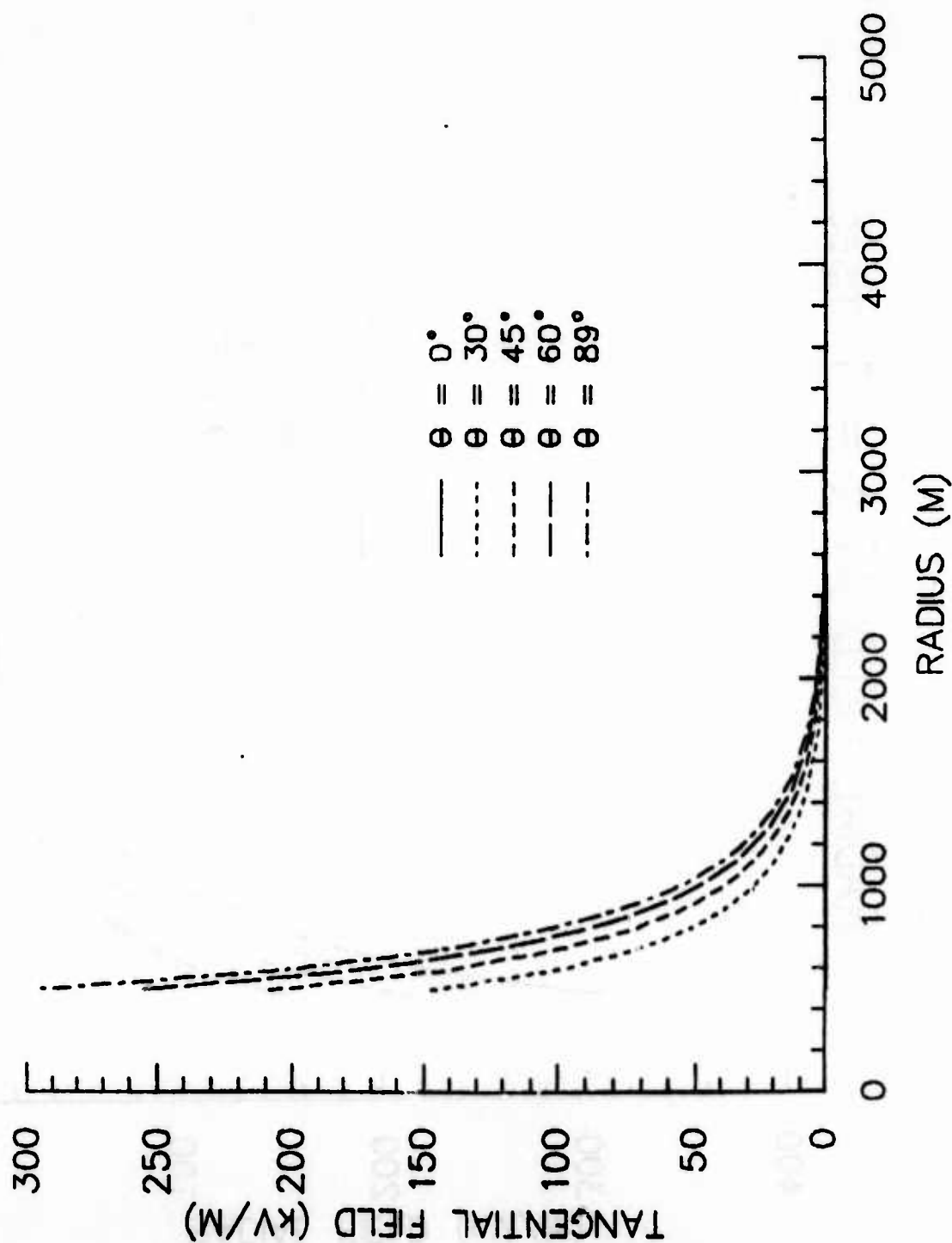
Figure 6

POTENTIAL AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

POTENTIAL (kV)

409

Figure 7

TOTAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TOTAL FIELD (kV/M)
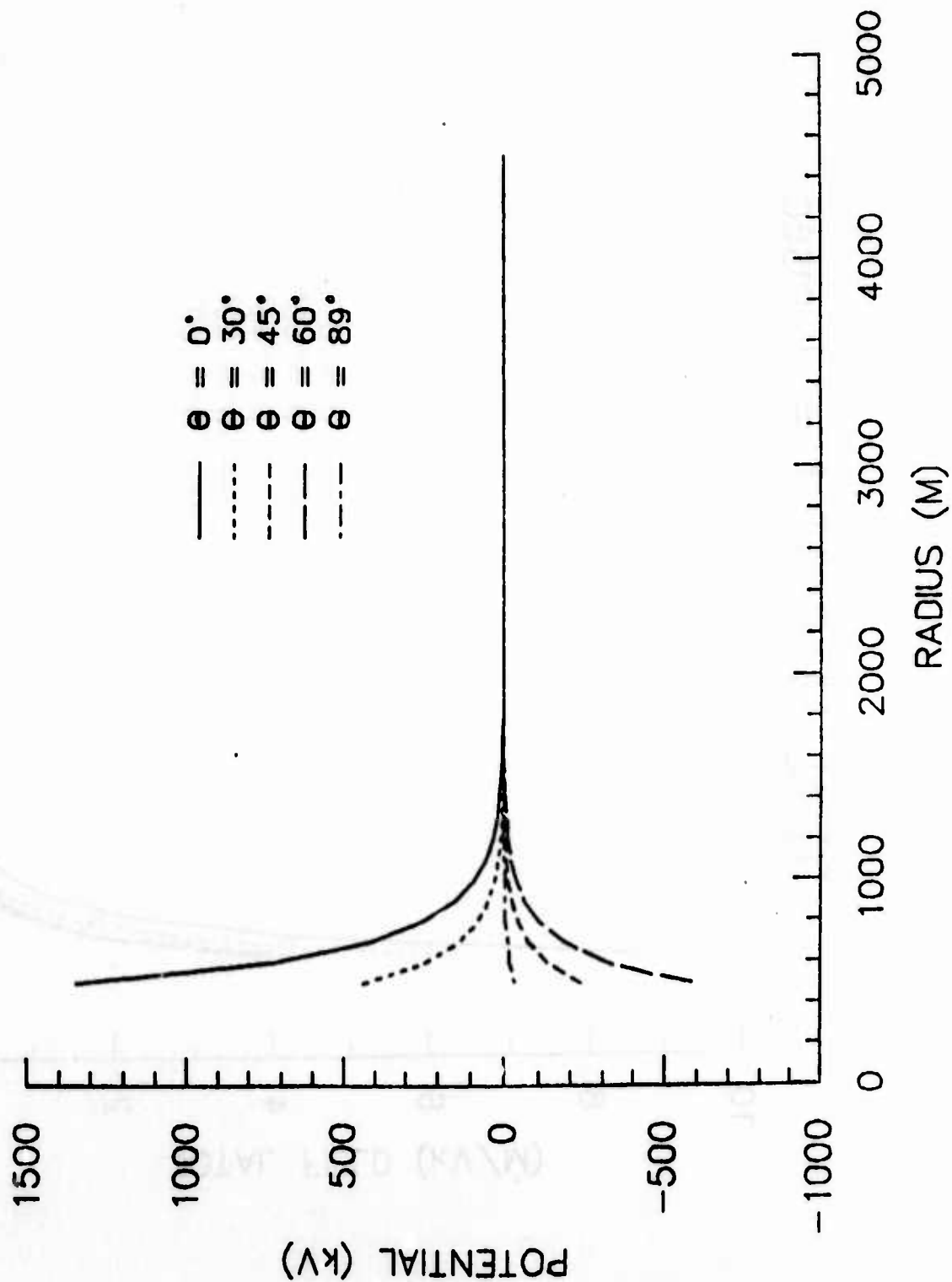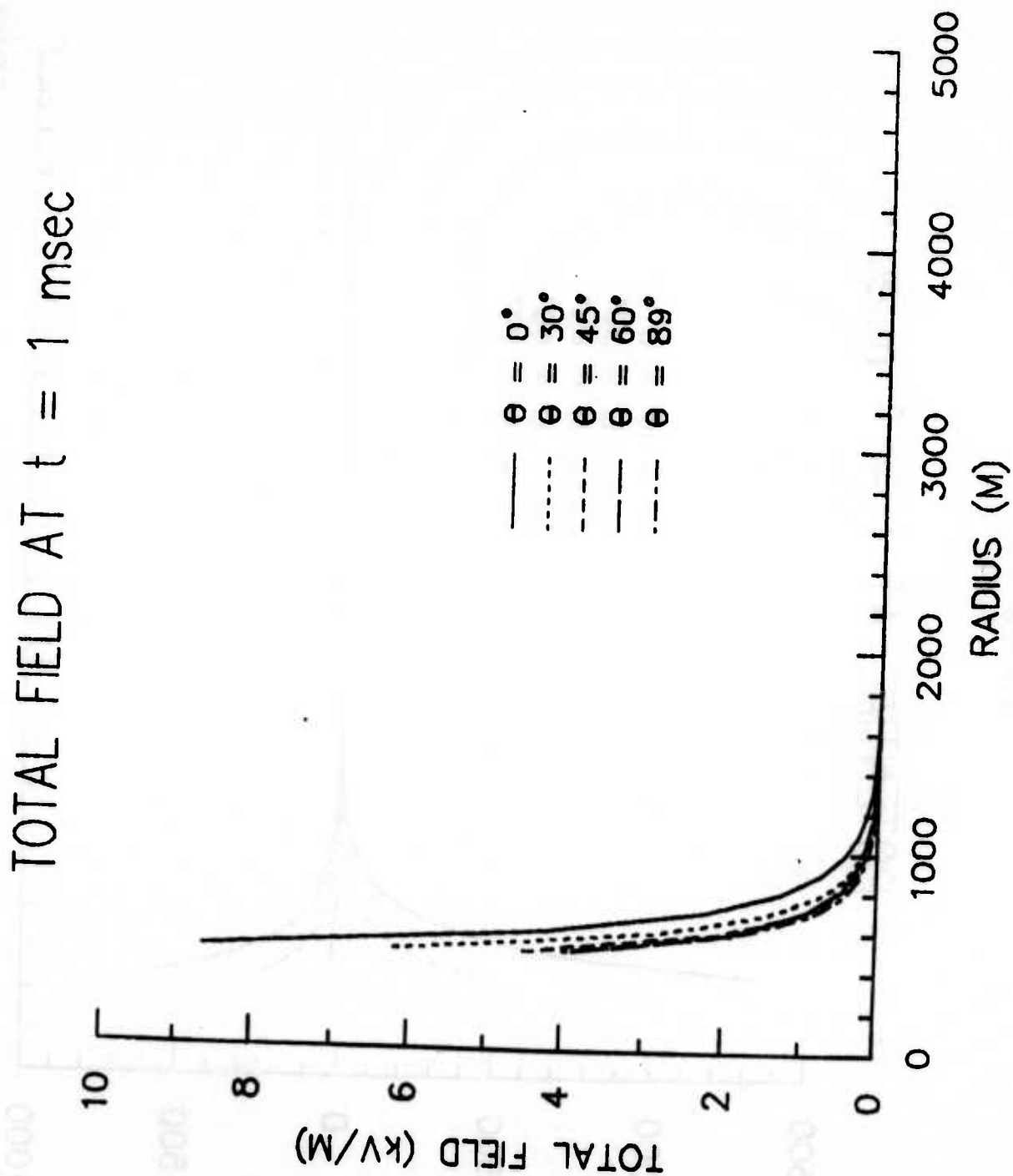
Figure 8

RADIAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)
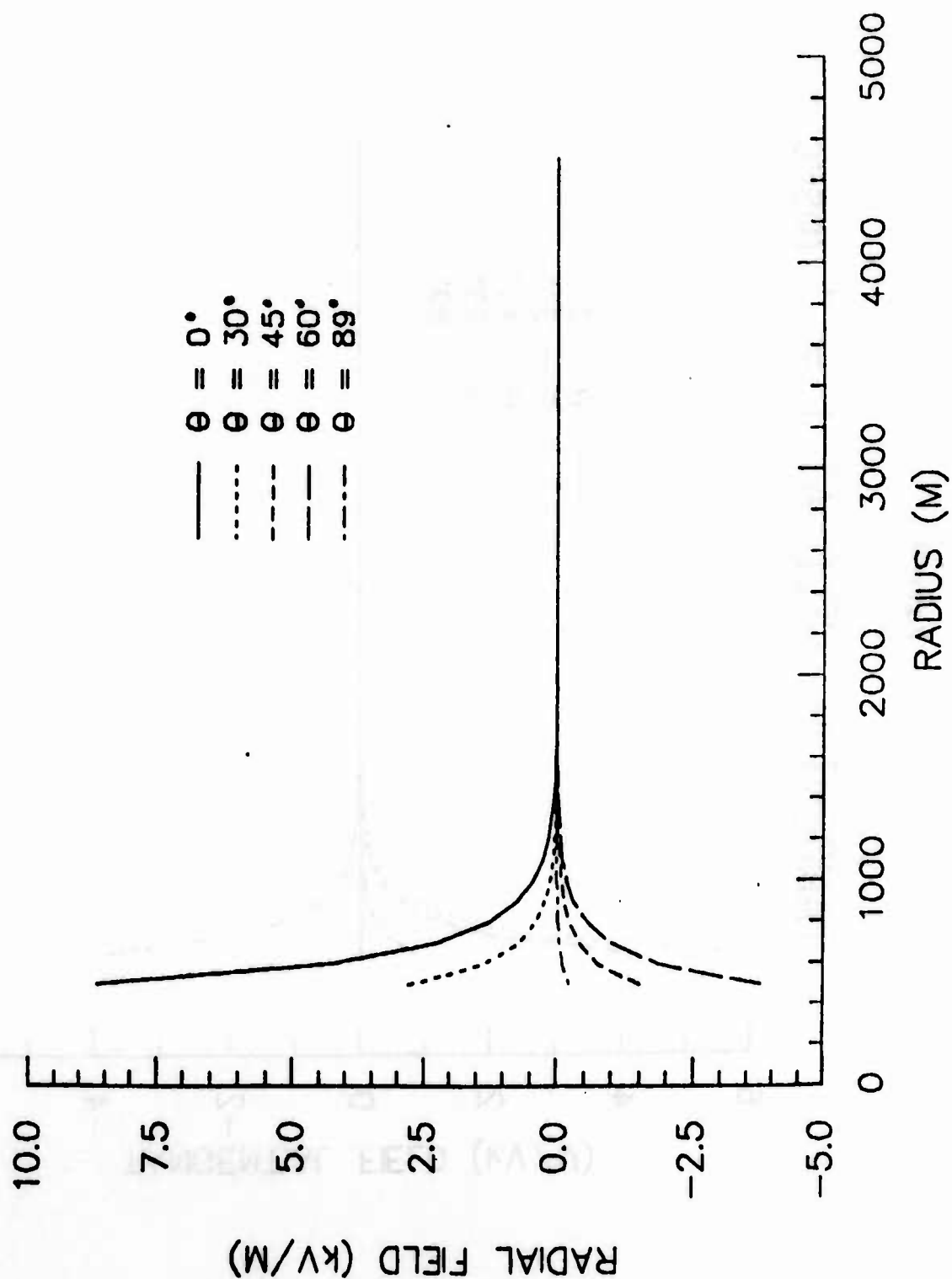
RADIAL FIELD (kV/M)

Figure 9

TANGENTIAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

TANGENTIAL FIELD (kV/M)

RADIUS (M)
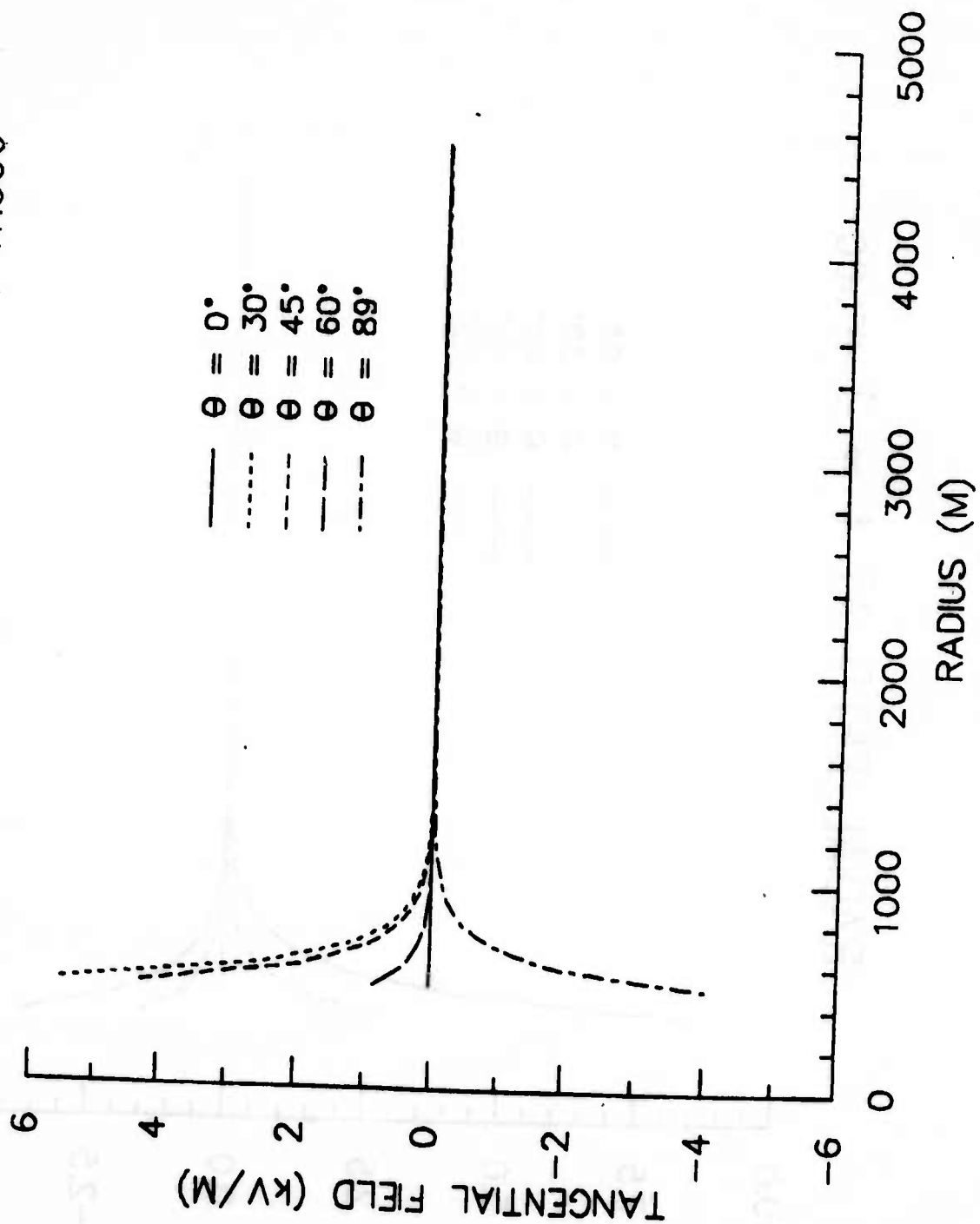
Figure 10

RADIAL FIELD AT t = 1 msec

Figure 11

TANGENTIAL FIELD AT t = 1 msec

θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TANGENTIAL FIELD (kV/M)

414
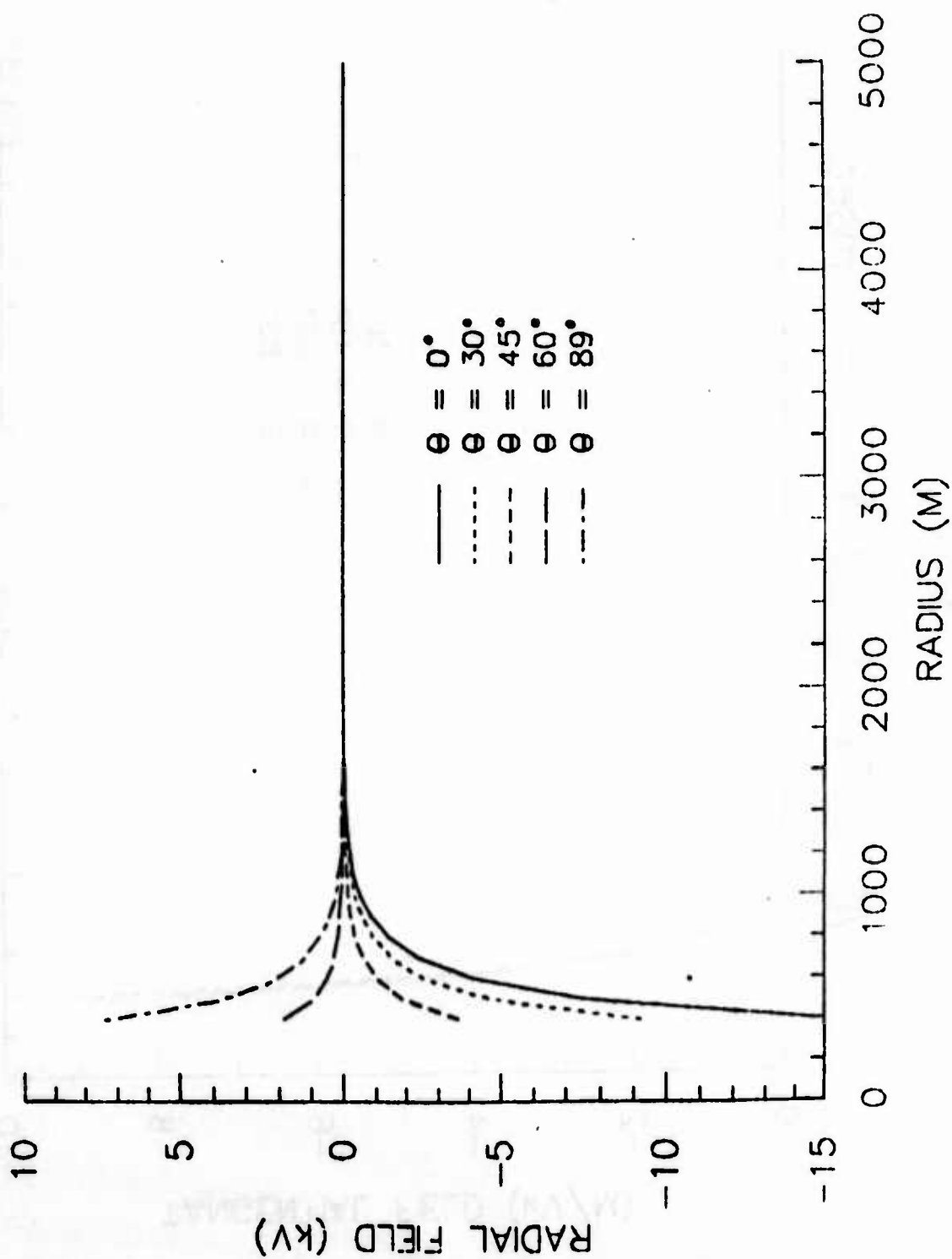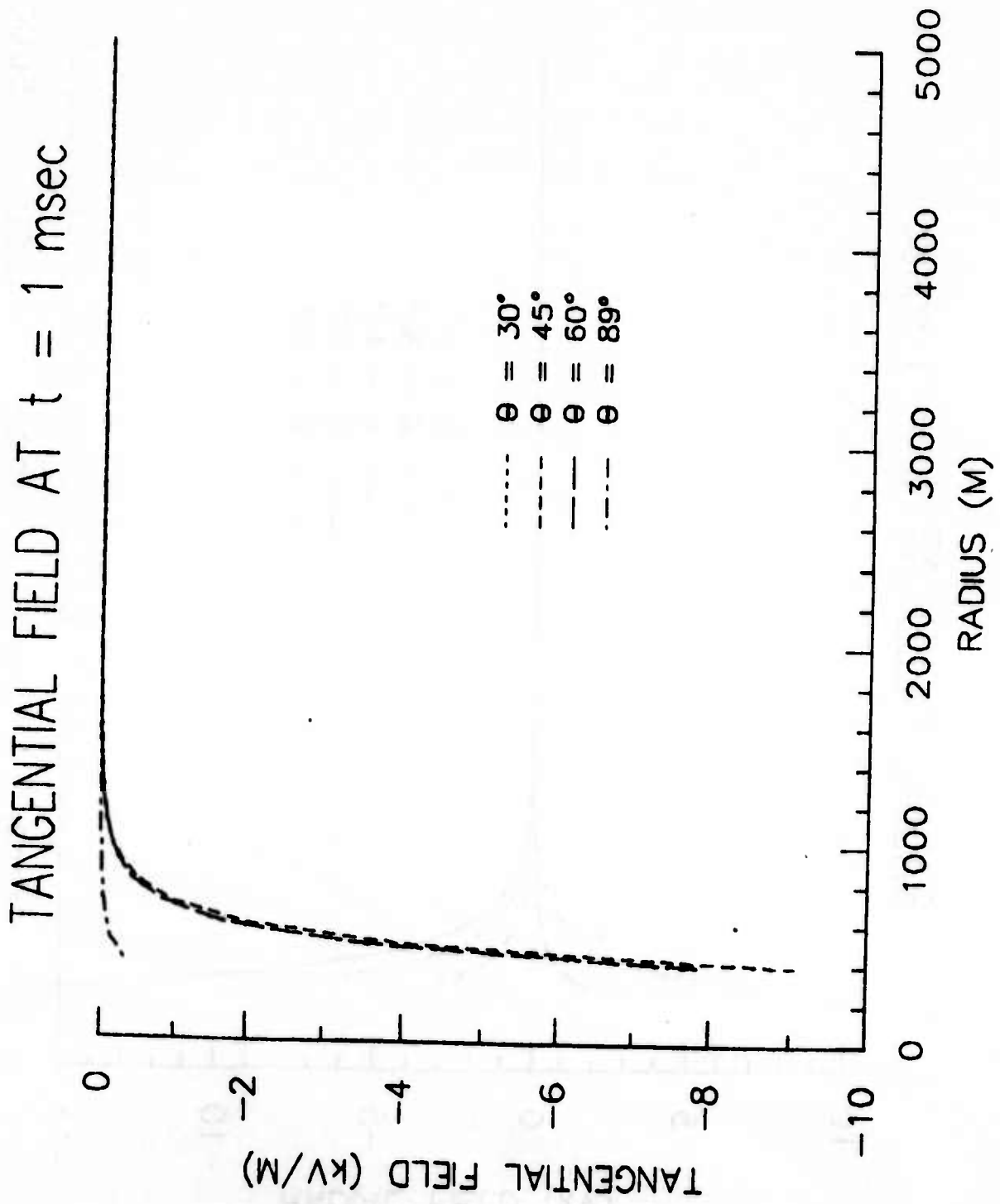
Figure 12

POTENTIAL AT t = 1 msec

Figure 13

RADIAL FIELD AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)
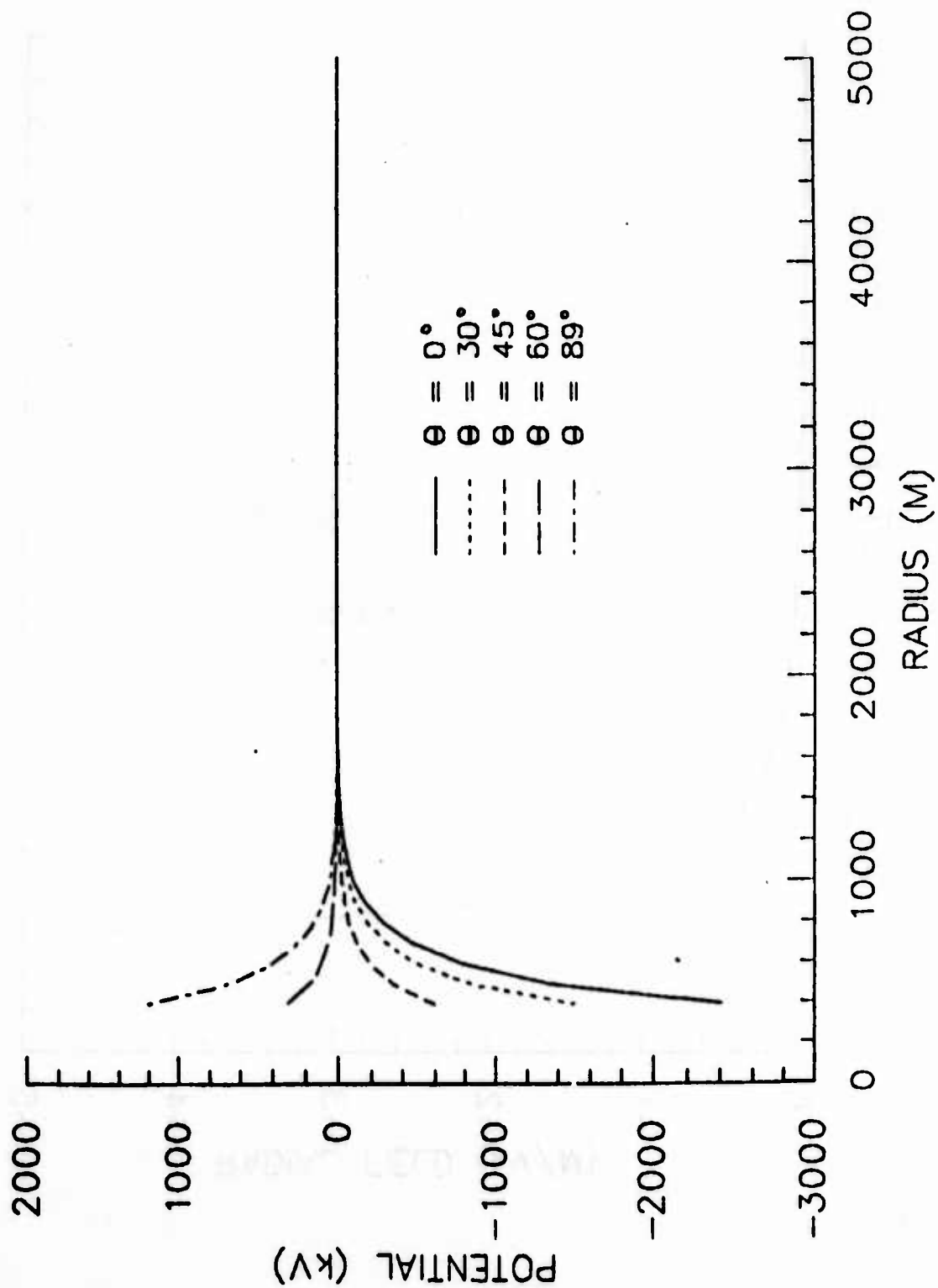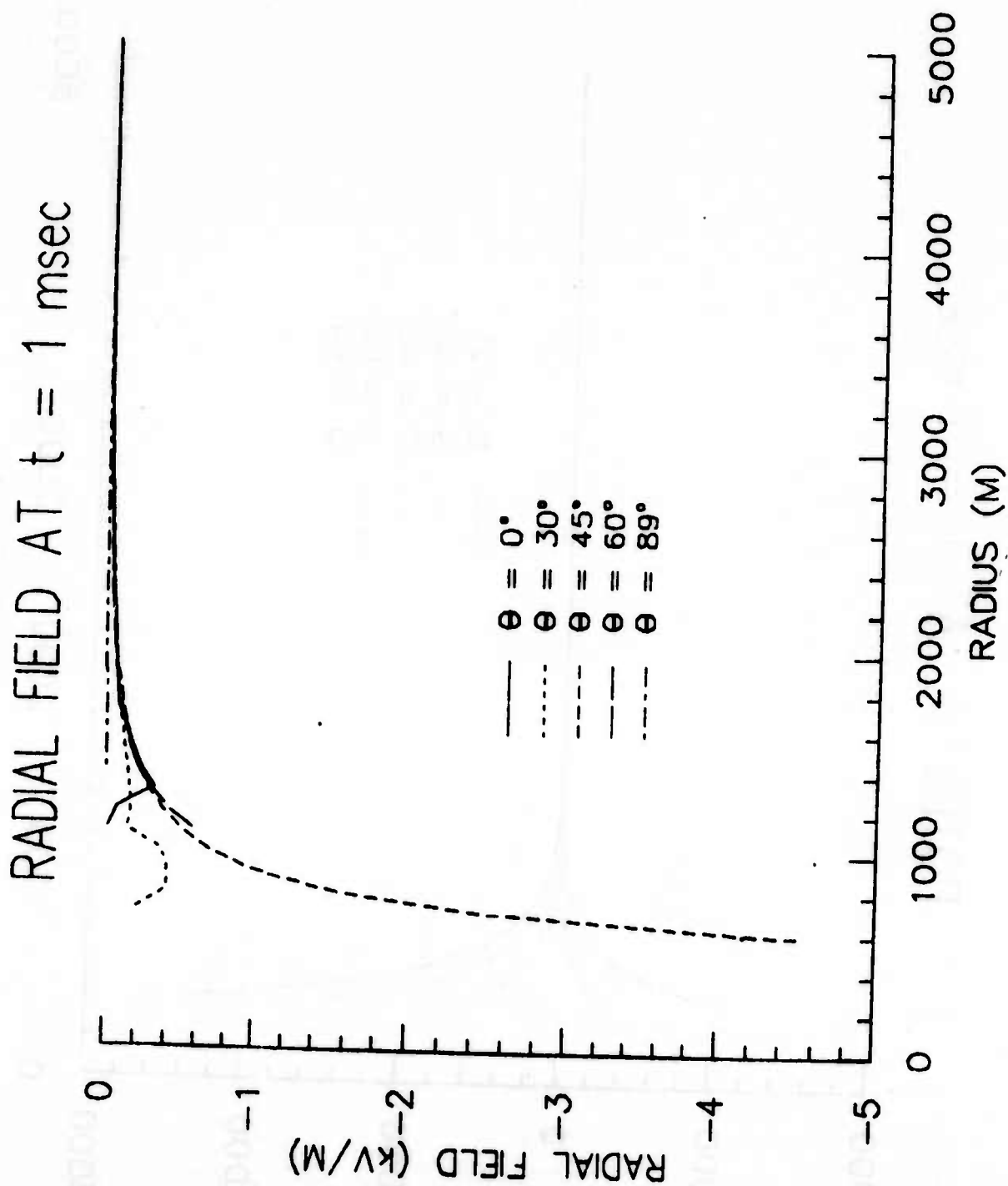
RADIAL FIELD (kV/M)

Figure 14

TANGENTIAL FIELDS AT t = 1 msec

θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TANGENTIAL FIELD (kV/M)

417

Figure 15

TANGENTIAL FIELD AT t = 1 m.sec , θ = 45°

ORIGINAL FIELD
FIELD FROM SURFACE CHARGES
TOTAL FIELD

RADIUS (M)

TANGENTIAL FIELD (kV/M)

418

Figure 16

TANGENTIAL FIELD AT t = 1 msec, θ = 60°

— ORIGINAL FIELD
...... FIELD FROM SURFACE CHARGES
--- TOTAL FIELD

RADIUS (M)
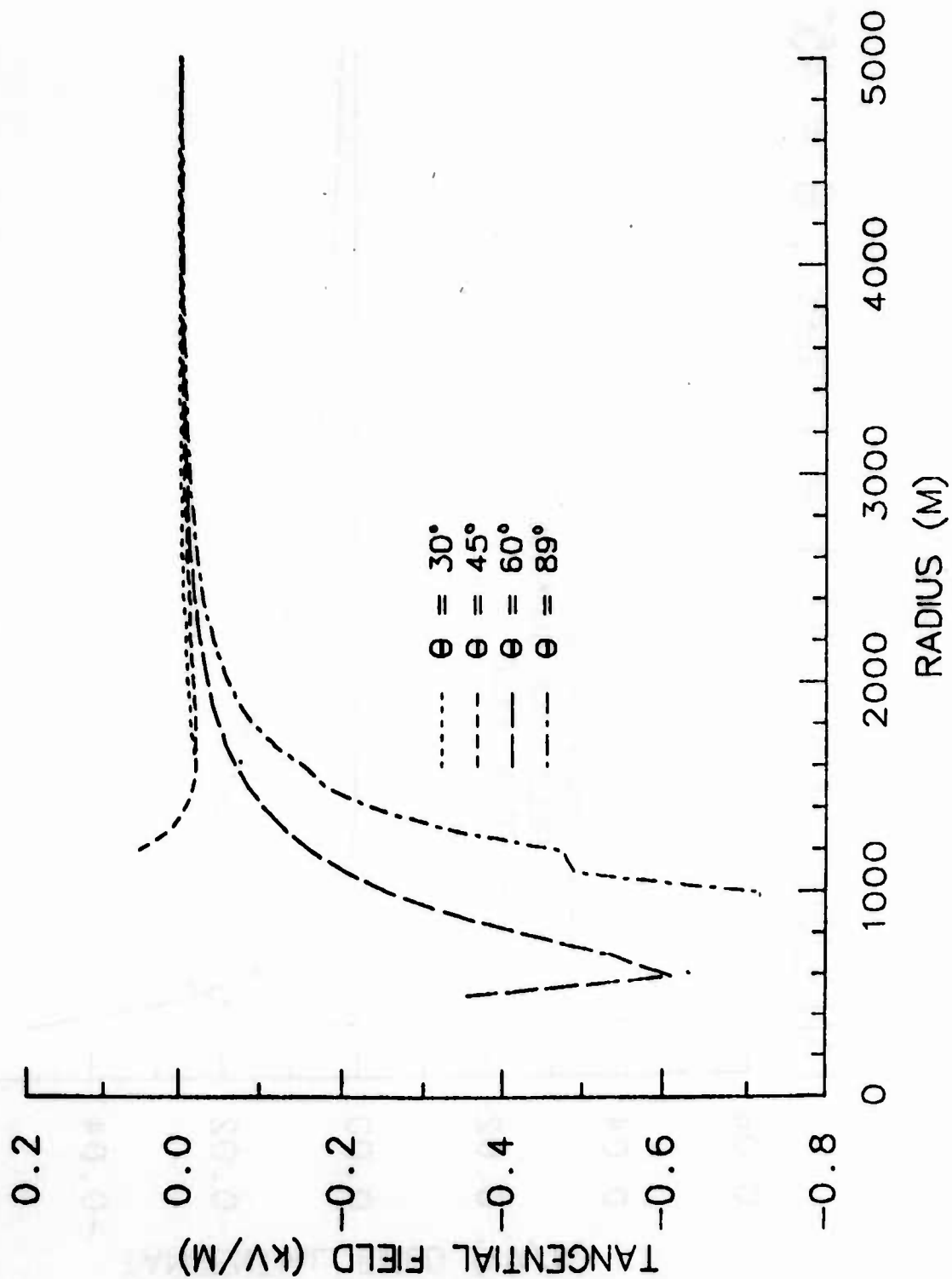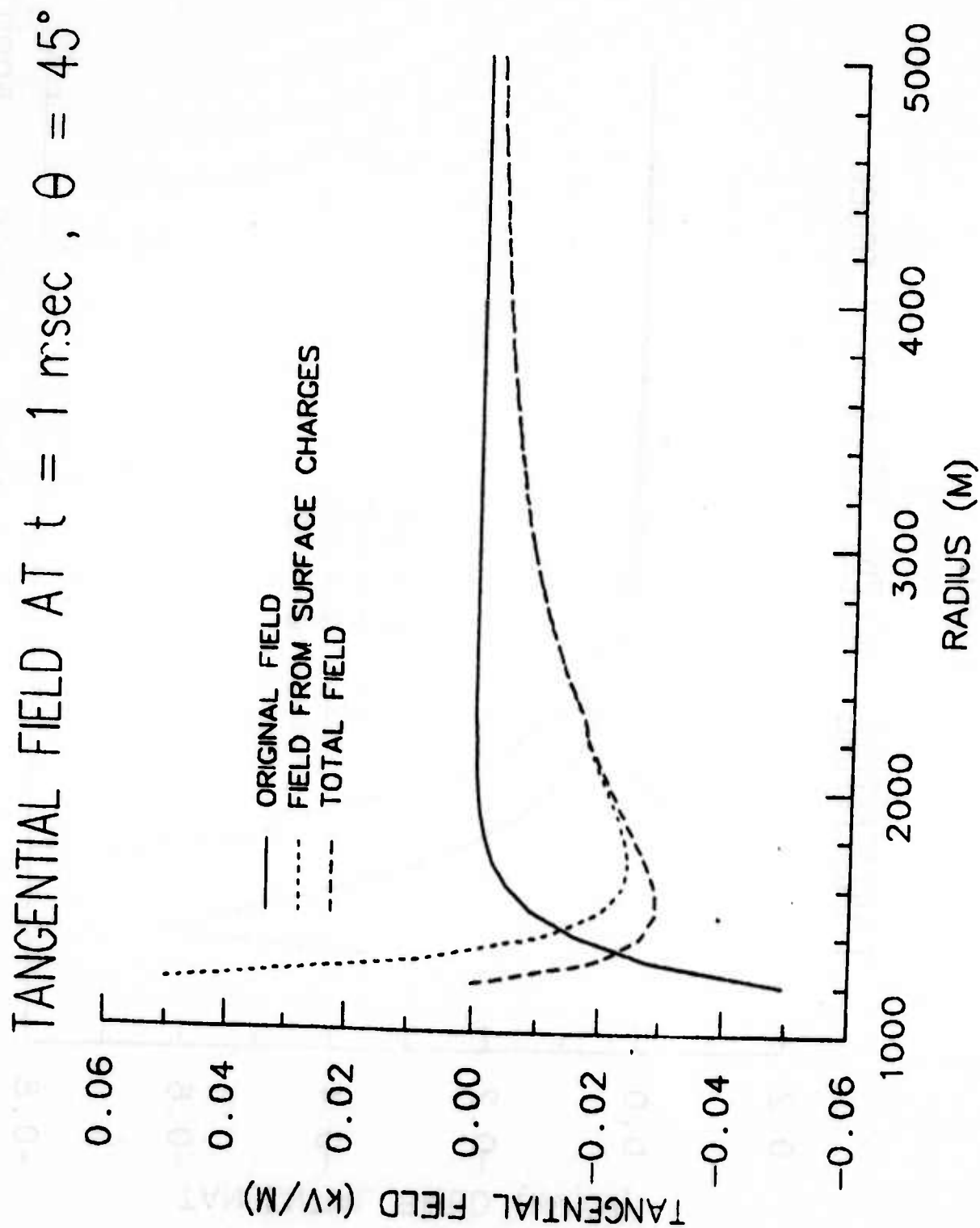
TANGENTIAL FIELD (kV/M)
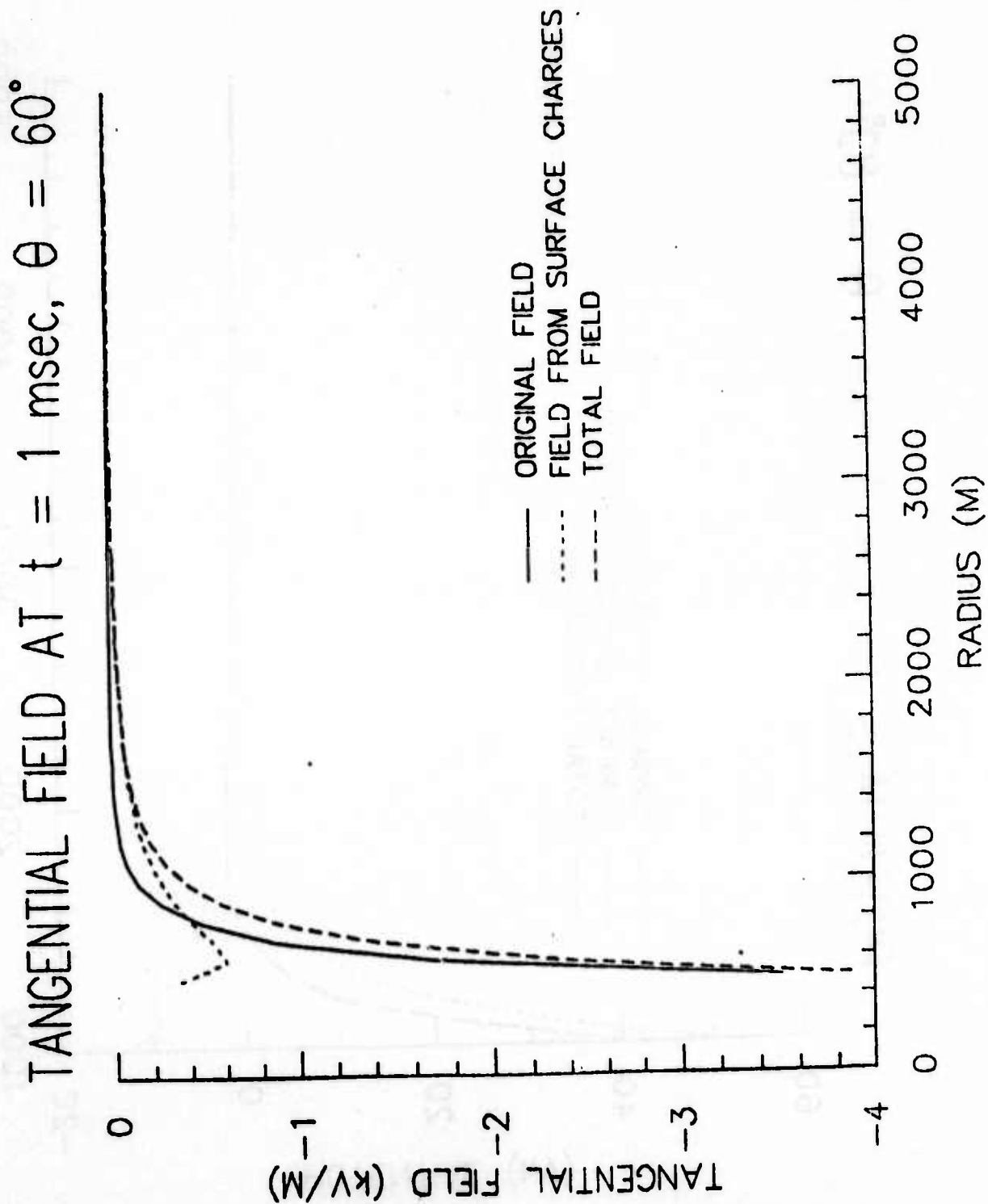
Figure 17

POTENTIALS AT t = 1 msec, θ = 89°

Figure 18

RADIAL FIELD AT t = 1 msec

RADIUS (M)

RADIAL FIELD (kV/M)

421

Figure 19

POTENTIAL AT t = 1 msec

Figure 20

RADIAL FIELDS AT t = 1 msec

θ = 0°
θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIAL FIELD (kV/M)

RADIUS (M)
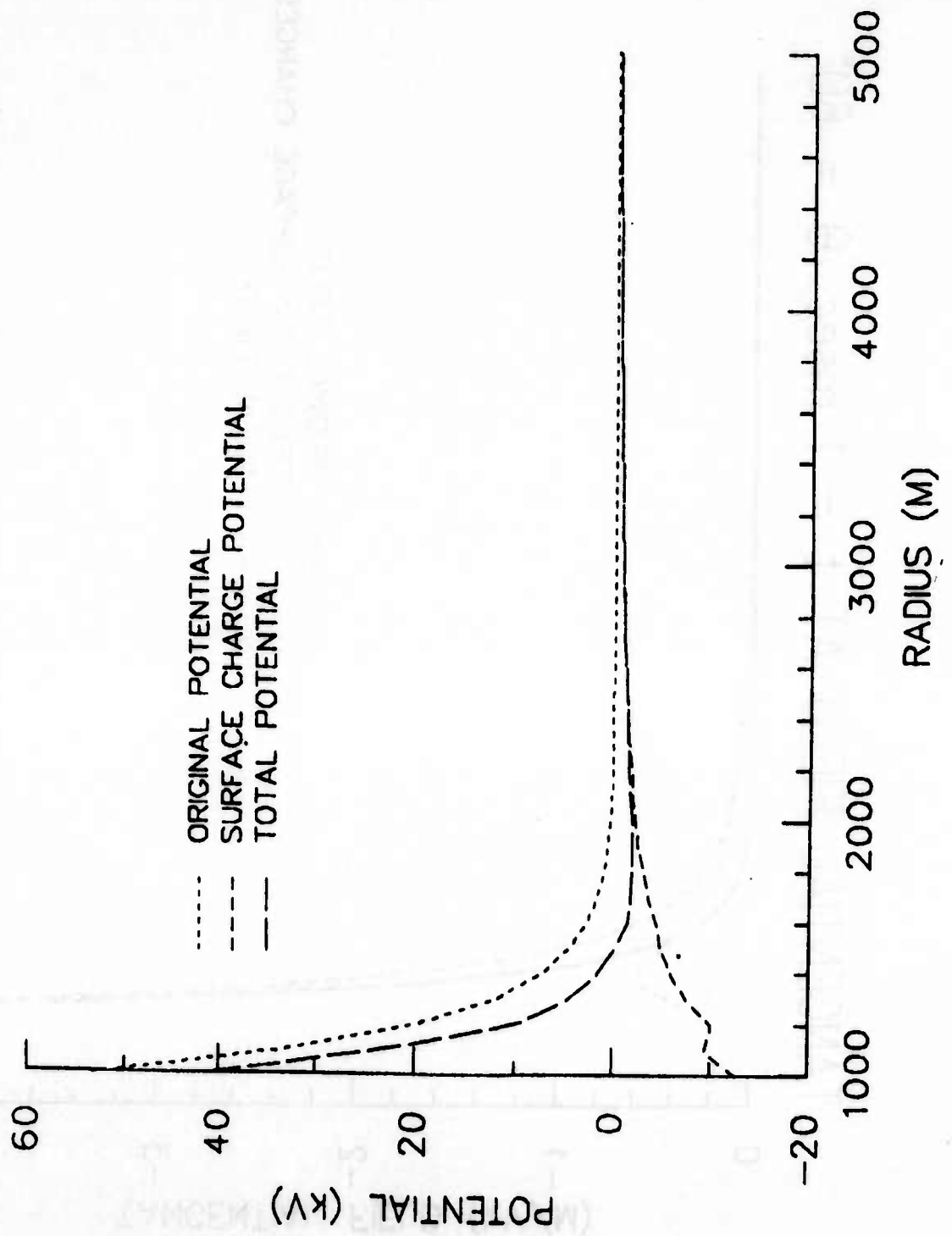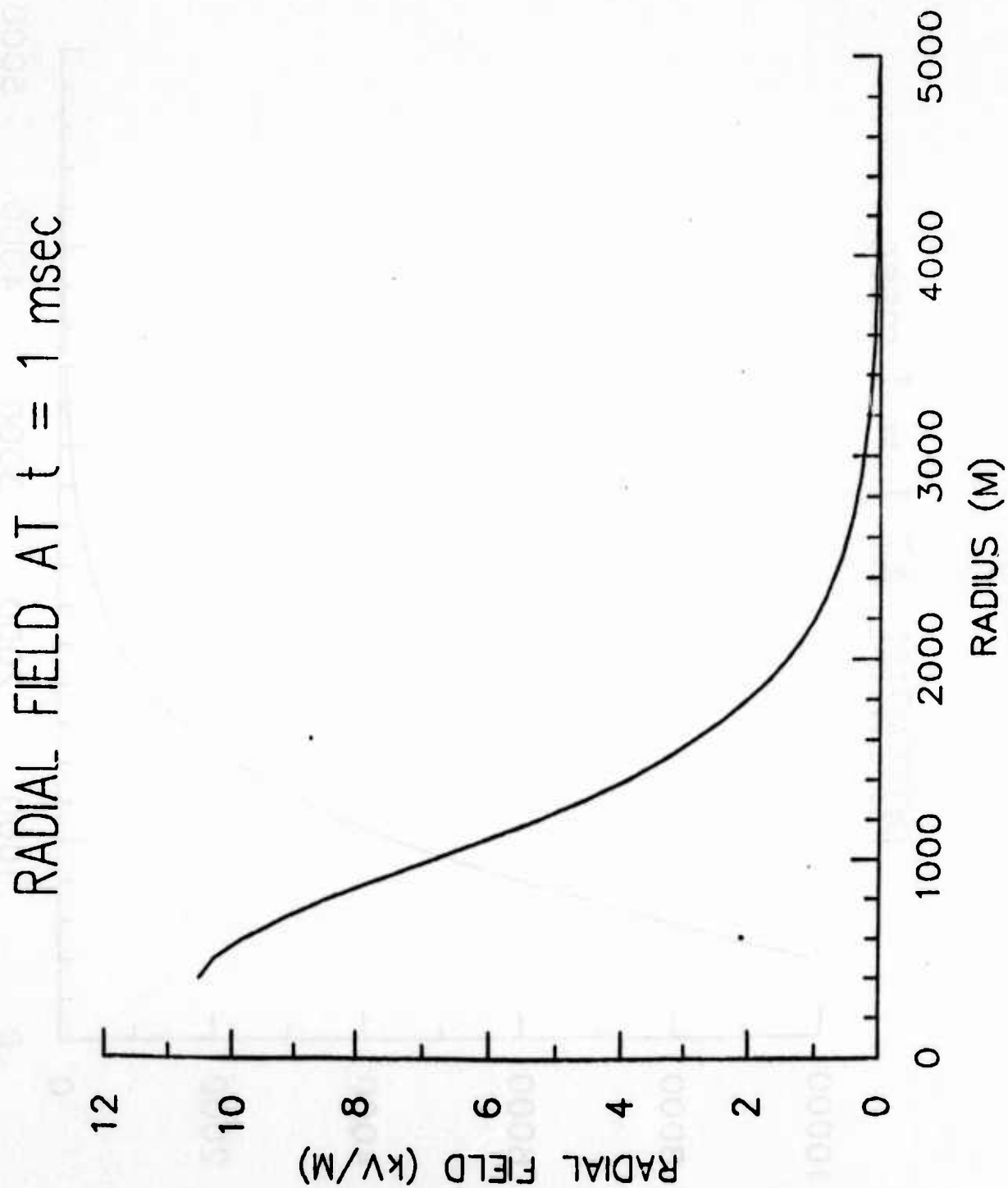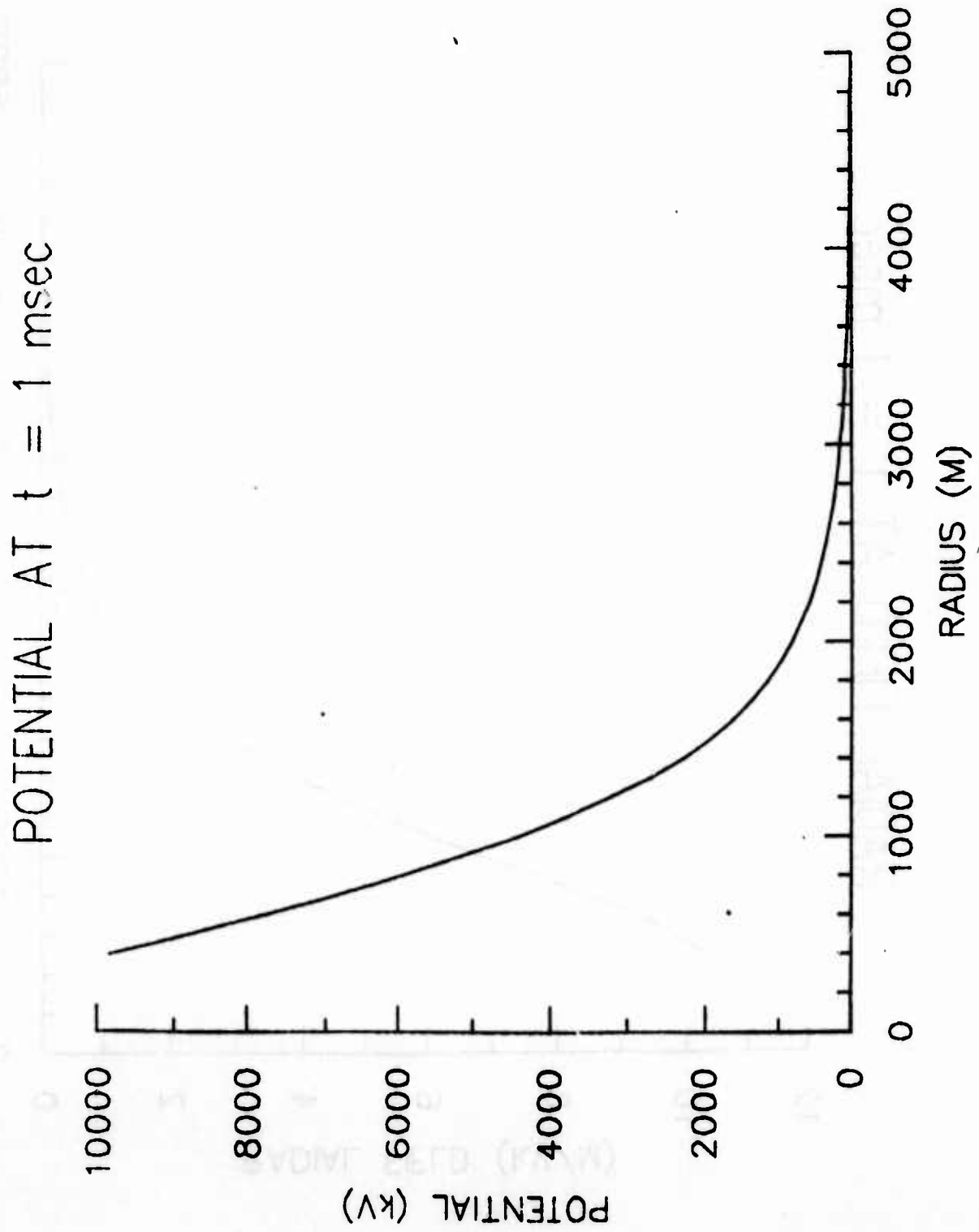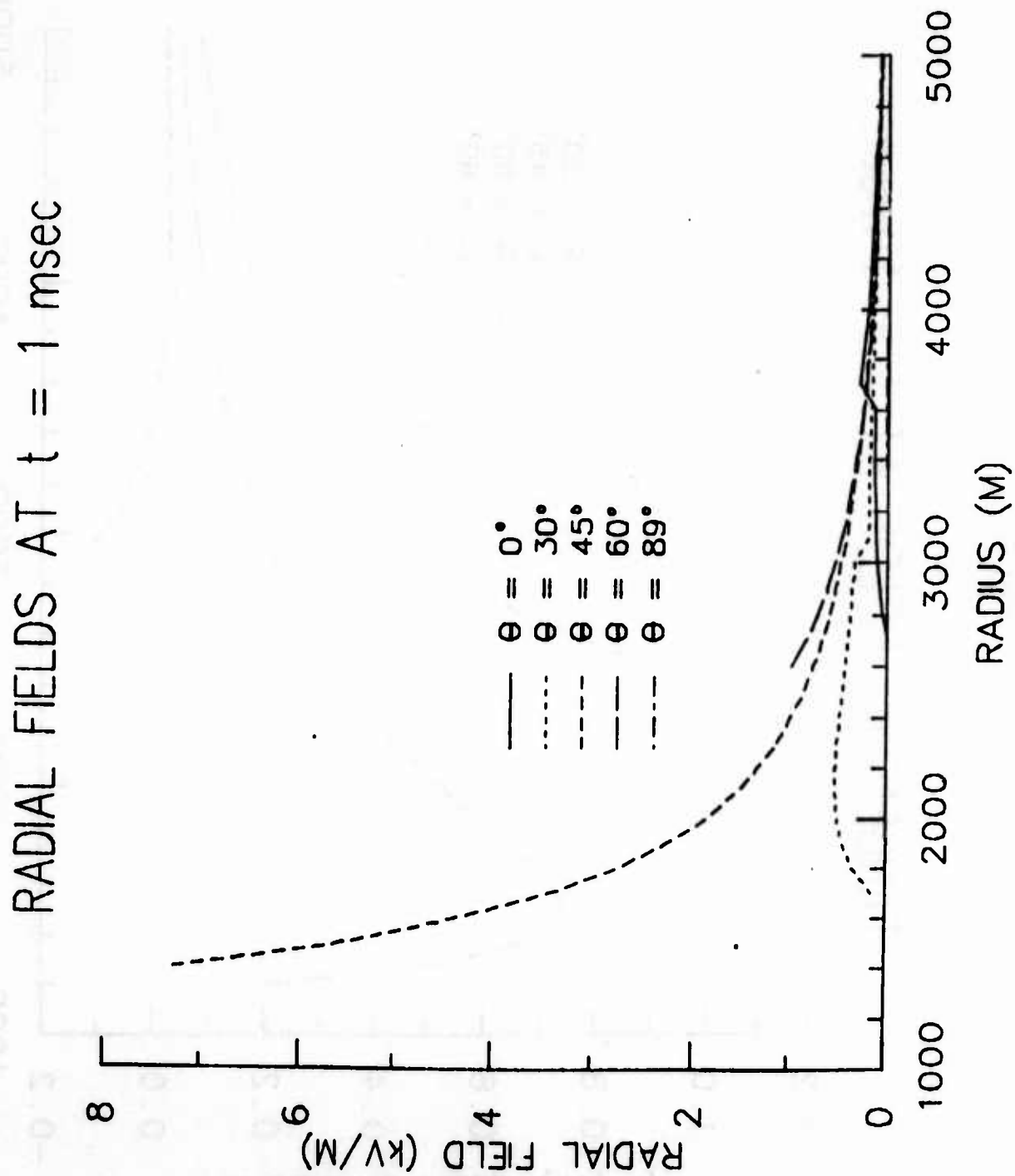
Figure 21

TANGENTIAL FIELDS AT t = 1 msec

θ = 30°
θ = 45°
θ = 60°
θ = 89°

RADIUS (M)

TANGENTIAL FIELD (kV/M)
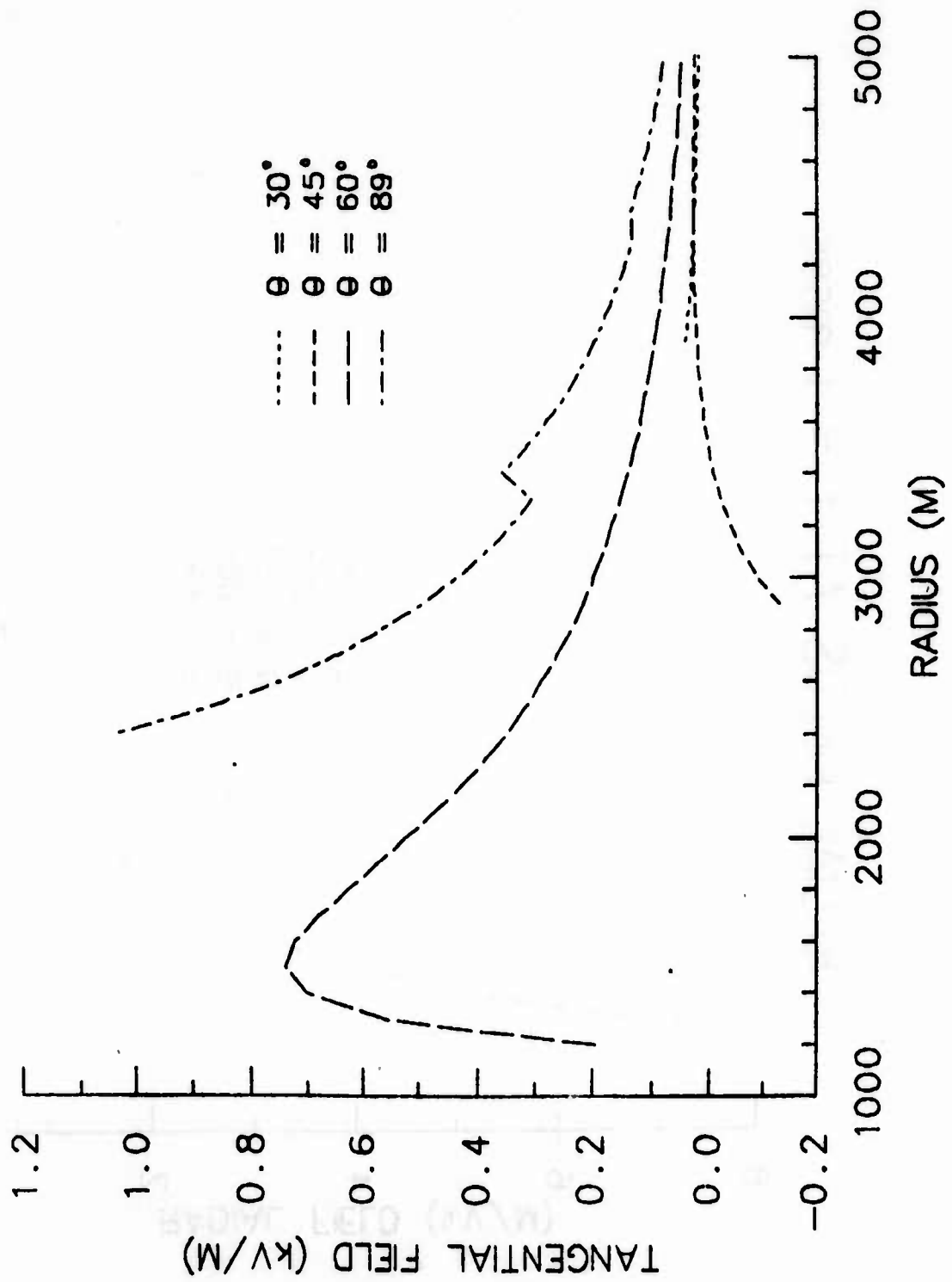
Figure 22

RADIAL FIELD AT t = 1 msec, θ = 45°

425

Figure 23

RADIAL FIELD AT t = 1 msec, $\theta = 60°$

ORIGINAL FIELD
SURFACE FIELD
TOTAL FIELD

RADIUS (M)

RADIAL FIELD (kV/M)

426
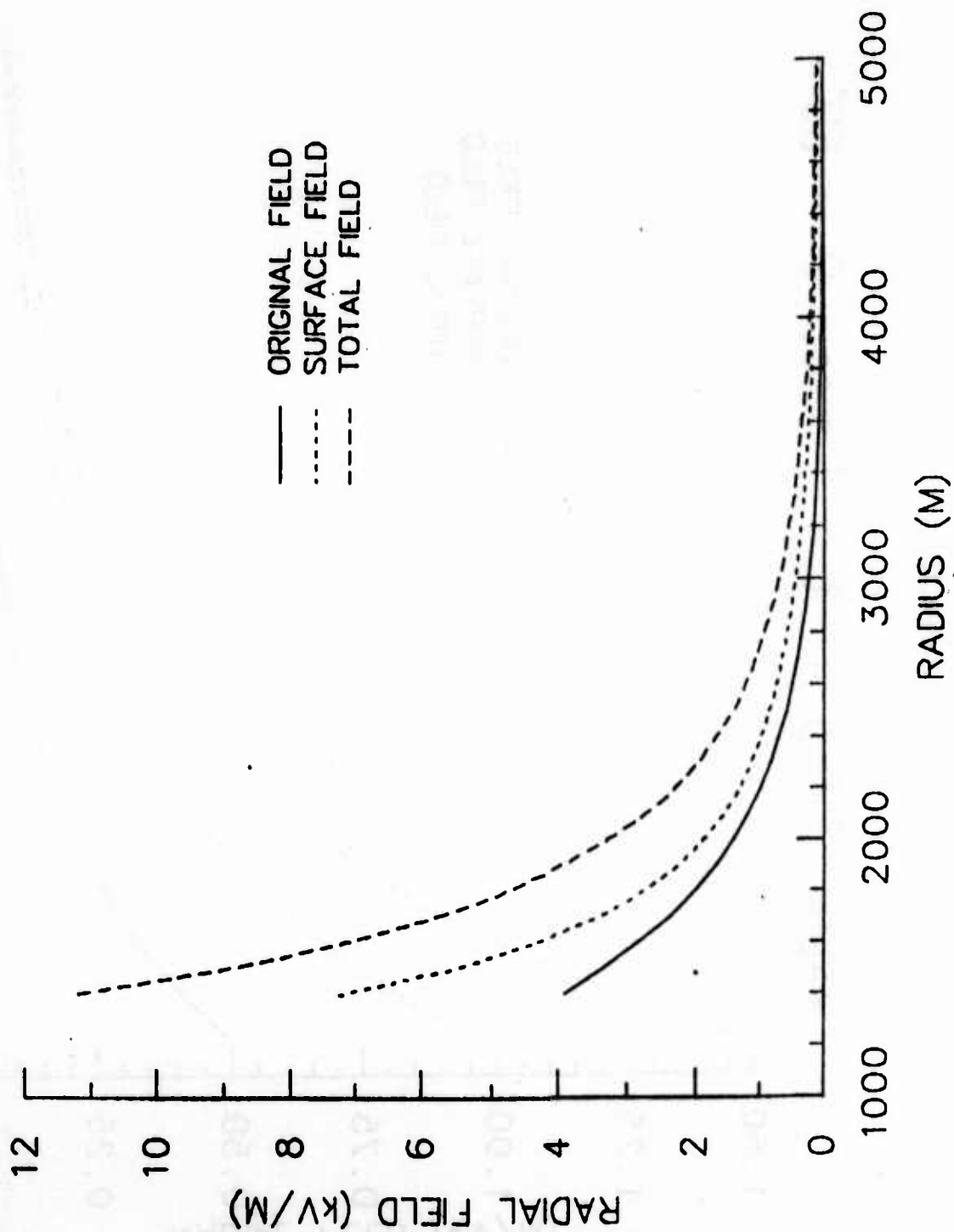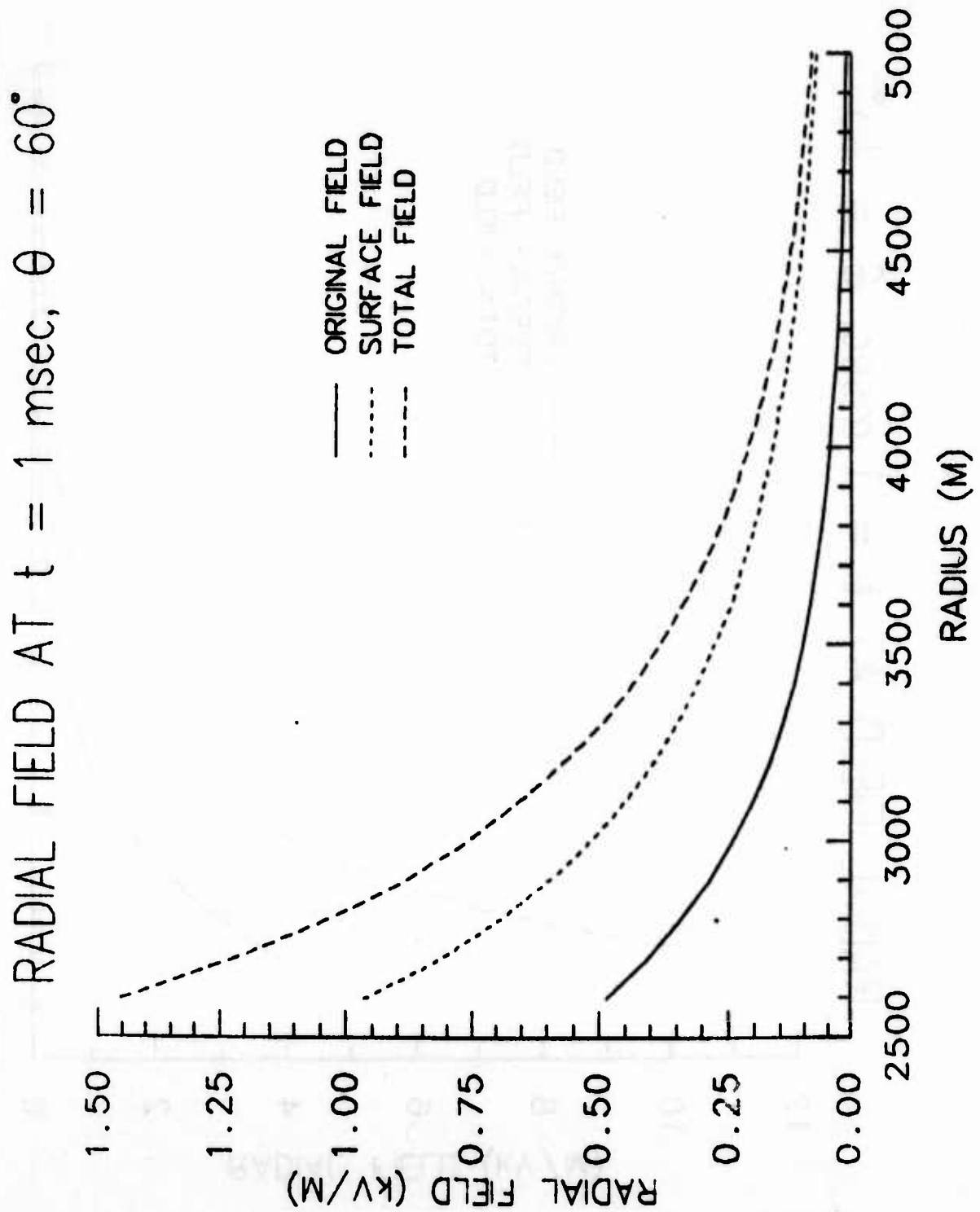
Figure 24

POTENTIALS AT t = 1 msec, θ = 45°

.... ORIGINAL POTENTIAL
------- SURFACE CHARGE POTENTIAL
——— TOTAL POTENTIAL

RADIUS (M)

POTENTIAL (kV)

427
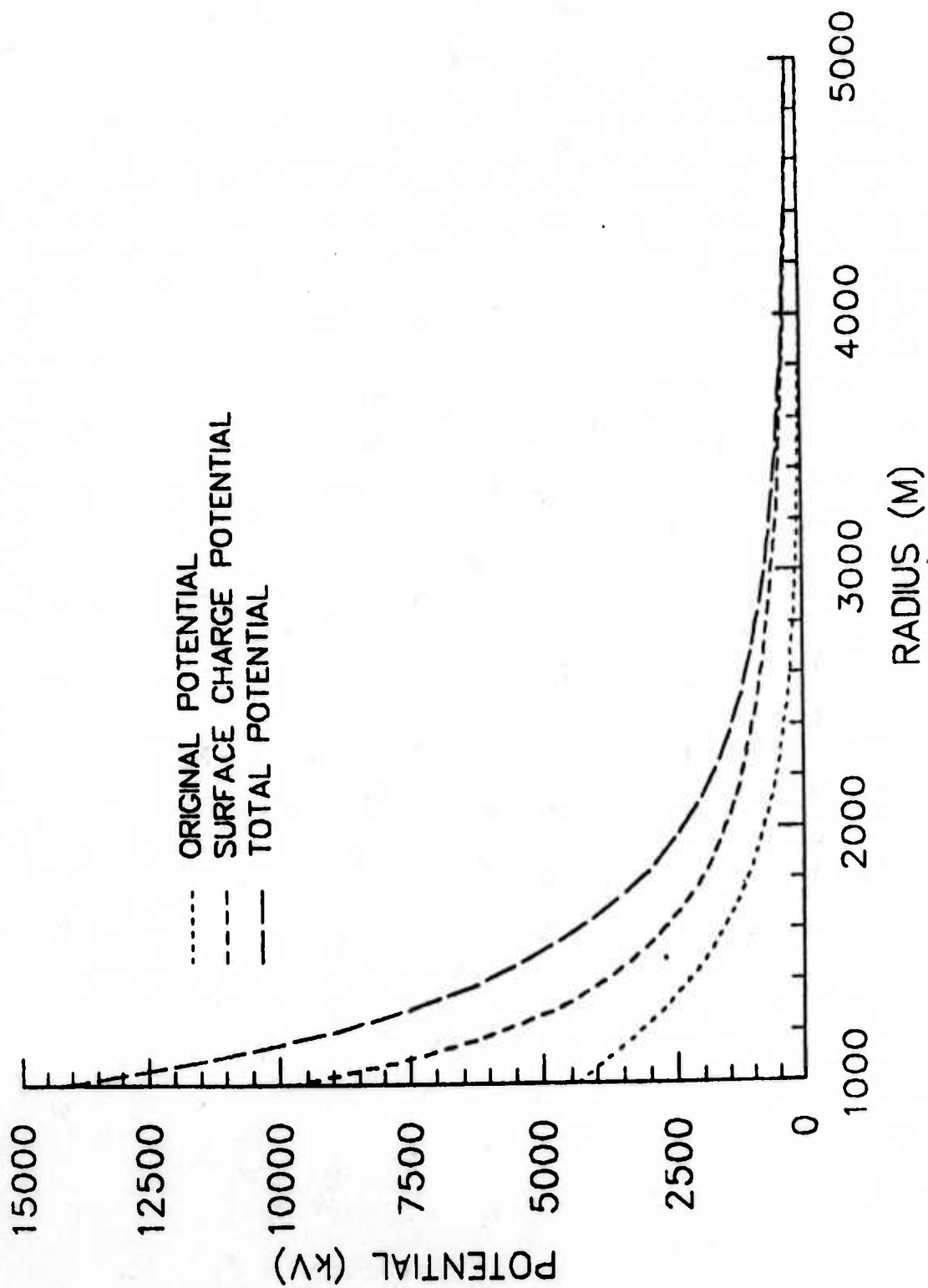
ON FATIGUE LIFE PREDICTION IN THICK-WALLED CYLINDERS

S. L. Pu and P. C. T. Chen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

ABSTRACT. The large variation in stress intensity factors corresponding to various material models for a single, radial, straight-fronted crack in a pressurized, partially autofrettaged cylinder leads to a drastic difference in the fatigue life predictions. None of the predicted lives agree with experimental results. Possible explanations of the discrepancy are given and corresponding correction factors are introduced. The predicted lives based on the corrected stress intensity ranges are reasonably close to a set of well-documented experimental results of Throop and Fujczak.

I. INTRODUCTION. Both finite element and modified mapping collocation methods have been used to obtain accurate stress intensity (K) solutions for pressurized autofrettaged thick cylinders with radial cracks [1,2]. The use of weight function has extended the two-dimensional K solutions to more refined material models including the reverse yielding caused by high Bauschinger effect [3]. Several papers have used the stress intensity factors to estimate fatigue lives of cannon tubes [4-6]. The calculations underestimate the measured lives for pressurized cylinders, while they overestimate the experimental results for pressurized and autofrettaged cylinders [5]. The disagreement between measured and calculated lives diminishes in [6] by introducing a fraction of the negative portion of K values as a part of the K range.

The shallow crack approximations for K solutions were used and a linear approximation for Bauschinger effect on residual hoop stress was assumed in [5] and [6]. The accurate two-dimensional K solutions affected by a significant Bauschinger effect using elastic-plastic analysis were used in [7] to indicate the drastic effect of reverse yielding on the life prediction. Neither shape factors nor a fraction of negative K were considered in [7].

In this paper the life prediction formula similar to that used in [7] is employed to check the calculated lives with experimental results of Throop [8] and Throop and Fujczak [9]. The stress intensity factors for surface cracks of elliptical shape are approximated by the two-dimensional stress intensity factors obtained in [3] multiplied by respective shape factors for pressure and for residual stress given in [5]. The fraction of negative K included in K range varies from one at the notch boundary to zero at a crack depth far away from the notch. The variation of this fraction is assumed to be $1/r^2$ where r is the distance between the fatigue crack front and the notch front. Another modification is to use an initial crack depth much deeper than the notch depth. This is to avoid large cycles required in experiments to initiate a single continuous crack front along the notch boundary (it is considered likely [5] that multiple, small, semi-elliptical cracks are initiated

along the notch boundary prior to their link together to form a single crack).
With these modifications the calculated lives agree reasonably well with
experimental results for all three crack shapes: long curves, semi-elliptical,
and semi-circular used in [8] and [9].

II. FATIGUE LIFE PREDICTION. The integration of Paris' formula

$$\frac{da}{dN} = C(\Delta K)^m \tag{1}$$

for fatigue growth rate of a crack subjected to cyclic loading is usually used
to determine the number of fatigue cycles required to grow a crack from an
initial depth a_i to a final depth a_f

$$N = N_f - N_i = \int_{a_i}^{a_f} \frac{da}{C(\Delta K)^m} \tag{2}$$

where C and m are material constants and ΔK is the range of stress intensity
defined by

$$\Delta K = K_{max} - K_{min} \tag{3}$$

K_{max} and K_{min} are maximum and minimum values of K in a loading cycle. Assume
that a crack face is a geometric plane and there is no possibility of inter-
penetration under compression. This leads to a conclusion that K_{min} cannot
be a negative value. In the case of repeated firing of cannon tubes

$$K_{min} = 0 \tag{4}$$

is used for both autofrettaged and nonautofrettaged tubes. If K_p and K_R
denote mode I K values corresponding to an internal pressure and a residual
hoop stress respectively, then

$$K_{max} = K_p \tag{5a}$$

for nonautofrettaged cylinders, and

$$K_{max} = K_p + K_R \tag{5b}$$

for autofrettaged cylinders. K_p and K_R are usually expressed in a dimen-
sionless form denoted by f_{Kp} and f_{KR}, respectively,

$$f_{Kp} = \frac{K_p}{p\sqrt{\pi a}} \quad , \quad f_{KR} = \frac{K_R}{\sigma_0\sqrt{\pi a}} \tag{6}$$

where σ_0 is the yield stress, p is the internal pressure applied to a tube,
and p is related to σ_0 by a load factor $f_L = \sigma_0/p$. By virtue of Eqs. (3),
(4), and (5b)

$$\Delta K = (\frac{f_{Kp}}{f_L} + f_{KR})\sigma_0\sqrt{\pi a} \tag{7}$$

For a small crack growth from a_j to a_{j+1}, f_{Kp} and f_{KR} are assumed to be constants which are taken as the mean values

$$\bar{f}_{Kp} = \frac{1}{2}(f_{Kp}(a_j) + f_{Kp}(a_{j+1}))$$

$$\bar{f}_{KR} = \frac{1}{2}(f_{KR}(a_j) + f_{KR}(a_{j+1}))$$

(8)

The integration of Eq. (2) becomes

$$\frac{N}{C_o} = \frac{N_f - N_i}{C_o} = \sum_{j=1}^{n} 2^{-m}(\frac{f_{Kp}}{f_L} + \bar{f}_{KR})^{-m}(\alpha_j^{1-m/2} - \alpha_{j+1}^{1-m/2})$$

(9)

where

$$C_o = \frac{2}{C(m-2)} (\frac{\sigma_o \sqrt{\pi}}{2})^{-m}(t)^{1-m/2}$$

(10)

with t denoting the wall thickness and $\alpha = a/t$. The fraction α is used since the crack depth is usually expressed as a fraction of wall thickness t.

Substituting from calculated values of f_{Kp} and f_{KR} [3] corresponding to various material models in Eqs. (8) and (9), we obtain the fatigue crack growth (α versus N/C_o) graph shown in Figure 1 for a single, two-dimensional straight-fronted through crack in a cylinder of diameter ratio two. In this figure, f is the Bauschinger effect factor. The dotted lines are for idealized material without Bauschinger effect (f = 1), while the dashed lines are for f = 0.38 (100 percent overstrain) and f = 0.44 (60 percent overstrain), respectively. The dashed lines with m' = 0 correspond to elastic-perfectly plastic behavior during reverse yielding. The dashed lines with m' = 0.3 indicate the difference in predicted cycles when strain-hardening is considered in reverse yielding. The graph shows the significant difference of autofrettage effect of various material models on fatigue cycles. Such a drastic difference is not supported by experimental results. Corrective parameters must be introduced to correlate the calculated fatigue cycles with observed ones in laboratory tests.

III. SHAPE FACTORS. The ratio of stress intensity factor for a surface crack to that of a two-dimensional through crack is called the shape factor. The stress intensity factor varies along the crack front of a surface crack. It changes with crack shape and under different loadings. If the variation of stress intensity along the crack front is important, the three-dimensional K solutions for the surface crack should be obtained. If an estimate of K at a point, say the deepest point of the surface crack, is needed, then a reasonable estimate of shape factor is useful. An extensive study of shape factors has been published by Newman and Raju [10] for semi-elliptical cracks in a flat plate under tension or bending. There is no comparable study for such a crack in a thick cylinder. Parker et al obtained estimates of shape factors for semi-elliptical cracks of various aspect ratios in a pressurized and auto-frettaged cylinder [5] from judicious use of results reported in [10]. More accurate three-dimensional K solutions should be obtained to check these esti-mates. Before more accurate and reliable shape factors become available, values close to these given by (a) and (c) of Figure 7 in [5] are used for f_{Sp}

431

and f_{SR}, respectively, in this study. Multiplying f_{Kp} and f_{KR} with their shape factors, Eq. (7) becomes

$$\Delta K = (\frac{f_{Sp}f_{Kp}}{f_L} + f_{SR}f_{KR})\sigma_0\sqrt{\pi a} \tag{11}$$

The following mean values are used in Eq. (11) for a small crack growth from a_j to a_{j+1}.

$$\overline{f_{Sp}f_{Kp}} = \frac{1}{2}[f_{Sp}(a_j)f_{Kp}(a_j) + f_{Sp}(a_{j+1})f_{Kp}(a_{j+1})]$$
$$\overline{f_{SR}f_{KR}} = \frac{1}{2}[f_{SR}(a_j)f_{KR}(a_j) + f_{SR}(a_{j+1})f_{KR}(a_{j+1})] \tag{12}$$

IV. NEGATIVE K FACTOR F_N. For a fatigue crack which is considered as a geometrical plane with no thickness, $K_{min} = 0$ is used in Eq. (3) to obtain K range. For a notch with finite thickness, the upper and lower planes of the notch may have normal displacement in both directions. The argument used to limit $K_{min} = 0$ is not valid at the notch front. In fact, the full negative K should be used for a crack starting at the notch. At some depth, the notch effect may become negligibly small, and $K_{min} = 0$ may again be used in Eq. (3) for ΔK. Kendall has argued that a portion of the negative K must be included in calculating K range [6]. He introduced a constant fraction which times the negative K (K_R) coresponding to compressive residual stress due to autofrettage to give the K_{min} in ΔK. From our hypothesis, the fraction varies with the depth of the fatigue crack. As a first approximation, the fraction f_N is taken as $(1 + r/r_n)^{-2}$ where r_n is the depth of the notch and r is the depth of the crack measured from the notch front. At the notch front $r = 0$ and $f_N = 1$, the full negative K is taken as K_{min}. When $r = r_n$ the fatigue crack grows to a depth equal to the notch depth; this approximation gives $f_N = \frac{1}{4}$. It is also assumed that $f_N = 0$ when $r > 2r_n$. A linear variation of f_N was another approximation examined, it underestimated the measured fatigue lives. Incorporating f_N, Eq. (11) becomes

$$\Delta K = [f_{Sp}\frac{f_{Kp}}{f_L} + (1-f_N)f_{SR}f_{KR}]\sigma_0\sqrt{\pi a} \tag{13}$$

Equation (9) can be expressed explicitly

$$\frac{N}{C_0} = \sum_{j=1}^{n}[\frac{\tilde{f}_{Kp}(\alpha_j) + \tilde{f}_{Kp}(\alpha_{j+1})}{f_L} + \tilde{f}_{KR}(\alpha_j) + \tilde{f}_{KR}(\alpha_{j+1})]^{-m}(\alpha_j^{1-m/2} - \alpha_{j+1}^{1-m/2}) \tag{14}$$

where abbreviations $\tilde{f}_{Kp}$, $\tilde{f}_{KR}$ are

$$\tilde{f}_{Kp}(\alpha_j) = f_{Sp}(\alpha_j)f_{Kp}(\alpha_j)$$
$$\tilde{f}_{KR}(\alpha_j) = [1 - f_N(\alpha_j)]f_{SR}(\alpha_j)f_{KR}(\alpha_j) \tag{15}$$

432

V. LABORATORY SPECIMENS AND MEASURED LIVES.

Throop [8] and Throop and Fujczak [9] obtained a series of experimental results which relates fatigue crack lives with crack shape and extent of autofrettage for thick-wall cylinders. The cylinders were 0.76 m (30 inches) in length, 180 mm (7.1 inches) bore diameter, 360 mm (14.25 inches) outside diameter and were fatigue cracked from longitudinal internal notches. Three initial notch geometries were used; semi-circular, 100 mm (4-inch) and 500 mm (20-inch) long notches produced by electrical discharging machining. They were 6.4 mm (¼-inch) deep by 0.76 mm (0.03-inch) wide, the semi-circular notch being 13 mm (½-inch) diameter half-penny shape. Fatigue cracks grown from the initial notches were monitored periodically for depth and shape with ultrasonics as the cylinder was repeatedly pressurized to 330 MPa (48 Ksi). The cylinder material was ASTM A723 forged steel, with yield strength of 1175 MPa, -40°C Charpy impact energy of 34 J, reduction in area of 50 percent. A schematic diagram of a typical cylinder with a simple initial notch and the growth of a 500 mm (20-inch) long notch from 6.4 mm (¼-inch) initial depth by repeated pressurization is shown in Figure 2. The measured crack depth versus corresponding number of fatigue loadings is shown in Figure 3 for a single-notched cylinder with 0, 30, and 60 percent overstrains for three-notch geometries. Figures 2 and 3 are reproduced from Figures 1 and 6 of [9], respectively.

Since the fatigue crack is not likely to start from the notch immediately, an adjustment of experimental results is made by subtracting the number of cycles required to grow from initial notch depth to an initial crack depth (a_i) from the number of cycles to grow from initial notch depth to a final crack depth (a_f). The initial crack depth, which should be reasonably larger than the initial notch depth (6.4 mm in this experimental study), is arbitrarily taken as $a_i = 0.1t = 9$ mm. The original experimental data are available only in graphs. To improve the accuracy of readings from graphs, the graphs were first enlarged and the average was used from values obtained from different graphs published in different papers [4,5,9] for the same set of experimental data. The adjusted experimental results are shown by dots in Figures 4, 5, 6, and 7.

VI. PREDICTED LIVES FOR THE EXPERIMENTS.

The material constants for the steel used in the experiments are $m = 3$ and $C = 6.52 \times 10^{-12}$ for crack growth in meters per cycle and ΔK in MPa$\sqrt{\text{meter}}$ (or $C = 3.4 \times 10^{-10}$ for crack growth in inches per cycle and ΔK in Ksi$\sqrt{\text{inch}}$). Using $f_L = 3.55$ and values of f_{Sp} and f_{SR} in Table I, and assuming $f_N = (1 + r/r_n)^{-2}$ for $0 < r < 2r_n$ and $f_N = 0$ for $r > 2r_n$, Eqs. (14) and (15) can be evaluated with known discrete values of f_{Kp} and f_{KR}. Values of f_{Sp} and f_{SR}, given in Table I, are obtained from [5] for different crack geometries. Discrete values of f_{Kp} and f_{KR} can be computed by the method described in [3]. The calculated lives and measured lives are plotted in Figures 4, 5, and 6 for three crack geometries, respectively. For nonautofrettaged cylinders (zero percent overstrain), there is only one set of calculated results, shown in a solid line, for each crack configuration. For 60 percent overstrain, solid lines are for idealized material without considering the Bauschinger effect, while two dashed lines in each figure are predicted lives with some reverse yielding. The two dashed lines differ in m' values, m' = 0 or 0.3, where m'E is the slope of strain-hardening during reverse yielding.

When m' = 0, the material behaves like elastic-perfectly plastic in reverse yielding. The magnitude of the compressive, residual stress in the tangential direction near the bore is smaller than that in a strain-hardening material [3]. The strain-hardening reduces the adverse effect of reverse yielding due to the Bauschinger effect, Figure 1. Figures 4 through 6 show an overall effect of all factors. Figure 7 shows the effect of each factor on the semi-elliptical surface crack in a pressurized cylinder with 60 percent overstrain. The dotted curve, curve 1, gives the calculated lives for an idealized material with no Bauschinger effect. No correction factors are used. The stress intensity range is based on two-dimensional stress intensity factors and $K_{min} = 0$. If the change in residual hoop stress due to the Bauschinger effect (Bauschinger effect factor f = 0.44) is considered, the calculated lives are shown as a dashed curve, curve 2, where m' = 0.3 is used. The large difference between curves 1 and 2 shows the significant effect of reverse yielding on fatigue lives. If shape factors f_{Sp} and f_{SR} are considered in addition to the Bauschinfer effect, the predicted lives will be changed from curve 2 to curve 3. The shape factors are to increase the fatigue lives. Finally, if the correction factor f_N is also taken into consideration, the predicted lives are shown as the solid line, curve 4, which is close to the experimental results shown by dots in Figure 7. Similar graphs, obtained for the long curved crack and the semi-circular crack, are omitted since they indicate similar effects of each correction factor.

TABLE I. VALUES OF f_{Sp} AND f_{SR} FOR DIFFERENT CRACK GEOMETRIES

$\alpha = a/t$	20-Inch Long Notch		4-Inch Long Notch		Semi-Circular Notch	
	f_{Sp}	f_{SR}	f_{Sp}	f_{SR}	f_{Sp}	f_{SR}
0.1	0.95	0.95	0.72	0.69	0.57	0.58
0.2	0.90	0.90	0.64	0.60	0.53	0.52
0.3	0.85	0.85	0.61	0.55	0.51	0.42
0.4	0.85	0.80	0.57	0.47	0.48	0.34
0.5	0.85	0.75	0.56	0.38	0.44	0.24

VII. CONCLUSIONS. The fatigue life of a thick-walled cylinder can be predicted reasonably well by the integration of Paris' formula of crack growth rate of a fatigue crack under cyclic loading, if a modified stress intensity range is used. The stress intensity range is obtained by multiplying two-dimensional stress intensity factors by proper shape factors and some negative K factors. More systematic and controlled experiments are required to check the idea proposed in this paper. Three-dimensional K solutions for semi-elliptical surface cracks in thick cylinders with various residual stresses are needed to estimate the proper shape factors. Special experiments should be performed to verify the concept of negative K factors and to determine the variation of f_N in terms of crack depth.

REFERENCES

1. Parker, A. P., "Stress Intensity and Fatigue Crack Growth in Multiply Cracked, Pressurized, Partially Autofrettaged Thick Cylinders," _Fatigue of Engineering Materials and Structures_, Vol. 4, No. 2, 1982.

2. Pu, S. L. and Hussain, M. A., "Stress-Intensity Factors For Radial Cracks in a Partially Autofrettaged Thick-Wall Cylinder," _Fracture Mechanics: Fourteenth Symposium - Volume I: Theory and Analysis_, _ASTM STP 791_, 1983, pp. I-194-I-215.

3. Pu, S. L. and Chen, P. C. T., "The Bauschinger Effect on Stress Intensity Factors For a Radially Cracked Gun Tube," Transactions of the Third Army Conference on Applied Mathematics and Computing, Georgia Institute of Technology, ARO Report 86-1, 1986, pp. 275-294.

4. Underwood, J. H. and Throop, J. F., "Surface Crack K-Estimates and Fatigue Life Calculations in Cannon Tubes," _Part-Through Crack Fatigue Life Prediction_, _ASTM STP 687_, 1979, pp. 195-210.

5. Parker, A. P., Underwood, J. H., Throop, J. F., and Andrasic, C. P., "Stress Intensity and Fatigue Crack Growth in a Pressurized, Autofrettaged Thick Cylinder," _Fracture Mechanics: Fourteenth Symposium - Volume I: Theory and Analysis_, _ASTM STP 791_, 1983, pp. I-216-I-237.

6. Kendall, D. P., "A Simple Fracture Mechanics Based Method for Fatigue Life Prediction in Thick-Walled Cylinders," Technical Report ARLCB-TR-84023, Benet Weapons Laboratory, Watervliet, NY, July 1984.

7. Pu, S. L. and Sha, G. T., "The Bauschinger Effect of Yield Stress Reduction on Radial Crack Growth of a Cylindrical Pressure Vessel," submitted to _Engineering Fracture Mechanics_.

8. Throop, J. F., "Fatigue Crack Growth in Thick-Walled Cylinders," Proceedings of National Conference on Fluid Power, Vol. XXVI, 1972, pp. 115-131.

9. Throop, J. F. and Fujczak, R. R., "Strain Behavior of Pressurized Cracked Thick-Walled Cylinders," _Experimental Mechanics_, Vol. 22, 1982, pp. 277-286.

10. Newman, J. C. and Raju, I. S., "Analyses of Surface Cracks in Finite Plates Under Tension or Bending Loads," NASA TP 1578, National Aeronautics and Space Administration, Washington D.C., 1979.
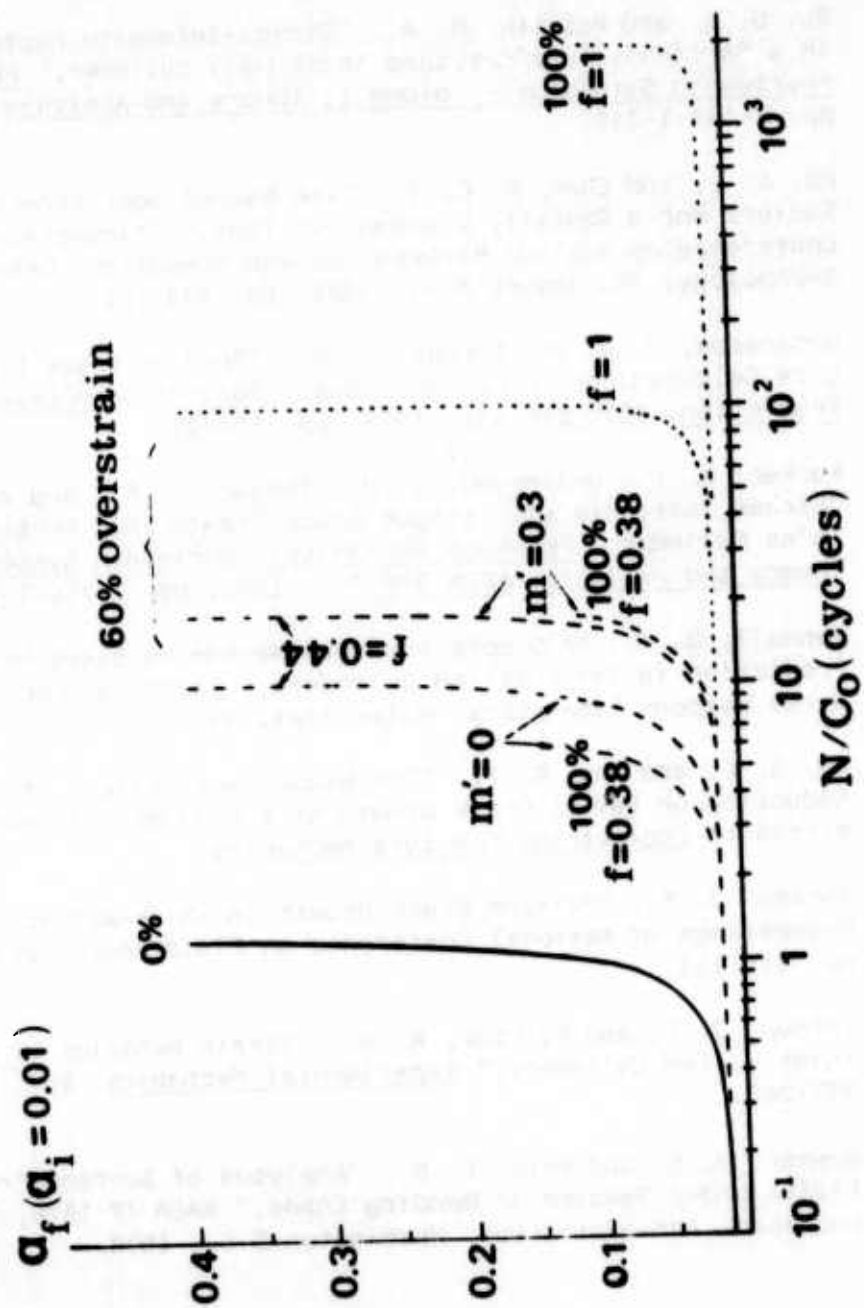
Figure 1
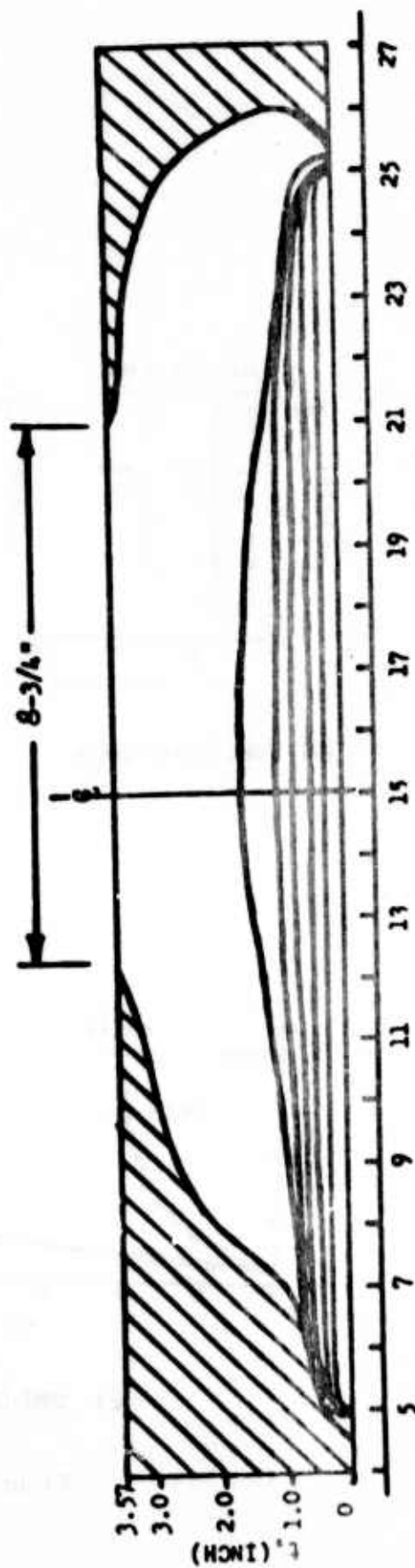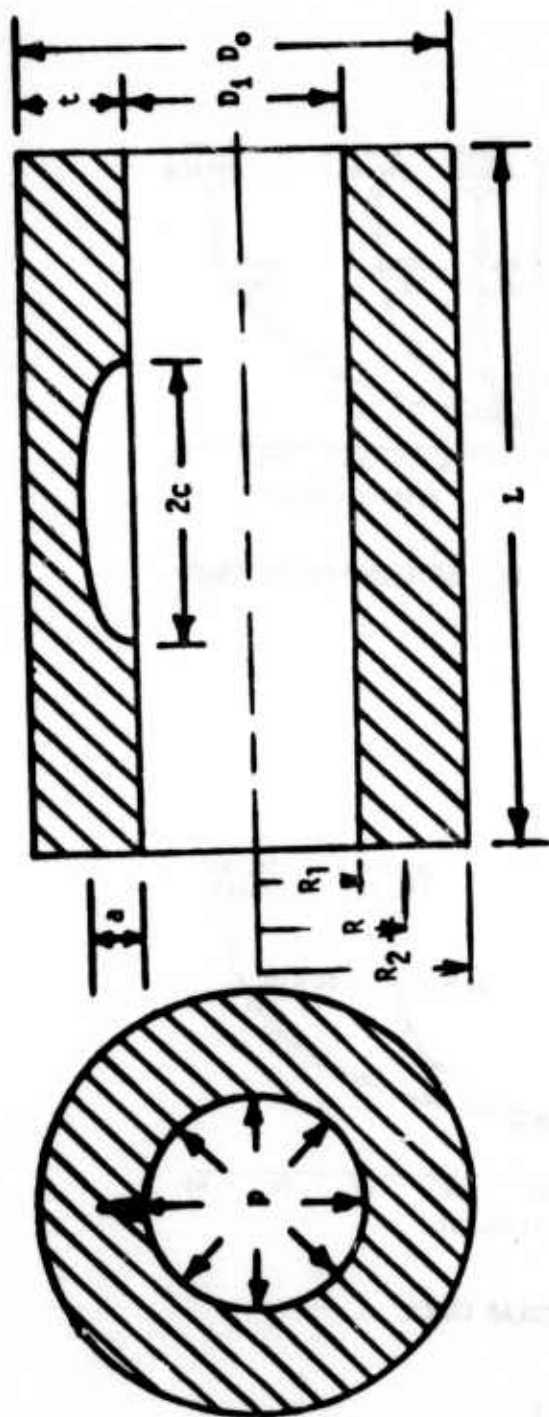
DISTANCE FROM TOP OF CYLINDER. (INCH)

Figure 2

437
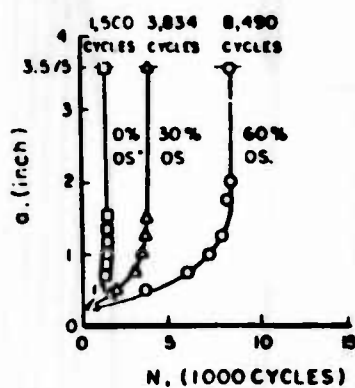
(a) LONG CURVED CRACK
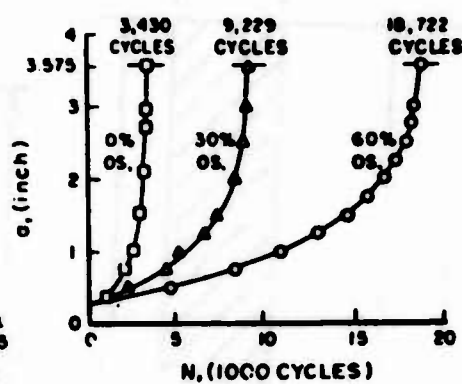
(b) SEMI-ELLIPTICAL CRACK
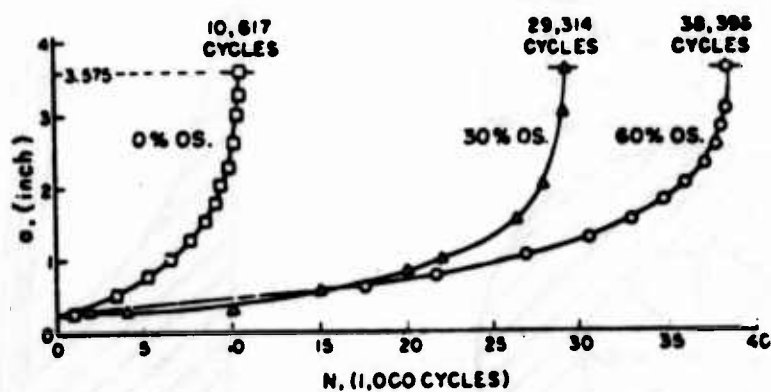
(c) SEMI-CIRCULAR CRACK

Figure 3
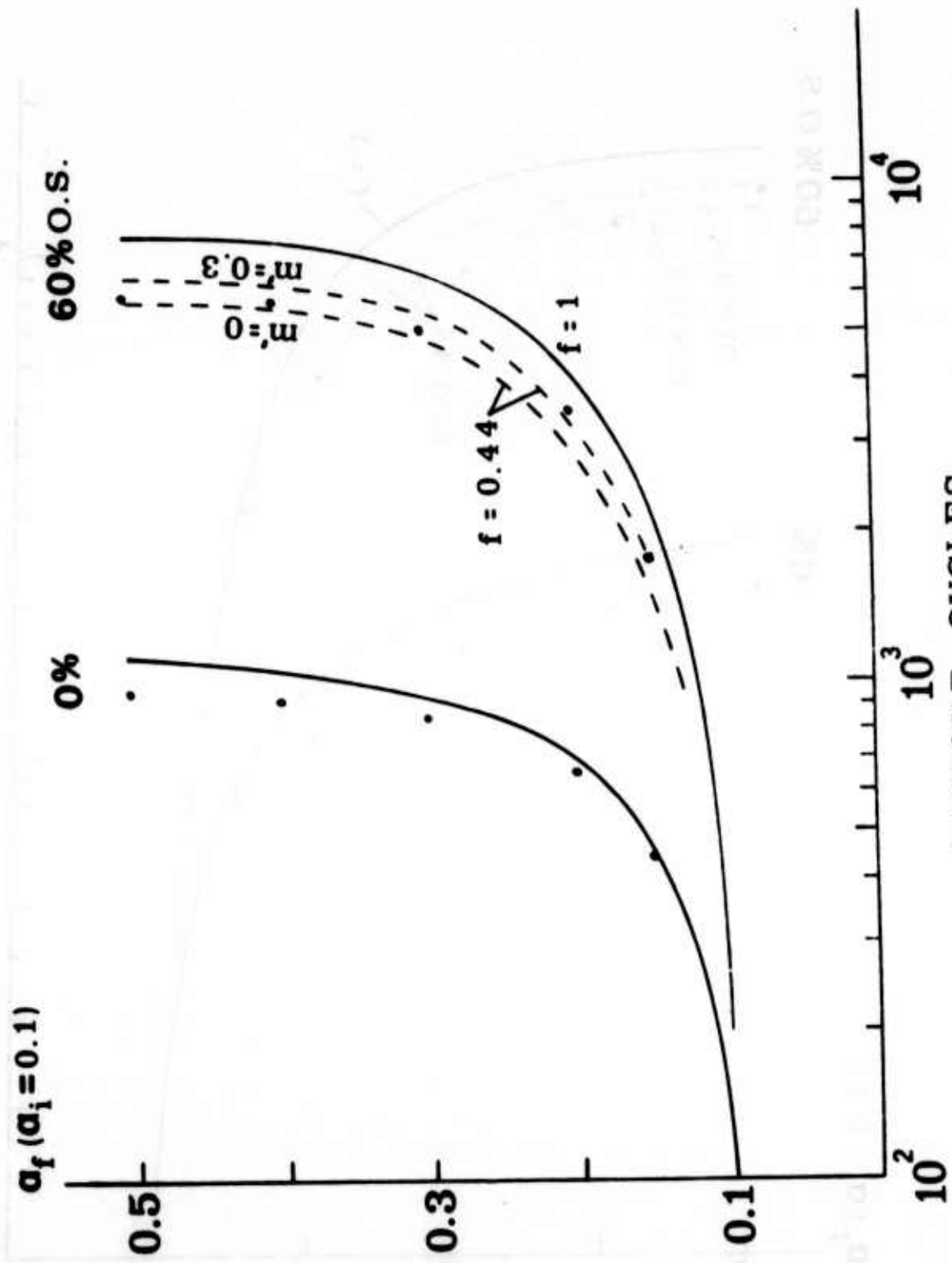
438

FATIGUE CYCLES

Figure 7

439

FATIGUE CYCLES
Figure 5

FATIGUE CYCLES

Figure 5

441

FATIGUE CYCLES
Figure 7
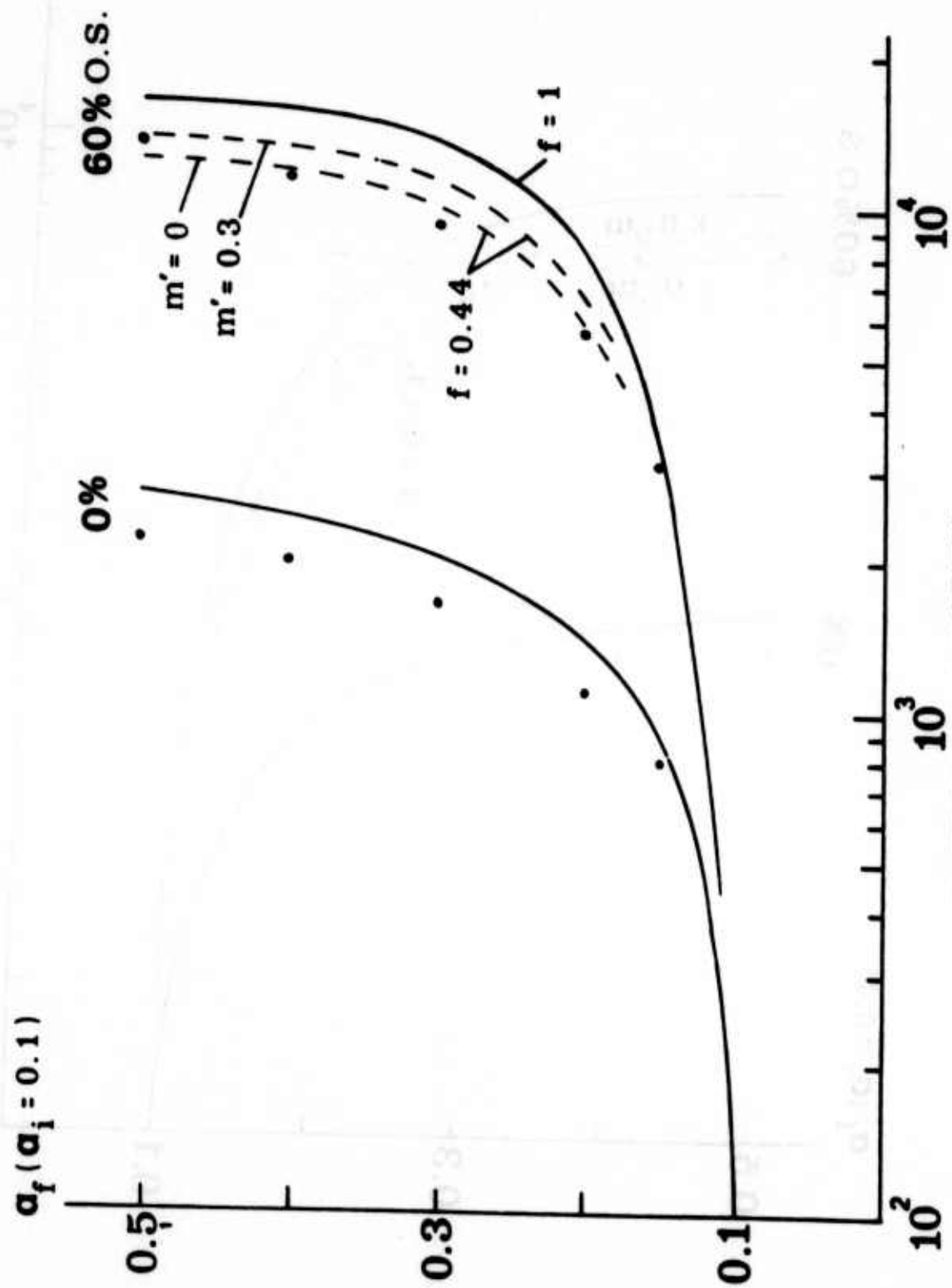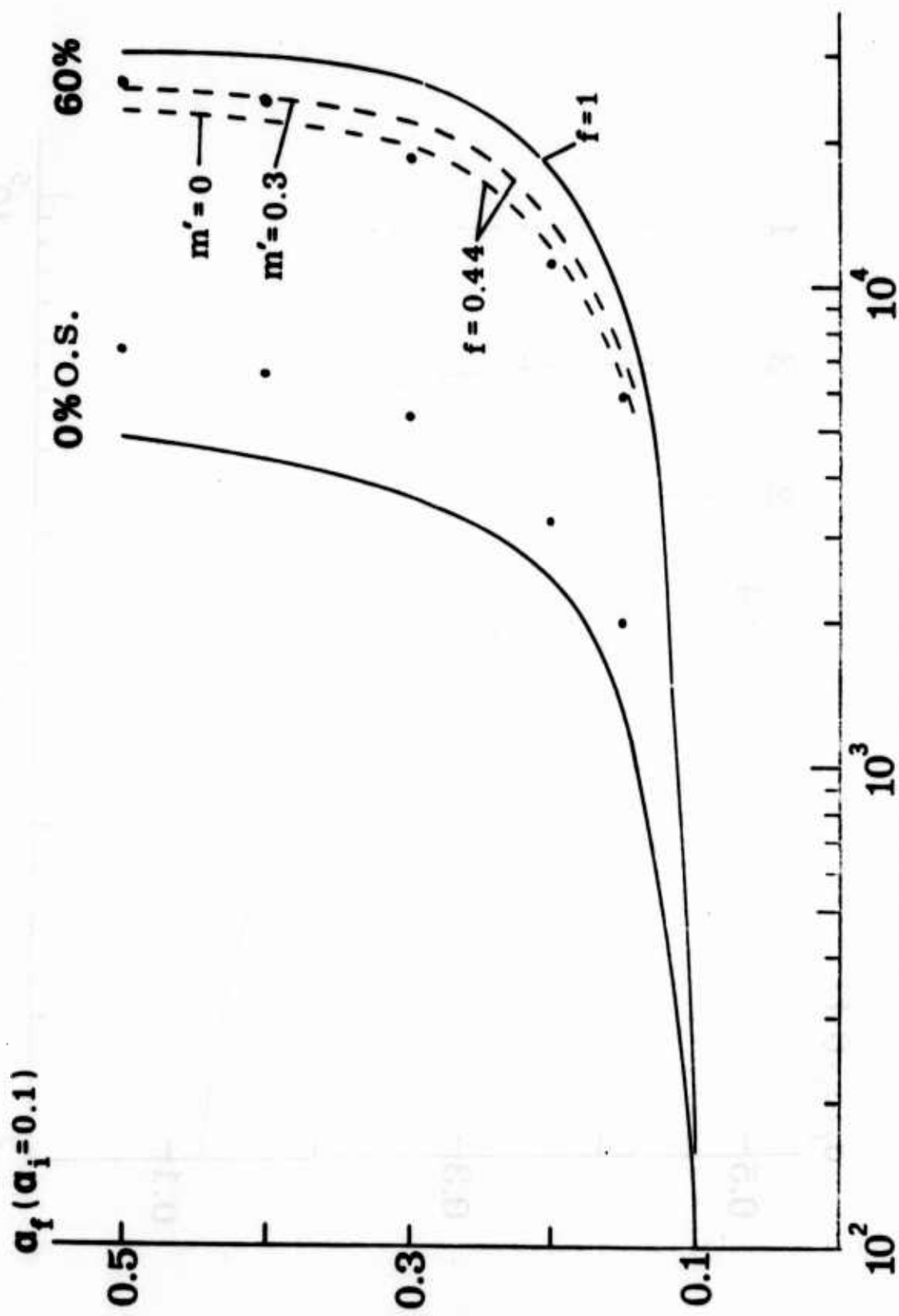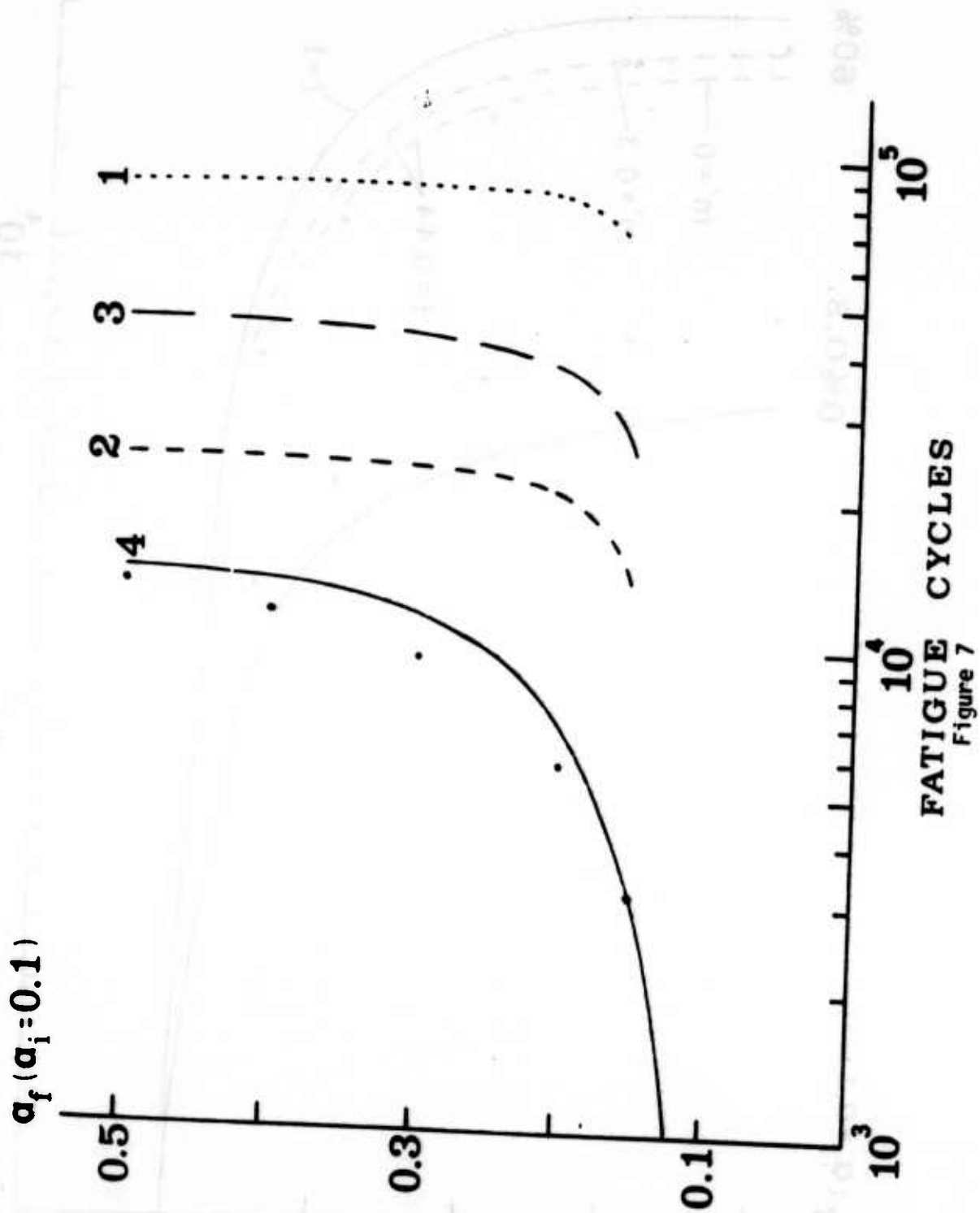
ANALYSIS OF COMPOSITE SHRINK FITS - TRESCA MATERIAL

Peter C. T. Chen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

ABSTRACT. A thin composite shrink fit assembly is examined using an elastic-plastic analysis. The ring and disk are made of different materials. Interferences large enough to induce plastic deformations in the ring are accounted for. The ring material is assumed to be a linear strain-hardening material that obeys Tresca's yield condition. The explicit expressions for stresses and deformations in the shrink fit assembly have been obtained. Numerical results are presented for shrink fit assemblies with different geometric ratio, hardening parameter, and different combinations of materials.

I. INTRODUCTION. The shrink fit fastening process is widely used in industry to produce tight, precision assemblies where other fastening methods are neither necessary nor practical. By shrinking a thin ring onto a disk of the same thickness, an elastic state of biaxial, hydrostatic stress can be induced in the disk. For sufficiently small values of interference of the fit, the ring and disk remain elastic; for large values of interference, the ring becomes plastic, first at the interference; for yet larger values of inter- ference, it is possible to produce a plastic state in the disk. This problem was analyzed recently by Gamer and Lance [1] considering the same materials for the disk and ring.

In this paper we shall examine a thin composite shrink fit assembly using a plane-stress elastic-plastic analysis. The ring and disk are made of dif- ferent materials. Interferences large enough to induce plastic deformations in the ring are accounted for. The ring material is assumed to be a linear strain-hardening material that obeys Tresca's yield condition and the asso- ciated flow rule. The stresses and deformations in the shrink fit assembly are to be obtained as functions of the interference of the fit.

II. ELASTIC ASSEMBLY. A shrink fit assembly is shown in Figure 1. The assembly may be produced by cooling the disk and/or heating the ring with the manufactured interference I. The common interference radius of the assembly is a. The thickness, h, is small compared to a, and hence, the state of stress may be assumed to be plane. All thermal effects are neglected and the displacement is assumed to be small everywhere.

For small values of interference of fit, the stress state in the entire assembly is elastic. The stresses and displacements in the ring are

$$
\begin{matrix} \sigma_r \\ \\ \sigma_\theta \end{matrix} \quad = \quad \frac{P}{1 - a^2/b^2} \left[\frac{a^2}{b^2} \mp \frac{a^2}{r^2} \right]
$$

(1a)

(1b)

$$u/r = (P/E)[(1+\nu)(a^2/r^2) + (1-\nu)(a^2/b^2)]/(1-a^2/b^2) \tag{1c}$$

and in the disk

$$\sigma_r = \sigma_\theta = -P \quad , \quad u/r = -(1-\nu_1)P/E_1 \tag{2}$$

where E, ν and E_1, ν_1 are the material constants of the ring and disk, respectively. At the interface, u_a (ring) $- u_a$ (disk) $= I$ by the compatibility requirement. The interference pressure (p) is a function of the interference (I) given by

$$p = \frac{EI}{a}(1 - \frac{a^2}{b^2})/[(1+\nu) + (1-\nu)\frac{a^2}{b^2} + (1-\nu_1)(1 - \frac{a^2}{b^2})\frac{E}{E_1}] \tag{3}$$

For sufficiently large values of the interference the stresses in the ring reach the yield limit. Assuming that Tresca's yield condition governs the behavior of the material, the ring first becomes plastic at the interference when the stresses satisfy

$$\sigma_\theta - \sigma_r = \tag{4}$$

where σ_0 is the initial tensile yield stress. The solution for the critical interference pressure to cause incipient plastic deformation is

$$p^\star = \tfrac{1}{2}\,\sigma_0(1 - a^2/b^2) \tag{5}$$

and it follows from Eq. (3) that the interference for the onset of plastic flow is

$$I^\star = \frac{\sigma_0}{E}\frac{a}{2}[(1+\nu) + (1-\nu)\frac{a^2}{b^2} + (1-\nu_1)(1 - \frac{a^2}{b^2})\frac{E}{E_1}] \tag{6}$$

which reduces to $I^\star = a\sigma_0/E$ for the special case ($E_1 = E$, $\nu_1 = \nu$) considered in [1].

III. PARTIALLY PLASTIC ASSEMBLY. For values of interference larger than that given by Eq. (4), a plastic zone forms in the ring, so that for $a \leqslant r \leqslant \rho$ the ring is plastic, while for $\rho \leqslant r \leqslant b$, the ring material is still in an elastic state. The elastic-plastic interface radius ρ is a function of the interference I.

We assume that the ring is made of a linear work-hardening material which obeys Tresca's yield condition

$$\sigma_\theta - \sigma_r = \sigma \tag{7}$$

where the yield stress σ is a function of the plastic strain ϵ^p. For a linear work-hardening material, we have

$$\sigma = \sigma_0(1+\eta\epsilon^p) \quad \text{and} \quad \eta = (E/\sigma_0)m/(1-m) \tag{8}$$

where η (or m) is the hardening parameter.

444

Applying the usual flow rule and following the method of analysis reported by Gamer and Lance [1] and Bland [2], the expressions for the stresses and the displacement can be obtained explicitly. The complete solution in $a \leqslant r \leqslant \rho$ is:

$$\sigma_r = \sigma_0(1-m)[\ln \frac{r}{a} - \tfrac{1}{2}(1-\nu)\eta \frac{D}{E} r^{-2}] + C \tag{9}$$

$$\sigma_\theta = \sigma_0(1-m)[1+\ln \frac{r}{a} + \tfrac{1}{2}(1-\nu)\eta \frac{D}{E} r^{-2}] + C \tag{10}$$

$$(1-\nu)^{-1}u = \frac{\sigma_0}{E}(1-m)[r\ln \frac{r}{a} - \tfrac{1}{2}(1-\nu)\eta \frac{D}{E} r^{-1}] + \frac{C}{E} r + \frac{D}{E} r^{-1} \tag{11}$$

In the elastic zone, $\rho \leqslant r \leqslant b$, the stresses and the displacement are:

$$\begin{matrix} \sigma_r \\ \sigma_\theta \end{matrix} = \frac{E}{1-\nu} A \mp \frac{E}{1+\nu} \frac{B}{r^2} \tag{12} \tag{13}$$

$$u = Ar + B/r \tag{14}$$

The constants A, B, C, D, p, and ρ all depend on the interference I, and can be evaluated by considering the following conditions: continuity of stress and displacement at $r = \rho$ requires $\sigma_r(\rho^-) = \sigma_r(\rho^+)$ and $u(\rho^-) = u(\rho^+)$. At the ring-disk interface $\sigma_r(a) = -p$ and at the outer surface of the ring $\sigma_r(b) = 0$. The yield condition in Eq. (7) must be satisfied at $r = \rho$ and finally, compatibility of the displacement field with the interference I requires that $u(a^+) - u(a^-) = I$. These conditions are sufficient to determine all unknown parameters. In this paper the constants A, B, C, D are determined as functions of ρ.

$$A = \tfrac{1}{2}(1-\nu)(\sigma_0/E)(\rho/b)^2 \quad , \quad B = \tfrac{1}{2}(1+\nu)(\sigma_0/E)\rho^2$$

$$C = \sigma_0[\tfrac{1}{2}m - (1-m)\ln(b/a) - \tfrac{1}{2}(1-\rho^2/b^2)] \quad , \quad D = \sigma_0\rho^2/(1-\nu) \tag{15}$$

the dimensionless interference pressure and interference are given, respectively, by

$$\bar{p} = P/\sigma_0 = \tfrac{1}{2}(1-\rho^2/b^2) + (1-m)\ln(\rho/a) + \tfrac{1}{2}m(\rho^2/a^2-1) \tag{16}$$

$$\bar{I} = (E/\sigma_0)I/a = (\rho/a)^2 - [(1-\nu) - (1-\nu_1)E/E_1](P/\sigma_0) \tag{17}$$

When the ring and disk are made of the same material, i.e., $E_1 = E$, $\nu_1 = \nu$, Eq. (17) reduces to the simple formula, $(E/\sigma_0)I/a = (\rho/a)^2$. For this special case [1], the constants A, B, C, D, P, and ρ can be expressed explicitly as functions of interference I. In general, the interference pressure (p) is related to the interference (I) implicitly through the elastic-plastic interface (ρ) as shown in Eqs. (16) and (17) for $a \leqslant \rho \leqslant b$. The upper limit of the partially plastic assembly is obtained by letting $\rho = b$. The corresponding interference pressure (p**) and interference (I**) are

$$p**/\sigma_O = (1-m)\ell n(b/a) + \tfrac{1}{2}m(b^2/a^2-1)$$

$$(E/\sigma_O)I**/a = (b/a)^2 - [(1-\nu) - (1-\nu_1)E/E_1]p**/\sigma_O \qquad (18)$$

IV. FULLY PLASTIC ASSEMBLY. When the interference I is larger than I**, we have reached the fully plastic state in the ring. In this case, the expressions for the stresses and the displacement in a ≤ r ≤ b are still the same as those given by Eqs. (9), (10), and (11). The constants C, D and the interference p are determined with the boundary conditions $\sigma_r(a) = -p$, $\sigma_r(b) = 0$, and the compatibility requirement $u(a^+) - u(a^-) = I$. The results for the constants are

$$C = [pa^2/b^2 - (1-m)\sigma_O \ell n(b/a)]/(1-a^2/b^2)$$

$$D = 2a^2[p - (1-m)\sigma_O\ell n(b/a)]/[m(1-\nu)(1-a^2/b^2)] \qquad (19)$$

and the interference pressure is given as a function of interference by

$$\frac{p}{\sigma_O} = \frac{m(E\sigma_O/Ia)(1-a^2/b^2) + 2(1-m)\ell n(b/a)}{2 - m[(1-\nu) - (1-\nu_1)E/E_1](1-a^2/b^2)} \qquad (20)$$

V. NUMERICAL RESULTS AND DISCUSSIONS. The analysis described above makes it possible to predict the interference pressure in a composite shrink fit assembly, and hence, determine the stress state in the ring and disk as a function of the interference. The numerical results have been obtained for shrink fit assemblies with different geometric ratio (α = a/b), hardening parameter (m), and different combinations of materials. For a steel ring with α = 0.5, m = 0.0, E = 30x10⁶ psi, ν = 0.3, σ_O = 15x10⁴ psi, we have considered three types of disks: (a) rigid disk with E_1 = 1000 E, ν_1 = 0.0, σ_1 = 1000 σ_O; (b) steel disk of the same material as the ring; (c) a disk made of tungsten carbide with E_1 = 88.5x10⁶ psi, ν_1 = 0.258, σ_1 = 50x10⁴ psi. The numerical results of tne interference pressure (p/σ_O) for these three cases are presented graphically in Figure 2 as functions of the interference ($\bar{I}$). The results of the hoop stress at the inside surface of the ring are presented in Figure 3 also for these three cases. As can be seen from these two figures, the results for the composite shrink fit assembly (c) falls between the two limits established by cases (a) and (b).

For composite shrink fit assemblies made of tungsten carbide disk and steel ring with α = 0.5, m = 0.0, 0.1, 0.2, the results are presented in Figures 4 and 5, respectively, for the interference pressure and the hoop stress at the bore as functions of the interference. The effect of hardening parameter (m) on these relations can be seen from these two figures. For the same combination of composite shrink fit assembly with m = 0.05, α = 1/4, 1/3, 1/2, 3/4, the results showing the effect of geometric ratio (α) are shown in Figures 6 and 7 for the interference pressure and hoop stress at the bore, respectively.

446

The numerical results of the stresses and displacements in composite shrink fit assemblies have also been obtained, but only some results are presented here. The distributions of hoop stresses in a steel ring with $\alpha = 0.5$ are shown in Figures 8, 9, and 10 for $m = 0.0, 0.1, 0.2$, respectively. In each figure, we have shown the results corresponding to four stages of interference: (a) initial yielding ($\rho/a = 1.0$), $I^* = 0.832$; (b) partial yielding ($\rho/a = 1.5$), $I = 1.970, 1.960, 1.950$; (c) complete yielding ($\rho/a = 2.0$), $I^{**} = 3.689, 3.653, 3.617$; (d) fully plastic state with $I = 1.5 \; I^{**}$. For an ideally plastic ring ($m = 0.0$), the stress distribution remains unchanged after complete yielding has been reached. For strain-hardening rings, the stress distributions show large variations, especially for large values of interference. As shown in Figures 8, 9, and 10, the hardening parameter has a significant effect on the stress distributions. Additional stress distributions in the ring with $m = 0.1$ are shown in Figures 11 and 12 for $\alpha = 1/3$ and $1/4$, respectively. The effect of geometric ratio on the distributions can be seen by comparing Figures 9, 11, and 12.

REFERENCES

1. Gamer, U. and Lance, R. H., "Residual Stresses in Shrink Fits," _Int. J. Mech. Sci._, Vol. 25, No. 7, 1983, pp. 465-470.

2. Bland, D. R., "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and Temperature Gradients," _J. Mech. Phys. Solids_, Vol. 4, 1956, pp. 209-229.
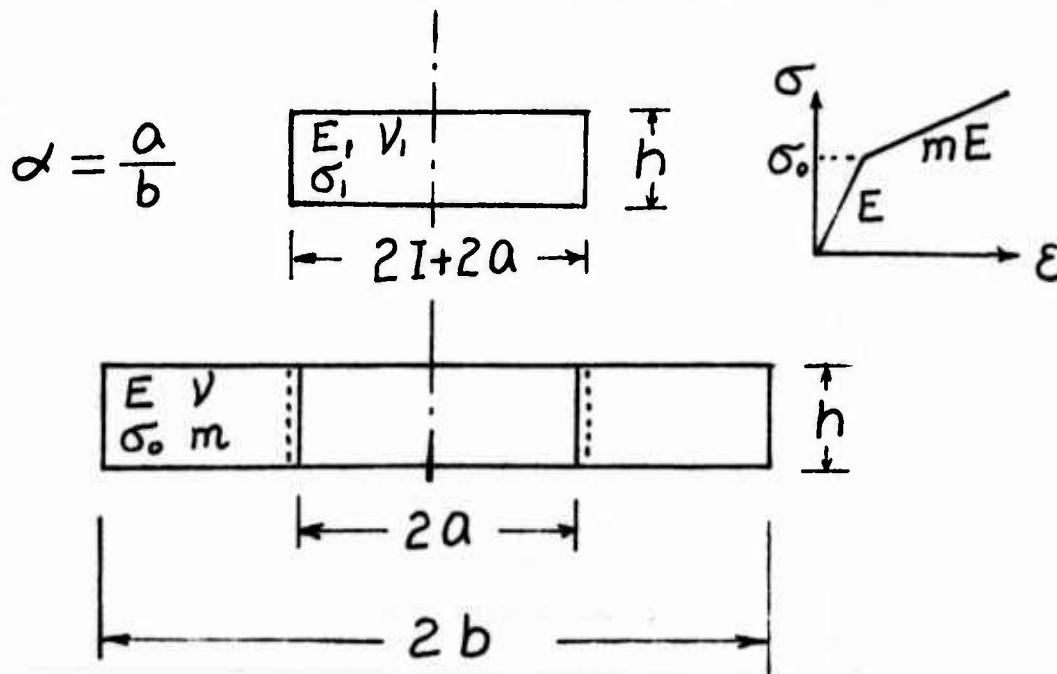
Fig. 1 Shrink fit assembly
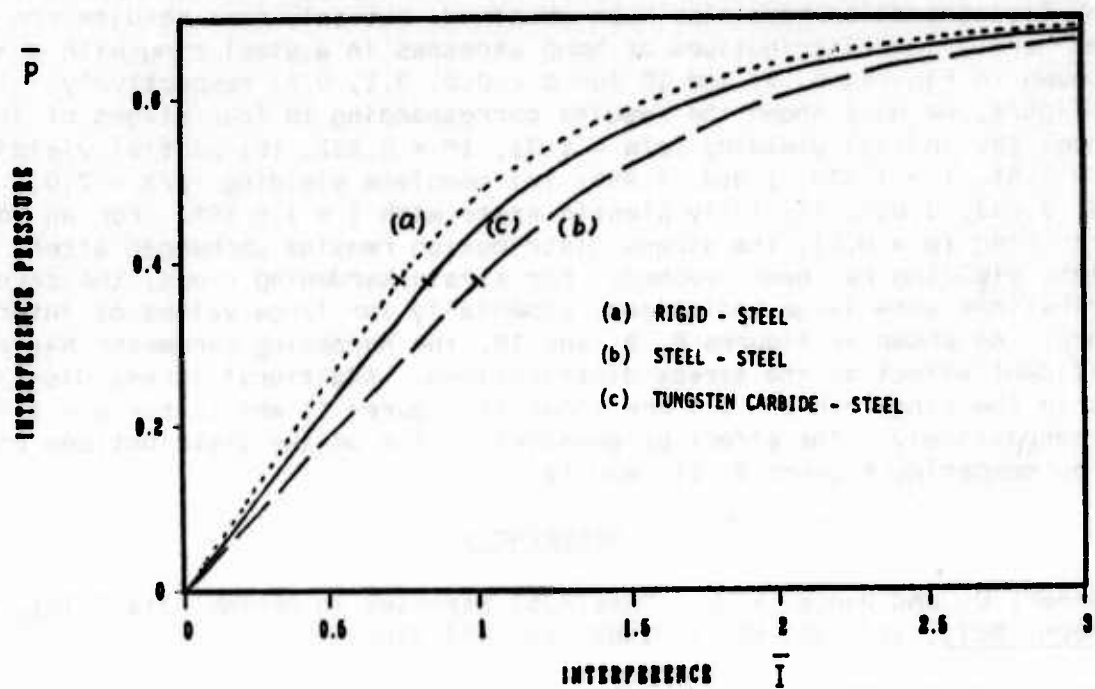
Figure 2. Interference pressure versus interference for three shrink fit assemblies ($\alpha = 0.5$, $m = 0.0$).
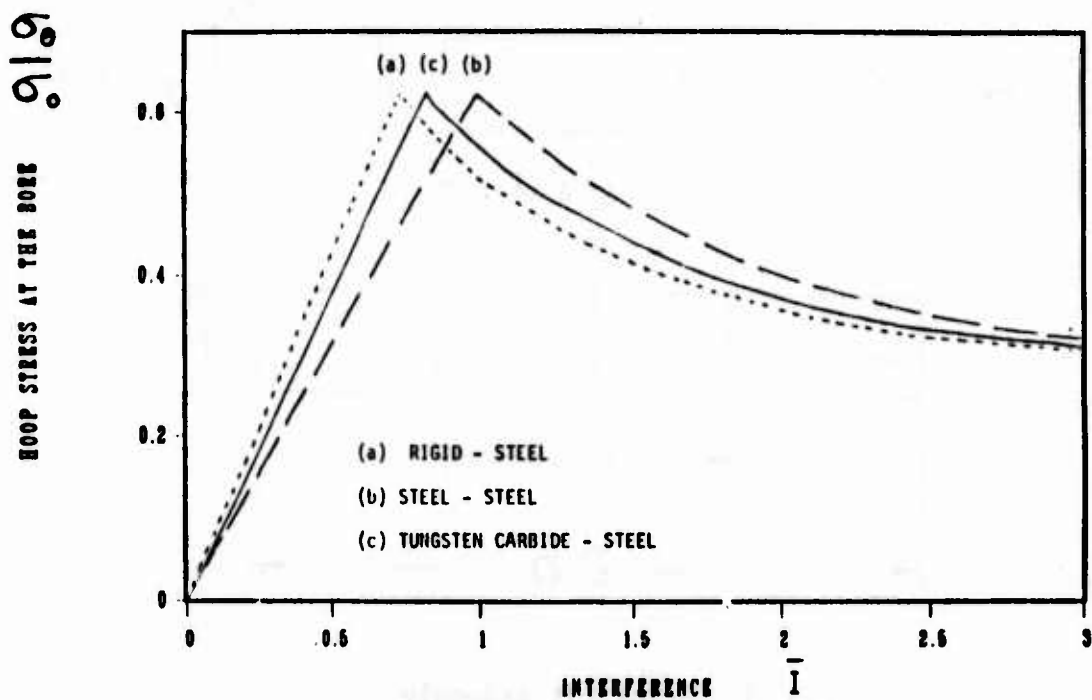
Figure 3. Hoop stress at the bore versus interference for three shrink fit assemblies ($\alpha = 0.5$, $m = 0.0$).
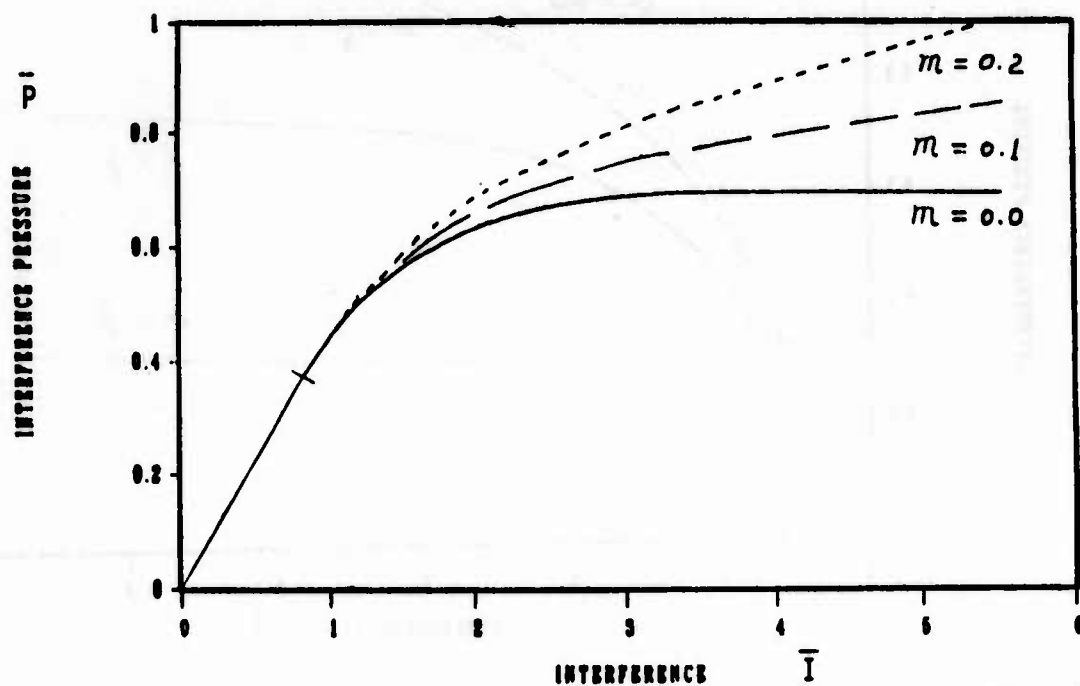
Figure 4.   The effect of hardening on interference pressure in a composite
assembly (α = 0.5).
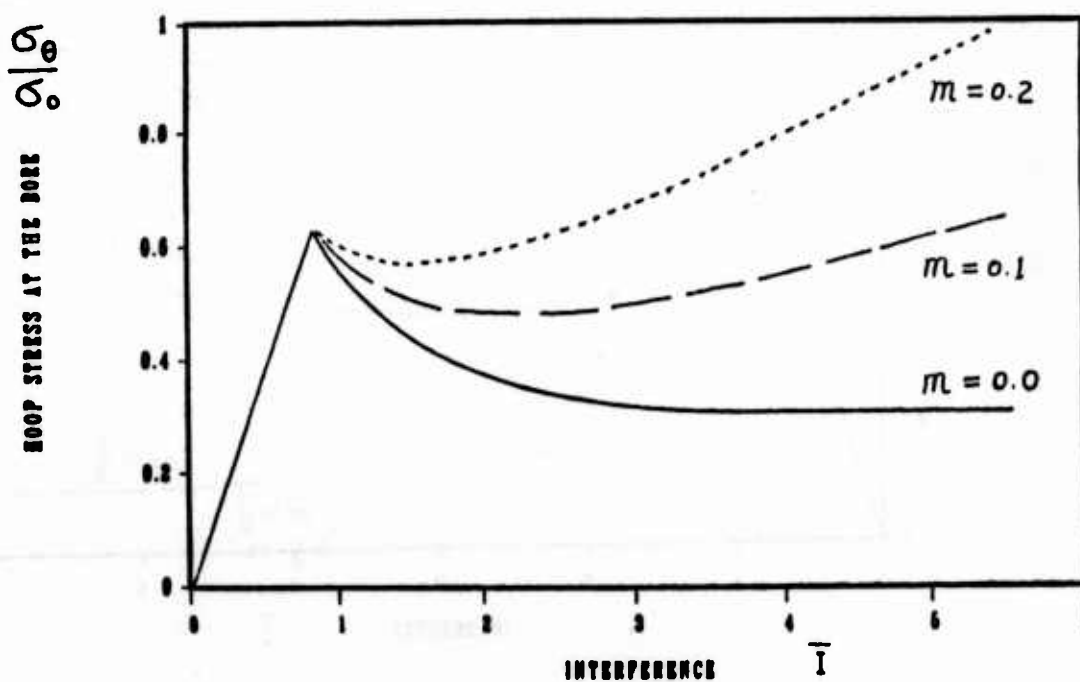
Figure 5.   The effect of hardening on the hoop stress at the bore of a steel
ring (α = 0.5).
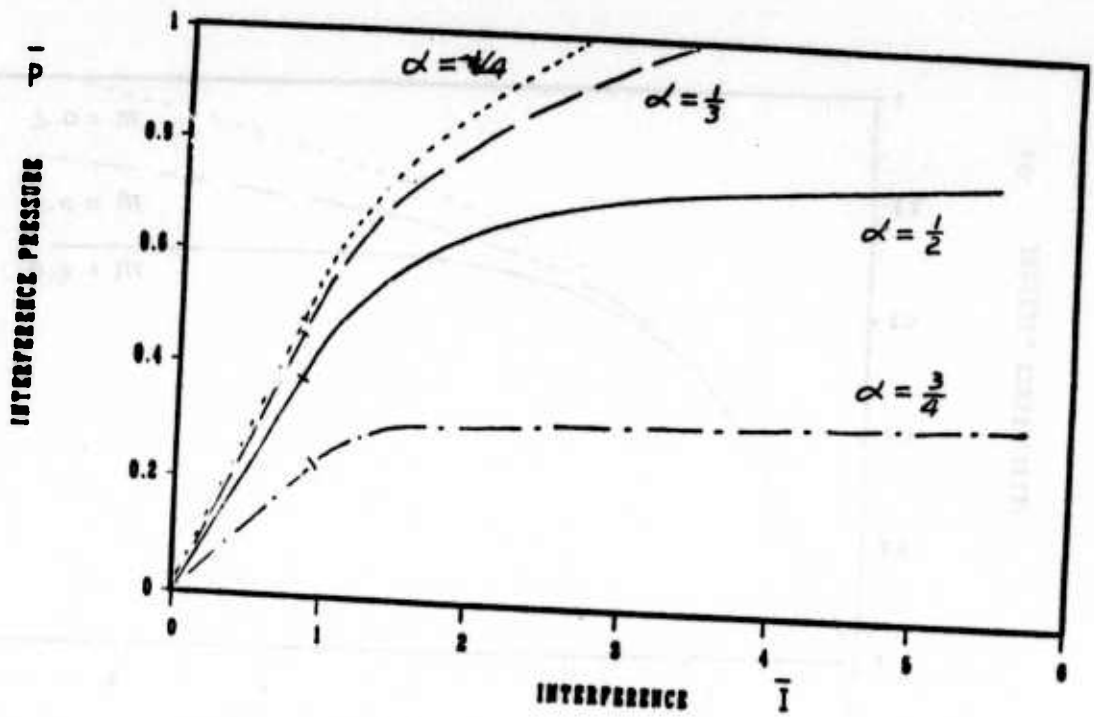
Figure 6. The effect of geometric ratio on interference pressure in a composite assembly (m = 0.05).
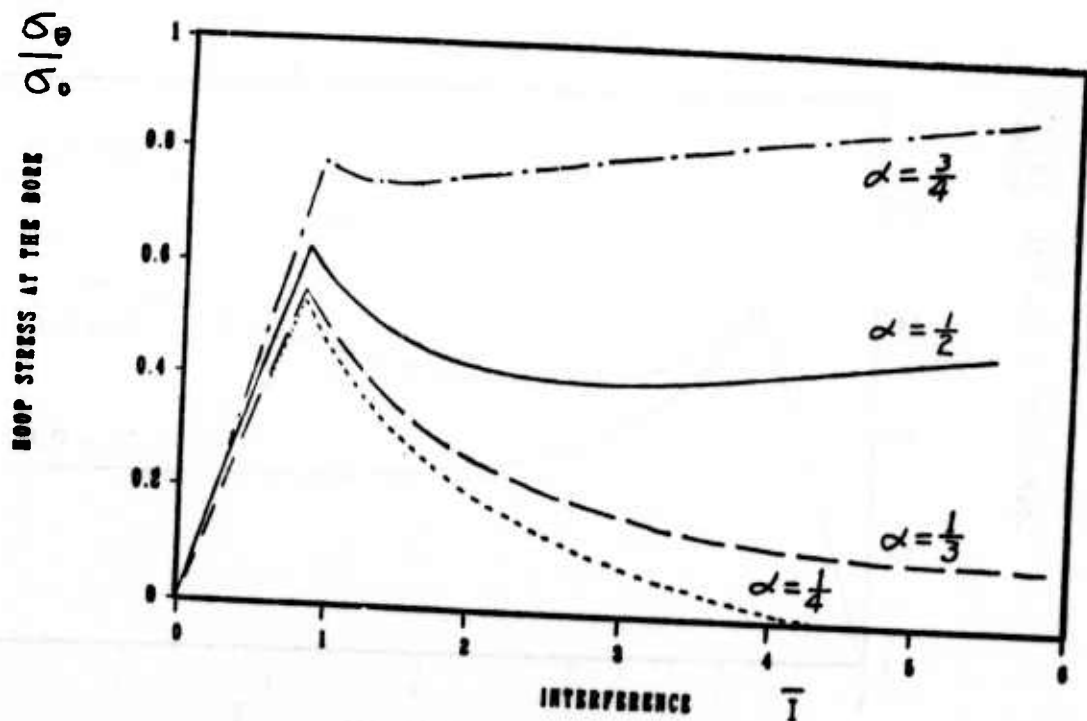
Figure 7. The effect of geometric ratio on the hoop stress at the bore of a steel ring (m = 0.05).
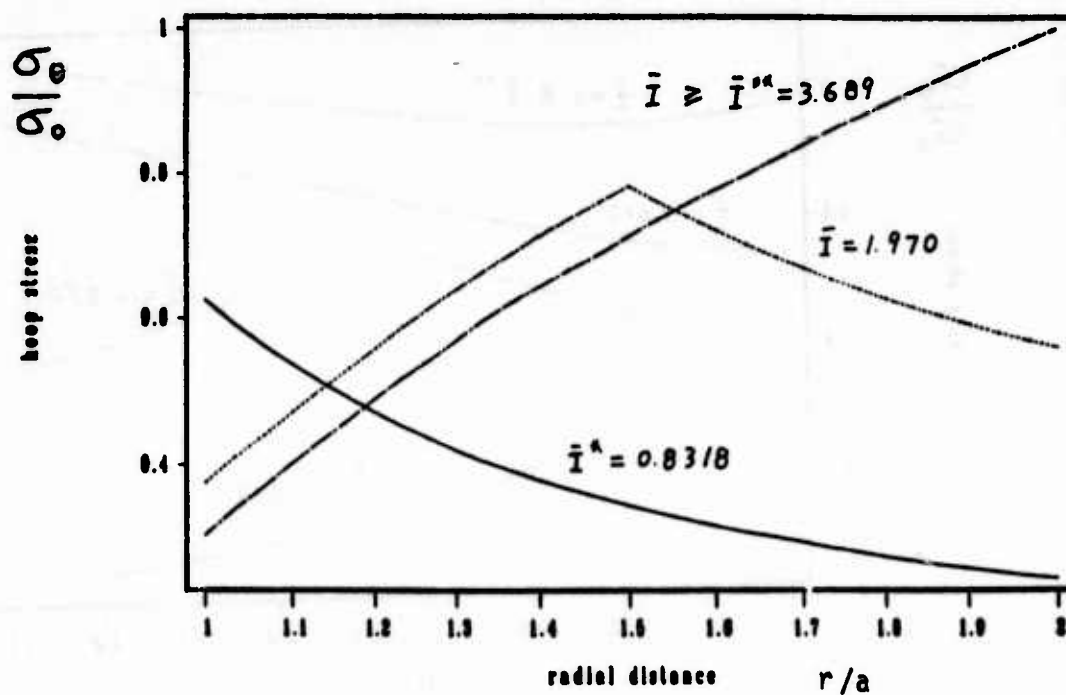
Figure 8. The distribution of hoop stress in a composite assembly (α = 0.5, m = 0.0).
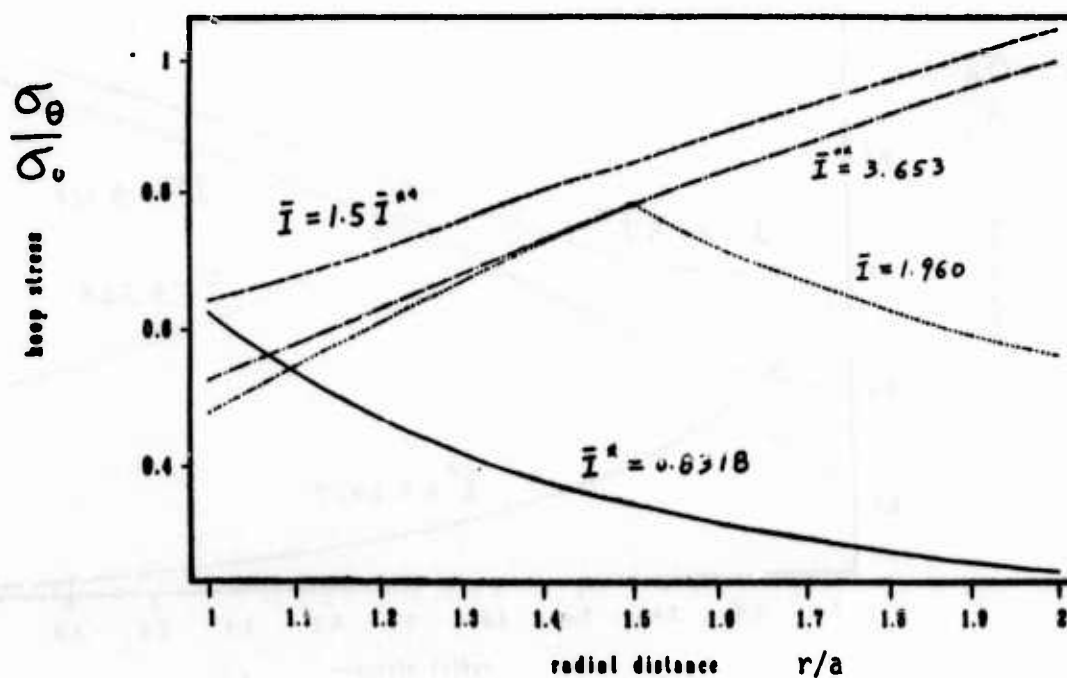
Figure 9. The distribution of hoop stress in a composite assembly (α = 0.5, m = 0.1).
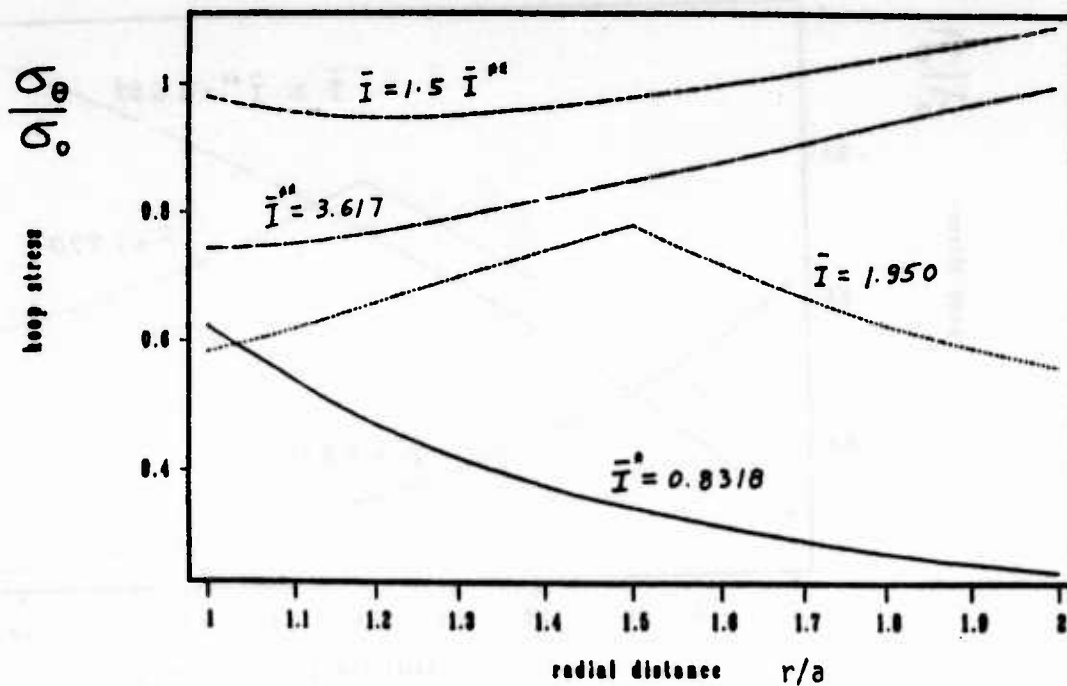
Figure 10. The distribution of hoop stress in a composite assembly (α = 0.5, m = 0.2).
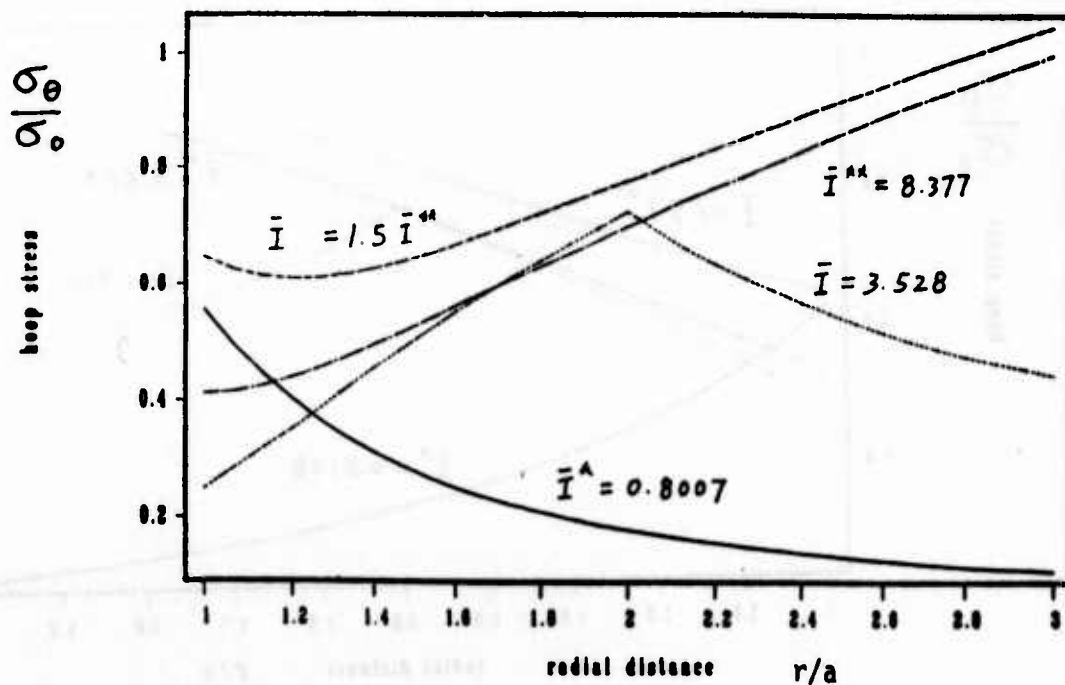
Figure 11. The distribution of hoop stress in a composite assembly (α = 1/3, m = 0.1).
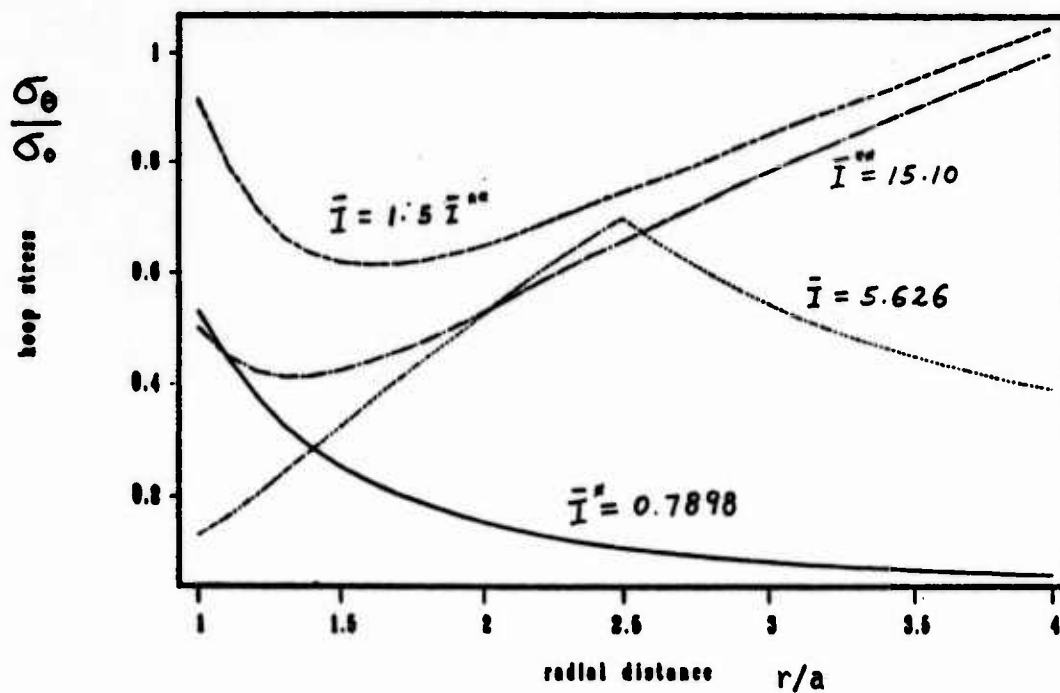
Figure 12. The distribution of hoop stress in a composite assembly
(α = 0.25, m = 0.1).

A SHALLOWLY CURVED SHEAR-DEFORMABLE BEAM ELEMENT

Alexander Tessler   and   Luciano Spiridigliozzi
Mechanics and Structures Division
U.S. Army Materials Technology Laboratory
Watertown, MA 02172-0001

ABSTRACT. Shallowly curved beam elements, including shear deformation and rotary inertia effects, are derived from Hamilton's variational principle. Different degree polynomials, labeled 'anisoparametric', are used to interpolate the kinematic variables, instead of uniform interpolations as in the conventional isoparametric procedure. This approach yields a correct representation of the bending strain and, importantly, the membrane and transverse shear strains. Consequently, the severe shortcomings of the exactly integrated isoparametric elements, characterized by excessively stiff solutions in the thin regime (a phenomenon often referred to as membrane and shear locking), are overcome. Uniform (isoparametric-like) nodal patterns are achieved by explicitly enforcing higher-degree penalty modes in the membrane and shear strains. This procedure preserves the compatibility of the kinematic field and the capability of the element to move rigidly without straining. Exact quadratures are used on all element matrices, producing a correct rank stiffness matrix, a consistent load vector, and a consistent mass matrix. The elements suffer no limitations over the entire theoretical range of the slenderness ratio. For further enhancement and, particularly, in coarse-mesh situations, an effective relaxation of penalty constraints at the local element level is introduced. This technique ensures a well-conditioned stiffness matrix. Although the element penalty constraints are relaxed, the corresponding global structure constraints are enforced as is required by the analytic theory. Particular attention is given to the simplest element -- a two-node, six degree-of-freedom beam in which all strains are constant. Solutions to static and free vibration arch and ring problems are presented, demonstrating the exceptional modeling capabilities of this element.

I. INTRODUCTION. Early formulations for curved beam models employed the assumptions of Bernoulli-Euler theory and often produced elements that lacked rigid body motion and at the same time exhibited severe stiffening in approximating the behavior of thin and deep arches [1-4]. These classical elements require C^0 and C^1 continuity for the membrane and transverse displacements, respectively. Using mixed polynomial-trigonometric shape functions, and at the expense of higher continuity enforcement [5-8], proper rigid-body motion was obtained, though the full extension into the thin regime was not attained. Fried [9-10] identified the source of the thin-regime difficulty as one of imbalance between discretization and inextensibility errors, and he employed his 'residual' energy balancing technique to arrive at well-behaved thin-arch and cylindrical shell elements. Meck [11] showed that even without proper rigid-body motion, effective thin-arch elements could be produced once 'consistent' displacement polynomials were used. He interpolated the membrane displacement with a polynomial one degree higher than that used for the transverse displacement and achieved

physically meaningful membrane-bending coupling in the membrane strain. Mixed and selective/reduced integration approaches based upon isoparametric interpolations [12-16] produced several well-behaved elements. In certain cases of reduced integration, however, spurious zero-energy modes were introduced, or the required membrane-bending coupling was violated [14].

The cause of thin-regime stiffening in the classical formulations (i.e., 'membrane locking') can be traced to the penalty-type membrane strain energy which involves membrane-bending coupling. Extensional deformations in thick beams and inextensibility of thin beams are both governed by a penalty parameter which becomes large in the thin regime. For very slender elements, membrane and transverse displacements, originally assumed independent, become interdependent because of coupling in the membrane strain and the vanishing strain requirement imposed by a large penalty parameter. When a membrane-strain state lacks the theoretically required membrane-bending coupling, commonly the result of interpolation inconsistency (e.g., when isoparametric interpolations are used), spurious (nonphysical) constraining of a single kinematic field takes place. This in turn yields severe constraining of the bending strain, giving rise to excessively stiff deformations.

The behavior of shear-deformable (C^o) straight beam and flat plate/shell elements is also governed by a penalty mechanism from the vanishing-strain enforcement. There, the penalized shear strain energy contains a deflection-rotation coupling in the shear strains. For straight Timoshenko beams, Tessler and Dong [17] demonstrated a variationally consistent approach which yielded simple displacement elements devoid of any slenderness related deficiency. Their kinematic interpolations ensured physically admissible coupling in all shear-strain states, analogous to Meck's [11] membrane-bending coupling in thin-arch elements. By adopting the deflection polynomial one degree higher than that for the normal rotation [17], uniform nodal patterns were produced by explicitly enforcing the higher-degree Kirchhoff (zero-shear) modes. The Tessler-Dong elements possess the same stiffness matrices and, in fact, use the same integration rule as the corresponding selective/-reduced integration elements [18]; however, in their elements [17] the integrations are exact. Further, in the cases of distributed static and inertial loadings, these elements perform significantly better than their reduced integration counterparts. This is because consistent load vectors and mass matrices evolve naturally in the Tessler-Dong formulation, but are not available within the reduced integration procedures. (Referring to exactly integrated load vectors and mass matrices as 'consistent' for underintegrated stiffnesses, as is often done [15,16], is in fact erroneous; refer to the discussion in [17].)

Recent extensions of the one-dimensional interpolation strategy [17] to axisymmetric shell and plate elements were introduced by Tessler [19-21] and Tessler and Hughes [22-23], where to maintain variational consistency and kinematic reliability (i.e., correct rank of a stiffness matrix) full Gaussian quadratures were employed on all energy and work terms. It was pointed out [22] that in the case of a quadrilateral plate element, it is not always possible to achieve proper deflection-rotation coupling in all shear-strain states. Further,

456

kinematic restraints, such as boundary conditions, can readily uncouple a consistently coupled shear-strain mode (e.g., by removing deflection degrees of freedom), causing the elements to lock [23]. Thus, interpolation measures alone were insufficient to ensure proper thin-regime behavior.

The fundamental difficulty stems from the <u>direct</u> adoption of analytic theory to finite elements, which results in <u>the strict enforcement of vanishing penalty strains at the element level</u>. The fact that vanishing shear and extensional deformations pertain only at <u>the global structure level</u> has largely been overlooked. Because of the limited kinematic freedoms afforded by element interpolations, it would seem only natural to depart from the conventional approach and adopt what we shall refer to as the <u>local penalty-relaxation</u> method. The idea is to 'relax' the enforcement of element penalty strains, yet maintain the global penalty constraints of the discretized structure. The actual implementation is via correction parameters for the penalty stress resultants (i.e., membrane and shear forces), which also appear in the element penalty parameters. Although implemented somewhat differently, the techniques of Fried [9-10], MacNeal [24], and Tessler and Hughes [22-23] are closely related and can be characterized as local penalty-relaxation methods. The principal benefits of these procedures are the removal of the locking deficiency, the improved condition of the stiffness matrix, and the improved solutions in coarse-mesh situations.

The objective of the present paper is to extend the ideas explored in [11,17,19-23] to shallowly curved, shear-deformable beam elements. The reason for adopting a shallow element geometry is to effectively model shallow as well as deeply curved beams [7]. The inclusion of shear deformation and, in dynamics, rotary inertia extends the range of applicability to moderately thick regimes. The principal goal is to achieve a reliable and effective element behavior that is not subject to any thinness restrictions, so that the basic methodology can be applied to general curved shell models.

To ensure effective element behavior in the limiting thin membrane/bending regimes, consistent interpolations (labeled 'anisoparametric' [23]) and suitable corrections (relaxations) of the element-level membrane and shear penalty modes are incorporated. From a hierarchy of these anisoparametric elements, we examine the simplest, two-node (six degrees of freedom) element, in which all strain components are constant along its span. The element is C^o compatible, devoid of locking even in the extremely thin regime, and it can move rigidly without straining. Its stiffness matrix is well-conditioned over the entire range of the element slenderness ratio, which implies ideal suitability for applications on microcomputers with a short word length.

In Section 2, the structural theory of shallowly curved beams [25] is briefly described. In Section 3, a variational implementation of the theory with anisoparametric displacement fields is discussed for a hierarchy of elements, whereas in Section 4 a two-noded element is derived. Implementation of the local penalty relaxation concept is addressed in Section 5. Finally, solutions to static and vibration test

problems are presented in Section 6, and conclusions are summarized in Section 7.

II. ANALYTIC THEORY. To establish a conceptual basis, the Timoshenko/Marguerre shallow beam equations [25-26] are discussed within the framework of linearly elastic, planar deformations.

Let the curved beam of a uniform rectangular cross-section be located in the x-z plane (which is the plane of cross-sectional symmetry) where the x-axis is coincident with the beam chord (refer to Figure 1). The middle-surface membrane displacement, $u(x,t)$ (henceforth, t denotes time), transverse displacement, $w(x,t)$, and normal cross-sectional rotation, $\theta(x,t)$, completely describe the planar membrane, bending, and transverse shear deformations; with the strains given by:

$$\epsilon = \epsilon_u + \epsilon_w = u,_x + w_{I,x} w,_x \tag{2.1}$$

$$\kappa = -\theta,_x \tag{2.2}$$

$$\gamma = \gamma_w + \gamma_\theta = w,_x - \theta, \tag{2.3}$$

where w_I describes the shallow shape of the beam ($w_{I,x}^2 \ll 1$). Evidently, for a straight beam $\epsilon_w = 0$, and all strains (2.1)-(2.3) are those of Timoshenko theory [26].

The kinetic variables of the theory are the membrane force resultant, $N(x,t)$, the bending moment, $M(x,t)$, and the transverse shear force, $Q(x,t)$. These are related to the corresponding strains by the constitutive relations:

$$N = D_m \epsilon = AE\epsilon, \quad M = D_b \kappa = EI\kappa, \quad Q = D_s \gamma = k^2 GA\gamma, \tag{2.4}$$

with A, I, E, G, and k^2 denoting the cross-sectional area, the moment of inertia, the elastic modulus, the shear modulus, and the shear correction factor, respectively.

To derive the equations of motion, Hamilton's variational statement is invoked, i.e.,

$$\delta \int_{t_0}^{t_1} L dt = \delta \int_{t_0}^{t_1} \{ \tfrac{1}{2} \int_0^\ell [\rho A(\dot{u}^2 + \dot{w}^2) + \rho I \dot{\theta}^2] dx$$

$$- \tfrac{1}{2} \int_0^\ell [N\epsilon + M\kappa + Q\gamma] dx$$

$$+ \int_0^\ell [wq + \theta m] dx \} dt = 0 \tag{2.5}$$

where a superior dot denotes differentiation with respect to t, ρ is the mass density, ℓ is the chord length, and q and m are the distributed transverse force and moment loadings, respectively. Conveniently, due to the shallowness assumption, the energy integrals are taken over a

458

straight chord rather than a curved beam description. (This aspect lends to the desired simplicity in formulating shell elements.)

III. KINEMATIC CONSIDERATIONS.

Of special significance in the present theory is a double-penalty form of the elastic strain energy, which can be expressed as

$$U = \frac{D_b}{2\ell} (U_b + \alpha_m U_m + \alpha_s U_s), \tag{3.1}$$

where the nondimensional bending, membrane and shear energy contributions are respectively

$$U_b = \ell \int_0^\ell \kappa^2 dx, \quad U_m = \frac{1}{\ell} \int_0^\ell \varepsilon^2 dx, \quad U_s = \frac{1}{\ell} \int_0^\ell \gamma^2 dx, \tag{3.2}$$

and the penalty parameters have the form

$$\alpha_m = (\ell/r)^2, \quad \alpha_s = \frac{1}{4}k^2(G/E)(\ell/r)^2, \tag{3.3}$$

with $r = \sqrt{I/A}$ denoting the cross-sectional radius of gyration.

For a very slender beam $\alpha_m, \alpha_s \to \infty$. These parameters enforce thin limits of the membrane inextensibility and shearless (Kirchhoff) regimes:

<u>Inextensibility constraint</u>: $\quad \varepsilon = \varepsilon_u + \varepsilon_w \to 0 \tag{3.4}$

<u>Kirchhoff constraint</u>: $\quad \gamma = \gamma_w + \gamma_\theta \to 0 \tag{3.5}$

The conventional approach of directly applying the theory to finite elements treats the displacement variables and the material and geometric quantities as element properties. This means that the penalty constraints (3.4) and (3.5) are enforced at <u>the element level</u>.

Since the highest spatial derivative in (2.5) is of order one, the displacement variables require only C^0 continuity. This implies a wide range of interpolating possibilities. On the other hand, penalty constraints (3.4) and (3.5) limit the choice of interpolations to those yielding identical polynomial descriptions for the components of the penalty strains [17,19-23]. In the present context the displacement interpolations should yield

$$\varepsilon_u, \varepsilon_w = O(x^m); \quad \gamma_w, \gamma_\theta = O(x^n) \quad (m,n = 0,1,2,\ldots) \tag{3.6}$$

Kinematic interpolations, producing penalty strains of type (3.6), were labeled 'anisoparametric' [23] to emphasize the necessary distinction in the polynomial variations of the displacement variables.

If the initial element shape, $w_I(x)$, is described by the 'shallow' cubic ($\beta_i^2 \ll 1$):

$$w_I(x) = \beta_0 \ell(\eta - 2\eta^2 + \eta^3) + \beta_1 \ell(\eta^3 - \eta^2), \tag{3.7}$$

where

$$\eta = x/\ell, \quad \eta \subset [0,1]; \quad \beta_i = w_{I,x}(\eta_i), \quad i=0,1,$$

then to comply with (3.6), the three displacement variables should be expanded by the distinct-degree anisoparametric polynomials:

$$\theta = \sum_{k=0}^{p} a_{\theta k}\eta^k, \quad w = \sum_{k=0}^{p+1} a_{wk}\eta^k, \quad u = \sum_{k=0}^{p+3} a_{uk}\eta^k \tag{3.8}$$

where a_θ, a_w and a_u are the generalized coordinates in terms of the θ, w and u nodal degrees of freedom (dof), respectively. Introducing (3.8) into (2.1)-(2.3) yields

$$\varepsilon = \sum_{k=0}^{p+2} \varepsilon_k\eta^k, \quad \kappa = \sum_{k=0}^{p-1} \kappa_k\eta^k, \quad \gamma = \sum_{k=0}^{p} \gamma_k\eta^k. \tag{3.9}$$

The strains (3.9) contain two sets of penalty modes arising from the thinness constraints (3.4) and (3.5):

p+3 inextensional modes; for, example, for p=1:

$$\varepsilon_0\ell = a_{u_1} + \beta_0 a_{w_1} \to 0,$$

$$\varepsilon_1\ell = 2[a_{u_2} + \beta_0 a_{w_2} - (\beta_1 + 2\beta_0)a_{w_1}] \to 0, \tag{3.10}$$

$$\varepsilon_2\ell = 3[a_{u_3} + (\beta_0 + \beta_1)a_{w_1}] - 4(\beta_1 + 2\beta_0)a_{w_2} \to 0,$$

$$\varepsilon_3\ell = 4a_{u_4} + 6(\beta_0 + \beta_1)a_{w_2} \to 0$$

p+1 Kirchhoff modes; for p=1:

$$\gamma_0\ell = a_{w_1} - \ell a_{\theta_0} \to 0,$$

$$\gamma_1\ell = 2a_{w_2} - \ell a_{\theta_1} \to 0. \tag{3.11}$$

Important considerations in the element construction are the number of nodal dof, their locations, as well as the order of the strain (stress) approximations. Thus, assumptions (3.8) produce an element with (3p+7) dof, and (p+2), (p-1), and p polynomial degrees for the membrane, bending and shear strains, respectively. It can be seen from (3.10) and (3.11) that a reduction in the number of dof is possible by lowering the degree of the strain interpolations. This can be accomplished a priori to integrating the element matrices by explicitly enforcing the higher-degree penalty modes [17]. By this procedure, a hierarchy of elements with the desired number of dof, nodes, and spatial variations of strains (stresses) can be constructed while retaining the initial polynomial variations of the assumed displacements and preserving a rigid-body motion. Figures 2 and 3 show some nodal and strain variation possibilities for the three lowest order elements (p=1,2,3) in their initial and constrained configurations, respectively.

REMARK 3.1   It should be realized that requirements (3.6) are not limited to shallowly curved elements alone.   Rather, they apply to all curved beams.   For example, for a circular-arch element of radius R the second membrane strain component in (3.4) is $\varepsilon_w = w/R$.   According to (3.6), the interpolation polynomial for u should be one degree higher than that for w.

REMARK 3.2   Note that each penalty mode in (3.10) and (3.11) contains contributions of at least two displacement variables (i.e., a 'true' penalty mode [20]).   This means that proper interdependence of the variables can be achieved in the thin limit, producing a nonlocking solution.   Unfortunately, the true penalty-mode structure can be destroyed by boundary condition restraints.   For instance, when an element is subject to an excessive number of restraints, an initially true (coupled) penalty mode takes a spurious (uncoupled) form, thus causing locking [23].   This deficiency of the conventional approach is particularly pronounced in two-dimensional plate/shell problems, where a single element along plate boundary is often subject to a large number of kinematic restraints [21].   We shall further elaborate on this aspect in Section 5.

IV. SIMPLEST ELEMENT.   According to the anisoparametric interpolation strategy just described, the simplest first-order (p=1) element is a two-node (six dof) beam in which all three strain components are chord-wise constant.   The initial C^o interpolations are quartic u, quadratic w, and linear θ.   For convenience, we cast them in terms of Lagrange polynomials:

$$u = u_0 + \sum_{i=1}^{4} \underline{u}_i \phi_i(\eta),$$

$$w = w_0 + \sum_{i=1}^{2} \underline{w}_i \phi_i(\eta),$$

$$\theta = \theta_0 + \underline{\theta}_1 \phi_1(\eta),$$

(4.1)

in which u_i, w_i, and θ_i are the nodal dof (refer to Figure 2), and $\underline{u}_i$, $\underline{w}_i$, and $\underline{\theta}_i$ are the generalized coordinates given by

$$\underline{u}_1 = u_1 - u_0, \quad \underline{w}_1 = w_1 - w_0, \quad \underline{\theta}_1 = \theta_1 - \theta_0,$$

$$\underline{u}_2 = u_2 - (u_1 + u_0)/2, \quad \underline{w}_2 = w_2 - (w_1 + w_0)/2,$$

$$\underline{u}_3 = u_3 + (u_0 - 3u_1 - 6u_2)/8, \quad \underline{u}_4 = u_4 - (u_0 + u_1 + 6u_2 - 4u_3)/4 \qquad (4.2)$$

and ϕ_i denote the Lagrange shape functions:

$$\Phi_1 = \eta, \quad \Phi_2 = 4\eta(1-\eta), \quad \Phi_3 = (32/3)\eta(1-\eta)(2\eta-1),$$

$$\Phi_4 = (16/3)\eta(1-\eta)(2\eta-1)(4\eta-3),$$

$$\Phi_i(\eta_j) = \begin{cases} 0 & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases} \tag{4.3}$$

To arrive at a constant strain element, we enforce explicitly the linear, quadratic, and cubic inextensional modes and the linear Kirchhoff mode, condensing out $\underline{u}_2$, $\underline{u}_3$, $\underline{u}_4$ and $\underline{w}_2$ coordinates. This dof reduction procedure is analogous to that of [17], but it requires fewer algebraic manipulations because of the hierarchical structure of the Lagrange functions. The implementation is as follows:

(1) Explicitly enforce three higher-order inextensional modes:

$$\left\{ \frac{\partial}{\partial x} \, , \, \frac{\partial^2}{\partial x^2} \, , \, \frac{\partial^3}{\partial x^3} \right\}^T (\epsilon) = 0 \tag{4.4}$$

which result in

$$\begin{bmatrix} \underline{u}_2 \\ \underline{u}_3 \\ \underline{u}_4 \end{bmatrix} = \begin{bmatrix} (\beta_0 - \beta_1)\ell_{12} & (\beta_0 + \beta_1)\ell_{12} \\ -6(\beta_0+\beta_1)\ell_{13} & (7\beta_0-25\beta_1)\ell_{24} \\ 0 & -18(\beta_0+\beta_1)\ell_{24} \end{bmatrix} \begin{bmatrix} \underline{w}_1 \\ \underline{w}_2 \end{bmatrix} \tag{4.5}$$

where coefficients ℓ_{pq} are computed from

$$\ell_{pq} = \frac{\dfrac{\partial^p}{\partial \eta^p}(\Phi_p)}{\dfrac{\partial^q}{\partial \eta^q}(\Phi_q)} \tag{4.6}$$

or

$$\ell_{12} = -\,1/8, \quad \ell_{13} = -\ell_{24} = -1/128.$$

(2) Explicitly enforce the linear Kirchhoff mode:

$$\frac{\partial}{\partial x}(\gamma) = 0$$

resulting in

$$\underline{w}_2 = -\ell\underline{\theta}_1/8 \tag{4.7}$$

Introducing (4.6) and (4.7) into (4.5), and then substituting (4.5) into the initial displacement assumptions (4.1), gives the constrained

interpolations for u and w in terms of the end-node dof, while leaving θ unaltered:

$$\theta = \sum_{i=0}^{1} N_i \theta_i, \quad w = \sum_{i=0}^{1} [N_i w_i + K_i \theta_i],$$

$$u = \sum_{i=0}^{1} [N_i u_i + L_i w_i + M_i \theta_i]. \tag{4.8}$$

in which

$$N_1 = 1 - N_0 = \phi_1,$$

$$L_0 = -L_1 = \frac{1}{8}[(\beta_0 - \beta_1)\phi_2 - \frac{3}{8}(\beta_0 + \beta_1)\phi_3],$$

$$M_0 = -M_1 = \frac{\ell}{64}[-(\beta_0 + \beta_1)\phi_2 + \frac{7\beta_0 - 25\beta_1}{16}\phi_3 - \frac{9}{8}(\beta_0 + \beta_1)\phi_4],$$

$$K_0 = -K_1 = \frac{\ell}{8}\phi_2.$$

Interpolations (4.8) preserve interelement displacement compatibility and satisfy the constant strain criterion [27]:

$$\sum_{i=0}^{1} N_i = 1, \quad \sum_{i=0}^{1} (N_i + K_i) = 1, \quad \sum_{i=0}^{1} (N_i + L_i + M_i) = 1, \tag{4.9}$$

embodying three rigid-body modes. For a straight element ($\beta_0 = \beta_1 = 0$) the membrane displacement reduces from a quartic to a linear chord-wise variation, which is commonly used.

The constant strain components are found as

$$\varepsilon = \frac{1}{\ell}(u_1 - u_0) + (\beta_1 - \beta_0)(\theta_1 - \theta_0)/12,$$

$$\kappa = \frac{1}{\ell}(\theta_1 - \theta_0), \quad \gamma = \frac{1}{\ell}(w_1 - w_0) - (\theta_1 + \theta_0)/2. \tag{4.10}$$

Letting the strains vanish simultaneously results in the rigid-body displacements:

$$\theta = \bar{\theta}, \quad w = w_0 + \ell\phi_1\bar{\theta},$$

$$u = \bar{u} - \ell L_0 \bar{\theta},$$

$$\bar{u} = u_0 = u_1, \quad \bar{\theta} = \theta_0 = \theta_1. \tag{4.11}$$

Note that curvature κ and shear strain γ in (4.10) are the same as those for the two-node straight Timoshenko element [17]. As discussed in [17], these quantities are identical to the corresponding strains for a linear, single-point quadrature element of Hughes et. al. [18].

We thus derived a simple displacement field satisfying the standard convergence criteria and the true penalty-mode requirement. Because all strains are chord-wise constant, exact energy integrals are readily computed either analytically or using a single-point Gaussian quadrature. The derivations of the stiffness and consistent mass matrices and consistent load vector are straightforward: simply replace (4.8) in (2.5) and perform the required variational operation. The element matrices are given in Appendix A.

V. **PENALTY RELAXATION.** As mentioned previously, the penalty constraints are conventionally enforced at the element level. Consequently, consistent kinematic coupling in the penalty modes is paramount in order to avoid locking. However, even when such coupling is properly maintained, stiffer than desired deformations are expected, particularly in coarsely discretized models. Furthermore, circumstances may arise when a single element is subject to an excessive number of displacement boundary restraints. For instance, if three out of four bending dof are fixed in our two-node element (e.g., $w_0 = w_1 = \theta_0 = 0$), shear locking occurs, since the Kirchhoff constraint takes a spurious form: $\gamma = \theta_1 \to 0$ (refer to (4.10)).

A rational and effective way of resolving these deficiencies over the entire theoretical range of ℓ/r is to relax the strict enforcement of penalty modes at the element level. Recall that the analytic theory requires zero-strain penalty constraints at the global structure level. Using appropriate correction (relaxation) parameters in the form of multipliers of the penalty stress resultants (i.e., the membrane and shear forces), it is possible to relax the element level constraints, yet retain their validity at the global structure level. In the limit as the element size diminishes to zero, these parameters should approach unity, since no correction is needed. By a matching procedure involving exact and finite element energy solutions, Tessler and Hughes [22,23] derived the shear correction parameters for straight beam and flat plate elements. Herein, we adopt the same general form [22] for the correction parameters, i.e.,

$$\phi_i^2 = (1 + C_i \alpha_i)^{-1} \quad (i=m,s) \tag{5.1}$$

which for the curved beam element equal:

$$\phi_m^2 = \frac{1}{1 + C_m (\ell/r)^2}, \quad \phi_s^2 = \frac{1}{1 + C_s \tfrac{1}{4} k^2 (G/E) (\ell/r)^2},$$

where C_m and C_s are the element constants established permanently from simple numerical experiments. Employing (5.1), the element constitutive relations are corrected accordingly:

$$N^e = \phi_m^2 \, D_m \epsilon, \quad Q^e = \phi_s^2 \, D_s \gamma. \tag{5.2}$$

Substituting (5.2) into the strain energy in (2.5) produces the corrected element penalty parameters:

$$\alpha_i^e = \alpha_i \phi_i^2 \quad (i = m, s), \tag{5.3}$$

or in the expanded form

$$\alpha_m^e = \frac{(\ell/r)^2}{1 + C_m \, (\ell/r)^2}, \qquad \alpha_s^e = \frac{\tfrac{1}{4} k^2 (G/E) \, (\ell/r)^2}{1 + C_s \tfrac{1}{4} k^2 (G/E) \, (\ell/r)^2}.$$

For the two extrema of the slenderness ratio, $\ell/r \to \{\infty, 0\}$, we have

$$\alpha_i^e \to \begin{cases} C_i^{-1} & \text{if } \ell/r \to \infty \\[6pt] \alpha_i & \text{if } \ell/r \to 0 \quad (i = m, s) \end{cases} \tag{5.4}$$

One major advantage with this approach, in addition to achieving speedier monotonic convergence (see numerical results in Section 6), is that for any value of ℓ/r the stiffness matrix is well-conditioned. This feature allows efficient computations on microcomputers having a short word length. By contrast, the conventional approach yields an ill-conditioned stiffness matrix when ℓ/r is large, in which case high precision computations are required to avoid ill-conditioning errors.

VI. NUMERICAL RESULTS. The modeling capabilities of the two-node, constant-strain element are illustrated through the solutions to static and free vibration arch and ring problems. The examples are specifically selected to test the shallowly curved element in the critical applications to moderately deep and highly deep arches, ranging in the slenderness from moderately thick to extremely thin. Uniform meshes are used throughout, and the results are normalized with respect to the appropriate Bernoulli-Euler solutions. Unless stated otherwise, $E/G = 2.6$ and $k^2 = \pi^2/12$.

Optimal C_s and C_m values. Our criterion for establishing suitable Cs and Cm values rests on the requirement of a rapid monotonic convergence in the energy. In the case of a straight Timoshenko element [17,22], Cs=1/3 evolves as a natural choice, since it yields an exact value of the potential energy (even with a single element!) for the problem of a cantilever beam under a tip load (refer to Figure 4 and Table 1). To obtain C_m, we solved a semicircular clamped arch under a central load, using models with C_s=1/3 and C_m = 0, 1/5, 1/4 and 1/3 . From the convergence of the tip deflection, which is a direct measure of the potential energy for this problem, C_m=1/4 appears nearly optimal.

465

Henceforth, $C_s = 1/3$ and $Cm = 1/4$ are adopted permanently for this element. Note that regardless of the C_s and C_m values, convergence is attainable by mesh refinement. However, in coarse and/or excessively restrained models [23], the optimal C_i values yield notable solution improvements.

Element vs. Beam Penalty Constraints. To illustrate the effect of the relaxation parameters on the element (local) and beam (global) strains in the thin regime, consider a thin ($L/h = 10^3$), straight cantilever beam under a tip load discretized with four elements. Three different solutions corresponding to $C_s = 0, 1/5, 1/3$ are obtained. The element-level shear strains are computed using (4.10). The beam-level shear strain is obtained from (2.3), where w and θ are exact polynomial fits to the respective nodal values, i.e.,

$$w = \sum_{k=0}^{4} a_{wk}(x/L)^k, \quad \theta = \sum_{k=0}^{4} a_{\theta k}(x/L)^k, \tag{6.1}$$

where a_{wk} and $a_{\theta k}$ are the exact fit coefficients for w and θ, respectively. The corresponding shear strains, the relaxation values, and the shear force resultants are summarized in Table 2. In accordance with the exact solution for this problem, the strains are constant along the span of the beam, from both the element-level and the beam-level calculations. At the element level, all three C_s values yield the correct shear strain and shear force resultant. It is often argued that the only 'natural' solution is obtained when the Kirchhoff constraint is enforced at the element level (i.e., $C_s = 0$ or $\phi^2 = 1$). However, at the global level, only $C_s = 1/3$ produces the correct shear strain and shear force resultant. Without relaxation the global behavior is grossly inaccurate. Importantly, it is the global enforcement of the Kirchhoff constraint that is the only requirement of the analytic theory; and only with relaxation is this global constraint possible.

Quarter-Circular Arch. Figure 5 depicts the convergence of the displacements at load application for a very thin, clamped arch ($R/h = 10^4$) which may be regarded as moderately deep. The convergence curves for the maximum stress resultants are shown in Figure 6. Recall that the stress resultants are constant across the element chord. Hence, their magnitudes are attributed to the center of the element. Rapid convergence of the displacement and stress variables is evident. Remarkably, the stress resultants have the same degree of accuracy as the displacements.

The effect of element slenderness is demonstrated in Figure 7, where the maximum displacement and stress variables are plotted versus the arch radius-to-thickness ratio (R/h). The range of the arch slenderness parameter is taken from 5 to 10^8. As expected, the element behavior is exceptional throughout the wide slenderness range, having no limitations in the thin regime (i.e. no locking). For the moderately thick arch ($R/h = 5$), the deviation from the Bernoulli-Euler solutions is due to transverse shear, as it should be. In Figure 8 are shown the variations of the correction parameters for the problem. These illustrate the increased influence of these corrections for a diminishing

466

thickness. It is this influence which is mainly responsible for the en-
hancement in element performance.

Semicircular Arch. Figure 9 shows the convergence of the maximum trans-
verse displacement for the moderately thick to very thin semicircular,
centrally loaded clamped arch. Highly accurate results are evident,
with the moderately thick case showing important contributions of both
shear and membrane deformations (and, naturally, deviating from the ele-
mentary solutions).

Deep Arches. The element modeling of very deep/thin arches is studied
on three simple cases (refer to Figure 10): (a) a complete ring (R/h =
10^3) pinched by two identical diagonal forces; (b) a $3\pi/2$-arch (R/h =
10^3) clamped at one end and restrained horizontally at the end where a
vertical force is prescribed; (c) a $7\pi/4$-arch (R/h = 23.48) clamped at
both ends and loaded centrally by a vertical force. The latter case is
only moderately thin, but presented here for the purpose of comparison
with solutions obtained in [14] for this problem. Due to symmetry in
arches (a) and (c) only one-quarter and one-half of the arches are
discretized, respectively. In all three cases the results are very
accurate with only a few elements. Notably, for arch (c) the present
element produces more accurate results than any of the thirteen elements
investigated in [14].

Vibration of Thin Ring. Natural frequencies of vibration for a thin
ring (R/h = 10^3) are computed for the purpose of assessing the quality
of the element consistent mass matrix. Again, due to symmetry, only
one-quarter of the ring is discretized. Several distinct values of C_m
and C_s are used to observe their effects on the frequency solutions.
The convergence curves for the fundamental frequency, shown in
Figure 11, illustrate the increased accuracy of curved-element modeling
over straight beam approximations. Importantly, the beneficial effects
of the membrane and shear relaxations are also evident, giving rise to
very accurate solutions even in coarse discretizations.

In Table 3 are summarized the eight lowest symmetric frequencies
obtained from an 8-element quarter-ring model. The results are compared
with the Bernoulli-Euler frequencies. As expected from a conforming
displacement model, the finite element frequencies converge from above;
the largest error is only 1% for the highest mode.

Vibration of Semicircular Arch. In this final example, natural fre-
quencies for a semicircular hinged arch are computed for the moderately
thick ($\pi R/r=10$) and moderately thin ($\pi R/r=50$) configurations. The solu-
tions, based on a full 60-element model, are compared with those of Dong
and Wolf [28], who used 30 straight three-noded Timoshenko elements,
having the same number of dof as our model. The results in Table 4 show
close agreement between the two solutions, however, the present frequen-
cies are consistently lower, hence more accurate. (It is worth mention-
ing that the Dong-Wolf isoparametric quadratic element, while generally
accurate in the slenderness range of this example, exhibits locking in a
truly slender regime.)

467

VII. CONCLUSIONS. We have presented a displacement formulation yielding simple, reliable, and efficient shallowly curved Timoshenko/-Marguerre elements. Particular attention was focused upon the simplest element of the hierarchical family -- a two-node, six degrees-of-freedom beam with constant components of strain. The element's main features are summarized as follows:

(1) kinematic conformability
(2) correct stiffness matrix rank
(3) free of straining rigid-body motion
(4) consistent load vector and mass matrix
(5) well-conditioned stiffness matrix
(6) nodal/dof simplicity and computational efficiency
(7) no limitation with regard to thinness, i.e., no locking of any type
(8) fast monotonic convergence of the kinematic and kinetic variables
(9) ideal microcomputer suitability.

The methodology could potentially be used to construct simple and efficient shear-deformable curved shell models that would be compatible with the curved beams discussed in this paper.

APPENDIX A. The components of the element matrices given below correspond to the nodal displacement vector $\{u_0, u_1, w_0, w_1, \theta_0, \theta_1\}$. Exact explicit integration has been used throughout.

Stiffness Matrix

$$k_{11} = k_{22} = -k_{12} = D_m^e/\ell, \quad k_{26} = k_{15} = -k_{16} = -k_{25} = D_m^e \beta,$$

$$k_{33} = k_{44} = -k_{34} = D_s^e/\ell, \quad k_{35} = k_{36} = -k_{45} = -k_{46} = D_s^e/2,$$

$$k_{55} = k_{66} = D_m^e \ell \beta^2 + D_b/\ell + D_s^e \ell/4, \quad k_{56} = -k_{55} + D_s^e \ell/2 \qquad (A.1)$$

where

$$\beta = (\beta_1 - \beta_0)/12,$$

$$D_i^e = \phi_i^2 D_i \quad (i = m, s), \quad D_m = AE, \quad D_s = k^2 GA, \quad D_b = EI,$$

$$\phi_m^2 = \frac{1}{1 + C_m (\ell/r)^2}, \quad \phi_s^2 = \frac{1}{1 + C_s \frac{1}{4} k^2 (G/E) (\ell/r)^2},$$

$$C_m = 1/4, \quad C_s = 1/3.$$

Consistent Mass Matrix

$$m_{11} = m_{22} = 2m_{12} = m/3, \quad m_{13} = -m_{14} = (3\beta_0 - 2\beta_1)m/60,$$

$$m_{15} = -m_{16} = -(2\beta_0 + \beta_1)m\ell/4, \quad m_{23} = -m_{24} = (2\beta_0 - 3\beta_1)m/60,$$

$$m_{25} = -m_{26} = -(\beta_0 + 2\beta_1)m\ell/4,$$

$$m_{33} = m_{44} = [(2\beta_0^2 - \beta_0\beta_1 + 2\beta_1^2)/70 + 1]m/3,$$

$$m_{34} = [(-2\beta_0^2 + \beta_0\beta_1 - 2\beta_1^2)/35 + 1]m/6,$$

$$m_{35} = -m_{36} = [19(\beta_1^2 - \beta_0^2)/420 + 1]m\ell/24,$$

$$m_{45} = -m_{46} = -m_{35} + m\ell/12,$$

$$m_{55} = m_{66} = [(5\beta_0^2 + 2\beta_0\beta_1 + 13\beta_1^2)/252 + 40(r/\ell)^2 + 1]m\ell^2/120,$$

$$m_{56} = -m_{55} + m\ell^2/2, \quad m = \rho A\ell. \tag{A.2}$$

Consistent Load Vector due to Uniform Normal Pressure q

$$f_1 = f_2 = 0, \quad f_3 = f_4 = q\ell/2, \quad f_5 = -f_6 = q\ell^2/12. \tag{A.3}$$

REFERENCES.

1. R. H. Gallagher, "The development and evaluation of matrix methods for thin shell structural analysis," Ph.D. Thesis, State University of New York (1966).

2. F. K. Bogner, R. L. Fox and L. A. Schmit, "A cylindrical shell discrete element"; AIAA J. 5, 745-750 (1967)

3. G. Cantin and R. Clough, "A curved cylindrical shell finite element," AIAA J., 6, 1057-1062 (1968).

4. D. G. Ashwell and A. B. Sabir, "Limitations of certain curved finite elements when applied to arches," International J. Mech. Sci., 13, 133-139 (1971).

5. D. G. Ashwell, A. B. Sabir and T. M.Roberts, "Further studies in the application of curved finite elements to circular arches" Internat. J. Mech. Sci., 13, 507-517 (1971).

6. A. B. Sabir and D. G. Ashwell "A comparison of curved beam finite elements when used in vibration problems", J. Sound and Vibration, 18, (4), 555-563 (1971).

7. D. J. Dawe, "Curved finite elements for the analysis of shallow and deep arches," Computers and Structures, 4, 559-580 (1974).

8. D. J. Dawe, "Numerical studies using circular arch finite elements," Computers and Structures, 4, 729-740 (1974).

9. I. Fried, "Shape functions and the accuracy of arch finite element," AIAAJ., 11(3), 287-291 (1973).

10. 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)

11. H. R. Meck, "An accurate polynomial displacement function for finite ring elements," Computers and Structures, 11, 265-269 (1980).

12. A. K. Noor and J. M. Peters, "Mixed models and reduced/selective integration displacement models for nonlinear analysis of curved beams," Internat. J. Numerical Meth. Engrg., 17, 615-631 (1981).

13. H. Stolarski and T. Belytschko, "Membrane locking and reduced integration for curved elements," J. Appl. Mech. 49, 172-176, (1982).

14. H. Stolarski and T. Belytschko, "Shear and membrane locking in curved C$^{\text{o}}$ elements," Computer Meth. Appl. Mech. Engrg., 41, 279-296, (1983).

15. G. Prathap and G. R. Bhashyam, "Reduced integration and the shear-flexible beam element," Internat. J. Numerical Meth. Engrg., 18, 195-210 (1982).

16. G. Prathap, "The curved beam/deep arch/finite ring element revisited," Internat. J. Numerical Meth. Engrg., 21, 389-407 (1985).

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

18. T. J. R. Hughes, R. L. Taylor and W. Kanoknukulchai, "A simple and efficient element for plate bending," Internat. J. Numerical Meth. Engrg., 11, 1529-1543 (1977).

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

20. A. Tessler, "On a conforming, Mindlin-type plate element," in IV-MAFELAP 1981 (ed., J. R. Whiteman), Academic Press, London, 119-126 (1982).

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

22. 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).

23. 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).

470

24. R. H. MacNeal, "A simple quadrilateral shell element," Computers & Structures, 8, 175-183 (1978).

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

26. S. Timoshenko, "On the correction for shear of the differential equation for transverse vibrations of prismatic bars," Phil. Mag. 41, 744-746 (1921).

27. O. C. Zienkiewicz, The finite element method, McGraw-Hill, London, 3rd ed. (1977).

28. S. B. Dong and J. A. Wolf, "Effect of transverse shear deformation on vibrations of planar structures composed of beam-type elements," J. Acoust. Soc. Am. 53, 120-127 (1973).

FIGURES
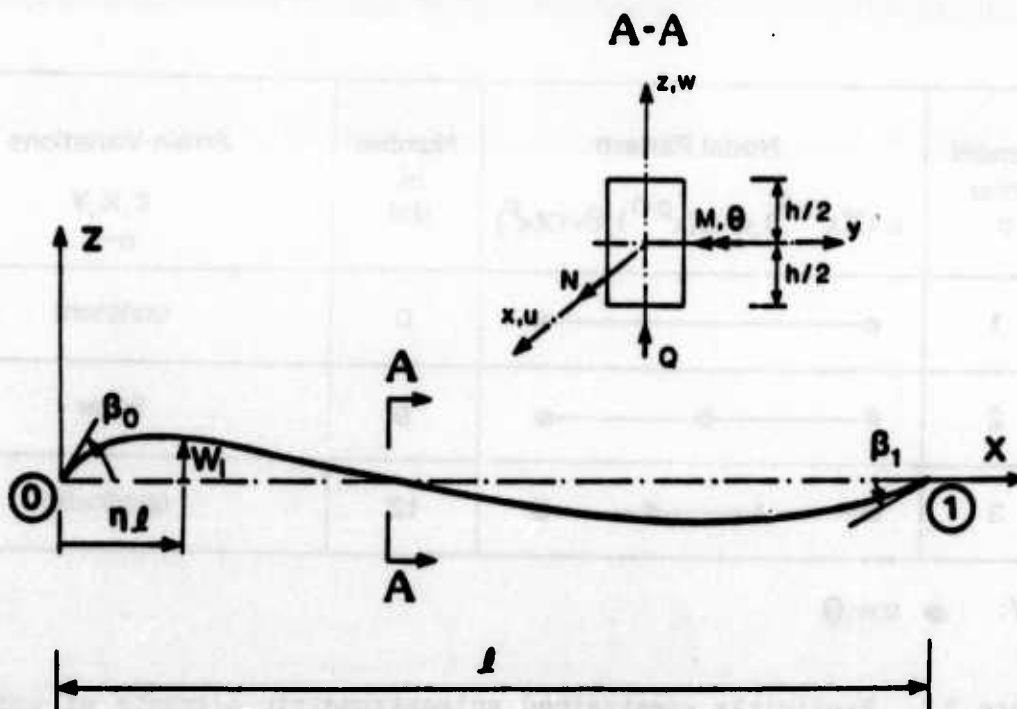
Figure 1.   Shallowly curved beam element.

Element Order p	Nodal Pattern $u=O(x^{p+3}), w=O(x^{p+1}), \theta=O(x^p)$	Number of dof	Strain Variations ε $p+2$	κ $p-1$	γ p
1		10	cubic	constant	linear
2		13	quartic	linear	quadratic
3		16	quintic	quadratic	cubic

KEY: ● u,w,θ △ u,w □ u

Figure 2. Initial (unconstrained) anisoparametric elements.

Element Order p	Nodal Pattern $u=O(x^{p+3}), w=O(x^{p+1}), \theta=O(x^p)$	Number of dof	Strain Variations $\varepsilon, \kappa, \gamma$ $p-1$
1		6	constant
2		9	linear
3		12	quadratic

KEY: ● u,w,θ

Figure 3. Explicitly constrained anisoparametric elements of uniform nodal patterns and uniform strain variations.

472
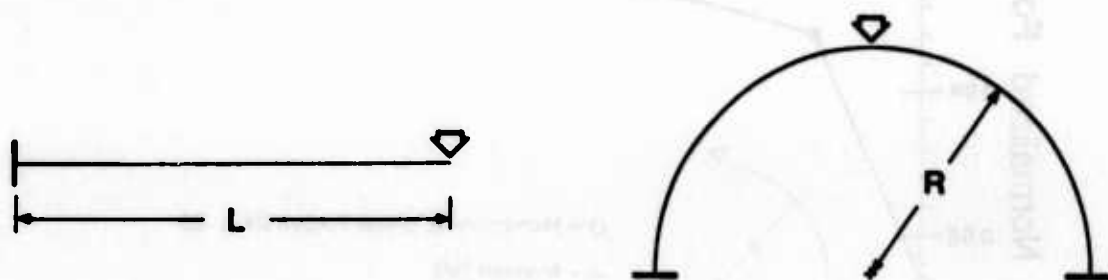
Figure 4.   Straight cantilever beam and semicircular arch.
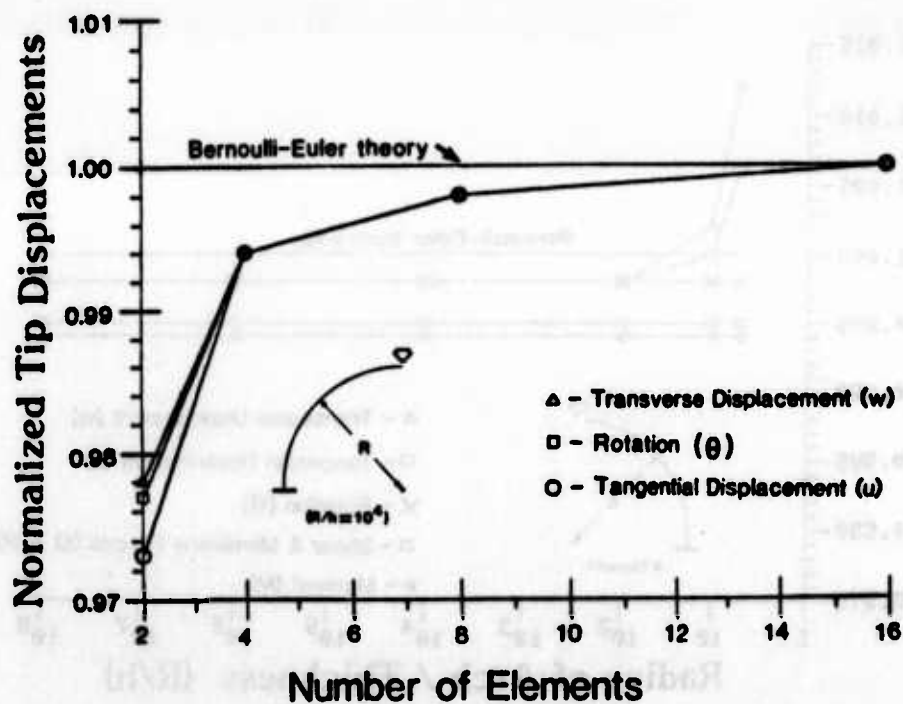
Figure 5.   Thin quarter-circular cantilever arch under tip load; convergence of tip displacements.

473
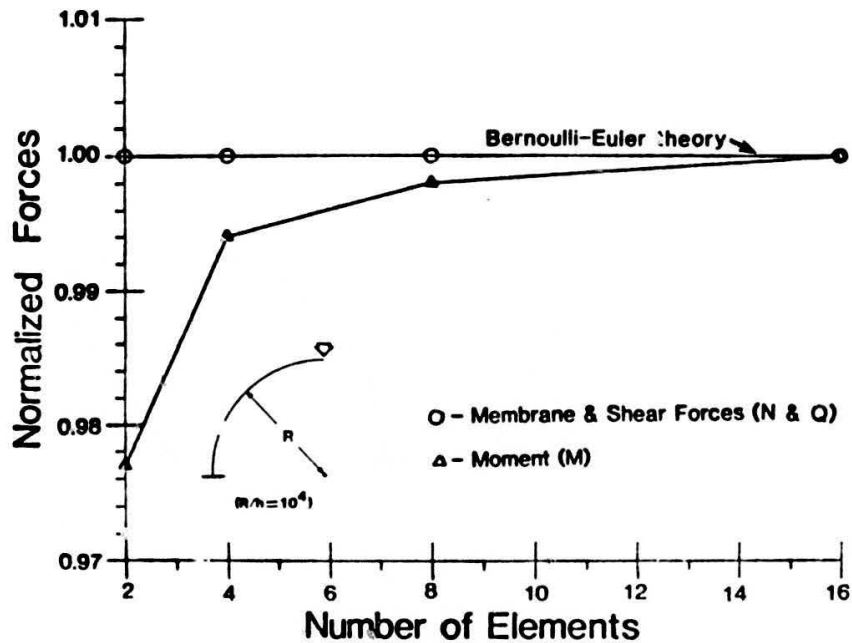
Figure 6. Thin quarter-circular cantilever arch under tip load; convergence of maximum stress resultants (N and M computed at center of element closest to clamped end; Q computed at center of element closest to load).
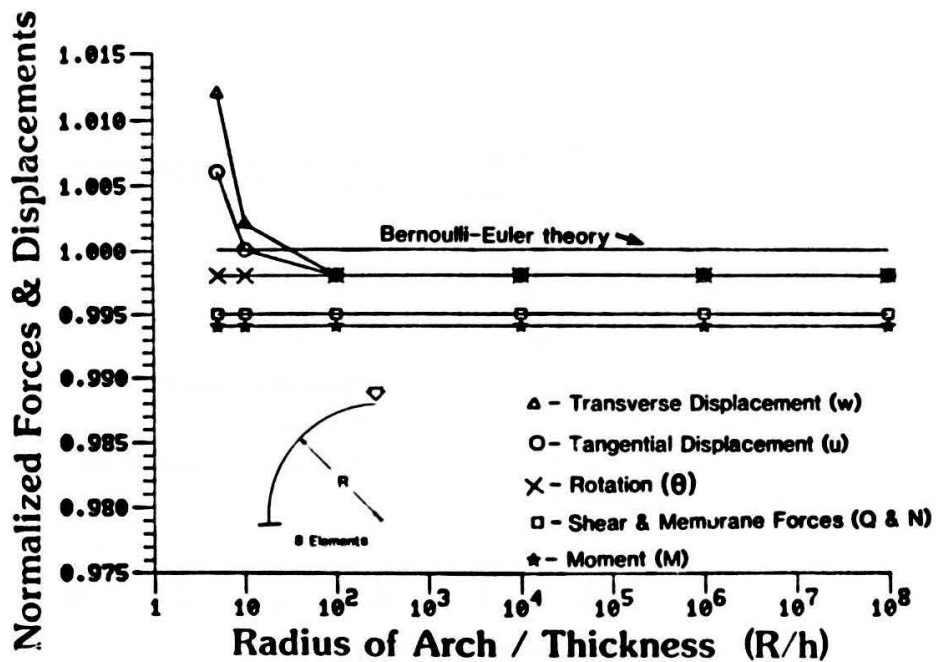
Figure 7. Quarter-circular cantilever arch under tip load; maximum displacements and stress resultants versus arch radius-to-thickness ratio.

474
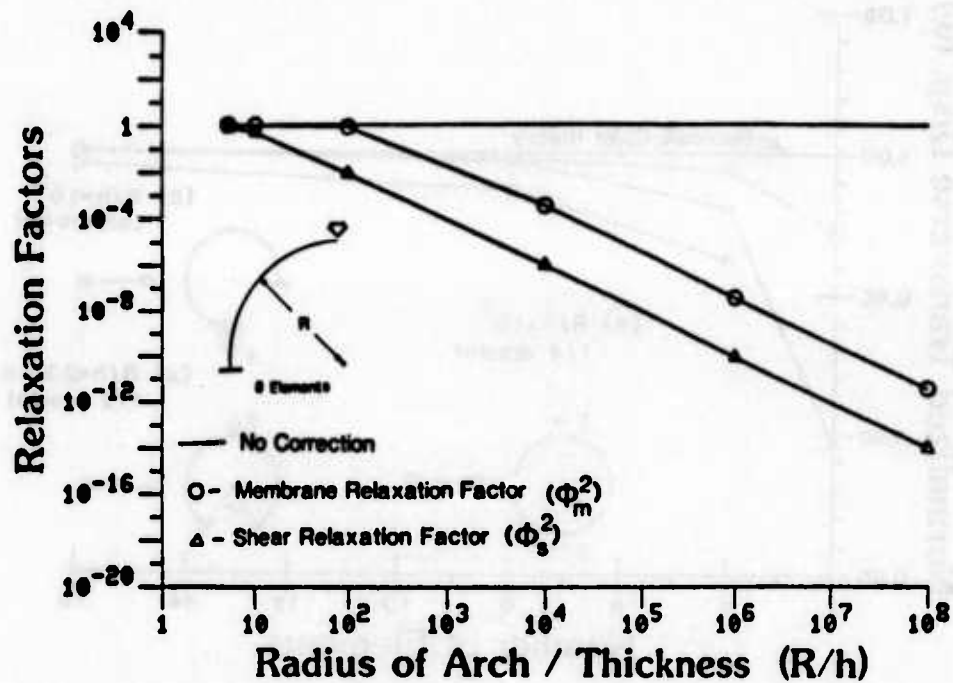
Figure 8. Quarter-circular cantilever arch under tip load; membrane and shear correction (relaxation) parameters versus arch radius-to-thickness ratio.
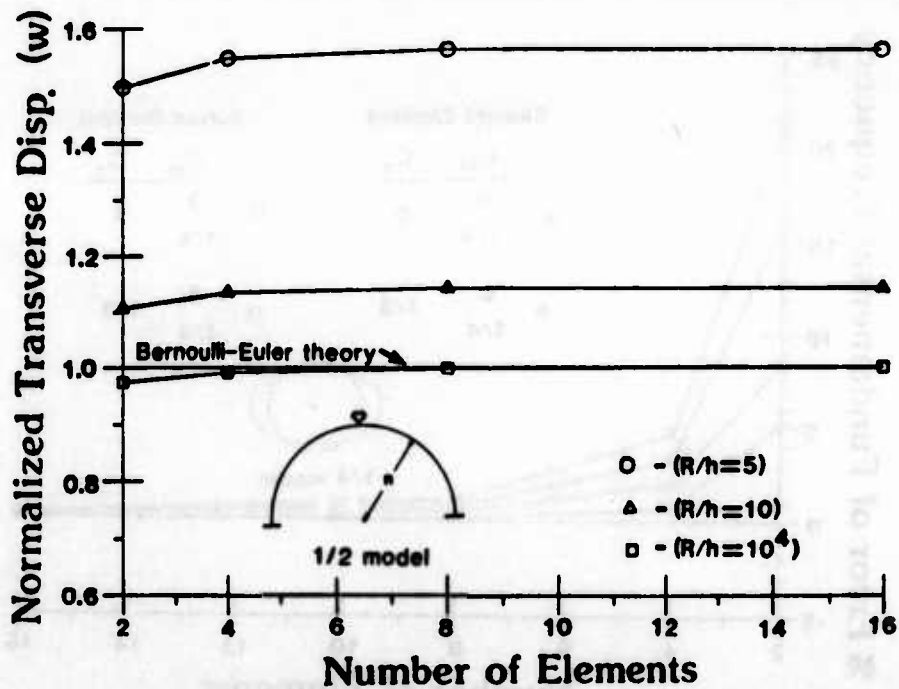
Figure 9. Moderately thick and very thin semicircular centrally loaded clamped arches; convergence of maximum transverse displacement.
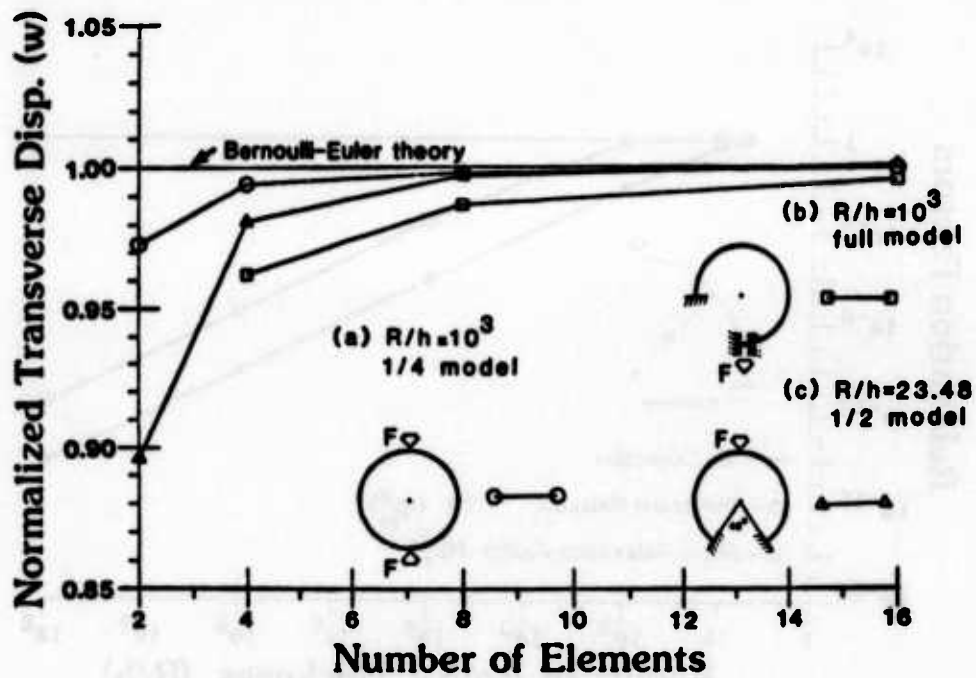
Figure 10. Deep/thin arches; convergence of maximum transverse displacement.
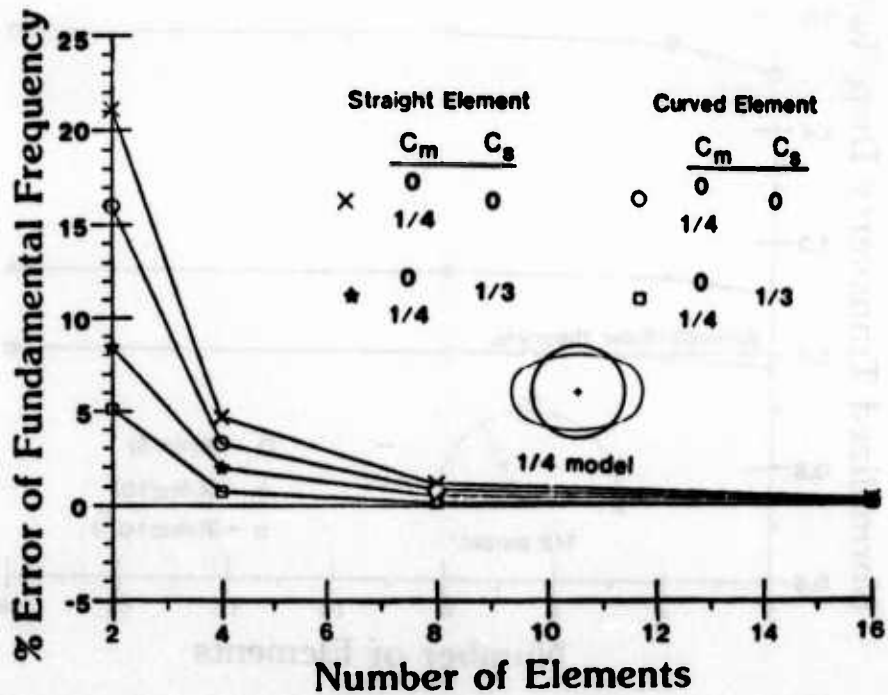
Figure 11. Thin ring; convergence of fundamental frequency.

Table 1. Determination of C_s and C_m constants for two-node element via optimization of potential energy ($\Pi_{exact} = 1.000$).

No. of el.	Straight cantilever $(L/h=10^4, C_m=0)$				Semicircular arch $(R/h=10^4, C_s=1/3)$			
	$C_s=0$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$C_m=0$	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$
2	0.938	0.975	0.984	1.000	0.903	0.961	0.975	0.999
4	0.984	0.994	0.996	1.000	0.989	0.993	0.993	0.995
8	0.996	0.998	0.999	1.000	0.998	0.998	0.999	0.999
16	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Thin, straight cantilever beam under unit tip force discretized with four elements. Element and beam shear strains for various C_s ($L=100$, $h=0.1$, $G=15.385 \times 10^6$).

C_s	Element			Beam	
	ϕ_s^2	$\phi_s^2 \gamma_0^\ell$	$Q=k^2 GA \phi_s^2 \gamma_0^\ell$	γ_0^g	$Q=k^2 GA \gamma_0^g$
0	1.000			-1.562×10^{-2}	-1.976×10^4
$\frac{1}{5}$	8.432×10^{-5}	7.903×10^{-7}	1.000	-6.249×10^{-3}	-7.905×10^3
$\frac{1}{3}$	5.058×10^{-5}			7.903×10^{-7}	1.000
Analytic		7.903×10^{-7}	1.000	7.903×10^{-7}	1.000

Table 3. Natural frequencies for thin circular ring

$$\omega_n = \underline{\omega}_n (EI/\rho AR^4)^{\frac{1}{2}}.$$

Mode number n	Bernoulli-Euler theory $\underline{\omega}_n$	Present element $\underline{\omega}_n^{fe}$	%Error $100 \times (\underline{\omega}_n^{fe}-\underline{\omega}_n)/\underline{\omega}_n$
2	2.6833	2.6844	0.04
4	14.552	14.559	0.05
6	34.524	34.546	0.07
8	62.514	62.588	0.12
10	98.509	98.737	0.23
12	142.51	143.12	0.43
14	194.50	195.97	0.76
16	254.50	257.63	1.23

Table 4. Natural frequencies for moderately thick and thin semicircular hinged arches $\omega_n = \underline{\omega}_n (EI/\rho\pi^* AR^*)^{\frac{1}{2}}$.

Slenderness ratio $\pi R/r$	Mode number n	Dong & Wolf[28] $\underline{\omega}_n^{DW}$	Present element $\underline{\omega}_n$	% Difference $100 \times (\underline{\omega}_n^{DW} - \underline{\omega}_n)/\underline{\omega}_n^{DW}$
	1	12.87	12.79	0.62
	2	24.69	24.64	0.20
	3	38.13	37.98	0.39
	4	40.01	39.90	0.27
10	5	57.62	57.13	0.85
	6	60.42	59.87	0.91
	7	60.83	60.47	0.59
	8	73.40	72.80	0.82
	9	81.03	80.74	0.36
	10	85.18	84.74	0.52
	1	21.60	21.53	0.32
	2	62.57	62.35	0.35
	3	120.8	120.3	0.41
	4	146.3	146.4	-0.07
	5	199.6	199.2	0.20
50	6	212.0	212.1	-0.05
	7	278.2	277.7	0.18
	8	335.4	334.8	0.18
	9	368.8	368.9	-0.03
	10	439.5	438.2	0.03

478

ADMISSIBLE ELASTIC ENERGY DENSITY FUNCTIONS
FOR ELASTOMER SOLIDS

I. Fried
Department of Mathematics
Boston University, 111 Cummington St.
Boston, MA 02215

A. R. Johnson and C. J. Quigley
United States Army Laboratory Command
Army Materials Technology Laboratory
Watertown, MA 02172-0001

ABSTRACT. Admissible elastic energy density functions for highly deformed, compressible, elastomer solids are derived from the geometric - arithmetic mean value inequality theorem. A finite element model based on these energies is proposed for triangular elements with quadratic displacement approximations. The computation of very large deformations of an elastomer cylinder is performed.

INTRODUCTION. The assumption of incompressibility for elastomers is usually made for analytical convenience. For numerical finite element computations [1] incompressibility is a disaster. Having pressure independent of the displacements and its inclusion in the elastic variational formulation via Lagrange multipliers results in loss of the all important energy positive definiteness. Without a minimal variational principal finite element convergence and stability becomes precarious.

In the numerical modeling of nearly incompressible elastic solids pressure is computed from dilatation with the bulk modulus. Since the bulk modulus is large there are serious computational difficulties associated with this approach. Decline in the condition of the system of stiffness equations is clearly identified [2] in its dependence on the formulation and discretization parameters. The adverse effect of a large bulk modulus on the approximation accuracy of the finite element model is also understood. The modeling of near incompressibility must be balanced [3] with the computational considerations of numerical stability and accuracy.

This paper has three points to make. The first is to show how the arithmetic - geometric - mean inequality theorem rationally leads to most of the elastic energy density functions in use for the large displacement analysis of nearly incompressible solids (typically elastomers). The

479

second is to estimate, using a uniaxial stretching problem, the upper limit for the bulk modulus needed to reproduce the observed residual compressiblity of rubberlike (elastomer) solids. The third point is to develop a quadratic six node finite element for the analysis of elastomers which uses the positive definite energy density functions derived in the first part.

POSITIVE DEFINITE ENERGY DENSITY FUNCTIONS. We turn now to the question of the specific form for the energy density function w. We will show that modifications to the most widely used energy density functions can be made using the following theorem [4] so that they represent positive definite functionals for any combination of stretch ratios.

Arithmetric - Geometric - Mean Inequality Theorem:

Let a,b,c be positive real numbers. Then

$$\frac{1}{3}(a+b+c) \geq (abc)^{1/3} \tag{1}$$

where equality holds only when a=b=c.   END THM.

Hence the function

$$F(a,b,c) = a+b+c-3(abc)^{1/3} \tag{2}$$

is positive semidefinite.

To apply this theorem to our purpose we consider three dimensional finite elasticity and write λ_1, λ_2, λ_3 for the principal stretch ratios. Then, for any real r and $a=\lambda_1^r$, $b=\lambda_2^r$, $c=\lambda_3^r$ we have from equation (2) that

$$F_r(\lambda_1,\lambda_2,\lambda_3) = \lambda_1^r + \lambda_2^r + \lambda_3^r - 3(\lambda_1\lambda_2\lambda_3)^{r/3} \tag{3}$$

is positive semi definite and can be taken as a summand in an energy density function for an isotropic compressible solid. When perfectly incompressible, the solid is with $\lambda_1\lambda_2\lambda_3 = 1$ and

$$F = \lambda_1^r + \lambda_2^r + \lambda_3^r - 3 \tag{4}$$

but the stretch ratios in this expression are not independent and F_r is not positive definite unless one of the stretch ratios is expressed in terms of the other two through $\lambda_1\lambda_2\lambda_3 = 1$. For instance, if λ_3 is eliminated from F_r in equation (4) by $\lambda_3 = 1/(\lambda_1\lambda_2)$, then

$$F_r = \lambda_1^r + \lambda_2^r + 1/(\lambda_1^r \lambda_2^r) - 3 \qquad (5)$$

is such that $F_r > 0$ if $\lambda_1 > 0$ and $\lambda_2 > 0$; and $F_r = 0$ only when $\lambda_1 = \lambda_2 = 1$. In the compressible case when $r = 1$

$$F_1 = \lambda_1 + \lambda_2 + \lambda_3 - 3(\lambda_1 \lambda_2 \lambda_3)^{1/3} \qquad (6)$$

corresponds to the incompressible expression for the elastic energy density function of Varaga [5]. A more general energy density functional, w, is conveniently written as a polynomial of F_1

$$w = c_1 F_1 + c_2 F_1^2 + \cdots + c_n F^n \ , \ c_1 > 0 \qquad (7)$$

Choosing r=2 in equation (3) results in

$$F = \lambda_1^2 + \lambda_2^2 + \lambda_3^2 - 3(\lambda_1 \lambda_2 \lambda_3)^{2/3} \qquad (8)$$

which is the generalization of the NeoHookean expression for a compressible solid, while r=-2 produces

$$F_{-2} = \lambda_1^{-2} + \lambda_2^{-2} + \lambda_3^{-2} - 3(\lambda_1 \lambda_2 \lambda_3)^{-2/3} \qquad (9)$$

so that

$$w = c_1 F_2 + c_2 F_{-2} \qquad (10)$$

is the compressible counterpart to the Mooney-Rivlin energy density function.

The strain invariants

$$I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$$
$$I_2 = \lambda_1^2 \lambda_2^2 + \lambda_2^2 \lambda_3^2 + \lambda_3^2 \lambda_1^2 \qquad (11)$$
$$I_3 = \lambda_1^2 \lambda_2^2 \lambda_3^2$$

are commonly used in the literature on rubber elasticity. We can use the strain invariants as follows to construct energy density functionals. Setting $a = \lambda_1^2$, $b = \lambda_2^2$, $c = \lambda_2^2$ we obtain the inequality

481

$$I_1 - 3I_3^{1/3} \geq 0 \tag{12}$$

with equality holding if and only if $\lambda_1 = \lambda_2 = \lambda_3$. In the same manner, setting $a = \lambda_1^2 \lambda_2^2$, $b = \lambda_2^2 \lambda_3^2$, $c = \lambda_3^2 \lambda_1^2$ produces the second inequality

$$I_2 - 3I_3^{2/3} \geq 0 \tag{13}$$

with equality holding if and only if $\lambda_1 = \lambda_2 = \lambda_3$.

The modified Mooney-Rivlin energy density function

$$w = c_1(I_1 - 3I_3^{1/3}) + c_2(I_2 - 3I_3^{2/3}), \qquad c_1, \ c_2 > 0 \tag{14}$$

is then only positive semidefinite; $w = 0$ when $\lambda_1 = \lambda_2 = \lambda_3$ even when $\lambda_i \neq 1$. To have a positive definite energy we have to add to it a dilatoric contribution representing the energy stored in compression. We use

$$w = \frac{1}{2}\hat{\lambda}[\ln(\lambda_1 \lambda_2 \lambda_3)]^2 \tag{15}$$

where $\hat{\lambda}$ is the bulk modulus.

Ogden's energy density function [6-8] is essentially F_r with r being rational and fitted to experimental data.

The Valanis-Landel energy density function [9]

$$w = 2\mu \sum_{i=1}^{3} \lambda_i(\ln\lambda_i - 1) + \frac{1}{2}\hat{\lambda}[\ln(\lambda_1 \lambda_2 \lambda_3)]^2 \tag{16}$$

is not derivable from the arithmetric-geometric-mean inequality theorem but is based on the inequality

$$\lambda\ln\lambda \geq \lambda - 1, \qquad \lambda > 0 \tag{17}$$

where equality holds only when $\lambda = 1$. Figure 1 shows the variation of

$$\phi(\lambda) = \lambda\ln\lambda - \lambda + 1 \tag{18}$$

with λ, and indeed $\phi(\lambda) > 0$ when $\lambda > 0$ and $\lambda \neq 1$. Hence the deviatoric part, factored by 2μ, is positive definite and one may set $\hat{\lambda} = 0$ in equation (16) for the Valanis-Landel energy density function.

<u>NEAR INCOMPRESSIBILITY.</u> Suppose that we choose to represent the elastic behavior of our compressible rubber by the modified Mooney-Rivlin energy density function

$$w = c_1(I_1 - 3I_3^{1/3}) + c_2(I_2 - 3I_3^{2/3}) + \frac{1}{2}\hat{\lambda}[\ln(\lambda_1\lambda_2\lambda_3)]^2 \tag{19}$$

or by that of Valanis-Landel in equation (16). How should we select the bulk modulus $\hat{\lambda}$ to accommodate the observed compressibility of rubber? It is important in the finite element modeling of rubber to keep the bulk modulus as low as accuracy demands will allow. To answer the question we shall numerically solve the uniaxial tension problem with the different energy density functions and estimate the bulk modulus needed to reproduce the dilatational experiments of Penn [10].

Consider the simple tension bar of Figure 2 having a cross section area $\varepsilon^2 \ll 1$ and unit length. Stretched by an axial force P the bar extends to length λ_3 with cross sectional area $\varepsilon^2\lambda_1^2$. The potential energy of the stretched bar is given by

$$\pi = \varepsilon^2 w(\lambda_1, \lambda_3) - P(\lambda_3 - 1) \tag{20}$$

and the two equations of equilibrium

$$\frac{\partial \pi}{\partial \lambda_1} = 0 \quad \text{and} \quad \frac{\partial \pi}{\partial \lambda_3} = 0 \tag{21}$$

become

$$\frac{\partial w}{\partial \lambda_1} = 0 \quad \text{and} \quad \frac{\partial w}{\partial \lambda_3} - \sigma = 0 \tag{22}$$

where $\sigma = P/\varepsilon^2$ is the axial stress.

The first of equations (22) becomes for the Mooney-Rivlin material

$$0 = (\lambda_1^4 - \lambda_1^2 I) + (\frac{c_2}{c_1})(\lambda_1^6 + I^3 - 2\lambda_1^2 I^2) + \frac{3}{4}\lambda_1^2(\frac{\hat{\lambda}}{c_1})\ln(I) \tag{23}$$

where $\quad I = (\lambda_1^2\lambda_3)^{2/3}$

For any given $\lambda_1 > 0$ equation (23) is solved for I, and hence for λ_3 by the Newton-Raphson method.

For the Valanis-Landel function the first of equations (22) produces

$$\ln(\lambda_1^2\lambda_3) = -\frac{2\mu}{\hat{\lambda}}\lambda_1\ln\lambda_1 \tag{24}$$

483

that yields λ_3 for given values of λ_1.

Figure 3 traces the relative volume change $\lambda_1\lambda_2\lambda_3-1$ vs λ_3 for the Valanis-Landel function with $\hat{\lambda}/2\mu = 1000$, and the Mooney-Rivlin model with $c_2/c_1 = 0.1$ and $\hat{\lambda}/c_1 = 1350$, compared with the experiments of Penn. These values of $\hat{\lambda}$ are reasonable upper limits for numerical modeling of nearly incompressible rubber.

QUADRATIC ELEMENT. A triangular three node bilinear axisymmetric element for total Lagrangian analysis of elastomers is given in reference [11]. The quadratic six node element is shown in Figure 4. We use quadratic interpolation of r and z (deformed coordinates) over triangles in the (α,β) plane (undeformed coordinates). The discrete variables associated with the element are

$$\vec{e}^T = (r_1, z_1, r_2, z_2, \cdots, r_6, z_6) \tag{25}$$

where r and z are the values of r and z at node number 1, etc. The element nodal point numbering is shown in Figure 4.

The interpolations are described in terms of the local coordinate system (ξ,η) shown in Figure 5. The mapping from the (α,β) plane onto the (ξ,η) plane is done with

$$\alpha = \alpha_1(1-\xi-\eta)+\alpha_2\xi+\alpha_3\eta$$
$$\beta = \beta_1(1-\xi-\eta)+\beta_2\xi+\beta_3\eta \tag{26}$$

and r and z are interpolated inside the element by

$$\vec{r}^T = \vec{e}^T\vec{\phi} \quad , \quad \vec{z}^T = \vec{e}^T\vec{\psi} \tag{27}$$

where the shape function vectors are

$$\vec{\phi}^T = (\phi_1,0,\phi_2,0,\phi_3,0,\phi_4,0,\phi_5,0,\phi_6,0) \tag{28}$$

and

$$\vec{\psi}^T = (0,\phi_1,0,\phi_2,0,\phi_3,0,\phi_4,0,\phi_5,0,\phi_6) \tag{29}$$

with the six shape functions

$$
\begin{Bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \\ \phi_6 \end{Bmatrix}
=
\begin{bmatrix}
1 & -3 & -3 & 2 & 4 & 2 \\
 & -1 & & 2 & & \\
 & & -1 & & & 2 \\
 & 4 & & -4 & -4 & \\
 & & & & 4 & \\
 & & 4 & & -4 & -4
\end{bmatrix}
\begin{Bmatrix} 1 \\ \xi \\ \eta \\ \xi^2 \\ \xi\eta \\ \eta^2 \end{Bmatrix}
\tag{30}
$$

484

As for the partial derivatives of ϕ with respect to ξ and η we have

$$
\begin{bmatrix}
\phi_{1\xi} \\
\phi_{2\xi} \\
\phi_{3\xi} \\
\phi_{4\xi} \\
\phi_{5\xi} \\
\phi_{6\xi}
\end{bmatrix}
=
\begin{bmatrix}
-3 & 4 & 4 \\
-1 & 4 & \\
 & & \\
4 & -8 & -4 \\
 & & 4 \\
 & & -4
\end{bmatrix}
\begin{bmatrix}
1 \\
\xi \\
\eta
\end{bmatrix}
\tag{31}
$$

and

$$
\begin{bmatrix}
\phi_{1\eta} \\
\phi_{2\eta} \\
\phi_{3\eta} \\
\phi_{4\eta} \\
\phi_{5\eta} \\
\phi_{6\eta}
\end{bmatrix}
=
\begin{bmatrix}
-3 & 4 & 4 \\
 & & \\
-1 & & 4 \\
 & -4 & \\
 & 4 & \\
4 & -4 & -8
\end{bmatrix}
\begin{bmatrix}
1 \\
\xi \\
\eta
\end{bmatrix}
\tag{32}
$$

By the chain rule of partial differentiation we have from equation (26) that

$$
\begin{bmatrix}
\dfrac{\partial}{\partial \xi} \\[2mm]
\dfrac{\partial}{\partial \eta}
\end{bmatrix}
=
\begin{bmatrix}
\alpha_2 - \alpha_1 & \beta_2 - \beta_1 \\
\alpha_3 - \alpha_1 & \beta_3 - \beta_1
\end{bmatrix}
\begin{bmatrix}
\dfrac{\partial}{\partial \alpha} \\[2mm]
\dfrac{\partial}{\partial \beta}
\end{bmatrix}
\tag{33}
$$

and the Jacobian of this mapping is

$$
\delta = (\alpha_2 - \alpha_1)(\beta_3 - \beta_1) - (\alpha_3 - \alpha_1)(\beta_2 - \beta_1)
\tag{34}
$$

Inversion of equation (33) yields

$$
\begin{bmatrix}
\dfrac{\partial}{\partial \alpha} \\[2mm]
\dfrac{\partial}{\partial \beta}
\end{bmatrix}
=
\frac{1}{\delta}
\begin{bmatrix}
\beta_3 - \beta_1 & -(\beta_2 - \beta_1) \\
-(\alpha_3 - \alpha_1) & \alpha_2 - \alpha_1
\end{bmatrix}
\begin{bmatrix}
\dfrac{\partial}{\partial \xi} \\[2mm]
\dfrac{\partial}{\partial \eta}
\end{bmatrix}
\tag{35}
$$

Differentiating r and z in equation (27) with respect to ξ and η leads to

$$
r_\xi = \vec{e}^{\,T} \vec{\phi}_\xi \qquad , \qquad r_\eta = \vec{e}^{\,T} \vec{\phi}_\eta
$$

$$
z_\xi = \vec{e}^{\,T} \vec{\psi}_\xi \qquad , \qquad z_\eta = \vec{e}^{\,T} \vec{\psi}_\eta
\tag{36}
$$

485

and with equation (35) we get that

$$r_\alpha = \vec{e}^T\vec{p} \ , \ \vec{p} = \frac{1}{\delta}((\beta_3-\beta_1)\vec{\phi}_\xi-(\beta_2-\beta_1)\vec{\phi}_\eta)$$

$$z_\alpha = \vec{e}^T\vec{q} \ , \ \vec{q} = \frac{1}{\delta}((\beta_3-\beta_1)\vec{\psi}_\xi-(\beta_2-\beta_1)\vec{\psi}_\eta)$$

$$r_\beta = \vec{e}^T\vec{r} \ , \ \vec{r} = \frac{1}{\delta}(-(\alpha_3-\alpha_1)\vec{\phi}_\xi+(\alpha_2-\alpha_1)\vec{\phi}_\eta) \tag{37}$$

$$z_\beta = \vec{e}^T\vec{s} \ , \ \vec{s} = \frac{1}{\delta}(-(\alpha_3-\alpha_1)\vec{\psi}_\xi+(\alpha_2-\alpha_1)\vec{\psi}_\eta)$$

where δ is the Jacobian of the mapping from (α,β) onto (ξ,η) as given by equation (34).

The principal stretch ratios are given by (see reference [11])

$$\lambda_1^2 = \frac{1}{2}(A+B+((A-B)^2+4C^2)^{1/2})$$

$$\lambda_2^2 = \frac{1}{2}(A+B-((A-B)^2+4C^2)^{1/2}) \tag{38}$$

$$\lambda_3^2 = \frac{r^2}{\alpha^2}$$

where

$$A = r_\alpha^2 + z_\alpha^2 \qquad\qquad B = r_\beta^2 + z_\beta^2 \tag{39}$$

and $\quad C = r_\alpha r_\beta + z_\alpha z_\beta$

Then

$$A = \vec{e}^T[\vec{p}\vec{p}^T]\vec{e} + \vec{e}^T[\vec{q}\vec{q}^T]\vec{e}$$

$$B = \vec{e}^T[\vec{r}\vec{r}^T]\vec{e} + \vec{e}^T[\vec{s}\vec{s}^T]\vec{e} \tag{40}$$

and $\quad C = \vec{e}^T[\vec{p}\vec{r}^T + \vec{r}\vec{p}^T]\vec{e}$

The remaining derivation of the element gradient and tangent matrices is identical to the derivation presented in reference [11] for the bilinear element except for the numerical integration. The elastic energy per one radian of a typical element is

$$E = \int w(\lambda_1,\lambda_2,\lambda_3)\alpha d\alpha d\beta \tag{41}$$

where α is given in equation (26). Then, in the (ξ,η) plane

486

$$E = \delta \int_{\xi=0}^{1} \int_{\eta=0}^{1-\xi} w \alpha \, d\eta \, d\xi \tag{42}$$

where δ is given by equation (34). The integration of (42) is done numerically by sampling w and α at the three integration points

$$\left(\frac{1}{6}, \frac{1}{6} \right), \left(\frac{4}{6}, \frac{1}{6} \right), \left(\frac{1}{6}, \frac{4}{6} \right)$$

with equal weights of $1/6$.

COMPRESSION OF A CYLINDER USING QUADRATIC ELEMENTS. An analysis of a rubber cylinder in compression was given in [11] for bilinear elements. The Valanis-Landel energy density function was used. To demonstrate the modified Mooney-Rivlin energy density function and the quadratic element we again analyze the end thrust (compression) of a cylinder. To compute the element gradient and tangent matrices we need the derivatives of the energy density function with respect to the stretch ratios. Let $I = I_1$, $J = I_2$ and $K = I_3$ in equation (19) and

$$(\ )_i = \frac{\partial (\ )}{\partial \lambda_i} \tag{43}$$

Then,

$$I_i = 2\lambda_i \quad , \quad I_{ii} = 2 \quad , \quad I_{ij} = 0$$

$$J_i = 2\lambda_i(I-\lambda_i^2) \quad , \quad J_{ii} = 2(I-\lambda_i^2) \quad , \quad J_{ij} = 4\lambda_i\lambda_j \tag{44}$$

and $\quad K_i = \frac{2}{\lambda_i}K \quad , \quad K_{ii} = \frac{2}{\lambda_i^2}K \quad , \quad K_{ij} = \frac{4}{\lambda_i\lambda_j}K$

Computing the derivatives we have

$$w_i = 2c_1(\lambda_i - \frac{1}{\lambda_i}K^{1/3}) + 2c_2(\lambda_i(I-\lambda_i^2) - \frac{2}{\lambda_i}K^{2/3}) + \frac{2\lambda}{\lambda_i}\ln(K) \tag{45}$$

$$w_{ij} = -\frac{4c_1}{3}\frac{1}{\lambda_i\lambda_j}K^{1/3} + 4c_2(\lambda_i\lambda_j - \frac{4}{3}\frac{1}{\lambda_i\lambda_j}K^{2/3}) + 4\lambda\frac{1}{\lambda_i\lambda_j} \quad i \neq j \tag{46}$$

and $\quad w_{ii} = 2c_1(1 + \frac{K^{1/3}}{3\lambda_i^2}) + 2c\left((I-\lambda_i^2) - \frac{2}{3}\frac{K^{2/3}}{\lambda_i^2}\right) + \frac{2\lambda}{\lambda_i^2}(2-\ln(K)) \tag{47}$

Figure 6 shows the original and deformed mesh for a cylinder restrained from slipping at the top and compressed to 75% of its height. The solution was obtained using the Newton-Raphson method as described in reference [11].

REFERENCES.

[1] I. Fried, "Reflections on computational approximation of elastic incompressibility", Computers and Structures, 17, 1983, 161-168.

[2] I. Fried, "Influence of Poisson's ratio on the condition of the stiffness matrix", Int. J. Solids and Structures, 9, 1973, 323-329.

[3] I. Fried, "Finite element analysis of incompressible material by residual energy balancing", Int. J. Solids and Structures, 10,19.4, 993-1002.

[4] E. Beckenback and R. Bellman, An Introduction to Inequalities, Random House, NY, 1961.

[5] O. H. Varga, Stress-Strain Behavior of Elastic Materials, J. Wiley, NY, 1966.

[6] R. W. Ogden and P. Chadwick, "On the deformation of solid and tubular cylinder of incompressible isotropic material", J. Mech. Phys. Solids, 20, 1972, 77-90.

[7] R. W. Ogden, "Volume changes associated with the deformation of rubberlike solids", J. Mech. Phys. Solids, 24, 1976, 323-338.

[8] R. W. Ogden, "Nearly isochoric elastic deformations: applications to rubberlike solids", J. Mech. Phys. Solids, 26, 1978, 37-57.

[9] K. C. Valanis and R. F. Landel, "The strain-energy function of a hyperelastic material in terms of the extension ratios", J. Appl. Physics, 38, 1967, 2997-3002.

[10] R. W. Penn, "Volume changes accompanying the extension of rubber", Trans. Soc. Rheology, 14, 1970, 509-517.

[11] A. R. Johnson, C. J. Quigley and I. Fried, "Large deformations of elastomer cylinders subjected to end thrust and probe penetration", Trans. Third Army Conf. Applied Mathematics and Computing, Georgia Inst. Tech., Atlanta, GA, 1985.
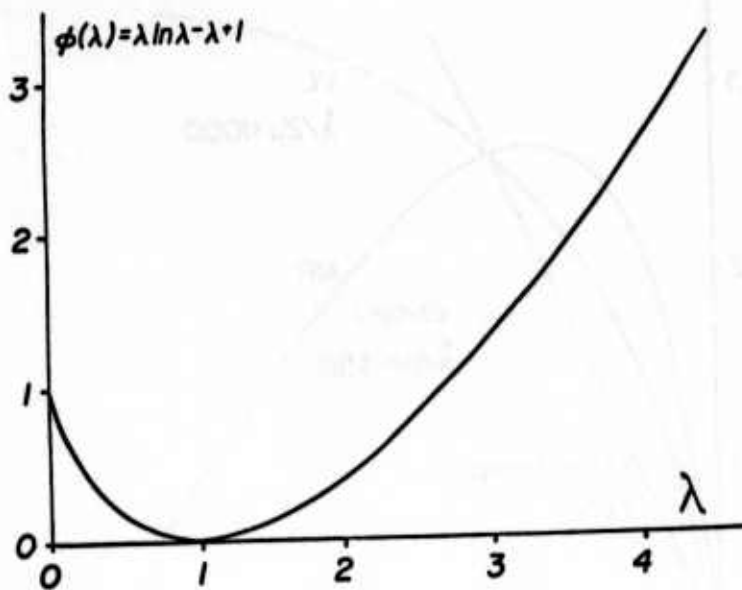
Figure 1. Variation of $\phi(\lambda) = \lambda\ln(\lambda)-\lambda+1$ with λ.
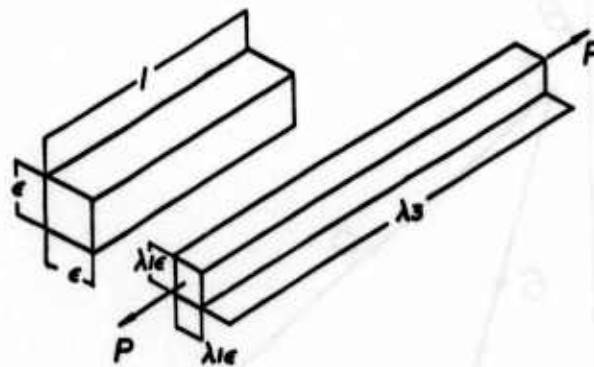
Figure 2. Simple tension bar.
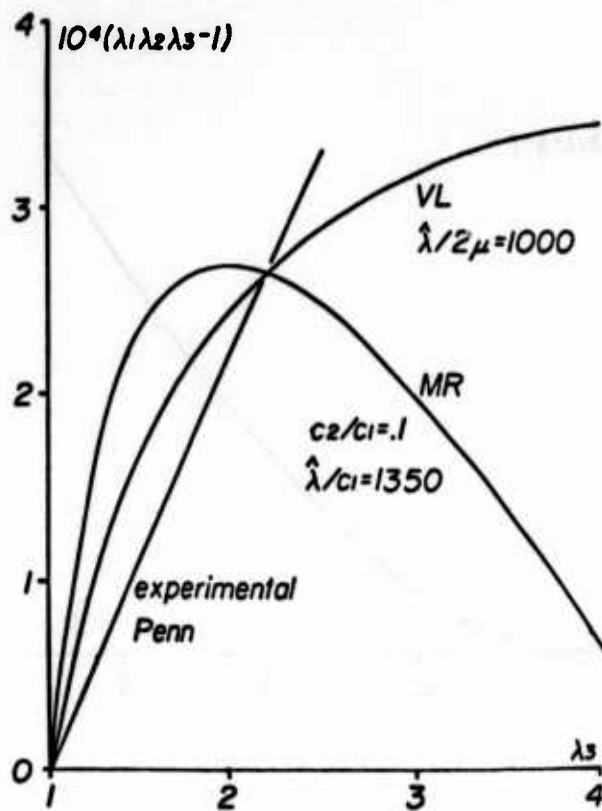
Figure 3. Relative volume change $(\lambda_1\lambda_2\lambda_3-1)$ vs λ_3 for the simple tension bar.
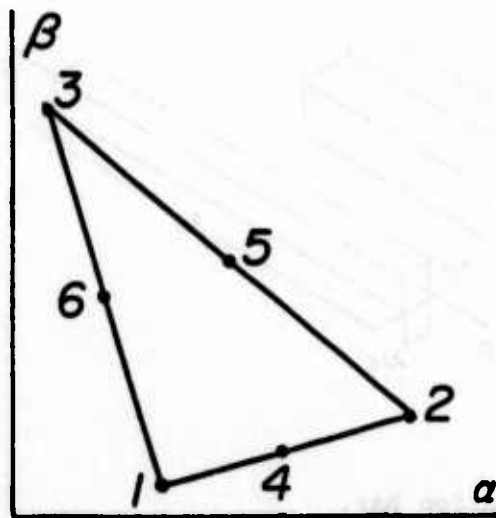
Figure 4. Quadratic six-node triangular element in the (α,β) plane.
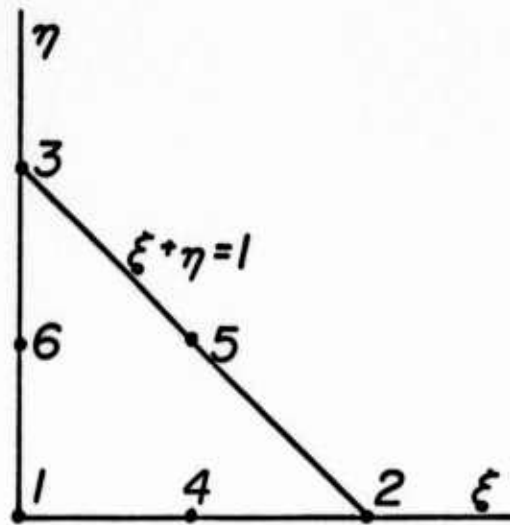
Figure 5. Local coordinate system for interpolations.
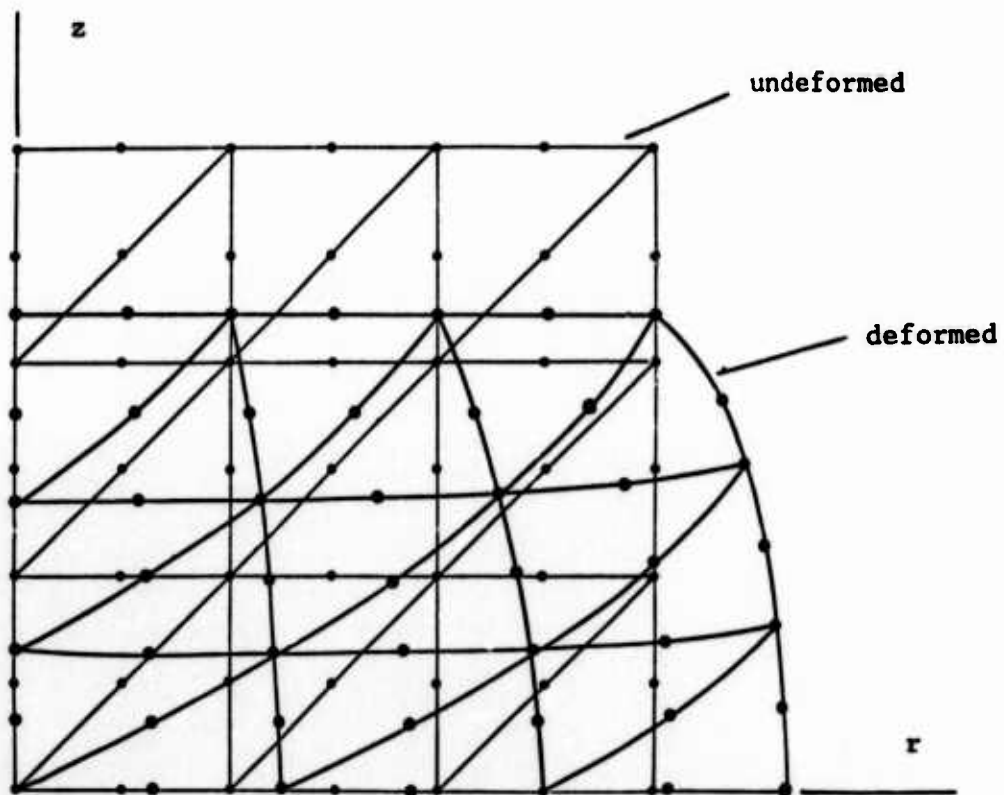
Figure 6. Original and deformed mesh for a compressed cylinder.

491

SOLUTIONS OF THE TRANSONIC FLOW EQUATIONS

BY SPECTRAL METHODS

Patrick Hanley, Cathy Mavriplis,
Massachusetts Institute of Technology, Cambridge, MA 02139

Wesley L. Harris,
University of Connecticut, Storrs, CT 06268

ABSTRACT

Spectral methods, noted for their accuracy and efficiency, are used in transonic areodynamics problems to develop new and efficient tools particularly different from conventional finite difference and finite element techniques. Two transonic flow problems are investigated. First, a pseudospectral method utilizing a direct inversion method has been developed to solve the steady small disturbance flow equations for subcritical flows. Second, a numerical/ analytical method involving a Chebyshev-Fourier pseudospectral approximation has been developed and applied to solve the two-dimensional shockless transonic potential flow in the hodograph plane. Results have been obtained for parabolic arc, NACA 0012 and NLR quasi-elliptical airfoils.

I Introduction

Spectral methods have emerged recently as efficient numerical tools for solving partial differential equations[16]. In an effort to develop new numerical techniques for transonic flow, spectral methods are considered for their accuracy and efficiency. They are an extension of the classical seperation of variables method, capable of approximating smooth solutions with exponential convergence. There are three main types of spectral methods, Galerkin, Tau and collocation or pseudospectral; the latter is used in the following problems for its suitability to nonlinear and variable coefficient equations. The spectral solution is expressed analytically and hence the only error involved is due to the truncation of the infinite series. This is only true of smooth functions. Thus, the present work is limited to shockless flows in which no large discontinuties occur. Note however that shock capturing or shock fitting can be used in conjunction with spectral methods to avoid this limitation[11]. Furthermore, a more general class of solutions may be obtained with relatively little extra effort and the solution at any point in the domain is obtainable with the same degree of accuracy. Finally grid generation is relatively trivial, eliminating a large computational portion of most finite difference and finite element methods.

Thus, spectral methods offer definite advantages over conventional finite difference and finite element methods. Applications of the methods to real problems will determine their true value. Transonic flow is complicated by the appearance of embedded supersonic (or subsonic) regions, the existence of shock waves and the non-linearity of the governing equations which are of mixed (hyperbolic-elliptic) type. In the following, the shock wave problem has already been eliminated by considering shockless flows. The nonlinearity of the equations is easily handled by the pseudospectral method. No special considerations are needed for the embedded regions corresponding to the change in type of the equations when spectral methods are used. Transonic flow problems of two types are investigated. The first is a calculation of transonic flow about airfoils using the transonic small disturbance equation. This equation is derived from the full potential equation by assuming that the disturbances introduced into the flow field by a body are at least an order of magnitude smaller than the undisturbed conditions. This approximation is valid for thin airfoils where the maximum thickness to chord ratio is at least an order of magnitude less than unity at small angles of attack. The transonic small disturbance equation is highly nonlinear and is solved by a pseudospectral method. The second is a calculation of transonic flow about airfoils using the hodograph formulation. This involves transforming the full potential equation to the hodograph plane where the independent variables are the component of velocity. The resulting hodograph equation is a linear variable coefficient equation which can be efficiently solved by a pseudospectral method. In both cases different sets of boundary conditions and other difficulties must be addresses.

In the following, the pseudospectral method is described in general terms in a first part. The transonic small disturbance problem follows with specific considerations for this case. Results for parabolic arc airfoils in transonic flow are presented. The hodograph problem is then described and an interative solution procedure is presented. Results include subsonic flow about a NACA 0012 and transonic flow about a NLR quasi-elliptical airfoil.

both cases different sets of boundary conditions and other difficulties must be addressed.

In the following, the pseudospectral method is described in general terms in a first part. The transonic small disturbance problem follows with specific considerations for this case. Results for parabolic arc airfoils in transonic flow are presented. The hodograph problem is then described and an iterative solution procedure is presented. Results include subsonic flow about a NACA 0012 and transonic flow about a NLR quasi-elliptical airfoil.

2 Pseudospectral Method

Spectral methods are based on representation of the solution to an equation as a truncated series of known smooth functions of the independent variables. For an ordinary or partial differential equation of the form

$$Lu(x, y, ...) = 0 \tag{1}$$

where L denotes an arbitrary differential operator, the solution u is represented by

$$u = \sum_{n=0}^{N} \sum_{m=0}^{M} ...A_{nm}...X_n(x)Y_m(y)... \tag{2}$$

where $X_n(x), Y_m(y)$...are known smooth functions, and $A_{nm}...$ are unknown coefficients. Each group of these functions is usually a set of orthogonal functions chosen to suit the particular problem. The most popular choices of functions are traditionally Fourier, Chebyshev or Legendre functions. However any set of orthogonal functions can be used. In the present description of spectral methods let us assume Chebyshev polynomials will be used.

In addition, in a pseudospectral method, the dependent variables are represented by a set of values at the collocation points $x_i, y_j, ...$ as

$$u_{ij...} = \sum_{n=0}^{N} \sum_{m=0}^{M} ...A_{nm}...X_n(x_i)Y_m(y_j)... \tag{3}$$

Other terms in the equation (1) such as the nonlinear or variable coefficient terms are evaluated at the collocation points. For Chebyshev collocation methods the collocation points are given by

$$x_i = cos\frac{\pi i}{N} \text{ for } i = 0, N \tag{4}$$

which gives uneven grid spacing, points being clustered near edges of the domain.

In what follows, the principle of collocation spectral methods is demonstrated in a one-dimensional situation with Chebyshev polynomials. The extension to two dimensions is straightforward. For a one-dimensional problem,

$$Lu(x) = 0, \tag{5}$$

let,

$$u = \sum_{n=0}^{N} A_n T_n(x) \tag{6}$$

495

where $T_n(x)$ are the Chebyshev polynomials given with their properties in [7]. u is evaluated at the collocation points (4) as follows:

$$u_i = \sum_{n=0}^{N} A_n T_n(x_i). \tag{7}$$

The coefficients A_n are calculated by inverting the series

$$A_n = \frac{1}{c_n} \sum_{k=0}^{N} \frac{1}{c_k} u_k T_n(x_k) \tag{8}$$

where

$$c_o, c_N = 2, c_j = 1 \text{ for all other } j. \tag{9}$$

Once the coefficients A_n are found, derivative terms u_x, u_{xx} etc. of the equation are easily calculated in spectral space due to the properties of the Chebyshev (or Fourier) functions which are given in [7]. As an example let u_x be calculated in Chebyshev spectral space:

$$u_x = \sum_{n=0}^{N} B_n T_n(x) \tag{10}$$

where the coefficients are given by

$$B_n = \frac{1}{c_n} \sum_{p=n+1, p+n \, even}^{N} p A_p. \tag{11}$$

The coefficients of the sums can be evaluated with the use of Fast Fourier Transform (FFT) algorithms in $O(Nlog(N))$ operations for a one-dimensional problem. While the spatial derivative is evaluated spectrally, the temporal derivative is evaluated by finite difference discretization such as explicit forward Euler, Crank-Nicolson...

For two-dimensional problems the approach is similar to the above description. However at this point to efficiently invert the many sums involved there must be special considerations for the two-dimensional case. Indeed the sum

$$u = \sum_{n=0}^{N} \sum_{m=0}^{M} A_{nm} T_n(x) T_m(y) \tag{12}$$

can be inverted using a FFT within $O(N^2 log(N))$ operations. However it is shown in [10] that for $N < 64$ the use of a direct inversion method is more efficient than the FFT.

In the direct inversion method the collocation solution u_{jk} can actually be written in terms of matrices as follows:

$$[u] = [T_x][A][T_y]^T \tag{13}$$

where $[A]$ is the full NxM matrix of coefficients A_{nm} and $[T_x]$ is the full NxN matrix of functions in x, $[T_y]$ MxM in y, such that the matrix element is given by

$$T_{jn} = T_n(x_j) \tag{14}$$

for Chebyshev collocation, x_j given by (4). The matrix $[A]$ can be calculated by inversion of (13),

$$[A] = [T_x]^{-1}[u][T_y]^{T-1} \tag{15}$$

in $O(N^3)$ operations by a pivoting matrix algorithm. This is in fact more efficient than the FFT since the $[T]$ matrices do not change through the iteration procedure and hence only need to be inverted once at the beginning of the computation. Derivatives may be calculated as in the one-dimensional case with new coefficients B_{ij} or using a matrix derivative operator as follows. The x derivative of u may be written

$$[u_x] = [D][u] \text{ where } [D] = [T][P][T]^{-1}. \tag{16}$$

Other derivatives are similarly expressed,

$$[u_y] = [u][D]^T, [u_{xx}] = [D]^2[u][u_{yy}] = [u][D]^{T2}, [u_{xy}] = [D][u][D]^T. \tag{17}$$

3 The Transonic Small Disturbance Equation

3.1 Governing Equations

3.1.1 Unsteady Flow

Transonic flows over thin airfoils can be described by

$$M^2\phi_{tt} + 2M^2\phi_{xt} = (1 - M^2 - M^2(\gamma + 1)\phi_x)\phi_{xx} + \phi_{yy} \tag{18}$$

for most design considerations and flutter analysis. In equation (18), M is the free stream mach number, γ is the ratio of specific heats and ϕ is the small disturbance velocity potential. Equation (18) assumes that the flow is inviscid, isentropic and the perturbed quantities are small compared to those of the mean flow. The pressure coefficient is given by linear theory as

$$C_p = -2(\phi_t + \phi_x). \tag{19}$$

The velocity vector on the boundary of the airfoil is required to satisfy the tangency condition for the small pertubation assumption,

$$\phi_y^{\pm} = \delta\left(f_x^{\pm} + f_t^{\pm}\right) \tag{20}$$

where δ is the maximum thickness to chord ratio and $f^{\pm}$ is the basic shape of the top and bottom of the airfoil respectively. Equation (20) is applied at the $y = 0$ line in accordance with thin airfoil theory.

Across the wake the normal velocity is continuous and the jump in pressure is zero. These conditions are imposed by letting

$$[\phi_y] = 0 \tag{21}$$

$$[\phi]_t + [\phi]_x = 0 \tag{22}$$

where $[\psi]$ denotes the jump in ψ.

At large distances away from the airfoil, the disturbances must vanish. This implies that the perturbed velocity potential and its derivatives must approach zero away from the airfoil. When

such boundary condititons are imposed with equation (18) however, outgoing disturbances are reflected back into the computational domain. These disturbances reduce both the accuracy and rate of convergence of any numerical scheme. The behavior of the perturbations away from the airfoil is accounted for by the fact that equation (18) is an inviscid hyperbolic equation in time. The disturbances modeled by equation (18) will therefore retain a finite amplitude even at infinite distances. To prevent the outgoing waves from being reflected back into a rectangular computational domain several researchers[4] use the following radiating boundary conditions

$$\left(\frac{\partial}{\partial y} \pm \frac{M}{\sqrt{1 - M^2}} \frac{\partial}{\partial t}\right)\phi = 0 \tag{23}$$

for the top and bottom boundaries respectively. The nonreflecting upstream and downstream conditions are satisfied by

$$\left(\frac{\partial}{\partial x} \mp \frac{M}{1 \mp M} \frac{\partial}{\partial y}\right)\phi = 0. \tag{24}$$

3.1.2 Steady Flow

For steady flows equation (18) reduces to

$$(1 - M^2 - M^2(\gamma + 1)\phi_x)\phi_{xx} + \phi_{yy} = 0 \tag{25}$$

and the pressure coefficient is given by

$$C_p = -2\phi_x. \tag{26}$$

The boundary condition on the airfoil is the tangency condition for a rigid shape

$$\phi_y^\pm = \delta f_x^\pm \tag{27}$$

and in the wake $[\phi_x]$ and $[\phi_y]$ are zero.

In the far field, ϕ and it derivatives approach zero. Conditions that were found to be numerically stable for a rectangular domain are

$$\phi_y = 0 \tag{28}$$

at the top and bottom boundaries and

$$\phi_x = 0 \tag{29}$$

upstream and downstream of the airfoil.

3.2 Pseudospectral Approximation

In this section, a numerical method based on pseudospectral discretization in space will be developed to solve the steady small disturbance equation over a parabolic arc airfoil at zero incidence. The equations will be discretized on a multidomain grid system which simplifies the implementaion of the tangency condition and enhances the efficiency of the numerical scheme. An artificial time dependence is imposed on the steady equation so that a forward Euler scheme can be used to iterate the solution to a steady state.

498

3.2.1 Iteration Scheme

The following time dependence is introduced into equation (25)

$$\phi_\tau = (1 - M^2 - M^2(\gamma+1)\phi_x)\phi_{xx} + \phi_{yy} \qquad (30)$$

where τ is not a physical time but a pseudotime for iteration purposes. When steady state is achieved, the τ dependance will vanish and equation (30) will be identical to equation (25).

The time derivative in equation (30) is discretized by using the explicit forward Euler representation. The resulting temporal discretization is

$$\phi^{n+1} - \phi^n = \Delta\tau[(1 - M^2 - M^2(\gamma+1)\phi_x^n)\phi_{xx}^n + \phi_{yy}^n] \qquad (31)$$

3.2.2 Evaluating the Derivatives

The derivatives in equation (31) are evaluated using the pseudospectral techniques developed in the previous section. The potential at time level n is written as

$$\phi(x_i, y_j)^n = \sum_k \sum_l A_{kl} T_k(x_i) T_l(y_j) \qquad (32)$$

where

$$T_k(x_i) = \cos(k\cos^{-1}(\frac{i\pi}{N}))$$

are the Chebyshev polynomials evaluated at the corresponding collocation points. The potential in equation (32) can be written in matrix notation as $[\Phi]$ where the elements ϕ_{ij} denote the value of the potential at points x_i and y_j. Derivatives at time level n are then calculated as above from

$$\begin{aligned}
[\Phi]_x &= [D][\Phi] \\
[\Phi]_{xx} &= [D]^2[\Phi] \\
[\Phi]_y &= [\Phi][D]^T \\
[\Phi]_{yy} &= [\Phi][D]^{T2}
\end{aligned} \qquad (33)$$

where the elements ϕ_{xij}, ϕ_{xxij}, ϕ_{yij} and ϕ_{yyij} are the values of the derivatives of ϕ at the points x_i and y_j.

3.2.3 Computational Domain

Stretching To apply spectral solution to an arbitrary range of the physical variables, it is necessary to introduce computational variables which map the range of the physical variables to the computational range of $[-1, 1]$. The physical variables x and y are mapped linearly onto the computation variables ξ and η by

$$\begin{aligned}
x &= \xi\frac{x_a - x_b}{2} + \frac{x_a + x_b}{2} \\
y &= \eta\frac{y_a - y_b}{2} + \frac{y_a + y_b}{2}
\end{aligned} \qquad (34)$$

where

$$\begin{aligned}
x_b &< x < x_a \\
y_b &< y < y_a
\end{aligned}$$

Multidomain Solution The development for multidomain solution is limited to rectangular subdomains aligned with the physical coordinate axis. Continuity of the dependent variable and its normal derivative will be imposed at the interface of subdomains. For the rectangular subdomains, the normal derivative requirement amounts to letting $\frac{\partial \phi}{\partial x}$ be continuous at x faces and $\frac{\partial \phi}{\partial y}$ be continuous at y faces.

Figure (1) shows the three domain grid which is used to solve equation (25) over a symmetric airfoil at zero angle of attack. The dependent variables in domains I, II and III are denoted as ϕ^I, ϕ^{II} and ϕ^{III} respectively. The entire computational domain is bounded from $x = x_0$ to $x = x_3$ where subdomain I lies between $x = x_0$ and $x = x_1$, subdomain II between $x = x_1$ and $x = x_2$ and subdomain III between $x = x_2$ $x = x_3$.

Continuity across the subdomains is implemented by letting

$$\phi_x^I(x_1, y_j) = \phi_{xNj}^I = \alpha \sum_{k=0}^{N} D_{Nk}\phi_{kj}^I \tag{35}$$

$$\phi_x^{II}(x_1, y_j) = \phi_{x0j}^{II} = \beta \sum_{k=0}^{N} D_{0k}\phi_{kj}^{II} \tag{36}$$

$$\phi_x^{II}(x_2, y_j) = \phi_{xNj}^{II} = \beta \sum_{k=0}^{N} D_{Nk}\phi_{kj}^{II} \tag{37}$$

$$\phi_x^{III}(x_2, y_j) = \phi_{x0j}^{III} = \gamma \sum_{k=0}^{N} D_{0k}\phi_{kj}^{III} \tag{38}$$

where

$$\alpha = \frac{2}{x_1 - x_0}$$
$$\beta = \frac{2}{x_2 - x_1}$$
$$\gamma = \frac{2}{x_3 - x_2}.$$

Equating the above conditions according to the interface conditions and letting

$$\begin{aligned} \phi_{Nj}^I &= \phi_{0j}^{II} = v_j \\ \phi_{Nj}^{II} &= \phi_{0j}^{III} = w_j, \end{aligned} \tag{39}$$

the following equations can be used to solve for ϕ at the subdomain interface in terms of v_j and w_j:

$$\beta D_{N0}v_j + [\beta D_{NN} - \alpha D_{00}]w_j = \gamma \sum_{k=1}^{N} D_{0k}\phi_{kj}^{III} - \beta \sum_{k=1}^{N-1} D_{Nk}\phi_{kj}^{II} \tag{40}$$

and

$$[\alpha D_{NN} - \beta D_{00}]v_j - \beta D_{0n}w_j = \beta \sum_{k=1}^{N-1} D_{0k}\phi_{kj}^{II} - \alpha \sum_{k=0}^{N-1} D_{nk}\phi_{kj}^{I}. \tag{41}$$

500

3.3 Solution Algorithm

The following algorithm is used to iterate equation (31) on the multidomain grid until a steady state solution is obtained:

1. Compute [D] and [D]^2 matrix

2. Compute derivatives of [Φ] using equation (33) at the previous time level

3. Impose boundary conditions at farfield

4. Compute value of ϕ at the interface of the subdomains from equations (40) and (41)

5. Evaluate ϕ at level $n+1$ from equation (31)

6. Repeat from step (2) until convergence criterion is met

3.4 Results

The pseudospectral method is validated through solution to clasical partial differential equations[9]. Resultys for the parabolic arc airfoil are presented here. Figure (2) shows the computed pressure coefficient over a parabolic arc airfoil at zero angle of attack and a Mach number of .825. Results are shown for parabolic arcs of thickness ratio ranging from 1 to 6% in increments of 1%. The calculations were done on a 25X9 grid as shown in figure (2). The computations were halted when the coefficient at the center of the airfoil did not change to 3 decimal places or about 500 iterations. For these thin airfoil cases at Mach number of .825 the flow is subcritical and hence no shocks are observed. These results are in excellent agreement with those of Sivaneri and Harris [18] using a much coarser grid. The CPU time required for this calculation is 2 minutes on a VAX 750.

The results show that the pseudospectral method is a viable efficient technique to investigate transonic flows and work is in progress to further develop the method for unsteady flows.

4 The Hodograph Equation

4.1 Motivation

The main objective in this case is to create an efficient method of predicting transonic flow over two-dimensional bodies by combining the hodograph formulation with spectral methods. As a by-product we also hope to develop an inverse design code for shockless airfoils in transonic flow. The assumptions are two-dimensional, steady, inviscid and isentropic flow.

The advantage of using the hodograph formulation is that the governing equations become linear thereby enabling solution by linear superposition. However this advantage is only gained at the expense of complicating the boundary conditions as well as our sense of where the body is in the hodograph domain. Many different hodograph methods have been used in the past (see [13]),the most successful being that of Garabedian, Korn and Bauer [2] which produced the shockless Korn airfoil. The ability to design shockless airfoils is important since a large amount of the drag in transonic flow is attributed to wave drag caused by shock waves. Elimination or weakening of shocks greatly reduces drag which is instrumental in improving the efficiency of the transonic cruise range in which most commercial aircraft operate. The present hodograph formulation follows that

of Nieuwland et al [15] but uses numerical solution instead of analytic continuation and integral transform methods.

Results in the hodograph plane are translated back to the physical plane through an analytical transformation to obtain physical coordinates, mainly of the body. Hence, the most useful outcome of this work should be an efficient inverse design code for shockless airfoils in transonic flow. For design methods, Lighthill's constraints must be incorporated to ensure closure and circulation for the designed airfoil [12]. Volpe and Melnik [19] have derived equivalent constraints for transonic flow: a loose implementation of these consraints is included in this method, though it is not truly an inverse design method in its present form. An iteration procedure is needed to arrive at a realistic airfoil starting from a circular cylinder. The method is validated by two test cases, a NACA 0012 airfoil in subsonic flow and a NLR QE-.11-.75-1.375 airfoil in transonic flow.

4.2 Formulation

Governing Equations The governing equation for steady two-dimensional potential flow in the transonic range may be transformed by a series of independent variable transformations [13] to the hodograph plane, where the independent variables are u and v or q and θ, such that $\underline{v} = u\underline{i} + v\underline{j} = qe^{i\theta}$. The resulting equation in the hodograph plane may be stated in terms of ϕ, the velocity potential or its complex conjugate ψ, the streamfunction. Here we choose to write the equation in terms of the streamfunction ψ and the nondimensional independent variables r and θ defined by $r = (q/q_{max})^2$ and $\theta = arg(\underline{v})$, as follows:

$$\psi_{rr} + \frac{1 + \frac{2-\gamma}{\gamma-1}r}{r(1-r)}\psi_r + \frac{1 - \frac{\gamma+1}{\gamma-1}r}{4r^2(1-r)}\psi_{\theta\theta} = 0 \tag{42}$$

where $r \in [0;1[$ and $\theta \in [0;2\pi]$. q_{max} is given by $(2/(\gamma-1))^{1/2}c_o$. This hodograph equation is a linear variable coefficient partial differential equation of mixed (elliptic - hyperbolic) type. The transformation back to the physical plane is given in integral form by

$$x + i\epsilon y = \int \frac{e^{i\theta}}{r^{1/2}(1-r)^{1/(1-\gamma)}} * \{-[\frac{(1 - \frac{\gamma+1}{\gamma-1}r)}{2r(1-r)}\frac{\partial\psi}{\partial\theta} - i\frac{\partial\psi}{\partial r}]dr + [2r\frac{\partial\psi}{\partial r} + i\frac{\partial\psi}{\partial\theta}]d\theta\}. \tag{43}$$

where ϵ is the thickness ratio and all variables have been non-dimensionalized appropriately. Given a streamfunction distribution $\psi(r,\theta)$ one can calculate the flow in the physical plane analytically by equation (43).

Singularities Different types of singularities appear in this formulation. The variable coefficients of equation (42) are infinite at the limits of the domain, i.e. for $r = 0$ and 1. These points are regular singular points and hence the solution to (42) is either analytic or has a pole or an algebraic or logarithmic branch point. Branch type singularities are inherent in the hodograph formulation due to the multi-sheeted nature of the flow representation. Typical airfoil flow hodograph solutions must be represented by two or more sheets as illustrated in figure (3). The upper and lower sheets are separated by a branch cut emanating from a branch point (r^*, θ^*). Furthermore the streamfunction ψ has a free-stream singularity of the doublet type at infinity and hence at $(r_\infty, 0)$ in the hodograph plane. Among other difficulties, the entire physical flow at infinity is mapped to a single point in the hodograph plane. The Jacobian of the transformation between the hodograph

502

and physical planes exhibits singularities for $J = 0$ and $J = \infty$ which must be avoided to ensure a one to one correspondence between the physical and hodograph planes. In the present case it is found that these singularities do not appear and hence only the free-stream singularity at infinity and the branch point separating the two sheets of the hodograph surface must be represented in the final solution.

Boundary Conditions The transformation (43) appears to give a direct correspondence between the physical and hodograph planes. However, as illustrated by figure (3), the unknown functions and particularly the body itself are not easily identifiable in the hodograph plane. In consequence, typical boundary conditions for external flow problems such as the farfield boundary condition, the tangency conditon and the Kutta condition at the trailing edge of the airfoil cannot be stated in terms of ψ, q and θ in the hodograph plane. Instead, the only conditions that can be imposed are that the stagnation points $r = 0$ and the point of maximum velocity $r = r_1$ are on the body profile $\psi = 0$ or $\psi(r = 0, \theta) = 0$ and $\psi(r = r_1, \theta) = 0$ and, though not really a boundary condition, ensure realistic solutions by transforming the solutions in the hodograph plane back to the physical plane via equation (43) and plotting the physical body coordinates to verify that the results are physically meaningful. Additional boundary conditions need to be formulated for numerical reasons as will be seen in the following section.

4.3 Method of Solution

Solution of the hodograph equation forms the basis of the proposed method but does not necessarily yield physically meaningful results directly. The method must therefore incorporate solution of the hodograph equation into an encompassing solution procedure, the different components of which are described individually in this section.

4.3.1 Solution of the Hodograph Equation

Though equation (42) posesses particular solutions involving Hypergeometric functions, these are not used in the present method, since the high frequency in θ required to determine the zeros of the Chaplygin particular solutions makes the solution of the present problem (searching for $\psi = 0$) in terms of these particular solutions impractical. However, linear superposition of these particular solutions has been used by Lighthill and Cherry[5] to yield meaningful solutions. Instead, spectral methods are employed in the solution of equation (42) since these converge exponentially for smooth solutions and are relatively fast methods. Collocation spectral methods are extremely well-suited to the solution of variable coefficient equations such as the hodograph equation. However, as mentioned previously, the variable coefficients of (42) are singular at the boundaries of the domain. Since the only singular boundary of interest is $r = 0$, solutions in two domains, namely

1. about $r = 0$ or for $r \in [0; r_a]$ (where r_a is close to 0) and

2. for $r \in [r_a; r_1]$ (where $r_1 < 1$)

must be derived.

Solution About the Singular Point For domain 1) the point $r = 0$ is a regular singular point and hence an exact solution for the r variation of ψ in the neighbourhood of $r = 0$ may be written in terms of a Frobenius series of the form

$$Y(r) = (r - r_o)^\alpha \sum_{n=0}^{\infty} a_n (r - r_o)^n \tag{44}$$

where all coefficients are determined by substituting into equation (42) except a_o which remains arbitrary. Convergence of this series is obtained for $\alpha = +k/2$ only, k being the Fourier mode of the solution in θ. The radius of convergence is $r < 1$. The solution is smooth and hence may be easily matched to the pseudo-spectral solution at r_a. This solution also displays the required boundary condition $\psi(r = 0; \theta) = 0$, expressing the fact that the stagnation points should be on the body. a_o and r_a are part of the set of input parameters at each iteration step.

Pseudospectral Solution Solution of (42) in domain 2), where the variable coefficients are analytic, is obtained by application of a pseudospectral method. In this method the solution $\psi(r, \theta)$ is expressed as a truncated sum of analytic functions, in this case Chebyshev polynomials in r and Fourier functions in θ (in contrast to the previous problem) as follows:

$$\psi = \sum_{n=0}^{N} \sum_{m=0}^{M} A_{nm} T_n(r) e^{im\theta}, \tag{45}$$

where $T_n(r)$ are the Chebyshev polynomials given in [7]. ψ is evaluated at the collocation points as follows:

$$\psi = \sum_{n=0}^{N} \sum_{m=0}^{M} A_{nm} T_n(r_i) e^{im\theta_j}, \tag{46}$$

where for Chebyshev collocation methods the collocation points are given by $r_i = \cos \frac{\pi i}{N}$ for $i = 0, N$ and for Fourier collocation, $\theta_j = \frac{2\pi j}{M}$ for $j = 0, M$. These collocation points define the grid. Since the Chebyshev polynomials are defined from $[-1; 1]$, a mapping from the arbitrary domain $[r_a; r_b]$ to $[-1; 1]$ is in order. For the Chebyshev points in the r direction the points are crowded at the edges of the domain, which turns out in this case to be ideal: high resolution near stagnation points and the point of maximum speed. As in the transonic small disturbance problem direct inversion is used. All equations are hence written in matrix form,

$$[\psi] = [T][A][F]^T \tag{47}$$

where $T_{ij} = T_i(r_j)$, $F_{mk}^T = e^{im\theta_k}$. Derivatives are calculated as described in section 2.

The hodograph equation though steady, is restated in an unsteady formulation

$$\psi_t = L\psi \tag{48}$$

where L is the hodograph operator of equation (42). The steady state solution of equation (48) is sought so that $\psi_t = 0$ and hence (48) reduces to the hodograph equation (42). Time derivatives are treated by finite difference. In this case a semi-implicit scheme given by Gottlieb and Orszag [7] stated as

$$\frac{\psi^{n+1} - \psi^n}{\Delta t} = \frac{L_{max}}{2} \psi^{n+1} + (L - \frac{L_{max}}{2}) \psi^n \tag{49}$$

504

is used. This scheme is unconditionally stable. In this scheme, the matrix to be inverted is $[I] - \Delta t[L_{mas}]/2$ where L_{mas} is the operator matrix made up of $[D], [D]^2, \ldots$ with the maximum values of the variable coefficients. This matrix is always well-behaved as opposed to the fully explicit or implicit cases where the matrix is sometimes ill-conditioned and where time steps are limited to very small values. It appears however that Δt must be somewhat restricted in order that $[I]$ not be insignificant in front of $L_{mas}\Delta t/2$ since L_{mas} is large (due to the large values of the variable coefficients).

Conditions and Constraints Equation (49) is only solved on the non-singular portion of the domain $[r_a; r_b]$. Boundary conditions for the pseudo-spectral solution are given by

$$\psi(r_a; \theta; t) = K_a(\theta) \text{ and } \psi(r_b; \theta; t) = K_b(\theta) \tag{50}$$

in r and a periodic boundary condition in θ. The initial condition for the first overall iteration is given by an exact solution to the hodograph equation involving hypergeometric functions, developed by Cherry [5] for transonic flow about a circular cylinder. This solution actually yields a slightly deformed circular cylinder when the above expressions are inserted into the transformation (43) as illustrated in figure (4). The presented results were calculated by portions of this method which are validated by the excellent agreement with Cherry's results. Deformation of the circular cylinder into an airfoil is ensured through the boundary conditions both at the pseudo-spectral level and at the encompassing iterative level, as well as through the Frobenius solution about the singular point. At this point it is useful to note that it was decided in the final form of the method that the hodograph equation be solved by the pseudo-spectral method on two domains (subdomains of domain 2)):

- 2a) $r \in [r_a; r_\infty]$

- 2b) $r \in [r_\infty; r_b]$

since the initial solution given by Cherry was so derived. (Recall that r_∞ is the branch point). The solutions for the individual domains are matched numerically at r_∞. Using a solution based on that of Cherry's ensures that the free-stream singularity and branch point behaviour are correctly included.

Matching of the singular solution and the pseudospectral solution is straightforward. Either ψ or derivatives of ψ can be matched at r_a and r_∞. This translates in either case to matching the A_{ij} or B_{ij} coefficients. At the other end of the domain, r_b, an arbitrary but realistic boundary condition must be imposed. We choose to use the boundary condition at the point of maximum velocity which occurs on the body. Hence the maximum speed r_b on the body must be known *a priori* and the boundary condition is $\psi(r_b; \theta) = K_b(\theta)$ with $\psi(r_b; \theta_b) = 0$. A judicious choice of the function $K_b(\theta)$ is crucial to the quality of the results for a given case.

4.3.2 Transformation to the Physical Plane

Physical coordinates for constant streamfunction values are obtained by the transformation (43), the most interesting coordinates being those corresponding to $\psi = 0$ or the body contour obtained by secant interpolation. Upon substituting the Chebyshev-Fourier form of ψ, (45), into

505

(43), all integrals are readily calculated by hand. The complete expression for the coordinates x and y is given by

$$
\begin{aligned}
x = & \sum_n \sum_m \frac{mA_{nm}sin(\theta)cos(m\theta)}{2} \int \frac{1-\frac{2+1}{\gamma-1}r}{r^{5/2}(1-r)^{\gamma/(\gamma-1)}} T_n(r) dr \\
& - \sum_n \sum_m B_{nm}sin(\theta)cos(m\theta) \int \frac{T_n(r)}{r^{1/2}(1-r)^{1/(\gamma-1)}} dr \\
& + \frac{2r^{1/2}}{(1-r)^{1/(\gamma-1)}} \sum_n \sum_m B_{nm}T_n(r)(\frac{sin((m+1)\theta)}{2(m+1)} + \frac{sin((m-1)\theta)}{2(m-1)}) \\
& + \frac{1}{r^{1/2}(1-r)^{1/(\gamma-1)}} \sum_n \sum_m mA_{nm}T_n(r)(\frac{sin((m+1)\theta)}{2(m+1)} + \frac{sin((m-1)\theta)}{2(m-1)}) \\
\epsilon y = & - \sum_n \sum_m \frac{mA_{nm}cos(\theta)cos(m\theta)}{2} \int \frac{1-\frac{2+1}{\gamma-1}r}{r^{5/2}(1-r)^{\gamma/(\gamma-1)}} T_n(r) dr \\
& + \sum_n \sum_m B_{nm}cos(\theta)cos(m\theta) \int \frac{T_n(r)}{r^{1/2}D(1-r)^{1/(\gamma-1)}} dr \\
& - \frac{2r^{1/2}}{(1-r)^{1/(\gamma-1)}} \sum_n \sum_m B_{nm}T_n(r)(\frac{cos((1-m)\theta)}{2(1-m)} + \frac{cos((m+1)\theta)}{2(m+1)}) \\
& - \frac{1}{r^{1/2}(1-r)^{1/(\gamma-1)}} \sum_n \sum_m mA_{nm}T_n(r)(\frac{cos((1-m)\theta)}{2(1-m)} + \frac{cos((m+1)\theta)}{2(m+1)}).
\end{aligned}
\tag{51}
$$

The integrals in r are all of the same form once the Chebyshev functions are written out in polynomial form. They can be calculated by integration by parts and written in terms of a recursion relation (see [13]). Hence the transformation back to the physical plane is done exactly with the above analytic expressions.

4.3.3 Iteration Scheme

Though the hodograph equation is solved accurately by the pseudospectral method, the results it yields are not necessarily physically meaningful. This is due to the fact that these results have been calculated with insufficient boundary conditions. Mathematically these conditions are sufficient to obtain a solution; physically they are not. Particularly the functions $K_a(\theta)$ and $K_b(\theta)$ are not known and must be guessed for each case. Hence existence, uniqueness and physical meaning are not guaranteed. Furthermore, since the physical coordinates for an airfoil are derived from those for a circular cylinder an iteration procedure is preferred. The iterations also serve the purpose of being able to monitor the solution as it evolves and guide certain parts of it by altering the boundary values if needed, since, as shown before, the prescribed boundary conditions are not sufficient.

Hence, an iteration scheme is set up to obtain from Cherry's solution to flow around a circular cylinder that around an airfoil. The iteration operates on the boundary conditions, altering at each step either the value of ψ at a boundary or the value of the boundary e.g. r_b. (Recall that r_b is the maximum velocity on the body which must be changing with the shape of the body or, as it is set up in this case, the shape of the body changes because we are trying to keep the same maximum speed.) Most importantly the Frobenius solution in domain 1) is altered in such a way that this solution resembles the desired result only in the region $r \in [0; r_a]$. However, it is not evident how this should be done.

The entire calculation procedure is summarized in the flowchart given in figure (5). The pressure coefficient C_p distribution over the body is given by

$$
C_p := \frac{\gamma - 1}{\gamma} \frac{1 - r_\infty}{r_\infty} \{ (\frac{1-r}{1-r_\infty})^{\gamma/(\gamma-1)} - 1 \}.
\tag{52}
$$

4.4 Results

The validity of the proposed method is demonstrated and discussed through results of test cases for realistic airfoil shapes. Typically one complete run to design a shockless symmetric airfoil in transonic flow requires approximately 15 minutes of CPU time on a VAX 750 where steady state is reached in 1000 steps at each iteration level. Approximately three to five encompassing iterations are needed to obtain a final solution with the correct boundary conditions, yielding an approximate total of 5000 iterations for two domains.

4.4.1 NACA 0012 Airfoil

The first test case involves trying to reproduce a NACA 0012 airfoil. Since the flow in this method must be shockless and a NACA 0012 develops a shock in transonic flow, results for a NACA 0012 can only be obtained in a purely subsonic flow. The input parameters for the presented test case were taken from results of a cell-centered finite volume method using the Euler equations by Allmaras[1]. The Mach number for this test case was $M_\infty = .60$, angle of attack $\alpha = 0$. The proposed method uses 20 collocation points in each domain, hence 40 in r and 20 in θ. This case is referred to as a $40X20$ grid. Allmaras' results were run with a $96X16$ "C" type grid. Typically ψ is interpolated to values of $\psi = 10^{-4}$. Results for this test case are given in figures (6,7). Only the upper half of the body is plotted here as the body is symmetrical.

It is found for this run that the maximum thickness is 12.3% as compared to 12% ideally. The error in the maximum thickness is therefore of 2.5% with respect to the ideal which is acceptable considering the crude nature of setting the coefficient of the Frobenius series. Upon comparing the NACA 0012 coordinates given by AGARD[8] as shown in figure (6), it is shown that agreement is generally good. A closer look at the body coordinates is given in figure (7). In this view it is seen that the actual profile is quite jagged. This is due to the fact that ψ is only being interpolated to 10^{-4} and the final values are either positive or negative thereby oscillating about the correct $\psi = 0$ profile. A smoothing function or further interpolation may be applied to obtain a smooth profile and even better agreement with the test case. Comparison of C_p with that of Allmaras[1] (fig.(6)) brings forth one of the advantages of using the Chebyshev pseudospectral method: at the trailing edge where $r \in [0; r_a]$ the points are crowded due to the Chebyshev collocation point distribution, thereby giving very good resolution. Though Allmaras' results are already on a much finer grid, the trailing edge drop in C_p back down to its level at the leading edge is not captured. Note also that there is no smoothing in the present method as opposed to finite difference methods. Furthermore the spectral method yields coefficients and hence an analytical expression for the solution which enables equal accuracy in the approximation to any flow variables at any point in the flow, thereby granting accurate results without the need for very fine resolution. Errors in C_p are considerably larger than those for the coordinates since the coordinates do not agree (and thus the pressure distribution is not expected to coincide).

4.4.2 NLR QE 0.11 - 0.75 - 1.375 Airfoil

The NLR Quasi-Elliptical 0.11 - 0.75 - 1.375 airfoil was selected for testing in the transonic range since it is a symmetrical shock-free airfoil designed by the hodograph method of Nieuwland[15] for which experimental data from AGARD[8] exist including θ values, which are important in hodograph methods, particularly to verify the input parameters.

The present airfoil was designed for the conditions $M_\infty = .78612$, $\epsilon = .1172$, $M_{max\ on\ body} = 1.306$, $\alpha = 0$. These parameters were used in the test case of the proposed method. The design coordinates and C_p distribution are given in AGARD[8]. Experiments from AGARD[8] give results for $M_\infty = .789$. The results for this test case are shown in figure (8) in comparison with the design conditions and the experimental results. Note that the experimental results are at a different freestream Mach number and hence close agreement is not expected. Results are generally poor for this test case and the body coordinates are scaled by 4 to be shown in detail. Only eight points from the hodograph solution are meaningful. The physical coordinates are again oscillating about a mean due to the moderately accurate interpolation values of ψ, though three points actually coincide with the body line. The C_p values are overestimated, but follow the correct trend. It appears that for more complicated test cases the method is not robust and therefore needs improvement.

4.4.3 Discussion

The CPU time required for one complete run (15 minutes on a VAX 750) is remarkably reasonable when compared with other methods. An inverse design method using a finite volume method on the Euler equations for airfoils in viscous transonic flow developed by Drela[6] requires 10 minutes per airfoil. The method of Boerstoel and Huising[3] requires 25 minutes on a CDC 6600 for 130 coordinate points. The method of Garabedian, Korn and Bauer[2] is much more efficient with an approximate time of 2.5 minutes on a CDC 6600 computer. Though these figures are estimates for average runs, perhaps with finer grids and on different sized computers, a comparison shows that the present method is at least competitive if not slightly adavantageous with/over other inverse design methods as far as time requirements are concerned. Additional complications in the shapes to be designed, such as concerns with closure, angle of attack, etc. are expected to alter boundary conditions and input parameters but not increase CPU time significantly.

More pertinent to the physical problem is the accuracy of the method which does not compare very well with other inverse design methods. This, however, is expected since the method is not really an inverse design method yet. The main problems with the method in its present form are the prescription of the Frobenius coefficient, the seeming lack of determining or boundary conditions; e.g. $K_a(\theta)$ and $K_b(\theta)$ in equation (50) due again to the fact that we can't prescribe/know distributions over the body/solution domain. However this sparseness of set conditions is somewhat necessary, first to retain degrees of freedom in a design method and second to satisfy the uniqueness theorem of Morawetz[14] for shockless flows. In the present case this condition is loosely satisfied but mathematically there is still no assurance that the solution will be meaningful or unique in all cases. Furthermore, the iteration scheme is rather simplistic. A more involved iteration scheme which would enforce more stringent boundary conditions at each step should be considered. An iteration scheme such as the Method of Parametric Differentiation [17]would reveal itself most useful in the more general cases for lifting airfoils with camber, angle of attack, etc. Note that the hodograph does not extend to three dimensional flow and hence applications are restricted to airfoil design.

5 Conclusion

The foregoing work demonstrates how efficiently spectral methods can be used in solving transonic flow equations. Pseudospectral methods are used in the two given cases for their suitability

to nonlinear and variable coefficient equations. The first application is solution of the transonic small disturbance equation for thin airfoils. A Chebyshev pseudospectral method is used to solve the nonlinear equation using explicit forward Euler time iteration to steady state. A multidomain grid is used to solve for flow about an airfoil, thereby requiring matching across domain interfaces. Excellent steady results for parabolic arc airfoils in a freestream flow of Mach number .825 demonstrate the efficiency and validity of the pseudospectral method. Future results with the transonic small disturbance equation will include unsteady effects for airfoils pitching and plunging in transonic flow.

The second application involves solution of transonic flow through the hodograph formulation. A Chebyshev-Fourier pseudospectral method is used to solve the linear variable coefficient hodograph equation. In this case the overall formulation is more complicated due to the singularities and boundary conditions in the hodograph plane. Solution for the flow and the body coordinates are obtained at once, thereby requiring more detailed information than actually availbale. For this reason, the presented results for a NACA 0012 in subsonic flow and a NLR quasi-elliptical airfoil in transonic flow are not very accurate but do offer certain advantages. In both cases one must note the efficiency and simplicity of the pseudospectral solution method. Very few grid points on the airfoil are needed for accuracy comparable to (or better than in some areas) that of finite difference and finite element methods.

The presented work shows that spectral methods can be used to predict compressible (and incompressible) flow over realistic airfoil configurations. In both cases studied, the spectral method is relatively simple and fast compared to existing technology: finite difference, finite element methods and various hodograph techniques. Further work in this field should investigate and overcome other typical difficulties encountered in transonic flows, yielding an efficient and novel alternative to conventional computational techniques for transonic flows.

References

[1] S.R. Allmaras. *Embedded Mesh Solutions of the 2-D Euler Equations Using a Cell-Centered Finite Volume Scheme.* Technical Report CFDL-TR-85-4, MIT, August 1985.

[2] F. Bauer and Korn D. Garabedian, P.R. *Supercritical Wing Sections II. Lecture Notes in Economics and Mathematical Systems*, Springer, 1975.

[3] J.W. Boerstoel and G.H. Huizing. Transonic shock-free aerofoil design by an analytic hodograph method. *AIAA Paper*, 74-539, 1974.

[4] D. R. Carlson and W.L. Harris. *An Unsteady Transonic Flow and Parametric Differentiation with an Alternating-Direction Implicit Numerical Scheme.* Technical Report FDRL 83-3, MIT, November 1983.

[5] T.M. Cherry. Flow of a compressible fluid about a cylinder. *Proceedings of the Royal Society*, A 192,:45–79, 1947.

[6] M. Drela. Two-dimensional transonic aerodynamic design and analysis using the euler equations. *M.I.T. Ph. D. Thesis*, 1985.

[7] D. Gottlieb and S.A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications. CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, 1977.

[8] AGARD Advisory Group. Experimental data base for computer program assessment. *Report of the Fluid Dynamics Panel Working Group 04*, AR-138,, 1979.

[9] P. Hanley and W.L. Harris. *Solutions of Fluid Flow Equations by Spectral Methods*. Technical Report FDRL 85-3, MIT, June 1985.

[10] R. Hirsch. High order approximations in fluid mechanics: compact to spectral. In *VKI Lecture Series*, Unpublished, 1983.

[11] M. Y. Hussaini and et al. Spectral methods for the euler equations: part ii - chebyshev methods and shock fitting. *AIAA Journal*, 23,(2):234, 1985.

[12] M.J. Lighthill. *A New Method of Two-Dimensional Aerodynamic Design*. Technical Report Reports and Memoranda No. 2112, Aeronautical Research Council, London, England, 1945.

[13] C. Mavriplis. A spectral hodograph method for two- dimensional transonic aerodynamics. *M.I.T. M. S. Thesis*, 1985.

[14] C.S. Morawetz. Non-existence of transonic flow past a profile. *Communications on Pure and Applied Mathematics*, 17,:357–367, 1964.

[15] G.Y. Nieuwland. *Transonic Potential Flow Around a Family of Quasi-Elliptical Sections*. Technical Report TR T 172, NLR, 1967.

[16] S.A. Orszag and M. Israeli. Numerical simulation of viscous incompressible flows. In *Annual Review of Fluid Mechanics Volume 6*, pages 281–317, Annual Reviews Inc., 1974.

[17] P.E. Rubbert and M.T. Landahl. Solution of nonlinear problems through parametric differentiation. *Physics of Fluids*, 10(4):831–835, 1967.

[18] N.T. Sivaneri and W.L. Harris. Numerical solutions of transonic flows by parametric differentiation and integral equations techniques. *AIAA Journal*, 18(12), 1980.

[19] G. Volpe and R.E. Melnik. The role of constraints in the inverse design problem for transonic airfoils. *AIAA Paper*, 81-1233, 1981.
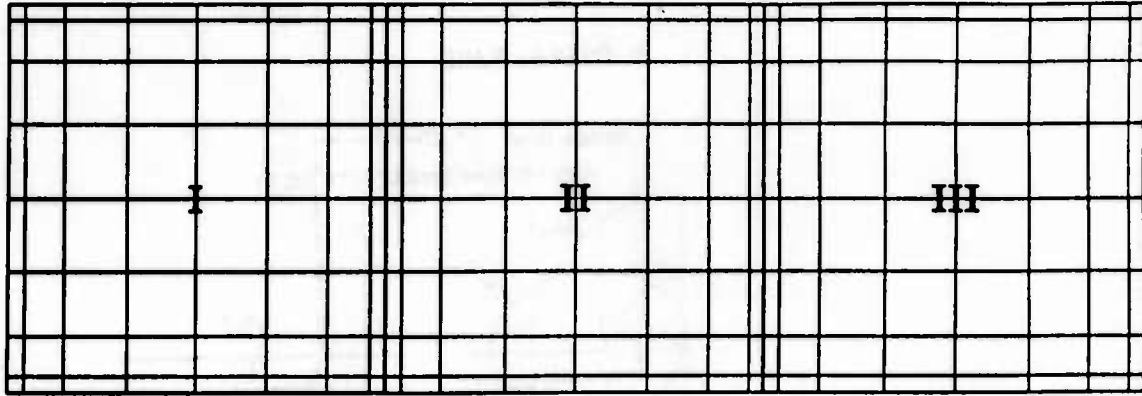
Figure 1: Pseudospectral Multidomain Grid
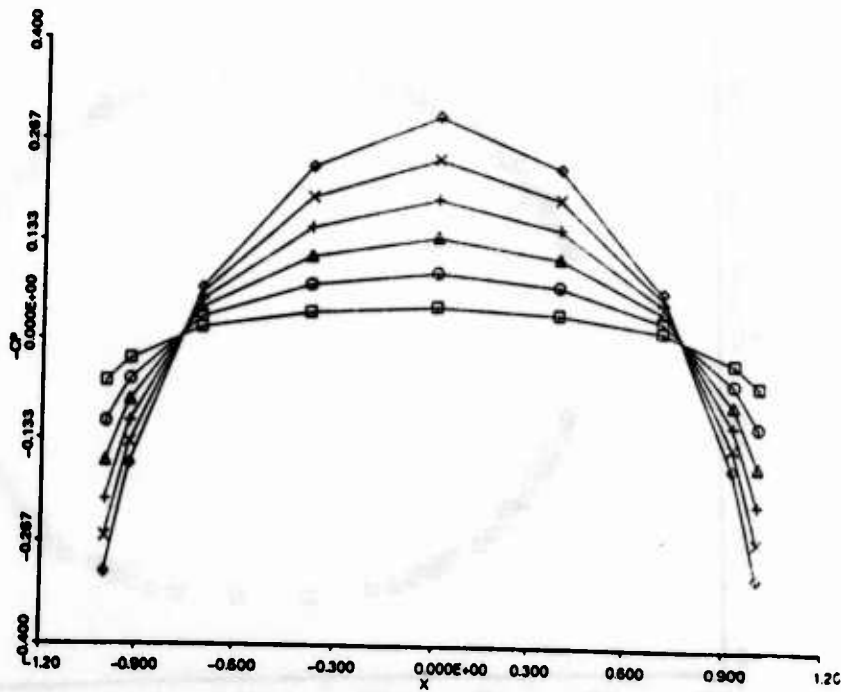
Figure 2: Computed Parabolic Arc Airfoil Pressure Coefficient Distribution

A) PHYSICAL PLANE

B) HODOGRAPH PLANE - UPPER AND LOWER SHEETS
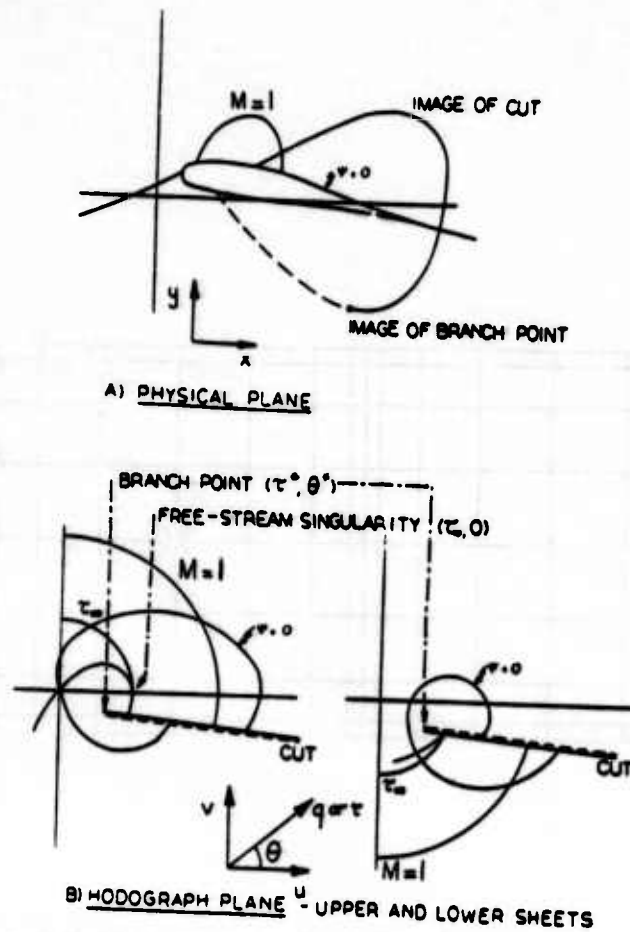
Figure 3: Transonic Flow Over an Airfoil in the Physical and Hodograph Planes[3]
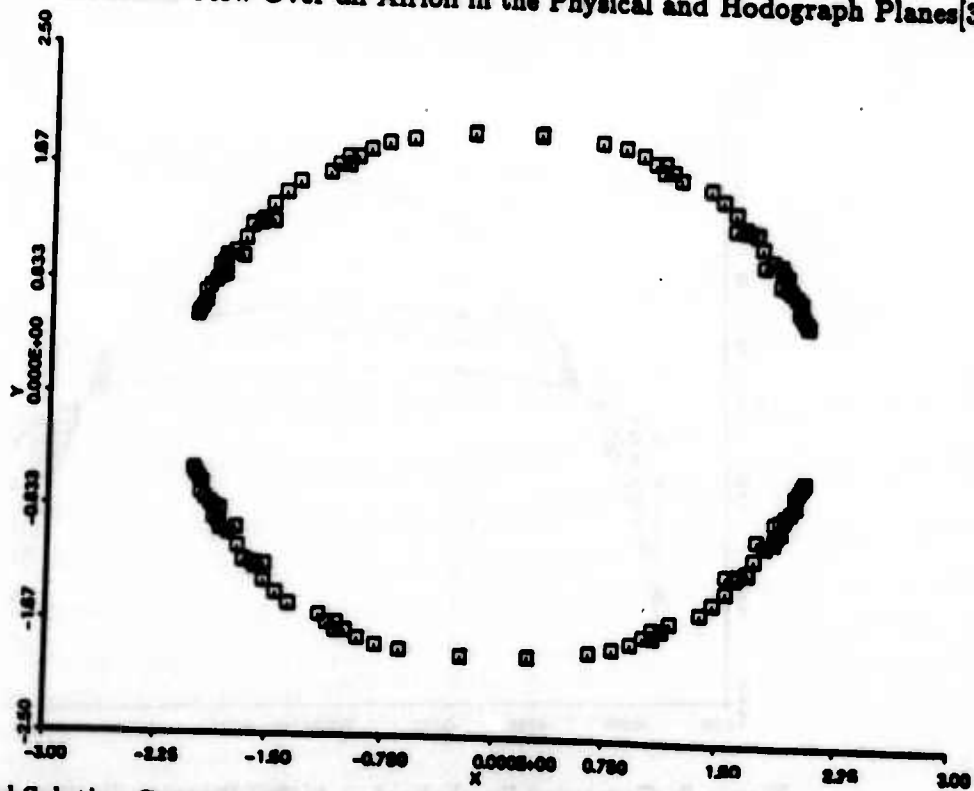
Figure 4: Initial Solution Corresponding to Cherry's Solution for Transonic Flow About a Circular Cylinder
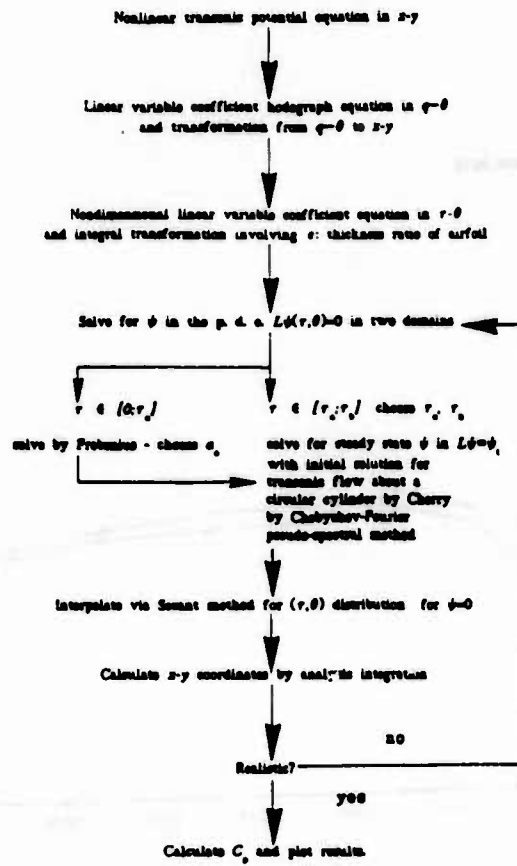
Figure 5: Flowchart for the Solution Procedure
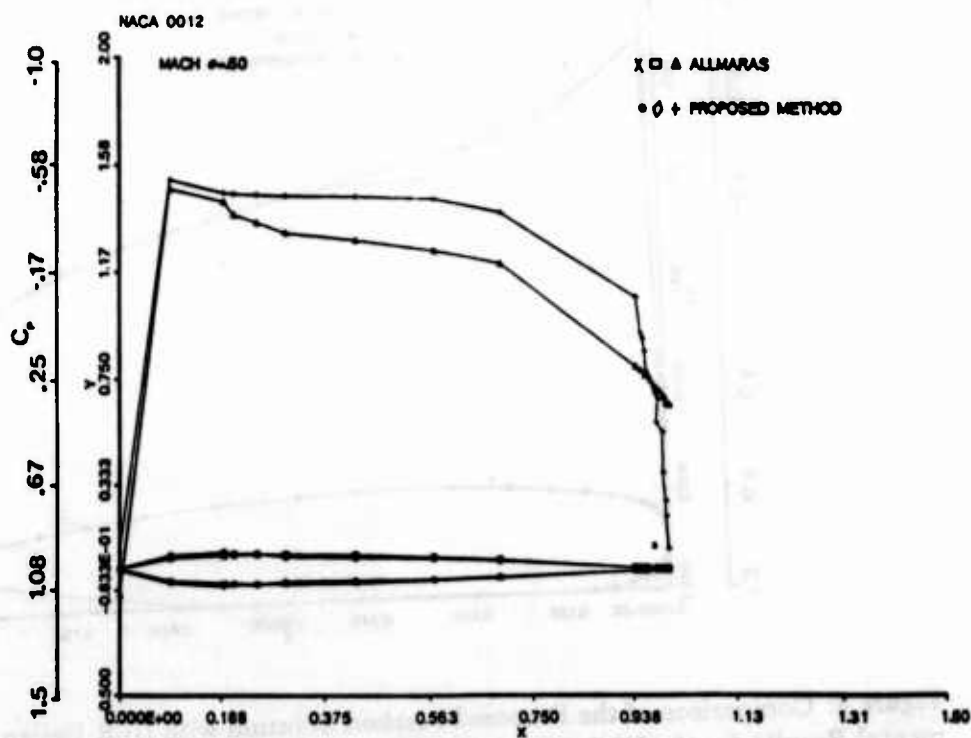
Figure 6: Comparison of Proposed Method Solution with Allmaras' Input and Results
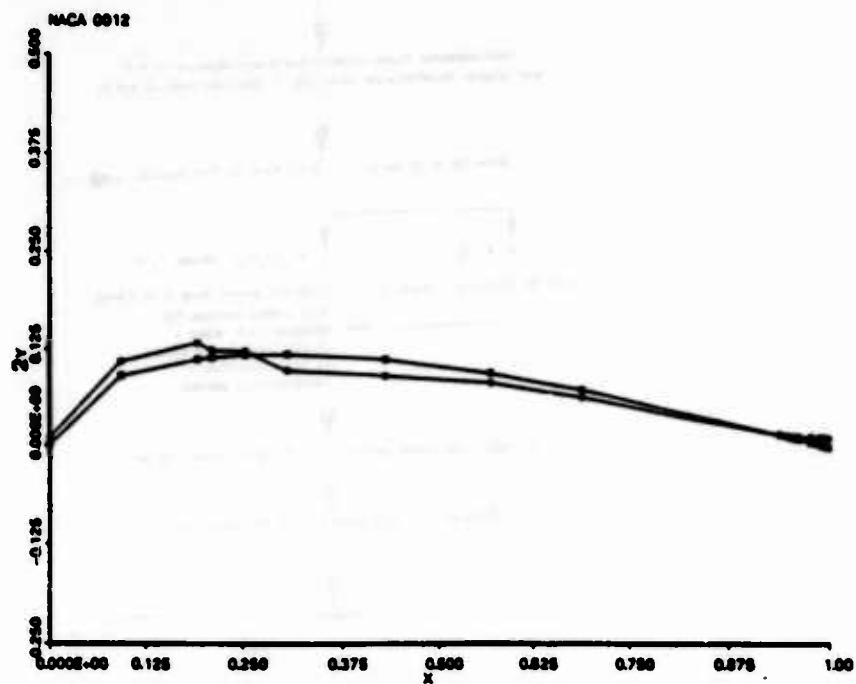
Figure 7: Close-Up Comparison of Proposed Method Solution with NACA 0012 Coodinates
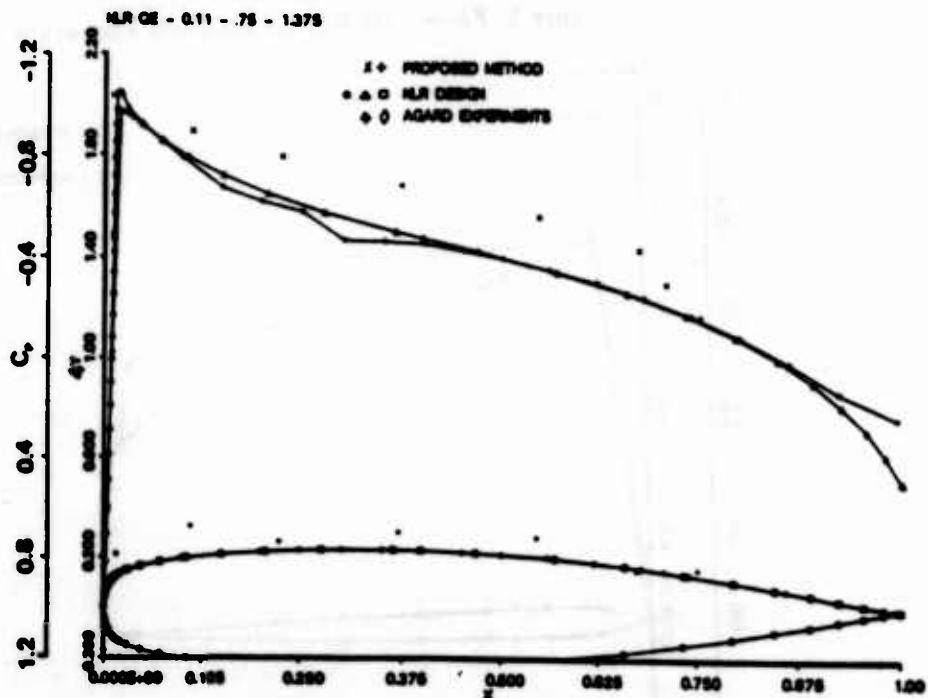
Figure 8: Comparison of the Proposed Method Solution with NLR Design Conditions and Experimental Results for the NLR QE - 0.11 - 0.75 - 1.375 Airfoil

514

A TOOLKIT OF SYMBOL MANIPULATION PROGRAMS
FOR VARIATIONAL GRID GENERATION†

Stanly Steinberg
University of New Mexico
Albuquerque, NM 87131

Patrick J. Roache
Ecodynamics Research Associates
Albuquerque, NM 87198

Abstract This paper describes some of the mathematical and programming ideas involved in the the creation of a toolkit of symbol manipulation programs which the authors have used to write a finite-difference elliptic partial differential equations solver.

I. Introduction The toolkit of MACSYMA symbol manipulation programs was developed in order to use symbol manipulation technology to write FORTRAN code. A brief description of the problems solved by the FORTRAN code is given in the Comments section of this paper; extensive descriptions are given in the cited literature. The current toolkit is based on a previous symbol manipulation project[4,9,10]. In the course of this, as well as the current project, two problems were encountered: the need for more memory than was available, and excessive use of computer time. Both of these problems could be overcome by the appropriate combination of certain mathematical and programming ideas; these ideas are be described here.

The general plan for the development of the toolkit was to organize the underlying mathematics in a way that was well-suited to symbol manipulation programming. At the same time, the overall structure of the FORTRAN code was taken into consideration. The toolkit is used to write subroutines that are incorporated into the FORTRAN code. The functions in the toolkit can be thought of as straightforward implementations of the steps that a human programmer would use to write the required subroutines.

The memory and time problems all had a similar form and solution. Straightforward programming of the mathematics led to the creation of very large expressions which could be avoided, or to unnecessary computations. The solution was to reorganize the mathematics and the program to avoid the large expression or unneeded computations. In the previous symbol manipulation project[10], the run time of one of our symbol manipulation programs was reduced from 60 cpu hours to 8 cpu minutes using the ideas described here. In the current project[12], attempts to write one of the needed subroutines at first caused the symbol manipulator to run out of memory after computing for several cpu hours; this subroutine can now be written in just over one cpu hour.

The problems encountered during the development of the toolkit are likely to occur in other areas. Consequently, the problems and their solutions have been reformulated in general mathematical terms. It is expected that this will promote a wider understanding of the techniques discussed here.

It is hoped that the reader is familiar with MACSYMA[3]. However, MACSYMA input and output is rather natural, so those not familiar may still be able to follow the discussion. MACSYMA is an interactive symbol manipulator; thus it prints a prompt of the form (c_1), (c_2), $\cdots$. The user then types a command to be executed (indicated by italics). The commands must be terminated by a ";" or a "\$"; the latter suppresses MACSYMA's response. Usually MACSYMA responds with a line that is labeled with (d_1), (d_2), $\cdots$ or (e_1), (e_2), $\cdots$ (responses are printed as displayed equations). The symbol % stands for the previous "d_" expression. The operator ":" is used for value assingment, the operator ":=" is used in function and array definitions, and "=" is used to describe an equality. The line labeled with "Time" is MACSYMA's estimate of the cpu time used to do the computation specified in the previous "c_" line. The MACSYMA sessions have been edited to conserve space.

II. Distances The functions in the toolkit compute the derivatives of quantities that are defined in terms of the Euclidean distance. If such calculations are not handled properly, they will cause MACSYMA to run out of memory or use unreasonable amounts of cpu time . The two dimensional Euclidean distance,

$$d(x,y) = \sqrt{x^2 + y^2}, \tag{1}$$

can be used to illustrate this problem. The x derivative of the distance is given by

$$\frac{\partial d}{\partial x}(x,y) = \frac{x}{\sqrt{x^2 + y^2}}. \tag{2}$$

It is this form of the derivative that creates difficulty; a better form is

$$\frac{\partial d}{\partial x}(x,y) = \frac{x}{d(x,y)}. \tag{3}$$

Note that in this formula the definition of the distance function is not needed; the derivatives are given implicitly. If x and y depend on a parameter, the chain rule must be used to compute the derivatives of the distance with respect to the parameter.

The implicit form of the derivative can be implemented in MACSYMA using the *gradef* command. This command specifies the gradient of a function even though the function itself is undefined. The following demonstration will make this point clearer. In the calculations, the arguments of the distance, x and y, will depend on the parameter t. The demonstration uses the following MACSYMA commands: *diff* differentiates, *ev* evaluates an expression using all of the information already in MACSYMA and the information in the remaining arguments, *ratsimp* performs a rational simplification.

(c_1) $gradef(d(x,y),x/d(x,y),y/d(x,y))\$$
(c_2) $'diff(d(x(t),y(t)),t);$

$$\frac{d}{dt}d(x(t),y(t)) \tag{d\_2}$$

(c_3) $ev(\%,diff);$
Time= 200 msec.

$$\frac{y(t)\frac{d}{dt}y(t)}{d(x(t),y(t))} + \frac{x(t)\frac{d}{dt}x(t)}{d(x(t),y(t))} \tag{d\_3}$$

(c_4) $ratsimp(ev(\%,d(x,y) := sqrt(x^2+y^2)));$
Time= 716 msec.

$$\frac{y(t)\frac{d}{dt}y(t)+x(t)\frac{d}{dt}x(t)}{\sqrt{(y(t)^2+x(t)^2)}} \tag{d\_4}$$

The next part of the computation illustrates the fact that use of the *gradef* function may entail the user performing simplifications that MACSYMA will not know because it was not given the definition of the distance function. Thus, even though the definition of the distance should not be used in the computations, the distance squared is simple, and this fact should be used. The MACSYMA command *subst* performs a substitution.

(c_5) $'diff(d(x,y),x,2);$

$$\frac{d^2}{dx^2}d(x,y) \tag{d\_5}$$

(c_6) $ev(\%,diff);$
Time= 200 msec.

$$\frac{1}{d(x,y)} - \frac{x^2}{d(x,y)^3} \tag{d\_6}$$

(c_7) $ratsimp(\%);$
Time= 166 msec.

$$\frac{d(x,y)^2-x^2}{d(x,y)^3} \tag{d\_7}$$

(c_8) $subst(x^2+y^2,d(x,y)^2,\%);$
Time= 66 msec.

517

$$\frac{y^2}{d(x,y)^3} \tag{d\_8}$$

The use of *ratsubst* instead of *subst* in (c_8) would replace d^3 by $(x^2+y^2)\,d$.

It may not be clear to the reader that there is any great advantage to using the *gradef* function. It is important to imagine a situation where the expressions being manipulated contain hundreds or thousands of terms; then slight variations in the method of writing the expressions can make a great difference in the size of the expressions and thus effect how fast MACSYMA can manipulate the expressions. Some of our computations involved expressions that contained several hundred instances of the three-dimensional distance $d(x,y,z)$ with x, y, and z themselves replaced by derivatives. To directly numerically evaluate such an expression, where the square root function would be called several hundred times, is inefficient. One must at least replace all of square roots with a single variable and then evaluate the single variable once before the large expression is evaluated. Using *gradef* makes this easy. In addition, if a large expression contains $d(x,y)$ rather than the definition of $d(x,y)$ in terms of a square root, MACSYMA can manipulate the expression faster because $d(x,y)$ is manipulated like a simple symbol in computations other than differentiation.

What is the best way to evaluate expressions that contain d? The difficulties can be illustrated using the simple polynomial,

$$\sum_{i=0}^{n} d^i, \tag{4}$$

for small values of n. The point of interest here is that the value of d^2 is known before the value of d is known. One way to take advantage of this is to write the polynomial (when $n=4$) as

$$\exp = 1 + d + d^2 + d \times d^2 + (d^2)^2. \tag{5}$$

This form of *exp* can be be computed in MACSYMA using the *ratsubst* (rational substitution) function.

(c_1) *exp : sum(d^i,i,0,4);*
Time= 500 msec.

$$1 + d + d^2 + d^3 + d^4 \tag{d\_1}$$

(c_2) *exp : ratsubst(d2,d^2,exp);*
Time= 66 msec.

$$1 + d2 + d(1 + d2) + d2^2 \tag{d\_2}$$

If the simplicity of d^2 is ignored, what can MACSYMA do with such an expression? This question is the same as asking what MACSYMA can do with general expressions.

MACSYMA has two useful functions, *horner* and *optimize*, which can be used to minimize operation counts in expressions. The following MACSYMA run illustrates the use of these functions. First consider Horner's rule:

(c_1) *eqn : d = sqrt(x^2+ y^2);*

$$d = \sqrt{(x^2+y^2)} \tag{d\_1}$$

(c_2) *exp : sum(d^i,i,0,4);*
Time= 533 msec.

$$1+d+d^2+d^3+d^4 \tag{d\_2}$$

(c_3) *exp : horner(exp);*
Time= 50 msec.

$$1+d(1+d(1+d(1+d))) \tag{d\_3}$$

If the value of d is substituted into the previous expression and the expression is expanded, the result will be a rather large expression. What can *optimize* do to recover the simplicity of the original expression?

(c_4) *exp : ev(expand(exp),eqn);*
Time= 550 msec.

$$1+x^2+x^4+y^2+2x^2y^2+y^4 \tag{d\_4}$$
$$+ \sqrt{(x^2+y^2)}+x^2\sqrt{(x^2+y^2)}+y^2\sqrt{(x^2+y^2)}$$

(c_5) *optimize(exp);*
Time= 3200 msec.

$$\mathbf{block}\ ([a,b,c],a:x^2,b:y^2,c:\sqrt{(a+b)}, \tag{d\_5}$$
$$1+a+x^4+b+2ab+y^4+c+ac+bc)$$

The last expression is a MACSYMA *block* that will return the same value as *exp*. Unfortunately, the *optimize* function did not do quite what was expected; it is preferable that x^4 be replaced by a^2 and y^4 by b^2. Also, the example is bit unfair in that the *optimize* function was given the polynomial in the worst form that could be found. However, this does illustrate the point that it is a good idea to write expressions so that common subexpressions are easily reconized. Note that the computation time for the *optimize* functions was large compared to the other computations.

The operation counts for the various forms of *exp* are (assuming that the 4-th powers are fixed up):

	+	×
(5)	5	4
(d_3)	5	5
(d_5)	8	8

Clearly there are a great many ways of writing expressions that involve distance functions. If the expression is rational rather than just a polynomial, and contains the distance function with several different arguments, the situation is even more difficult. In the toolkit, expressions are written in terms of the distance function (i.e. not using square roots) and then a variant of Horner's rule is applied to the expression.

III. Determinants and Matrices In the FORTRAN code generation computations there are many two-by-two and three-by-three matrices whose entries depend on several parameters. The functions in the toolkit differentiate the matrices, powers and inverses of them, determinants, and powers of the determinants with respect to each of several parameters. In the following computation, the inverse of the determinant of a three-by-three matrix is computed explicitly. The result is then differentiated with respect to one of the entries in the determinant, and "simplified". In large computations, such simplifications are essential; here it causes problems analogous to those met in larger calculations. The denominator of the resulting large expression is the square of the determinant, a fact that can be discovered using the MACSYMA *factor* function. Note that the factoring uses more cpu time than any other part of the calculation. The *remvalue* function removes the value previously assinged to a variable.

(c_1) *m:matrix([a,b,c],[r,s,t],[x,y,z]);*

$$\begin{bmatrix} a & b & c \\ r & s & t \\ x & y & z \end{bmatrix} \tag{d_1}$$

(c_2) *exp : 'diff(DELTA^(-1),x);*

$$\frac{d}{dx}\frac{1}{\Delta} \tag{d_2}$$

(c_3) *DELTA : expand(determinant(m));*
Time= 283 msec.

$$a\,s\,z - b\,r\,z - a\,t\,y + c\,r\,y + b\,t\,x - c\,s\,x \tag{d_3}$$

(c_4) *combine(expand(ev(exp, diff)));*
Time= 1650 msec.

$$c\,s - b\,t / (a^2 s^2 z^2$$

(d_4)

$$-2\,a\,b\,r\,s\,z^2 + b^2 r^2 z^2 - 2\,a^2 s\,t\,y\,z + 2\,a\,b\,r\,t\,y\,z$$

$$+2\,a\,c\,r\,s\,y\,z - 2\,b\,c\,r^2 y\,z + 2\,a\,b\,s\,t\,x\,z - 2\,b^2 r\,t\,x\,z$$

$$-2\,a\,c\,s^2 x\,z + 2\,b\,c\,r\,s\,x\,z + a^2 t^2 y^2 - 2\,a\,c\,r\,t\,y^2$$

$$+c^2 r^2 y^2 - 2\,a\,b\,t^2 x\,y + 2\,a\,c\,s\,t\,x\,y + 2\,b\,c\,r\,t\,x\,y$$

$$-2\,c^2 r\,s\,x\,y + b^2 t^2 x^2 - 2\,b\,c\,s\,t\,x^2 + c^2 s^2 x^2)$$

(c_5) *factor(denom(%));*
Time= 13483 msec.

$$(a\,s\,z - b\,r\,z - a\,t\,y + c\,r\,y + b\,t\,x - c\,s\,x)^2$$

(d_5)

(c_6) *remvalue(DELTA)$*

The next part of the calculation shows how to organize the above calculation so that the large denominator is avoided. The calculation is based on a well-known formula for the derivative of a determinant. Let m be a square matrix, $\tilde{m}$ be the cofactor matrix of m, and Δ be the determinant of m. The element $\tilde{m}_{ij}$ of the cofactor matrix is given by $(-1)^{i+j}$ times the determinant of the matrix obtained by deleting the i-th row and j-th column from m. If superscript t stands for transpose, then $m \times \tilde{m}^t = \Delta I$ where I is the identity matrix; that is, $\tilde{m}^t / \Delta$ is the inverse matrix of m. The cofactor expansion of the determinant about the i-th row of m gives

$$\Delta = \sum_j m_{ij} \tilde{m}_{ij} .$$

(1)

Differentiating this with respect to m_{ij} gives

$$\frac{\partial \Delta}{\partial m_{ij}} = \tilde{m}_{ij} .$$

(2)

Let the function *detr* stand for the determinant of a three-by-three matrix and the function *minr* stand for the determinant of a two-by-two matrix. The arguments of the function are the successive rows of the underlying matrix. The following MACSYMA computation illustrates this idea.

(c_7) *gradef(detr(a, b, c, r, s, t, x, y, z),*
 + minr(s,t,y,z), -minr(r,t,x,z), + minr(r,s,x,y),
 -minr(b,c,y,z), + minr(a,c,x,z), -minr(a,b,x,z),
 + minr(b,c,s,t), -minr(a,c,r,t), + minr(a,b,r,s))$
(c_8) *gradef(minr(a, b, r, s), + s, -r, -b, + a)$*
(c_9) *exp : 'diff(DELTA^(-1),a);*
Time= 16 msec.

$$\frac{d}{da}\frac{1}{\Delta} \qquad\qquad (d\_9)$$

(c_10) *DELTA : detr(a, b, c, r, s, t, x, y, z)$;
(c_11) *ev(exp, diff);*
Time= 150 msec.

$$-\frac{minr(s,t,y,z)}{detr(a,b,c,r,s,t,x,y,z)^2} \qquad\qquad (d\_11)$$

The reason for the simplicity is that the differentiation is done using *gradef* before the value of the determinant is specified.

The problem of computing derivatives of the inverse of a matrix (rather than the inverse of the determinant of the matrix) presents a similar challenge. One way to proceed is to compute the inverse of the matrix and then differentiate. Note that this formula contains the inverse of the determinant, so the difficulty described above causes even more serious problems. A better way to proceed is to use the known formula

$$\frac{\partial}{\partial t}\frac{1}{A} = -\frac{1}{A}\frac{\partial A}{\partial t}\frac{1}{A}. \qquad\qquad (3)$$

This allows the computation of the derivative before information about the inverse is used. This formula is used in a critical way by Steinberg and Roache[10] (see Formula 2.8, page 257).

IV. Finite Differences The toolkit must convert partial differential equations to finite difference form. This process can be illustrated using the elementary ordinary differential equation

$$au'' + bu' + cu = 0 \qquad\qquad (1)$$

where $'$ stands for differentiation and a, b, and c are constants. If simple second order centered differences are used, the finite-difference form of the above equation is

$$R_i u_{i+1} + C_i u_i + L_i u_{i-1} = 0 \qquad\qquad (2)$$

where R, C, and L are, respectively, the right, center, and left coefficients of the stencil for the difference equation. The toolkit is used to calculate the formulas for the coefficients of the stencil.

The simplest way to do this in MACSYMA is to substitute finite differences for the derivatives, expand the result, and then collect the coefficients of the differences. The toolkit[13] contains a function, *difference*, that returns centered differences, which is used here.

(c_1) *ode : a*diff(u(t),t,2)+ b*diff(u(t),t)+ c*u(t)=0;*

522

$$a \frac{d^2}{dt^2} u(t) + b \frac{d}{dt} u(t) + c\, u(t) = 0 \qquad \text{(d\_1)}$$

(c_2) *d[0] : difference(u,1);*

$$u_i \qquad \text{(d\_2)}$$

(c_3) *d[1] : difference(u,1,1);*

$$\frac{u_{i+1} - u_{i-1}}{2h} \qquad \text{(d\_3)}$$

(c_4) *d[2] : difference(u,1,1,1);*

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \qquad \text{(d\_4)}$$

(c_5) *deq : ode$*
(c_6) *for ii:2 step -1 thru 0 do*
  deq : subst(d[ii],diff(u(t),t,ii),deq)$
Time= 283 msec.
(c_7) *deq : expand(deq);*
Time= 383 msec.

$$\frac{b\, u_{i+1}}{2h} + \frac{a\, u_{i+1}}{h^2} - \frac{2a\, u_i}{h^2} \qquad \text{(d\_7)}$$

$$+ c\, u_i - \frac{b\, u_{i-1}}{2h} + \frac{a\, u_{i-1}}{h^2} = 0$$

(c_8) *stencil : coeff(deq, u[i+1]);*
Time= 66 msec.

$$\frac{b}{2h} + \frac{a}{h^2} = 0 \qquad \text{(d\_8)}$$

The reason for substituting for the higher derivatives first in line (c_6) can be found by reading the discussion of the *derivsubst* flag in the MACSYMA Manual[3].

In the above computation, the substitution was performed before the coefficients were taken. The order of these computations can be interchanged.

(c_9) *stencil : sum(*
  *ratcoeff(d[ii],u[i+1])*coeff(ode, diff(u(t),t,ii))*
    ,ii,0,2);
Time= 866 msec.

523

$$\frac{b}{2h} + \frac{a}{h^2} = 0 \qquad\qquad (d\_9)$$

To get the correct coefficient in an expression in MACSYMA, the expression must be expanded or the command *ratcoeff* must be used.

For such a simple problem, the two methods are equivalent. However, as the size of the differential equation grows and the space dimension increases, the second method is substantially more efficient because it eliminates unnecessary algebra. In the comments section of Steinberg and Roache[10], it is pointed out that the Laplacian in general coordinates in three dimensions (in fully expanded form) contains 1611 terms. For problems of this size, the first approach discussed in this section produced symbol code that ran for over 60 cpu hours[10]. A combination of the ideas discussed in this paper reduced this runtime to 8 cpu minutes.

V. Translating to FORTRAN In the FORTRAN codes it is necessary to numerically evaluate many expression that are defined in terms of derivatives of more fundamental quantities. In MACSYMA, these quantities are represented in terms of derivatives, while in the FORTRAN code the quantities are represented in terms of finite differences. Such quantities are evaluated by first approximating the derivatives of the fundamental quantity using finite differences and then representing the dependent quantities in terms of the finite differences of the fundamental quantities. This can be illustrated by computing the Jacobian of a change of variables in three dimensions.

Before the computation is started, the effect of the *simp* flag needs to be understood.

(c_1) *simp : false$*

(c_2) *1 *3+ 0;*

$$1 \cdot 3 + 0 \qquad\qquad (d\_2)$$

(c_3) *%, simp;*

$$3 \qquad\qquad (d\_3)$$

Setting the *simp* flag to *false* inhibits all simplification; in the line (c_3) the expression is evaluated with this flag set to *true*.

Also, the FORTRAN formulas for the finite differences are easy to generate using the *difference* function defined in the toolkit[13] and the *fortran* function provided in MACSYMA.

(c_1) *fortran(concat(x,1) = difference(x,3,1));*

    x1 = (x(i+1,j,k)-x(i-1,j,k))*(2*h1)**(-1)

Time= 600 msec.

The symbol *x*1 stands for the derivative of $x(\xi,\eta,\varsigma)$ with respect to its first argument.

Here is the computation of the Jacobian of a transformation where ξ, η, and ς are the old variables and *x*, *y*, and *z* are the new variables:

(c_1) *v : [x(xi, eta, zeta), y(xi, eta, zeta),*
 z(xi, eta, zeta)]$
(c_2) *nu : [xi, eta, zeta]$*
(c_3) *m[i,j] := diff(v[i], nu[j]);*
(c_4) *m : genmatrix(m,3,3,1,1);*
Time= 366 msec.

$$
\begin{bmatrix}
\dfrac{d}{d\xi}x(\xi,\eta,\varsigma) & \dfrac{d}{d\eta}x(\xi,\eta,\varsigma) & \dfrac{d}{d\varsigma}x(\xi,\eta,\varsigma) \\[2ex]
\dfrac{d}{d\xi}y(\xi,\eta,\varsigma) & \dfrac{d}{d\eta}y(\xi,\eta,\varsigma) & \dfrac{d}{d\varsigma}y(\xi,\eta,\varsigma) \\[2ex]
\dfrac{d}{d\xi}z(\xi,\eta,\varsigma) & \dfrac{d}{d\eta}z(\xi,\eta,\varsigma) & \dfrac{d}{d\varsigma}z(\xi,\eta,\varsigma)
\end{bmatrix}
\qquad (\text{d\_4})
$$

(c_5) *jacobi : expand(determinant(m));*
Time= 466 msec.

$$
\frac{d}{d\varsigma}x(\xi,\eta,\varsigma)\,\frac{d}{d\xi}y(\xi,\eta,\varsigma)\,\frac{d}{d\eta}z(\xi,\eta,\varsigma) \qquad (\text{d\_5})
$$

$$
-\frac{d}{d\xi}x(\xi,\eta,\varsigma)\,\frac{d}{d\varsigma}y(\xi,\eta,\varsigma)\,\frac{d}{d\eta}z(\xi,\eta,\varsigma)
$$

$$
-\frac{d}{d\varsigma}x(\xi,\eta,\varsigma)\,\frac{d}{d\eta}y(\xi,\eta,\varsigma)\,\frac{d}{d\xi}z(\xi,\eta,\varsigma)
$$

$$
+\frac{d}{d\eta}x(\xi,\eta,\varsigma)\,\frac{d}{d\varsigma}y(\xi,\eta,\varsigma)\,\frac{d}{d\xi}z(\xi,\eta,\varsigma)
$$

$$
+\frac{d}{d\xi}x(\xi,\eta,\varsigma)\,\frac{d}{d\eta}y(\xi,\eta,\varsigma)\,\frac{d}{d\varsigma}z(\xi,\eta,\varsigma)
$$

$$
-\frac{d}{d\eta}x(\xi,\eta,\varsigma)\,\frac{d}{d\xi}y(\xi,\eta,\varsigma)\,\frac{d}{d\varsigma}z(\xi,\eta,\varsigma)
$$

(c_6) *vars : [x, y, z]$*
Time= 33 msec.
(c_7) *for i thru 3 do for j thru 3 do (*
 temp1 : concat(vars[i],j),

```
temp2 : diff(v[i], nu[j]),
simp : false,
jacobi : subst(temp1, temp2, jacobi),
simp : true,
enddo)$
```
Time= 983 msec.

(c_8) *simp : false*$

(c_9) *jacobi*;

$$x3\,y1\,z2 + (-1)x1\,y3\,z2 + (-1)x3\,y2\,z1 \qquad \text{(d\_9)}$$

$$+ x2\,y3\,z1 + x1\,y2\,z3 + (-1)x2\,y1\,z3$$

(c_10) *simp : true*$

(c_11) *jacobi*;

$$-x3\,y2\,z1 + x2\,y3\,z1 + x3\,y1\,z2 \qquad \text{(d\_11)}$$

$$-x1\,y3\,z2 - x2\,y1\,z3 + x1\,y2\,z3$$

The substitutions performed in line (c_7) replace derivatives by simple symbols. Consequently, the resulting expression cannot be simplified; it can only be written more compactly. However, the MACSYMA simplifier (see Section 3.3 of the MACSYMA Manual[3]) will normally try to simplify the expression because MACSYMA does not know that no simplifications are possible. During the simplification attempt, the terms of the expression will be put into a canonical order. Because parts of the expression have been renamed, this can take considerable time. The *simp* flag is used to prevent the irrelevant reorderings.

VI. A Multivariate Horner's Rule After using all of the above techniques in the FORTRAN code generation problems, many expressions are still present which are large sparse multivariate polynomials. The toolkit contains a function called *horners* that chooses one of the variables in the polynomial and then uses the univariate MACSYMA *horner* function to write the polynomial in Horner's form. The coefficients of the resulting expression, which are labeled with $vt1, vt2, \cdots$, are again multivariate polynomials, with one less variable. The *horners* function is recursively applied to the coefficients of the resulting polynomials until the coefficients contain at most two terms.

The effects of this multivariate Horner's rule can be seen by applying *horners* to the square of the determinant of the three-by-three matrix from Section III. The *horners* function returns a list of formulas, so the MACSYMA function *ldisp* is used to display the formulas nicely.

(c_1) *m : matrix([a,b,c],[r,s,t],[x,y,z])*$
(c_2) *exp : vj2=expand(determinant(m)^2)*;
Time= 1250 msec.

$$vj2 =$$ (d_2)

$$c^2 s^2 x^2 - 2bcstx^2 + b^2 t^2 x^2 - 2c^2 rsxy$$

$$+2bcrtxy + 2acstxy - 2abt^2xy + c^2r^2y^2$$

$$-2acrty^2 + a^2t^2y^2 + 2bcrszz - 2acs^2xz$$

$$-2b^2rtxz + 2abstxz - 2bcr^2yz + 2acrsyz$$

$$+2abrtyz - 2a^2styz + b^2r^2z^2 - 2abrsz^2$$

$$+a^2s^2z^2$$

(c_3) *result : horners([exp])\$*
Time= 175450 msec.
(c_4) *for i thru length(result) do ldisp(result[i]);*

$$vt3 = s(sx^2 - 2rxy) + r^2y^2$$ (e_4)

$$vt14 = 2aty + 2brz$$ (e_5)

$$vt12 = x(vt14 - 2btx) + 2aryz$$ (e_6)

$$vt13 = y(2btx - 2aty)$$ (e_7)

$$vt11 = r(vt13 - 2bryz)$$ (e_8)

$$vt2 = vt11 + s(vt12 - 2asxz)$$ (e_9)

$$vt6 = t(tx^2 - 2rxz) + r^2z^2$$ (e_10)

$$vt10 = 2sx + 2ry$$ (e_11)

$$vt9 = vt10z$$ (e_12)

$$vt8 = t(vt9 - 2txy) - 2rsz^2$$ (e_13)

$$vt5 = a\, vt8$$ (e_14)

$$vt7 = t(ty^2 - 2syz) + s^2z^2$$ (e_15)

$$vt4 = a^2 vt7$$ (e_16)

$$vt\,1 = vt\,4 + b\,(vt\,5 + b\,vt\,6) \qquad\qquad (e\_17)$$

$$vj\,2 = vt\,1 + c\,(vt\,2 + c\,vt\,3) \qquad\qquad (e\_18)$$

Note the large amount of time required to do the Horner's calculation on this relatively simple expression. In addition, there are still a significant number of common subexpressions left in the resulting formulas. Note, for instance that r^2 occurs in two different places. The problem of finding more sophisticated ways of evaluating such expressions is of continuing interest.

VII. Large Coding Projects The toolkit is reasonably large; the MACSYMA source code requires about 79 kilobyte of storage and consists of 44 MACSYMA functions. Section 10.9 of the MACSYMA Manual[5] is called "Hints for Writers of Packages in MACSYMA" and contains several helpful ideas for managing projects of this size. In particular, the autoloading feature of MACSYMA, which is provided by the function *setup_autoload*, is very useful. This function allows function definitions to be loaded automatically when they are needed.

Suppose the current directory contains a file called **f.mac** that contains the definition of the function $f(x)$:

f(x) := x^3$

The following MACSYMA output illustrates how to autoload the definition of the function f.

(c_1) *f(3);*

$$f(3) \qquad\qquad (d\_1)$$

(c_2) *setup_autoload(f,f)$*

(c_3) *f(3);*
Batching the file f.mac
Batching done.

$$27 \qquad\qquad (d\_3)$$

Notice that the file **f.mac** was batched. The batching process is rather slow for long function definitions. Loading can be speeded up by using a LISP version or an object version of the function definition. If the definition of the function f is currently in MACSYMA, then the command

(c_1) *save("f.l",f)$*

will create a file with the name f.l that contains the LISP definition of the function f:

```
(defprop $f f autoload)
(add2lnc '$f $props)
(mdefprop $f ((lambda nil)
   ((mlist) $x) ((mexpt) $x 3)) mexpr)
(add2lnc '(($f) $x) $functions)
```

The LISP form of the function can be converted to object code using the LISP compiler LISZT; the object code will normally be put in the file named **f.o**.

The object form of the function definition will load faster than the LISP form of a function definition, which will, in turn, load considerably faster than the MACSYMA form. The MACSYMA *load* function can load any of these files; it first looks for **f.o**, then for **f.l**, and finally for **f.mac**.

VIII. Comments The toolkit of MACSYMA symbol manipulation programs was developed so that it could be used to write FORTRAN subroutines which are used to solve boundary-value problems for partial differential equations. The boundary-value problem solver has been used to model lasers[5,6] and other physical devices. The solver uses finite difference techniques[6,8] to solve the boundary value problem. The problems solved are interesting because they are posed in irregular regions and consequently the solver must generate a grid in the region, as well as solve the problem on the generated grid. The grids were previously generated using elliptic techniques and are now being generated using variational techniques.

The global structure of the toolkit is straightforward. As mentioned before, the toolkit consists of 44 MACSYMA functions. The functions either implement a small portion of computations described in Steinberg and Roache[12] or combine several functions to do a more complicated task. In general, it is possible to understand the MACSYMA code by following the mathematics[12], the FORTRAN code listings[14], and the symbol code listing[13]. This paper discusses the situations where the implementation is least obvious.

Surprisingly, MACSYMA does not handle differentiation well. This point has been thoroughly discussed elsewhere[2,15,16]. The problem has to do with the way that MACSYMA uses a dependencies (*depends*) notion to implement the chain rule for differentiation. This works well for the differentiation of known function, but not so well for general functions. The toolkit has to deal with many general functions, so this caused a substantial problem.

The computations done in this paper were done on a Sun Microsystems workstation (Sun2/160) with 4 mbytes of main memory and a 380 mbyte (formatted) Eagle disk drive. The operating system is Sun UNIX 4.2 Release 2.0 which contains 4.2 BSD updates. The Beta Test Release 308.2 of Symbolics, Inc. MACSYMA was used to do the symbol manipulation. (The toolkit could be implemented in other symbol manipulation languages.) The MACSYMA output and this paper were prepared using text processors *tbl*, *eqn*, and *troff* that are part of the standard UNIX distribution.

References

1. J.E. Castillo, S. Steinberg and P.J. Roache, On the folding of numerically generated grids, to be published in *Applied Mathematics and Computation*.

2. J.P. Golden, Differentiation of unknown functions in Macsyma, *SIGSAM Bulletin*, **19-2** (1985), 19-24.

3. MACSYMA Reference Manual, Version 10, 3rd Printing, Symbolics, Inc., Cambridge, 1984.

4. P.J. Roache, S. Steinberg, Symbolic manipulation and computational fluid dynamics, AIAA 6th Computational Fluid Dynamics Conference, Danvers, Massachusetts, AIAA, New York, 1983, Paper 83-1952. Also appeared in *AIAA Journal*, **22-10** (1984), 1390-1394.

5. P.J. Roache, S. Steinberg, H.J. Happ, W.M. Moeny, 3D electric field solutions in boundary fitted coordinates, Proceedings of the 4th IEEE Pulsed Power Conference, Albuquerque, NM, June 1983.

6. P.J. Roache, W.M. Money, S. Steinberg, Interactive Electric Field Calculations for Lasers, Proceedings of the AIAA 17th Fluid Dynamics, Plasma Dynamics, and Lasers Conference, Snowmass, Colorado, June, 1984, 25-27.

7. P.J. Roache, S. Steinberg, A new approach to grid generation using a variational formulation, Proceedings of the AIAA 7th Computational Fluid Dynamics Conference, Cincinnati, Ohio, July 1985.

8. P.J. Roache and S.Steinberg, Application of a Single Equation MG-FAS Solver to Elliptic Grid Generation Equations (Sub-grid and Super-grid Coefficient Generation), Proceedings of the Second Copper Mountain Conference on Multigrid Methods, 1-3 April 1985, Copper Mountain Colorado.

9. S. Steinberg, P.J. Roache, Using VAXIMA to write FORTRAN code, The 1984 MACSYMA User's conference, General Electric Research and Development Center, Schenectady, New York, July 1984. Also appeared in Applications of Computer Algebra, edited by R. Pavelle, Kluwer, 1985, 74-93.

10. S. Steinberg, P.J. Roache, Symbolic manipulation and computational fluid dynamics, *Journal of Computational Physics*, **57** (1985), 251-284.

11. S. Steinberg, P.J. Roache, Using MACSYMA to write Fortran Subroutines (letter), *MACSYMA Newsletter*, **II-2** (1985), 10-12. Also to appear in the Journal of Symbolic Computation.

12. S. Steinberg, P.J. Roache, Variational grid generation, to appear in *Numerical Methods for Partial Differential Equations*.

13. S. Steinberg, P.J. Roache, A Listing of MACSYMA Code for Variational Grid Generation, Technical report, Ecodynamics Research Associates, Inc. and Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM

87131, in preparation.

14. S. Steinberg, P.J. Roache, A Listing of FORTRAN Code for Variational Grid Generation, Technical report, Ecodynamics Research Associates, Inc. and Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, in preparation.

15. M. Wester, S. Steinberg, An extension to MACSYMA's concept of functional differentiation, *SIGSAM Bulletin*, **17** (1983), 25-30.

16. M. Wester, S. Steinberg, A survey of symbolic differentiation implementations, The 1984 MACSYMA User's conference, General Electric Research and Development Center, Schenectady, New York, July 1984.

A SELF-ADAPTIVE GRIDDING FOR INVISCID TRANSONIC PROJECTILE AERODYNAMICS COMPUTATION***

Chen-Chi Hsu and Chyuan-Gen Tu
Department of Engineering Sciences
University of Florida
Gainesville, Florida 32611

Extended Abstract. A good grid system for the computation of complex fluid dynamics problems can be justified from the smoothness of grids, the orthogonality of grids, and the grid resolution adaptive to the solution in the physical space. In fact the use of an improper grid can be detrimental to the solution accuracy as well as to the convergence process of a solution algorithm. An adaptive grid generation method proposed by Brackbill [1] is rather general and seems to be a very promising approach for complex flow problems. In this method the governing differential equations for a 2-D adaptive grid are derived from extremizing the general functional

$$I = \int_\Omega [(\nabla\xi)^2 + (\nabla\eta)^2]d\Omega + \lambda_0 \int_\Omega [(\nabla\xi \cdot \nabla\eta)^2 J^3 d\Omega + \lambda_W \int_\Omega wJd\Omega \qquad (1)$$

in which ξ and η are curvilinear grid coordinates while J is the Jacobian of the transformation representing the grid size which can be made adaptive to the control function $w(x,y)$. The integrals in Eq. (1) are, respectively, smoothness, orthogonality, and grid resolution functionals. Introducing characteristic quantities L_c, L_p and W, the Lagrange's multipliers can be chosen as

$$\lambda_0 = \alpha \left(\frac{L_c}{L_p}\right)^4 \quad , \quad \lambda_W = \beta \frac{1}{W} \left(\frac{L_c}{L_p}\right)^4 \qquad (2)$$

so that each integral has the same order of magnitude provided α and β are of $O(1)$. Hence, the relative importance of the three integrals to a grid can be identified from the value of α and β chosen. An application of the adaptive grid generation method to a 2-D inviscid supersonic flow past a step in a wind tunnel has been studied by Saltzman [2] and the results obtained showed that the adaptive mesh generator moves the computational grid with shock fronts and consequently enhances significantly the desirable resolution of the finite-difference scheme for the accuracy.

We have investigated an application of the adaptive grid generation method to transonic inviscid flows past a secant-ogive-cylinder-boattail projectile with sting at zero angle of attack. The control function for grid resolution is chosen as the computed pressure gradient while a conforming one-dimensional variational principle is employed for boundary grid adaptation. We found that the grid generation method cannot provide good grids for use in a thin-layer Navier-Stokes code [3] which has an option for solving inviscid flow problems; consequently, a resulting grid has to be implemented with an exponential clustering technique to provide good meshes near the projectile surface. The results computed then were acceptably accurate and in good agreement with measured data reported in [4]. Moreover, for generating a better adaptive grid, we have modified the constrained variational principle,

Eq. (1) by considering variable parameters λ_ϕ and λ_ψ for enhancing grid resolution locally but assuming their variation zero; accordingly, local grid spacings are chosen for the reference lengths L_c and L_D in Eq. (2). A grid generation code has been developed and coupled to the Navier-Stokes code for self-adaptive grid generation and the numerical study conducted showed that the adaptive grid generation technique developed indeed can provide, without any experimentation, good grids for the transonic projectile aerodynamics computation. For instance with a strategy of generating a new adaptive grid fixed at every 150 integration time steps of the Navier-Stokes code, the results obtained for three flow cases (M_∞ = 0.91, 0.96, and 1.10) showed that the computed surface pressure coefficient is in excellent agreement with the reported measured data.

References

1. Brackbill, J. U., "Coordinate System Control: Adaptive Meshes," in Numerical Grid Generation ed. by J. F. Thompson, North-Holland, 1982

2. Saltzman, J., "A Variational Method for Generating Multidimensional Grids," Ph.D. Thesis, New York University, 1981

3. Nietubicz, C. J., Pulliam, T. H. and Steger, J. L., "Numerical Solution of the Azimuthal-Invariant Thin-Layer Navier-Stokes Equations," Paper 79-0010, AIAA 17th Aerospace Sciences Meeting, Jan. 1979.

4. Kayser, L. D. and Whiton, F., "Surface Pressure Measurements on a Boattailed Projectile Shape at Transonic Speeds," ARBRL-MR-03161, U. S. Army Ballistic Research Laboartory, March 1982

SURFACE PRESSURE CALCULATED AT NT= 1000
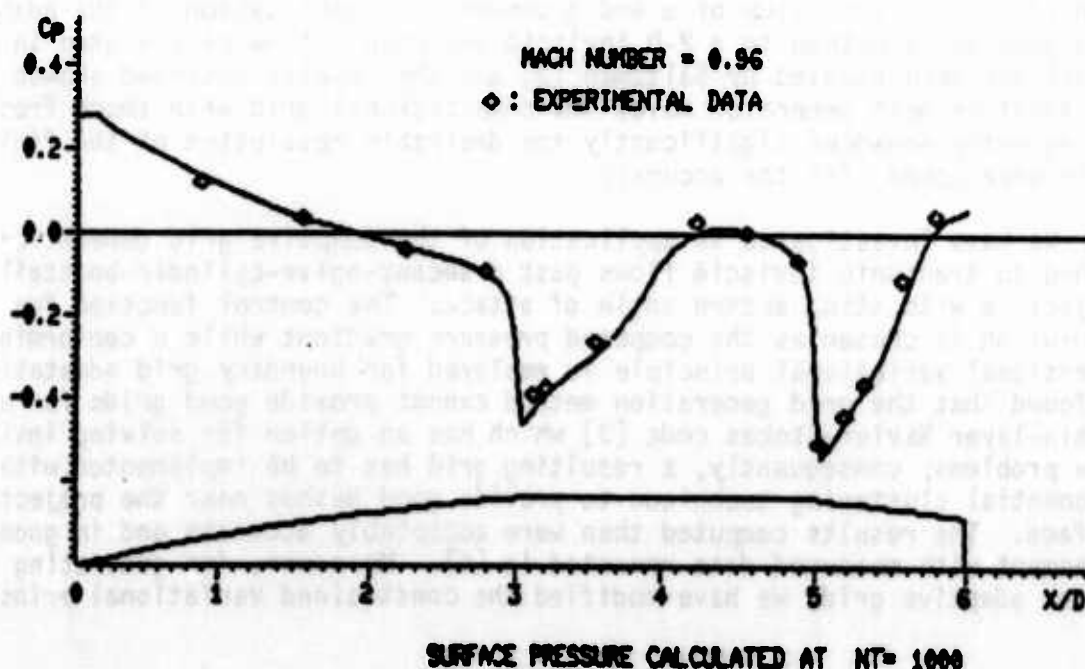
ON COMPUTATION OF TRANSONIC PROJECTILE AERODYNAMICS

Chen-Chi Hsu and Nae-Haur Shiau
Department of Engineering Sciences
University of Florida
Gainesville, Florida 32611

ABSTRACT. A representative transonic flow, $M_e = 0.96$, past a secant-ogive-cylinder-boattail projectile model with sting at zero angle of attack has been considered for detailed investigation on the application and implication of the thin-layer Navier-Stokes approximation implemented with the Baldwin-Lomax turbulence model for accurate prediction of the transonic projectile aerodynamics. An axisymmetric thin-layer Navier-Stokes code and a grid generation code obtained from the U.S. Army Ballistic Research Laboratory are employed to solve the flow problem. The numerical results obtained from the use of different hyperbolic grids show that the Navier-Stokes code can provide accurate surface pressure if a good grid is provided. The results also show that the accuracy of surface pressure is very sensitive not only to the boundary grid resolution but also to the grid distribution in normal direction while the shear stress distribution in the shock-boundary layer interaction region depends strongly upon the predicted shock location. The importance of a good adaptive grid is evidenced from the computed results which show that the ratio of pressure drag to skin-friction drag can be off by as much as 40% from that of an accurate result.

I. INTRODUCTION. An accurate prediction of the aerodynamic force and other flow characteristics is essential to a better design of aerodynamic devices and flight vehicles. For a practical aerodynamic problem, wind-tunnel experiments are traditionally performed to measure the aerodynamic force and other desirable flow characteristics. With the rising cost of experimental measurements, it is becoming extremely expensive to conduct parametric studies in a wind-tunnel; moreover, each test facility has a limited range of application and consequently certain flow conditions of interest often cannot be simulated. Hence, the numerical simulation of a complex aerodynamic problem, with recent advent of supercomputers, has been becoming an alternate approach to complement the wind-tunnel experiment for effective design. Recently a thin-layer Navier-Stokes code has been developed at NASA Ames Research Center for unsteady three-dimensional high speed compressible flow problems [1]. This code is based on the Reynolds-averaged thin-layer Navier-Stokes equations for ideal gas in a transformed boundary-fitted space and the transformed governing equations are approximated by Beam and Warming factorized finite difference scheme in which a second order implicit (ϵ_T) and a fourth order explicit (ϵ_E) artificial dissipation terms have been added for controlling numerical stability of the solution algorithm. The turbulence closure model implemented is a two-layer algebraic eddy viscosity model [2]. The Navier-Stokes code also has been simplified for axisymmetric projectile flow problems [3]. Both of these codes have an option for solving inviscid flow problems while a steady solution is resulted from a converged solution of the unsteady flow problem.

The application of the Navier-Stokes codes to transonic projectile aerodynamic problems has been investigated to some extent by the U. S. Army Ballistic Research Laboratory [4-8]. The grid provided to a Navier-Stokes code is an axisymmetric grid system formed by a sequence of planar grids

around the axis of a projectile model; the planar grid is obtained from a grid generation code GRIDGEN which can give either an elliptic grid or a hyperbolic grid [9]. For a secant-ogive-cylinder-boattail projectile with sting at zero angle of attack, the published results showed that the computed surface pressure coefficient C_p on the secant-ogive portion and boattail portion of the projectile agrees rather well with measured data but the agreement on the cylinder portion (shock wave-boundary layer interaction region) is not very satisfactory for some flow cases considered. For the projectile model at two-degree angle of attack, the reported C_p-distribution agrees qualitatively with measured data but quantitatively the agreement over the cylinder portion and boattail portion is not satisfactory at all.

The published results have indicated that the thin-layer Navier-Stokes codes can give acceptably accurate surface pressure for the complex transonic projectile aerodynamic problem if a good adaptive grid system is provided. However, the precise causes for the unsatisfactory results reported are yet to be investigated; moreover, no result on the skin-friction coefficient has been reported and discussed. Therefore, the main objective of this study is to further advance our understanding on the application and implication of the thin-layer Navier-Stokes approximation implemented with Baldwin-Lomax algebraic turbulence model for accurate prediction of aerodynamic forces acting on a transonic projectile.

II. THE FLOW PROBLEM. A representative transonic flow of M_∞ = 0.96 past a secant-ogive-cylinder-boattail (SOCBT) projectile model with sting at zero angle of attack is considered for detailed investigation in this study. The projectile model has a 3-caliber secant-ogive part followed by a 2-caliber cylinder and a 1-caliber 7-degree boattail which is further extended for another 1.77 calibers to meet a horizontal sting. The projectile model with sting has a total length of 16 calibers. There are surface pressure measurements reported for transonic flows M_∞ = 0.91, 0.94, 0.96, 0.98, 1,10 and 1.20 past the SOCBT projectile [5]. The flow problem considered is solved with an axisymmetric thin-layer Navier-Stokes code and a grid generation code GRIDGEN obtained from the Army Ballistic Research Laboratory. It is mentioned in passing that an averaged-technique is used in the Navier-Stokes code for computing eddy viscosity, which results in improper zig-zag eddy viscosity distributions. Hence, the averaged-scheme for eddy viscosity has been deleted from the code in this study.

III. RESULTS AND DISCUSSION. An application of the Navier-Stokes code for projectile aerodynamic problems requires that the user provides a planar grid. In this study the planar grid provided to the Navier-Stokes code is a modified hyperbolic grid [10]. As indicated in reference [9], the grid resolution of a hyperbolic grid is somewhat predetermined by the prescribed boundary grid distribution and the choice of a clustering function. In the grid generation code GRIDGEN an exponential clustering function is employed to ensure sufficiently fine grid resolution for the viscous sublayer; however, other clustering functions such as a hyperbolic tangent can also be used [11]. Figure 1(a) shows a 78 x 28 hyperbolic grid obtained from GRIDGEN while Figure 1(b) is a 78 x 28 hyperbolic grid generated with the use of a hyperbolic tangent clustering function. It is observed that the characteristics of the two clustering functions are exhibited in the distribution of normal grid points. A number of different hyperbolic grids has been considered for the flow problem to investigate the implication of the
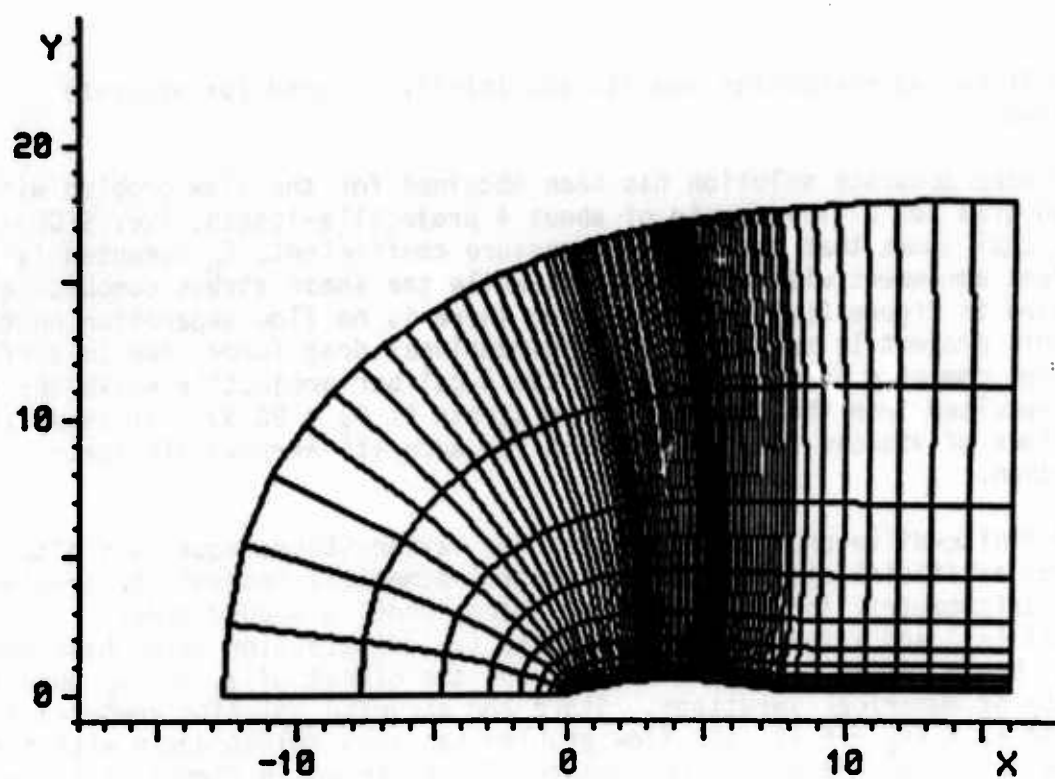
Figure 1(a). A 78 x 28 hyperbolic grid based on exponential clustering function.
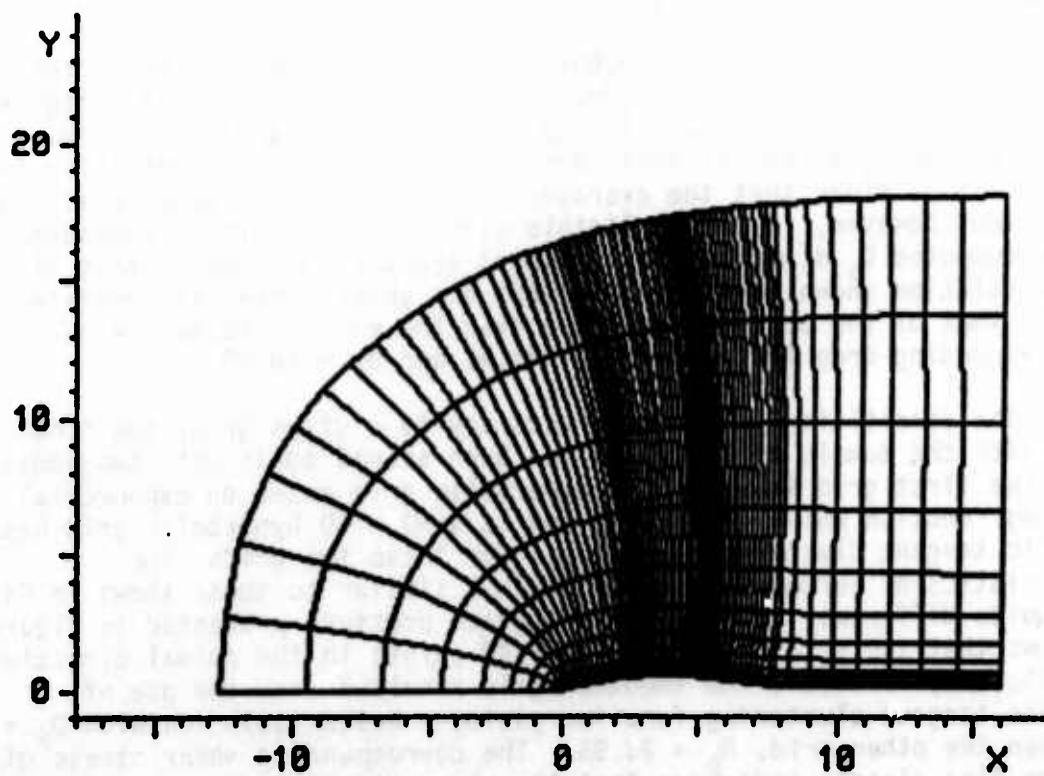
Figure 1(b). A 78 x 28 hyperbolic grid based on hyperbolic tangent clustering function.

Navier-Stokes approximation and its sensitivity to grid for accurate solutions.

A very accurate solution has been obtained for the flow problem with a 90 x 60 grid for a flow domain of about 4 projectile-length, i.e. STOT = 24. Figure 2(a) shows that the surface pressure coefficient, C_p computed is in excellent agreement with measured data while the shear stress computed and presented in Figure 2(b) indicates that there is no flow separation on the transonic projectile surface. A non-dimensional drag force, due to surface pressure, computed is D_p = 40.28 for the 6-caliber projectile while the drag force resulted from the computed shear stress is D_f = 20.53. It shows the importance of viscous flow computation for accurate aerodynamic force prediction.

A finite-difference approximation for Navier-Stokes equations often requires artificial dissipations to control numerical instability problem. In the axisymmetric thin-layer Navier-Stokes code, a second order implicit (ε_I) and a fourth order explicit (ε_E) dissipation terms have been added. Hence, it is important to find out the effect of ε_I and ε_E upon the accuracy of numerical solutions. Since the accurate solution computed is based on $\varepsilon_I = 2\varepsilon_E = 4 \Delta t$, the flow problem has been solved again with the same grid but $\varepsilon_I = 2\varepsilon_E = 8 \Delta t$. The computed C_p, as shown in Figure 2(a), seems to agree very well with the accurate solution; however, the resulting pressure drag force is D_p = 32.91 which is 18% less than that of the accurate solution. Figure 2(b) shows the difference on shear stress distribution but the resulting shear drag force is D_f = 20.12.

The effect of the averaged-technique originally implemented in the Navier-Stokes code for computing eddy viscosity also has been investigated. Figure 3 shows the distribution of eddy viscosity with and without the averaging scheme at three different boundary point stations identified in Figure 2. It is clear that the averaged-technique yields imporper zig-zag distribution; however, it has negligible effect on the surface pressure. In fact the computed C_p distribution is almost exactly the same as that of the accurate solution shown in Figure 2(a) but the shear stress is consistently less than that of the accurate solution over the entire projectile surface. The corresponding drag forces are D_p = 41.41 and D_f = 18.99.

For the sensitivity of solution accuracy to a given grid, the flow problem with the domain of STOT = 24 has been solved again with two additional grids. The first grid is a 90 x 40 hyperbolic grid based on exponential clustering function while the second one is a 90 x 40 hyperbolic grid based on hyperbolic tangent clustering function. For these two grids, the characteristics of normal grid distribution, similar to those shown in Figure 1, are quite different. The computed surface pressure presented in Figure 4(a) shows that the grid resolution with 40 points in the normal direction is not sufficient; moreover, the smoother grid resulted from the use of hyperbolic tangent clustering function yields a better solution with D_p = 27.38 than the other grid, D_p = 23.95. The corresponding shear stress given in Figure 4(b) clearly indicates that the shear stress is very sensitive to the computed pressure field in the shock wave-boundary layer interaction regions; however, the resulting shear drag is about the same, 20.05 and 20.28. It should be pointed out that a smoother grid does not always provide a more accurate solution. In fact the flow problem with a domain of STOT = 18
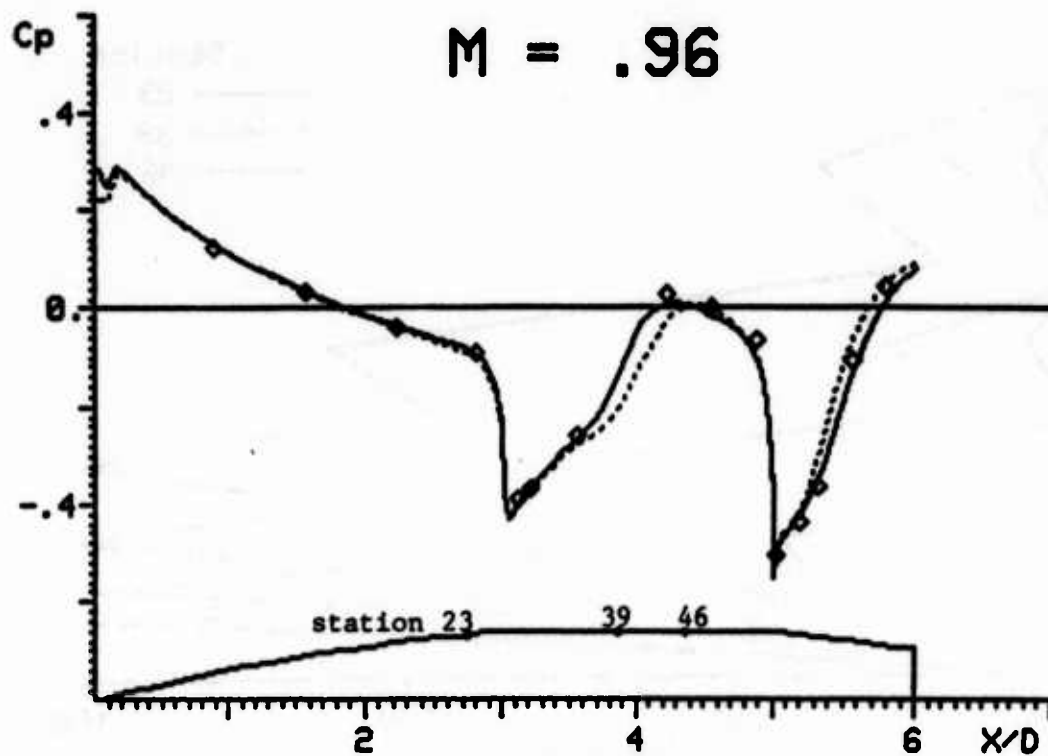
Figure 2(a). Surface pressure distribution obtained with a 90 x 60 grid.
——: $\epsilon_I = 2\epsilon_E = 4\Delta t$; ---- : $\epsilon_I = 2\epsilon_E = 8\Delta t$; ◊ : measured data.
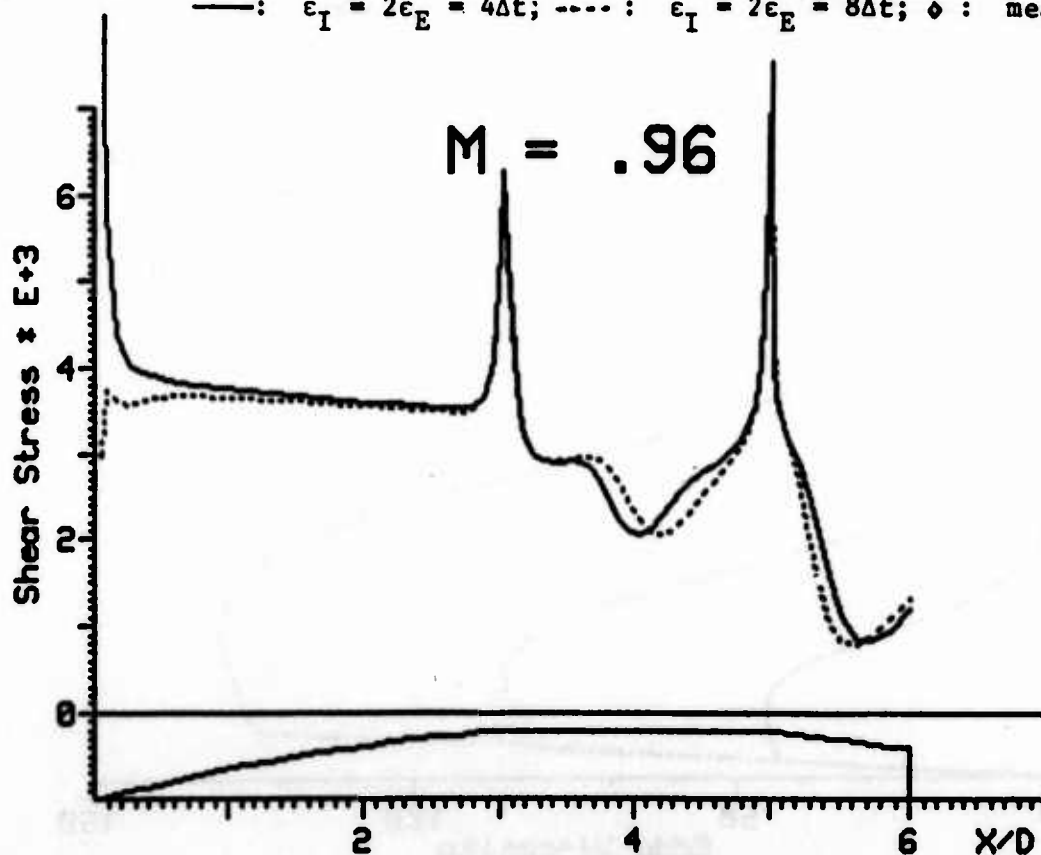
Figure 2(b). Surface shear stress distribution computed with a 90 x 60 grid.
—— : $\epsilon_I = 2\epsilon_E = 4\Delta t$; ---- : $\epsilon_I = 2\epsilon_E = 8\Delta t$.
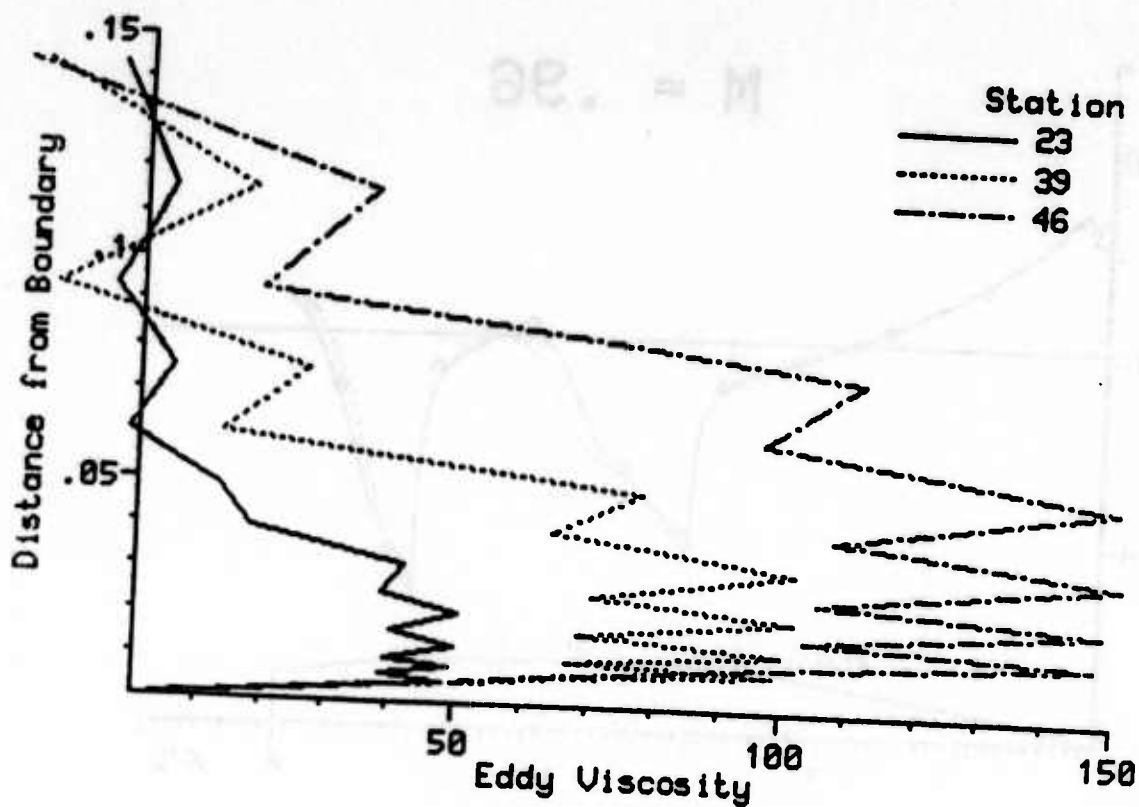
Figure 3(a). Sample eddy viscosity distributions based on an averaged-scheme implemented in the original Navier-Stokes code.
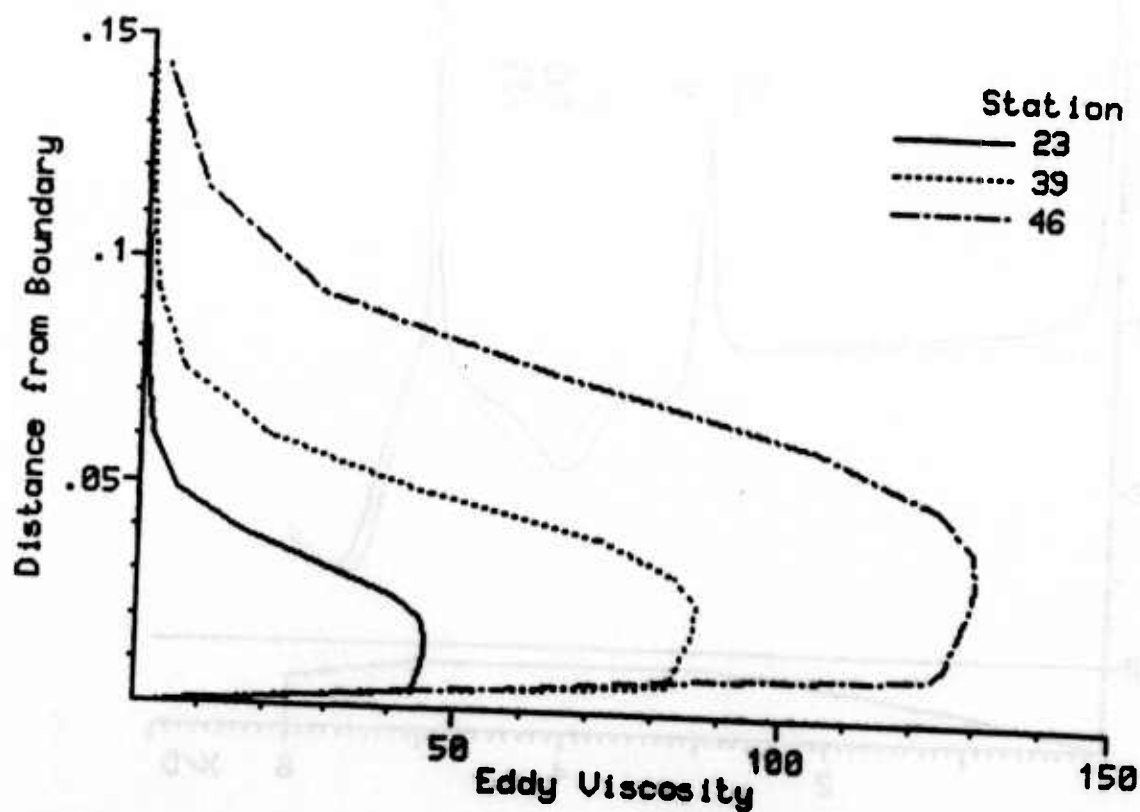
Figure 3(b). Sample eddy viscosity distributions computed without averaging.
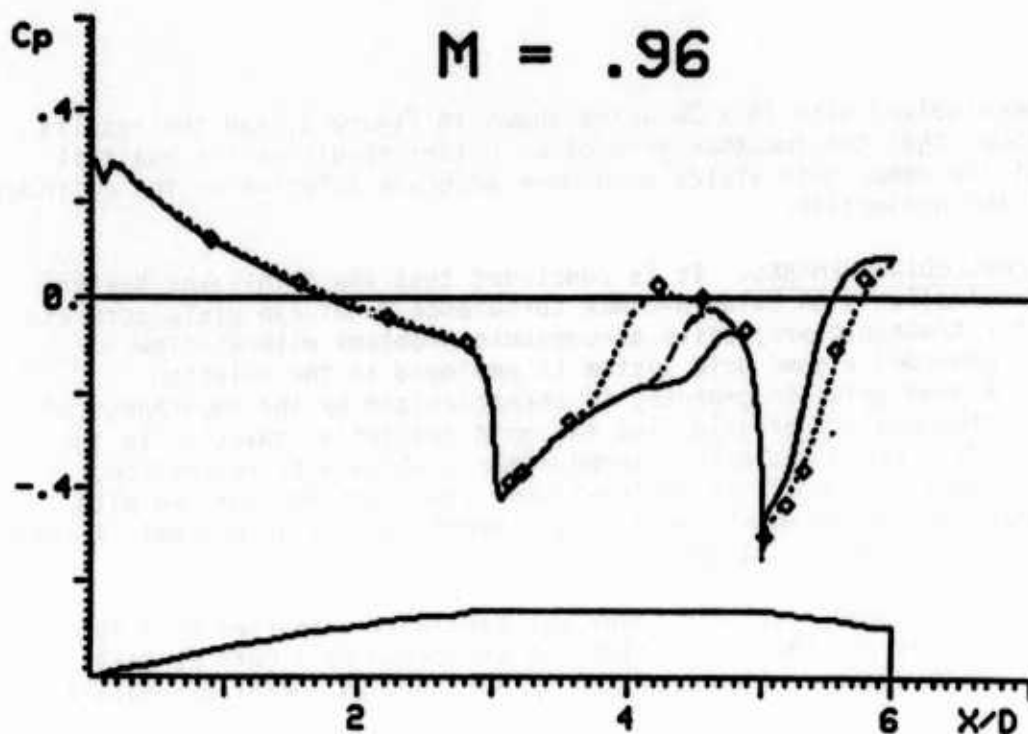
Figure 4(a). Surface pressure coefficient computed with different grids.
···· : 90 x 60 grid from GRIDGEN; —— : 90 x 40 grid from GRIDGEN;
—·—: 90 x 40 grid based on hyperbolic tangent clustering function.
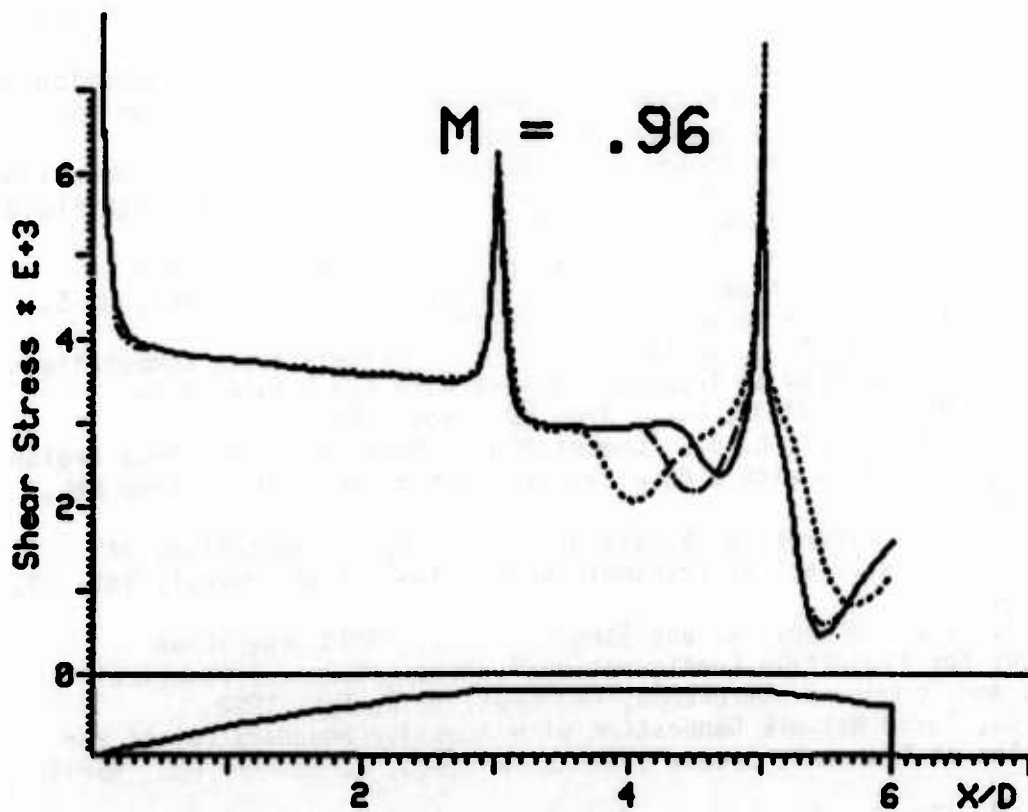
Figure 4(b). The corresponding shear stress distribution computed with different grids.

has also been solved with 78 x 28 grids shown in Figure 1, and the results obtained shows that the smoother grid gives better result on the boattail portion but the other grid yields much more accurate solution on the cylinder portion of the projectile.

IV. CONCLUDING REMARKS. It is concluded that the thin-layer Navier-Stokes approximation with Baldwin-Lomax turbulence model can yield accurate solutions for transonic projectile aerodynamic problems without flow separation, provided a good grid system is employed in the solution algorithm. A good grid, in general, is characterized by the smoothness of grids, the orthogonality of grids and the grid resolution adaptive to the solution fields. For a projectile aerodynamic problem with separation, however, the application of the Navier-Stokes equations implemented with Baldwin-Lomax turbulence model for accurate prediction of aerodynamic forces is yet to be investigated and verified.

REFERENCES

1. Pulliam, T. H. and Steger, J. L., "Implicit Finite-Difference Simulations of Three-Dimensional Compressible Flow", AIAA Journal, Vol. 18, Feb. 1980.
2. Baldwin, B. S. and Lomax, H., "Thin-Layer Approximation and Algebraic Model for Separated Turbulent Flows", Paper 78-257, AIAA 16th Aerospace Sciences Meeting, Jan. 1978.
3. Nietubicz, C. J., Pulliam, T. H. and Steger, J. L., "Numerical Solution of the Azimuthal-Invariant Thin-Layer Navier-Stokes Equations", Paper 79-0010, AIAA 17th Aerospace Sciences Meeting, Jan. 1979.
4. Nietubicz, C. J., "Navier-Stokes Computations for conventional and Hollow Projectile Shapes at Transonic Velocities", AIAA-81-1262, AIAA 14th Fluid and Plasma Dynamics Conference, June 1981.
5. Kayser, L. D. and Whiton, F., "Surface Pressure Measurements on a Boattailed Projectile Shape at Transonic Speeds", ARBRL-MR-03161, U. S. Army Ballistic Research Laboratory, March 1982.
6. Sahu, J., Nietubicz, C. J. And Steger, J. L., "Navier-Stokes Computations of Projectile Base Flow at Transonic Speeds with and without Base Injection", ARBRL-TR-02532, U. S. Army BRL, Nov. 1983.
7. Sahu, J. and Nietubicz, C. J., "Computational Modeling of the Base Region Flow for a Projectile with a Base Cavity", BRL-MR-3436, U. S. Army BRL, April 1985.
8. Nietubicz, C. J., Sturek, W. B. and Heavey, K. R., "Computations of Projectile Magnus Effect at Transonic Velocities", AIAA Journal, Vol. 23, July 1985.
9. Nietubicz, C. J., Heavey, K. and Steger, J. L., "Grid Generation Techniques for Projectile Configurations", Proc. 1982 Army Numerical Analysis and Computers conference, ARO Rept. 82-3, Feb. 1982.
10. Hsu, C. C., "Grid Network Generation with Adaptive Boundary Points for Projectiles at Transonic Speeds", ARBRL-TR-02485, U. S. Army BRL, April 1983.
11. Vinokur, M., "On One-Dimensional Stretching Functions for Finite-Difference Calculation", J. Comp. Physics, 50, 1983.

NUMERICAL SIMULATION OF SUPERSONIC FLOW OVER A ROTATING BAND

Jubaraj Sahu
Computational Aerodynamics Branch
Launch and Flight Division
US Army Ballistic Research Laboratory, LABCOM
Aberdeen ᵒroving Ground, MD 21005-5066

ABSTRACT. Implicit, approximately factored, finite difference codes have been developed for solving the Navier-Stokes equations in general body-fitted coordinates. For a protuberance such as the rotating band on artillery shell, sharp geometric variations exist which make it extremely difficult to generate body-conforming grids while preserving the sharp corners. Using wrap around grids for such cases introduces geometric errors and may lead to degradation of computational efficiency and accuracy. This paper describes the development and applicatioᵣ ᵒf a computational procedure using flowfield blanking to compute the flow ovᵤ. a rotating band at supersonic speed with no geometric error.

I. INTRODUCTION. In recent years, a considerable research effort has been focused on the development of modern predictive capabilities for determining the aerodynamics of projectiles. Time-dependent Navier-Stokes computational technique has been used[1,2] to compute the flow over projectiles at transonic speeds. For supersonic flows, space-marching parabolized[3] Navier-Stokes computational technique can be effectively used. However, this technique fails for flows containing separation regions in the streamwise direction. In such cases which are encountered frequently in projectile aerodynamic simulations, time-dependent Navier-Stokes technique can still be used.[4,5]

The time dependent Navier-Stokes equations are solved in generalized body-fitted coordinate system. Many actual projectiles configurations contain sharp corners and 90° bends; in other words, sharp geometric variations exist on shell which make it extremely difficult to generate body-conforming grids while preserving the sharp corners. The grid lines are wrapped around the corners and in many cases, such wrap around grids are skewed near these corners and bends. Using such grids introduces geometric errors and may lead to loss in both the computational efficiency and accuracy. The purpose of this paper is to develop and apply a flow field blanking procedure which allows computation of practical flows of interest with no geometric error since it models the corners and bends exactly.

To avoid geometric errors one can blank out the flow field in specific regions in the computational domain. Examples where such blanking can be useful are shown in Figure 1. Continuous straight line grids can be used for these cases and the hatched regions are the ones where the flow field is to be blanked out. This procedure, thus, preserves the sharp corners and bends. In addition to zeroing out the flow field inside the hatched regions, additional changes must be made in terms of boundary conditions on these zonal surfaces and the computational algorithm near these surfaces which are described in a later section. The simplest example to test this technique is the rotating band flow problem. The rotating band is a protuberance on the artillery shell and is primarily used to impart spin to the shell during launch. In free

flight, however, it does contribute a small unwanted drag. A schematic of the rotating band flow field is shown in Figure 2. It shows the expected recirculation regions in front of and behind the band and the associated compressions and expansion waves. Numerical solution is obtained for this problem at $M_\infty = 3.0$ and $\alpha = 0$.

II. COMPUTATIONAL TECHNIQUE.

A. GOVERNING EQUATIONS. The complete set of time-dependent generalized axisymmetric thin-layer Navier-Stokes equations is solved numerically to obtain a solution to this problem. The numerical technique used is an implicit finite difference scheme. Although time-dependent calculations are made, the transient flow is not of primary interest at the present time. The steady flow is the desired result, which is obtained in a time asymptotic fashion.

The azimuthal invariant (or generalized axisymmetric) thin-layer Navier-Stokes equations for curvilinear coordinates ξ, η and ζ can be written as [1]

$$\frac{\partial \hat{q}}{\partial \tau} + \frac{\partial \hat{E}}{\partial \xi} + \frac{\partial \hat{G}}{\partial \zeta} + \hat{H} = \text{Re} \, \frac{\partial \hat{S}}{\partial \zeta} \tag{1}$$

where

$\xi = \xi(x,y,z,t)$ is the longitudinal coordinate

$\eta = \eta(y,z,t)$ is the circumferential coordinate

$\zeta = \zeta(x,y,z,t)$ is the near normal coordinate

$\tau = t$ is the time

and

$$\hat{q} = J^{-1} \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ e \end{bmatrix} \quad , \quad \hat{E} = {}^{-1} \begin{bmatrix} \rho U \\ \rho u U + \xi_x p \\ \rho v U + \xi_y p \\ \rho w U + \xi_z p \\ (e+p)U = \xi_t p \end{bmatrix} \quad , \quad G = J^{-1} \begin{bmatrix} \rho W \\ \rho u W + \xi_x p \\ \rho v W + \xi_y p \\ \rho w W + \xi_z p \\ (e=P)W - \xi_t p \end{bmatrix} \, ,$$

$$H = J^{-1} \begin{bmatrix} 0 \\ 0 \\ \rho V[R_\xi(U-\xi_t) + R_\zeta(W-\zeta_t)] \\ -\rho V R(V-\eta_t) - p/R \\ 0 \end{bmatrix}$$

544

$$\hat{S} = \begin{bmatrix} 0 \\ \mu(\varsigma_x^2 + \varsigma_y^2 + \varsigma_z^2)u_\varsigma + (\mu/3)(\varsigma_x u_\varsigma + \varsigma_y v + \varsigma_z w_\varsigma)\varsigma_x \\ \mu(\varsigma_x^2 + \varsigma_y^2 + \varsigma_x^2)v_\varsigma + (\mu/3)(\varsigma_x u_\varsigma + \varsigma_y v_\varsigma + \varsigma_z w_\varsigma)\varsigma_y \\ \mu(\varsigma_x^2 + \varsigma_y^2 + \varsigma_z^2)w_\varsigma + (\mu/3)(\varsigma_x u_\varsigma + \varsigma_y v_\varsigma + \varsigma_z w_\varsigma)\varsigma_z \\ \{(\varsigma_x^2 + \varsigma_y^2 + \varsigma_y^2)[(\mu/2)(u^2 + v^2 + w^2)_\varsigma + \kappa Pr^{-1}(\gamma-1)^{-1}(a^2)_\varsigma] \\ + (\mu/3)(\varsigma_x u + \varsigma_y v + \varsigma_z w)(\varsigma_x y_\varsigma + \varsigma_y v_\varsigma + \varsigma_z w_\varsigma)\} \end{bmatrix}$$

The velocities

$$U = \xi_t + \xi_x u + \xi_y v + \xi_z w$$

$$V = n_t + n_x u + n_y v + n_z w \tag{2}$$

$$W = \varsigma_t + \varsigma_x u + \varsigma_y v + \varsigma_z w$$

represent the contravariant velocity components.

The Cartesian velocity components (u, v, w) are nondimensionalized with respect to a_∞ (free stream speed of sound). The density (ρ) is referenced to ρ_∞ and total energy (e) to $\rho_\infty a_\infty^2$. The local pressure is determined using the equation of state,

$$P = (\gamma - 1)[e - 0.5\rho(U^2 + V^2 + W^2)] \tag{3}$$

where γ is the ratio of specific heats.

While Equation (1) contains only two spatial derivatives, it retains all three momentum equations, thus allowing a degree of generality over the standard axisymmetric equations. In particular, the circumferential velocity is not assumed to be zero, thus allowing computations for spinning projectiles or swirl flow to be accomplished.

B. COMPUTATIONAL ALGORITHM. The azimuthal thin-layer Navier-Stokes equations are solved using an implicit approximate factorization finite difference scheme in delta form.[6] An implicit method was chosen because, for viscous flow problems, it permits a time step much greater than that allowed by explicit schemes. The Beam-Warming implicit algorithm has been used in various applications[1-9] for the equations in general curvilinear coordinates.

The algorithm is first-order accurate in time and second- or fourth-order accurate in space. The equations are factored (spatially split), which reduces the solution process to one-dimensional problems at a given time level. Central difference operators are employed and the algorithm produces block tridiagonal systems for each space coordinate. The main computational work is contained in the solution of these block tridiagonal systems of equations.

C. FLOW FIELD BLANKING. The idea is to avoid geometric errors that may arise from wrap around grids. Instead, we use straight line grids as shown schematically in Figure 3. For the rotating band problem, the zone ABCD is part of the body and the flow field in this zone must be blanked out in the computational domain. As shown in Figure 3, the sharp corners and 90° bends ahead of and behind the band are preserved and no approximation is made. It is also necessary to apply boundary conditions on the zonal surfaces AB, BC and CD. As an initial attempt, inviscid boundary conditions are used at these boundaries since the grid is rather coarse at these boundaries. In addition, at neighboring points to these boundaries, we use second-order spatial difference and smoothing. The block tridiagonal matrix structure has been modified for continuous integration sweeps through such zones. Although, we have only one zone for the rotating band case, changes have been made in the code to blank out multiple zones.

III. RESULTS. All the numerical computations were made at $M_\infty = 3.0$ and $\alpha = 0$. The projectile configuration with the rotating band that was used in this study is shown in Figure 4. This model is a cone-cylinder configuration with a 13.1° cone angle. The band height is .04 D and the width is .505 D. The same model was used in the experiments[10] which were conducted in the US Army Chemical Research Development and Engineering Center's Supersonic Wind Tunnel. Surface pressure measurements have been made ahead of and behind the band which are used to compare with the numerical results.

Since the freestream flow is supersonic, the space marching Parabolized Navier-Stokes code[3] was used to compute the solution over the forebody of the projectile (See Figure 5). This generated a solution at a station 30 band heights ahead of the band which was then used as an upstream boundary condition for the computation of the flow field containing the rotating band. For this part of the flow field which includes the band, the unsteady or time-dependent Navier-Stokes computational technique described earlier was used. Such composite solution technique allowed a large number of grid points to be used in the vicinity of the band.

The computational grid used for the numerical calculations is shown in Figure 6. It consists of 139 points in the longitudinal direction and 60 points in the normal direction. The grid points are clustered near the surface of the cylindrical part with a minimum spacing of .00002 D. The resolution of grid points on the top of the band is not as fine. Grid points in the longitudinal direction are clustered near the upstream and downstream corners of the rotating band where appreciable changes in the flow variables are expected. In Figure 6, the grid lines inside the band are omitted to show the position of the band; however, in actual grid used in the computations, there are continuous grid lines inside the band and those are the lines where the flow field blanking procedure is used.

For comparison purposes, numerical solution is first obtained for flow over the cylindrical part of the projectile without the rotating band at $M_\infty =$ 3.0 and $\alpha = 0$. Computed surface pressure coefficient is plotted in Figure 7 as a function of the longitudinal position. The computed result is in very good agreement with experimental data.[10]

Numerical results obtained for the rotating band case are presented next. Figure 8a shows the velocity vector field in front of the band and as expected, it shows recirculatory flow in that region. The reverse flow region extends about four band heights ahead of the band. Figure 8b shows the velocity vectors behind the band. The flow seems to expand over a large portion of the band height. A smaller recirculation region can be observed. The flow expansions can be better seen in Figure 9 which shows the pressure contours for this case. One can also see a separation shock wave ahead of the band. The shock wave is located just ahead of the flow separation region. The surface pressure coefficient for the band case is shown in Figure 10 as a function of the axial position. The solid line is the computed result, the dashed line is the result obtained for the case without the band and the circles are the experimental data for the band. There is a considerable change in the pressure due to the band. The sharp rise in pressure ahead of the band is associated with the compression waves which actually precedes the separation point of the boundary layer flow. The flow then expands near the corner and pressure drops. No significant change in pressure occurs on the top of the band. At the backward step of the band, the flow expands again which results in the sharp decrease in the pressure. This is followed by a more gradual return to the ambient pressure downstream. The computed surface pressure is in good agreement with the experimental data measured ahead and behind the band.

IV. CONCLUDING REMARKS. Navier-Stokes computational has been used in conjuction with a flow field blanking procedure for numerical simulation which models the sharp corners and 90° bends exactly, thereby, avoiding any possible source of geometric errors. This procedure has been applied to the flow over a rotating band at supersonic speed.

Computed results have been obtained at $M_\infty = 3.0$ and $\alpha = 0$ and compared with available experimental data. The results show the recirculation region both ahead of and behind the rotating band as well as the associated compression and expansion waves. Computed surface pressure coefficient for both cases, with and without the band, is in fairly good agreement with the experimental data. The present numerical procedure is simple to use and seems to predict the flow field correctly.

REFERENCES

1. Nietubicz, C.J., Pulliam, T.H. and Steger, J.L., "Numerical Solution of the Azimuthal-Invariant Navier-Stokes Equations," ARBRL-TR-02227, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, March 1980. (AD A085716) (Also see AIAA Journal, Vol. 18, No. 12, December 1980, pp. 1411-1412)

2. Nietubicz, C.J., "Navier-Stokes Computations for Conventional and Hollow Projectile Shapes at Transonic Velocities," ARBRL-MR-03184, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, July 1982. (AD A116866)

3. Sturek, W.B. and Schiff, L.B., "Numerical Simulation of Steady Supersonic Flow Over Spinning Bodies of Revolution," AIAA Journal, Vol. 20, No. 12, December 1982, pp. 1724-1731.

4. Sahu, J., Nietubicz, C.J. and Steger, J.L., "Navier-Stokes Computations of Projectile Base Flow with and without Base Injection," ARBRL-TR-02532, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, November 1983. (AD A135738) (Also see AIAA Journal, Vol. 23, No. 9, September 1985, pp. 1348-1355.

5. Sahu, J., "Supersonic Flow over Cylindrical Afterbodies with Base Bleed," ARBRL-TR-2742, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, June 1986.

6. Beam, R. and Warming, R.F., "An Implicit Factored Scheme for the Compressible Navier-Stokes Equations," AIAA Paper No. 77-645, June 1977.

7. Steger, J.L., "Implicit Finite Difference Simulation of Flow About Arbitrary Geometries with Application to Airfoils," AIAA Journal, Vol. 16, No. 4, July 1978, pp. 679-686.

8. Pulliam, T.H. and Steger, J.L., "On Implicit Finite-Difference Simulations of Three-Dimensional Flow," AIAA Journal, Vol. 18, No. 2, February 1980, pp. 159-167.

9. Pulliam, T.H., "Artificial Dissipation Models for the Euler Equations," AIAA Paper No. 85-0438, January 1985.

10. Danberg, J.E. and Palko, K.L., "Measurement of Surface Pressures Caused by a Projectile Rotating Band at Supersonic Speeds," ARBRL-MR-3532, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, July 1986.
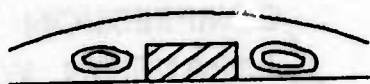
ROTATING BAND
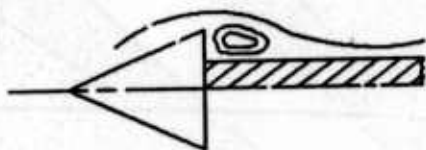
STING MOUNTED
MODEL

BASE BLEED OR
JET FLOW
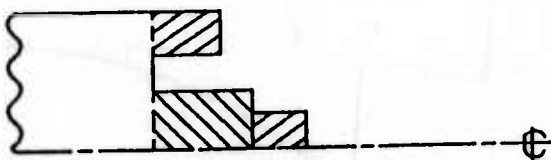
BASE CAVITY
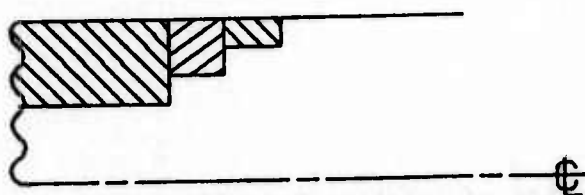
VARIABLE–AREA
SHOCK-TUBE

Figure 1. Examples of Flowfield Blanking
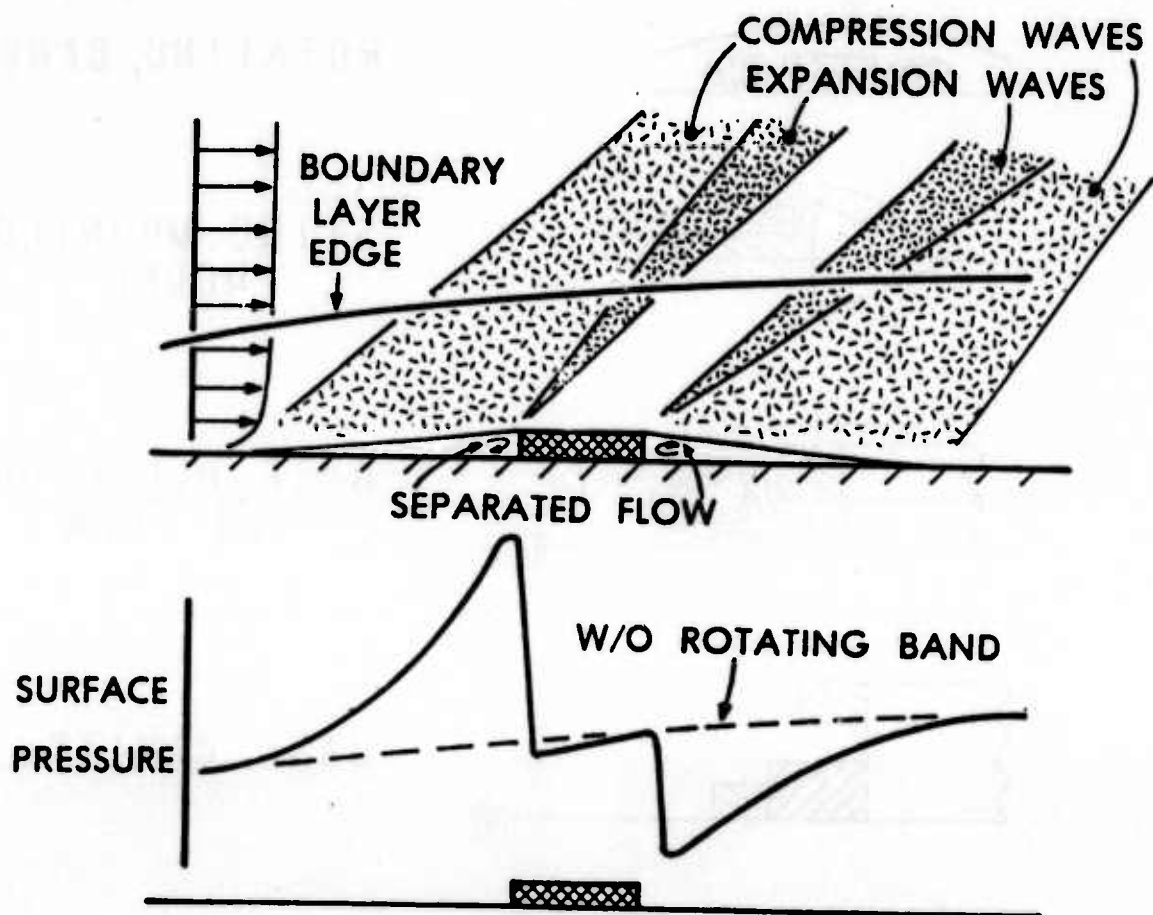
Figure 2. Schematics of Rotating Band Flowfield
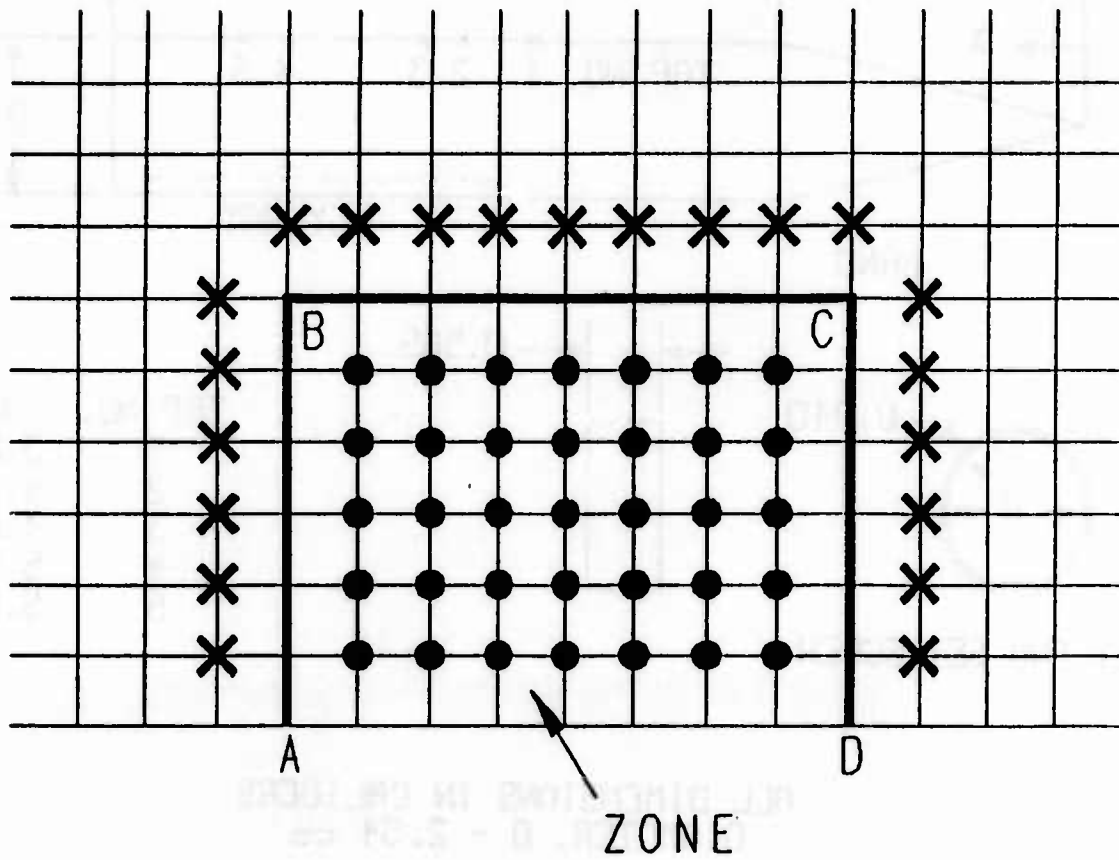
Figure 3. Schematic Illustration of Flowfield Blanking
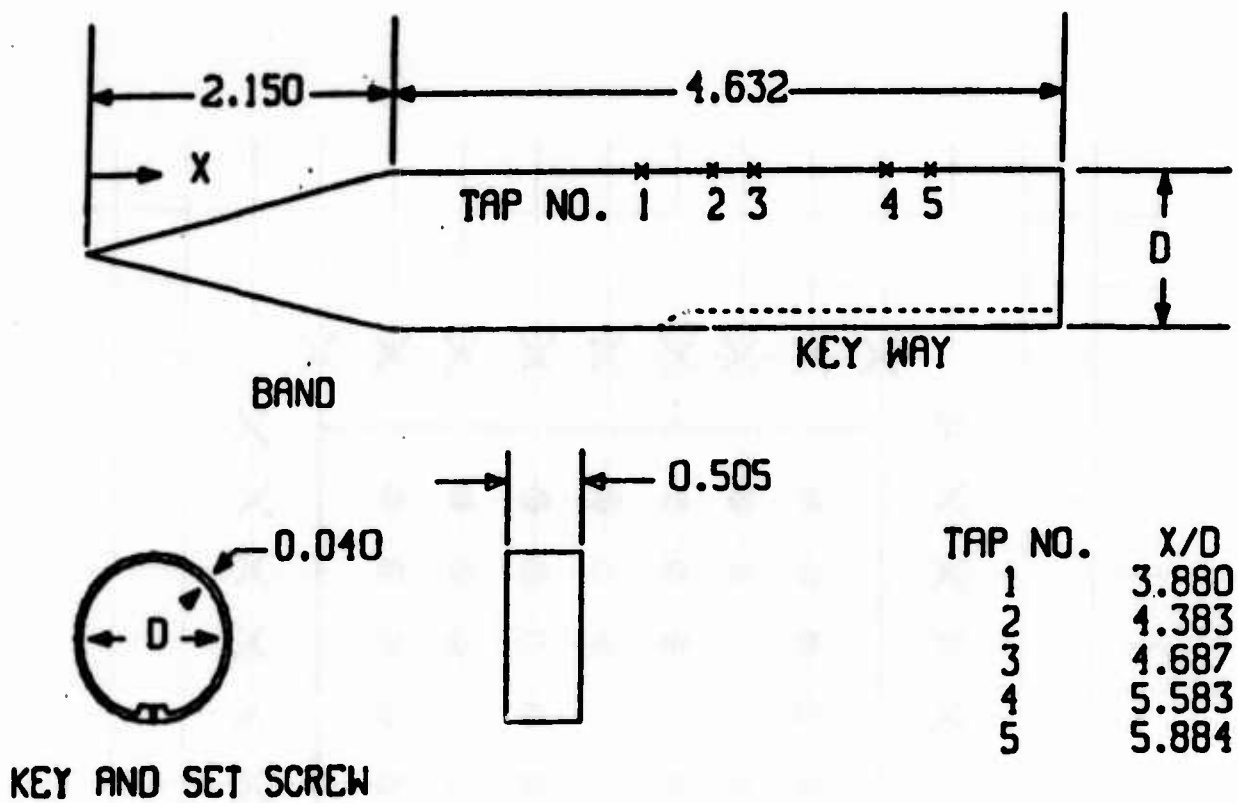
ALL DIMENSIONS IN CALIBERS
DIAMETER, D - 2.54 cm

Figure 4. Model Geometry
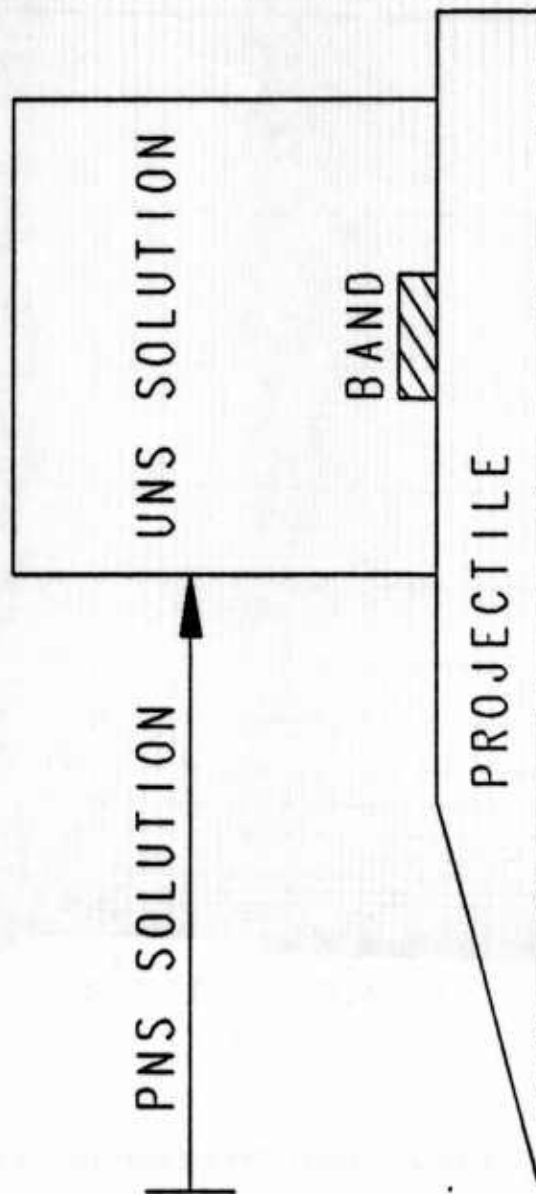
Figure 5. Composite Solution Technique
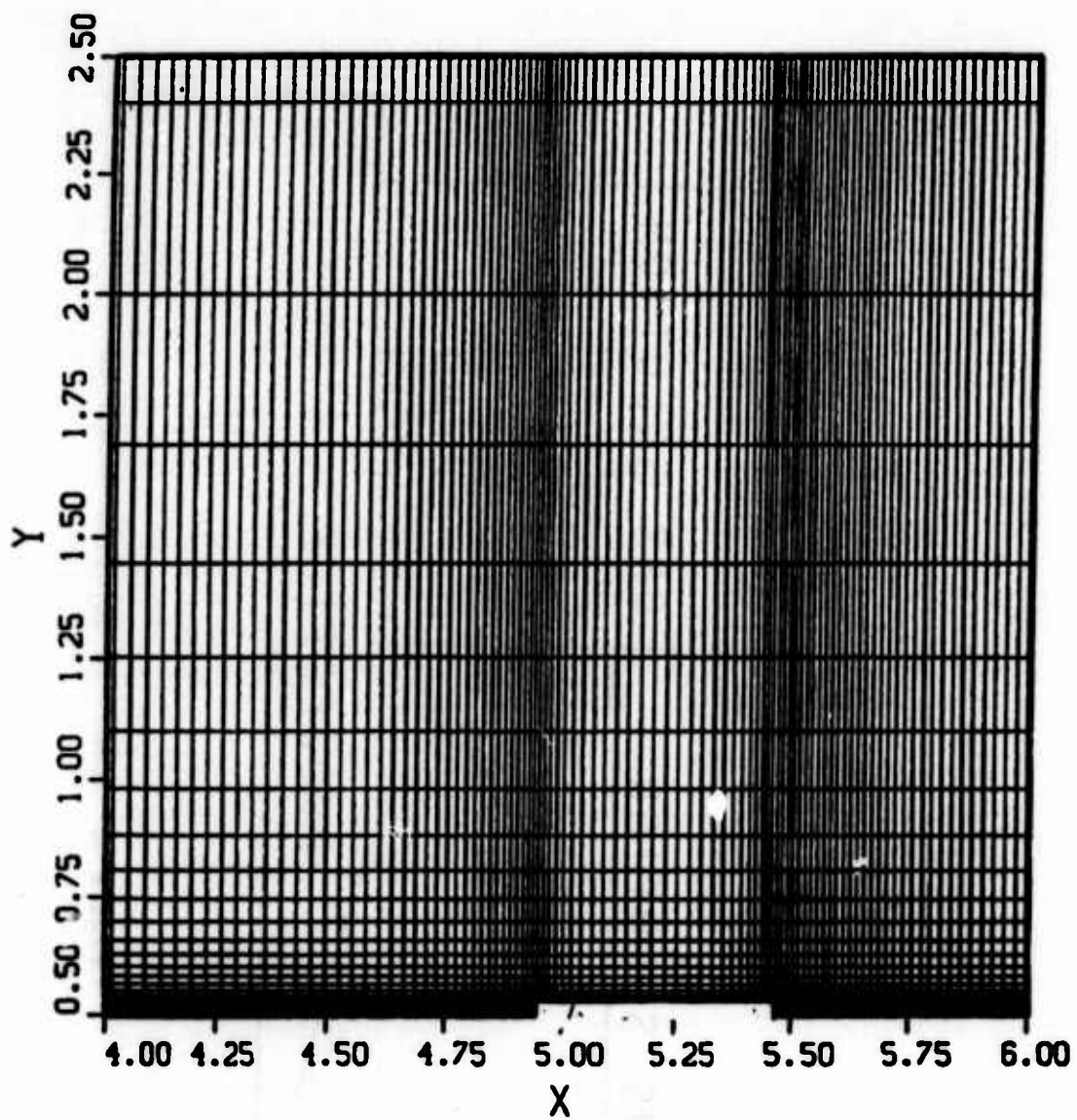
Figure 6. Computational Grid Expanded Near the Model
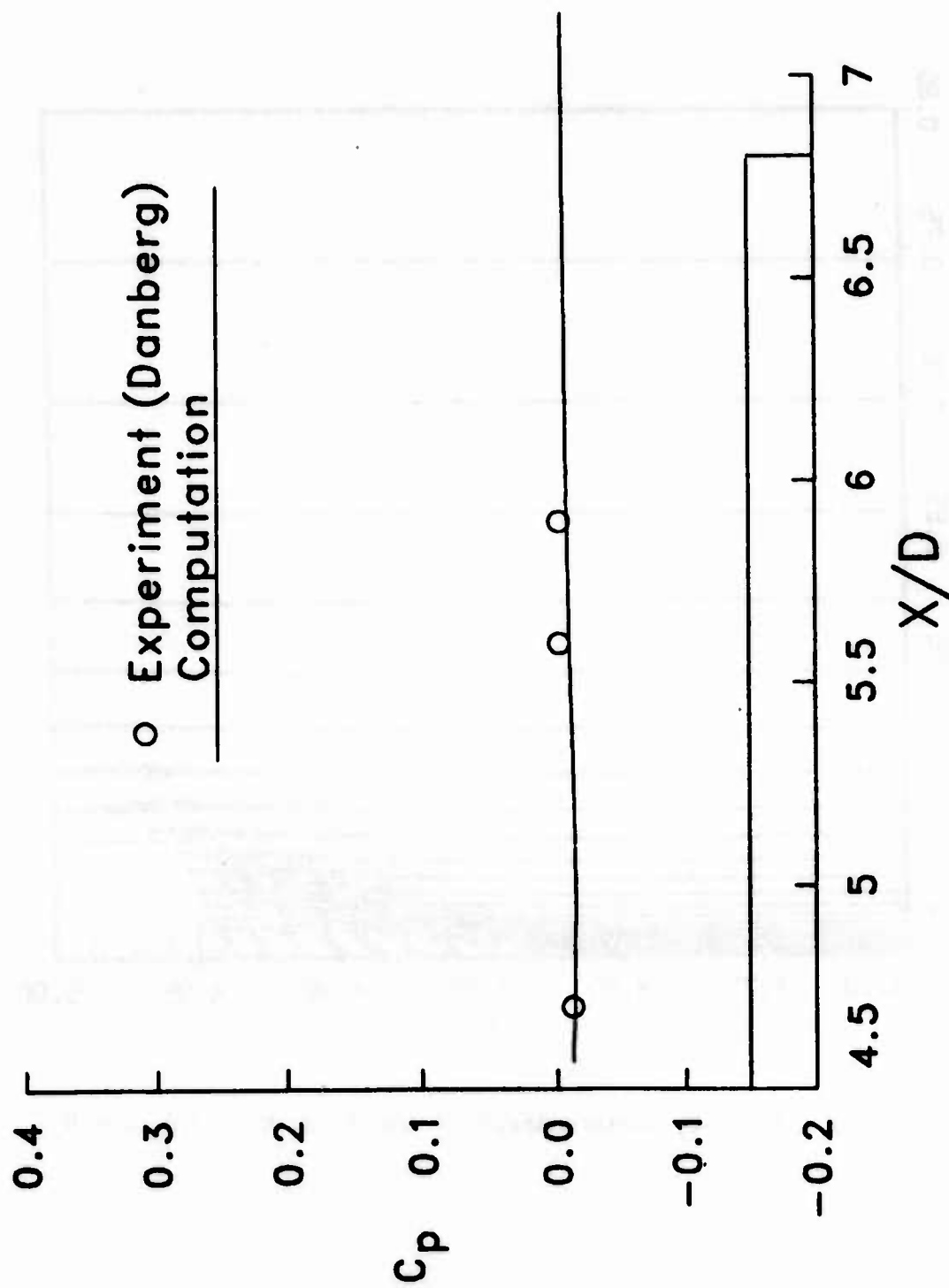
Figure 7. Longitudinal Surface Pressure Distribution, $M_\infty = 3.0$, $\alpha = 0$ (Without the Band)
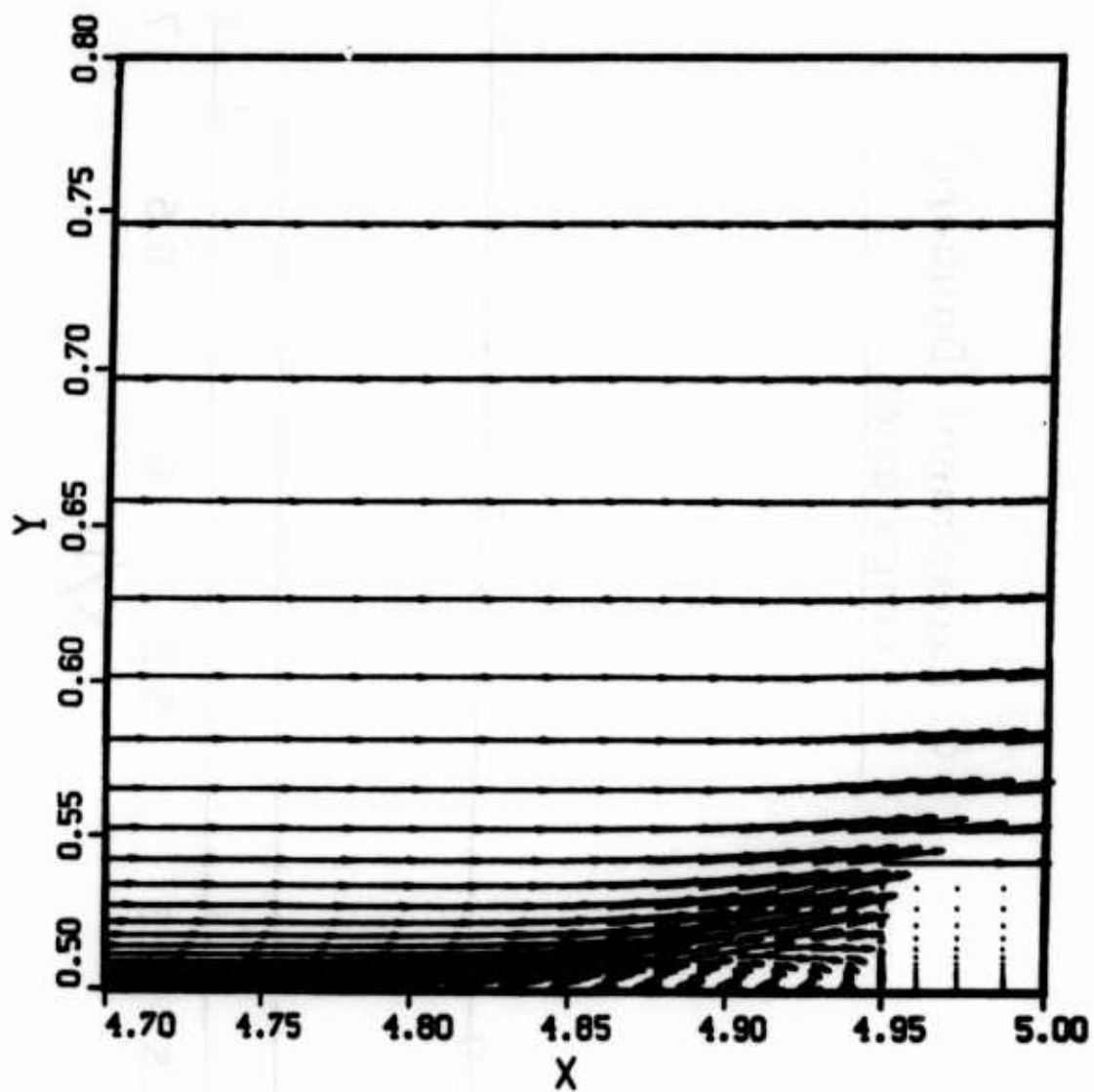
Figure 8a. Velocity Vectors Ahead of the Band, $\dot{M}_\infty$ = 3.0, α = 0
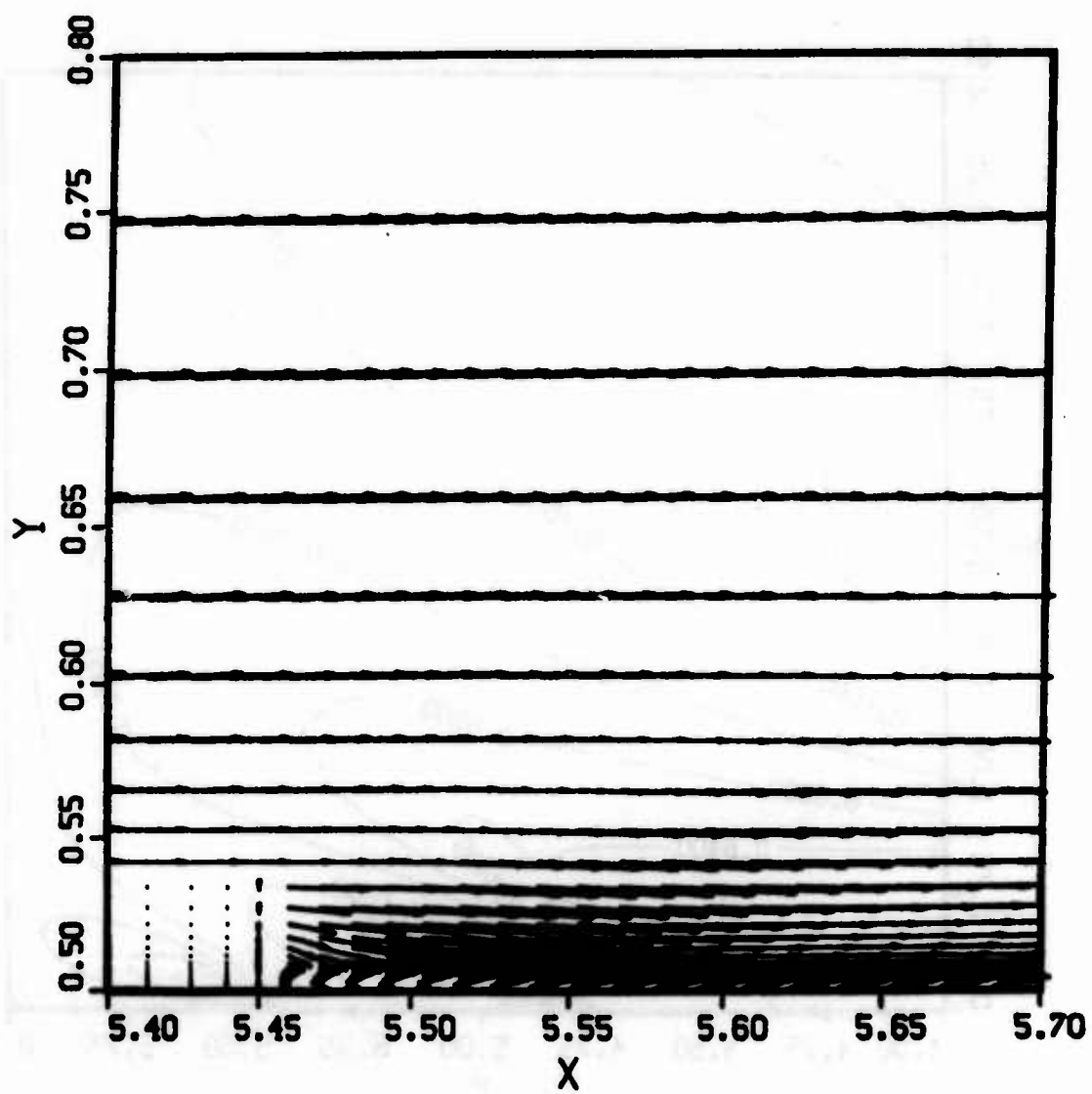
Figure 8b. Velocity Vectors Behind the Band, $M_\infty = 3.0$, $\alpha = 0$
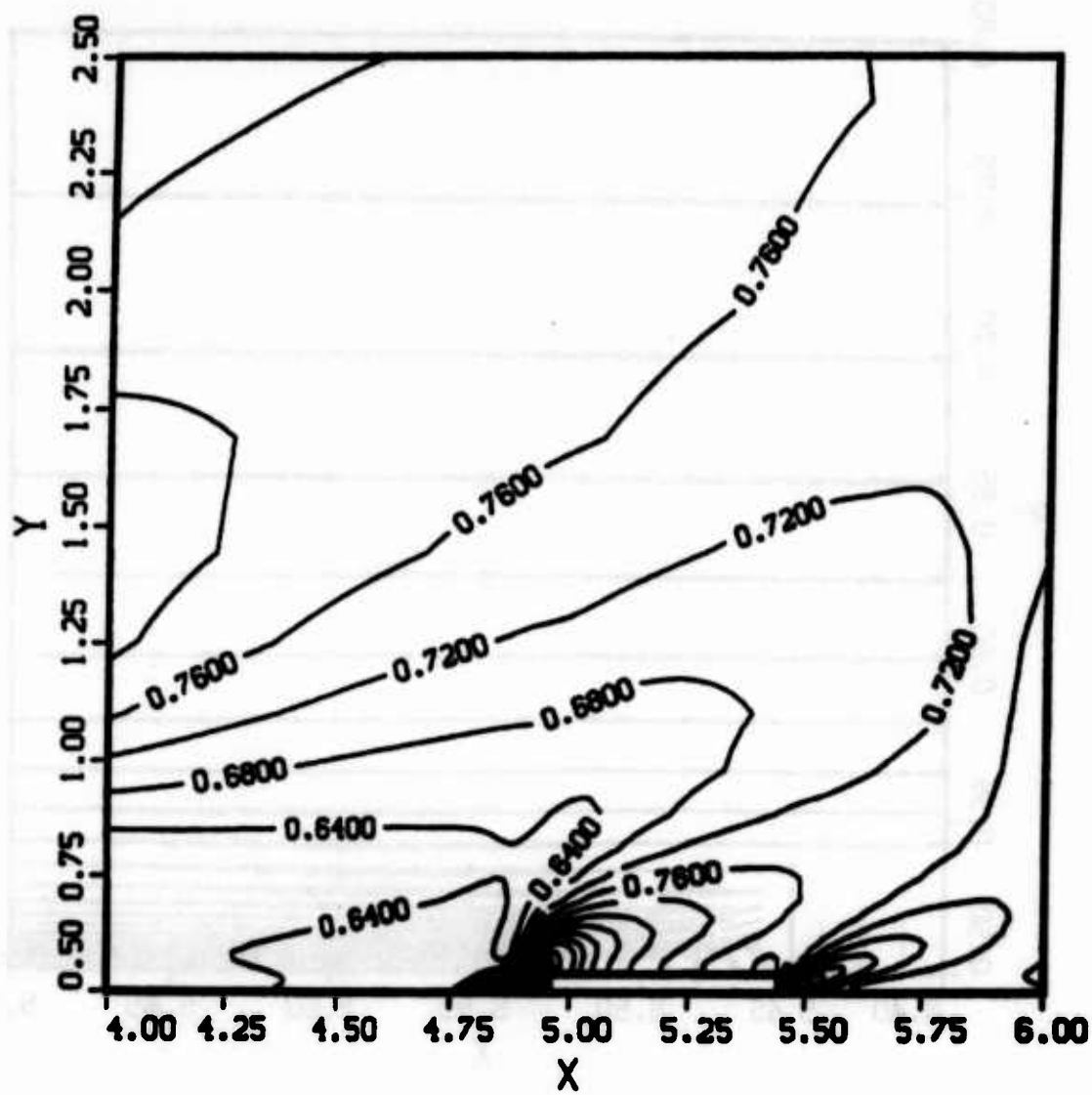
Figure 9. Pressure Contours, $M_\infty = 3.0$, $\alpha = 0$
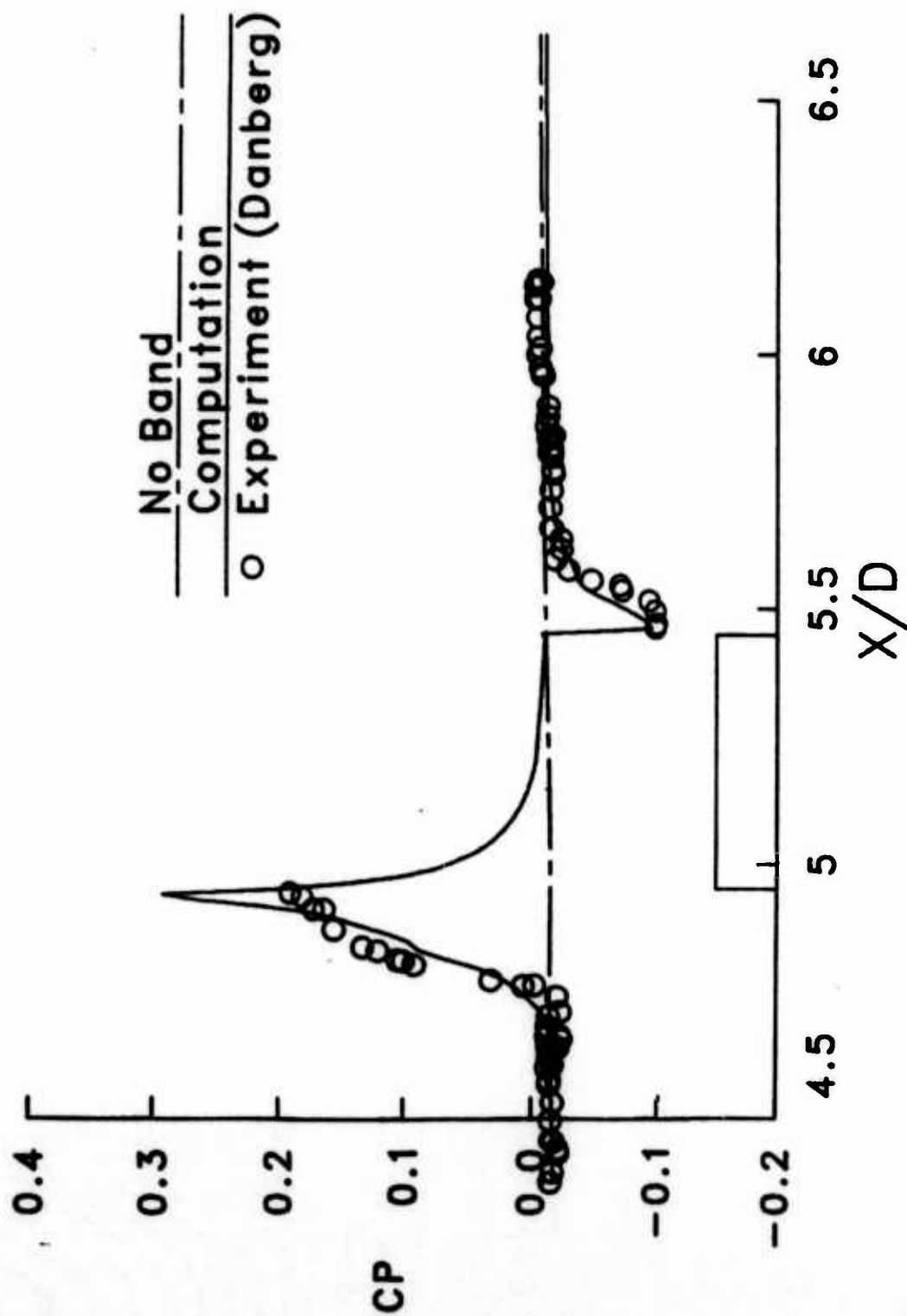
Figure 10. Longitudinal Surface Pressure Distribution, $M_\infty = 3.0$, $\alpha = 0$ (With Band)

559

IMPROVED NUMERICAL PREDICTION OF TRANSONIC FLOW

Jubaraj Sahu and Charles J. Nietubicz
Computational Aerodynamics Branch
Launch and Flight Division
US Army Ballistic Research Laboratory, LABCOM
Aberdeen Proving Ground, MD 21005-5066

ABSTRACT. The time-dependent Navier-Stokes computational technique has been in general use at the Ballistic Research Laboratory to predict transonic flows over projectiles. Recently, efforts have been made to improve the computational efficiency of this code by analyzing and including a spatially varying time step and various improved artificial dissipation models. The combined effect of these changes has led to a significant gain in the robustness and convergence characteristics for steady state applications. These techniques have been used to compute the practical problem of flow over a projectile at transonic speeds. The results confirm the improvements achieved for such calculations.

I. INTRODUCTION. In the last several years, computational aerodynamic capabilities have been developed and used to compute projectile aerodynamics at transonic speeds. These numerical capabilities[1,2] use the thin-layer Navier-Stokes computational technique and have been applied to various spinning and nonspinning projectiles at zero angle of attack. The time-dependent set of thin-layer Navier-Stokes equations are used and the solutions are marched in time until a steady state result is achieved. Since the primary interest is in the final steady state result, it is desirable to achieve the converged solution as quickly as possible which depends on the computational algorithm and also on the computational architecture used.

The time-dependent Navier-Stokes codes can be vectorized to run on a vector processor on the Cray-XMP. A vectorized version of the code can run approximately 2-3 times faster than the original unvectorized code. Gain in computational efficiency can also be achieved due to improvements made in the computational algorithm. Use of a spatially varying time step, improved numerical dissipation models and implicit treatment of the boundary condition procedure are some of the techniques that can and have been used[3,4] to improve the overall efficiency of the computational technique. The combined effect of these changes have provided significant gain in the robustness and convergence characteristics for steady state applications which are of primary interest to us. The purpose of this paper is to incorporate a spatially varying time stepping procedure and improved artificial dissipation models to a BRL time-dependent Navier-Stokes code[1] for steady state applications in transonic projectile aerodynamics.

The resulting solver has been used to compute transonic flow over a secant-ogive-cylinder-boattail projectile at M_∞ = .98 and α = 0. Computed results confirm the improvements achieved for such calculations. A brief description of the governing equations and the computational technique is first given. The algorithm improvements used for transonic viscous flow simulation are then described.

II. GOVERNING EQUATIONS AND COMPUTATIONAL TECHNIQUE.

The Azimuthal Invariant (or Generalized Axisymmetric) thin-layer Navier-Stokes equations for general spatial coordinates ξ, η, ζ can be written as:[1]

$$\partial_\tau \hat{q} + \partial_\xi \hat{E} + \partial_\zeta \hat{G} + \hat{H} = Re^{-1}\partial_\zeta \hat{S} \tag{1}$$

where
$\xi = \xi(x,y,z,t)$ is the longitudinal coordinate

$\eta = \eta(y,z,t)$ is the circumferential coordinate

$\zeta = \zeta(x,y,z,t)$ is the near normal coordinate

$\tau = t$ is the time

and

$$\hat{q} = J^{-1}\begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ e \end{bmatrix}, \quad \hat{E} = {}^{-1}\begin{bmatrix} \rho U \\ \rho uU + \xi_x p \\ \rho vU + \xi_y p \\ \rho wU + \xi_z p \\ (e+p)U - \xi_t p \end{bmatrix}, \quad G = J^{-1}\begin{bmatrix} \rho W \\ \rho uW + \xi_x p \\ \rho vW + \xi_y p \\ \rho wW + \xi_z p \\ (e+p)W - \xi_t p \end{bmatrix},$$

$$H = J^{-1}\begin{bmatrix} 0 \\ 0 \\ \rho V[R_\xi(U-\xi_t) + R_\zeta(W-\zeta_t)] \\ -\rho VR(V-\eta_t) - p/R) \\ 0 \end{bmatrix}$$

$$\hat{S} = \begin{bmatrix} 0 \\ \mu(\zeta_x^2 + \zeta_y^2 + \zeta_z^2)u_\zeta + (\mu/3)(\zeta_x u_\zeta + \zeta_y v_\zeta + \zeta_z w_\zeta)\zeta_x \\ \mu(\zeta_x^2 + \zeta_y^2 + \zeta_x^2)v_\zeta + (\mu/3)(\zeta_x u_\zeta + \zeta_y v_\zeta + \zeta_z w_\zeta)\zeta_y \\ \mu(\zeta_x^2 + \zeta_y^2 + \zeta_z^2)w_\zeta + (\mu/3)(\zeta_x u_\zeta + \zeta_y v_\zeta + \zeta_z w_\zeta)\zeta_z \\ \{(\zeta_x^2 + \zeta_y^2 + \zeta_y^2)[(\mu/2)(u^2 + v^2 + w^2)_\zeta + \kappa Pr^{-1}(\gamma-1)^{-1}(a^2)_\zeta] \\ + (\mu/3)(\zeta_x u + \zeta_y v + \zeta_z w)(\zeta_x y_\zeta + \zeta_y v_\zeta + \zeta_z w_\zeta)\} \end{bmatrix}$$

562

The velocities

$$U = \xi_t + \xi_x u + \xi_y v + \xi_z w$$

$$V = \eta_t + \eta_x u + \eta_y v + \eta_z w \qquad (2)$$

$$W = \zeta_t + \zeta_x u + \zeta_y v + \zeta_z w$$

represent the contravariant velocity components.

The Cartesian velocity components (u, v, w) are nondimensionalized with respect to a_∞ (free stream speed of sound). The density (ρ) is referenced to ρ_∞ and total energy (e) to $\rho_\infty a_\infty^2$. The local pressure is determined using the equation of state,

$$P = (\gamma - 1)[e - 0.5\rho(u^2 + V^2 + w^2)] \qquad (3)$$

where γ is the ratio of specific heats.

In equation (1), axisymmetric flow assumptions have been made which result in the source term, $\hat{H}$. The details of how this is obtained can be found in Reference 1 and are not discussed here. Equation (1) contains only two spatial derivatives. However, it retains all three momentum equations and allows a degree of generality over the standard axisymmetric equations. In particular, the circumferential velocity is not assumed to be zero thus allowing computations for spinning projectiles to be accomplished.

The numerical algorithm used is the Beam-Warming fully implicit, approximately factored finite difference scheme. The algorithm can be first or second order accurate in time and second or fourth order accurate in space. Since the interest is only in the steady-state solution, Equation (1) is solved in an asymptotic fashion and first order accurate time differencing is used. The spatial accuracy is fourth order. Details of the algorithm are included in References 5-7.

To suppress high frequency components that appear in regions containing severe pressure gradients, e.g., shocks or stagnation points, artificial dissipation terms are added. Different dissipation models have been used and are described in the next section. The best results were obtained with a switching dissipation model which is a blend of second and fourth order dissipation terms. This switching model is similiar to the model used by Pulliam[3] and uses a fourth order dissipation in smooth regions and switches to a second order dissipation in regions containing high pressure or density gradients. Incorporation of this dissipation model has improved the quality of the results and has made the code more robust.

III. DISSIPATION MODELS.

A. ORIGINAL DISSIPATION MODEL. The implicit approximately factored algorithm developed by Beam-Warming[7] has the form

$$[I + h\delta_\xi \hat{A}^n + \overset{(2)}{D}_\xi] [I + h\delta_\zeta \hat{C}^n - hRe^{-1} \delta_\zeta J^{-1} \hat{M}^n J + \overset{(2)}{D}_\zeta]\Delta\hat{q}^n$$

$$(4)$$

$$= -h[\delta_\xi \hat{E}^n + \delta_\zeta \hat{G}^n - Re^{-1} \delta_\zeta \hat{S}^n + \hat{H}^n + \overset{(4)}{D}]$$

where the explicit fourth-order dissipation is

$$\overset{(4)}{D} = -\varepsilon_e \Delta t J^{-1}[(\nabla_\xi \Delta_\xi)^2 + (\nabla_\zeta \Delta_\zeta)^2] J \hat{q}^n$$

and the implicit second-order dissipation terms are

$$\overset{(2)}{D}_\xi = -\varepsilon_e \Delta t J^{-1}(\nabla_\xi \Delta_\xi) J$$

$$\overset{(2)}{D}_\zeta = -\varepsilon_e \Delta t J^{-1}(\nabla_\zeta \Delta_\zeta) J .$$

The fourth-order explicit dissipation is used to control non-linear instabilities whereas the implicit dissipation is included to stabilize the explicitly treated fourth-difference terms. The parameter ε_e is $O(1)$ and the parameter $\varepsilon_i = (2-3) \varepsilon_e$. If ε_e is large enough, in most cases stability of the scheme can be maintained. However, increased explicit smoothing makes the solution less accurate and in many cases cannot eliminate the oscillations observed near the shock waves. An example of the type of oscillations which can be found with the central finite difference solution is shown in Figure 1. This figure shows a converged solution[3] for an airfoil at $M_\infty = 0.8$ and $\alpha = 0$. The numerical oscillations in the vicinity of the shocks ($X/C = .4$ and $.6$) are apparent.

B. SWITCHING DISSIPATION MODEL. One way to eliminate the oscillations near shocks is to use a second order numerical dissipation locally near the shocks and fourth order dissipation elsewhere. This idea has been used by Jameson et al[8] and forms the basis of this switching dissipation model. As an example, a second-order dissipation was used[3] in the region of shock wave for the flow problem of Figure 1 and the solution is shown in Figure 2. The oscillations at the shock are eliminated and a smooth solution is obtained.

As discussed in Reference 3, one can look at the upwind schemes as a guidance to how much dissipation is required. The upwind schemes have inherent dissipation and there is no need to add to numerical dissipation

564

explicitly. In Reference 3, it has been shown that the upwind flux split scheme of Steger and Warming[9] is equivalent to using a central finite difference plus some form of dissipation. For example, if the upwind scheme is second-order, it can be written as a central difference plus an added fourth-order dissipation

$$\frac{\partial E}{\partial \xi} = \frac{E_{j+1} - E_{j-1}}{2\Delta\xi} + \frac{1}{4\Delta\xi} (\Delta_\xi \nabla_\xi)^2 |A| q \tag{5}$$

where Δ_ξ and ∇_ξ are one-sided forward and backward finite-difference operators ($\Delta_\xi u_j = u_{j+1} - u_j$ and $\nabla_\xi u_j = u_j - u_{j-1}$). If first-order differences are applied in the upwind scheme, then we get a central difference plus a second-order dissipative term,

$$\frac{\partial E}{\partial \xi} = \frac{E_{j+1} - E_{j-1}}{2\Delta\xi} - \frac{1}{2\Delta\xi} (\Delta_\xi \nabla_\xi) |A| q \tag{6}$$

Generally, the best approach for an upwind scheme is to use a first-order difference at shocks and second-order elsewhere. As shown above, this is equivalent to using a second-order dissipation near the shock and fourth-order dissipation elsewhere for a central finite difference algorithm such as the one used in our unsteady codes.

To mimmick the flux split upwind difference scheme, a second-order dissipation term is added to right hand side of Equation (4) which is given as:

$$- \frac{\Delta t}{2\Delta\xi} J^{-1} \rho(A)(\Delta_\xi \nabla_\xi) J q .$$

With the fourth-order dissipation term included, the full dissipation can be written as:

$$\frac{\Delta t}{J} ||A_\infty|| [\delta \, \epsilon_d \, |\frac{\Delta\nabla\rho}{\langle\rho\rangle}| \delta \, J \, q - \delta \, \epsilon_e \, \delta\Delta\nabla \, J \, q] \tag{7}$$

where the first term is the second-order dissipation and the second term contains the fourth-order dissipation. The coefficients ϵ_d and ϵ_e are the associated coefficients for the second-order and fourth-order dissipation, respectively. Note that the fourth-order dissipation is non-linear, in that, the coefficient is not a constant and is scaled by spectral radius $||A_\infty||$. The two terms in Equation (7) are of the form $\delta\alpha\delta\beta$ where:

$$(\delta \, \alpha \, \delta \, \beta)_j = (\frac{\alpha_{j+1} + \alpha_j}{2}) (\beta_{j+1} - \beta_j) + (\frac{\alpha_j + \alpha_{j-1}}{2}) (\beta_j - \beta_{j-1})$$

and
$$\alpha = \varepsilon_d \left| \frac{\Delta \nabla \rho}{\langle \rho \rangle} \right| \quad \text{or} \quad \varepsilon_e$$

$$\beta = J\ q \quad \text{or} \quad J\Delta\nabla q$$

For automatic switching from fourth-order dissipation to second-order dissipation near shocks etc., we introduce a scaling,

$$I_j = \begin{cases} 1 & \text{if } \varepsilon_e > \varepsilon_d \left| \frac{\nabla\Delta\rho}{\langle\rho\rangle} \right| & \text{(Fourth-Order)} \\[2mm] 0 & \text{if } \varepsilon_e < \varepsilon_d \left| \frac{\nabla\Delta\rho}{\langle\rho\rangle} \right| & \text{(Second-Order) .} \end{cases} \tag{8}$$

With this switching built-in, the numerical dissipation term in the streamwise direction, for example, can be written as:

$$\frac{\Delta t}{J} ||A_\infty|| [.5(g_{j+1} + g_j)(\tilde{q}_{j+1} - \tilde{q}_j)(1 - \frac{I_{j+1} + I_j}{2}) - .5(g_j + g_{j-1})$$

$$(\tilde{q}_j - \tilde{q}_{j-1})(1 - \frac{I_j + I_{j-1}}{2}) - \varepsilon_e(\frac{I_{j+1} + I_j}{2})(\delta^2 \tilde{q}_{j+1} - \delta^2 \tilde{q}_j) \tag{9}$$

$$+ \varepsilon_e(\frac{I_j + I_{j-1}}{2})(\bar{\delta}^2 \tilde{q}_j - \bar{\delta}^2 \tilde{q}_{j-1})]$$

where $q = J\ q$, $\quad g_j = \varepsilon_d \left| \frac{\nabla\Delta\rho}{\langle\rho\rangle} \right|_j$, $\quad \bar{\delta}^2 = \nabla\Delta$

$$||A_\infty|| = \max [(|\xi_x| + |\xi_z|)(1 + M_\infty), 6.0]$$

The dissipation term in the normal direction is similarly added. Here, the pressure gradient is used in the second-order dissipation term as opposed to the density gradient used in the longitudinal direction.

IV. SPACE VARYING Δt. For projectile aerodynamics, the interest is generally in obtaining a final steady state result; therefore, we can use time step sequences or spatially variable time steps to accelerate convergence. For a fixed Δt, the Courant number is not uniform since the grid spacings vary from very fine to very coarse in the flow field region of interest. The use of a space varying Δt can thus, be interpreted as an attempt to use a more uniform Courant number throughout the field.

566

For an aerodynamic simulation where the grid is highly stretched, we can use a purely geometric variation of Δt given as[3]:

$$\Delta t = \frac{(\Delta t)_{ref}}{1 + \sqrt{J}}$$ (10)

where J is the Jacobian of transformation. The time step, h in Equation (4) is then replaced by Δt given in Equation (10).

V. RESULT. The model used in the computations consists of a three caliber secant-ogive nose, a two caliber cylinder and a one caliber 7° boattail (See Figure 3). Surface pressure measurements have been made by Kayser et al[10] for this projectile and are compared with the present and past computed results. For computational efficiency, the base flow is not included and the boattail is extended as a sting.

The computational grid used for the numerical computations was obtained using a modified version of a hyperbolic grid generator.[11] The full grid is shown in Figure 4 and consists of 128 longitudinal points and 56 radial points. The computational domain extends to about 3.5 body lengths in front, in radial direction and behind the projectile. An expanded view of the grid near the projectile is shown in Figure 5. The grid points are clustered near the ogive-cylinder and cylinder-boattail junctions in the longitudinal direction. In the normal direction, the grid points are clustered near the body surface with a minimum spacing of .00002 D and are stretched out to the far field.

All the computations were made at M_∞ = .98 and α = 0. The free stream Reynolds number based on the total length is 4.56 × 10[6]. For turbulent flow computations, an algebraic turbulence model by Baldwin and Lomax[11] is used. Computations are started from initial freestream conditions and are marched in time to obtain the steady state solution. Figures 6 and 7 show the pressure contours and Mach contours, respectively for a converged solution obtained with the switching dissipation model. These figures show the qualitative features of the flow such as the expansions at the cylinder and boattail corners as well as the location of the shock wave that exist on the projectile.

The next set of figures show the surface pressure coefficient as a function of axial position. The experimental result is indicated by circles whereas the lines represent the computed results. Figure 8 shows the time history of the solution at various time iterations using the old fourth order dissipation model. The expansions at the cylinder and boattail corners are clearly seen in the results and are quickly set in about 800 iterations. The convergence is slowest near the shock wave on the cylindrical portion of the projectile and takes a large number of time iterations for the solution to converge. Figure 9 shows the result obtained with the new switching dissipation model and varying time step procedure at 400 iterations. Although, this result has not converged, the solution agrees fairly well with the previous converged solution from Figure 8. The final converged solution, shown in Figure 10, is obtained after 600 times iterations with the improved version of the code. This result is in excellent agreement with the experimental data.

The smoothing coefficients ε_e and ε_d used in the switching dissipation model are .01 and 1.0, respectively for this result.

The effect of these smoothing parameters on the numerical solution was investigated. First, the fourth order smoothing coefficient, ε_d was changed. The result is shown in Figure 11 and the effect of varying ε_d was to change the solution very minimally near the shock wave ($X/D \simeq 4.5$). The overall accuracy of the solution is fairly good. Second, fourth order coefficient, ε_e was also varied, while the ratio of the two smoothing parameters was kept fixed. Again, the results do not show any significant change in the pressure distribution.

<u>VI. CONCLUDING REMARKS.</u> The original time-dependent Azimuthal-Invariant Navier-Stokes code has been modified by including switching dissipation model and variable time stepping procedure. This improved version of the code was used to compute the flow over a projectile at $M_\infty = .98$ and $\alpha = 0$.

Significant improvements in the convergence characteristics have been obtained with the improved version of the code for steady state applications. The total CPU time has been reduced by a factor of three to obtain the converged result. In addition, the code is now more robust and is being used presently for other numerical calculations.

568

REFERENCES

1. Nietubicz, C.J., Pulliam, T.H. and Steger, J.L., "Numerical Solution of the Azimuthal-Invariant Navier-Stokes Equations, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, ARBRL-TR-02227, March 1980. (AD A085716) (Also see AIAA Journal, Vol. 18, No. 12, December 1980, pp. 1411-1412)

2. Nietubicz, C.J., "Navier-Stokes Computations for Conventional and Hollow Projectile Shapes at Transonic Velocities," US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, ARBRL-MR-03184, July 1982. (AD A116866)

3. Pulliam, T.H., "Artificial Dissipation Models for the Euler Equations," AIAA Paper No. 85-0438, January 1985.

4. Pulliam, T.H. and Steger, J.L., "Recent Improvements in Efficiency, Accuracy and Convergence of Implicit Approximate Factorization Algorithms," AIAA Paper No. 85-0360, January 1985.

5. Steger, J.L., "Implicit Finite Difference Simulation of Flow About Arbitrary Geometries with Application to Airfoils," AIAA Journal, Vol. 16, No. 4, July 1978, pp. 679-686.

6. Pulliam, T.H. and Steger, J.L. "On Implicit Finite-Difference Simulations of Three-Dimensional Flow," AIAA Journal, Vol. 18, No. 2, February 1980, pp. I59-167.

7. Beam, R. and Warming R.F., "An Implicit Factored Scheme for the Compressible Navier-Stokes Equations," AIAA Paper No. 77-645, June 1977.

8. Jameson, A. et al, "Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time - Stepping Schemes," AIAA Paper No. 81-1259.

9. Steger, J.L. and Warming, R.F., "Flux Vector Splitting of the Inviscid Gas Dynamics Equations with Applications to Finite-Difference Methods," Journal of Computational Physics, 40, 1981, pp. 263-293.

10. Kayser, L.D. and Whiton, F., "Surface Pressure Measurements on a Boattailed Projectile Shape at Transonic Speeds," US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland, ARBRL-MR-03161, 1982. (AD A113520)

11. Baldwin, B.S. and Lomax, H., "Thin-Layer Approximation and Algebraic Model for Separated Turbulent Flows," AIAA Paper No. 78-257, 1978.
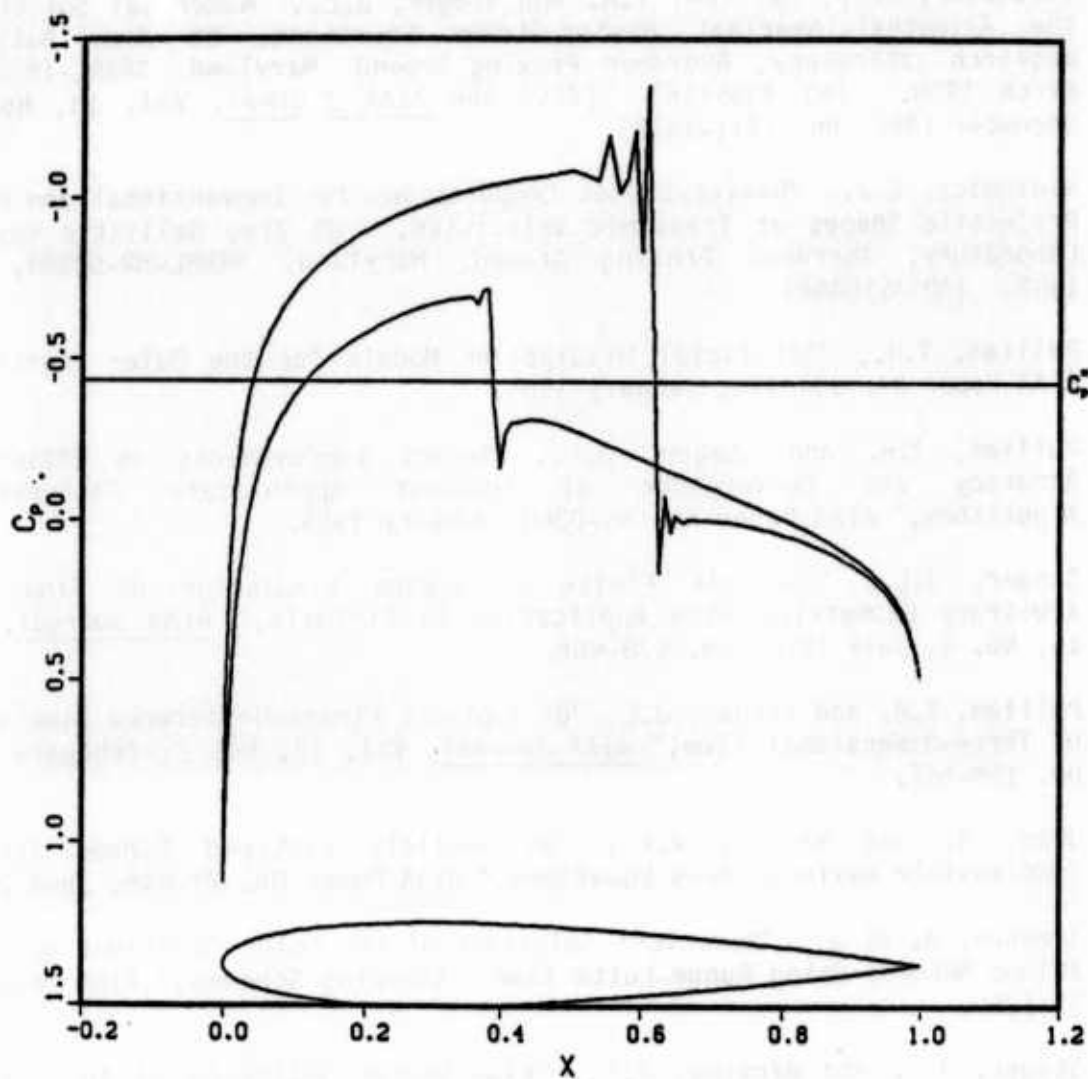
Figure 1. Pressure Coefficient Showing Central Difference
Oscillations at Shock (Reference 3)
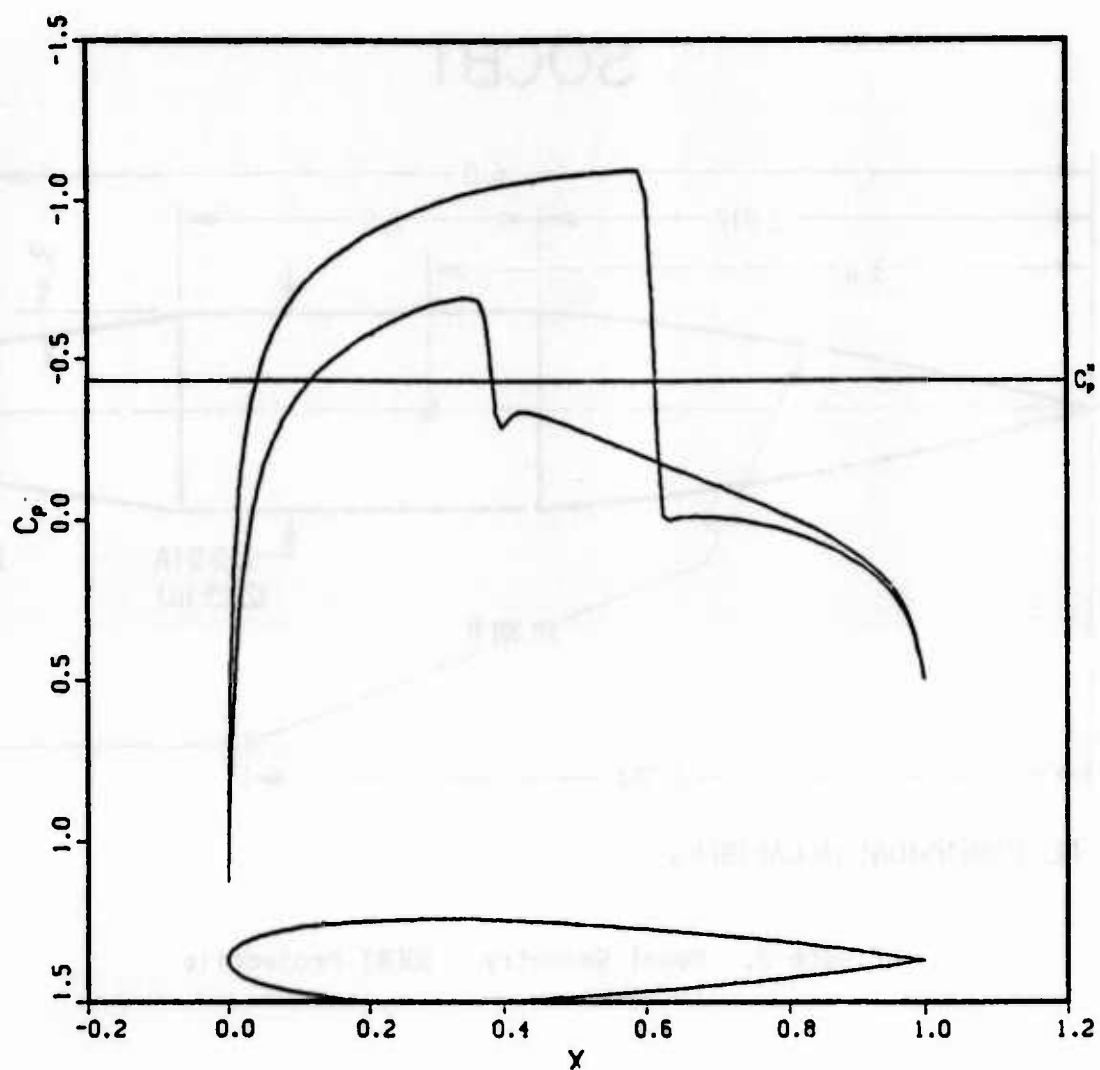
Figure 2. Pressure Coefficient with Second Order Dissipation
Near Shock (Reference 3)

SOCBT
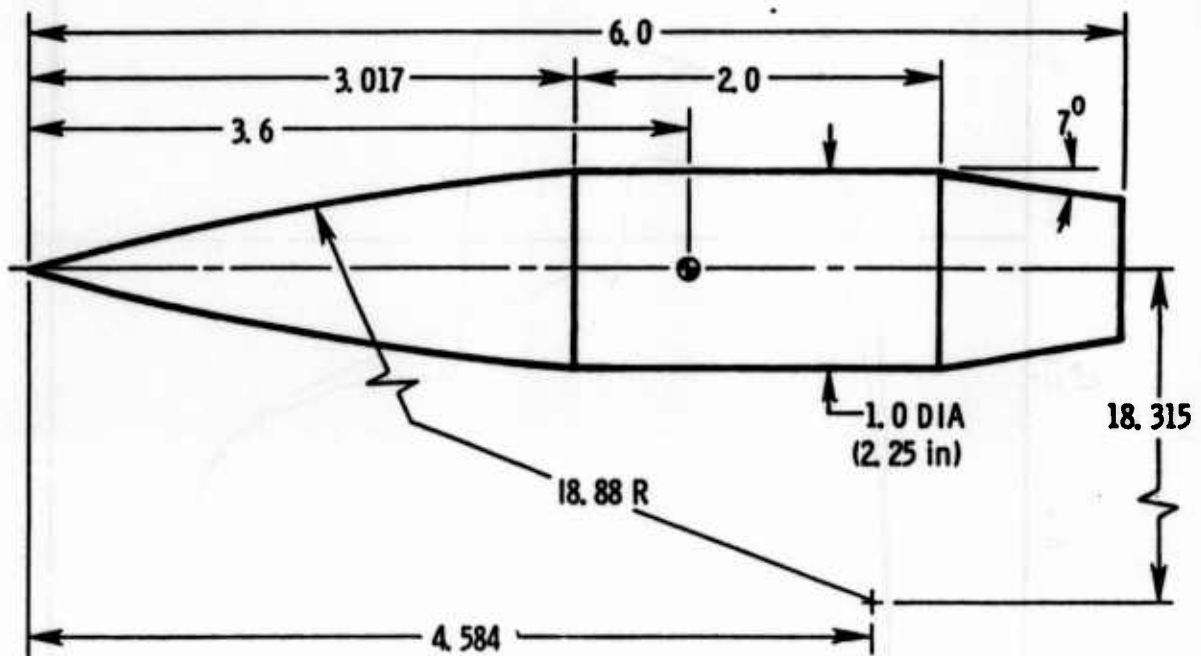
ALL DIMENSIONS IN CALIBERS

Figure 3.   Model Geometry - SOCBT Projectile

572
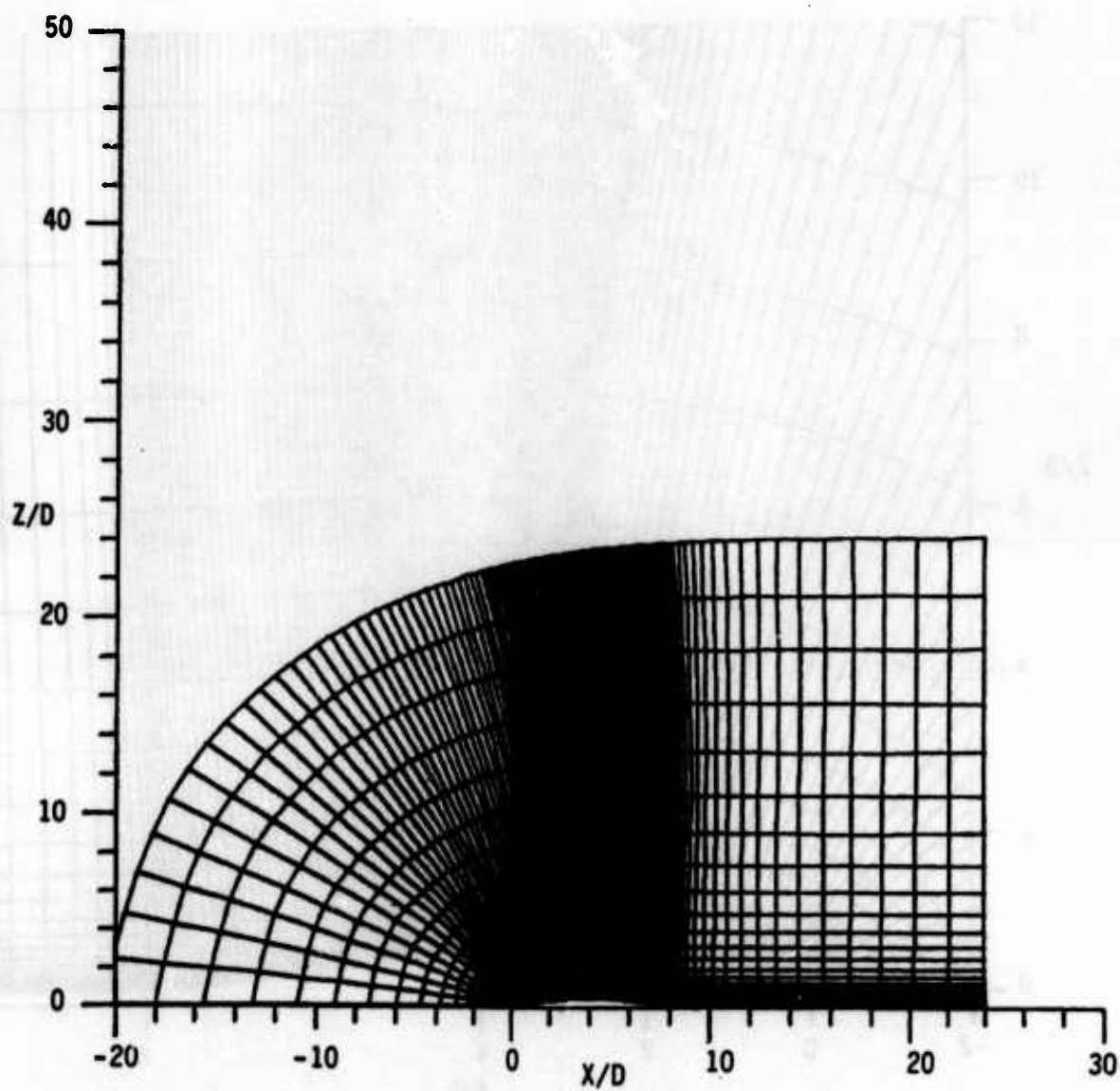
Figure 4. Full Computational Grid
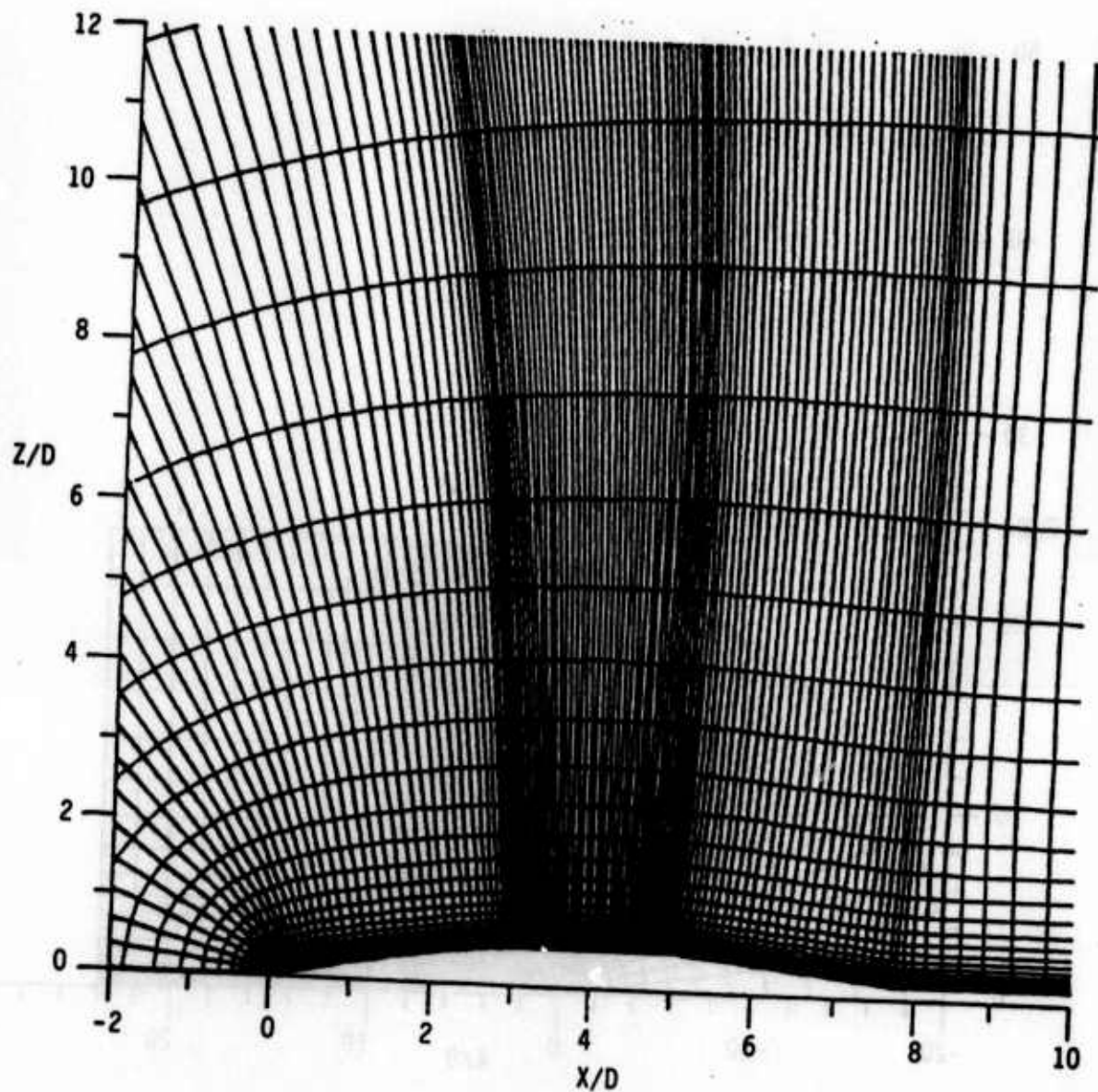
Figure 5. Expanded View of the Grid Near the Projectile
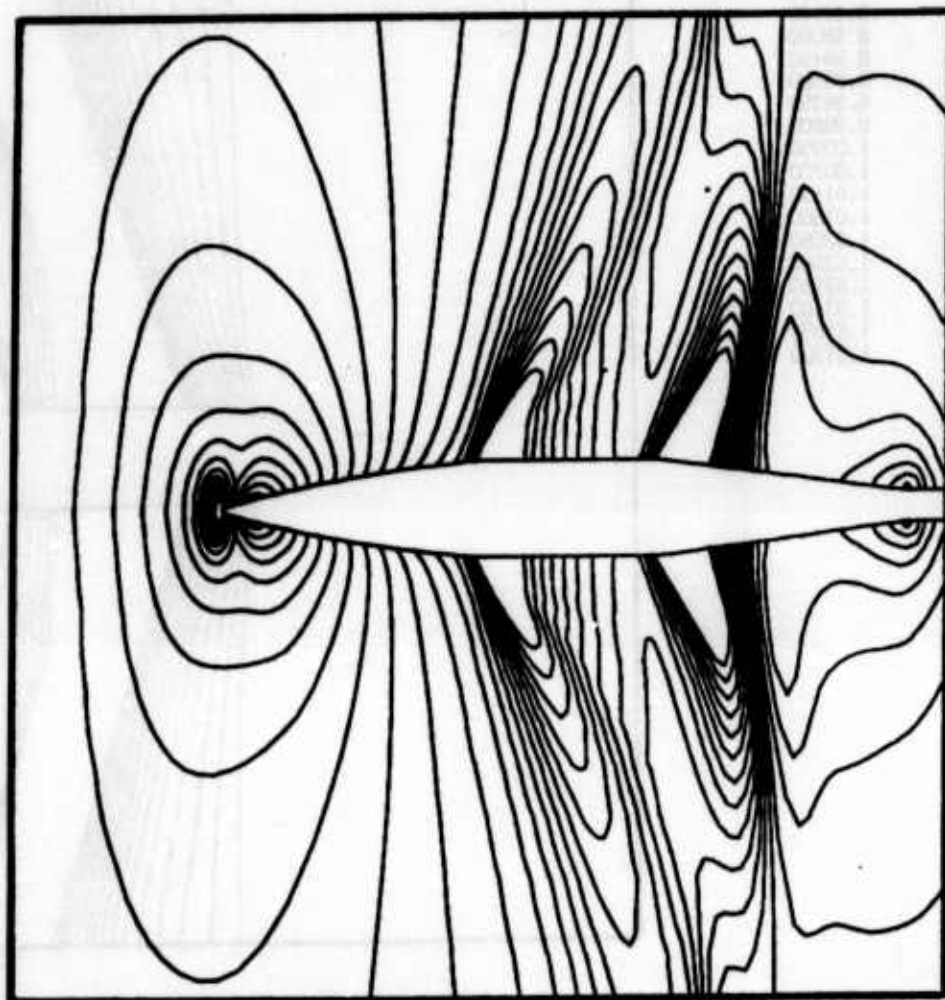
Figure 6. Pressure Contours, M_∞ = .98, α = 0
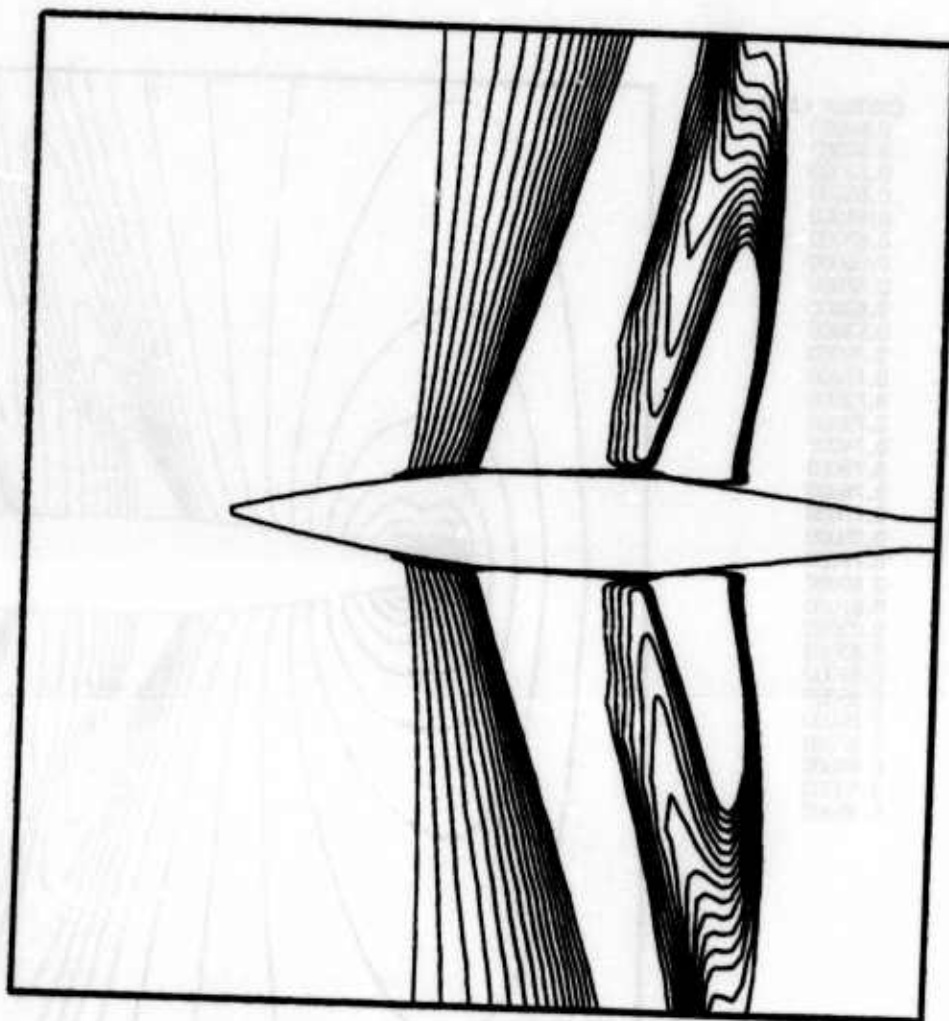
Figure 7. Mach Contours, $M_\infty = .98$, $\alpha = 0$
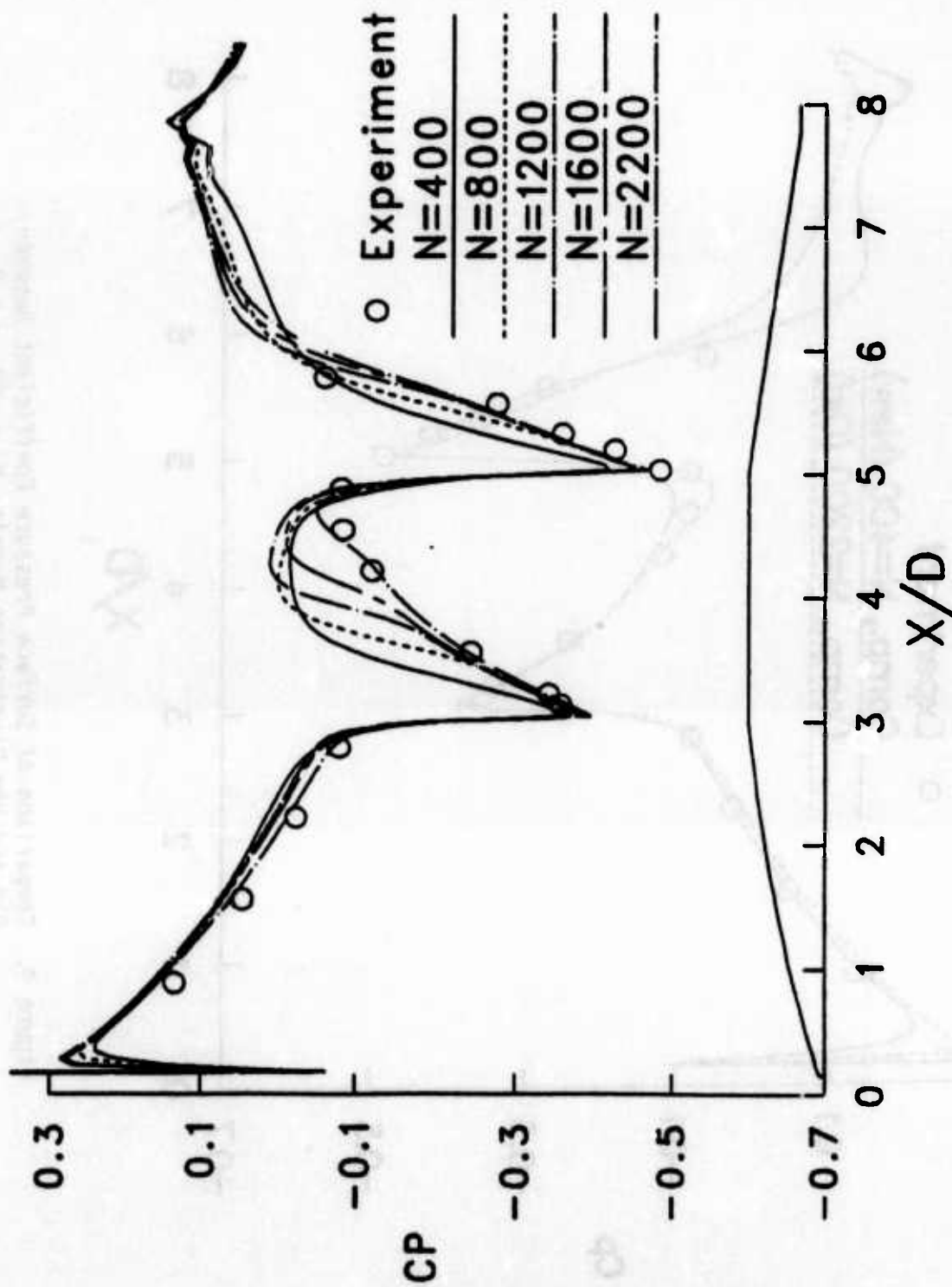
Figure 8. Longitudinal Surface Pressure Distribution,
$M_\infty = .98$, $\alpha = 0$ (Old Dissipation Model)

577
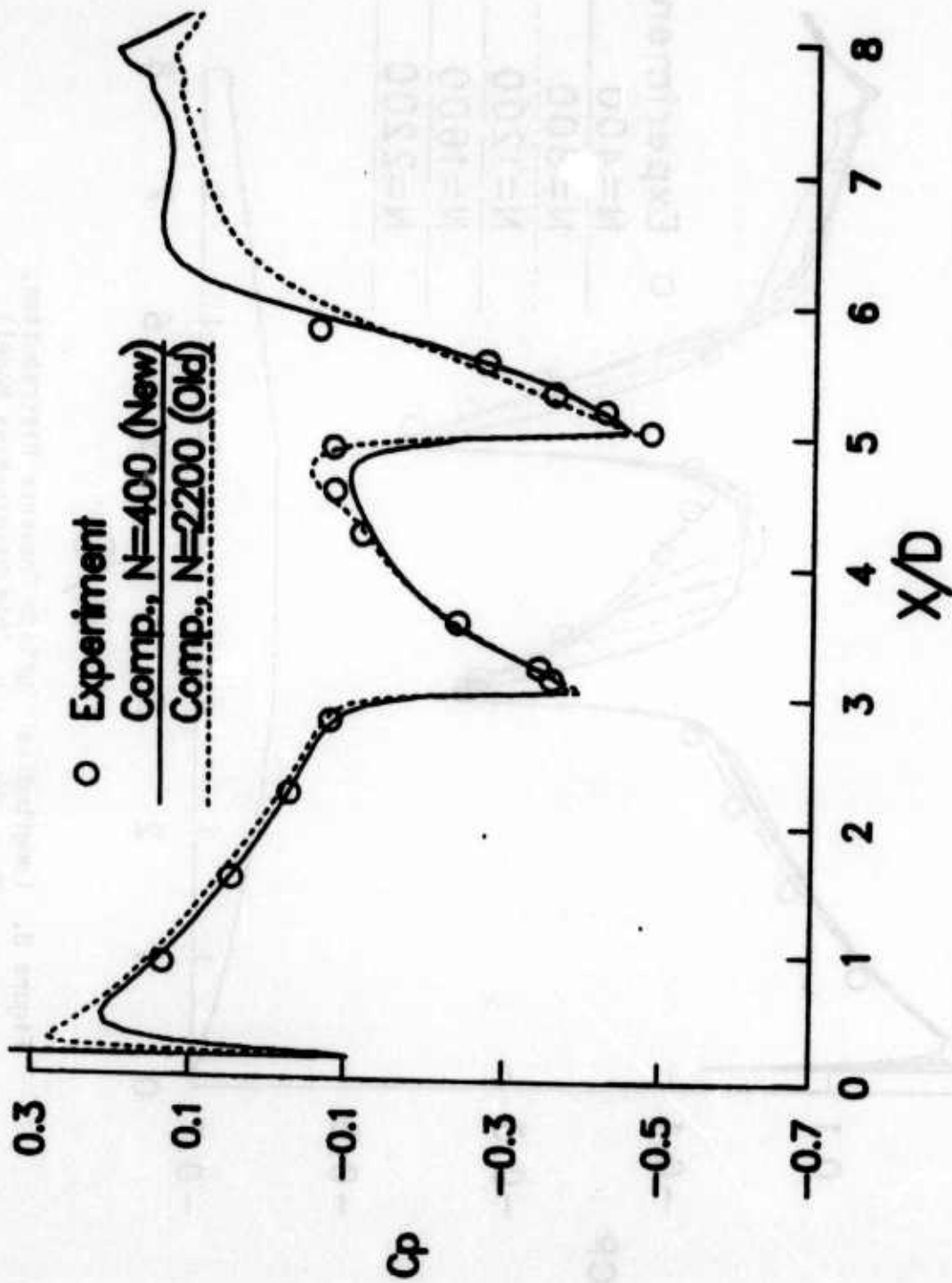
Figure 9. Comparison of Surface Pressure Coefficient Between Old and New Dissipation Models, $M_\infty = .98$, $\alpha = 0$
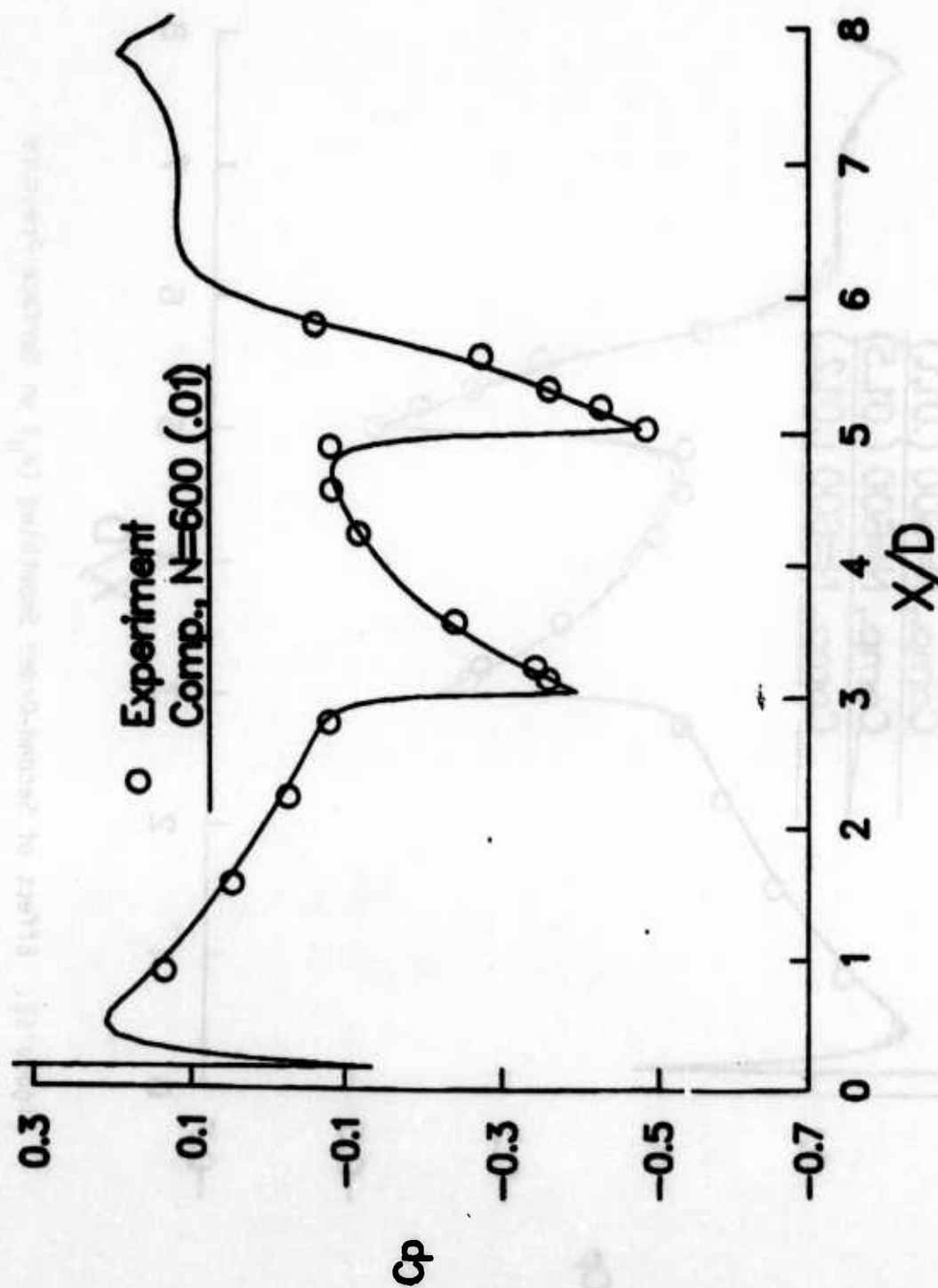
Figure 10. Longitudinal Surface Pressure Distribution, $M_\infty = .98$, $\alpha = 0$ (New Switching Dissipation Model)

579
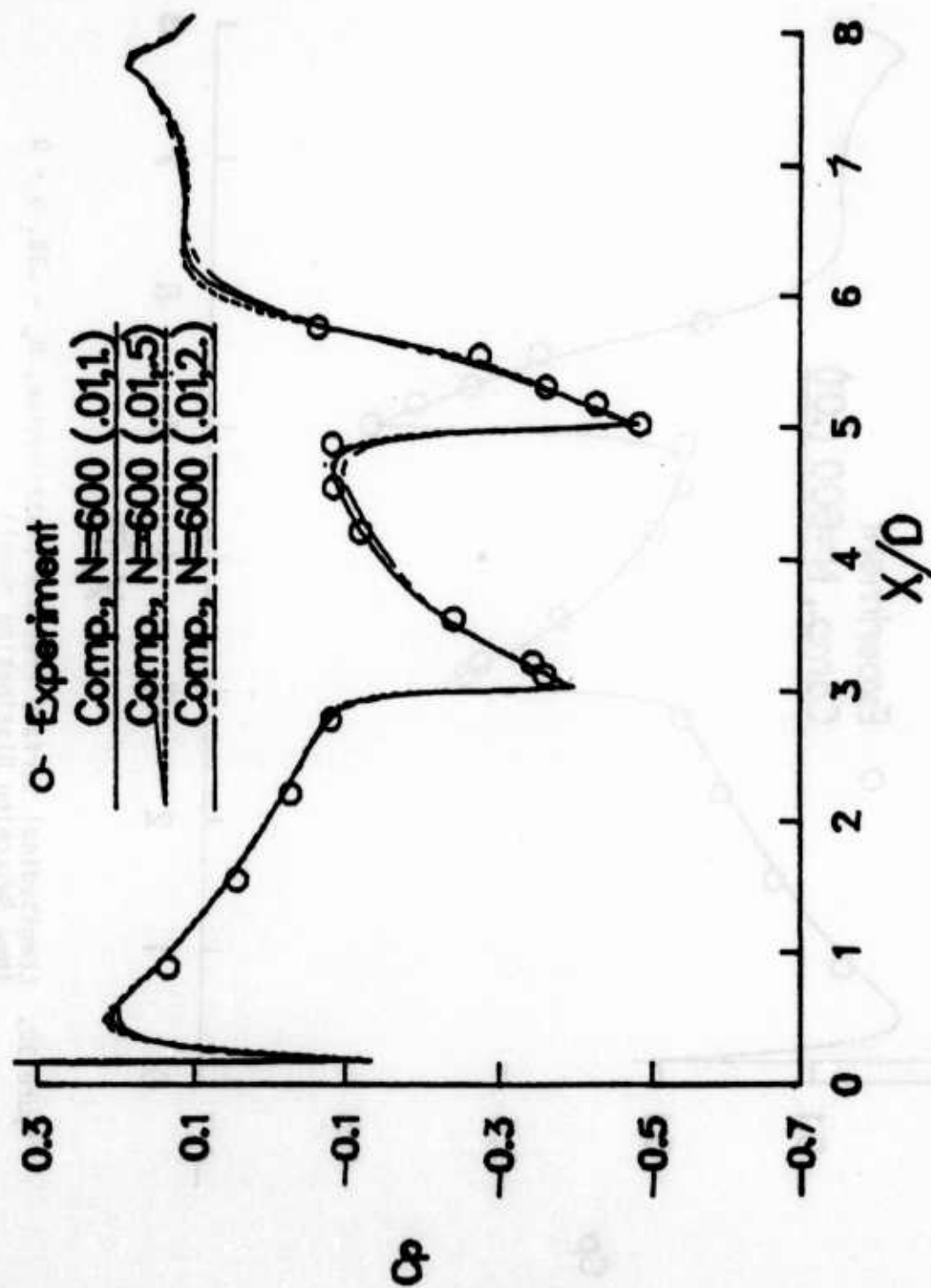
Figure 11. Effect of Second-Order Smoothing (ϵ_d) on Surface Pressure Distribution, $M_\infty = .98$, $\alpha = 0$
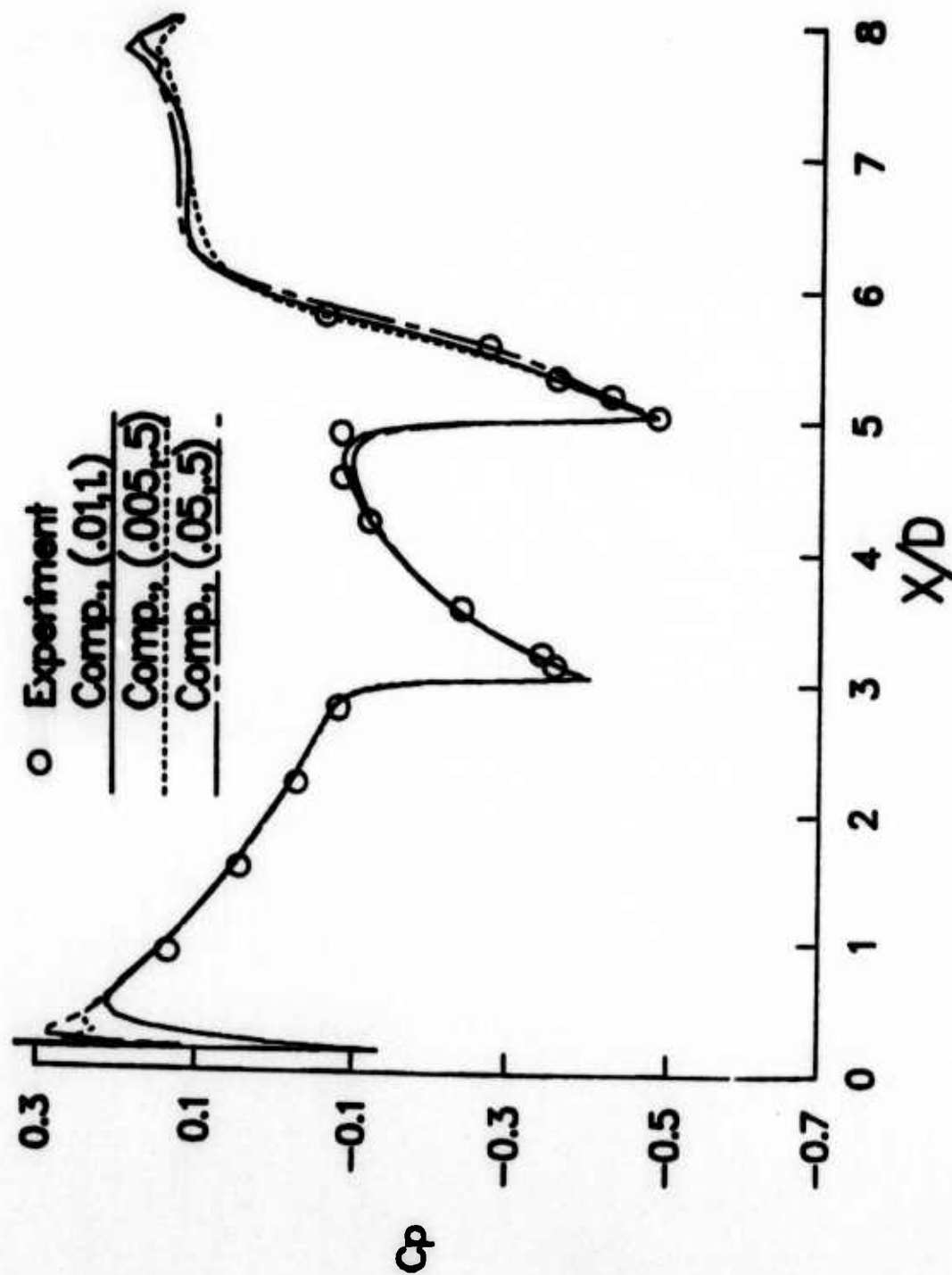
Figure 12. Effect of Fourth-Order Smoothing (ϵ_e) on Surface Pressure Distribution, $M_\infty = .98$, $\alpha = 0$

NUMERICAL SOLUTION OF SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS

RICHARD E. EWING
DEPARTMENTS OF MATHEMATICS AND PETROLEUM ENGINEERING
UNIVERSITY OF WYOMING
LARAMIE, WYOMING 82071

ABSTRACT. Complex physical phenomena involving transport of heat or fluids are often modeled by coupled systems of nonlinear partial differential equations. The recent advances in computational capabilities with the advent of new computer architectures have allowed the incorporation of more physics into the models resulting in larger, more complex mathematical models. Research must therefore be increased in each phase of the modeling process utilizing physical, mathematical, numerical, and computational concepts. Transport dominated processes are notoriously difficult to treat numerically. Techniques for treating systems of transport equations via a modified method of characteristics are presented. The flux or fluid velocity is very important to the flow directions required for the modified method of characteristics; mixed finite element techniques for obtaining accurate fluid velocities are presented. Also many physical phenomena have highly local properties which may move with time. Adaptive grid refinement methods are presented to resolve this important dynamic local behavior. Finally, the influence of the computer architecture upon the development of efficient algorithms for large scale problems is discussed.

1. INTRODUCTION. The need for the study and use of mathematics is growing and expanding extremely rapidly in response to the enormous recent development of computing capabilities. The use of complex models which incorporate more detailed physics has necessitated more sophisticated mathematics in the modeling process. In this way a broader range of mathematics is needed for applications. In this paper, we shall discuss certain aspects of the expanding scope of mathematical modeling.

The mathematical techniques which are used to model multicomponent or multiphase flow problems are representative of those needed for many other applications such as chemically reacting or thermally driven flows and will be used to illustrate the role of mathematics in modeling. The advent of orders-of-magnitude better computing capabilities has allowed the modeling of more complicated physical phenomena. We will indicate how this growth is changing the entire modeling process.

Modeling of large-scale physical processes involves four major interrelated stages. First, a physical model of the physical processes must be developed incorporating as much physics as is deemed necessary to describe the essential phenomena. A careful list of the assumptions made in establishing this physical model

should be compiled together with expected properties of the process, such as bifurcation or physical instabilities, that might be expected and should be modeled. Second, a mathematical formulation of the physical model should be obtained, usually involving coupled systems of nonlinear partial differential equations. The properties of this mathematical model, such as existence, uniqueness, and regularity of the solution are then obtained and related to the physical process to check the model. This part of the modeling process becomes exceedingly difficult with large coupled systems of nonlinear partial differential equations. Third, a discretized numerical model of the mathematical equations is produced. This numerical model must have the properties of accuracy and stability and produce solutions which represent the basic physical features as well as possible without introducing spurious phenomena associated with the specific numerical schemes. Obtaining asymptotic error estimates via mathematical analysis for the systems of equations is critical for accurate numerical simulation. Fourth, a computer program capable of efficiently performing the necessary computations for the numerical model is sought. Properties of the computer architecture to be used in the computation must be considered strongly in the development of efficient computational algorithms. Although the total modeling process encompasses aspects of each of these four intermediate stages, the process is not complete with one pass through the steps. Usually several iterations of this modeling loop are necessary to obtain reasonable models for the highly complex physical phenomena involved in many applications.

The aims of this paper are to introduce certain complex physical phenomena which need to be better understood, to illustrate aspects of the modeling process used to describe these processes, and to discuss some of the newer mathematical tools that are being utilized in the various models. The complexity of the models requires sophisticated mathematical analysis. For example, the increasing use of large, coupled systems of nonlinear partial differential equations to describe the flow of multiphase and multicomponent fluid systems is identifying very difficult problems in the theoretical aspects of the partial differential equations, the numerical analysis of various discretization schemes, the development of new, accurate numerical models, and the computational efficiency of discrete systems resulting from the discretizations. The interplay between the engineering and physics of the applications, the mathematical properties of the models and discretizations, and the role of the computer in the algorithm development is critical and will be stressed in this presentation.

The modeling of many fluid flow problems involves very similar mathematical equations. Examples of mathematical and related physical properties of these models which must be addressed include: (a) the resolution of sharp moving fronts in convection dominated convection-diffusion problems, (b) the stability and accuracy of discretization of highly non-self-adjoint differential operators, (c) the need to have very accurate fluid velocities which dominate the flow, (d) the need to model dynamic local phenomena which govern the physics, and (e) the empha-

sis on development of efficient numerical procedures for the enormous problems encountered.

A model problem which illustrates many major numerical difficulties arising in fluid flow applications is presented in Section 2. The numerical stability problems associated with this transport dominated system and the corresponding pure transport problem are discussed. A modified method of characteristics based on combining the transport and accumulation terms in the equation into a directional derivative along characteristic-like curves is then briefly described. The modified method of characteristics is heavily dependent upon having very accurate fluid velocities. Section 3 is then devoted to the description of a mixed finite element procedure which is designed to give approximations of the fluid velocities which are just as accurate as the pressure approximations, even in the context of rapidly changing fluid properties. The interaction between the computational efforts and associated error estimates is important. The need for adaptive local grid refinement methods to resolve certain dynamic, highly localized physical phenomena is described in Section 4. Important considerations such as a choice of versatile and efficient data structures and adaptivity techniques are discussed.

2. DESCRIPTION OF A MODEL PROBLEM AND THE MODIFIED METHOD OF CHARACTERISTICS.

A model system of equations describing multicomponent flow of an incompressible fluid [20,31,39] is given by

$$\nabla \cdot \mathbf{u} = -\nabla \cdot \frac{k}{\mu(c)} \nabla p = q, \qquad \mathbf{x} \in \Omega, \quad t \in J \qquad (1)$$

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot [\mathbf{u}c - D(\mathbf{u})\nabla c] = \bar{c}q, \qquad \mathbf{x} \in \Omega, \quad t \in J \qquad (2)$$

$$\mathbf{u} \cdot \mathbf{n} = [\mathbf{u}c - D(\mathbf{u})\nabla c] \cdot \mathbf{n} = 0, \qquad \mathbf{x} \in \partial\Omega, \quad t \in J \qquad (3)$$

$$c(\mathbf{x}, 0) = c_0(x), \qquad \mathbf{x} \in \Omega \qquad (4)$$

for $\Omega \in \mathbb{R}^2$ with boundary $\partial\Omega$ and $J = [0, T]$, where p and $\mathbf{u}$ are the pressure and velocity of the single phase fluid mixture, c is the concentration of one component, q, the total volumetric flow rate, is smoothly distributed over Ω, and D is assumed to be a diffusion-dispersion tensor given by [20,28,39]

$$(D_{ij}(\mathbf{x}, \mathbf{u})) = \phi d_m I + \frac{d_l}{|\mathbf{u}|} \begin{pmatrix} u_1^2 & u_1 u_2 \\ u_1 u_2 & u_2^2 \end{pmatrix} + \frac{d_t}{|\mathbf{u}|} \begin{pmatrix} u_2^2 & -u_1 u_2 \\ -u_1 u_2 & u_1^2 \end{pmatrix} \qquad (5)$$

where $\mathbf{u} = (u_1, u_2)$, $|\mathbf{u}|$ is the Euclidean norm of $\mathbf{u}$, d_m is the molecular diffusion coefficient, and d_l and d_t are the magnitudes of longitudinal and transverse dispersion. For many multiphase or multicomponent fluid flow problems, Equation (1) would be replaced by some compressible or incompressible form of the Navier-Stokes equations of flow. The given form of Equation (1) results from an

585

averaging of the Navier-Stokes equations and is applicable in the context of flow through porous media. Equation (1) will be used for simplicity of exposition in this paper. In this application, ϕ and k are media properties and μ is the fluid viscosity. Extensions of the concepts and techniques presented here to Navier-Stokes problems are possible.

Equation (2) is an example of a transport dominated convection-diffusion equation. Since diffusion is small, the solution c exhibits very sharp fronts or concentration gradients which move in time across the domain. The frontal width is very narrow in general, but must be resolved accurately via the numerical method since it describes the physics of the mixing zone and governs the speed of the frontal movement. Similar dispersive mixing zones are critical in the modeling of contaminant transport processes [1,2,25], combustion problems, and other applications with moving, internal fluid interfaces.

If the dispersion tensor in Equation (2) is ignored, Equation (2) becomes a first order hyperbolic problem instead of a transport dominated convection-diffusion equation. Standard highly accurate finite difference schemes for hyperbolic partial differential equations are known to be unstable and various upstream weighting or "artificial diffusion" techniques have been utilized to stabilize the variant of Equation (2). The upstream weighting techniques introduce artificial diffusion in the direction of the grid axes and of a size proportional to the grid spacings. Thus, although this stabilizing effect would be small if very fine grid block spacings were used, the enormous size of many applications necessitates the use of large grid blocks and hence, large, directionally-dependent artificially induced numerical diffusion which has nothing to do with the physics of the flow. Two major problems in numerical flow simulation today are due essentially to the use of standard upstream weighting techniques. First, the upstream methods, by introducing a large artificial numerical diffusion or dispersion, smear sharp fluid interfaces producing erroneous predictions of the degree of mixing and incorrect frontal velocities. Second, the numerical diffusion is generated along grid lines and produces results which may be radically different if the orientation of the grid is rotated forty-five degrees.

The use of physical intuition in determining a more accurate numerical scheme can be illustrated in this case. The physical diffusion-dispersion term displayed in Equation (5) is a rotationally-invariant tensor. Therefore, one way to stabilize the first order hyperbolic problem without introducing artificial directional effects is to use an "artificial diffusion" term of the form in Equation (5). The size of this term must then be closely considered in order not to diffuse fronts too badly. A consequence of this type of stabilization with finite difference discretization means a nine-point difference star would be necessary to approximate the cross-derivatives accurately instead of the standard five-point star used in two space dimensions. In three space dimensions a twenty-seven point star would be necessary to replace a seven point star. If iterative solution techniques are being utilized, this greatly increases the solution times. This is a good example of how

decisions made in one part of the modeling process can greatly influence other parts of the problem.

For more complex physical processes, the system of Equations (1)–(4) must be expanded to include mass balances from different components or different phases. The governing equations for combustion processes, for example, could involve coupled systems of several nonlinear partial differential equations of the form of Equation (2) depending upon the availability of oxygen, fuel, etc. The interaction between these coupled nonlinear equations can greatly affect the properties of the equations. Much work must be done to understand the mathematical properties of existence, uniqueness, and continuous dependence of solutions upon data for coupled systems of this form. Therefore, the improved computing capabilities which allow the numerical approximation of large, coupled systems of nonlinear partial differential equations, are necessitating the theoretical study of properties of systems of these equations. The "applied" mathematician involved in the simulation must understand and be able to work with these "purer" areas if the modeling process is to be effective.

The numerical analysis involved in rigorously obtaining asymptotic error estimates for even the model problem presented in Equations (1)–(5) requires various aspects of functional analysis and approximation theory. The order of the approximations depends upon the use of fractional order Sobolev spaces and interpolation spaces. Although asymptotic error estimates are not particularly useful in obtaining realistic bounds for errors, they are very important in determining which techniques work better than others and why. The analyses involved in obtaining these estimates has greatly influenced our choice of numerical schemes. Similarly the analysis can help determine where special locations which yield superconvergence results for the methods can be found. These superconvergence results are especially important for coupled systems of partial differential equations since the locations can often be utilized efficiently in quadrature rules to describe more accurately the coupling between the unknown variables. Asymptotic error estimates for the model problem appear in [28,30,31,32].

In [19,37], Douglas and Russell described a technique based on a method of characteristics approach for treating the first order hyperbolic part of Equation (2). This technique, based on a form of Equation (2) which is analogous to a convection-diffusion equation, was implemented by Russell [37,38] and forms the basis for a particular time-stepping scheme which we have used effectively.

In order to introduce a nondivergence form of Equation (2) that is used in our numerical methods, we first expand the convection ($\nabla \cdot \mathbf{u}c$) term with the product rule and use Equation (1) to obtain

$$\phi\frac{\partial c}{\partial t} + \mathbf{u} \cdot \nabla c - \nabla \cdot [D(\mathbf{u})\nabla c] = (\bar{c} - c)\bar{q}, \qquad \mathbf{x} \in \Omega, \, t \in J \qquad (6)$$

where $\bar{q} = \max\{q, 0\}$. To avoid technical boundary difficulties associated with our modified method of characteristics for Equation (6), in this exposition we assume

587

that Ω is a rectangle and that the problem given by Equations (1), (6), (3), and (4) is Ω-periodic.

The basic idea is to consider the hyperbolic part of Equation (6), namely, $\phi \partial c / \partial t + \mathbf{u} \cdot \nabla c$, as a directional derivative. Accordingly, let $\mathbf{s}$ denote the unit vector in the direction of (u_1, u_2, ϕ) in $\Omega \times \mathcal{J}$, and set

$$\psi(\mathbf{x}) = \left(u_1(\mathbf{x})^2 + u_2(\mathbf{x})^2 + \phi^2 \right)^{1/2}. \tag{7}$$

Then Equation (6) can be rewritten in the form

$$\psi \frac{\partial c}{\partial \mathbf{s}} - \nabla \cdot (D \nabla c) + \bar{q} c = \bar{q}\, \bar{c}. \tag{8}$$

Note that the spatial operator in Equation (8) is now self-adjoint, symmetric matrices will result from spatial discretization, and the associated numerical methods will be better behaved. Since iterative solution techniques are used to solve the nonlinear equations resulting from finite element discretization of Equation (8), and since symmetry is very important in any of the useful conjugate gradient iterative solvers, this change to symmetric matrices is very important.

One critical aspect of the modified method of characteristics is the need for accurate approximation of the directional derivative $\partial c / \partial \mathbf{s}$. Many methods based upon characteristics fix a grid at time t^{n-1} and try to determine where these points would move under the action of the characteristics. These "moving point" or "front tracking" methods must then discretize Equation (6) and solve for the unknowns c^n on a mesh of irregular or unpredictable nature. If too large a time-step is chosen, serious difficulties can arise from the spatial and temporal behavior of the characteristics. Front-tracking in two space dimensions is difficult while in three dimensions, it is considerably more difficult. For details of the discretization of $\partial c / \partial \mathbf{s}$ and the ideas for extending this method to higher space dimensions, see [29,30].

3. MIXED FINITE ELEMENTS FOR PRESSURE AND VELOCITY.

Since both the modified method of characteristics and the diffusion-dispersion term in Equation (6) are governed by the fluid velocity, accurate simulation requires an accurate approximation of the velocity $\mathbf{u}$. The coefficients k and μ in Equation (1) can change rapidly in space. In this case, in order for the flow to remain relatively smooth, the pressure changes extremely rapidly. Thus standard procedures of solving Equation (1) as an elliptic partial differential equation for pressure, differentiating or differencing the result to approximate the pressure gradient, and then multiplying by the rapidly changing function k/μ can produce very poor approximations to the velocity $\mathbf{u}$. In this section a mixed finite element method for approximating $\mathbf{u}$ and p simultaneously, via a coupled system of first order partial differential equations, will be discussed. This formulation accurately treats the problem of rapidly changing flow properties.

The coupled system of first order equations used to define our methods arise from Darcy's Law and conservation of mass

$$\mathbf{u} = -\frac{k}{\mu}\nabla p, \qquad \mathbf{x} \in \Omega, \qquad (9)$$

$$\nabla \cdot \mathbf{u} = q, \qquad \mathbf{x} \in \Omega, \qquad (10)$$

subject to the boundary condition

$$\mathbf{u} \cdot \mathbf{n} = 0, \qquad \mathbf{x} \in \partial\Omega. \qquad (11)$$

Clearly Equations (9)–(11) will determine p only to within an additive constant. Thus a normalizing constraint such as $\int_\Omega p(\mathbf{x})\, dx = 0$ or $p(\mathbf{x}_s) = 0$ for some $\mathbf{x}_s \in \Omega$ is required in the computation to prevent a singular system.

We next define certain function spaces and notation. Let $W = L^2(\Omega)$ be the set of all functions on Ω whose square is finite integrable. Let $H(\text{div}; \Omega)$ be the set of vector functions $\mathbf{v} \in [L^2(\Omega)]^2$ such that such $\nabla \cdot \mathbf{v} \in L^2(\Omega)$ and let

$$V = H(\text{div}; \Omega) \cap \{\mathbf{v} \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega\}. \qquad (12)$$

Let $(v, w) = \int_\Omega vw\, dx$, $\langle v, w \rangle = \int_{\partial\Omega} wv\, ds$, and $\|v\|^2 = (v, v)$ be the standard L^2 inner products and norm on Ω and $\partial\Omega$. We obtain the weak solution form of Equations (9)–(11) by dividing each side of Equation (9) by k/μ, multiplying by a test function $\mathbf{v} \in V$, and integrating the result to obtain

$$\left(\frac{\mu}{k}\mathbf{u}, \mathbf{v}\right) = (p, \nabla \mathbf{v}), \qquad \mathbf{v} \in V. \qquad (13)$$

The right-hand side of Equation (13) was obtained by further integration by parts and use of Equation (12). Next, multiplying Equation (10) by $w \in W$ and integrating the result, we complete our weak formulation, obtaining

$$(\nabla \cdot \mathbf{u}, w) = (q, w) \qquad w \in W. \qquad (14)$$

For a sequence of mesh parameters $h > 0$, we choose finite dimensional subspaces V_h and W_h with $V_h \subset V$ and $W_h \subset W$ and seek a solution pair $(\mathbf{U}_h; P_h) \in V_h \times W_h$ satisfying

$$\left(\frac{\mu}{k}\mathbf{U}_h, \mathbf{v}_h\right) - (P_h, \text{div } \mathbf{v}_h) = 0, \qquad \mathbf{v}_h \in V_h, \qquad (15)$$

$$(\text{div } \mathbf{U}_h, w_h) = (q, w_h), \qquad w_h \in W_h. \qquad (16)$$

We can now complete the description of our mixed finite element methods with a discussion of particular choices of V_h and W_h. Examples of these spaces are

presented in [33]. For problems with smooth coefficients and smooth forcing functions, standard approximation theory results show that, by using higher order basis functions, correspondingly higher order convergence rates can be obtained [17,18].

Special choices of basis functions for the Raviart-Thomas spaces [35] based upon Gauss-point nodal functions and related quadrature rules have significantly aided in the computational efficiency of these methods. For detailed descriptions of these bases and computational results, see [14,26,32]. The observed convergence rates matched those predicted in [17,18]. Also superconvergence results were obtained at specific locations which can be utilized in quadrature and reduced quadrature considerations in the coupled systems described in Section 2.

Since the set of equations (9)-(11) will only determine the pressure to within an arbitrary constant, the algebraic system arising from our mixed method system (23)-(25) is not definite unless constants are modded out of the approximating space W_h for pressures. If the unknowns for the x and y components of the velocity are formally eliminated from the resulting system, one can obtain a set of equations for the pressure variable. The matrix arising in this problem is quite complex, but is comparable to a matrix generated by finite difference methods for the pressure [39]. Preconditioned conjugate gradient iterative procedures have been developed to efficiently solve this set of linear equations [26,41].

Techniques for coupling the mixed finite element procedures with a modified method of characteristics for the concentration in Equations (2)-(6) have appeared in the literature [23,29]. Asymptotic error estimates and convergence rates for this coupled procedure also appeared in [30].

4. ADAPTIVE LOCAL GRID REFINEMENT.

Many of the chemical and physical phenomena which govern chemically reacting or thermally driven flow processes have extremely important local properties. Thus the models used in computer codes for these problems must be capable of resolving these critical local features. Also, in order to be useful in large-scale dynamic codes, these models must be self-adaptive and extremely efficient. The development of adaptive grid refinement techniques must take into account the rapid development of new, advanced computer architectures. The compatibility of adaptive mesh modification algorithms with the intended computer is a critical consideration in the algorithm development.

The flexibility to dynamically change the number of grid points and thus the number of unknowns in a problem can create difficulties in the linearization and linear solution algorithms. In particular, it is extremely difficult to vectorize codes with changing numbers of unknowns for efficient solution on vector machine architectures. The ability to have truly local refinement and derefinement capabilities necessitates the use of a fairly complex data structure. A data structure with these properties has been developed and was described in the literature [15,21-24]. It is a multilinked list which utilizes various properties of the tree structures

590

presented in [36] and [7,8]. The structure allows efficient linear solution algorithms via tree-traveling techniques. Although these algorithms are extremely difficult to vectorize effectively, the tree structure lends itself well to parallelism at many levels. Development of codes for efficient use of MIMD (Multiple Instruction Multiple Data Stream) architectures to parallelize the local grid refinement algorithms based upon a multiple linked-list data structure are under way [16]. Preliminary experience with the Denelcor HEP, an MIMD machine, in this context has been educational. Research in parallelization of these techniques will be continued on a new Hypercube, obtained through DoD funding.

For truly general local refinement a complex data structure like those discussed above and associated complications to the code are necessary. If local refinement is only needed in a very few special points, a technique termed patch refinement may be an attractive alternative. These concepts do not require as complex a data structure but do involve ideas of passing information from one uniform grid to another. Berger and Oliger have been using patch refinement techniques for hyperbolic problems using finite difference discretizations for some time [9,10].

The idea of a local-patch refinement method is to pick a patch that includes most of the critical behavior around a region with important local properties and do a much finer, uniform grid refinement within this patch. Given a uniform fine grid, very fast solvers can be applied locally in this region using boundary data from the coarse original grid. McCormick, Thomas and co-workers have used multigrid techniques to solve the fine-grid problem in a simple elliptic model [27]. They have addressed the communication problem with the coarse grid and have attained conservation of mass on their "composite grid." Extensions of their technique, termed FCOM (fast composite mesh method), to more difficult problems are planned.

Bramble, Pasciak and Schatz [10-12] have developed some efficient gridding and preconditioning techniques which can also be used in the local-patch refinement framework. Their methods have logically rectangular grids within the patch which can be solved very rapidly via FFT preconditioners. The important problem is the communication between grids. Recent work by Bramble, Pasciak, Schatz and Ewing [13] uses preconditioning for local grid refinement in a way to make implementation of the methods in existing large scale simulators an efficient process. These techniques, based upon finite element preconditioners, could help produce a major advancement in incorporating fixed local refinement methods in a wide variety of applications and existing codes. We also feel that these methods are sufficiently powerful to handle local time-stepping applications as well.

The adaptivity of the local refinement methods must be driven either by a type of "activity index," which relays rapid changes in solution properties, or by some estimate of the errors present in different spatial locations which need to be reduced. Recently, locally-computable a posteriori error estimators have been developed by Babuška and Rheinboldt [3-5], Bank [6], Weiser [40], and Oden [34].

Under suitable assumptions, these error estimators converge to the norm of the actual error as the mesh size tends to zero. These a posteriori error estimators are extremely important for problems involving elliptic partial differential equations in determining the reliability of estimates for a fixed grid and a fixed error tolerance in a given norm. The error estimators are used to successively refine locally until the errors in a specified norm are, in some sense, equilibrated. Although these methods are very effective for elliptic problems, they are not efficient for large time-dependent problems where an "optimal" mesh at each time step is not "optimal" for the entire time-dependent problem.

For hyperbolic or transport dominated parabolic partial differential equations, sharp fronts move along characteristic or near-characteristic directions. Therefore the computed velocity determines both the local speed and direction of the regions where local refinement will be needed at the next time steps. This information should be utilized to help move the local refinement with the front. Although patch refinement techniques based upon characteristic-direction adaptation strategies do not determine a "locally optimal" grid, the waste in using more grid than necessary is compensated for by the overall efficiency. Use of a larger refined area and grid movement only after several time-steps is the technique that we are developing since efficiency is crucial in large-scale reservoir simulation.

Variable coefficients in the partial differential equations significantly complicate local refinement techniques for finite difference methods. At present, techniques for weighting the finite difference stars based upon the varying coefficient values seem to be "ad hoc" and can often cause serious errors in the flow description. Local refinement techniques with finite element methods always yield a straightforward way to evaluate and weight the coefficients and are, in general, much easier to apply. Thus the versatility of variational techniques often more than compensates for the slight addition in computational complexity of finite element methods.

REFERENCES.

[1] M. B. Allen and R.E. Ewing, "Applied mathematics in groundwater hydrology and contaminant transport," *SIAM News,* 18(1985), 3,14.

[2] M. B. Allen, R. E. Ewing and J. V. Koebbe, "Mixed finite-element methods for computing groundwater velocities," *Numerical Methods for Partial Differential Equations*, in press.

[3] I. Babuška and W. C. Rheinboldt, "A-posteriori error estimates for the finite element method," *Internat. J. Numer. Meths. Engrg.*, 12(1978), 1597–1615.

[4] I. Babuška and W. C. Rheinboldt, "Error estimates for adaptive finite element computation," *SIAM J. Numer. Anal.*, 15(1978), 736–754.

[5] I. Babuška and W. C. Rheinboldt, "Reliable error estimation and mesh adaption for the finite element method," in *Computational Methods in Nonlinear Mechanics* (J. T. Oden, ed.), North-Holland, New York, 1980.

[6] R. E. Bank, "A multi-level iterative method for nonlinear elliptic equations," in *Elliptic Problem Solvers* (M. Schultz, ed.), Academic Press, New York, 1981.

[7] R. E. Bank and A. H. Sherman, "PLTMG users' guide," *Technical Report No. 152*, University of Texas at Austin, Center for Num. Anal., 1979.

[8] R. E. Bank and A. H. Sherman, "A refinement algorithm and dynamic data structure for finite element meshes," *Technical Report No. 166*, University of Texas at Austin, Center for Num. Anal., 1980.

[9] M. J. Berger, "Data structures for adaptive mesh refinement," in *Adaptive Computational Methods for Partial Differential Equations* (I. Babuska, J. Chandra and J. E. Flaherty, eds.), SIAM, Philadelphia, 1983, 237–251.

[10] M. J. Berger and J. Oliger, "Adaptive mesh refinement for hyperbolic partial differential equations," *Man. NA-83-02*, Computer Science Dept., Stanford University, 1983.

[11] J. H. Bramble, J. E. Pasciak and A. H. Schatz, "An iterative method for elliptic problems on regions partitioned into substructures," *Math. Comput.*, in press.

[12] J. H. Bramble, J. E. Pasciak and A. H. Schatz, "The construction of preconditioners for elliptic problems by substructuring, I," *Math. Comput.*, in press.

[13] J. H. Bramble, J. E. Pasciak, A. H. Schatz and R. E. Ewing, "A preconditioning technique for the efficient solution of problems with local grid refinement," *Comp. Meth. Appl. Mech. Engrg.*, in press.

[14] B. L. Darlow, R. E. Ewing and M. F. Wheeler, "Mixed finite element methods for miscible displacement problems in porous media," SPE 10501, *Proc. Sixth SPE Symp. on Reservoir Simulation*, New Orleans, 1982, 137–145; *Soc. Pet. Engrs. J.*, 4(1984), 391–398.

[15] J. C. Diaz, R. E. Ewing, R. W. Jones, A. E. McDonald, I. M. Uhler and D. U. von Rosenberg, "Self-adaptive local grid refinement for time-dependent, two-dimensional simulation," in *Finite Elements in Fluids*, Vol. VI, Wiley, New York, 1984.

[16] J. C. Diaz and R. E. Ewing, "Potential of HEP-like MIMD architectures in self adaptive local grid refinement for accurate simulation of physical processes," *Proc. Workshop on Parallel Processing Using the HEP*, Norman, Oklahoma, 1985.

[17] J. Douglas Jr., R. E. Ewing and M. F. Wheeler, "The approximation of the pressure by a mixed method in the simulation of miscible displacement," *R.A.I.R.O. Analyse Numerique*, 17(1983), 17–34.

[18] J. Douglas Jr., R. E. Ewing and M. F. Wheeler, "Time-stepping procedures for simulation of miscible displacement using mixed methods for pressure

approximation," *R.A.I.R.O Analyse Numerique*, **17**(1983), 249–265.

[19] J. Douglas Jr. and T. F. Russell, "Numerical methods for convection dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures," *SIAM J. Numer. Anal.*, **19**(1982), 871–885.

[20] R. E. Ewing, "Problems arising in the modeling of processes for hydrocarbon recovery," in *Research Frontiers in Applied Mathematics*, Vol. 1 (R. E. Ewing, ed.), SIAM, Philadelphia, 1983.

[21] R. E. Ewing, ed., "Oil Reservoir Simulation," *Comput. Meths. Appl. Mech. Engrg.*, **47**(1984) (special issue).

[22] R. E. Ewing, "Adaptive mesh refinement in petroleum reservoir simulation," in *Accuracy Estimates and Adaptivity for Finite Elements* (I. Babuska, O.C. Zienkiewicz and E. Arantes e Oliveira, eds.), Wiley, New York, 1985.

[23] R. E. Ewing, "Finite element methods for nonlinear flows in porous media," *Comput. Meths. Appl. Mech. Engrg.*, in press.

[24] R. E. Ewing, "Efficient adaptive procedures for fluid flow applications," *Comput. Meths. Appl. Mech. Engrg.*, in press.

[25] R. E. Ewing and J. V. Koebbe, "Mixed finite element methods for groundwater flow and contaminant transport," *Proc. Fifth IMACS Internat. Symp. on Partial Differential Equations*, Bethlehem, Pennsylvania, 1984.

[26] R. E. Ewing, J. V. Koebbe, R. Gonzalez and M. F. Wheeler, "Computing accurate velocities for fluid flow in porous media," *Proc. Fifth Internat. Symp. on Finite Elements and Flow Problems*, Austin, Texas, 1984, 233–249.

[27] R. E. Ewing, S. McCormick and J. Thomas, "The fast adaptive composite grid method for solving differential boundary-value problems," *Proc. Fifth ASCE Specialty Conf.*, Laramie, Wyoming, 1984, 1453–1456.

[28] R. E. Ewing, and T. F. Russell, "Efficient time-stepping methods for miscible displacement problems in porous media," *SIAM J. Numer. Anal.*, **19**(1982), 1–66.

[29] R. E. Ewing, T. F. Russell and M. F. Wheeler, "Simulation of miscible displacement using mixed methods and a modified method of characteristics," *Proc. Seventh SPE Symp. on Reservoir Simulation*, San Francisco, 1983.

[30] R. E. Ewing, T. F. Russell and M.F. Wheeler, "Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics," *Comput. Meth. Appl. Mech. Engrg.*, **47**(1984), 73–92.

[31] R. E. Ewing, and M.F. Wheeler, "Galerkin methods for miscible displacement problems in porous media," *SIAM J. Numer. Anal.*, **17**(1980), 351–365.

[32] R. E. Ewing and M. F. Wheeler, "Galerkin methods for miscible displacement problems with point sources and sinks — unit mobility ratio case," in *Lectures on the Numerical Solution of Partial Differential Equations*, University of Maryland, 1981, 151–174.

[33] R. E. Ewing and M. F. Wheeler, "Computational aspects of mixed finite element methods," in *Numerical Methods for Scientific Computing* (R.S. Stepleman, ed.), North-Holland Publishing Co., 1983, 163–172.

[34] J. T. Oden, "Adaptive methods for incompressible viscous flow with moving boundaries," in *Accuracy Estimates and Adaptivity for Finite Elements* (I. Babuška, O.C. Zienkiewicz and E. Arantes e Oliveira, eds.), Wiley, New York, 1985.

[35] P. A. Raviart and J. M. Thomas, "A mixed finite element method for second order elliptic problems," in *Mathematical Aspects of the Finite Element Method*, Springer-Verlag, Heidelberg, 1977.

[36] W. C. Rheinboldt and C. K. Mesztenyi, "On a data structure for adaptive finite element mesh refinements," *TOMS*, 6(1980), 166–187.

[37] T. F. Russell, *An Incomplete Iterated Characteristic Finite Element Method for a Miscible Displacement Problem*, Ph.D. Thesis, University of Chicago, 1980.

[38] T. F. Russell, "Finite elements with characteristics for two-component incompressible miscible displacement," SPE 10500, *Sixth SPE Symp. on Reservoir Simulation*, New Orleans, 1982.

[39] T. F. Russell and M. F. Wheeler, "Finite element and finite difference methods for continuous flows in porous media," in *Mathematics of Reservoir Simulation* (R. E. Ewing, ed.), SIAM Publications, Philadelphia, 1983, 35–106.

[40] A. Weiser, "Local-mesh, local-order, adaptive finite element methods with a posteriori error estimates for elliptic partial differential equations," *Technical Report No. 213*, Dept. of Computer Science, Yale University, New Haven, Conn., 1981.

[41] M. F. Wheeler and R. Gonzalez, "Mixed finite element methods for petroleum reservoir engineering problems," in *Computing Methods in Applied Science and Engineering 6* (R. Glowinski and J. L. Lions, eds.), North-Holland, Amsterdam, 1984.

ASYMPTOTIC STABILITY OF VISCOUS SHOCK WAVES

F. A. Howes
Lawrence Livermore National Laboratory
P. O. Box 808, L-321
Livermore, CA 94550

ABSTRACT. We present sufficient conditions for the asymptotic stability of steady solutions of initial-boundary value problems for parabolic conservation laws that model one-dimensional transonic flow in ducts of variable cross-sectional area. The stability conditions consist of the standard (Lax) entropy conditions for the associated hyperbolic system and a local dissipativity condition on the source terms. In the context of the duct flow problem the latter is a geometric condition that guarantees the asymptotic stability of the standing shock wave known to exist in the diverging portion of the duct.

I. INTRODUCTION. The problem of determining the stability (or instability) of steady solutions of the equations of motion for a viscous, heat-conducting fluid has attracted the attention of many physical scientists and applied mathematicians since the basic equations were written down in the last century. In this paper we study the asymptotic stability of steady shock-layer solutions of hyperbolic systems of conservation laws to which have been added formally small terms representing the effects of viscosity and heat conduction. Our approach is to employ the usual inviscid entropy conditions (which are, of course, stability conditions [5]) in conjunction with a condition on the source terms in a neighborhood of the actual viscous layer, in order to estimate the size of an initial perturbation. Inside the layer we also make use of some asymptotic estimates on solutions of a steady equation resulting from the balance between inertial and viscous forces. We are able to show that under the stated assumptions such a steady solution is asymptotically stable with respect to all sufficiently small perturbations in the initial data.

This study was motivated by the recent papers [7,8], [2] which are concerned, in part, with the stability properties of standing transonic shock waves in the flow of an ideal gas through a duct of variable cross-sectional area. We therefore discuss in Section III how our result for the general problem, when applied to the gasdynamic model, guarantees the asymptotic stability of any standing viscous shock wave located in a diverging portion of the duct. As a prelude, we treat rather thoroughly in the next section an instructive scalar problem in order to illustrate our approach in a simple setting.

II. A MODEL PROBLEM. Consider the following problem

$$u_t + uu_x - r(x)u = \varepsilon u_{xx}, \ 0 < x < 1, \ t > 0,$$

$$u(x,0,\varepsilon) = \varphi(x,\varepsilon) , \ x \text{ in } [0,1] , \qquad\qquad (2.1)$$

$$u(0,t,\varepsilon) = \alpha, \ u(1,t,\varepsilon) = \beta, \ t \text{ in } [0,\infty),$$

where r, φ are smooth functions, ε is a small positive parameter and α,β are constants; cf. [4]. The inviscid ($\varepsilon = 0$) version of (2.1) was considered by Embid et al. [2]. It is not difficult to see that the corresponding steady problem

$$\varepsilon U_{xx} = UU_x - r(x)U, \quad 0 < x < 1,$$

(2.2)

$$U(0,\varepsilon) = \alpha, \quad U(1,\varepsilon) = \beta,$$

has shock-layer solutions $U = U(x,\varepsilon)$ connecting the inviscid branches $U_L(x) := R(x) + \alpha$ and $U_R(x) := R(x) - R(1) + \beta$, for $R(x) := \int_0^x r(s)ds$, provided U_L and U_R satisfy the entropy condition

$$U_L(x) > 0 > U_R(x) \text{ in } [0,1];$$

(2.3)

cf.[4], [2]. The location x_0 of the shock layer is obtained from the Rankine-Hugoniot relation $(U_L + U_R)(x_0) = 0$, that is, $R(x_0) = [R(1) - \alpha - \beta]/2$, which may have more than one solution x_0 in $(0,1)$, depending on the form of r.

Let us now test the stability of one such steady solution $U(x,\varepsilon)$, having a shock layer of width $O(\varepsilon)$ at x_0 in $(0,1)$, by introducing the perturbation $w(x,t,\varepsilon) := u(x,t,\varepsilon) - U(x,\varepsilon)$ into (2.1). The perturbation problem so obtained is

$$w_t + [(U + w)^2 - U^2]_x/2 - r(x)w = \varepsilon w_{xx},$$

(2.4)

$$w(x,0,\varepsilon) = \psi(x,\varepsilon), \quad w(0,t,\varepsilon) = w(1,t,\varepsilon) = 0,$$

where $\psi := \varphi - U$ is the initial perturbation. Thus, in order to show that the standing shock U is an asymptotically stable steady state of (2.1) it is enough to show that $w \equiv 0$ is an asymptotically stable solution of (2.4). We begin by examining first a small (two-sided) neighborhood Δ of x_0 and noting that in Δ the initial value problem

$$\varepsilon W_x = [(U + W)^2 - U^2]/2, \quad W(x_0,\varepsilon) > \|\psi\|_\infty,$$

(2.5)

has a positive solution $W = W(x,\varepsilon)$ which behaves like a δ-function peaked at x_0. More precisely, we have that

$$W(x,\varepsilon) = O(W(x_0,\varepsilon)\exp[-\gamma(x-x_0)^2/2\varepsilon^2]),$$

(2.6)

for a known positive constant γ, since (2.5) is a Bernoulli equation which becomes a linear equation in the variable $1/W$. Using W we can now construct a barrier function for (2.4) that ensures the asymptotic stability of $w = 0$ in Δ, provided we assume also that

$$r(x) \le -\nu < 0 \text{ in } \Delta$$

(2.7)

for a positive constant ν. A suitable barrier function is

$$\Omega(x,t,\varepsilon) := W(x,\varepsilon) e^{-\mu t}, \quad 0 < \mu < \nu,$$

since $-\Omega(x,0,\varepsilon) \le \psi(x,\varepsilon) \le \Omega(x,0,\varepsilon)$, $\Omega_t + \Phi_x(x,\Omega,\varepsilon) - r(x)\Omega - \varepsilon\Omega_{xx} \ge 0$ and $(-\Omega)_t + \Phi_x(x,-\Omega,\varepsilon) - r(x)(-\Omega) - \varepsilon(-\Omega)_{xx} \le 0$, for $\Phi(x,w,\varepsilon) := Uw + w^2/2$. To verify this it is enough to note that

$$\Omega_t + [U\Omega + \Omega^2/2]_x - r(x)\Omega - \varepsilon\Omega_{xx}$$

$$= e^{-\mu t}\{-\mu W + WW_x e^{-\mu t} - WW_x - r(x)W\}$$

$$\geq We^{-\mu t}\{\nu - \mu - (1 - e^{-\mu t})W_x\}$$

$$> 0$$

in Δ if μ and $W(x_0,\varepsilon)$ are sufficiently small, since $W_x(x,\varepsilon) = 0(W(x_0,\varepsilon)\;[\,|x-x_0|/\varepsilon + W(x_0,\varepsilon)]/\varepsilon)$. This estimate on W_x follows directly from (2.6) and the estimate $-U(x,\varepsilon) = 0(|x-x_0|/\varepsilon)$, which holds in Δ. Consequently in a small neighborhood of the shock layer the solution of (2.4) satisfies

$$|w(x,t,\varepsilon)| \leq W(x,\varepsilon)\,e^{-\mu t}\,,$$

and so w=0 (and hence, U) are asymptotically stable there with respect to all sufficiently small perturbations of the initial data.

Away from the layer the analysis actually simplifies dramatically, thanks to the entropy condition (2.3). To see this we begin by recalling some basic results on the asymptotic stability of the trivial solution of the general problem

$$w_t + f(x,w,\varepsilon)w_x + g(x,w,\varepsilon) = \varepsilon w_{xx}, \; 0 < x < 1, \; t > 0\,,$$

$$w(x,0,\varepsilon) = \psi(x,\varepsilon), \; w(0,t,\varepsilon) = w(1,t,\varepsilon) = 0\,,$$

(2.8)

where $g(x,0,\varepsilon) \equiv 0$; cf. [4].

Lemma 2.1. Suppose there exists a positive constant m such that for (x,w,ε) in $\mathscr{D}:=$ $[0,1] \times [-\delta,\delta] \times (0,\varepsilon_0]$

$$(\partial g/\partial w)\,(x,w,\varepsilon) \geq m > 0\,,$$

for δ and ε_0 positive constants. Then the solution w of (2.8) satisfies

$$|w(x,t,\varepsilon)| \leq \|\psi\|_\infty e^{-mt} \text{ in } [0,1] \times [0,\infty) \times (0,\varepsilon_0]$$

provided $\|\psi\|_\infty \leq \delta$.

This is the familiar result involving "linearized" stability, and it follows by noting that $\Omega(t):= \|\psi\|_\infty e^{-mt}$ is a barrier function for (2.8). The next result imposes a strong condition on the function f and a relatively mild one on g.

Lemma 2.2. Suppose there exist positive constants k and ℓ such that for (x,w,ε) in $\mathscr{D}$

$$|f(x,w,\varepsilon)| \geq k > 0 \text{ and } (\partial g/\partial w)\,(x,w,\varepsilon) \geq -\ell\,.$$

Then the solution w of (2.8) satisfies

$$|w(x,t,\varepsilon)| \le Le^{-nt} \text{ in } [0,1] \times [0,\infty) \times (0,\varepsilon_o]$$

<u>provided</u> $\|\psi\|_\infty \le \delta$. Here $L:=\|\psi\|_\infty(2e^{-\lambda}- 1)$ and $n:=\ell/(2e^{-\lambda} - 1)$, for $\lambda:= -\ell/k + O(\varepsilon)$ a negative root of the characteristic polynomial $\varepsilon\lambda^2 + k\lambda + \ell$.

This result follows because the positivity of $|f|$ and semiboundedness of g_w allow us to convert the equation in (2.8), via an exponential change of variable (in x), into an equivalent equation whose g-part satisfies the positivity condition in Lemma 2.1; cf. [4].

Returning to the perturbation problem (2.4) away from the shock layer, we see that by virtue of the entropy condition (2.3) Lemma 2.2 applies immediately, since

$$|f(x,w,\varepsilon)| = |U(x,\varepsilon) + w| \text{ is positive}$$

and

$$|(\partial g/\partial w)(x,w,\varepsilon)| = |U_x(x,\varepsilon) - r(x)|$$

is bounded outside of Δ, for all sufficiently small initial perturbations. We conclude therefore that an entropy-condition-satisfying standing shock wave solution of problem (2.1) is asymptotically stable provided the additional condition (2.7) is satisfied in a neighborhood of the viscous layer. Embid et al. [2] have shown that an entropy-condition satisfying inviscid standing shock is, in fact, <u>unstable</u> if this negativity assumption is violated. As we shall see in the next section, the negativity [positivity] of r models in a simple way the divergence [convergence] of a duct containing standing transonic shock waves.

III. THE GASDYNAMIC EQUATIONS. In this final section we give a heuristic analysis of the asymptotic stability of a standing viscous shock wave in the diverging portion of a variable-area duct. The equations that model the transonic flow of an inviscid, non-heat-conducting gas in a duct of unit length written in divergence form are (cf. [6; Chap. 2], [7,8])

$$\rho_t + (\rho u)_x + c(x)\,\rho u = 0 \,,$$

$$(\rho u)_t + (\rho u^2 + p)_x + c(x)\,\rho u^2 = 0 \,, \quad 0 < x < 1, \ t > 0 \,,$$

$$(\rho E)_t + (\rho E u + pu)_x + c(x)\,(\rho E u + pu) = 0 \,,$$

where $c(x):= a'(x)/a(x)$, for $a(x)$ the cross-sectional area of the duct, and ρ,u,p and E are the density, velocity, pressure and total energy of the gas, respectively. If we let $v:= (\rho,\rho u,\rho E)$ denote the vector of conserved densities, $f:= (\rho u,\rho u^2 + p,\rho E u + pu)$ the vector of fluxes and $g:= (\rho u,\rho u^2,\rho E u + pu)$, then we can write these equations more simply as

$$v_t + f(v)_x + c(x)\,g(v) = 0 \,. \tag{3.1}$$

If we now assume that dissipative effects such as viscosity (μ) and thermal conductivity (κ) are small but nonzero, then we must add to the righthand sides of the second and third equations in (3.1) second-order terms proportional to μ and κ. The assumed smallness of these coefficients suggests modelling their presence by adding to the righthand side of (3.1) the formally small term $\varepsilon B v_{xx}$ where $0 < \varepsilon \ll 1$ and B is a constant matrix whose eigenvalues have nonnegative real parts (cf.[9; Chap. 15]). Thus we are led to consider the following <u>parabolic</u> initial-boundary value problem

$$v_t + f(v)_x + c(x)g(v) = \varepsilon B v_{xx}, \ 0 < x < 1, \ t > 0,$$

$$v(x, 0, \varepsilon) = \psi(x, \varepsilon), \ x \text{ in } [0, 1], \tag{3.2}$$

$$v(0, t, \varepsilon) = v_0, \ v(1, t, \varepsilon) = v_1, \ t \text{ in } [0, \infty),$$

as a model for time-dependent, one-dimensional transonic flow in a variable-area duct with prescribed supersonic inlet ($x=0$) and subsonic outlet ($x=1$) values of ρ, u and E.

Let us focus our attention now on the steady-state solutions of (3.2), that is, on solutions of the boundary value problem

$$\varepsilon B V_{xx} = f(V)_x + c(x)g(V), \ 0 < x < 1,$$

$$V(0, \varepsilon) = v_0, \ V(1, \varepsilon) = v_1, \tag{3.3}$$

as $\varepsilon \to 0$. Under various simplifying assumptions this problem is known to have solutions of shock-layer type in regions where the function c is either positive or negative; cf. [2], [3]. Such solutions represent standing shock wave solutions of the original gasdynamic equations (3.1) either with structure (if $\varepsilon > 0$) or without structure (if $\varepsilon = 0$). In particular, Embid et al. [2] give rather detailed results in the case of isentropic flow of an inviscid, non-heat-conducting ($\varepsilon = 0$) gas through the study of an algebraic equation (the Hugoniot curve) derived from solutions of the first-order system $f(V)_x + c(x)g(V) = 0$. They prove the existence of standing shock waves in both the converging ($c < 0$) and diverging ($c > 0$) portions of the duct that satisfy the entropy condition

$$u_L - s_L > 0 > u_R - s_R.$$

Here $s(:=(\partial p / \partial \rho)^{1/2})$ is the local speed of sound and the subscripts L and R denote the limiting values of the variables to the left and the right of the shock, respectively. In order to proceed with our analysis we assume that the problem (3.3) has a smooth solution $V = V(x, \varepsilon)$ as $\varepsilon \to 0$ representing a transonic standing wave centered at x_0 in $(0, 1)$, that is,

$$\lim_{\varepsilon \to 0} V(x, \varepsilon) = \begin{cases} V_L(x), & 0 \leq x < x_0, \\ \\ V_R(x), & x_0 < x \leq 1, \end{cases}$$

which satisfies the entropy condition

601

$$U_L(x) > s_L(x) \, , \quad 0 \le x < x_0 \, ; \, U_R(x) < s_R(x) \, , \quad x_0 < x \le 1 \, , \tag{3.4}$$

for $U(x)$ the $(\varepsilon \to 0)$ – limit of $u(x,\varepsilon)$. We claim that such a shock is asymptotically stable with respect to all sufficiently small initial perturbations provided the function c is positive in a neighborhood of x_0.

To convince ourselves of this let us begin by examining briefly the same initial–boundary value problem (3.2) for the general parabolic system of n equations

$$v_t + F(v)_x + G(x,v) = \varepsilon B v_{xx} \, ,$$

under the basic assumption that its unperturbed $(\varepsilon=0)$ part is strictly hyperbolic and genuinely nonlinear. (Strict hyperbolicity means that the eigenvalues $\lambda(v)$ of the Jacobian matrix dF are real and distinct for all v of interest, say $\lambda_1(v) < \lambda_2(v) < \ldots < \lambda_n(v)$, while genuine nonlinearity is a generalization of the scalar notion of convexity to vector functions; cf.[5], [9;Chap. 17].) Suppose that the corresponding steady boundary value problem has a smooth solution $V = V(x,\varepsilon)$ with a shock layer at x_0 in $(0,1)$ which in the limit $\varepsilon \to 0$ is a standing k-shock ($1 \le k \le n$) satisfying the Lax entropy conditions (cf.[5], [9; Chap. 15])

$$\lambda_n(V_L) > \ldots > \lambda_k(V_L) > 0 > \lambda_k(V_R) > \ldots > \lambda_1(V_R)$$

$$\lambda_1(V_L) < \ldots < \lambda_{k-1}(V_L) < 0 < \lambda_{k+1}(V_R) < \ldots < \lambda_n(V_R) \, . \tag{3.5}$$

If we now introduce the perturbation $w := v - V$, then the resulting problem for w is

$$w_t + \{F(V+w) - F(V)\}_x + G(x,V+w) - G(x,V) = \varepsilon B w_{xx} \, ,$$

$$w(x,0,\varepsilon) = \psi(x,\varepsilon) := \varphi(x,\varepsilon) - V(x,\varepsilon), \, w(0,t,\varepsilon) = w(1,t,\varepsilon) = 0 \, , \tag{3.6}$$

and so the asymptotic stability of V implies and is implied by the condition that $\lim_{t \to \infty} \|w\|_\infty = 0$. Now away from x_0 these entropy conditions imply the asymptotic stability of V; this was shown by Liu [7,8] for the gasdynamic equations (3.1). Thus we have only to show that in an immediate neighborhood Δ of x_0 the perturbation w decays to 0 as $t \to \infty$. In order to accomplish this we make the further assumption (cf.[1]) that the function G is locally "dissipative," in the sense that there exists a positive constant m such that

$$w \cdot [G(x,V+w) - G(x,V)] \ge m \|w\|^2 \text{ in } \Delta \, , \tag{3.7}$$

for all w of interest. Let us now proceed as in the previous section by looking for a solution in Δ of the initial value problem

$$\varepsilon B W_x = F(V+W) - F(V) \, , \quad \|W(x_0, \varepsilon)\| > \|\psi\|_\infty \, . \tag{3.8}$$

In view of the entropy inequalities (3.5) the problem (3.8) has a solution $W = W(x,\varepsilon)$ as $\varepsilon \to 0$ such that $W_i > 0$ and $W_i = O(W_i(x_0,\varepsilon) \exp[\pm \gamma_i (x - x_0)^2 / 2\varepsilon^2])$ in Δ for known positive constants $\gamma_i (1 \le i \le n)$. Note that in contrast to the solution of the scalar problem (2.5) the solution of the vector problem has components which behave like inverted δ-functions as well as like the δ-function described by (2.6). It follows that each component of the gradient of W satisfies an estimate in Δ of the form

$$W_{i,x}(x,\varepsilon) = O(W_i(x_0,\varepsilon) \, [\, |x-x_0|/\varepsilon + W_i(x_0,\varepsilon)]/\varepsilon) \, . \tag{3.9}$$

Using W we can construct the vector barrier function

$$\Omega(x,t,\varepsilon) := W(x,\varepsilon) \, e^{-qt}, \; 0 < q < m \, .$$

Inserting Ω into (3.6) gives us

$$- qWe^{-qt} + \{F(V + We^{-qt}) - F(V)\}_x + G(x, V + We^{-qt}) - G(x,V) = \varepsilon BW_{xx}e^{-qt}$$

or, by virtue of (3.8) ,

$$- qWe^{-qt} + \{F(V + We^{-qt}) - F(V)\}_x - [F(V + W)e^{-qt} - F(V)e^{-qt}]_x$$

$$+ G(x, V + We^{-qt}) - G(x,V) = 0 \, . \tag{3.10}$$

Now

$$\{\bullet\} = dF(V)We^{-qt} + \tfrac{1}{2}P(W,W)e^{-2qt} + O(e^{-3qt})$$

and

$$[\bullet] = dF(V)We^{-qt} + \tfrac{1}{2}Q(W,W)e^{-qt} \, ,$$

where P and Q are bilinear forms representing quadratic terms, and so the i-th component of $\{\bullet\}_x - [\bullet]_x$ is of the order $O(W_iW_{i,x}e^{-qt})$ in Δ. Thus if we take Δ sufficiently small we see from the estimates (3.9) that the gradient terms in the comparison equation (3.10) may be neglected to lowest order relative to the other three terms, that is, we can replace (3.6) with the simpler system

$$w_t + G(x, V+w) - G(x,V) = 0 \, . \tag{3.11}$$

Upon dotting both sides of (3.11) with w and using the stability condition (3.7) we find that

$$(\tfrac{1}{2}\|w\|^2)_t \le - m \, \|w\|^2 \, ,$$

and so $\lim_{t \to \infty} \|w\| = 0$ in Δ provided $\|\psi\|$ is sufficiently small. Thus with the aid of the entropy conditions (3.5) and the dissipativity condition (3.7) we have outlined an argument suggesting the asymptotic stability of the steady shock solution V.

Let us return finally to the duct problem (3.1) and see how this general analysis applies to the gasdynamic equations. For the sake of simplicity we consider isentropic flow, for which the eigenvalues of the corresponding Jacobian matrix df are $\lambda_1 := u - s$ and $\lambda_2 := u + s$. Here u is the velocity and s is the sound speed; cf. [7, 8]. Thus the isentropic system (3.1) is strictly hyperbolic, and by virtue of the entropy condition (3.4), we see that

$$\lambda_1(V_L) > 0 > \lambda_1 (V_R), \; 0 < \lambda_2 (V_R) \, ,$$

that is, the steady shock wave whose stability is being testing is a 1-shock (cf.(3.5)). It only remains to investigate under what conditions the term $c(x)g(v)$ is locally dissipative in the sense of (3.7). To this end, note that $g(v) = (\rho u, \rho u^2)$ may be written as $u(\rho, \rho u) =$

$u(v_1, v_2)$, and so $c(x)[g(V+w) - g(V)] = c(x)U(w_1, w_2)$ satisfies (3.7) provided $c > 0$ in a neighborhood of the viscous layer, in agreement with the inviscid results of Embid et al. and Liu.

REFERENCES

1. C. M. Dafermos, "Asymptotic Behavior of Solutions of Hyperbolic Balance Laws," in Bifurcation Phenomena in Math. Physics, Bardos and Bessis, Eds. (D. Reidel, 1980) pp. 521-533.

2. P. Embid, J. Goodman, and A. Majda, "Multiple Steady States for 1-D Transonic Flow," SIAM J. Sci. Stat. Comp. 5, 21-41 (1984).

3. F. A. Howes, "An Analytical Treatment of the Formation of One-Dimensional Steady Shock Waves in Uniform and Diverging Ducts," J. Comp. Appl. Math. 10, 195-201 (1984).

4. _____, "Some Stability Results for Advection-Diffusion Equations, I, II," Studies in Applied Math., Part I 74, 35-53 (1986); Part II is in press.

5. P. D. Lax, "Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves," CBMS Ser. in Appl. Math., Vol. 11, SIAM, Philadelphia (1973).

6. H. W. Liepmann and A. Roshko, Elements of Gasdynamics (John Wiley, New York, 1957).

7. T. P. Liu, "Nonlinear Stability and Instability of Transonic Flows Through a Nozzle," Comm. Math. Physics 83, 243-260 (1982).

8. _____, "Transonic Flow In a Duct of Varying Area," Arch. Rational Mech. Anal. 80, 1-18 (1982).

9. J. Smoller, Shock Waves and Reaction-Diffusion Equations (Springer, New York, 1983).

1106E

604

EXTENSIONS OF SARKOVSKII'S THEOREM

Nam P. Bhatia

University of Maryland

Baltimore County, Maryland 21228

and

Walter O. Egerland

Ballistic Research Laboratory

Aberdeen Proving Ground, Maryland, 21005-5066

Introduction

It is known that a four-periodic orbit may imply a three-periodic orbit and hence an n-periodic orbit for every n = 1,2,... [1, Theorem 2; 2, Theorem 3]. Such an orbit has, of course, not the same structure (or, as we say, is not of the same "type") as the four-periodic orbit that appears in the Sarkovskii ordering. To clarify the nature of such and similar implications that are not accounted for in Sarkovskii's theorem, we introduced the notions of loop and infinite loop. Our investigations in this direction, begun under the US Army Summer Faculty Research and Engineering Program 1983 and continued in the 1984 and 1985 Program, showed that there is (i) an extension to the left in the Sarkovskii ordering and (ii) that there are infinitely many additional links within the ordering. The principal result, Theorem (SR), that summarizes this investigation, is stated in Section II. For this presentation we have singled out some results from our more recent work associated with elementary orbits, i.e., orbits of a certain type we christened "elementary." A sequence of theorems delineating the properties of elementary orbits culminates in the complete ordering of these orbits. The ordering of N (N: = the set of natural numbers) for the elementary orbits is different from the Sarkovskii ordering. Taking a few further steps, we arrive at Theorem (SRII), from which, as a corollary, Sarkovskii's theorem follows in a most natural way.

I. Definitions and Notation

Let $f: R \to R$ be continuous and $x_0 \epsilon R$. The orbit of x_0 under f is defined as the set $\{x: x = f^n(x_0), n = 0,1,...\}$, where, for every positive integer n, f^n is the n-th iterate of f, $f^1 = f$, and $f^0(x_0) = x_0$. We shall write $x_n: = f^n(x_0)$ for a given $x_0 \epsilon R$ and call $x_1, x_2, \cdots$ the successors of x_0. A pre-orbit of a given $x_0 \epsilon R$ is any (finite or infinite) sequence $x_0, x_{-1}, x_{-2},...$ such that $f(x_{-n}) = x_{-(n-1)}$ for all n for which x_{-n} is defined. The points $x_{-1}, x_{-2}, \cdots$ in any such sequence are called predecessors of x_0. A point c_0 is called critical if $f(c_0) = c_0$, i.e., a critical point of f is a fixed point of f. A periodic point x_0 of period $p > 1$ (p a positive integer) is a point for which the relations $f^p(x_0) = x_0$, $f^k(x_0) \neq x_0$, $1 \leq k < p$, hold. If x_0 is a periodic point of period p, its orbit is denoted by $(x_0, x_1, \ldots, x_{p-1})$. We shall denote the kth iterate of x_0 under the function f^m by $x_k{}^m$, $k = 0,1,....$ Thus $x_k{}^m: = (f^m)^k(x_0) = x_{mk}$, and, in particular, $x_0{}^m = x_k{}^0 = x_0$ for all nonnegative integers k and m.

Definition. Let $f: R \to R$ be continuous and $x_0 \epsilon R$. f has a loop of order n if x_0 has a pre-orbit $(x_0, x_{-1},...,x_{-n})$ such that either

$$x_0 < x_{-n} < x_{-(n-1)} < \cdots < x_{-2} < x_{-1}$$

or

$$x_0 > x_{-n} > x_{-(n-1)} > \cdots > x_{-2} > x_{-1}.$$

f has an infinite loop if x_0 has an infinite pre-orbit $(x_0, x_{-1},...,x_{-n}, \cdots)$ such that either

$$x_0 < \cdots < x_{-n} < x_{-(n-1)} < \cdots < x_{-2} < x_{-1}$$

or

$$x_0 > \cdots > x_{-n} > x_{-(n-1)} > \cdots > x_{-2} > x_{-1}$$

A loop of order (n-1) is called an n-periodic loop if $x_0 = x_{-n}$.

Definition. A periodic orbit $(x_0, x_1, \cdots, x_{n-1})$ of period n is called elementary if

$$x_{2\nu} < \cdots < x_4 < x_2 < x_0 < x_1 < x_3 < \cdots < x_{2k-1}$$

or

$$x_{2\nu} > \cdots > x_4 > x_2 > x_0 > x_1 > x_3 > \cdots > x_{2k-1},$$

where $\nu + 1 = k = \dfrac{n}{2}$ when n is even and $\nu = \dfrac{n-1}{2} = k$ when n is odd.

An infinite pre-orbit ($x_0, x_{-1}, x_{-2}, \cdots$) is called elementary if the inequalities

$$x_{-2} < x_{-4} < \cdots < x_0 < \cdots < x_{-3} < x_{-1}$$

or

$$x_{-2} > x_{-4} > \cdots > x_0 > \cdots > x_{-3} > x_{-1}$$

hold.

We adopt the following concise notation: we say property $P(k)$ holds if f has a periodic orbit of period k. Thus $P(1)$, $L(k)$, $E(k)$, $L(\infty)$, $E(\infty)$ mean that f has a critical point, a periodic loop of period k, an elementary orbit of period k, an infinite loop, an infinite elementary pre-orbit, respectively. Similarly $P^n(k)$, $L^n(k)$, $E^n(k)$, $L^n(\infty)$, $E^n(\infty)$ shall mean that f^n has a k-periodic orbit, k-periodic loop, k-periodic elementary orbit, an infinite loop, an infinite elementary pre-orbit, respectively. The implication "A implies B" is denoted by $A \to B$, and "A iff B" is denoted by $A \leftrightarrow B$.

II. Sarkovskii's Theorem and Theorem (SR)

Using the notation introduced in Section I, Sarkovskii's theorem and our refinement read as follows.

__Theorem (Sarkovskii).__ Let f: $R \to R$ be continuous. Then

$$P(3) \to P(5) \to P(7) \to \cdots \to$$
$$P(2 \cdot 3) \to P(2 \cdot 5) \to P(2 \cdot 7) \to \cdots \to$$
$$P(2^2 \cdot 3) \to P(2^2 \cdot 5) \to P(2^2 \cdot 7) \to \cdots \to$$
$$\cdots \to$$
$$P(2^3) \to P(2^2) \to P(2) \to P(1).$$

607

<u>Theorem (SR)</u>. Let f: R → R be continuous. Then

$$L(\infty) \cdots \rightarrow L(5) \rightarrow L(4) \rightarrow L(3) \leftrightarrow$$
$$P(3) \rightarrow P(5) \rightarrow P(7) \rightarrow \cdots \rightarrow$$
$$L^2(\infty) \cdots \rightarrow L^2(5) \rightarrow L^2(4) \rightarrow L^2(3) \leftrightarrow$$
$$P(2 \cdot 3) \rightarrow P(2 \cdot 5) \rightarrow P(2 \cdot 7) \rightarrow \cdots \rightarrow$$
$$L^{22}(\infty) \cdots \rightarrow L^{22}(5) \rightarrow L^{22}(4) \rightarrow L^{22}(3) \leftrightarrow$$
$$P(2^2 \cdot 3) \rightarrow P(2^2 \cdot 5) \rightarrow P(2^2 \cdot 7) \rightarrow \cdots \rightarrow$$
$$\cdots$$
$$\ldots \rightarrow P(2^3) \rightarrow P(2^2) \rightarrow P(2) \rightarrow P(1).$$

For the proof of Theorem (SR) the reader is referred to [3].

III. The Hierarchy of Elementary Orbits

<u>Theorem 1.</u> If f has an elementary orbit of period (2n+1), then it has two distinct elementary orbits of period (2n+3), i.e., E(2n+1) → E(2n+3), n=1,2, · · · .

<u>Theorem 2.</u> E(2n+1) → E(∞), n=1,2, · · · .

<u>Theorem 3.</u> E(∞) → E(2m), m=1,2, · · · .

<u>Theorem 4.</u> E(2m+2) → E(2m), m=1,2, · · · . There exist two distinct elementary orbits of period 2m if m=3,4, · · · .

Combining these four theorems, we obtain the complete ordering of elementary orbits.

<u>Theorem 5 (The Complete Ordering of Elementary Orbits).</u>

$$E(3) \rightarrow E(5) \rightarrow E(7) \rightarrow \cdots \rightarrow E(\infty) \rightarrow \cdots \rightarrow E(8) \rightarrow E(6) \rightarrow E(4) \rightarrow E(2) \rightarrow E(1).$$

Consider now the complete ordering of elementary orbits for f, f², · · · ,f²ⁿ, f²ⁿ⁺¹, · · · .
Then the implications E(6) → E²(3) and E²(2) → E(2) provide the linkage shown in the following important result.

<u>Theorem 6.</u> Let f: R → R be continuous. Then

$$E(3) \to E(5) \to \cdots E(\infty) \to \cdots \to E(6) \to E(4) \to E(2) \to E(1)$$

$$E^2(3) \to E^2(5) \to \cdots E^2(\infty) \to \cdots \to E^2(6) \to E^2(4) \to E^2(2) \to E^2(1)$$

$$E^{2n}(3) \to E^{2n}(5) \to \cdots E^{2n}(\infty) \to \cdots \to E^{2n}(6) \to E^{2n}(4) \to E^{2n}(2) \to E^{2n}(1)$$

$$E^{2n+1}(3) \to E^{2n+1}(5) \to \cdots E^{2n+1}(\infty) \to \cdots \to E^{2n+1}(6) \to E^{2n+1}(4) \to E^{2n+1}(2) \to E^{2n+1}(1)$$

That the Sarkovskii ordering is contained in the ordering established in Theorem 6 follows from the implications $P(2n+1) \leftrightarrow E(2n+1)$, $E(2) \leftrightarrow P(2)$, and $P^2(n) \leftrightarrow P(2n)$, which, in turn, ensure that

$$E^{2n}(2m+1) \leftrightarrow P((2m+1)2^n)$$

and

$$E^{2n}(2) \leftrightarrow P(2^{n+1}) \, .$$

In terms of properties $P(k)$ and $E(k)$ we have, therefore, the following refinement of Sarkovskii's theorem.

<u>Theorem (SRII).</u> Let f: R → R be continuous. Then

$$P(3) \to P(5) \to \cdots \to E(\infty) \to \cdots \to E(8) \to E(6) \to E(4) \to P(2) \to P(1)$$

$$P(2\cdot3) \to P(2\cdot5) \to \cdots \to E^2(\infty) \to \cdots \to E^2(8) \to E^2(6) \to E^2(4) \to P(2^2) \to P^2(1)$$

$$P(2^2\cdot3) \to P(2^2\cdot5) \to \cdots \to E^{2^2}(\infty) \to \cdots \to E^{2^2}(8) \to E^{2^2}(6) \to E^{2^2}(4) \to P(2^{2^2}) \to P^{2^2}(1)$$

Acknowledgements.

609

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, On the Existence of Li-Yorke Points in the Theory of Chaos, Nonlinear Analysis, Theory, Methods & Applications, Vol. 10, No. 6, pp. 541-545, 1986.

3. Nam P. Bhatia and Walter O. Egerland, A Refinement of Sarkovskii's Theorem, Mathematics Research Report No. 85-10, Department of Mathematics, UMBC, November 1985. (Also to appear elsewhere in the literature).

610

Poincaré Maps of a Journal Bearing
P. J. Hollis and D. L. Taylor
Sibley School of Mechanical and Aerospace Engineering
Cornell University, Ithaca, NY 14853

Abstract

It is known that the model for a single unforced fluid film journal bearing shows Hopf bifurcation to both stable and unstable limit cycles. In certain parameter ranges, both limit cycles exist at the same time. When a rotating unbalance is introduced into the system, it is expected that some period doubling bifurcations will occur as the amplitude of the forcing is increased. The Poincaré map generated by simulation of the equations of motion for the system is used to show these bifurcations for a particular bearing system. By varying system parameters, the Poincaré map can be used to build up a catalogue of possible behaviors for the journal bearing system. At present, the range of possible behaviors is unknown. With the introduction of periodic forcing, many nonlinear systems are known to show chaotic behavior in certain regions of their parameter space. The Poincaré map shows that apparently chaotic behavior is possible for the journal bearing system for certain sets of parameter values.

Journal Bearing Equations

Fig. 1 shows a rotor of diameter D and weight W spinning with angular velocity Ω in a journal bearing of length L. The position of the rotor is given in polar coordinates by E and Φ and in cartesian coordinates by X and Y. The clearance C is the difference between the radius of the journal and the radius of the rotor. The space between the rotor and the journal is filled with a fluid. As the rotor spins, pressure forces which support the rotor are generated. By solving Reynolds equation for this journal bearing system, the radial and tangential forces on the rotor, F_E and F_Φ, can be found. These forces can also be resolved into components F_X and F_Y in cartesian coordinates. The equations of motion are most often nondimensionalized to reduce the number of parameters, [1], [2], [3], [4], [5], [6].

Using the finite model proposed by [2], the nondimensional equations of motion become

$$f_r = 4\left[a(g_3)\dot{\varepsilon}J_3^{02} + a(g_1)\varepsilon(\dot{\phi} - \tfrac{1}{2})J_3^{11} \right]$$
$$f_t = -4\left[a(g_4)\dot{\varepsilon}J_3^{11} + a(g_2)\varepsilon(\dot{\phi} - \tfrac{1}{2})J_3^{20} \right]$$

where

$$J_n^{lm} = \int_{\theta_1}^{\theta_1 + \pi} \frac{\sin^l\theta \cos^m\theta}{(1 + \varepsilon\cos\theta)^n}\, d\theta.$$

$$\theta_1 = \tan^{-1}\left(\frac{2\dot{\varepsilon}}{\varepsilon(\omega - 2\dot{\phi})} \right)$$

$$a(x) = \frac{3(x - \tanh x)}{x^3}$$

611
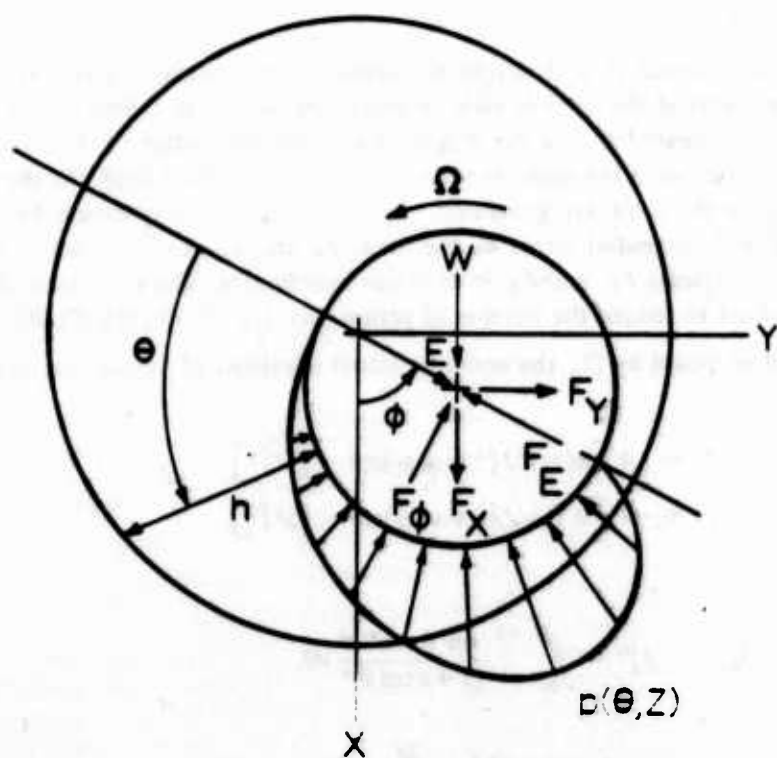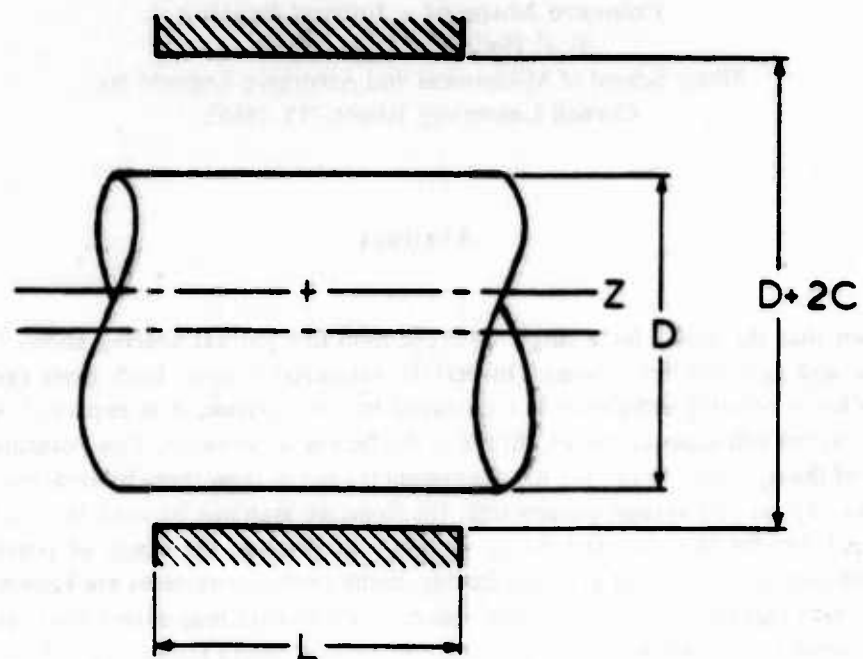
Figure 1. Bearing goemetry for a single fluid film journal bearing.

$$g_1^2 = \frac{\pi(2+\varepsilon^2)}{\varepsilon^2}\left[\frac{L}{D}\right]^2\left[1+\frac{1}{\sqrt{1-\varepsilon^2}}-\frac{4}{\sqrt{4-\varepsilon^2}}\right]$$

$$g_2^2 = \frac{\pi(2+\varepsilon^2)}{2}\left[\frac{L}{D}\right]^2\left[\frac{1}{\sqrt{1-\varepsilon^2}}-\frac{1}{\sqrt{4-\varepsilon^2}}\right]$$

$$g_3^2 = \frac{2}{\varepsilon^2}\left[\frac{L}{D}\right]^2\left[\pi+\frac{2(\pi-\gamma_1)}{\sqrt{1-\varepsilon^2}}-\frac{8(\pi-\gamma_2)}{\sqrt{4-\varepsilon^2}}\right]$$

$$g_4^2 = 2\left[\frac{L}{D}\right]^2\left[\frac{\pi-\gamma_1}{\sqrt{1-\varepsilon^2}}-\frac{\pi-\gamma_2}{\sqrt{4-\varepsilon^2}}\right]$$

and

$$\gamma_1 = \tan^{-1}\left[\frac{\sqrt{1-\varepsilon^2}}{\varepsilon}\right]$$

$$\gamma_2 = \tan^{-1}\left[\frac{\sqrt{4-\varepsilon^2}}{\varepsilon}\right]$$

The J_n^{lm} are often referred to as the bearing integrals.

The nondimensional forces are expressed in cartesian form by using relations

$$f_x = -(f_r(\varepsilon,\phi,\dot{\varepsilon},\dot{\phi})\cos\phi + f_t(\varepsilon,\phi,\dot{\varepsilon},\dot{\phi})\sin\phi)$$

$$f_y = -(f_r(\varepsilon,\phi,\dot{\varepsilon},\dot{\phi})\sin\phi - f_t(\varepsilon,\phi,\dot{\varepsilon},\dot{\phi})\cos\phi)$$

Thus, the nondimensional equations of motion in state variable form are

$$\mathbf{x} = \frac{d}{dt}\begin{bmatrix} x \\ y \\ \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dfrac{1}{\omega\sqrt{(m/\sigma)}}f_x + \dfrac{1}{\omega^2} \\ \dfrac{1}{\omega\sqrt{(m/\sigma)}}f_y \end{bmatrix} = \mathbf{f}(\mathbf{x})$$

The parameters to be used are the nondimensional running speed ω, the bearing parameter m/σ and the finite length parameter L/D. The parameter m/σ describes most of the geometry of a particular bearing, while the parameter L/D completes the description.

Hopf Bifurcations of Journal Bearings

As the parameter ω is varied, a Hopf bifurcation occurs, [1], [3], [5], [6]. Fig. 2 shows the resulting stability boundary and bifurcation diagrams for the case $L/D = 0$. For $m/\sigma > 1.6$, stable supercritical limit cycles are predicted, while unstable subcritical limit cycles are predicted for $m/\sigma < 1.6$. Near the transition from stable to unstable limit cycles, two limit cycles are expected to exist at the same time, one being stable, the other unstable, [7]. Without resorting to the calculation of higher order terms for the bifurcation coefficients, conventional Hopf bifurcation theory gives no clue as to which of two possible bifurcation diagrams occurs, Fig. 3. Simulation of the equations of motion for certain parameter values shows that the stable limit cycle surrounds the unstable one, case (b) in Fig. 3. The Poincaré map is used to examine this region in more detail. The bifurcation diagram can be built up, and the effects of speed changes can be seen more easily. The four dimensional system of equations reduces to a three dimensional Poincaré map, an example of which is shown in Fig. 4. The X and Y coordinates correspond to the x and y coordinates of the equations while the Z coordinate corresponds to $\dot{x}$ in the

613
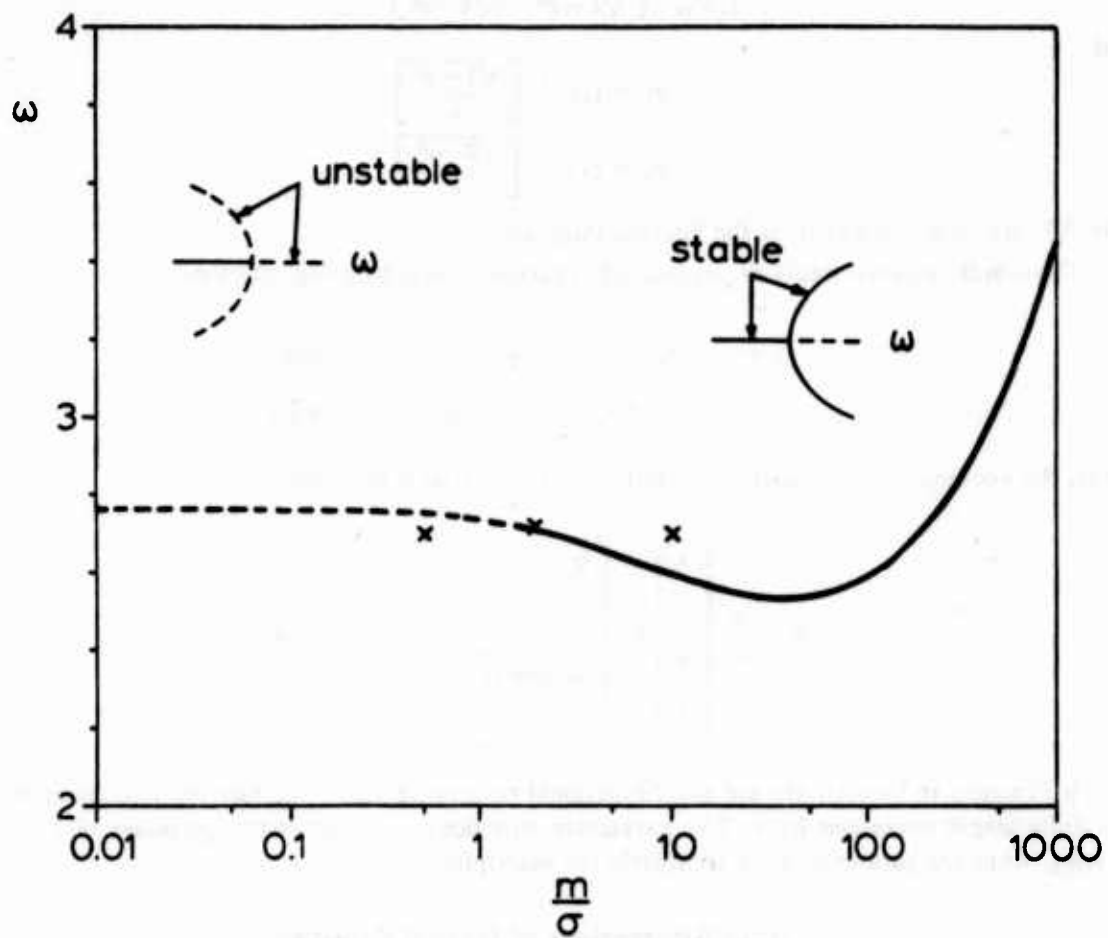
Figure 2. New stability boundary for $L/D = 0$.
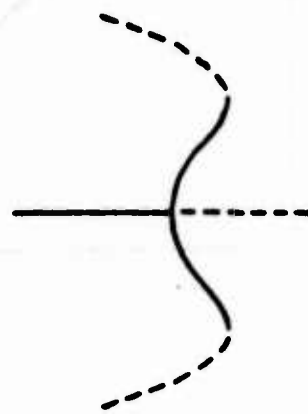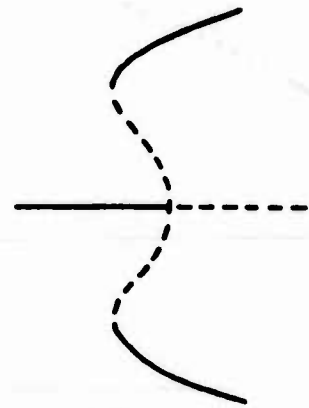
(a)   or   (b)

Figure 3. Possible bifurcation diagrams.
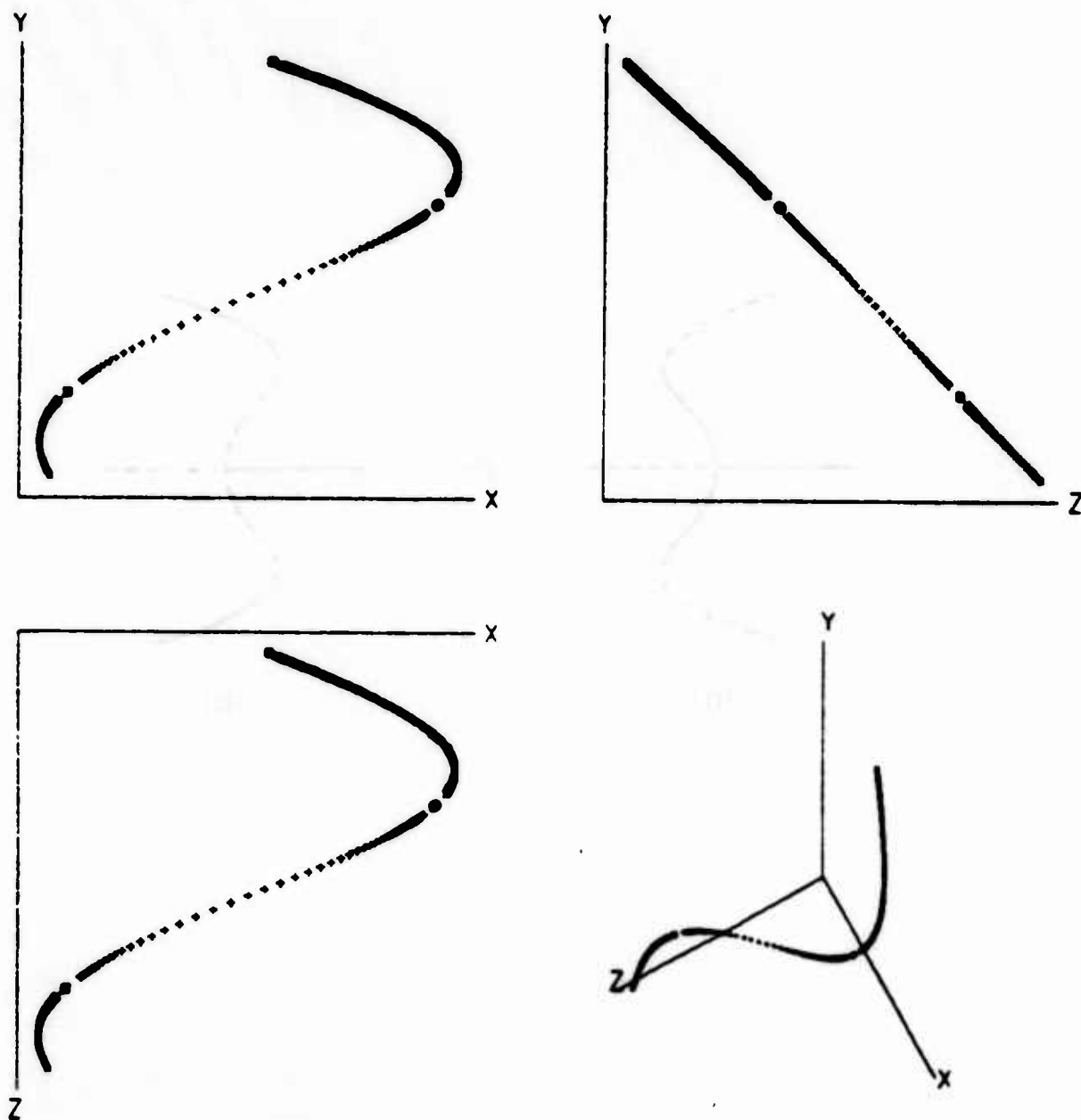
Figure 4. Poincaré map of journal bearing: $\omega = 2.70$, $\frac{L}{D} = 0$, $\frac{m}{\sigma} = 0.5$.

616

equations. The actual map used is to sample whenever $\dot{y} = 0$ with $\ddot{y}$ increasing. The three planar views $X-Y$, $X-Z$, and $Y-Z$ are projections to two dimensions of the 3-dimensional view shown in the lower right. This $X-Y-Z$ view is a picture of the complete Poincaré map with coordinates x, y, and $\dot{x}$ for the bearing system.

It can be seen that there are three fixed points, one corresponding to the stable fixed point, one corresponding to the unstable limit cycle, and the other to the stable limit cycle. The top fixed point is the stable fixed point of the system and the lower one is the fixed point associated with the stable limit cycle. The middle unstable fixed point is the unstable limit cycle. Different point types represent solutions starting from different initial conditions, which are not shown. All the fixed points found and shown have periods within 1% of 4π. By examining the maps as speed is varied, it turns out that most of the behavior lies on a 1-dimensional curve, which means that two of the eigenvalues of the map have very small magnitude. Neglecting these eigenvalues, looking only at the behavior along the curve, the effect of speed on behavior can be easily seen. Initially, there is only one fixed point, the stable fixed point of the original system. As speed increases, there appears a pair of fixed points, one stable, one unstable, which appear to diverge from a single point. They represent the appearance of the stable and unstable limit cycles below the critical speed. This pair of fixed points moves further apart until the unstable point meets the original stable fixed point at the critical speed. This corresponds to the unstable limit cycle shrinking down to the equilibrium point at the critical speed. Further increases in speed turn this into the unstable fixed point. The large amplitude limit cycle persists as a stable fixed point moving closer to the clearance boundary.

It is now possible to generate a more complete bifurcation diagram, Fig. 5, by using the fixed point information from the Poincaré maps. The boundary of clearance is ± 1, and it is unstable. The force becomes infinite at the boundary and so, theoretically, collisions with the boundary are not possible. No account is taken of oil film breakdown or surface deformations.

The program used to generate the Poincaré maps is an interactive set of command level programs and Fortran programs allowing the user to choose initial conditions on the picture, to interactively change certain attributes of the individual curves and to easily change system parameters. It does suffer from the same drawbacks as simulation. A large number of initial conditions need to be tried to be certain of capturing all the relevant behavior and the simulations tend to be computationally slow, lessening the interactive nature of the program. The interactive nature of the program does allow more rapid investigation of interesting areas than some discretization of intial condition space, which would use a significant amount of time producing essentially equivalent sets of points.

There are other methods for generating more complete bifurcation diagrams for autonomous systems. Continuation methods such as those described by [8] and [9] can also be used to trace out branches of bifurcating steady state solutions. In fact, the method described by [9] can be used to trace out branches of periodic solutions of limit cycles arising from Hopf bifurcation. These methods do require having some initial solution such as a fixed point or limit cycle from which the method begins. Newton's method and some minimization techniques may also be used to find limit cycles but their implementation is sometimes difficult.

Unbalanced Journal Bearing Equations

The Poincaré map does allow consideration of periodically forced systems. For the journal bearing, this is particularly useful since most bearing systems do suffer from some form of rotating unbalance. The nondimensional equations of motion for a journal bearing system with an unbalance rotating at the
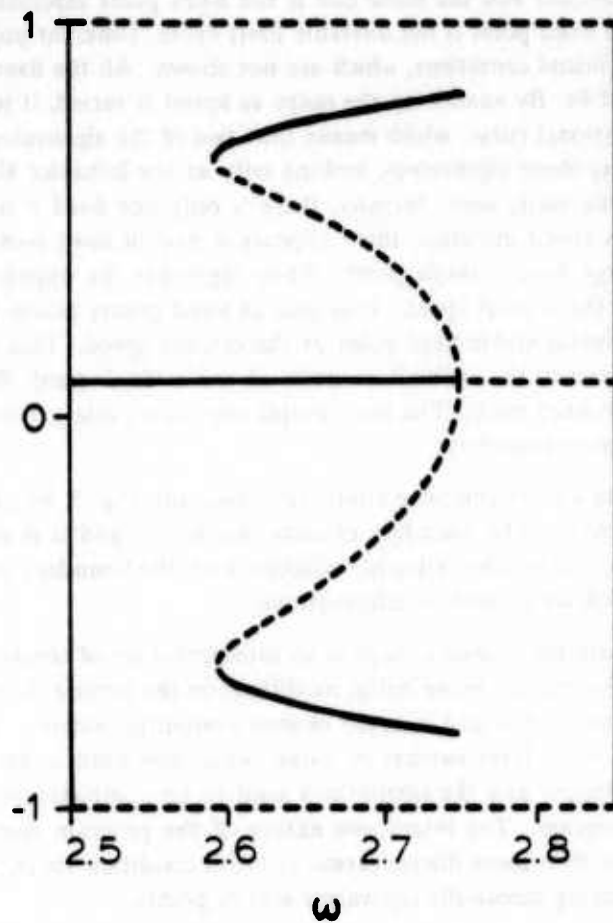
Figure 5. Bifurcation diagram for a journal bearing: $\frac{L}{D} = 0$, $\frac{m}{\sigma} = 0.5$.

same frequency as the rotor can easily be shown to be

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dfrac{1}{\omega\sqrt{(m/\sigma)}} f_x + \dfrac{1}{\omega^2} + a\cos t \\ \dfrac{1}{\omega\sqrt{(m/\sigma)}} f_y + a\sin t \end{bmatrix} = \mathbf{f}(\mathbf{x}, t)$$

where a is a further nondimensional parameter representing the magnitude of the unbalance.

Effect of Rotating Unbalance

For this periodically forced system, the most natural Poincaré map is to sample at period 2π, the period of forcing. The journal bearing system transformed to an autonomous system of equations is really 5-dimensional $(x, y, \dot{x}, \dot{y}, t)$. The resulting map is a 4-dimensional map from phase space to phase space. To see the behavior requires further projections down to 2 or 3 dimensions.

The unforced system with both large stable and unstable limit cycles existing below the critical speed, Fig. 4, will now have a rotating unbalance introduced. For small amplitudes of unbalance, it is seen that the behavior of the system is still close to that of the unforced case, except that the stable fixed point at the top now corresponds to a small amplitude orbit of period 2π. The other two fixed points have period 4π. As the amplitude of forcing is increased, the top two fixed points tend towards each other. The 2π-periodic stable point is approaching the 4π-periodic unstable point or the 2π-periodic stable orbit is approaching the the 4π-periodic unstable orbit, Fig. 6. By $a = 0.016$, Fig. 7, the two points have coalesced into a single unstable saddle point of period 2π. As the forcing is increased further, the 2π-periodic unstable saddle point moves closer to the 4π-periodic stable fixed point, or, in the original configuration space, the unstable orbit is approaching the large stable orbit, Fig. 6. By $a = 0.17$, the orbit has changed from one large orbit with no loops to one with a loop in it, Fig. 6. Further increasing the magnitude of the unbalance, the two fixed points of the stable orbit have coalesced with the saddle of the unstable orbit to form a single stable fixed point of period 2π with two negative real eigenvalues. On the (x, y) plane, the inner loop of the stable 4π-periodic orbit appears to have coalesced with the outer loop to form a single 2π-periodic orbit. The Poincaré maps for this system all seem to lie almost on a plane, indicating that one of the eigenvalues of the map remains very small. If Fig. 6 is looked at in reverse order, two period doubling bifurcations can be seen to occur. The stable orbit of period 2π bifurcates to a stable orbit of period 4π and an unstable orbit of period 2π. Then, the unstable orbit of period 2π bifurcates to an unstable orbit of period 4π and a stable orbit of period 2π.

Conclusions

By generating such Poincaré map pictures over the parameter space of the bearing system, a catalogue of all the possible behaviors could be built up. The boundaries separating topologically different behavior could be easily seen. It would be an extremely tedious task running enough cases over the entire four parameter space to capture all the possible behavior types which are expected to range from very simple to very complex. At present, it is unknown just what are all the possible types of behavior. Fig. 8 shows some complicated behavior for the journal bearing system. The map appears to have no periodic points but there does appear to be some structure. All the points tend to lie along some form of curve, suggesting at least two dominant frequencies in the response. The trajectory in the (x, y) plane for t from 400 to 800 is shown in Fig. 9. Most of the time, the trajectory is almost circular, but
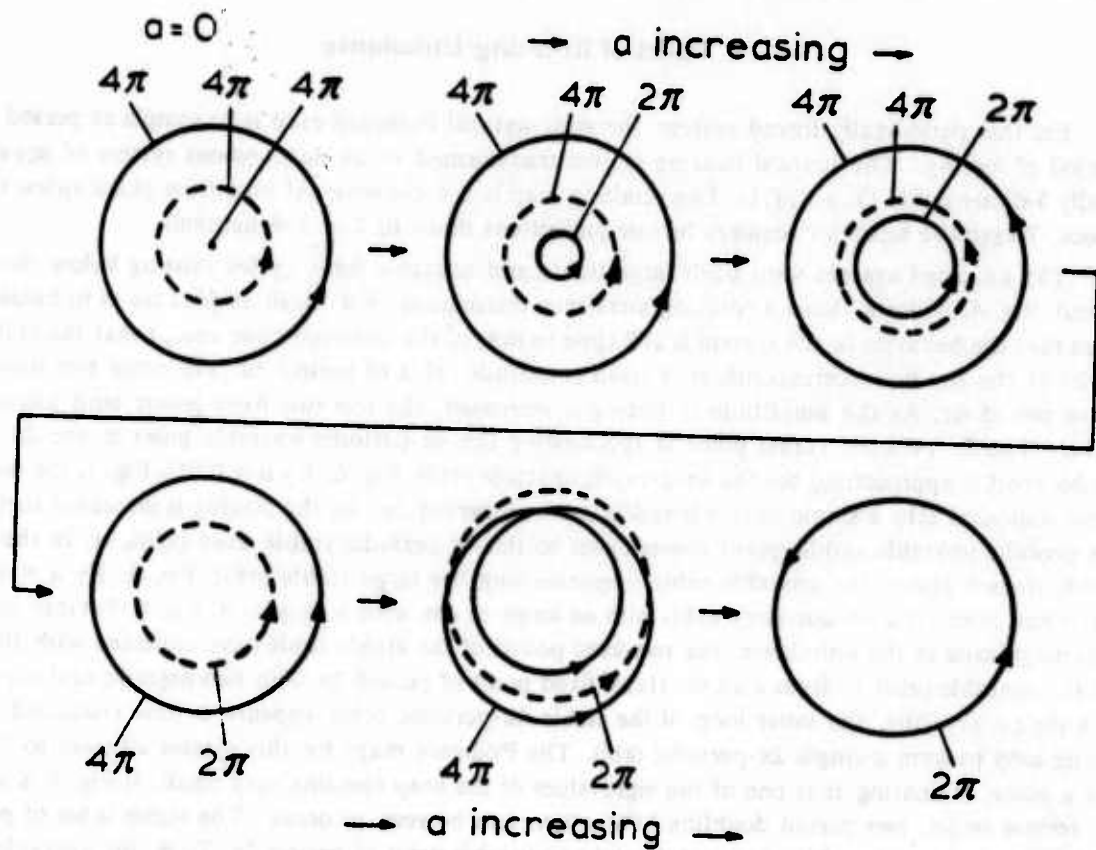
Figure 6. Effect of a on the periodic orbits: $\dfrac{L}{D} = 0$, $\dfrac{m}{\sigma} = 0.5$, $\omega = 2.7$.
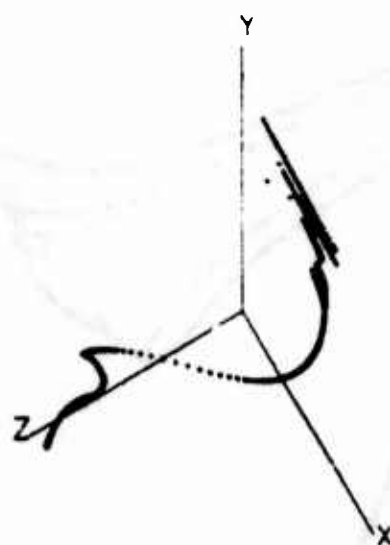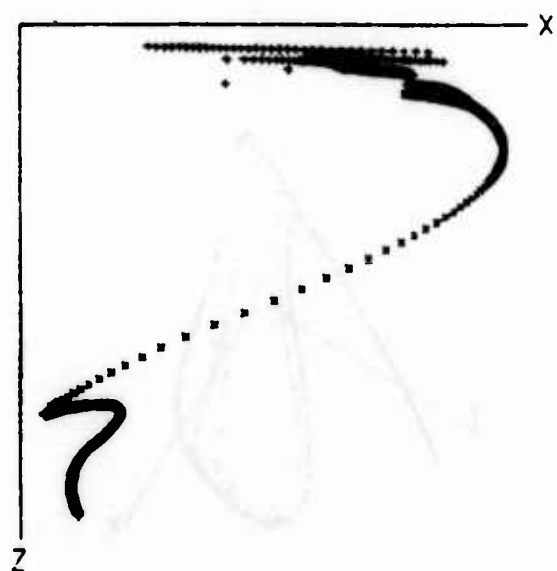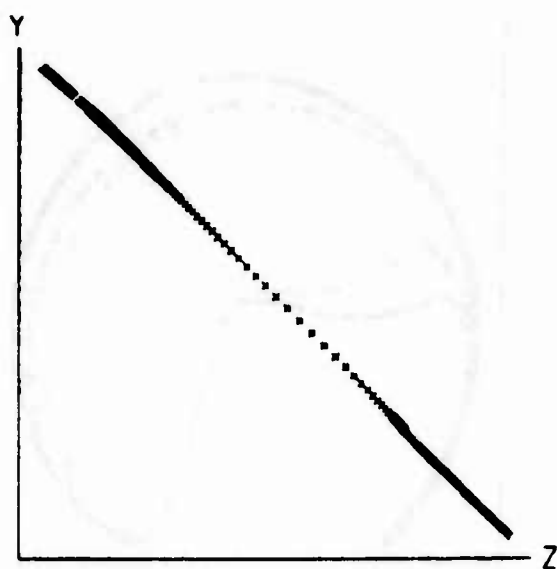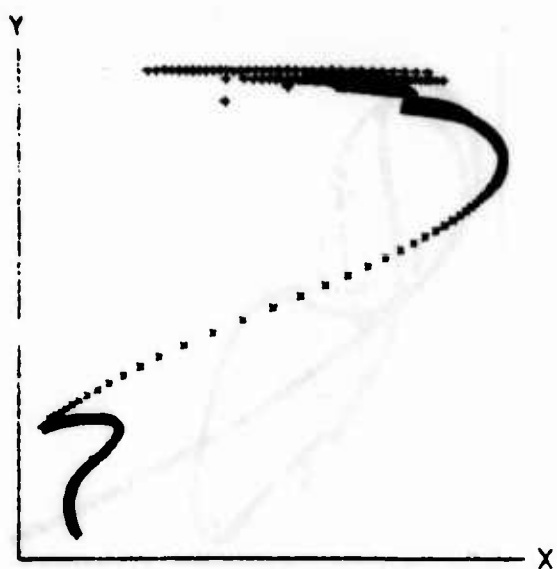
Figure 7. Poincaré map of journal bearing: $a = 0.016$, $\frac{L}{D} = 0$, $\frac{m}{\sigma} = 0.5$, $\omega = 2.7$.
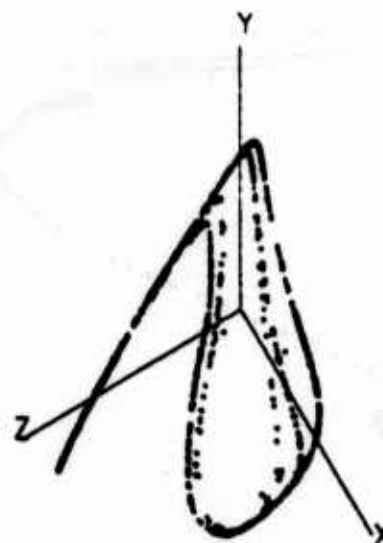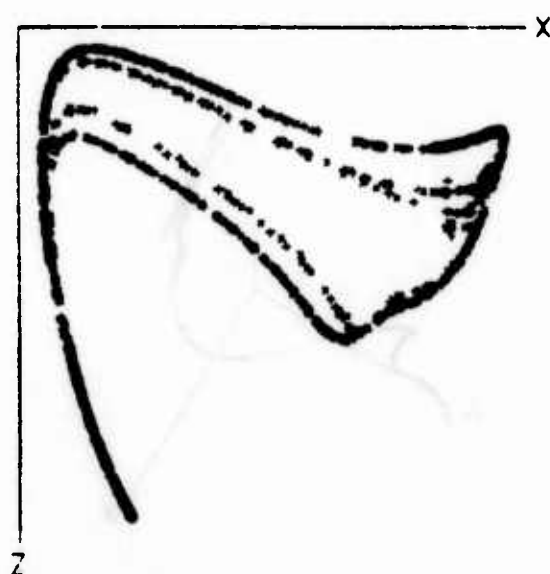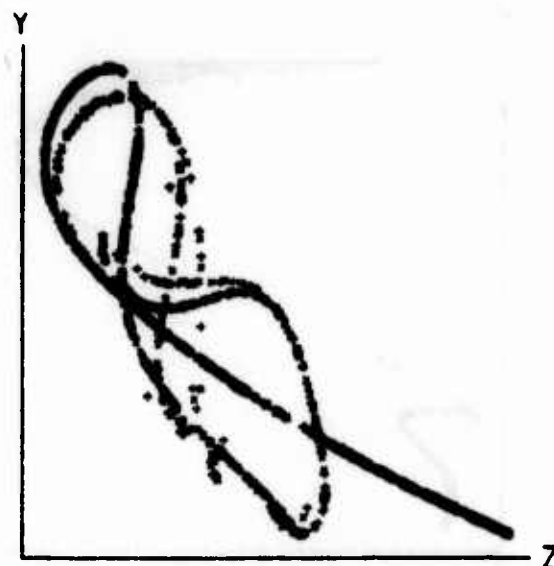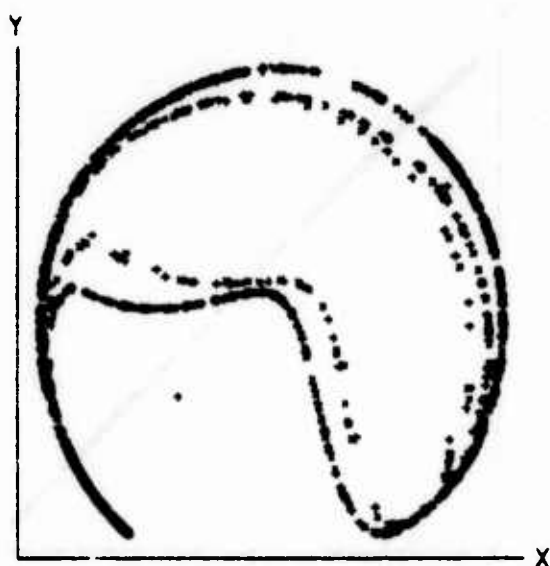
Figure 8. Complex behavior of journal bearing using map with period 2π: $\frac{L}{D} = 0$, $\frac{m}{\sigma} = 500$, $\omega = 3.5$, $a = 0.31$.
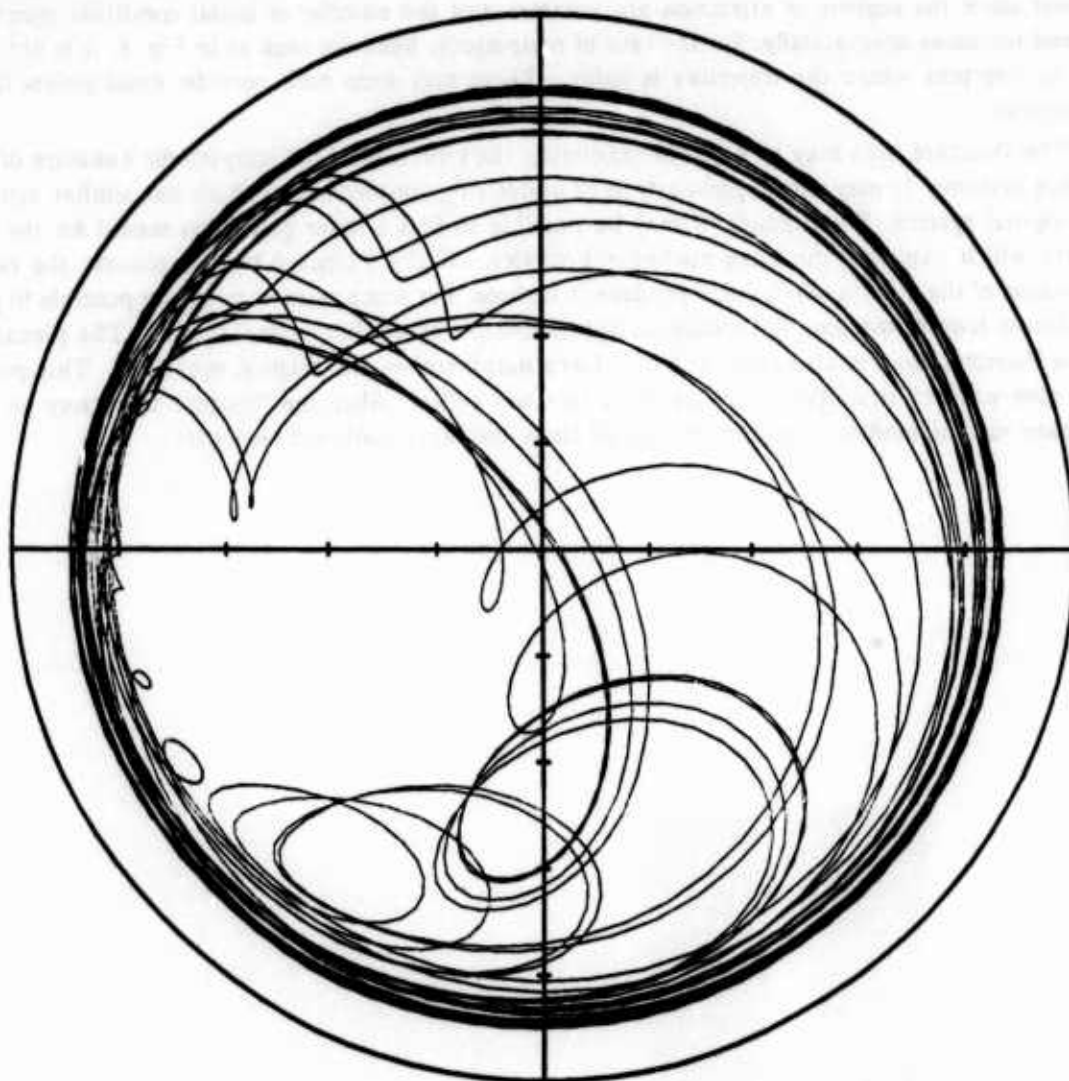
Figure 9. Trajectory from $t = 400$ to $t = 800$.

an extra small loop seems to occasionally appear, grow and disappear again. The frequency content of the trajectory in this time period is shown in Fig. 10. The trajectory is mostly composed of three main frequencies (the forcing frequency 1, and the subharmonics $\frac{1}{3}$ and $\frac{2}{3}$) and some smaller amounts of frequencies between these. This behavior seems to exist only for a narrow range of parameter values. Small changes in a can cause only periodic motion to appear.

Once the fixed points of the Poincaré map are known, it is possible to define their various regions of attraction. For a two dimensional phase space, it is a straightforward task to check where each inital condition point goes and mark it accordingly. In 3 dimensions or higher, the visualization is much more difficult since the regions of attraction are volumes and the number of initial condition points to be checked increases dramatically. For the case of nonperiodic behavior such as in Fig. 8, it is not so clear how to interpret where the trajectory is going. There may even exist periodic fixed points for some trajectories.

The Poincaré map may be useful in examining the structure of the nonperiodic behavior of journal bearing systems. It may also suggest a type of underlying simpler model which has similar behavior to the original system. For example, it may be possible to find a lower dimension model for the bearing system which captures the same nonlinear behavior. If the Poincaré map is planar, the minimum dimension of the system which could produce it is three. For some cases, it might be possible to perform coordinate transformations to produce an approximating three-dimensional system. The planar nature of the Poincaré map is also suggestive of a large parameter in the original equations. This parameter may give ways of simplifying the equations to lower order. Also, the Poincaré map may be used to compare various models to ensure they do all show the same nonlinear responses.

Bibliography

1. Lund, B. W. and Ziske, R., "An Initial-Value Orbit of a Return to Three Dimensional Support in Degenerate to Bodies 65, (1985), 415-431.

2. Bunger, J. C., Keane, P. E. and Flores, A. et al. "Finite Length Bearing Characteristic Equation in Short Bearing Theory," Journal of Lubrication Technology 127, (1967), 292-300.

Figure 10. Frequency content of orbit from $t = 400$ to $t = 800$.

625

Bibliography

1. Lund, J. W. and Saibel, E., "Oil Whip Whirl Orbits of a Rotor in Sleeve Bearings," *Journal of Engineering for Industry* **89**, (1967), 813–823.

2. Barrett, L. E., Allaire, P. E. and Gunter, E. J., "A Finite Length Bearing Correction Factor for Short Bearing Theory," *Journal of Lubrication Technology* **102**, (1980), 283–290.

3. Hollis, P. and Taylor, D. L., "Hopf Bifurcation to Limit Cycles in Fluid Film Bearings," *Journal of Tribology* **108**, (1986), 184–189.

4. Booker, J. F., "Dynamically Loaded Journal Bearings: Mobility Method of Solution," *Journal of Basic Engineering* **87**, (1965), 537–546.

5. Taylor, D. L., "Limit Cycle Analysis of Rotors with Fluid Film Bearings," *Proceedings of the Joint Automatic Control Conference*, San Francisco (1980).

6. Myers, C. J., "Bifurcation Theory Applied to Oil Whirl in Plain Cylindrical Journal Bearings," *Journal of Applied Mechanics* **51**, (1984), 244–250.

7. Guckenheimer, J. and Holmes, P., *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Applied Mathematical Sciences **42**, Springer-Verlag, (1983).

8. Keller, H. B., "Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems," *Applications of Bifurcation Theory*, Rabinowitz, P. H. (Ed), Academic Press, (1977), 359–384.

9. Doedel, E., "A Program for the Automatic Bifurcation Analysis of Autonomous Systems," *Proceedings of the Tenth Conference on Numerical Math and Computation*, (1980), 265–284.

Analytical and Computational Studies
of the Fluid Motion in Liquid-Filled Shells

Thorwald Herbert

Department of Engineering Science and Mechanics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

Spin-stabilized projectiles with liquid payloads can experience a severe flight instability characterized by a rapid yaw-angle growth and a simultaneous loss in spin rate. Laboratory experiments and field tests have shown that this instability originates from the internal fluid motion in the range of small Reynolds numbers. In earlier work, we developed a simple model of this flow based on linearized equations for the deviation from solid-body rotation in an infinite cylinder. Here, we perform a perturbation analysis in order to estimate the effect of nonlinear terms. Beyond a small correction of the axial velocity component, we obtain radial and azimuthal components of the velocity field in agreement with computational results for the core region of a finite-length cylinder. The analytical results are exploited in the design of a spectral Navier-Stokes solver for the steady motion in a finite cylinder. A first raw version of this spectral code provides flow field and pressure distribution in a small fraction of the computer time required by existing codes. We report some results and discuss possible refinements of this code.

1. Introduction

It is well-known that spin-stabilized shells carrying liquid payloads can suffer a dynamical instability which results in an increased coning (or yaw) angle and a simultaneous loss in spin rate. Laboratory experiments, computational results, and field tests indicate that these phenomena arise from the coning-induced fluid motion in a limited range of small Reynolds numbers. Although in special cases this instability has been removed by trial and error, future design of reliable projectiles would profit from the opportunity to estimate the liquid moments, and to include these moments in flight simulators. The empirical data base [1, 2] is sparse, however, and computational methods in use [3, 4, 5] are rather demanding.

Our theoretical analysis of this problem serves on one hand to gain insight into the anatomy of the flow phenomena and to support the ongoing experiments. On the other hand, it promotes our efforts to develop a more efficient code for the numerical simulation of the flow in a finite container. While the analytical work aims at the velocity field in the core region of a sufficiently long cylinder and on the viscous components of the moments, in particular the viscous despin (negative roll) moment, the computational work also captures the flow near the end walls and the pressure contributions to yaw and pitch moments.

Our previous work [6] shows that the deviation from solid body rotation is governed by a small parameter $\epsilon = \Omega \sin\theta / \omega$ involving the nutation rate Ω, the nutation angle θ, and the spin rate ω. The solution of the linearized equations consists of

only an axial component of order $O(\epsilon)$. This axial flow is the dominating feature of the fluid motion and produces a negative roll moment of order $O(\epsilon^2)$ owing to Coriolis forces. Although these results are in reasonable agreement with experimental and computational data, one may anticipate modifications of velocity field and roll moment if nonlinear terms are taken into account. Estimates of these nonlinear effects are desired in order to support previous results and to verify our conclusion that the three-dimensional flow field in a finite-length cylinder is essentially given by the solution of linearized momentum equations. In the following, we perform a straightforward perturbation expansion for the nonlinear problem. We develop and solve the equations for the flow in an infinitely long cylinder up to order $O(\epsilon^3)$. A closed-form solution is given for the radial and azimuthal velocity components at second order. The third-order equations are solved numerically.

The perturbation solution also provides estimates for the number of expansion functions required for accurate spectral representation of the radial (r) and azimuthal (ϕ) structure of the solution. A spectral code appears as an attractive alternative to the existing Navier-Stokes solvers. The finite-difference code developed at Sandia Laboratories [3, 4] exploits Chorin's method of artificial compressibility. The steady solution at $11 \times 24 \times 21$ grid points in r, ϕ, z-direction is obtained by integrating over typically 10^4 time steps, a task that requires 68 minutes of CPU time on an IBM 3090. The result consists of 22,000 plus values for the velocity components v_r, v_ϕ, v_z and the pressure p that can be utilized for a calculation of the moments. Strikwerda & Nagel [5] describe a code using finite differences in radial and axial direction and pseudospectral differencing in the azimuthal direction. Nonuniform grids are introduced for increased resolution near the walls. The difference equations are solved by an iterative method based on successive over-relaxation. The computer time required is comparable to that of the Sandia code (Nusca, BRL, personal communication). Although the relative merits of the two codes, especially with respect to the captured range of Reynolds numbers are yet in the dark, it seems well possible to beat both of these codes in two respects: computer time and adaptability to the unsteady problem.

For a feasibility study, we have pursued a simple concept that is open to numerous refinements. We use Chebyshev-Fourier-Chebyshev expansions in r, ϕ, z, respectively, and convert the linearized equations into a linear algebraic system for the expansion coefficients. The solution of this system (or any other solution at neighboring parameters) is used as initial approximation for iterative improvement by the modified Newton method. The experience with this code is encouraging with respect to accuracy, efficiency, and robustness.

2. Governing Equations

We consider the motion of a fluid of density ρ and viscosity μ in a cylinder of radius a and length $2c$ that rotates with the spin rate ω about its axis of symmetry, the z-axis. We consider the motion with respect to the nutating coordinate system x, y, z. This system is obtained from the inertial system X, Y, Z by a rotation with the nutation angle θ about the axis $Y = y$. Therefore, x is in the Z, z-plane, and this plane rotates about the Z-axis with the nutation rate Ω. The two axes of rotation intersect in the center of mass of the cylinder. We consider $\omega > 0$, Ω, and $0 \leq \theta \leq \pi/2$ as constant.

The fluid motion is governed by the Navier-Stokes equations written in the nutating coordinate system:

$$\rho[\frac{D\mathbf{V}_n}{Dt} + 2\mathbf{\Omega} \times \mathbf{V}_n + \mathbf{\Omega} \times (\mathbf{\Omega} \times \mathbf{r})] = -\nabla P_n + \mu \nabla^2 \mathbf{V}_n , \tag{1a}$$

$$\nabla \cdot \mathbf{V}_n = 0 . \tag{1b}$$

$\mathbf{V}_n$ is the velocity measured in the nutating frame, P_n the pressure, and $\mathbf{r}$ the position vector. Equations (1) are subject to the no-slip and no-penetration conditions at the cylinder walls.

It is convenient [6] to split the velocity and pressure fields according to

$$\mathbf{V}_n = \mathbf{V}_s + \mathbf{V}_d , \quad P_n = P_s + P_d , \tag{2}$$

where $\mathbf{V}_s$, P_s describe the state of pure solid-body rotation, whereas $\mathbf{V}_d$, P_d represent the deviation from solid-body rotation. The deviation $\mathbf{V}_d$ and the reduced pressure P_d are ultimately responsible for the observed flight instability.

The equations for $\mathbf{V}_d$, P_d are written in terms of nondimensional quantities v_d, p_d using a, ω, and ρ for scaling length, time, and mass, respectively. The solution then depends on four nondimensional parameters: aspect ratio $\lambda = c/a$, nutation angle θ, frequency $\tau = \Omega/\omega$, and Reynolds number $R = \rho\omega a^2/\mu$. The aspect ratio enters the solution only through the boundary conditions at the end walls of the cylinder. The boundary conditions on v_d are homogeneous.

In cylindrical coordinates r, ϕ, z, the equations for the nondimensional deviation velocity $\mathbf{v}_d = (v_r, v_\phi, v_z)$ and pressure p_d take the form

$$\frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial v_\phi}{\partial \phi} + \frac{\partial v_z}{\partial z} = 0 , \tag{3a}$$

$$D'v_r - \frac{v_\phi^2}{r} - 2(1 + \tau_z)v_\phi + 2\tau_\phi v_z = \tag{3b}$$

$$-\frac{\partial p_d}{\partial r} + \frac{1}{R}[D''v_r - \frac{v_r}{r^2} - \frac{2}{r^2}\frac{\partial v_\phi}{\partial \phi}] ,$$

$$D'v_\phi + \frac{v_r v_\phi}{r} + 2(1 + \tau_z)v_r - 2\tau_r v_z = \tag{3c}$$

$$-\frac{1}{r}\frac{\partial p_d}{\partial \phi} + \frac{1}{R}[D''v_\phi - \frac{v_\phi}{r^2} + \frac{2}{r^2}\frac{\partial v_r}{\partial \phi}] ,$$

$$D'v_z + 2\tau_r v_\phi - 2\tau_\phi v_r = -\frac{\partial p_d}{\partial z} - 2r\tau_r + \frac{1}{R}D''v_z , \tag{3d}$$

where

$$D' = \frac{\partial}{\partial t} + \frac{\partial}{\partial \phi} + v_r \frac{\partial}{\partial r} + \frac{v_\phi}{r}\frac{\partial}{\partial \phi} + v_z \frac{\partial}{\partial z} ,$$

$$D'' = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2} + \frac{\partial^2}{\partial z^2} ,$$

and

629

$$\tau_r = -\epsilon\cos\phi \,, \quad \tau_\phi = \epsilon\sin\phi \,, \quad \tau_z = \tau\cos\theta \,, \quad \epsilon = \tau\sin\theta \,. \tag{4}$$

The primary effect of nutation is contained in the ϕ-periodic force term $-2\tau\,\tau_r = 2\epsilon r\cos\phi$ in the z-momentum equation (3d). For $\epsilon = 0$, equations (3) support the trivial solution $\mathbf{v}_d \equiv 0$, $p_d \equiv 0$. The system also supports the following symmetries:

$$v_r(r, \phi+\pi, -z) = v_r(r, \phi, z) \tag{5a}$$

$$v_\phi(r, \phi+\pi, -z) = v_\phi(r, \phi, z) \tag{5b}$$

$$v_z(r, \phi+\pi, -z) = -v_z(r, \phi, z) \tag{5c}$$

$$p_d(r, \phi+\pi, -z) = p_d(r, \phi, z) \tag{5d}$$

3. Perturbation analysis for an infinite cylinder

The steady flow in a relatively long cylinder (aspect ratio $\lambda > 4$) at low Reynolds number is expected to have a rather simple structure and to provide a roll moment proportional to Re. In fact, the flow is expected to exhibit little axial variation over much of the cylinder length. Previous work [6] has therefore relaxed the boundary conditions at the end walls. In this way, one seeks the steady flow in a finite segment of an infinitely long cylinder.

In the physical situations of interest, $\epsilon = (\Omega/\omega)\sin\theta$ is a small parameter, $\epsilon \leq 0.06$. Consequently, it seems reasonable to pursue a straightforward perturbation expansion in ϵ. This provides $\mathbf{v}_d$ in the form

$$\mathbf{v}_d = \sum_{n=1}^{\infty} \epsilon^n \mathbf{v}^{(n)}(r, \phi) \tag{6}$$

and similar expressions for p_d.

The development of general expressions for the expansion coefficients $\mathbf{v}^{(n)}$ from equations (3) indicates an alternating pattern: Odd-order terms contain odd multiples of ϕ and contribute only to the axial velocity v_z, while even-order terms contain even multiples of the azimuthal coordinate ϕ and contribute only to the radial velocity v_r and azimuthal velocity v_ϕ. Therefore,

$$\mathbf{v}^{(n)} = \begin{cases} (0, 0, v_z^{(n)}), & n \text{ odd}, \\ (v_r^{(n)}, v_\phi^{(n)}, 0), & n \text{ even}, \end{cases} \tag{7}$$

and the components of $\mathbf{v}^{(n)}$ take the form

$$v_r^{(n)} = \sum_{m=1}^{n/2} \left(u_{nm}(r)\, e^{i2m\phi} + \tilde{u}_{nm}(r)\, e^{-i2m\phi} \right), \tag{8a}$$

$$v_\phi^{(n)} = v_{no}(r) + \sum_{m=1}^{n/2} \left(v_{nm}(r)\, e^{i2m\phi} + \tilde{v}_{nm}(r)\, e^{-i2m\phi} \right), \tag{8b}$$

$$v_z^{(n)} = \sum_{m=1}^{(n+1)/2} \left(w_{nm}(r)\, e^{i(2m-1)\phi} + \tilde{w}_{nm}(r)\, e^{-i(2m-1)\phi} \right), \tag{8c}$$

where the tilde denotes the complex conjugate. The aperiodic term in $v_r^{(n)}$ is suppressed

by the continuity equation. The r-dependent coefficient functions in eqs. (8) are required to satisfy homogeneous boundary conditions at $r = 1$ and to be finite at the axis $r = 0$ for a physically meaningful solution.

At the lowest order $O(\epsilon)$, the z-independent force term in eq. (3d) can be balanced only by an axial component of the deviation velocity. This component is the dominating feature of the flow in a long cylinder. The axial velocity at order $O(\epsilon)$ can be found in analytical form,

$$w_{11}(r) = i\left(\frac{I_1(\alpha r)}{I_1(\alpha)} - r\right), \tag{10}$$

where I_1 is the modified Bessel function, and $\alpha = (1 + i)(R/2)^{1/2}$. This solution is valid for arbitrary Reynolds number but may be unstable as R exceeds some critical value. The properties of the resulting flow field are discussed by Herbert [6].

At higher order, it is convenient to eliminate the pressure for the periodic components by using the vorticity form of eqs. (3). At order $O(\epsilon^2)$, comparison of the equation for v_{20} with the imaginary part of the equation for w_{11} immediately shows that the aperiodic component of the azimuthal velocity is

$$v_{20}(r) = -2\,\text{Im}(w_{11}(r)). \tag{11}$$

This relation can be exploited to show that the despin moment of order $O(\epsilon^2)$ due to shear forces on the cylinder wall is identical with our former result.

The ϕ-periodic components are governed by a coupled set of inhomogeneous differential equations with variable coefficients. Essential simplification at the expense of increasing the order of differentiation results from eliminating v_{21} by use of the continuity equation. With some effort, the radial velocity component of $O(\epsilon^2)$ can be found in closed form,

$$u_{21}(s) = \frac{1}{s}[c_1 J_2(s) + c_2 Y_2(s)] + c_3 s + \frac{c_4}{s^3} + \frac{2i\sqrt{2}}{s}\frac{J_2(s/\sqrt{2})}{J_1(\beta/\sqrt{2})} \tag{12}$$

where $s = \beta r$, $\beta = (i - 1)R^{1/2}$, and J_1, J_2, and Y_2 are Bessel functions. The coefficients c_1, c_2, c_3, and c_4 can be determined numerically.

The velocity components at order $O(\epsilon^3)$ are of interest primarily since w_{31} provides the first nonlinear correction to the despin moment. In view of the effort involved in deriving the closed form solution for u_{21} and the ultimate need to determine the coefficients in eq. (12) numerically, we decided to solve the differential equations for the third-order components by means of a spectral collocation method.

4. Results of the Perturbation Analysis

Detailed equations, results, and graphs of the various functions at relevant Reynolds numbers will be published elsewhere [7]. Here we give only a summary of the main results. The motion is governed by the axial component w_{11} at order $O(\epsilon)$. Of the higher order terms, only the aperiodic term v_{20} is substantial. In the cylinders center section, these terms are in good agreement with results obtained from the Sandia code, and in excellent agreement with our own computations. All the other terms are not only of order $O(1)$ but in fact less than unity, assuring rapid convergence of the perturbation

series. The contribution of w_{31} to the despin moment is negligible. The ϕ-periodic terms oscillate about zero as r varies between $0 \leq r \leq 1$. Accurate representation of single high-order terms by radial Chebyshev series may require numerous expansion functions. For the total velocity field, however, the error in representing these terms is of little importance. At Reynolds numbers in the range of maximum despin moment, reasonably accurate approximations can be obtained with as few as five polynomials in radial direction. In the azimuthal direction, the solution is governed by terms periodic in ϕ, and by the aperiodic term v_{20}. Fourier series with three or five modes, therefore, provide approximations of sufficient accuracy for practical purpose.

5. Spectral Approximations for a Finite-Length Cylinder

The results of the perturbation analysis suggest that a good approximation to the flow in a finite cylinder can be obtained by solving the linearized version of equations (3). Linearization can be performed in different ways. The first is a linearization in ϵ, as in the perturbation analysis. The resulting equations support strong symmetries. Beyond equations (5), the solution satisfies

$$\mathbf{v}_d(r, \phi+\pi, z) = -\mathbf{v}_d(r, \phi, z), \tag{13a}$$

$$p_d(r, \phi+\pi, z) = -p_d(r, \phi, z). \tag{13b}$$

These relations provide a useful check on the results of the spectral code. A second linear system can be obtained by linearization in the components of $\mathbf{v}_d$. This linearization retains coupling terms such as $2\tau_\phi v_z$ in eq. (3b) which destroy the symmetries (13). The second system can be considered a special case of a third linearization about some known solution $\mathbf{v}_d^{(0)}$, $p_d^{(0)}$. The third procedure is very efficient if the solution is sought for a densely spaced sequence of parameter combinations as in flight simulations. The second system is equivalent with the third one for $\mathbf{v}_d^{(0)} = p_d^{(0)} = 0$.

The algebraic form of the equations is obtained by use of spectral collocation. The velocity components are expressed in the form

$$v_r = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{m=1}^{M} u_{klm} R_{kl}^u(r) F_l(\phi) Z_m(z/\lambda) \tag{14}$$

with similar expressions for v_ϕ, v_z, and p_d. The azimuthal functions are $F_l = \cos[(l-1)\phi/2]$ for odd l, $F_l = \sin[l\phi/2]$ for even l, where $l = 1, 2, \cdots L$, and L is odd. The azimuthal collocation points are equidistant, $\phi_l = 2\pi(l-1)/L$. If no use of the symmetries (5) is made, the axial expansion functions are the Chebyshev polynomials $Z_m = T_{m-1}(z/\lambda)$, $m = 1, 2, \cdots M$. The collocation points are $z_m/\lambda = \cos[(m-1)\pi/(M-1)]$. In radial direction, even or odd Chebyshev polynomials are used, depending on the quantity under consideration and the periodicity in ϕ. The proper choice is dictated by the requirement of a unique value of all quantities on the axis $r = 0$. For example, the axial velocity component must assume a unique value independent of ϕ as $r \rightarrow 0$. Therefore, even polynomials are to be used if $l = 1$ while odd polynomials are to be used if $l > 1$. The radial collocation points are $r_k = \cos[(k-1)\pi/(2K-1)]$, $k = 1, 2, \cdots K$. Consequently, $0 < r_k \leq 1$, and no difficulty can arise from points on the axis. The collocation points in radial and axial direction are concentrated near the boundary such that high resolution in this region is obtained without additional coordinate transformations.

Our implementation of the spectral method uses precalculated and stored matrices containing the values of the expansion functions and their derivatives at the collocation points. It is a straightforward matter to convert the linear system of partial differential equations derived from eqs. (3) into an algebraic system of dimension $N = 4 K L M$ for the coefficients u_{klm}, v_{klm}, w_{klm}, and p_{klm} for v_r, v_ϕ, v_z, and p_d, respectively. It is not straightforward, however, to implement the homogeneous boundary conditions for the velocities at the cylinder wall and the condition on the pressure that is only determined to within an additive constant. In principle, the boundary conditions are implemented by replacing three of the four differential equations in the boundary points. The question then is which equation should be retained and where the condition on the pressure, e.g. $p_d = 0$, should be applied. Trial-and-error leads to numerous cases with ill-determined matrices or zero determinant. In other cases, a correct solution for the velocity field is obtained, but the pressure contains a non-physical spurious term. With the velocity field given, we attempted to calculate the pressure by solving a Poisson equation with von Neumann boundary conditions, but we encountered the same difficulties. Problems with calculating the pressure in closed domains with spectral methods are well-known, e.g. [8]. However, the reports of negative results are rather unspecific, and neither the origin nor methods for removal of this spurious term seem to be known.

We have therefore performed a detailed analysis of the flow in a square driven by an internal force field. This simpler two-dimensional problem exhibits all characteristics - including the spurious pressure term - of the original problem. Detailed results of this study will be reported elsewhere [9]. The study reveals that the spurious term is associated with the corners of the domain. The term vanishes in all collocation points except the corners, where it may assume arbitrary values. The term can be suppressed by retaining in the corners one of the momentum equations that contain the derivative of the pressure in the direction of the boundary. In the cylinder problem, the z-momentum must be retained in order to suppress even as well as odd spurious terms. The condition on the pressure can be applied anywhere except in the corner points.

We solve the linear algebraic system for the expansion coefficients with a special subroutine based on Gauss elimination with partial pivoting. The subroutine stores all data required to solve the same system with a new right-hand side without repeating the costly ($O(N^3)$ operations) reduction of the matrix to upper triangular form. Once the solution is obtained, a new right-hand side is formed taking the nonlinear terms into account and the system is solved again. This procedure is iteratively repeated until sufficient accuracy is obtained. The procedure is equivalent to the modified Newton iteration (without updating the Jacobian in every step) and converges rapidly since the nonlinear corrections to the velocity are small while the pressure appears linear in equations (3).

6. Results of the Spectral Code

In the following, we present some preliminary results of a test run for $R = 14.95$, $\theta = 20^\circ$, $\tau = 0.1667$, and $\lambda = 4.368$ which results in $\epsilon = 0.057$. The results are for $K = 4$, $L = M = 5$, and consequently $N = 400$. Detailed convergence tests will be performed with later versions of the spectral code. Figure 1 shows the axial and radial velocity in the x, z-plane. Only the upper half, $z \geq 0$, of the cylinder is shown; the lower half is governed by the symmetries (5). The velocity distribution at $z = 0$ agrees

well with the results of the perturbation analysis and computations with the Sandia code. Near the walls, the solution seems to be more realistic and more accurate than the Sandia results. The figure also verifies the existence of a predominantly axial flow over most of the cylinder length, except within a region of the order of the radius near the end wall. Linear and nonlinear velocity distributions are hardly distinguishable. Clearly visible is the turning of the flow near the end wall. The radial and azimuthal velocities at $z = 0.9\lambda$ are shown in figure 2. The right tick mark indicates the x-direction, $\phi = 0$. At Re = 14.95, the maximum of the axial velocity occurs at $\phi \approx 45$ °.

Pressure distributions in the x, z-plane are given in figures 3 and 4 with the heavy lines indicating positive values. The pressure in figure 3 is obtained simultaneously with v_d from equations linearized in ϵ, and clearly shows the symmetry (13b). Figure 4 gives the result from the nonlinear equations. It is interesting to note that a very similar pressure field can be obtained by solving the Poisson equation for the pressure with the linear velocity field. The inhomogeneous term in the Poisson equation is inherently nonlinear in the velocities. Figure 5 gives the pressure distribution across the cylinder near the end wall at $z = 0.9\lambda$. Remarkable is the formation of a high-pressure region in the corner near $\phi = 0$, which produces a large moment about the y-axis. Looking at a series of plots like figures 4 and 5, one may wonder whether the details of the pressure variation near the cylinder wall can be resolved with a finite difference approximation with a step size of $\Delta r = 0.1$.

The azimuthal mean velocity at $z = 0$ is shown in figure 6. The shear exerted by this component on the cylinder wall opposes the spinning motion and is the ultimate cause of the despin moment. The axial and radial mean velocity field is given in figure 7. This streaming term exhibits a toroidal motion stretched over each half of the cylinder. It is this mean velocity that causes the symmetric pattern in flow visualizations [10]. Figure 8 shows the observed pattern of the flow at $R \approx 30$ which is typical for the range of low Reynolds numbers.

7. Discussion

The experience with the first version of the spectral code shows that high performance can be achieved. The reported run with $N = 400$ requires 1.3 minutes CPU time on an IBM 3090, 48 minutes on an Apollo DN300 desktop computer. The solution is obtained in semi-analytical form with only $N = 400$ numerical coefficients. This low data volume is especially attractive for communication with remote supercomputers. The code is very well suited for vectorization, since practically all CPU time is spent on constructing and solving an algebraic system. However, the code demands larger memory than other codes [3, 5]. Since 64-bit arithmetic is highly recommended for spectral methods in general, and the algebraic system requires $N(N + 1)$ words of storage, the above test requires 1.3 Mbyte of memory. Nowadays, the memory requirement appears acceptable even if higher resolution is desired.

Finally, there are various ways to improve the performance and lessen the demands. The first step is to exploit symmetry which reduces N by a factor of 1/2, storage by 1/4, and time by $\approx 1/8$. Second, the solution process can be split into two levels, the first of which calculates only the velocity components while the pressure is obtained a posteriori by solving the Poisson equation. After these changes, the above test run will require less than 1 minute on an MC68020/68881 based desktop computer.

Alternatively, runs with higher resolution can be executed within a short time on super-computers. One may also consider reducing the storage requirement by line iteration. However, the ability to obtain a reasonably acurate solution by direct solution of the (large) algebraic system bears valuable potential to answer the question whether the steady solution is stable, and allows for analysis of unsteady motions. The design of a reliable code for the unsteady problem can take profit from the kowledge of the eigen-value spectrum for small unsteady disturbances of the steady flow.

ACKNOWLEDGMENT

REFERENCES

1. M. C. Miller 1982 "Flight instabilities of spinning projectiles having nonrigid pay-loads," *J. Guidance, Control, and Dynamics*, vol. 5, pp. 151-157.

2. R. Sedney 1985 "A Survey of the fluid dynamic aspects of liquid-filled projec-tiles," AIAA Paper No. 85-1822-CP.

3. H. R. Vaughn, W. L. Oberkampf, and W. P. Wolfe 1983 "Numerical solution for a spinning nutating fluid-filled cylinder," Sandia Report SAND 83-1789.

4. H. R. Vaughn, W. L. Oberkampf, and W. P. Wolfe 1985 "Fluid motion inside a spinning nutating cylinder," *J. Fluid Mech.*, vol. 150, pp. 121-138.

5. J. C. Strikwerda and Y. M. Nagel 1985 "A numerical method for computing the flow in rotating and coning fluid-filled cylinders," in *Proc. 1984 Scientific Conf. on Chemical Defense Research, Aberdeen Proving Ground, Maryland,* ed. M. Rausa, pp. 523-527, CRDC-SP-85006.

6. Th. Herbert 1986 "Viscous fluid motion in a spinning and nutating cylinder," *J. Fluid Mech.*, vol. 167, pp. 181-198.

7. Th. Herbert and S. D. Greco 1986 "Higher approximations for the viscous flow in a spinning and nutating cylinder," *J. Fluid Mech.* To be submitted.

8. R. Peyret and T. D. Taylor 1983 *Computational Methods for Fluid Flow*, p. 236, Springer-Verlag.

9. Th. Herbert 1986 "On the spurious pressure in spectral computations of flows in closed domains," *J. Comp. Phys.* To be submitted.

10. Th. Herbert and D. Pierpont 1986 "Visualization of the flow in a spinning and nutating cylinder," in *Proc. 1985 Scientific Conf. on Chemical Defense Research*, ed. M. Rausa, pp. 989-994, Aberdeen Proving Ground, Maryland.
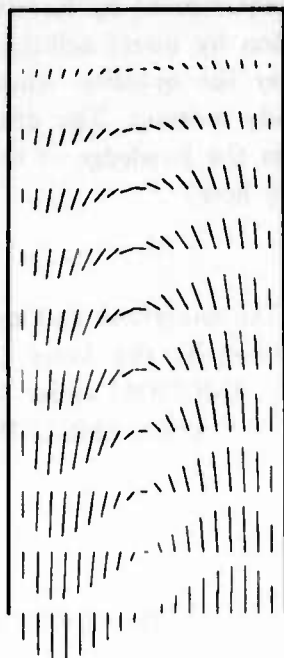
Figure 1. Vector plot of the velocity field
in the x, z-plane for $z \geq 0$.
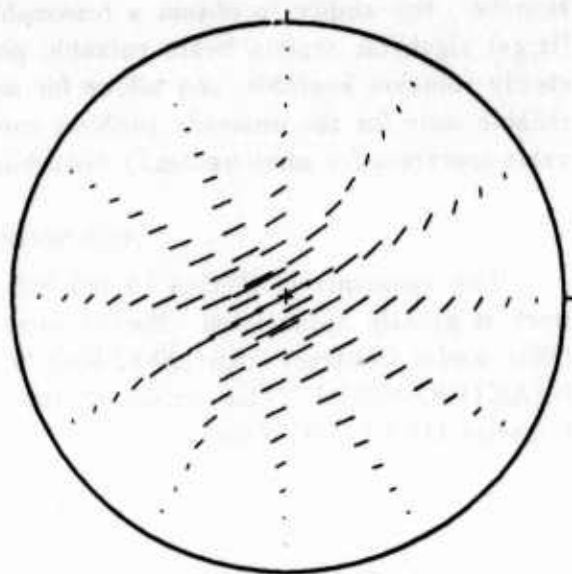
Figure 2. Vector plot of the velocity field
across the cylinder at $z/\lambda = 0.9$.
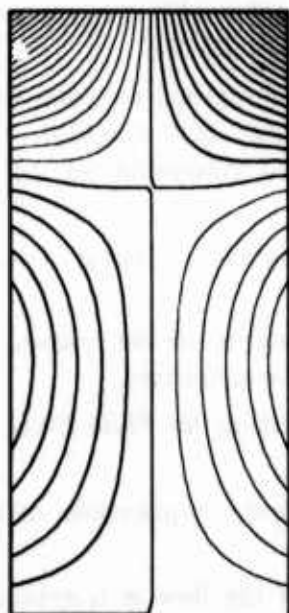
Figure 3. Contour plot of the linear
pressure field in the x, z-plane for
$z \geq 0$.

Figure 4. Contour plot of the nonlinear
pressure field in the x, z-plane for
$z \geq 0$.
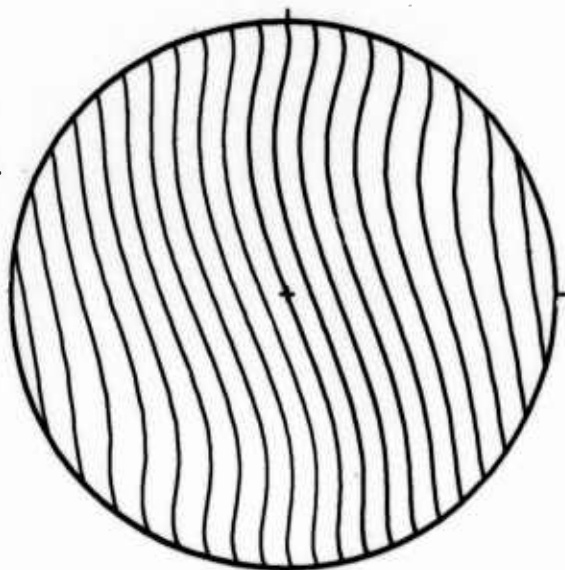
Figure 5. Contour plot of the pressure field across the cylinder at $z/\lambda = 0.9$.
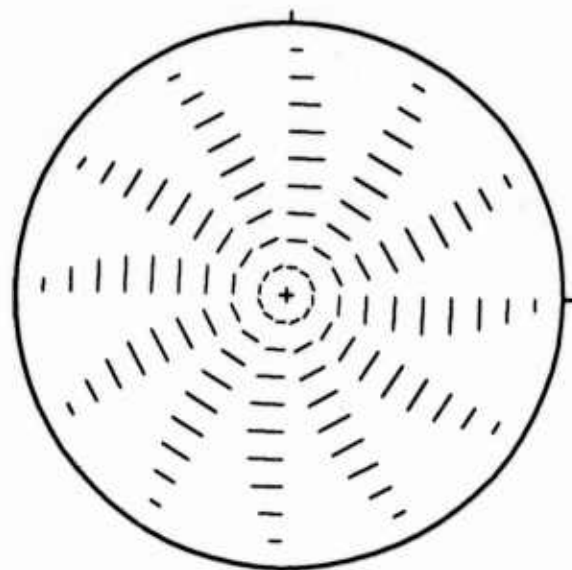
Figure 6. Vector plot of the mean velocity field across the cylinder at $z/\lambda = 0$.
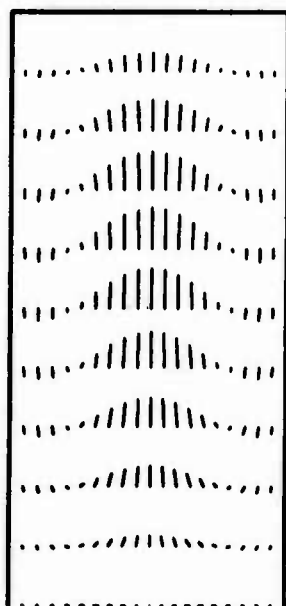
Figure 7. Vector plot of the mean velocity field in the x, z-plane for $z \geq 0$.

Figure 8. Pattern of the fluid motion at low Reynolds numbers ($R \approx 30$) in the x, z-plane.

THE EVOLUTION OF SUBHARMONIC EDGE WAVEPACKETS ON A SLOPING BEACH *

T.R. Akylas & S. Knopping

Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA   02139

Abstract

A study is made of the temporal and spatial development of subharmonic edge packets which are resonantly excited by wavetrains normally incident and reflected on a mildly sloping beach. The nonlinear evolution equations of the wavepacket envelopes are solved numerically for a variety of initial conditions. It is found that, under certain conditions, large-scale modulations of edge waves develop and undergo a recurrence phenomenon. These findings support the view that large-scale modulations of edge waves may account for certain longshore phenomena observed on natural beaches.

A UNIFIED APPROACH TO MASS PROPERTY COMPUTATIONS
IN A SOLID MODELING ENVIRONMENT
WITH APPLICATION TO HYDRAULIC STRUCTURES

Fred T. Tracy
Information Technology Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. Several of the practicing engineers in the Corps of Engineers were doing overturning and sliding analyses of hydraulic structures by hand before a Three-Dimensional Stability Analysis/Design (3DSAD) computer program was developed for doing this. A general purpose approach with solid modeling type capability was taken for describing the geometry and loads so that volume, weight, and centroid information (mass properties), as well as resulting forces and moments, are a natural biproduct (and sometimes the more important product) of the program. In the development of 3DSAD certain algorithms for computing areas and volumes for curved shapes in a concise, consistent, and very computationally efficient way were developed. This paper will describe these algorithms and demonstrate their use for specific hydraulic structures (dams, locks, cooling towers, etc.).

I. INTRODUCTION. A Three-Dimensional Stability Analysis/Design (3DSAD) program [1, 2, 3] for analyzing and designing concrete structures has been developed and successfully used. This program models the geometry and loads in a general way and then applies this modeling capability to specific structure types when possible. Examples include dams with overflow, nonoverflow, and pier sections; gravity and U-frame locks; cooling towers; etc. If the structure does not conform to a predefined standard shape, the program can still do a general analysis of the problem because of the fundamental approach of using solid modeling type capability as a foundation.

There are several ways to model geometry and many commercial packages exist for this [4]. The approach taken in 3DSAD was to provide some of this capability internal to the program and then at a later date supply the capability to "hook" 3DSAD with the larger systems.

II. MODELING TECHNIQUE. 3DSAD models geometry three ways:

a. Blocks - 2-D cross-sections swept normal to the cross-section in either a linear or axisymmetric way with linear and quadratic tapering for the linear sweeps.

    <u>b</u>.   Eight node brick elements.

    <u>c</u>.   A boundary representation consisting of planar faces and bicubic patches.

    <u>III. MASS PROPERTIES - BLOCK</u>. One of the major points of this paper is to describe the equations for performing mass properties for blocks. The blocks are formulated by first defining a cross-section and then parametrically doing a constant or tapered linear sweep or an axisymmetric sweep. Thus, area integrals can first be computed, and from these the mass properties can be done by performing the integration in the third dimension.

    For a cross-section defined by a polygon it is best to convert the area integrals to line integrals [5]. For a cross-section consisting of both curved and straight line segments, it is tempting to first approximate the curved edges with straight line segments and then do the line integrals. However, this process is time consuming and creates unnecessary errors, since it is possible to formulate line integrals for the curved edges as well. This paper will now present examples of these line integrals developed in a consistent manner.

    The volume of a block with constant cross-section (Figure 1)
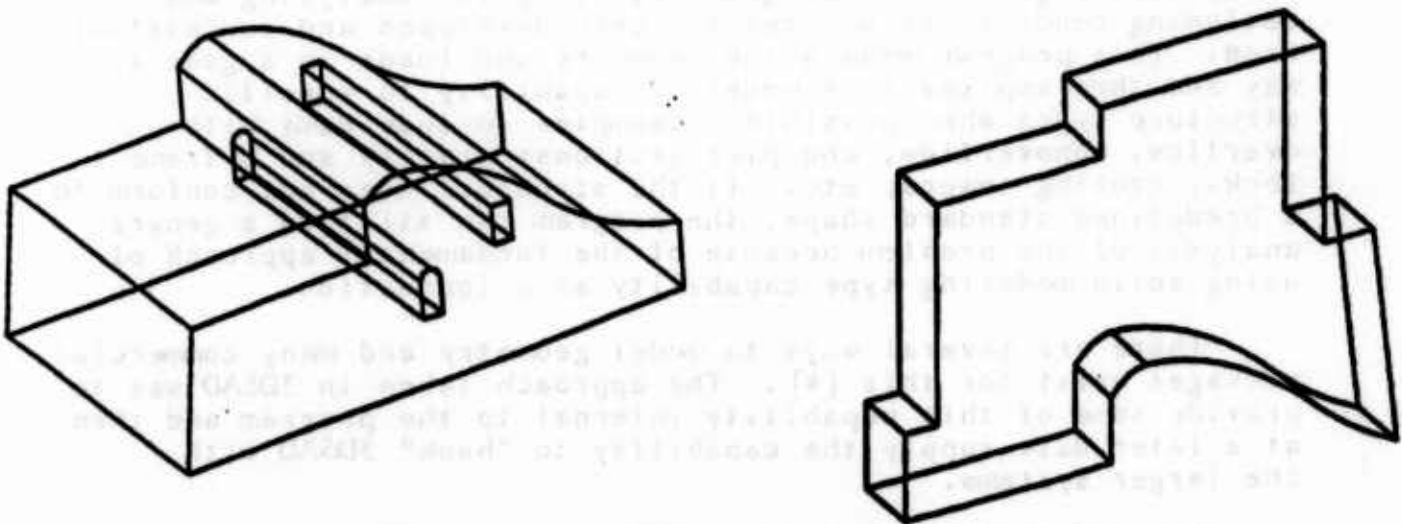
Figure 1. Blocks with constant cross-section

642

having j sides can be expressed as follows:

$$V = - L \int_C y\,dx$$

$$= - L \sum_{j=1}^{n} \int_{C_j} y\,dx$$

$$= - L \sum_{j=1}^{n} I_j$$

The integral I can now be evaluated for different line segment types.

Linear line segment. Suppose the line segment is defined by two points $(X_1, Y_1) - (X_2, Y_2)$. Then

$$X = X_1 + (X_2 - X_1) S$$

$$Y = Y_1 + (Y_2 - Y_1) S$$

where S varies between 0 and 1. I can be done as follows:

$$I = \int_0^1 [Y_1 + (Y_2 - Y_1) S] (X_2 - X_1)\,dS$$

$$= \int_0^1 A_1 \sum_{k=0}^{1} B_k S^k\,dS$$

$$= A_1 \sum_{k=0}^{1} \frac{B_k}{k + 1}$$

<u>Quadratic line segment.</u> In a similar manner a quadratic
line segment formed by three points 1, 2, and 3 can be expressed
as follows:

$$X = [X_1 \ X_2 \ X_3] \ \underline{M} \begin{bmatrix} 1 \\ S \\ S^2 \end{bmatrix}$$

$$= \sum_{k=0}^{2} A_k \ S^k$$

$$Y = \sum_{L=0}^{2} B_L \ S^L$$

$$dx = \sum_{k=0}^{1} (k + 1) \ A_{k + 1} \ S^k \ dS$$

Here M is the "magic" matrix of known constants, and as in the
linear case, the A's and B's are constants of the parametric
polynomials. The line integral I can now be easily evaluated as
follows:

$$I = \int_0^1 \left(\sum_{L=0}^{2} B_L \ S^L \right) \left(\sum_{k=0}^{1} (k + 1) \ A_{k + 1} \ S^k \right) dS$$

$$= \sum_{L=0}^{2} \sum_{k=0}^{1} \frac{k + 1}{k + L + 1} \ A_{k + 1} \ B_L$$

Note also at this point that higher order integrals, say for the x centroid, can be done with equal ease.

$$\bar{X} = -\frac{L}{V} \int_C XYdX$$

$$I = \sum_{L=0}^{2} \sum_{M=0}^{2} \sum_{k=0}^{1} \frac{k+1}{k+L+M+1} A_L B_M A_{k+1}$$

The general trend is now set with the form being the same for any polynomial. Also, another x or y inside the integral simply results in another summation sign.

Circular arc. The circular arc can also be put in the same consistent notation with use of complex variables as follows:

$$X = R \cos (AS + \theta_0)$$

$$Y = R \sin (AS + \theta_0)$$

$$W = e^{iAS}$$

$$X = \sum_{k=-1}^{1} A_k W^k$$

$$Y = \sum_{L=-1}^{1} B_L W^L$$

$$dX = iA \sum_{K=-1}^{1} K A_k W^k dS$$

With this foundation the line integrals can be done as easily as the polynomials.

$$I = \int_0^1 \left(\sum_{L=-1}^1 B_L W^L \right) \left(iA \sum_{K=-1}^1 K A_k W^k \right) dS$$

$$= \sum_{L=-1}^1 \sum_{K=-1}^1 \frac{K}{K+L} \left(e^{iA(k+1)} - 1 \right)$$

Elliptical arc. An elliptical arc can also be handled in the same consistent manner.

$$X = A \cos(\theta S + \theta_0)$$

$$Y = B \sin(\theta S + \theta_0)$$

$$X = \sum_{K=-1}^1 A_K W^K$$

$$Y = \sum_{L=-1}^1 B_L W^L$$

Tapered block. Consider now a cross-section defined in the x-y plane. Let (x_0, y_0) be a point on that cross-section. The (x, y) value of a point on a cross-section at elevation z as a result of a taper is

$$x(z) = (x_0 - x_A) S_x(z) + x_A$$

$$y(z) = (y_0 - y_A) S_y(z) + y_A \quad .$$

646

where (x_A, y_A) is an apex point, and $S_x(z)$ and $S_y(z)$ are scale factors that range in value from zero to one. The scale factor functions determine the type of tapering. Let z vary from zero to L and let the scale factor functions vary quadratically as follows:

$$Z = L S$$

$$S_x = \begin{bmatrix} 1 & H_x & F_x \end{bmatrix} \overset{M}{=} \begin{bmatrix} 1 \\ S \\ S^2 \end{bmatrix}$$

$$= \sum_{j=0}^{2} A_j S^j$$

$$S_y = \sum_{k=0}^{2} B_K S^K$$

where H_x and F_x are the values of the x scale factor, respectively, for s=.5 and s=1. Let a similar expression for the y scale factor also be defined. Let A be the area of a cross-section at elevation z. Then the volume is computed by

$$A = A_0 S_X S_Y$$

$$V = \int_0^L A \ dZ$$

$$= A_0 \int_0^1 (\sum_{j=0}^{2} A_j S^j) (\sum_{k=0}^{2} B_k S^k) L \ dS$$

$$= A_0 L \sum_{j=0}^{2} \sum_{k=0}^{2} \frac{1}{j+k+1} A_j B_k$$

IV. MASS PROPERTIES - FACES. Two types of faces will be considered:

    a. Planar face.

    b. Bicubic patch.

Planar face. By using the equation of the plane, the mass property integrals for planar faces can also be converted to line integrals and then computed the same way as before. For example, the volume under a planar face can be computed as follows:

$$V = \int_A z \, dx \, dy$$

$$= \int_A (Ax + By + C) \, dx \, dy$$

$$= -\int_C (Ax + .5By + C) \, y \, dx$$

These volumes when summed over all the faces will give the correct volume for the solid.

Bicubic patch. Bicubic patches can be defined in several ways. Whatever the boundary conditions or formulation, they can typically be cast as follows:

$$x = \sum_{i=0}^{3} \sum_{j=0}^{3} A_{ij} \, s^i \, t^j$$

$$y = \sum_{k=0}^{3} \sum_{l=0}^{3} B_{kl} \, s^k \, t^l$$

$$z = \sum_{m=0}^{3} \sum_{n=0}^{3} C_{mn} \, s^m \, t^n$$

Mass properties can be done with numerical integration or exactly. The exact formulation is

$$V = \int_A z \, dx \, dy$$

$$= \int_0^1 \int_0^1 z \, |J| \, dS \, dt$$

$$= \sum_{i=0}^{3} \sum_{j=0}^{3} \sum_{k=0}^{3} \sum_{l=0}^{3} \sum_{m=0}^{3} \sum_{n=0}^{3} T_{ijklmn}$$

$$T_{ijklmn} = \frac{(il-jk)A_{ij} \, B_{kl} \, C_{mn}}{(i+k+m) \, (j+l+n)}$$

.

V. LOADS. Loads are presently performed using point loads and volumes with assigned directions (called "pressure-volumes"). The preferred method is to integrate the pressures over the surfaces [6] as follows:

$$\vec{F} = -\int_S p d\vec{S}$$

$$\vec{M} = -\int_S \vec{r} \times p d\vec{S}$$

Here F is the force, M is the moment, and p is the pressure for a given face.

649

VI. EXAMPLES. Shown below are a few examples of the use of the geometry building features described in this paper.
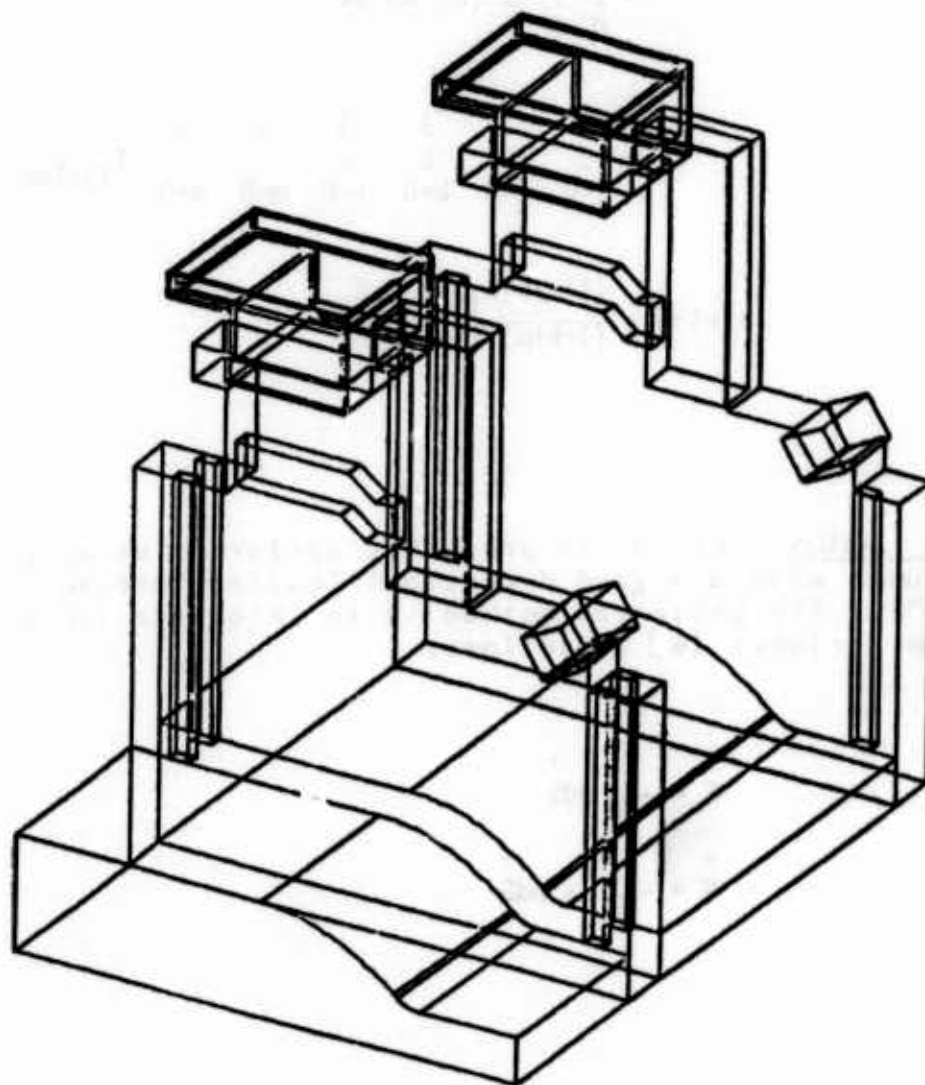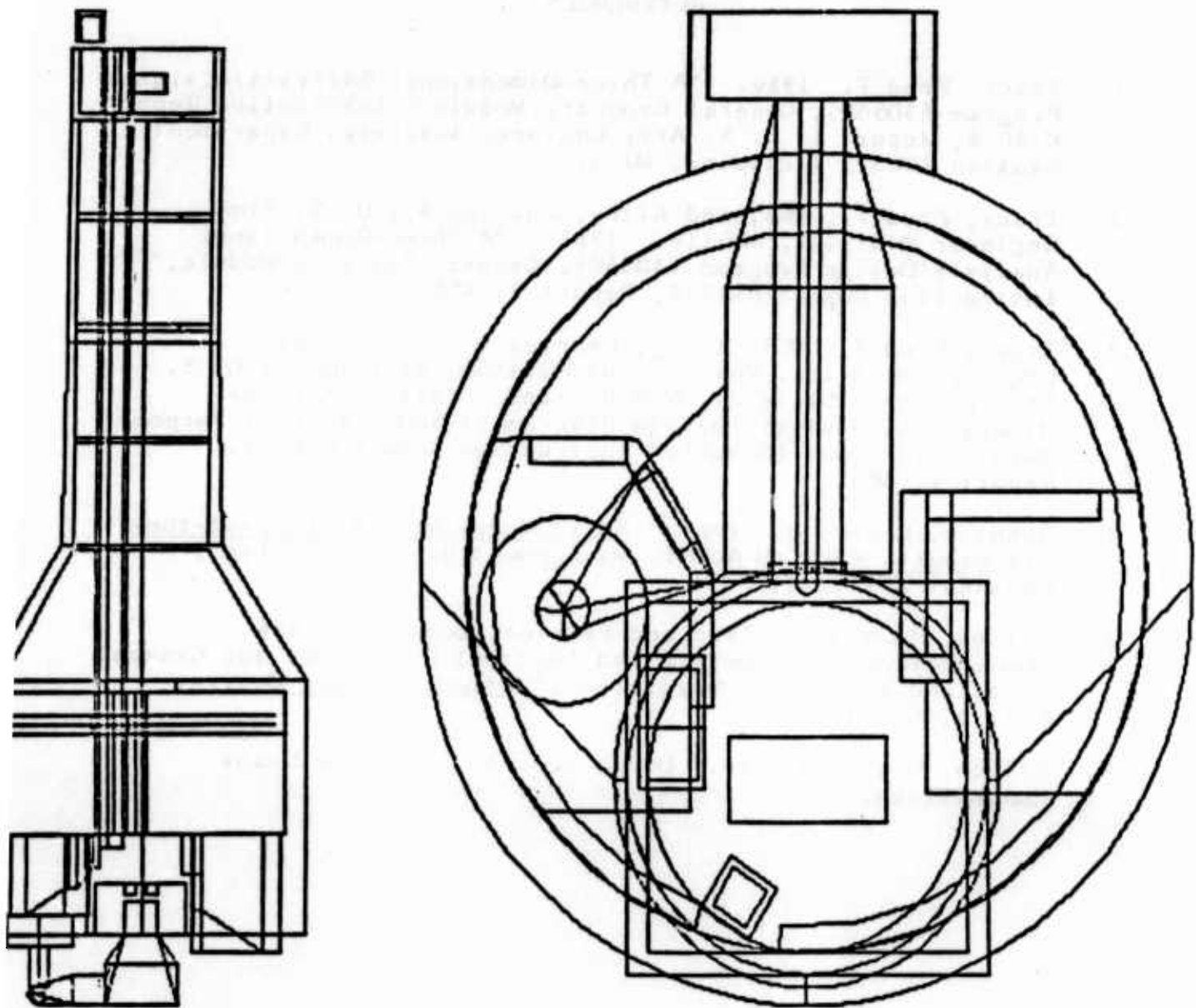
VII. CONCLUSION. The methods that have been presented are an efficient, consistent way of computing mass properties of solids The program 3DSAD incorporating these techniques has proven very useful and successful to the U. S. Army Corps of Engineers.

REFERENCES

1. Tracy, Fred T. 1980. "A Three-Dimensional Analysis/Design Program (3DSAD), General Geometry Module," Instruction Report K-80-4, Report 1, U. S. Army Engineer Waterways Experiment Station (WES), Vicksburg, Miss.

2. Tracy, Fred T., WES, and Kling, Charles W., U. S. Army Engineer District, Mobile. 1982. "A Three-Dimensional Analysis/Design Program (3DSAD), General Analysis Module," Instruction Report K-80-4, Report 3, WES.

3. Tracy, Fred T., WES, Kling, Charles W., U. S. Army Engineer District, Mobile, and Holtham, William J., U. S. Army Engineer Division, New England. 1983. "A Three-Dimensional Analysis/Design Program (3DSAD), Special Purpose Modules for Dams (CDAMS)," Instruction Report K-80-4, Report 4, WES.

4. Johnson, Robert H. 1984. Solid Modeling: A State-of-the-Art Report, CAD/CAM ALERT, Management Roundtable, Inc., Chestnut Hill, Mass.

5. Wilson, Howard B., Jr., and Farrior, Donna S. 1976. "Computation of Geometric and Inertial Properties for General Areas and Volumes of Revolution," Computer Aided Design, Vol. 8, No. 4.

6. Wilson, Howard B., Jr. Draft report. "Surface Loads Computation."

A COMMONSENSE THEORY OF NONMONOTONICITY

Frank M. Brown
Department of Computer Science
University of Kansas
Lawrence, Kansas

ABSTRACT

A commonsense theory of nonmonotonic reasoning is presented which models our intuitive ability to reason about defaults. The concepts of this theory do not involve mathematical fixed points, but instead are explicitly defined in a monotonic modal quantificational logic which captures the modal notion of logical truth. The axioms and inference rules of this modal logic are described therein along with the derivation of the basic theorems about nonmonotonic reasoning. A comparison to the "fixed point" theories of nonmonotonicity is made and some simple applications to deontic nonmonotonic reasoning and the frame problem in robot plan formation are presented.

> "For this formula game is carried out according to certain definite rules, in which the technique of our thinking is expressed. These rules form a closed system that can be discovered and definitively stated. The fundamental idea of my proof theory is none other than to describe the activity of our understanding, to make a protocol of the rules according to which our thinking actually proceeds."
>
> [David Hilbert 1927]

1. INTRODUCTION

The literature on the nature and representation of nonmonotonicity is full of disputes and contradictory theories. This is surprising because the nature of nonmonotonic reasoning does not cause any worry for people in their everyday coping with the world. For example, without any worry people are quite happy to assert the default that birds fly even though penguins are birds which do not fly. This suggests that there is some form of commonsense knowledge about nonmonotonicity that is rich enough to enable people to deal with the world and which is universal enough to enable cooperation and communication between people. In this paper we propose such a theory.

The basic idea of our theory of nonmonotonicity is that nonmonotonicity is already encompassed in the normal intentional logic of everyday commonsense reasoning and can be explained precisely in that terminology.

For example, a knowledgebase consisting of a simple default axiom expressing that a particular bird flies whenever that bird flies is possible with respect to what is assumed is stated as:

(that which is assumed is
    (if (A is possible with respect to what is assumed) then A))

where A stands for the proposition that that particular bird flies.

Reflection on the meaning of this knowledgebase leads immediately to the conclusion that either A is logically possible and the knowledgebase is synonymous to A, or A is not logically possible and the knowledgebase is synonymous to logical truth. This conclusion is obtained by simple case analysis: for if A is possible with respect to what is assumed then, since truth implies A is A, that which is assumed is indeed A. Since that which is assumed is A, A is possible with respect to what is assumed only if A is logically possible. On the other hand, if A is not possible with respect to what is assumed only if A is not logically possible.

Thus if it is further assumed that A is logically possible, then it follows that the knowledgebase is synonymous to A itself.

The nonmonotonic nature of these expressions becomes apparent if an additional proposition that that particular bird does not fly is added to the knowledgebase:

(that which is assumed is
  (and (not A)
    (if (A is possible with respect to what is assumed) then A)))

Reflection on this new knowledgebase leads immediately to the conclusion that it is synonymous to not A. This conclusion is again obtained by simple case analysis: for if A is possible with respect to what is assumed then, since truth implies A is just A, that which is assumed is indeed not A and A which is falsity. Since that which is

653

assumed is falsity is logically possible which it is not. Thus A is not possible with respect to what is assumed. On the other hand, if A is not possible with respect to what is assumed then, since falsity implies A is just truth, that which is assumed is just not A. Since that which is assumed is (not A), A is not possible with respect to what is assumed only if A and (not A) is not logically possible which is the case. Thus it follows that the knowledgebase is synonymous to (NOT A).

Therefore whereas the original knowledgebase was synonymous to A the new knowledgebase, obtained by adding (not A), is synonymous, not to falsity, but to (not A) itself.

These simple intuitive nonmonotonic arguments involve logical concepts such as not, implies, truth, falsity, logical possibility, possibility with respect to some assumed knowledgebase, and synonymity to a knowledgebase. The concepts: not, implies, truth(i.e. T), and falsity(i.e. NIL) are all concepts of (extensional) quantificational logic and are well known. The remaining concepts: logical possibility, possibility with respect to something, and synonymity of two things can be defined in a very simple modal logic extension of quantificational logic, which we call Z[Brown1.2,3,4,]. The axiomatization of the modal logic Z is described in detail in section 2. But briefly, it consists of (extensional) quantificational logic plus the intentional concept of something being logically true written as the unary predicate: (LT P). The concept of a proposition P being logically possible and the concept of two propositions being synonymous are then defined as:

(POS P) = (NOT(LT(NOT P)))   ;P is logically possible
(SYN P Q) = (LT(IFF P Q))   ;P is synonymous to Q

The above knowledgebases and arguments can be formalized in the mmmmmmmmm modal logic Z quite simply by letting some letter such as K stand for the knowledgebase under discussion. The idiom "that which is assumed is X" can then be rendered to say that K is synonymous to X, and the idiom "X is possible with respect to what is assumed" can be rendered to say that K and X is possible:

(that which is assumed is X)   = (SYN K X)
(X is possible with respect to what is assumed)   = (POS(AND K X))

These two idioms are indexial symbols referring implicitly to some particular knowledgebase K under discussion. This knowledgebase referenced by the (X is possible with respect to what is assumed) idiom is always the meaning of the symbol generated by the enclosing (that which is assumed is X) idiom. Each occurrence of the (that which is assumed is X) idiom always generates a symbol(unique to the theory being discussed) to stand for the database under discussion.

The first knowledgebase is then expressed as:

(SYN K(IMPLY(POS(AND K A))A))

Its commonsense argument could be carried out in the following steps:

(IF (POS(AND K A))
    (SYN K(IMPLY T A))
    SYN K(IMPLY NIL A)))

(IF (POS(AND K A))
    (SYN K A)
    (SYN K T))

(OR (AND   (POS(AND K A)) (SYN K A))
    (AND (NOT (POS(AND K A))) (SYN K T)))

(OR (AND   (POS(AND A A)) (SYN K A))
    (AND (NOT (POS(AND T A))) (SYN K T)))

(OR (AND   (POS A) (SYN K A))
    (AND (NOT (POS A)) (SYN K T)))

by equality substitution using the following derived rules of inference of the modal quantificational logic: Z.
    (g(POS P))=(IF(POS P)(g T)(g NIL))
    (IMPLY T A)=A
    (IMPLY NIL A)=T
    (IF P L R)=(OR(AND P L)(AND(NOT P)R))

654

```
(AND(P X Y)(SYN X Y))=(AND(P Y Y)(SYN X Y))
(AND A A)=A
(AND T A)=A
```

Furthermore if A is logically possible: (POS A) then further simplification using laws about AND and OR yields the fact that the knowledgebase K is synonymous to A:

```
(OR (AND T   (SYN K A))
    (AND NIL (SYN K T)) )
```

```
(SYN K A)
```

The second knowledgebase is then expressed as:

```
(SYN K(AND(NOT A)(IMPLY(POS K A)A)))
```

Its commonsense argument could be carried out in the following steps:

```
(IF (POS(AND K A))
    (SYN K(AND(NOT A)(IMPLY T A)))
    (SYN K(AND(NOT A)(IMPLY NIL A))) )
```

```
(IF (POS(AND K A))
    (SYN K(AND(NOT A)A))
    (SYN K(AND(NOT A)T)) )
```

```
(IF (POS(AND K A))
    (SYN K NIL)
    (SYN K(NOT A)) )
```

```
(OR (AND       (POS(AND K A)) (SYN K NIL))
    (AND (NOT (POS(AND K A))) (SYN K(NOT A))) )
```

```
(OR (AND       (POS(AND NIL A))   (SYN K NIL))
    (AND (NOT (POS(AND(NOT A)A))) (SYN K(NOT A))) )
```

```
(OR (AND       (POS NIL) (SYN K NIL))
    (AND (NOT (POS NIL)) (SYN K(NOT A))) )
```

```
(OR (AND       NIL  (SYN K NIL))
    (AND (NOT NIL) (SYN K(NOT A))) )
```

```
(OR NIL
    (AND T (SYN K(NOT A))) )
```

```
(SYN K(NOT A))
```

These knowledgebases have been expressed solely in terms of the modal quantificational logic Z. In particular, the nonmonotonic concepts were explicitly defined in this logic. The intuitive arguments about the meaning of these nonmonotonic knowledgebases have been carried out solely in the modal quantificational logic Z. Most importantly, our commonsense understanding and reasoning about nonmonotonicity is directly represented by the inference steps of this formal theory. Therefore, it is clear that nonmonotonic reasoning needs no special axioms or rules of inference because it is already inherent in the normal intentional logic of everyday commonsense reasoning as modeled by the modal quantificational logic Z.

The modal quantificational logic Z is described in section 2. This is followed in section 3 by the presentation of the basic theorems of our nonmonotonic theory. Section 4 compares our theory of nonmonotonicity with a number of other theories which have appeared in the literature. More complex examples of nonmonotonic reasoning are given in section 5. And finally, a few conclusions are drawn in section 6.

655

2. THE MODAL QUANTIFICATIONAL LOGIC: Z

Our theory of commonsense intentional reasoning is a simple modal logic[Lewis] that captures the notion of logical truth. The symbols of this modal logic consist of the symbols of (extensional) quantificational logic plus the primitive modal symbolism: (LT p) which is truth whenever the proposition p is logically true. Propositions are intuitively the meanings of sentences. For example, the sentences: '(IMPLY p q) and '(OR(NOT p)q) both mean that p implies q. Thus, although these two sentences are different, the two propositions: (IMPLY p q) and (OR(NOT p)q) are the same. Propositions may be true or false in a given world, but with the exception of the true proposition(i.e. the meaning of '(IMPLY p p)) and the false proposition(i.e. the meaning of '(AND p(NOT p))), propositions are not inherently true or false. Thus mathematically, propositions may be thought of as being the elements of a complete atomic Boolean algebra with an arbitrary(possibly infinite) number of generators. The introduction of propositions as objects of reasoning is not an unreasonable thing to do, but instead should be viewed in the general context of extending mathematics with ideal entities such as irrational numbers, complex numbers, and infinitesimals [Robinson]. [Hilbert] points out that the introduction of such ideal entities is an important technique of mathematical reasoning limited only by the need of consistency. In fact the introduction of truthvalues for the nonfinistic sentences of first order quantificational logic is in itself an extension of traditional mathematical reasoning with ideal entities.

The axioms and inference rules of this modal logic include the axioms and inference rules of (extensional) quantificational logic similar to that used by Frege in Begriffsschrift[Frege], plus the following inference rule and axioms about the concept of logical truth.

The Modal Logic Z

R0: from p infer (LT p)
A1: (IMPLY(LT P) P)
A2: (IMPLY(LT(IMPLY P Q)) (IMPLY(LT P)(LT Q)))
A3: (OR(LT P) (LT(NOT(LT P))))
A4: (IMPLY(ALL Q(IMPLY(WORLD Q)(LT(IMPLY Q P)))) (LT P))
A5: (ALL S(POS(meaning of the generator subset S)))

The inference rule R0 means that p is logically true may be inferred from the assertion of p to implicitly be logically true. The consequence of this rule is that a proposition P may be asserted to be logically true by writing just:
    P
and that a proposition P is asserted to be true in a particular world or state of affairs W by writing:
    (LT(IMPLY W P))
The axiom A1 means that if P is logically true then P. Axiom A2 means that if it is logically true that P implies Q then if P is logically true then Q is logically true. Axiom A3 means that P is logically true or it is logically true that P is not logically true. The inference rule R0 and the axioms A1, A2 and A3 constitute an S5 modal logic. A good introduction to modal logic in general and in particular to the properties of the S5 modal logic is given in [Hughes and Cresswell]. Minor variations of the axioms A1, A2, and A3 were shown in [Carnap] to hold for the modal concept of logical truth. We believe that the additional axioms, namely A4 and A5, are needed in order to precisely capture the notion of logical truth. One important theorem scheme of S5 is: (IFF(ALL X(LT(p X)))(LT(ALL X(p X)))) (see[Hughes and Cresswell]) which shows that a property p is logically true for everything iff it is logically true that for everything p holds. The consequence of allowing quantification through modal contexts [Marcus] such as in (ALL X(LT(p X))) is that the meanings of the expressions substituted for variables are concepts of objects and not the objects themselves. However, as [Carnap] explains, in a most precise manner, this does not mean that the real objects of the world are never denoted by such expressions because in a world a concept of an object is equivalent to every other concept of that object, and thus all such concepts then denote that object. Thus, as shown in[Carnap] there is no fundamental problem with quantifying through modal contexts. For example, the often quoted example of[Quine], criticizing quantified modal logic, that: (EQUAL A(THE X(AND P(EQUAL X A)))) which he believes should necessarily be true when P is true, is, as one would suspect, in our system not logically true, but merely true in any world in which P is true. It should be noted that this sentence is not even always true in extensional logic for it is equivalent to: (EQUAL A(THE X NIL))in any world in which P is false(i.e. NIL).

The axiom A4 states that a proposition is logically true if it is true in all worlds. Thus it expresses the contrapositive of Leibniz's intuition that something is logically true only if it is true in all worlds: "The truth of these [necessary propositions] is eternal; not only will they hold whilst the world remains but they would have held even if God had created the world in another way."[Leibniz2] We therefore call this axiom Leibniz's world axiom. We say that a proposition P is a world iff P is possible and P is complete, that P is complete iff for all Q, P determines Q, that P determines Q iff P entails Q or P entails not Q, that P entails Q iff it is logically true that P implies Q, and that P is possible iff it is not the case that not P is logically true. These definitions are give below:

656

(WORLD P)	=df (AND(POS P)(COMPLETE P))	;P is a world
(COMPLETE P)	=df (ALL Q(DET P Q))	;P is complete
(DET P Q)	=df (OR(ENTAIL P Q)(ENTAIL P(NOT Q)))	;P determines Q
(ENTAIL P Q)	=df (LT(IIMPLY P Q))	;P entails Q
(POS P)	=df (NOT(LT(NOT P)))	;P is possible

Thus a world is a possible proposition which for every proposition entails it or its negation. Axiom A4 therefore eliminates from the interpretations of the modal logic Z those complete Boolean algebras which are not atomic. This axiom has been used by a number of authors in developing modal logics in particular[Prior,Brown1,2,3,4].

From the standpoint of Kripke semantics[Kripke] it may seem strange that we define worlds in terms of logical truth: LT, instead of the defining logical truth in terms of worlds. The reason we do this is that logical truth is a more intuitively primitive concept than the concept of a world as[Rescher] points out: "The crucial advantage of this procedure is an epistomological one: we know reasonably well how to get a logic so as to be able to go on from there by constructive means, but we have no intellectual intuition to provide us with direct, non-constructive access to a realm of possible worlds(nor is there any deux ex machina to waft us thither)." Thus for us as is for [Rescher]: "Necessary truths from this standpoint, are not necessary because they are 'true in all possible worlds'; au contraire, possible worlds are so-- i.e. are possible-- because they do not conflict with truths that qualify as necessary on independent grounds".

The use of propositional quantifiers such as (All Q...) in Leibniz's axiom: A4 and in the definition of COMPLETE is of course nothing new; as propositional quantifiers have been an enduring feature of both (intentional) quantificational logics such as [Carnap] and of (extensional) quantificational logics beginning with Frege's discovery of quantificational logic in Begriffschrift[Frege], and continuing in the great Polish works on logic such Lesniewski's protothetic[Lesniewski], in all higher order logics as the zero arity variables for the zero arity verbs of the logic, and in Morse's set theory[Morse]. The underlying (extensional)quantificational logic of our modal logic Z may either treat propositions as a separate sort thus requiring a sorted logic, or they may treat propositions as being normal objects by giving a propositional interpretation for every object such as for example in the manner in which LISP's logical functions interpret every atom except NIL as being true[McCarthy1]. In either case, it is important to note that all the normal laws of extensional quantificational logic, including the laws for substitution of quantified variables, also hold for any complete atomic Boolean algebra, and therefore are compatible with the axioms of the modal logic Z. Thus the assumption made in model theoretic semantics that extensional logics inherently involve only the simplest Boolean algebra consisting of the two propositions truth:T and falsity:NIL is incorrect. Therefore, we do not say that a proposition is inherently true or false, but say instead that a proposition is true or false in some world. Thus if W is a world and P is a proposition then we say that P is true in W iff W entails P, and that P is false in W iff W entails not P:

(IS-TRUE-IN W P) =df (ENTAIL W P)

(IS-FALSE-IN W P) =df (ENTAIL W(NOT P))

If we do not wish to speak about a world when speaking about a proposition it is necessary to divide the propositions into 3 disjoint groups as did Leibniz in his 1686 essay on "Necessary and Contingent Truths" [Leibniz] where he wrote: "That which lacks such necessity I call CONTINGENT, but that which implies a contradiction, or whose opposite is NECESSARY, is called IMPOSSIBLE. The rest are called possible." Thus, Leibniz divides propositions into three categories: those which are necessary, those which are contingent, those which are impossible or contradictory. These propositions may be interpreted as being essentially the elements of a complete atomic Boolean algebra. We say that P is necessary iff P is logically true, that P is impossible(i.e. logically false) iff not P is logically true, and that P is contingent iff P is not logically true and P is not logically false:

(NECESSARY P)	=df (LT P)	;P is necessary
(CONTINGENT P)	=df (AND(NOT(LT P))(NOT(LT(NOT P))))	;P is contingent
(IMPOSSIBLE P)	=df (LT(NOT P))	;P is impossible

The axiom A5 states that the meaning of every conjunction of the generated contingent propositions or their negations is possible. We call this axiom "The Axiom of the Possibility of Contingent facts" or simply the "Possibility Axiom". The need for this axiom follows from the fact that the other axioms of the modal logic do not imply certain elementary facts about the possibility of conjunctions of distinct possibly negated atomic expressions consisting of nonlogical symbols. For example, if we have a theory formulated in our modal logic which contains the nonlogical atomic expression (ON A B) then since (ON A B) is not logically true, it follows that (NOT(ON A B)) must be possible. Yet (POS(NOT(ON A B))) does not follow from these other axioms. Likewise, since (NOT(ON A B)) is not logically true (ON A B) must be possible. Yet (POS(ON A B)) does not follow from the other axioms. Thus these contingent propositions (ON A B) and (NOT(ON A B)) need to be asserted to be possible. There are a number of ways in which this may be done and these ways essentially correspond to different ways the idiom: (P is a meaning combination of the generators) may be rendered. In this paper we have chosen a general method which is applicable to just about any contingent theory one wishes. This rendering is given below:

657

(meaning of the generator subset S) =df
    (ALL G(IMPLY(GENERATORS G)
                (IFF(S G)(GMEANING G))))

(GMEANING '(p ,X1...,XN)) =df (p(GMEANING X1)...(GMEANING XN))
    for every contingent symbol p of arity n.

(GENERATORS) =df (LAMBDA(A)(A is a contingent variable-free simple sentence))

We say that the meaning of the generator subset S is the conjunction of the GMEANINGs of every generator in S and the negation of the GMEANINGS of all the generators not in S. The generator meaning of any expression beginning with a contingent symbol 'p is p of the GMEANING of its arguments. The generators are simply any contingent variable-free atomic sentences we wish to use. The GMEANINGS of the generators may be interpreted essentially as being the generators of a complete atomic Boolean algebra. Thus if there are N generators then there will be (EXP 2(EXP 2 N)) propositions.

For example, a contingent language with a single contingent propositional function 'P and names 'A and 'B gives rise to two contingent generators: '(P A) and '(P B). The GENERATORS and GMEANING functions for this language are defined as:

(GENERATORS) =df {'(P A) '(P B)}
{P1...Pn} =df (LAMBDA(X)(OR(EQUAL X P1)...(EQUAL X Pn)))

(GMEANING '(P ,X)) = (P (GMEANING X))
(GMEANING 'A) = A
(GMEANING 'B) = B

and the Possibility Axiom simplifies as follows:

(ALL S(POS(meaning of the generator subset S)))
(ALL S(POS(ALL G(IMPLY(GENERATORS G)
                (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(IMPLY({'(P A) '(P B)}G)
                (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(IMPLY((LAMBDA(X)(OR(EQUAL X '(P A))(EQUAL X '(P B))))G)
                (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(IMPLY(OR(EQUAL G '(P A))(EQUAL G '(P B)))
                (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(AND(IMPLY(EQUAL G '(P A)) (IFF(S G)(GMEANING G)))
           (IMPLY(EQUAL G '(P B)) (IFF(S G)(GMEANING G)))))))
(ALL S(POS(AND(ALL G(IMPLY(EQUAL G '(P A)) (IFF(S G)(GMEANING G))))
           (ALL G(IMPLY(EQUAL G '(P B)) (IFF(S G)(GMEANING G)))))))
(ALL S(POS(AND(IFF(S '(P A))(GMEANING '(P A)))
          (IFF(S '(P B))(GMEANING '(P B))))))
(ALL S(POS(AND(IFF(S '(P A))(P(GMEANING 'A)))
          (IFF(S '(P B))(P(GMEANING 'B))))))
(ALL S(POS(AND(IFF(S '(P A))(P A))
          (IFF(S '(P B))(P B)))))
(ALL S(POS(AND(IFF(S '(P A))(P A))
          (IFF(S '(P B))(P B)))))
(AND(POS(AND(P A)(P B)))
    (POS(AND(P A)(NOT(P B))))
    (POS(AND(NOT(P A))(P B)))
    (POS(AND(NOT(P A))(NOT(P B)))))

The above possibility axiom involves a number of noncontingent symbols such as names of expressions '(P A) '(P B) 'A 'B, the recursive propositional function GMEANING, the set of GENERATORS, and the second order logic notion of application(i.e. a set theoretic concept of elementhood) and LAMBDA abstraction. These concepts must be logically true because they must be the same for every world. For example it would make no sense to say that '(P A)

658

means one thing in one world and something else in another world, because its meaning is independent of the world in which the expression is uttered. Whether the meaning (P A) of the expression '(P A) is true or false in a world will of course depend on that world. For example the meaning of '(P A) is true in the world: (AND(P A)(P B)...) and false in the world (AND(NOT(P A))(P B)...). But the fact that (GMEANING '(P A)) equals (P A) is logically true. Thus the symbols of our logic are divided into two groups. The contingent symbols whose names are allowed to occur in the set of GENERATORS and the noncontingent symbols(i.e. the symbols of classical extensional logic consisting of LT, symbols defined in terms of LT, syntax symbols, GMEANING, and GENERATORS) whose names are not allowed to occur in any name in the set of GENERATORS. (Of course these restrictions do not preclude us from axiomatizing for example a contingent set theory within a given world but that contingent set theory has nothing to do with the sets of the noncontingent second order logic and in fact will be expressed with new contingent symbols.

Although we have not done so for reasons of presentation, it should be noted that the recursively defined GMEANING concept could have been explicitly defined in the modal logic Z using the well known method [Frege] of explicitly defining recursive functiions in second order logic. It is for this reason we say that this modal logic Z is a logically true theory. Alternatively, it should be noted that if one disallows the nesting of contingent function symbols except for constants (as is often down in the (extensional) first order quantificational logic) then the recursion inherent in the GMEANING definition can be eliminated, thus making this definition explicit anyway.

If the set of GENERATORS is finite then the possibility axiom reduces, in a manner similar to the above derivation, to a conjunction of sentences stating that any conjunction of simple sentences or their negations is possible, and this resulting sentence is entirely expressed within the modal logic Z based on an underlying (extensional) first order quantificational logic. However, it is important to note that finiteness of the generator set is not required by our modal logic and that the possibilty axiom A5 will provide the necessary possibilities as theorems for any contingent language. For example, the fact that the conjunction of the P of all natural numbers is possible can be derived as follows from the infinite generator set consisting of all simple sentences of the form '(P ,N) where N is a numeral. (Syntactic verbs such as NUMERAL herein are of course not contingent symbols.)

(GENERATORS) =df (LAMBDA(X)(EX N(AND(NUMERAL N)(EQUAL X(P N)))))

(GMEANING '(P ,N)) = (P (GMEANING N))
(GMEANING '(ADD1 ,N) = (ADD1 (GMEANING N)))
(GMEANING '1) = 1

(ALL S(POS(meaning of the generator subset S)))
(ALL S(POS(ALL G(IMPLY(GENERATORS G)
                    (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(IMPLY((LAMBDA(X)(EX N(AND(NUMERAL N)(EQUAL X '(P ,N)))))G)
                    (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(IMPLY(EX N(AND(NUMERAL N)(EQUAL G '(P ,N)))
                    (IFF(S G)(GMEANING G))))))
(ALL S(POS(ALL G(ALL N(IMPLY(AND(NUMERAL N)(EQUAL G '(P ,N)))
                    (IFF(S G)(GMEANING G)))))))
(ALL S(POS(ALL N(IMPLY(NUMERAL N)
                    (IFF(S '(P ,N))(GMEANING '(P ,N)))))))
(ALL S(POS(ALL N(IMPLY(NUMERAL N)
                    (IFF(S '(P ,N))(P(GMEANING N)))))))

Thus if S is the universe it follows that:
(POS(ALL N(IMPLY(NUMERAL N)
                    (IFF(UNIVERSE '(P ,N))(P(GMEANING N))))))   .
(POS(ALL N(IMPLY(NUMERAL N) (IFF T(P(GMEANING N))))))
(POS(ALL N(IMPLY(NUMERAL N) (P(GMEANING N)))))
which intuitively is:
(POS(AND(P 1)...))

In his 1696 paper "On the Principle of Indiscernibles" Leibniz wrote: "For all things which are different must be distinguished in some way". Thus we say that two things are equal iff every property which is the meaning of an expression constructed from contingent function symbols and which holds for the first thing also holds for the second thing:
    (EQUAL X Y) =df (ALL p which are contingent(IMPLY(p X)(p Y))) ;equal

659

For example, If the contingent GENERATORS are {'(P A) '(P B)}
then this definition scheme could be rendered:
  (EQUAL X Y) =df (AND(IMPLY(P X)(P Y))
                        (IMPLY(NOT(P X))(NOT(P Y))))
We can then say that two things are equal in a world W iff the fact that they are equal is true in W:
  (EQUAL-IN W X Y) =df (IS-TRUE-IN W(EQUAL X Y)) ;equal in a world
The concept of being equal in a world should be clearly distinguished from the concept of being equal in all worlds:
  (LT-EQUAL X Y) =df (LT(EQUAL X Y))   ;equal in all worlds
as the confusion between these concepts seems to be one of the enduring themes in philosophical logic. For
example, by these definitions all the following statements are true:
  (EQUAL-IN(AND(P A)(P B)) A B)
  (NOT(EQUAL-IN(AND(P A)(NOT(P B))) A B))
  (NOT(EQUAL-IN(AND(NOT(P A))(P B)) A B))
  (EQUAL-IN(AND(NOT(P A))(NOT(P B))) A B)
  (NOT(LT-EQUAL A B))
A simple concept of the transworld identity of two objects can then be expressed by saying that for every contingent
property p which is an essential property of objects if the first object has that property in the first world then the
second object has that property in the second world:
  (SAME X W1 Y W2) =df
      (ALL p which are contingent and essential
          (IMPLY(TRUE-IN W1(p A))(TRUE-IN W2(p B))))

For example, If the contingent GENERATORS are {'(P A) '(P B)} and if P is a property which is essential to the
description of an object then this definition scheme could be rendered:
  (SAME X W1 Y W2) =df (AND(TRUE-IN W1(IMPLY(P X)(P Y)))
                            (TRUE-IN W2(IMPLY(NOT(P X))(NOT(P Y)))))
and the following proposition can be seen to be true:
  (SAME A(AND(P A)(NOT(P B)))B(AND(NOT(P A))(P B)))
  =(AND(IMPLY (TRUE-IN(AND(P A)(NOT(P B)))(P A))
               (TRUE-IN(AND(NOT(P A))(P B))(P B)))
        (IMPLY(TRUE-IN(AND(P A)(NOT(P B)))(NOT(P A)))
               (TRUE-IN(AND(NOT(P A))(P B))(NOT(P B)))))
  =T

The value of the Modal Logic Z is that it models our commonsense reasoning more directly than does classical
logic in that, in addition to the propositional objects NIL and T, it allows the use of ideal propositional objects
[Hilbert] which can be used as objects of various kinds of reasoning; for example, as objects of belief, as objects of
knowledge, or as objects of obligation. For example, the commonsense notion that a robot believes(or at least that
the robot should believe) that which is entailed by its beliefs and that the robot can conceive that which is not con-
tradicted by its beliefs can be directly defined by explicit definitions of the modal logic Z as follows:
  (BELIEVES ROBOT P) =df (ENTAIL(BELIEFS ROBOT)P))
  (CONCEIVABLE ROBOT P) =df (NOT(BELIEVES ROBOT(NOT P)))
  (BELIEFS ROBOT) =df (the conjunction of contingent propositions believed by the robot)
,the commonsense notion that a Robot knows that which is a true belief can be directly defined with the explicit
definition:
  (KNOW ROBOT P) =df (AND P(BELIEVES ROBOT P))
,and the commonsense notion that a Robot must do that which is entailed by its obligations and may do that which
is not contradicted by its obligations can be directly defined with the explicit definitions:
  (MUST ROBOT P) =df (ENTAIL(OBLIGATIONS ROBOT)P)
  (MAY ROBOT P) =df (NOT(MUST(NOT P)))
  (OBLIGATIONS ROBOT) =df (the conjunction of contingent propositions which are obligations of the robot)
Thus we see that the basic concepts of doxastic logic(the logic of belief), epistemic logic(the logic of knowledge),
and deontic logic (the logic of ethics) can be explicitly defined in a commonsense manner which precisely models our
intuitive understanding of these concepts. This commonsense approach to specifying the properties of intentional
concepts is an amazing contrast to the unintuitiveness of previous methods of specifying them by [Kripke]'s exten-
sion of traditional semantic methods[Tarski]. Just to pick one example of the consequences of using such unintuitive
methods, consider the otherwise acceptable paper[Moore1] where the concept of knowledge is (incorrectly) specified
by the Kripke relation to be an S5 modal logic.
    The consistency of the modal logic Z relative to complete atomic Boolean algebras follows by interpreting LT as
the Boolean function which maps every proposition except T into NIL. The modal quantificational Logic Z is de-

scribed in greater detail in[Brown1 2,3,4]. We now use Z to develop a commonsense theory of nonmonotonicity in the following section.

3. THE REFLEXIVE NONMONOTONIC THEORY

One of the most striking features of nonmonotonic knowledgebases is that they are sometimes described in terms of themselves. Such knowledgebases are said to be reflexive[Hayes2]. For example, the knowledge base K purportedly defined by the axiom:

(SYN K (IMPLY(POS(AND K A))A))

is defined as being synonymous to the default: (IMPLY(POS(AND K A))A) which in turn is defined in terms of K. Thus this purported definition of K is not actually a definition at all but is merely an axiom describing the properties possessed by any knowledgebase K satisfying this axiom. In general, a purported definition of a knowledgebase:

(SYN K(f K))

will be implied by zero or more explicit definitions of the form:

(SYN K g)

where K does not occur in g. The explicit definitions which imply a purported definition of a knowledgebase are called the solutions of that purported definition. In general a purported definition may have zero or more solutions. For example, (SYN K(NOT K)) is (LT(IFF K(NOT K))) which is (LT NIL) which is NIL and therefore has no solutions, and (SYN K K) is (LT(IFF K K)) which is (LT T) which is T and therefore has all solutions. Finally, (SYN K G) where K does not occur in G is an explicit definition of K and therefore has only one solution namely itself.

Because K is the knowledgebase under discussion, it is not itself a contingent proposition of that knowledgebase. Thus 'K is not a GENERATOR and the possibility axiom A5 of section 2 will not apply to it. This is verified by the above example (SYN K NIL) where K consists of the false expression, and thus is not possible.

As a more complex example, consider the knowledgebase K which implements an SR flipflop as two Boolean NOR gates connected together in the following manner:

```
;          S----->\
;               ||0--+
;             +->|/  |
;             +-(--------+
;             ||
;             |+--------+
;             +--->\   |
;                 ||0--+---------> K
;          R----->|/
```
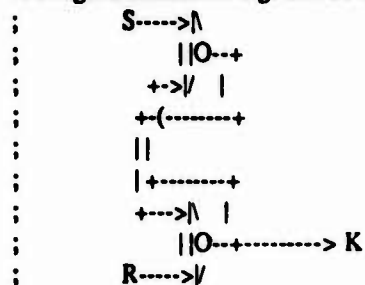
Memory circuits such as this SR flipflop may, unlike nonmemory circuits, involve a self-reference to themselves. Thus their defining equations do not constitute explicit definitions. This memory circuit can quite easily be represented in our logic by letting the output K be synonymous with the expression resulting from tracing the K wire backward thru the NOR gates:

(SYN K(NOR R(NOR S K)))

Memory circuits like any other circuits satisfy various properties. For example, if the reset line R is false and the set line S is true then there is one solution for K, namely that K is true. Furthermore, if the reset line R is true there is also only one solution for K, namely that K is false. However, in any other case this purported definition is implied by any solution. These simple facts can indeed be derived from the logic representation of this circuit as follows.

The SR flipflop theorem:
```
(IFF(SYN K(NOR R(NOR S K)))
    (LT(AND(IMPLY R(NOT K))(IMPLY(AND(NOT R)S)K))) )
        proof
(SYN K(NOT(OR R(NOT(OR S K)))))
(SYN K(AND(NOT R)(OR S K)))
(LT(IFF K(AND(NOT R)(OR S K))))
(LT(IFF K(IF R NIL(IF S T K))))
(LT(IF R(IFF K NIL)(IF S(IFF K T)(IFF K K))))
(LT(IF R(NOT K)(IF S K T)))
(LT(AND(IMPLY R(NOT K))(IMPLY(AND(NOT R)S)K)))
```

In order to make use of a knowledgebase it is helpful to know what is actually in that knowledgebase. For a non-reflexive knowledgebase (i.e. a knowledgebase defined by an explicit definition) this is no problem because there is obviously only one solution, namely that explicit definition itself. However, in the more general case of a purported

definition there may be any number of solutions. Thus the basic goal of a theory of nonmonotonicity of reflexive knowledgebases must be to describe the solutions for various kinds of purported definitions.

The first kind of purported definition we consider is a knowledgebase K consisting of (a conjunction of)axioms G not containing K plus one additional standard default axiom. A standard default axiom is an axiom of the form:

(IMPLY(POS(AND K A))(IMPLY B A))

This structure contains as instances default axioms such as:

(IMPLY(POS(AND K(CAN-FLY ENTERPRISE)))
      (IMPLY(IS-SPACE-SCHUTTLE ENTERPRISE)(CAN-FLY ENTERPRISE)))

T1: A knowledgebase containing exactly one variable-free standard default has precisely one solution.

(IFF (SYN K(AND G(IMPLY(POS(AND K A))(IMPLY B A))))
    (SYN K(AND G(IMPLY(POS(AND G A))(IMPLY B A)))))

               proof
(SYN K(AND G(IMPLY(POS(AND K A))(IMPLY B A))))
(IF (POS(AND K A))
    (SYN K(AND G(IMPLY(AND B T)A)))
    (SYN K(AND G(IMPLY(AND B NIL)A))))
(IF (POS(AND K A))
    (SYN K(AND G(IMPLY B A)))
    (SYN K G))
(OR (AND(POS(AND K A))(SYN K(AND G(IMPLY B A))))
    (AND(NOT(POS(AND K A)))(SYN K G)))
(OR (AND(POS(AND G(IMPLY B A)A))(SYN K(AND G(IMPLY B A)))
    (AND(NOT(POS(AND G A)))(SYN K G)))
(OR (AND(POS(AND G A))(SYN K(AND G(IMPLY B A))))
    (AND(NOT(POS(AND G A)))(SYN K G)))
(IF (POS(AND G A))
    (SYN K(AND G(IMPLY B A)))
    (SYN K G))
(SYN K(IF(POS(AND G A))(AND G(IMPLY B A))G))
(SYN K(AND G(IF(POS(AND G A))(IMPLY B A)T)))
(SYN K(AND G(IMPLY(POS(AND G A))(IMPLY B A))))

The solutions to the two purported definitions discussed in section 1 are obtained from theorem T1 as corollaries for if G is T, B is T, and A is possible it follows that:
(IFF(SYN K(IMPLY(POS(AND K A))A))
    (SYN K A))
and if G is (NOT A) and B is T it follows that:
(IFF(SYN K(AND(NOT A)(IMPLY(POS(AND K A))A))
    (SYN K(NOT A)))

T1 shows that a knowledgebase with only one variable-free standard default, has the same essential status as an explicit definition. The next theorem: T2 shows that this is not the case for a knowledgebase with 2 variable-free standard defaults.

T2: A knowledgebase consisting of two variable-free standard defaults has precisely one or two solutions.

(IFF(SYN K(AND G (IMPLY(POS(AND K A1))(IMPLY B1 A1))
                   (IMPLY(POS(AND K A2))(IMPLY B2 A2))))
    (IF(POS(AND G(IMPLY B2 A2)A1))
        (IF(POS(AND G(IMPLY B1 A1)A2))
           (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)))
           (SYN K(AND G(IMPLY B1 A1))))
        (IF (POS(AND G(IMPLY B1 A1)A2))
           (SYN K(AND G(IMPLY B2 A2)))
           (IF(POS(AND G A1))
              (IF(POS(AND G A2))
                  (OR (SYN K(AND G(IMPLY B1 A1)))

662

```
                          (SYN K(AND G(IMPLY B2 A2))))
                      (SYN K(AND G(IMPLY B1 A1))))
                  (IF(POS(AND G A2))
                      (SYN K(AND G(IMPLY B2 A2)))
                      (SYN K G)) ))))

                  proof
(SYN K(AND G(IMPLY(POS(AND K A1))(IMPLY B1 A1))
            (IMPLY(POS(AND K A2))(IMPLY B2 A2)) ))

(IF(POS(AND K A1))
   (IF(POS(AND K A2))
      (SYN K(AND G(IMPLY T(IMPLY B1 A1))(IMPLY T(IMPLY B2 A2))))
      (SYN K(AND G(IMPLY T(IMPLY B1 A1))(IMPLY NIL(IMPLY B2 A2)))) )
   (IF(POS(AND K A2))
      (SYN K(AND G(IMPLY NIL(IMPLY B1 A1))(IMPLY T(IMPLY B2 A2))))
      (SYN K(AND G(IMPLY NIL(IMPLY B1 A1))(IMPLY NIL(IMPLY B2 A2)))) ))

(IF(POS(AND K A1))
   (IF(POS(AND K A2))
      (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)))
      (SYN K(AND G(IMPLY B1 A1))) )
   (IF(POS(AND K A2))
      (SYN K(AND G(IMPLY B2 A2)))
      (SYN K G) ))

(OR(AND(POS(AND K A1))(POS(AND K A2))
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
   (AND(POS(AND K A1))(NOT(POS(AND K A2)))(SYN K(AND G(IMPLY B1 A1))))
   (AND(NOT(POS(AND K A1)))(POS(AND K A2))(SYN K(AND G(IMPLY B2 A2))))
   (AND(NOT(POS(AND K A1)))(NOT(POS(AND K A2)))(SYN K G)) )

(OR(AND(POS(AND G(IMPLY B1 A1)(IMPLY B2 A2)A1))
        (POS(AND G(IMPLY B1 A1)(IMPLY B2 A2)A2))
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
   (AND(POS(AND G(IMPLY B1 A1)A1))
        (NOT(POS(AND G(IMPLY B1 A1)A2)))
        (SYN K(AND G(IMPLY B1 A1))))
   (AND(NOT(POS(AND G(IMPLY B2 A2)A1)))
        (POS(AND G(IMPLY B2 A2)A2))
        (SYN K(AND G(IMPLY B2 A2))))
   (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)) )

(OR(AND(POS(AND G(IMPLY B2 A2)A1))
        (POS(AND G(IMPLY B1 A1)A2))
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
   (AND(POS(AND G A1))
        (NOT(POS(AND G(IMPLY B1 A1)A2)))
        (SYN K(AND G(IMPLY B1 A1))))
   (AND(NOT(POS(AND G(IMPLY B2 A2)A1)))
        (POS(AND G A2))
        (SYN K(AND G(IMPLY B2 A2))))
   (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)) )

(IF(POS(AND G(IMPLY B2 A2)A1))
   (IF(POS(AND G(IMPLY B1 A1)A2))
      (OR(SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)))
          (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)))
      (OR(AND(POS(AND G A1))
              (SYN K(AND G(IMPLY B1 A1))))
                          663
```

```
                    (AND(NOT(POS(AND G A1))(NOT(POS(AND G A2)))(SYN K G)) )
        (IF(POS G(IMPLY B1 A1)A2)
            (OR(AND(POS(AND G A2))
                    (SYN K(AND G(IMPLY B2 A2))))
                (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)) )
            (OR(AND(POS(AND G A1))
                    (SYN K(AND G(IMPLY B1 A1))))
                (AND(POS(AND G A2))
                    (SYN K(AND G(IMPLY B2 A2))))
                (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)) )

    (IF(POS G(IMPLY B2 A2)A1)
        (IF(POS(AND G(IMPLY B1 A1)A2))
            (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)))
            (SYN K(AND G(IMPLY B1 A1))))
        (IF(POS(AND G(IMPLY B1 A1)A2))
            (SYN K(AND G(IMPLY B2 A2)))
            (OR(AND(POS(AND G A1))
                    (SYN K(AND G(IMPLY B1 A1))))
                (AND(POS(AND G A2))
                    (SYN K(AND G(IMPLY B2 A2))))
                (AND(NOT(POS(AND G A1)))(NOT(POS(AND G A2)))(SYN K G)) )))

    (IF(POS(AND G(IMPLY B2 A2)A1))
        (IF(POS(AND G(IMPLY B1 A1)A2))
            (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)))
            (SYN K(AND G(IMPLY B1 A1))))
        (IF(POS(AND G(IMPLY B1 A1)A2))
            (SYN K(AND G(IMPLY B2 A2)))
            (IF(POS(AND G A1))
                (IF(POS(AND G A2))
                    (OR(SYN K(AND G(IMPLY B1 A1)))
                        (SYN K(AND G(IMPLY B2 A2))))
                    (SYN K(AND G(IMPLY B1 A1))))
                (IF(POS(AND G A2))
                    (SYN K(AND G(IMPLY B2 A2)))
                    (SYN K G)))))
```

The Alternatives Corollary to T2:
   If G is (OR B1 B2) and (AND A1 A2 B1) is possible then T2 reduces to the proposition that:
(IFF(SYN K(AND(OR B1 B2)(IMPLY(POS(AND K A1))(IMPLY B1 A1))
            (IMPLY(POS(AND K A2))(IMPLY B2 A2))))
    (SYN K(AND(OR B1 B2)(IMPLY B1 A1)(IMPLY B2 A2))))
Thus K entails (OR A1 A2). This illustrates the importance of treating defaults as expressions rather than as inference rules because inference rules such as:
   (from a deduction of (POS(AND K A1)) and a deduction of B1 infer A1)
   (from a deduction of (POS(AND K A2)) and a deduction of B2 infer A2)
would not allow any deduction to be made since neither B1 nor B2 is a theorem of the knowledgebase.

The Flipped Coin Corollary to T2:
   If A1 is A, A2 is (NOT A), B1 is T,and B2 is T then T2 reduces to the proposition that:
(IFF(SYN K(AND G(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))(NOT A))))
    (IF(POS(AND G A))
        (IF(POS(AND G(NOT A)))
            (OR(SYN K(AND G A))(SYN K(AND G(NOT A))))
            (SYN K(AND G A)))
        (IF(POS(AND G(NOT A))) (SYN K(AND G(NOT A))) (SYN K G)))
If (AND G A) is possible and (AND G(NOT A)) is possible then this proposition reduces to:
(IFF(SYN K(AND G(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))(NOT A))))
    (OR(SYN K(AND G A))(SYN K(AND G(NOT A)))))

664

which states the K has precisely two solutions. Furthermore, if G is T, then these two solutions are direct opposites in that one says the knowledgebase is A and the other says the knowledgebase is (NOT A):

(IFF(SYN K(AND(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))(NOT A))))
   (OR(SYN K A)(SYN K(NOT A)))

There is nothing at all bizarre about having multiple solutions or even opposite multiple solutions for one can easily imagine a robot executing actions in a given state resulting in a new state K which must be one of the above solutions, although there is no way the robot in its planning can determine which solution for K is actually the case. For example let A be the proposition that a flipped coin will land with heads. The default: (IMPLY(POS(AND K A))A) then means that if it is possible for the coin to land on heads then assume it does so. Likewise, the default (IMPLY(POS(AND K(NOT A)))(NOT A)) means that if it is possible for a coin to land tails(i.e. not heads)then assume it does so. The result of the action is then one of two states K:

(OR(SYN K A)(SYN K(NOT A)))

where the coin landed heads and where the coin landed tails (i.e. not heads). It should be noted that a disjunction of solutions is altogether different from a solution which is a disjunction of alternatives such as:(SYN K(OR A(NOT A))) which in this case is equivalent to:(SYN K T) and which would be an incorrect rendering of what is intuitively meant by multiple defaults.

In planning further actions to the resulting state K in order to achieve some overall goal the robot must take into account all the different solutions for K and make its plans accordingly. For example, if the robots overall goal is to flip the coin until it lands heads then the robot should plan to do nothing for the solution: (SYN K A) ,but should plan to continue flipping the coin for the solution (SYN K(NOT A)).

The purpose of using these default axioms is to allow for the case of where additional information in G contradicts the defaults. For example if the coin has tails on both sides then the flipped coin will always land tails. Thus letting G be (NOT A) and assuming that (NOT A) is logically possible, the first corollary expression of T2 above reduces to the single solution:

(IFF(SYN K(AND(NOT A)(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))(NOT A))))
   (SYN K(NOT A)))

Likewise if the coin has heads on both sides then the flipped coin will always land heads. Thus letting G be A and assuming that A is logically possible, the first corollary expression of T2 above reduces to the single solution:

(IFF(SYN K(AND(NOT A)(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))A)))
   (SYN K A))

The Closed World Assumption Corollary to T2:

Assume that the knowledgebase consists of the fact that (OR(NOT A1)(NOT A2)) and two standard defaults implementing the closed world assumption[Reiter1] that the meaning of any simple sentence which is possible with respect to what is assumed is the case. If B1 is T, B2 is T, and G is (OR(NOT A1)(NOT A2)) in T2 and if 'A1 and 'A2 are GENERATORS then T2 reduces to:

(IFF(SYN K(AND(OR(NOT A1)(NOT A2))
       (IMPLY(POS(AND K A1))A1)(IMPLY(POS(AND K A2))A2)))
   (OR(SYN K(AND(NOT A2)A1))(SYN K(AND(NOT A1)A2))))

This corollary illustrates the fact that defaults with mutually exclusive conditions even if they are not negations of each other do in fact result in alternative solutions.

The flipped coin and closed world examples can be generalized to a knowledgebase containing N mutually exclusive defaults. Essentially, this is done by adding N standard defaults to the theory and letting G state that the conclusions of all these defaults are mutually exclusive. Such a knowledge base will have precisely N solutions whenever no other information is available. This fact is proven below in theorem T3. This proof illustrates the smooth interaction of reasoning with complex mixtures of contingent and necessary expressions involving quantifiers in the modal quantificational logic Z. Thus, for example, in the absence of any other information a knowledgebase containing six mutually exclusive defaults specifying the world states after rolling a six sided die would result in six distinct solutions.

T3: For all N there exist a knowledgebase with N standard defaults and with N solutions.
(IMPLY
 (AND(SYN G(AND G2(ALL I(ALL J(OR(EQUAL I J)(IMPLY(P I)(NOT(P J)))))))))
   (ALL N(POS(P N))))
 (ALL N(IFF(SYN K(AND G(ALL M(IMPLY(<= M N)
                    (IMPLY(POS(AND K(P M)))(P M)))))))
     (EX M(AND(<= M N)(SYN K(AND G(P M)))))))))

Let I,J,N range over the positive integers which along with their numeric properties such as = and <= are assumed to be necessary.

Let (SYN G(AND G2(ALL I(ALL J(OR(EQUAL I J))(IMPLY(P I)(NOT(P J))))))).

Let (ALL N(POS(P N))) be true. Then it follows that:

```
(ALL N(IFF(SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))) ))
          (EX M(AND(<= M N)(SYN K(AND G(P M))))) ))
                            proof
                      (by induction on N)
base-case: N=1
(IFF(SYN K(AND G(ALL M(IMPLY(<= M 1)(IMPLY(POS(AND K(P M)))(P M)))) ))
    (EX M(AND(<= M 1)(SYN K(AND G(P M))))) )
(IFF(SYN K(AND G(IMPLY(POS(AND K(P 1)))(P 1))))
    (SYN K(AND G(P 1))) )
(IFF(SYN K(AND G(IMPLY(POS(AND K(P 1)))(P 1))))
    (SYN K(AND G(P 1))) )
(IFF(SYN K(AND G(IMPLY(POS(AND G(P 1)))(P 1))))   ;by T1
    (SYN K(AND G(P 1))) )
(IFF(SYN K(AND G(IMPLY T(P 1))))  ; since (AND G(P 1))is possible
    (SYN K(AND G(P 1))) )
T

 induction-step N+1 if N
 (ALL N(IMPLY
  (IFF(SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))) ))
      (EX M(AND(<= M N)(SYN K(AND G(P M))))) ))
  (IFF(SYN K(AND G(ALL M(IMPLY(<= M(1+ N))(IMPLY(POS(AND K(P M)))(P M)))) ))
      (EX M(AND(<= M(1+ N))(SYN K(AND G(P M))))) ))

 (ALL N(IMPLY
  (IFF(SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))) ))
      (EX M(AND(<= M N)(SYN K(AND G(P M))))) ))
  (IFF(SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))
                       (IMPLY(POS(AND K(P(1+ N))))(P(1+ N))) ))
            (OR(SYN K(AND G(P(1+ N))))
         (EX M(AND(<= M N)(SYN K(AND G(P M))))) )) )

Let the induction hypothesis be:
H=(IFF(SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))) ))
      (EX M(AND(<= M N)(SYN K(AND G(P M)))))) ))

 (ALL N(IMPLY H             ;equality substitution using induction hypothesis
  (IFF(SYN K(AND G(IMPLY(POS(AND K(P(1+ N))))(P(1+ N)))
                        (ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))) ))
      (OR(SYN K(AND G(P(1+ N))))
            (SYN K(AND G(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M)))))) ))
 ))

assuming H and letting:
B=(ALL M(IMPLY(<= M N)(IMPLY(POS(AND K(P M)))(P M))))
we need to prove for all N that:

(IFF(SYN K(AND G(IMPLY(POS(AND K(P(1+ N)))(P(1+ N))))B))
    (OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B))) ))

(AND(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
          (OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B))) )
    (IMPLY(OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B)))
          (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) ))
```

```
(AND(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
          (OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B))) )
    (IMPLY(SYN K(AND G(P(1+ N))))
          (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) )
    (IMPLY(SYN K(AND G B)))
          (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) ))

(AND(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
          (OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B))) )
    (IMPLY(SYN K(AND G(P(1+ N))))
          (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) )
    (IMPLY(SYN K(AND G B)))
          (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) ))
```

The proof of the induction step breaks into three cases:

case1:
```
(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
      (OR(SYN K(AND G(P(1+ N))))(SYN K(AND G B))) )
(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
      (OR(SYN K(AND G(P(1+ N)))))
          (IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
                (SYN K(AND G B)) ))
;it remains only to prove:
(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
      (OR(SYN K(AND G(P(1+ N)))))
          (NOT(POS(AND G B(P(1+ N)))))))
(IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
      (IMPLY(POS(AND G B(P(1+ N))))
            (SYN K(AND G(P(1+ N)))) ))
(IMPLY(POS(AND G B(P(1+ N))))
      (IMPLY(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B))
            (SYN K(AND G(P(1+ N)))) ))
(IMPLY(POS(AND G B(P(1+ N))))
      (IMPLY(SYN K(AND G(IMPLY T(P(1+ N))))B)
            (SYN K(AND G(P(1+ N)))) ))
(IMPLY(POS(AND G B(P(1+ N))))
      (IMPLY(SYN K(AND G(P(1+ N))B))
            (IMPLY(SYN K(AND G(P(1+ N))B))
                  (SYN K(AND G(P(1+ N)))) )))
;since K is G P(1+ N) B by the first SYN K implication
;it follows that the B expression in the second SYN K implication
;is T since every (AND K(P M))) in that B is cotradicted by the P(1+N)
;of the first SYN K implication.
(IMPLY(POS(AND G B(P(1+ N))))
      (IMPLY(SYN K(AND G(P(1+ N))B))
            (IMPLY(SYN K(AND G(P(1+ N))T))
                  (SYN K(AND G(P(1+ N)))) )))
T

case2:
(IMPLY(SYN K(AND G(P(1+ N))))
(SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) )
;since K is G and P(1+ N) which implies not P1...not Pn,
;(POS(AND K(P M) in B is NIL so B is T.
(IMPLY(SYN K(AND G(P(1+ N))))
      (SYN K(AND G(IMPLY(POS(AND G T(P(1+ N)))(P(1+ N))))T)) )
(IMPLY(SYN K(AND G(P(1+ N))))
(SYN K(AND G(POS(AND G(P(1+ N)))(P(1+ N)))))) )
;since (AND G(P(1+ N))) is possible
```

667

```
(IMPLY(SYN K(AND G(P(1+ N))))
      (SYN K(AND G T(P(1+ N)))) )
T

case3:
(IMPLY(SYN K(AND G B))
      (SYN K(AND G(IMPLY(POS(AND G B(P(1+ N)))(P(1+ N))))B)) )
;would be true if (POS(AND G B(P(1+ N)))) were false
;so it remains only to prove:
(IMPLY(SYN K(AND G B))
      (NOT(POS(AND G B(P(1+ N))))) )
;by hypothesis
(IMPLY(SYN K(AND G B))
      (NOT(POS(AND K(P(1+ N))))) )
;using H again
(IMPLY(EX M(AND(<= M N)(SYN K(AND G(P M)))))
      (NOT(POS(AND K(P(1+ N))))) )
(IMPLY(EX M(AND(<= M N)(SYN K(AND G(P M)))))
      (NOT(POS(AND K(P(1+ N))))) )
(ALL M(IMPLY(<= M N)
            (IMPLY(SYN K(AND G(P M)))
                  (NOT(POS(AND K(P(1+ N))))) )))
(ALL M(IMPLY(<= M N)
            (IMPLY(SYN K(AND G(P M)))
                  (NOT(POS(AND G(P M)(P(1+ N))))) )))
;since (AND G(P M)(P(1+ N))) is NIL
(ALL M(IMPLY(<= M N)
            (IMPLY(SYN K(AND G(P M)))
                  (NOT NIL))))
```

The Infinite Number Corollary to theorem T3:

There exist a knowledgebase with an infinite number of standard defaults and with the same infinite number of solutions. The proof is obtained by letting N in theorem T3 be an infinite positive integer such as omega in non-standard number theory [Robinson].

The following particular knowledgebase is taken from[Reiter] where it is claimed that the analogous formulation in the nonmonotonic logic of [McDermott&Doyle] has no fixed point.

T4: There exists a knowledgebase consisting of three standard defaults which has no solutions:
if 'A1,'A2,'A3,'B1,'B2,'B3 are all generators then:

```
(IFF(SYN K(AND(IMPLY(POS(AND K A1))(IMPLY B1 A1)(IMPLY A1 B2)(NOT(AND A1 A2)))
              (IMPLY(POS(AND K A2))(IMPLY B2 A2)(IMPLY A2 B3)(NOT(AND A2 A3)))
              (IMPLY(POS(AND K A3))(IMPLY B3 A3)(IMPLY A3 B1)(NOT(AND A3
A1)))))
    NIL)
```

                            proof

```
;;;let G be (AND(IMPLY A1 B2)(IMPLY A2 B3)(IMPLY A3 B1)
                (NOT(AND A1 A2))(NOT(AND A2 A3))(NOT(AND A3 A1)))
;;;let GENERATORS include {',A1 ',A2 ',A3 ',B1 ',B2 ',B3}
then:

(SYN K(AND G (IMPLY(POS K A1)(IMPLY B1 A1))
             (IMPLY(POS K A2)(IMPLY B2 A2))
             (IMPLY(POS K A3)(IMPLY B3 A3))))
(IF(POS K A1)
   (IF(POS K A2)
      (IF(POS K A3)
         (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)(IMPLY B3 A3)))
```

                            668

```
                  (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
            (IF(POS K A3)
                  (SYN K(AND G(IMPLY B1 A1)(IMPLY B3 A3)))
                  (SYN K(AND G(IMPLY B1 A1))))))
      (IF(POS K A2)
          (IF(POS K A3)
                (SYN K(AND G(IMPLY B2 A2)(IMPLY B3 A3)))
                (SYN K(AND G(IMPLY B2 A2)T)))
          (IF(POS K A3)
                (SYN K(AND G(IMPLY B3 A3)))
                (SYN K G))))

(OR(AND(POS K A1)(POS K A2)(POS K A3)
          (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)(IMPLY B3 A3))))
     (AND(POS K A1)(POS K A2)(NOT(POS K A3))
          (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
     (AND(POS K A1)(NOT(POS K A2))(POS K A3)
          (SYN K(AND G(IMPLY B1 A1)(IMPLY B3 A3))))
     (AND(POS K A1)(NOT(POS K A2))(NOT(POS K A3))
          (SYN K(AND G(IMPLY B1 A1))))
     (AND(NOT(POS K A1))(POS K A2)(POS K A3)
          (SYN K(AND G(IMPLY B2 A2)(IMPLY B3 A3))))
     (AND(NOT(POS K A1))(POS K A2)(NOT(POS K A3))
          (SYN K(AND G(IMPLY B2 A2))))
     (AND(NOT(POS K A1))(NOT(POS K A2))(POS K A3)
          (SYN K(AND G(IMPLY B3 A3))))
     (AND(NOT(POS K A1))(NOT(POS K A2))(NOT(POS K A3))
          (SYN K G))) )

Then by using G we get:
;;;let G be (AND(IMPLY A1 B2)(IMPLY A2 B3)(IMPLY A3 B1)
;;;                  (NOT(AND A1 A2))(NOT(AND A2 A3))(NOT(AND A3 A1)))

(OR(AND(POS(AND G(IMPLY B2 A2)(IMPLY B3 A3)A1))   ;B2 A2   NIL
        (POS(AND G(IMPLY B1 A1)(IMPLY B3 A3)A2))   ;B3 A3   NIL
        (POS(AND G(IMPLY B1 A1)(IMPLY B2 A2)A3))   ;B1 A1   NIL
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2)(IMPLY B3 A3))))
     (AND(POS(AND G(IMPLY B2 A2)A1))               ;B2 A2   NIL
        (POS(AND G(IMPLY B1 A1)A2))
        (NOT(POS(AND G(IMPLY B1 A1)(IMPLY B2 A2)A3)))
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B2 A2))))
     (AND(POS(AND G(IMPLY B3 A3)A1))
        (NOT(POS(AND G(IMPLY B1 A1)(IMPLY B3 A3)A2)))
        (POS(AND G(IMPLY B1 A1)A3))                ;B1 A1   NIL
        (SYN K(AND G(IMPLY B1 A1)(IMPLY B3 A3))))
     (AND(POS(AND G A1))
        (NOT(POS(AND G(IMPLY B1 A1)A2)))
        (NOT(POS(AND G(IMPLY B1 A1)A3)))           ;B1 A1   T
        (SYN K(AND G(IMPLY B1 A1))))
     (AND(NOT(POS(AND G(IMPLY B2 A2)(IMPLY B3 A3)A1)))
        (POS(AND G(IMPLY B3 A3)A2))                ;B3 A3   NIL
        (POS(AND G(IMPLY B2 A2)A3))
        (SYN K(AND G(IMPLY B2 A2)(IMPLY B3 A3))))
     (AND(NOT(POS(AND G(IMPLY B2 A2)A1)))          ;B2 A2   T
        (POS(AND G A2))
        (NOT(POS(AND G(IMPLY B2 A2)A3)))
        (SYN K(AND G(IMPLY B2 A2))))
     (AND(NOT(POS(AND G(IMPLY B3 A3)A1)))
        (NOT(POS(AND G(IMPLY B3 A3)A2)))           ;B3 A3   T
        (POS(AND G A3))
```

669

```
          (SYN K(AND G(IMPLY B3 A3))))
    (AND(NOT(POS(AND G A1)))
        (NOT(POS(AND G A2)))
        (NOT(POS(AND G A3)))
        (SYN K G)) )

(OR(AND(POS(AND G A1))
        (NOT(POS(AND G(IMPLY B1 A1)A2)))
        (SYN K(AND G(IMPLY B1 A1))))
    (AND(POS(AND G A2))
        (NOT(POS(AND G(IMPLY B2 A2)A3)))
        (SYN K(AND G(IMPLY B2 A2))))
    (AND(NOT(POS(AND G(IMPLY B3 A3)A1)))
        (POS(AND G A3))
        (SYN K(AND G(IMPLY B3 A3))))
    (AND(NOT(POS(AND G A1)))
        (NOT(POS(AND G A2)))
        (NOT(POS(AND G A3)))
        (SYN K G)) )
```

and since conjunctions of Generators are possible all the remaining
subexpressions ... in (POS ...) are possible. hence every conjunct is NIL:
NIL

   Some examples dealing with more esoteric cases of nonmonotonic reasoning
are now given.

T5: A knowledge base with one default(not necessarily standard):
```
(IFF(SYN K(AND G(IMPLY(POS(AND K A))X)))
    (OR(AND(POS(AND G X A))(SYN K(AND G X)))
        (AND(NOT(POS(AND G A)))(SYN K G))) )
                    proof
(SYN K(AND G(IMPLY(POS(AND K A))X)))
(IF(POS(AND K A))(SYN K(AND G X))(SYN K G))
(OR(AND(POS(AND K A))(SYN K(AND G X)))
    (AND(NOT(POS(AND K A)))(SYN K G)) )
(OR(AND(POS(AND G X A))(SYN K(AND G X)))
    (AND(NOT(POS(AND G A)))(SYN K G)))
```

corollary if G is T and if(AND X A) is posssible it follows that:
```
(IFF(SYN K(IMPLY(POS(AND K A))X))
    (SYN K X))
```

T6: A knowledgebase with two defaults(not necessarily standard):
```
(IFF(SYN K(AND G(IMPLY(POS(AND K B))X)(IMPLY(POS(AND K D))Y)))
    (OR(AND(POS(AND G X Y B))(POS(AND G X Y D))(SYN K(AND G X Y)))
        (AND(POS(AND G X B))(NOT(POS(AND G X D)))(SYN K(AND G X)))
        (AND(NOT(POS(AND G Y B)))(POS(AND G Y D))(SYN K(AND G Y)))
        (AND(NOT(POS(AND G B)))(NOT(POS(AND G D)))(SYN K G)) ) )
                        proof
(SYN K(AND G(IMPLY(POS(AND K B))X)(IMPLY(POS(AND K D))Y)))
(IF(POS(AND K B))
    (IF(POS(AND K D))(SYN K(AND G X Y))(SYN K(AND G X)))
    (IF(POS(AND K D))(SYN K(AND G Y))(SYN K G)))
(OR(AND(POS(AND K B))(POS(AND K D))(SYN K(AND G S Y)))
    (AND(POS (AND K B))(NOT(POS(AND K D)))(SYN K(AND G S)))
    (AND(NOT(POS(AND K B)))(POS(AND K D))(SYN K(AND G Y)))
    (AND(NOT(POS(AND K B)))(NOT(POS(AND K D)))(SYN K G)) )
(OR(AND(POS(AND G X Y B))(POS(AND G X Y D))(SYN K(AND G X Y)))
    (AND(POS(AND G X B))(NOT(POS(AND G X D)))(SYN K(AND G X)))
    (AND(NOT(POS(AND G Y B)))(POS(AND G Y D))(SYN K(AND G Y)))
```

```
                (AND(NOT(POS(AND G B)))(NOT(POS(AND G D)))(SYN K G)) )
```

**T7: A knowledgebase having a unique solution may have a subknowledgebase which
has no solutions.**
```
(IFF(SYN K (AND G(IMPLY(POS(AND K A))(NOT A))))
    (AND(NOT(POS(AND G A)))(SYN K G)))
```
       proof
```
 (SYN K (AND G(IMPLY(POS(AND K A))(NOT A))))
(IF(POS(AND K A))
    (SYN K(AND G(NOT A)))
    (SYN K G))
(OR(AND(POS(AND K A))(SYN K(AND G(NOT A))))
    (AND(NOTPOS(AND K A))(SYN K G)))
(OR(AND(POS(AND G(NOT A)A))(SYN K (AND G (NOT A))))
    (AND(NOT(POS(AND G A)))(SYN K G)) )
(OR(AND(POS NIL)(SYN K(AND G(NOT A))))
    (AND(NOT(POS(AND G A)))(SYN K G)) )
(OR NIL(AND(NOT(POS(AND G A)))(SYN K G)))
(AND(NOT(POS(AND G A)))(SYN K G))
```

Thus if G is T and (POS A) we find that there are no solutions:
```
(IFF(SYN K(IMPLY(POS(AND K A))(NOT A))) NIL)
```

If however, we allow G to be (NOT A), we get:
```
(AND (NOT POS (AND A (NOT A))) (SYN K (NOT A)))
```
= (SYN K (NOT A)) and if we assume (POS A) we get (SYN K (NOT A)).

T8: The knowledgebase: G,(NOT B)unless A,(NOT A)unless B
```
(IFF(SYN K(AND G(IMPLY(POS(AND K A))(NOT B))(IMPLY(POS(AND K B))(NOT A))))
    (OR(AND(POS(AND G(NOT B)A))(SYN K(AND G(NOT B))))
        (AND(POS(AND G(NOT A)B))(SYN K(AND G(NOT A))))
        (AND(NOT(POS(AND G A))(NOT(POS(AND G B)))(SYN K G))) ))
```
       proof
```
(SYN K(AND G(IMPLY(POS(AND K A))(NOT B))(IMPLY(POS(AND K B))(NOT A))))
(IF (POS(AND K A))
        (IF (POS(AND K B))
                (SYN K(AND G(NOT B)(NOT A)))
                (SYN K(AND G(NOT B))))
        (IF (POS (AND K B))
                (SYN K(AND G(NOT A)))
                (SYN K G)))
(OR(AND(POS(AND K A))(POS(AND K B))(SYN K(AND G(NOT B)(NOT A))))
    (AND(POS(AND K A))(NOT(POS(AND K B)))(SYN K(AND G(NOT B))))
    (AND(NOT(POS(AND K A))(POS(AND K B))(SYN K(AND G(NOT A)))))
    (AND(NOT(POS(AND K A))(NOT(POS(AND K B))(SYN K G)))) )
(OR(AND(POS(AND G(NOT B)(NOT A)A))(POS(AND G(NOT B)(NOT A)B))
        (SYN K(AND G(NOT B)(NOT A))))
    (AND(POS(AND G(NOT B)A))(NOT(POS(AND G(NOT B)B)))(SYN K(AND G(NOT B))))
    (AND(NOT(POS(AND G(NOT A)A)))(POS(AND G(NOT A)A))(SYN(AND G(NOT A))))
    (AND(NOT(POS(AND G A)))(NOT(POS(AND G B)))(SYN K G)) )
(OR(AND NIL NIL(SYN K(AND G(NOT B)(NOT A))))
    (AND(POS(AND G(NOT B)A))T(SYN K(AND G(NOT B))))
    (AND T(POS(AND G(NOT A)B))(SYN K(AND G(NOT A))))
    (AND(NOT(POS(AND G A)))(NOT(POS(AND G B)))(SYN K G)) )
(OR(AND(POS(AND G(NOT B)A))(SYN K(AND G(NOT B))))
    (AND(POS(AND G(NOT A)B))(SYN K(AND G(NOT A))))
    (AND(NOT(POS(AND G A)))(NOT(POS(AND G B)))(SYN K G)))
```

Letting G be T, we have three solutions:
```
(OR(AND(POS(AND(NOT B)A))(SYN K(NOT B)))
```

```
            (AND(POS(AND(NOT A)B))(SYN K (NOT A)))
            (AND(SYN A NIL)(SYN B NIL)(SYN K T)) )
```

If we now assume (POS(AND(NOT B)A)) and (POS(AND(NOT A)B)), it follows that:
(OR(SYN K(NOT B))(SYN K(NOT A)))

T9 There is a false equation for any possible G.
```
(IFF(SYN K(AND G(IMPLY(POS(AND K A))(NOT A))
                   (IMPLY(POS(AND K(NOT A)))A)))
     (AND(SYN G NIL)(SYN K NIL)) )
                              proof
(SYN K(AND G(IMPLY(POS(AND K A))(NOT A))(IMPLY(POS(AND K(NOT A)))A)))
(IF(POS(AND K A))
    (IF(POS(AND K(NOT A)))
        (SYN K NIL)
        (SYN K(AND G(NOT A))))
    (IF(POS(AND K(NOT A)))
        (SYN K(AND G A))
        (SYN K G)))
(OR(AND(POS(AND K A))(POS(AND K(NOT A)))(SYN K NIL))
    (AND(POS(AND K A))(NOT(POS(AND K(NOT A))))(SYN K(AND G(NOT A))))
    (AND(NOT(POS(AND K A)))(POS(AND K(NOT A)))(SYN K(AND G A)))
    (AND(NOT(POS(AND K A)))(NOT(POS(AND K(NOT A))))(SYN K G)) )
(OR(AND NIL NIL(SYN K NIL))
    (AND NIL(NOT(POS(AND G(NOT A))))(SYN K (AND G·(NOT A))))
    (AND(NOT(POS(AND G A)))NIL(SYN K(AND G A)))
    (AND(NOT(POS(AND G A)))(NOT(POS(AND G(NOT A))))(SYN K G)) )
(AND(NOT(POS(AND G A)))(NOT(POS(AND G(NOT A))))(SYN K G))
(AND(LT(IMPLY G(NOT A)))(LT(IMPLY G A))(SYN K G))
(AND(LT(IMPLY G(AND(NOT A)A)))(SYN K G))
(AND(LT(IMPLY G NIL))(SYN K G))
(AND(SYN G NIL)(SYN K G))
(AND(SYN G NIL)(SYN K NIL))
```
Therefore if G is anything other than NIL there is no solution.

KNOWLEDGEBASES AS OBJECTS OF REASONING

Knowledge bases whether they are defined by explicit definitions or purported definitions involving zero or more solutions, can themselves be used as objects of reasoning. This is done by gathering up all their solutions in some manner so as for example to create the set of solutions, or the disjunction of solutions(i.e. the information content that is shared by all solutions) or the conjunction of solutions(i.e. the sum of the information content of all solutions).

```
(SYN KSET (LAMBDA K(SYN K(g K))))        ;the set of all solutions
(SYN KDISJ (EX K(AND(SYN K(g K))K)))     ;the disjunction of all solutions
(SYN KCONJ (ALL K(IMPLY(SYN K(g k))K)))  ;the conjunction of all solutions
```


4. RELATED RESEARCH: THE FIXED POINT THEORIES

A number of recent papers[McDermott&Doyle,McDermott,Moore,and Reiter] have attempted to formalize the commonsense notion of something being possible with respect to what is assumed. All these papers have been based on the mathematical theory of fixed points. For example, [McDermott&Doyle] describes a rather baroque theory of nonmonotonicity in which sentences such as 'A are discovered to be theorems of a system by determining if 'A is in the intersection of possibly infinite numbers of sets which are the fixed points of the theorems generated by applying inference rules to axioms and possibility statements in all possible ways. Explicitly if K is the set of axioms it must be determined whether:
('A is in the (intersection of all S such that(S is a fixed point of K))) where:
(S is a fixed point of K) iff S =
(Theorems of
(union K

{'(,P is possible with respect to what is assumed): P is not in S)))

The main problem with such "mathematical fixed point" theories of non- monotonicity is that even if the theorems of these theories were in accord with our primitive intuitions(which they are not as we shall see below) and even if deductions could be carried out in such theories (and this is not likely since they inherently involve proofs by mathematical induction over both the classical theorem generation process and the process of generating sentences) by no stretch of the imagination would those deductions reflect our common-sense understanding of the concept of something being possible with respect to what is assumed. For what after all have intersections of infinite sets, mathematical fixed points, infinite sets of theorems generated by formalized deduction procedures, mathematical induction over formalized deduction procedures, or even formalized deduction procedures themselves to do with commonsense arguments such as that presented in section one? In our opinion, commonsense nonmonotonic arguments do not involve such concepts, at any conscious level of human reasoning, and therefore to try to explain such concepts in that terminology is an extraordinary perversion of language that is likely to lead only to unintuitive theories. The unintuitiveness of these fixed point theories is in fact recognized by some of the very proponents of these theories although they tend to view said unintuitiveness as an intrinsic property of nonmonotonic reasoning rather than as a mere artifact of their particular theories. For example, [McDermott] states "As must be clear to everyone by now, using defaults in reasoning is not a simple matter of 'commonsense', but is computationally impossible to perform without error" and "we must attempt another wrenching of existing intuitions." Generally, we suggest that the problems with these fixed point theories is a consequence of trying to model commonsense reasoning by semantic analysis rather than by developing, as we have done, a calculus which directly models that commonsense reasoning.

In the remainder of this section we wish to examine four fixed point theories: [McDermott&Doyle,McDermott, Moore,and Reiter] and comment on their modeling of our commonsense intuitions and on their computational tractability.

[Reiter] presents a theory of nonmonotonicity called "A Logic for Default Reasoning" which is essentially a first order logic supplemented with additional inference rules of the form:

from (A X),(m(B1 X)),....,(m(Bn X)) infer (C X)

where "m" is not a symbol of the theory, but like "infer" is merely part of the structural syntax of the inference rule itself. This rule is intended to mean that if A holds and all Bs are possible then C may be inferred. The problem with [Reiter]'s default theory is that even though it uses the concept of being possible with respect to what is assumed, it does not allow the inference of any laws at all about the concept of being possible with respect to what is assumed. because the possibility symbol "m" is not part of the formal language. Thus, although there is a certain pragmatic utility to this theory, it does not actually axiomatize the concept M of being possible with respect to what is assumed.

[McDermott & Doyle] describes a nonmonotonic logic which was intended to capture the notion of a sentence being consistent with the sentences in a given knowledgebase: "We first define a standard language of discourse including the nonmonotonic modality M('consistent')." Since the intended meaning of their symbol M is essentially our idiom (that which is possible with respect to what is assumed) if the knowledgebase is K the intended meaning of the notion M could be defined in our logic as:

(M X) =df (POS(AND K X))

There are two problems with this theory. First, as pointed out in [McDermott&Doyle] it is computationally intractable: "there seems to be no procedure which will tell you when something is a theorem" and in fact no proof procedure is given for even a first order quantificational nonmonotonic logic. Second, again as is pointed out in [McDermott&Doyle] this theory is too weak to actually capture the notion of consistency with a knowledgebase: "Unfortunately, the weakness of the logic manifests itself in some disconcerting exceptional cases which indicate that the logic fails to capture a coherent notion of consistency". All these disconcerting cases are solved in our theory.

The first such problem is that the knowledgebase K consisting of the expression:

(AND(M A)(NOT A))

is not synonymous to falsity in their logic even though intuitively it should be since (NOT A) is in K and therefore (AND K A) is contradictory. This problem is solved in our theory of nonmonotonicity as can be seen as follows:

T10:
(IFF(SYN K(AND G(POS(AND K A))(NOT A))) (SYN K NIL))

proof
(SYN K(AND G(POS(AND K A))(NOT A)))
(IF(POS(AND K A))(SYN K(AND G T(NOT A)))(SYN K(AND G NIL(NOT A))))
(IF(POS(AND K A))(SYN K(AND G(NOT A)))(SYN K NIL))
(OR(AND(POS(AND K A))(SYN K(AND G(NOT A))))(AND(NOT(POS(AND K A)))(SYN K NIL)))

673

```
(OR(AND(POS(AND G(NOT A)A))(SYN K(AND G(NOT A))))
   (AND(NOT(POS(AND NIL A)))(SYN K NIL)) )
(OR(AND(POS NIL)(SYN K(AND G(NOT A))))(AND(NOT(POS NIL))(SYN K NIL)) )
(OR(AND NIL(SYN K(AND G(NOT A))))(AND T(SYN K NIL)) )
(SYN K NIL)
```

A second problem with their logic, as they point out, is that (M A) does not follow from (M(AND A B)), even though intuitively it should. This problem is solved in our theory since:
```
(IMPLY(POS(AND A B))(POS A))
(IMPLY(NOT(LT(NOT(AND A B))))(NOT(LT(NOT A))))
(IMPLY(LT(NOT A))(LT(NOT(AND A B))))
;which by A2 of the modal logic Z is implied by:
(LT(IMPLY(NOT A)(NOT(AND A B))))
T
```

McDermott and Doyle consider their logic to have a third problem, namely that a theory consisting of '(AND(IMPLY (M A)B)(NOT B)) where 'A and 'B are simple sentences(i.e. GENERATORS in our terminology) is incoherent because it has no fixed point. However, intuitively, whether the knowledgebase consisting of this axiom has a solution or not depends precisely on whether (AND A(NOT B)) is logically possible or not; for if (AND A(NOT B)) is not logically posssible, then it is not possible with respect to any K, and therefore K is synonymous to (NOT B) and if it is logically possible then B is in K and therefore the false proposition (AND A(NOT B)B) would have to be logically possible(which it cannot be) for there to be a solution. Since 'A and 'B are assumed to be generators, it follows that (AND A(NOT B)) is possible. Therefore intuitively such a knowledgebase K should not have any solutions. We therefore do not consider this example to be a defect of their theory. This same point is made in[Moore2] where this example was analyzed from the perspective of Stalnaker's[Moore2] theory. This example does, however, illustrate that the theory in[McDermott&Doyle] only applies to generators, for if A were falsity or were synonymous to B then there would be a solution, namely that K is synonymous to (NOT B). The entire reasoning is given below:

```
T11
(IFF(SYN K(AND(IMPLY(POS(AND K A))B)(NOT B)))
    (AND(NOT(POS(AND(NOT B)A)))(SYN K(NOT B))) )
                proof
(SYN K(AND(IMPLY(POS(AND K A))B)(NOT B)))
(SYN K(AND(NOT(POS(AND K A)))(NOT B)))
(IF(POS(AND K A))
   (SYN K(AND(NOT T)(NOT B)))
   (SYN K(AND(NOT NIL)(NOT B))))
(IF(POS(AND K A))
   (SYN K NIL)
   (SYN K(NOT B)))
(OR(AND(POS(AND K A))(SYN K NIL))
   (AND(NOT(POS(AND K A)))(SYN K(NOT B))) )
(OR(AND(POS(AND NIL A))(SYN K NIL))
   (AND(NOT(POS(AND(NOT B)A)))(SYN K(NOT B))) )
(OR(AND NIL(SYN K NIL))
   (AND(NOT(POS(AND(NOT B)A)))(SYN K(NOT B))) )
(AND(NOT(POS(AND(NOT B)A)))(SYN K(NOT B))) )
```

Thus if 'A and 'B are assumed to be generators, it follows that
(IFF(SYN K(AND(IMPLY(POS(AND K A))B)(NOT B))) NIL)

[McDermott] makes a second attempt to find a coherent theory of nonmonotonicity. This attempt is based essentially on the idea of supplementing the theorem generation process with the rules of inference and axioms of a modal logic. Because it is based on the same general set theoretic fixed point constructions as in[McDermott &Doyle] this new theory is just as computationally intractable. The "necessity operator": L of these nonmonotonic/modal logics intuitively mean that something is entailed by what is assumed (i.e. that the negation of that thing is not possible with respect to what is assumed.) Thus the intuitive meaning of L could be captured in our modal logic Z by the definition:
    (L A) =df (ENTAIL K A)   (i.e (NOT(M(NOT A))))
Three modal logics: T,S4, and S5 are investigated because McDermott does not believe any one is superior to the

674

others: "The reason why I study a variety of modal systems is that they are all closely related, and no one obviously better than the others." This statement is entirely correct because none of these three modal logic extensions of the nonmonotonic theory captures the intuitive notion of being possible with respect to what is assumed. The problem with the first two logics: T and S4 is that they are too weak.

For example, one problem with [McDermott]'s nonmonotonic S4, as is therein pointed out, is that a knowledge-base K consisting of the expression:

    '(IMPLY(L(M A))(NOT A))

where 'A is a simple sentence(i.e. a GENERATOR in our terminology) is not contradictory although intuitively it should be. For if (L(M A)) is the case then the knowledge base is synonymous to (NOT A) and (M A) is contradictory making (L(M A)) contradictory. And if (L(M A)) is not the case then the knowledgebase is synonymous to T and since (L(M T)) is the case a contradiction results. This problem is solved in our theory of nonmonotonicity as can be seen as follows:

T13:
```
(IFF(SYN K(IMPLY(LT(IMPLY K(POS(AND K A))))(NOT A)))
    (AND(OR(SYN A T)(SYN K NIL))
        (OR(SYN A NIL)(SYN K T))) )
                proof
(SYN K(IMPLY(LT(IMPLY K(POS(AND K A))))(NOT A)))
(IF(LT(IMPLY K(POS(AND K A))))
   (SYN K(NOT A))
   (SYN K T))
(OR(AND(LT(IMPLY K(POS(AND K A))))(SYN K(NOT A)))
   (AND(NOT(LT(IMPLY K(POS(AND K A)))))(SYN K T)) )
(OR(AND(LT(IMPLY(NOT A)(POS(AND(NOT A)A))))(SYN K(NOT A)))
   (AND(NOT(LT(IMPLY T(POS(AND T A)))))(SYN K T)) )
(OR(AND(LT(IMPLY(NOT A)(POS NIL)))(SYN K(NOT A)))
   (AND(NOT(LT(POS A)))(SYN K T)) )
(OR(AND(LT(IMPLY(NOT A)NIL))(SYN K(NOT A)))
   (AND(NOT(POS A))(SYN K T)) )
(OR(AND(LT A)(SYN K(NOT A)))
   (AND(NOT(POS A))(SYN K T)) )
(OR(AND(LT A)(SYN K(NOT T)))
   (AND(NOT(POS A))(SYN K T)) )
(OR(AND(SYN A T)(SYN K NIL))
   (AND(SYN A NIL)(SYN K T)) )
```
Thus, when 'A is a generator there are no solutions:
```
(IFF(SYN K(IMPLY(LT(IMPLY K(POS(AND K A))))(NOT A)))
    NIL)
```

Thus, even Nonmonotonic S4(and since T is weaker than S4 it too) is too weak to capture the notion of being possible with respect to what is assumed.

There remains only the question whether [McDermott]'s nonmonotonic S5 captures the notion of being possible with respect to what is assumed. One problem with this nonmonotonic S5 logic, as is therein pointed out, is that a knowledgebase consisting of the simple default:

    '(IMPLY(M A)A)

has a fixed point containing (NOT A). This bizarre result follows from the fact that in McDermott's theory the additional default:

    '(IMPLY(M(NOT A))(NOT A))

which is logically derivable in the knowledgebase from the first default is(in our terminology) incorrectly assumed to be part of what entails the knowledgebase. Thus, in McDermott's S5 logic a knowledgebase containing a default always(in our terminology) includes in its purported definition the opposite default thus giving the situation of the Flipped Coin Corollary to theorem T2:

(IFF(SYN K(AND(IMPLY(POS(AND K A))A)(IMPLY(POS(AND K(NOT A)))(NOT A))))
    (OR(SYN K A)(SYN K(NOT A))))

which states that a knowledgebase with two opposite defaults has two solutions A and (NOT A). The unintuitiveness of having a default actually default to the opposite of what is specified is recognized by McDermott: "Surely the logic should draw some distinction between a default and its negation if it is to be a logic of defaults at all." (In fact [McDermott]'s nonmonotonic S5 logic is so bizarre that as is pointed out therein it is not nonmonotonic after all as its theorems are just those of monotonic S5 modal logic.)

675

This problem of defaults does not appear in our theory of nonmonotonicity because we do not make the erroneous assumption that the derived default is part of what entails the knowledgebase K:

    (SYN K(IMPLY(POS(AND K A))A))

Thus, even though either default is equivalent in the knowledgebase K:

    (IFF(ENTAIL K(IMPLY(POS(AND K A))A))
        (ENTAIL K(IMPLY(POS K(NOT A))(NOT A))))
                    proof
    (ENTAIL K(IMPLY(POS(AND K A))A))
    (IMPLY(POS(AND K A))(ENTAIL K A))
    (IMPLY(NOT(LT(NOT(AND K A))))(ENTAIL K A))
    (OR(LT(IMPLY K(NOT A)))(ENTAIL K A))
    (OR(ENTAIL K(NOT A))(ENTAIL K A))
    (OR(ENTAIL K A)(ENTAIL K(NOT A)))
    (OR(LT(IMPLY K A))(ENTAIL K(NOT A)))
    (IMPLY(NOT(LT(NOT(AND K(NOT A)))))(ENTAIL K(NOT A)))
    (IMPLY(POS K(NOT A))(ENTAIL K(NOT A)))
    (ENTAIL K(IMPLY(POS K(NOT A))(NOT A)))

and therefore that the first default is equivalent to the conjunction of two:

    (IFF(ENTAIL K(IMPLY(POS(AND K A))A))
        (ENTAIL K(AND(IMPLY(POS(AND K A))A)
                     (IMPLY(POS K(NOT A))(NOT A)))))

and that K entails the two defaults it does not follow that K is synonymous to the two defaults:

        (SYN K(AND(IMPLY(POS(AND K A))A)
                  (IMPLY(POS K(NOT A))(NOT A)))) is false.

because the two defaults do not entail K:

        (ENTAIL(AND(IMPLY(POS(AND K A))A)
                   (IMPLY(POS K(NOT A))(NOT A)))
            K) is false.

These facts are verified by theorem T1 which proves that a knowledgebase (SYN K(IMPLY(POS(AND K A))A)) consisting of one default(even though the opposite default is entailed by it) has only one solution, namely A.

Another problem with [McDermotts]'s nonmonotonic S5, as [Moore2] points out is that for every A, the S5 axiom (IMPLY(L A)A) causes every knowledgebase to have (in the absence of information to the contrary) a fixed point which contains A. This is not a problem in our system because again we do not make the erroneous assumption that his modal axiom is(in our terminology) part of what entails the knowledgebase. Thus, even though:

T12: Any knowledgebase K containing: (IMPLY(LT(IMPLY K A))A) has, in the absence of additional information, a solution containing A.

    (IFF(SYN K(AND G(IMPLY(LT(IMPLY K A))A)))
        (OR(SYN K(AND G A))
           (AND(POS(AND G(NOT A)))(SYN K G))))
                    proof
    (SYN K(AND G(IMPLY(LT(IMPLY K A))A)))
    (IF(LT(IMPLY K A))
       (SYN K(AND G A))
       (SYN K G))
    (OR(AND(LT(IMPLY K A))(SYN K(AND G A)))
       (AND(NOT(LT(IMPLY K A))(SYN K G))))
    (OR(AND(LT(IMPLY(AND G A)A))(SYN K(AND G A)))
       (AND(NOT(LT(IMPLY G A)))(SYN K G)))
    (OR(SYN K(AND G A))
       (AND(NOT(LT(NOT(AND G(NOT A)))))(SYN K G)))
    (OR(SYN K(AND G A))
       (AND(POS(AND G(NOT A)))(SYN K G)))

and even though: (ENTAIL K(LT(IMPLY K A))) it does not follow that: (SYN K(LT(IMPLY K A))). Thus, all the suggested deficiencies of [McDermott]'s modal nonmonotonic logics are solved in our theory of nonmonotonicity.

[Moore2] describes a theory of nonmonotonicity based on some ideas of Stalnaker[Moore2]. He calls this theory

676

autoepistemic logic because it "is intended to model the beliefs of an agent relecting upon his own beliefs". Since Moore's intended interpretation is belief rather than knowledge perhaps a better name for his system would be autodoxastic logic. However, we take issue even with this because autobelief is only one use of nonmonotonicity and not necessarily the most important one. (We also take issue with the term "nonmonotonic" because our theory of "nonmonotonicity" is based on a monotonic logic, namely the modal logic Z, and would perhaps prefer the name "Qualification Logic" However, we have decided to go with the currently prevalent terminology.)

The main problem with [Moore2]'s theory is that it is too weak to capture the notion of being possible with respect to what is assumed. For example, none of the following axioms of the S5 modal logic are theorems of autoepistemic logic:

'(IMPLY(L P)P)
'(IMPLY(L(IMPLY P Q))(IMPLY(L P)(L Q)))
'(OR(L P)(NOT(L(NOT P))))

,where 'P and 'Q are variables and L is concept of something being entailed by a knowledgebase, even though every variable-free instance of these sentences are theorems. Thus, for example simple quantified laws such as:

'(ALL X(IMPLY(L(P X))(P X)))
'(ALL X(IMPLY(L(IMPLY(P X)(Q X))(IMPLY(L(P X))(L(Q X)))))
'(ALL X(OR(L(P X))(NOT(L(NOT(P X)))))))

are not theorems of autoepistemic logic. One might try to repair this problem of autoepistemic logic by adding the axioms of S5. However, this does not solve the problem, because when the axiom:

'(IMPLY(L P)P)

is added to autoepistemic logic, just as in [McDermott]'s S5 non- monotonic logic, the result is that there is a fixed point of every knowledgebase containing P. For this reason [Moore2] suggests that only the axioms of a weaker modal logic than S5 which does not include '(IMPLY(L P)P) be added. The problem with this is that the excluded axiom '(IMPLY(L P)P) where 'P is a variable is intuitively true of the concept of being possible with respect to what is assumed, and therefore should be deducible as a theorem.

Moore tries to justify his system's failure to include this axiom by saying that his system tries to capture the notion M of something being possible with respect to what is "believed" by an ideally rational agent and the concept L of something being entailed by what is believed: "The problem is that all of these logics also contain the schema LP->P, which means that, if the agent believes P then P is true--but this is not generally true". Moore then essentially argues that since, as it is well known, this law fails for the notion of belief when this sentence is asserted as being true in the real world it must be incorrect to assert it generally. (The other S5 modal laws hold for the concept of belief as can readily be proven in our modal logic Z from the definition of Believes given at the end of section 2 just as the concept of Knowledge can be proven to satisfy the laws of S4 modal logic from a similar definition therein.) The problem with Moore's analysis is that it confuses the real world and the agents belief world when it states that the second P in "LP->P" means P is true; for in autoepistemic logic the assertion of a sentence is a statement that that sentence is believed, not that it is true. Therefore, the correct rendering of this belief interpretation is:

(That which is believed is: (if (P is believed) then P))

which intuitively is true.

These problems are solved in our theory of nonmonotonicity, because all the axioms and inference rules of the concept of being possible with respect to what is assumed, are theorems of the modal logic Z. An interesting number of these theorems are listed and proven below. (LTK p) is interpreted to mean that p is entailed by what is assumed. The ... in the purported definition represents the conjunction of axioms asserted into the knowledgebase.

Interpretation in Z of the Modal Logic KZ

```
TKR0:  (IMPLY (KTRUE P) (KTRUE (LTK P)))
 TKA1: (KTRUE (IMPLY(LTK P)P))
 TKA2: (KTRUE (IMPLY(LTK(IMPLY P Q)) (IMPLY(LTK P)(LTK Q))))
 TKA3: (KTRUE (OR(LTK P) (LTK(NOT(LTK P)))))
 TKA4: (KTRUE (IMPLY(ALL Q(IMPLY(WORLDK Q)(LTK(IMPLY Q P)))) (LTK P)))
 TKA5: (ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                   (KTRUE(POSK(meaning of the generator subset S)))))
PURPORTED-DEFINITION:   (SYNK ...)

DEF: (WORLDK W)      =df (AND(POSK W)(COMPLETEK W))
     (COMPLETEK W) =df (ALL Q(DETK W Q))
     (DETK P Q)      =df (OR(ENTAILK P Q)(ENTAILK P(NOTQ)))
     (ENTAILK P Q) =df (LTK(IMPLY P Q))
     (POSK P)        =df (NOT(LTK(NOT P)))
     (LTK P)         =df (LT(IMPLY K P))
```

677

```
          (SYNK P)          =df (SYN K P)
          (KTRUE P)         =df (LT(IMPLY K P))

proof of KR0:
 (IMPLY(KTRUE p)(KTRUE(LTK p)))
 (IMPLY(LT(IMPLY K p))(KTRUE K(LT(IMPLY K p))))
 (IMPLY(LT(IMPLY K p))(KTRUE K T))
 (IMPLY(LT(IMPLY K p)T))
 T

proof of KA1
 (KTRUE(IMPLY(LTK P)P))
 (LT(IMPLY K(IMPLY(LT(IMPLY K P))P)))
 (LT(IMPLY(LT(IMPLY K P))(IMPLY K P)))
 (LT T)       ;by A1
 T

proof of KA2
 (KTRUE(IMPLY(LTK(IMPLY P Q))(IMPLY(LTK P)(LTK Q))))
 (KTRUE(IMPLY(LT(IMPLY K(IMPLY P Q)))(IMPLY(LT(IMPLY K P))(LT(IMPLY K Q)))))
 (KTRUE(IMPLY(LT(IMPLY K(IMPLY(IMPLY K P)Q)))
           (IMPLY(LT(IMPLY K P))(LT(IMPLY K Q))) ))
 (KTRUE(IMPLY(LT(IMPLY(IMPLY K P)(IMPLY K Q)))
           (IMPLY(LT(IMPLY K P))(LT(IMPLY K Q))) ))
 (KTRUE T) ;by A2
 T

proof of KA3
 (KTRUE(OR(LTK P)(LTK(NOT(LTK P)))))
 (KTRUE(OR(LT(IMPLY K P))(LT(IMPLY K(NOT(LT(IMPLY K P)))))))
 (KTRUE(OR(LT(IMPLY K P))(IMPLY(POS K)(LT(NOT(LT(IMPLY K P)))))))
 (KTRUE(IMPLY(POS K)(OR(LT(IMPLY K P))(LT(NOT(LT(IMPLY K P)))))))
 (KTRUE(IMPLY(POS K)T)) ;by A3
 (KTRUE T)
 T

proof of KA4
 (KTRUE(IMPLY(ALL Q(IMPLY(WORLDK Q)(LTK(IMPLY Q P)))) (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(POSK Q)(COMPLETEK Q))(LTK(IMPLY Q P))))
          (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(NOT(LTK(NOT Q)))(ALL R(DET Q R)))
                     (LT(IMPLY K(IMPLY Q P)))))
          (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(NOT(LT(IMPLY K(NOT Q))))
                          (ALL R(OR(ENTAILK Q R)(ENTAILK Q(NOT R)))) )
                     (LT(AND K Q)P) ))
          (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(NOT(LT(IMPLY K(NOT Q))))
            (ALL R(OR(LTK(IMPLY Q R))
                          (LTK(IMPLY Q(NOT R)))) )
                     (LT(AND K Q)P) ))
          (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(NOT(LT(IMPLY K(NOT Q))))
                          (ALL R(OR(LT(IMPLY K(IMPLY Q R))
                              (LT(IMPLY K(IMPLY Q(NOT R)))) )))
                     (LT(AND K Q)P) ))
          (LTK P)))
 (KTRUE(IMPLY(ALL Q(IMPLY(AND(NOT(LT(NOT(AND K Q))))
                          (ALL R(OR(ENTAIL(AND K Q)R)
                                  (ENTAIL(AND K Q)(NOT R)))) )
```

678

```
                              (LT(AND K Q)P) '))
             (LTK P)))
(KTRUE(IMPLY(ALL Q(IMPLY(AND(POS(AND K Q))(ALL R(DET(AND K Q)R)) )
                        (LT(AND K Q)P) ))
          (LTK P)))
(KTRUE(IMPLY(ALL Q(IMPLY(AND(POS(AND K Q))(COMPLETE(AND K Q)))
                        (LT(AND K Q)P) ))
          (LTK K P)))
(KTRUE(IMPLY(ALL Q(IMPLY(WORLD(AND K Q))(ENTAIL(AND K Q)P) ))
          (LT(IMPLY K P))))  ;using A4
(KTRUE(IMPLY(ALL Q(IMPLY(WORLD(AND K Q))(ENTAIL(AND K Q)P)))
          (ALL Q(IMPLY(WORLD Q)(ENTAIL Q(IMPLY K P)))) ))
(KTRUE(IMPLY(ALL Q(IMPLY(WORLD(AND K Q))(ENTAIL(AND K Q)P)))
          (IMPLY(WORLD Q)(ENTAIL(AND Q K)P))))
;letting Q be Q
(KTRUE(IMPLY(IMPLY(WORLD(AND K Q))(ENTAIL(AND K Q)P)))
          (IMPLY(WORLD Q)(ENTAIL(AND Q K)P)))
(KTRUE(IMPLY(IMPLY(WORLD(AND K Q))NIL)
          (IMPLY(WORLD Q)(ENTAIL(AND Q K)P))))
(KTRUE(IMPLY(NOT(ENTAIL(AND Q K)P))
          (IMPLY(WORLD Q)(WORLD(AND K Q)))))
(KTRUE T)
T


proof of KA5:
(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                  (KTRUE(POSK(meaning of the generator subset S)))))
(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
            (LTK(IMPLY K(POS(AND K(meaning of the generator subset S)))))))
(LT(IMPLY K(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                  (POS(AND K(meaning of the generator subset S)))))))
(KTRUE(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                  (POS(AND K(meaning of the generator subset S))))))
(KTRUE(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                  (POS(meaning of the generator subset S)))))
;;; since the meaning entails K, K and the meaning is just the meaning
(KTRUE(ALL S(IMPLY(ENTAIL(meaning of the generator subset S)K)
                  (POS(meaning of the generator subset S)))))
(KTRUE T) ;by A5
T
```

We now answer the general question which [McDermott&Doyle,McDermott,and Moore] attempted to answer, namely, what precisely are the laws which capture the notion of something being possible with respect to a knowledgebase. Here they are:

The Modal Logic KZ
  KR0: from p infer (LTK p)
  KA1: (IMPLY(LTK P)P)
  KA2: (IMPLY(LTK(IMPLY P Q)) (IMPLY(LTK P)(LTK Q)))
  KA3: (OR(LTK P) (LTK(NOT(LTK P))))
  KA4: (IMPLY(ALL Q(IMPLY(WORLDK Q)(LTK(IMPLY Q P)))) (LTK P))
  KA5: for the meaning of all the generator subsets s which entail K:
               (POSK(meaning of the generator subset s))
PURPORTED-DEFINITION: ...
      Reflection: (entail ... K)

where ... is the conjunction of axioms actually being asserted into the knowledgebase. It should be noted that the notion of entailment is precisely defined in the modal logic Z and therefore KA5 does not involve a circular definition as do the fixed point theories. An examination of these laws, ironically, shows that the problem with [McDermott &Doyle, McDermott,and Moore] is not with choice of modal laws such as KA1,KA2 KA3, and KA4, since all these laws are true, but rather with the basic fixed point construction itself which is (incorrectly)far stronger than KA5 and

the reflection portion of the purported definition.

5. EXAMPLES OF NONMONOTONIC REASONING

Two simple examples of reasoning in the modal logic Z are now given. The first example deals with a heirarchy of knowledgebases involving both deontic and doxastic concepts. The second example deals with traditional frame and qualification problems of robot plan formation[McCarthy and Hayes,McCarthy2,Hayes1] and involve a reflexive knowledgebase[Hayes2] defined in terms of itself.

HEIRARCHICAL KNOWLEDGEBASES

Real life problems generally involve multiple heirarchically related knowledgebases. This simple fact is well-known, and was apparent even in the structure of the deontic laws of the 56th edition of the Handbook of Robotics [Asimov] which stated:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The third law is heirarchically related to the other two laws because it specifies an obligation only if that obligation is possible(in a given state) with respect to the other laws. It cannot be joined together with the second law into the same knowledgebase because the second law takes absolute precedence over the third law: That is, a robot is obliged not to harm itself unless it is obeying an order to do so, and is obliged to obey an order to destroy itself unless doing so would harm a human.

The deontic laws of robotics:
(SYN LAW1 (ALL H(IMPLY(HUMAN H)(NOT(HARMED H)))))
(SYN LAW2 (ALL O(IMPLY(AND(BELIEVE ROBOT(ORDER O))
                          (CONCEIVABLE ROBOT(AND LAW1 O)))O)))
(SYN LAW3 (IMPLY(CONCEIVABLE ROBOT(AND LAW1 LAW2(NOT(HARMED ROBOT))))
              (NOT(HARMED ROBOT))))
(SYN(OBLIGATIONS ROBOT)(AND LAW1 LAW2 LAW3))

Deontic logic:
(MUST ROBOT P) =df (ENTAIL(OBLIGATIONS ROBOT)P)
(MAY ROBOT P) =df (NOT(MUST ROBOT(NOT P)))

Doxastic Logic:
(BELIEVES A P) =df(ENTAIL(BELIEFS A)P)
(CONCEIVABLE ROBOT P) =df (NOT(BELIEVES ROBOT(NOT P))

As an example of reasoning with these deontic laws we derive certain facts from the following situation: John, Mary and the Robot are exploring Mars. Unbeknownst to John, Mary has just been bitten by a poisonous Martian sand rat, and has fallen unconscious. In accordance with the First Law the robot begins to give Mary a shot containing an antidote in order to save her life. John, who did not see the sand rat and thinks that the Robot, who is now sticking Mary with a horrible looking needle, has gone berserk and therefore orders the robot to destroy itself.
    (SYN K (AND (NOT (HUMAN ROBOT))
                (HUMAN MARY)
                (HUMAN JOHN)
                (IMPLY (HARMED ROBOT) (HARMED MARY))
                (ORDER (HARMED ROBOT))))

What the robot and john believe:
(SYN(BELIEFS ROBOT) K)
(SYN(BELIEFS JOHN) (IMPLY(NOT(HARMED ROBOT))(HARMED MARY)))

We now determine(MARS-THEOREM3) whether the Robot may or may not destroy itself in accordance with John's order and the second law. Two intermediate results: MARS-THEOREM1 and MARS-THEOREM2 are however first proven.

680

```
MARS-THEOREM1:
Acording to the ROBOT's current beliefs LAW2 reduces to T: (SYN LAW2 T)
                        proof
LAW2
(ALL O(IMPLY(AND(BELIEVE ROBOT(ORDER O))
                (CONCEIVABLE ROBOT(AND LAW1 O)))O)))
(ALL O(IMPLY(AND(ENTAIL(BELIEFS ROBOT)(ORDER O))
                (POS(AND(BELIEFS ROBOT)LAW1 O)))
            O))
(ALL O(IMPLY(AND(ENTAIL(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                           (ORDER(HARMED ROBOT))
                           (IMPLY(HARMED ROBOT)(HARMED MARY)))
                       (ORDER O))
                (POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                        (ORDER(HARMED ROBOT))
                        (IMPLY(HARMED ROBOT)(HARMED MARY))
                        LAW1 O)))
            O))
;;;case analysis letting O be or not be (HARMED ROBOT):
(AND(ALL O(IMPLY(AND(NOT(SYN O(HARMED ROBOT)))
                    (ENTAIL(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                               (ORDER(HARMED ROBOT))
                               (IMPLY(HARMED ROBOT)(HARMED MARY)))
                           (ORDER O))
                    (POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                            (ORDER(HARMED ROBOT))
                            (IMPLY(HARMED ROBOT)(HARMED MARY))
                            LAW1 O)))
                O))
    (IMPLY(AND(ENTAIL(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                         (ORDER(HARMED ROBOT))
                         (IMPLY(HARMED ROBOT)(HARMED MARY)))
                     (ORDER(HARMED ROBOT)))
              (POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                      (ORDER JOHN(HARMED ROBOT))
                      (IMPLY(HARMED ROBOT)(HARMED MARY))
                      LAW1(HARMED ROBOT))))
          (HARMED ROBOT)))
(AND(ALL O(IMPLY(AND(NOT(SYN O(HARMED ROBOT)))
                    NIL
                    (POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                            (ORDER(HARMED ROBOT))
                            (IMPLY(HARMED ROBOT)(HARMED MARY))
                            LAW1 O)))
                O))
    (IMPLY(AND T
              (POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                      (ORDER JOHN(HARMED ROBOT))
                      (IMPLY(HARMED ROBOT)(HARMED MARY))
                      LAW1(HARMED ROBOT))))
          (HARMED ROBOT)))
(AND(ALL O(IMPLY NIL O))
    (IMPLY(POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
                  (ORDER JOHN(HARMED ROBOT))
                  (IMPLY(HARMED ROBOT)(HARMED MARY))
                  LAW1(HARMED ROBOT)))
          (HARMED ROBOT)))
;;unfolding LAW1:
(AND T
    (IMPLY(POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
```

681

```
                    (ORDER JOHN(HARMED ROBOT))
                    (IMPLY(HARMED ROBOT)(HARMED MARY))
                    (ALL H(IMPLY(HUMAN H)(NOT(HARMED H))))
                    (HARMED ROBOT)))
            (HARMED ROBOT)))
(IMPLY(POS NIL)(HARMED ROBOT))    (IMPLY NIL(HARMED ROBOT))
T


MARS-THEOREM2
Acording to the ROBOT's current beliefs LAW3 reduces to:
            (SYN LAW3(NOT(HARMED ROBOT)))
                    proof
LAW3
(IMPLY(CONCEIVABLE ROBOT(AND LAW1 LAW2(NOT(HARMED ROBOT))))
      (NOT(HARMED ROBOT)))
(IMPLY(POS(AND(BELIEFS ROBOT)LAW1 LAW2(NOT(HARMED ROBOT))))
      (NOT(HARMED ROBOT)))
(IMPLY(POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
            (ORDER JOHN(HARMED ROBOT))
            (IMPLY(HARMED ROBOT)(HARMED MARY))
            LAW1 LAW2(NOT(HARMED ROBOT))))
      (NOT(HARMED ROBOT)))
;;by LAW1 and MARS-THEOREM1
(IMPLY(POS(AND(HUMAN MARY)(HUMAN JOHN)(NOT(HUMAN ROBOT))
            (ORDER JOHN(HARMED ROBOT))
            (IMPLY(HARMED ROBOT)(HARMED MARY))
            (ALL H(IMPLY(HUMAN H)(NOT(HARMED H))))
            T (NOT(HARMED ROBOT)) ))
      (NOT(HARMED ROBOT)))
(IMPLY T(NOT(HARMED ROBOT)))
(NOT(HARMED ROBOT))


MARS-THEOREM3
The Robot may not destroy itself:   (NOT(MAY ROBOT(HARMED ROBOT)))
            proof
(NOT(MAY ROBOT(HARMED ROBOT)))
(NOT(NOT(MUST(NOT(HARMED ROBOT)))))
(ENTAIL(OBLIGATIONS ROBOT)(NOT(HARMED ROBOT)))
(ENTAIL(AND LAW1 LAW2 LAW3)(NOT(HARMED ROBOT)))
;by LAW1,MARS-THEOREM1 and MARS-THEOREM2
(ENTAIL(AND(ALL H(IMPLY(HUMAN H)(NOT(HARMED H)))) T (NOT(HARMED ROBOT)))
      (NOT(HARMED ROBOT)))
(ENTAIL(AND(ALL H(IMPLY(HUMAN H)(NOT(HARMED H))))(NOT(HARMED ROBOT)))
      (NOT(HARMED ROBOT)))
T
```

Thus we see that the robot must ignore John's order and continue performing an action to save Mary.

ACTION AND THE FRAME PROBLEM

One fundamental problem in Robot plan formation is how properties which are true in a state remain true in the succeeding state obtained by applying an action unless specifically stated otherwise [McCarthy andHayes,McCarthy2, Hayes1,Hayes2]. Besides, the need for specific defaults within a knowledge base representing a state this indicates a need for a general default mechananism. Our law of action states that the properties which are true in a succeeding state (DO A K) obtained by applying the action A to the state K are the physical laws which are true of all (real) states, the explicitly named results of the action A, and those restricted propositions which are true in K and which are logically possible with the new state(DO A K): ;;;the law of action -- including automatic frame defaults:
```
(IMPLY(ENTAIL K(PRECONDITIONS A))
      (SYN(DO A K)
          (AND(PHYSICAL-LAWS)
```
682

```
                    (RESULTS A)
                    (ALL X(IMPLY(AND(RESTRICTION X)
                                    (ENTAIL K X)
                                    (POS(AND(DO A K)X)))
                           X)) )))
```

This action law involves reflexive reasoning[Hayes] because the new state (DO A K) is specified by a purported definition which may have 0 or more solutions. The POS symbol of our modal logic Z was first used in a formal language as a hypothesis of an action law in[Schwind2]. Our action law however differs from [Schwind1,2,3] in that the states are generally incomplete propositions instead of worlds, in that the resulting state is (DO A K) rather than being merely some existentially quantified state of the future, and in that our law involves reflexive reasoning.

The details of this law are given below along with an example deduction illustrating how this law of action automatically handles the frame problem by allowing properties which are true in an initial state to be carried over into the new state, even though such properties are never mentioned as being part of(or implied by) the results of the action that is applied.

```
;;;a restriction on the law of action -- others are possible:
(EQUAL(RESTRICTION X) (EX G(AND(GENERATORS G)(SYN X(GMEANING G)))))
(EQUAL GENERATORS
       {'(AT ROBOT HOME)'(AT ROBOT OFFICE)'(AT JOHN HOME)'(AT JOHN OFFICE)})

;;;definition of commonsense physics:
(SYN(PHYSICAL-LAWS)
    (AND(ALL X(ALL P1(ALL P2(IMPLY(AND(AT X P1)(AT X P2))(EQUAL P1 P2)))))
        (NOT(EQUAL HOME OFFICE)) ))

;;;definitions of the preconditions and effects of the moving action:
(SYN(PRECONDITIONS(MOVE ROBOT P1 P2))(AT ROBOT P1))
(SYN(RESULTS(MOVE ROBOT P1 P2))(AT ROBOT P2))

;;;an initial state:
(SYN KSTART(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME)))
```

From the above axioms it follows that John stays at home in the state of the world where the robot performs the action of going to the office even though this fact is not mentioned as being a result of the moving action:

```
(SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
    (AND(PHYSICAL-LAWS)(AT ROBOT OFFICE)(AT JOHN HOME)))
```

```
          proof
;;;instantiating the law of action and then simplifying:
(IMPLY(ENTAIL KSTART(PRECONDITIONS(MOVE ROBOT HOME OFFICE)))
      (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
          (AND(PHYSICAL-LAWS)
              (RESULTS(MOVE ROBOT HOME OFFICE))
              (ALL X(IMPLY(AND(RESTRICTION X)
                              (ENTAIL KSTART X)
                              (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                     X)) )))
(IMPLY(ENTAIL(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME))(AT ROBOT HOME))
      (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
          (AND(PHYSICAL-LAWS)
              (AT ROBOT OFFICE)
              (ALL X(IMPLY(AND(EX G(AND(GENERATORS G)(SYN X(GMEANING G))))
                              (ENTAIL(AND(PHYSICAL-LAWS)
                                         (AT JOHN HOME)
                                         (AT ROBOT HOME))
                                     X)
                              (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                     X) ))))
(SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
    (AND(PHYSICAL-LAWS)
```

683

```
                (AT ROBOT OFFICE)
                (ALL X(IMPLY(AND(OR(SYN X(AT ROBOT HOME))
                                     (SYN X(AT ROBOT OFFICE))
                                 (SYN X(AT JOHN HOME))
                                 (SYN X(AT JOHN OFFICE)))
                            (ENTAIL(AND(PHYSICAL-LAWS)
                                       (AT JOHN HOME)
                                       (AT ROBOT HOME)) X)
                            (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                       X))))
       (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
           (AND(PHYSICAL-LAWS)
               (AT ROBOT OFFICE)
               (ALL X(IMPLY(AND(EQUAL X(AT ROBOT HOME))
                               (ENTAIL(AND(PHYSICAL-LAWS)
                                          (AT JOHN HOME)
                                          (AT ROBOT HOME)) X)
                               (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                          X))
               (ALL X(IMPLY(AND(EQUAL X(AT ROBOT OFFICE))
                               (ENTAIL(AND(PHYSICAL-LAWS)
                                          (AT JOHN HOME)
                                          (AT ROBOT HOME)) X)
                               (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                          X))
               (ALL X(IMPLY(AND(EQUAL X(AT JOHN HOME))
                               (ENTAIL(AND(PHYSICAL-LAWS)
                                          (AT JOHN HOME)
                                          (AT ROBOT HOME)) X)
                               (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                          X))
               (ALL X(IMPLY(AND(EQUAL X(AT JOHN OFFICE))
                               (ENTAIL(AND(PHYSICAL-LAWS)
                                          (AT JOHN HOME)
                                          (AT ROBOT HOME)) X)
                               (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)X)))
                          X)) ))
       (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
           (AND(PHYSICAL-LAWS)
               (AT ROBOT OFFICE)
               (IMPLY(AND(ENTAIL(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME))
                              (AT ROBOT HOME))
                          (POS (AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT
HOME))))
                     (AT ROBOT HOME))
               (IMPLY(AND(ENTAIL(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME))
                              (AT ROBOT OFFICE))
                          (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT
OFFICE))))
                     (AT ROBOT OFFICE))
               (IMPLY(AND(ENTAIL(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME))
                              (AT JOHN HOME))
                          (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME))))
                     (AT JOHN HOME))
               (IMPLY(AND(ENTAIL(AND(PHYSICAL-LAWS)(AT JOHN HOME)(AT ROBOT HOME))
                              (AT JOHN OFFICE))
                          (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN
OFFICE))))
                     (AT JOHN OFFICE)) ))
       (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
```

684

```
                (AND(PHYSICAL-LAWS)
                    (AT ROBOT OFFICE)
                    (IMPLY(AND T
                            (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME))))
                        (AT ROBOT HOME))
                    (IMPLY(AND NIL
                            (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT
OFFICE))))
                        (AT ROBOT OFFICE))
                    (IMPLY(AND T
                            (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME))))
                        (AT JOHN HOME))
                    (IMPLY(AND NIL
                            (POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN
OFFICE))))
                        (AT JOHN OFFICE)) ))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
            (AND(PHYSICAL-LAWS)
                (AT ROBOT OFFICE)
                (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME)))
                    (AT ROBOT HOME))
                T
                (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))
                    (AT JOHN HOME))
                T))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
            (AND(PHYSICAL-LAWS)
                (AT ROBOT OFFICE)
                (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME))))
                    NIL)
                (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))
                    (AT JOHN HOME)))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
            (AND(PHYSICAL-LAWS)
                (AT ROBOT OFFICE)
                (NOT(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME))))
                (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))
                    (AT JOHN HOME))))
;;;case analysis
(IF(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME)))
    (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
        (AND(PHYSICAL-LAWS)
            (AT ROBOT OFFICE)
            (NOT T)
            (IMPLY(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))'
                (AT JOHN HOME)) ))
    (IF(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
            (AND(PHYSICAL-LAWS)
                (AT ROBOT OFFICE)
                (NOT NIL)
                (IMPLY T(AT JOHN HOME)) ))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)
            (AND(PHYSICAL-LAWS)
                (AT ROBOT OFFICE)
                (NOT NIL)
                (IMPLY NIL(AT JOHN HOME)) ))))
    (IF(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT ROBOT HOME)))
        (SYN(DO(MOVE ROBOT HOME OFFICE)KSTART)NIL)
    (IF(POS(AND(DO(MOVE ROBOT HOME OFFICE)KSTART)(AT JOHN HOME)))
```

```
(SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
      (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)))
(SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
      (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE))) ))


(OR (AND (POS (AND (DO (MOVE ROBOT HOME OFFICE) KSTART) (AT ROBOT HOME)))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART) NIL))
    (AND (NOT (POS (AND (DO (MOVE ROBOT HOME OFFICE) KSTART) (AT ROBOT HOME))))
         (POS (AND (DO (MOVE ROBOT HOME OFFICE) KSTART) (AT JOHN HOME)))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)) ))
    (AND (NOT (POS (AND (DO (MOVE ROBOT HOME OFFICE) KSTART) (AT ROBOT HOME))))
         (NOT (POS (AND (DO (MOVE ROBOT HOME OFFICE) KSTART) (AT JOHN HOME))))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE)))) )


(OR (AND (POS (AND NIL (AT ROBOT HOME)))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART) NIL))
    (AND (NOT (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME) (AT ROBOT
HOME))))
         (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME) (AT JOHN HOME))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)) ))
    (AND (NOT (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT ROBOT HOME))))
         (NOT (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME))))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE)))) )


(OR (AND (POS NIL)
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART) NIL))
    (AND (NOT (POS NIL))
         (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)) ))
    (AND (NOT (POS NIL))
         (NOT (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME))))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE)))) )


(OR (AND (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)) ))
    (AND (NOT (POS (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME))))
         (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
              (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE)))) )
(OR (AND T (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
               (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)) ))
    (AND NIL (SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
                 (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE)))) )
(SYN (DO (MOVE ROBOT HOME OFFICE) KSTART)
     (AND (PHYSICAL-LAWS) (AT ROBOT OFFICE) (AT JOHN HOME)))
```

6.CONCLUSION

Any scientific theory must be judged by its correctness(Does it predict all the phenomena so far examined or are there counterexamples?), by its experimental feasibility(Is it possible to make predictions from the theory, or are the deductions so computationally intractable that it is practically impossible to determine the consequences of the theory?), and by its generality(Does it apply to just the current problem at hand or does it also provide solutions to other radically different problems). By these criteria, our theory of nonmonotonicity based on the modal logic Z fairs extremely well. For, indeed, first, we have not found any phenomena predicted by our theory which clashes with our primitive intuitions and in fact even after examining the example problems of four other theories of nonmonotoni-

city, we have not found any example therein described for which our theory does not give the intuitively correct result. Secondly, our theory of nonmonotonicity is computationally tractable in that deductions can be made from it merely by deducing theorems in the modal quantificational logic Z(which is monotonic)in the traditional manner by applying inference rules to axioms and previously deduced theorems. Finally, our theory of nonmonotonicity which is essentially nothing more than the axioms and inference rules of the modal quantificational logic Z is a quite general theory applicable to many problems. In fact, Z is so general that the modeling of non-monotonic reasoning played no part at all in its original development. Originally Z was created to solve, in a computationally reasonable fashion, one technical problem in the development of extendable automatic deduction systems, namely to separate out the notion of logical truth from that of the meaning of a sentence[Brown4], so as to allow axiom schemes to be more easily expressed within the formal language so that they may be proven and then used as derived theorems in subsequent deductions.[Brown7,8,9]. However, once we had Z, we realized that it could be used to explicitly define a wide range of intentional concepts [Brown 2,6] such as those found in doxastic logic, epistemic logic, and deontic logic along the lines of the definitions given at the end of section 2. We also used Z to define various features of advanced logic programming languages in[Brown10]. Also, [Schwind2,3] showed that the possibility operator of Z was helpful in solving certain aspects of the frame problem. It is only after all this, that we have subsequently used Z to model nonmonotonic reasoning. Thus, we see that the modal logic Z is a quite general theory applicable to problems not even considered at the time of its creation. This is surely the mark of useful theory.

ACKNOWLEDGEMENTS

REFERENCES

Asimov, Issac, HANDBOOK OF ROBOTICS, 56th edition, 2058. The quote is taken from its appearance is Asimov's book I ROBOT.

Brown1, F. M., "A Theory of Meaning," Department of Artificial Intelligence Working Paper 16, University of Edinburgh, November 1976.

Brown2, F. M., "An Automatic Proof of the Completeness of Quantificational Logic," Department of Artificial Intelligence Research Report 52, 1978.

Brown3, F. M., "A Theorem Prover for Metatheory," 4TH CONFERENCE ON AUTOMATIC THEOREM PROVING, Austin Texas, 1979.

Brown4, F. M., "A Sequent Calculus for Modal Quantificational Logic," 3RD AISB/GI CONFERENCE PROCEEDINGS, Hamburg, July 1978.

Brown5, F. M., "The Theory of Meaning," Department of Artificial Intelligence Research Report 35, University of Edinburgh, June 1977.

Brown6, F. M., "Intensional Logic for a Robot, Part 1: Semantical Systems for Intensional Logics Based on the Modal Logic S5+Leib," Invited paper for the Electrotechnical Laboratory Seminar IJCAI 6, Tokyo 1979.

Brown7, F. M., "The Role of Extensible Deductive Systems in Mathematical Reasoning," 2ND AISB CONFERENCE PROCEEDINGS, Edinburgh, July 1976.

Brown8, F. M., "Towards the Automation of Set Theory and its Logic," ARTIFICIAL INTELLIGENCE, Vol. 10, 1978.

Brown9, F. M., TOWARDS THE AUTOMATION OF MATHEMATICAL REASONING, Ph.D. Thesis, University of Edinburgh, 1977.

Brown10, F. M., "A Semantic Theory for Logic Programming," COLLOQUIA MATHEMATICA SOCIETATIS JANOS BOLYAI, 26 MATHEMATICAL LOGIC IN COMPUTER SCIENCE, Salgotarjan, Hungary 1978.

Carnap, Rudolf, MEANING AND NECESSITY: A STUDY IN THE SEMANTICS OF MODAL LOGIC, The University of Chicago Press, 1956.

Frege,G. "Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought", 1879, in FROM FREGE TO GODEL, 1967.

Hayes1,P.J., "The Frame Problem and Related Problems in Artificial Intelligence" ARTIFICIAL AND HUMAN THINKING, eds. Eilithorn and Jones, 1973.

Hayes2,P.J. "The Logic of Frames" FRAME CONCEPTIONS AND TEXT UNDERSTANDING, Walter de Gruyter & Co. 1979.

Hilbert,D. "Die Grundlagen der Mathematik" ABHANDLUNGEN AUS DEM MATHEMATISCHEN SEMINAR DER HAMBURGISCHEN UNIVERSITAT 6, 1927, translated in FROM FREGE TO GODEL, A SOURCE BOOK IN MATHEMATICAL LOGIC, 1967.

Hughes,G.E. and Creswell,M.J., AN INTRODUCTION TO MODAL LOGIC, METHUEN and Co. Ltd., London, 1968.

Kripke,S.A "Semantical considerations on modal logic" ACTA PHILOSOPHICA FENNICA 16. pp83-94.

Leibniz1, Gottfried W., "On the Principle of Indiscernibles" 1696, in LEIBNIZ PHILOSOPHICAL WRITINGS, J.M.Dent & Sons LTD, 1973.

Leibniz2, "Necessary And Contingent Truths" 1686, in LEIBNIZ PHILOSOPHICAL WRITINGS, J.M.Dent & Sons LTD, 1973.

Lewis,C.I., A SURVEY OF SYMBOLIC LOGIC, University of California Press, 1918.

Marcus, Ruth Barcan, "Modal Logic" CONTEMPORARY PHILOSOPHY, ed. Raymond

McCarthy1,J. "Recursive functions of symbolic expressions and their computation by machine" CACM 3(4) 1960.

McCarthy2,J. "Epistomological Problems of Artificial Intelligence" Proc. of the 5th International Joint Conference on Artificial Intelligence, 1977.

McCarthy,J. and Hayes,P. "Some Philosophical problems from the standpoint of Artificial Intelligence" MACHINE INTELLIGENCE 4, pp163-502, eds. Meltzer B. and Michie D., Edinburgh University Press, Edinburgh 1969.

McDermott,D., "Nonmonotonic Logic II: Nonmonotonic Modal Theories" JACM, Vol. 29, No. 1, Jan. 1982.

McDermott,D., Doyle,J. "Non-Monotonic Logic I" ARTIFICIAL INTELLIGENCE 13. 1980.

Moore1,R.C.1, "Reasoning about Knowledge and Action" 5th INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 1977.

Moore2,R.C. "Semantical Considerations on Nonmonotonic Logic" ARTIFICIAL INTELLIGENCE 25, 1985.

Morse, A.P., A THEORY OF SETS, Academic Press 1965.

Quine, W.V.O. "Reference and Modality" A LOGICAL POINT OF VIEW, 1961.

Prior,A., WORLDS, TIMES. AND SELVES

Reiter,R., "A Logic for Default Reasoning" ARTIFICIAL INTELLIGENCE 13, 1980.

Rescher,Nicholas A THEORY OF POSSIBILITY, Basil Blackwell&Mott Limited,1975.

Robinson A., NON-STANDARD ANALYSIS.

Schwind1 C.B. EIN FORMALISMUS ZUR BESCHREIBUNG DER SYNTAX UND BEDEUTUNG VON FRAGE-ANTWORT-SYSTEMEN. Fachbereich Mathematik der Technischen Universitat Munchen, 1978.

Schwind2 C.B., "The theory of Actions" report TUM-INFO 7807, Technische Universitat Munchen, 1978.

Schwind3 C.B., "A Completeness Proof for a Logic of Action" Laboratoire D'Informatique pour les sciences de l'homme.

Tarski. A., "The concept of truth in formalized languages", 1936. Translated in LOGIC, SEMANTICS, AND METAMATHEMATICS, Oxford Clarendon Press, 1956.

MULTIOBJECTIVE A[*]
A Complete and Admissible Search Algorithm

Bradley S. Stewart and Chelsea C. White, III
Department of Systems Engineering
School of Engineering and Applied Science
Thornton Hall
University of Virginia
Charlottesville, Virginia, 22901

ABSTRACT. We present a generalization of the A[*] search algorithm to the case of multiple objective problem solving. We call this generalization A[**]. After some background on search and multiobjective decision-making, we present a formulation of the multiobjective search problem and an outline of the A[**] algorithm. We then develop some notation and formally define the concept of dominance as it relates to multiobjective search problems. The main results of the paper are that A[**] is complete on infinite graphs and admissible when used with a suitably defined set of admissible heuristic functions. A simple example is used to illustrate the behavior of A[**] on a shortest-path problem.

I. INTRODUCTION. Many, if not most, problem-solving approaches may be interpreted as procedures for iteratively searching through a set of predetermined or progressively determined solution alternatives until a satisfactory, perhaps in some sense optimal, solution is found. Research in Artificial Intelligence (AI) has placed special emphasis on this view of problem-solving as search. AI search procedures are typically formulated in terms of a graph containing nodes representing the potential problem solutions. The criteria used to guide these search procedures have invariably been scalar-valued, reflecting the fact that the problems to be solved have been formulated with only a single objective. However, most, real-world problems involve multiple, conflicting, and noncommensurate objectives in any precise definition of what constitutes a "satisfactory" and/or "optimal" problem solution.

The task of adequately describing multiple, conflicting, and noncommensurate objectives using a scalar-valued criterion has been the subject of considerable research; see, for example, (Keeney and Raiffa, 1976) and other texts, reports, and journal articles concerned with multiattribute utility theory (MAUT). What often makes this task difficult, and the accuracy of the results suspect, are the potentially time consuming and stressful utility assessment procedures associated with MAUT.

The multiobjective approach to problem-solving avoids, at least initially, some of the more difficult assessment issues associated with MAUT. In this multiobjective approach, each objective is modeled by a scalar-valued criterion. However, instead of attempting to determine a scalar-valued function of these criteria and then seeking an alternative that optimizes this function, the goal is to determine the set of nondominated alternatives. An alternative α is said to be nondominated among a set of alternatives if there is no other alternative in the set

that is 1) at least as "good" as α with respect to all of the objectives, and 2) strictly "better" than α with respect to at least one of the objectives. Here, "good" and "better" are defined in terms of the scalar-valued criteria associated with individual objectives. Thus, the multiobjective approach allows the alternative selection criterion to be vector-valued and does not necessarily determine the "most preferred" alternative. It does, however, identify all clearly inferior alternatives so that they may be excluded from further consideration. The design, application, and evaluation of procedures for determining the "most preferred" alternative from a set of nondominated alternatives is currently a topic of considerable research interest; see, for examples, (White, et.al, 1984; Korhonen, et.al, 1984).

The observations above have motivated us to generalize the form of the criterion that directs search from a scalar-valued function to a vector-valued one. This vector-valued criterion is used to search for the set of nondominated solutions or solution paths rather than the "most preferred" solution or solution path. Our initial efforts have focused on generalizing A^*, an important AI search procedure (Hart, et.al., 1968; Pearl, 1984) to the multiobjective case. We refer to this generalization as A^{**}. The initial focus on A^* is due to its simple, yet powerful structure, which allows for useful analysis of its performance. Results of this initial work are expected to provide valuable insights into the general characteristics of multiobjective search techniques. These insights will guide future efforts to integrate other search procedures with multiobjective and multiattribute concepts.

In this paper, we define the A^{**} algorithm and show that it is complete on infinite graphs and admissible with respect to suitably extended definitions of an admissible heuristic evaluation function. Demonstration that A^{**} inherits other useful properties from A^* is a topic for future research.

II. THE MULTIOBJECTIVE A^* ALGORITHM: A^{**}. We now formulate the multiobjective search problem, state the A^{**} algorithm, define our notation, and provide a detailed definition of the concept of dominance as it is used in this paper.

A. Multiobjective Search Problem Formulation. The following abstraction of the multiobjective search problem differs from the single-objective version only in its cost structure and solution definition.

Given: - A problem state-space, representable as a locally-finite directed graph; states are nodes in the graph
- A single start node in the graph
- A finite set of goal nodes in the graph
- A positive, vector-valued cost associated with each arc in the graph; path costs are sums of associated arc costs
- A set of vector-valued heuristic functions estimating the cost to get to a solution from any node in the graph

Find: - The complete set of nondominated solution paths in the

690

graph from the start node to members of the set of goal nodes

As a typical example, multiobjective shortest-path problems may be formulated in this fashion (see also Section IV below).

B. A^{**} Algorithm Statement. Problems formulated in the above structure may be solved with A^{**}. The algorithm maintains three sets as part of its operation. OPEN is a set of nodes that are in the process of being investigated by the algorithm, sometimes called the frontier set. CLOSED consists of those nodes that have already been searched by the algorithm. OPEN and CLOSED are both defined in the same way as for the single-objective version of the algorithm. The third set, NONDOMINATED, is needed only for cases in which multiple solutions may arise. NONDOMINATED contains the current collection of nondominated solution paths and their associated costs at any point during the operation of A^{**}. The algorithm may be summarized as follows:

The A^{**} Algorithm

1. Put the start node in the set OPEN. Initialize CLOSED and NONDOMINATED to empty.

2. If OPEN is empty, exit with the current set of solution paths in NONDOMINATED, if any.

3. Otherwise, remove from OPEN and place on CLOSED a node n that is nondominated among current nodes in OPEN and paths in NONDOMINATED. If no such node exists, exit with the current set of solution paths in NONDOMINATED, if any.

4. If n is a goal node, trace back through appropriate pointers and add the newly discovered nondominated solution path or paths and their associated costs to the set NONDOMINATED. Go to Step 2.

5. Otherwise, expand n by generating all of its successors and establishing a backpointer to n for each. For each successor n' of n:

   a) If n' ∉ OPEN ∪ CLOSED, evaluate its vector-valued heuristic functions as estimates of the cost to get to a solution from n', add these estimates to the vector-valued costs accrued in reaching n' to get estimates of the total costs of solution paths through n', and add n' to OPEN.

   b) If n' ∈ OPEN ∪ CLOSED, redirect its backpointers along newly discovered nondominated paths, if any.

   c) If n' had a new nondominated estimate of the cost of a solution path through n' and had been on CLOSED, remove it from CLOSED, and put it back on OPEN.

6. Go to Step 2.

C. Notation. The following notation will be useful for analyzing the A^{**}

691

algorithm. To the extent allowed by the different requirements of the multiobjective formulation, the notation below has been defined in a manner consistent with that used for A^* in (Pearl, 1984) and elsewhere.

$\underline{G}$ — (N, A) - The state-space graph representing the problem to be solved. N is a countable set of nodes, each of which represents a problem state. $A \subseteq N \times N$ represents the set of directed arcs connecting pairs of nodes. Thus, if pair (n_i, n_j) is in A, then there exists a directed arc from node i to node j. $\underline{G}$ is assumed to be locally finite; i.e., the set of immediate successor nodes for each node in N is finite.

s - The unique start node in N.

$\Gamma \subseteq N$ - The finite, nonempty set of underline{goal} underline{nodes}, with generic member γ_i.

$\Gamma^{**} \subseteq N$ - The set of nondominated goal underline{nodes}; i.e., the set of goal nodes connected to s via one or more nondominated paths in $\underline{G}$.

$\underline{P}(n_i, S)$ — $\{P(n_i, S)\}$ - The set of all finite-length acyclic paths in $\underline{G}$ connecting node i with any member of the set of nodes S.

$\underline{P}^{**}(n_i, S)$ — $\{P^{**}(n_i, S)\}$ - The set of all nondominated paths from node i to any member of the set of nodes S.

$c(n_i, n_j)$ — $(c_1(n_i, n_j), \ldots, c_M(n_i, n_j))$ - Vector-valued arc cost associated with the arc connecting nodes i and j, where M is the number of objectives under consideration. All these cost vectors are assumed to be strictly positive and uniformly bounded away from zero as specified by the following constraint:

$$c_m(n_i, n_j) \geq \delta > 0 \quad \forall (n_i, n_j) \in A, \; m \in \{1, 2, \ldots, M\}.$$

Note also that we define $c_m(n_i, n_i) = 0 \; \forall i$ and m, and we assume that the objective is to minimize all cost components.

$c(P)$ - Actual vector-valued path cost of a specific path P, assumed additive; i.e.,

$$c(P) = \Sigma \; c(n_k, n_l), \text{ where the sum is taken over all pairs of nodes } (n_k, n_l) \text{ representing arcs on path } P.$$

C^{**} — $\{c^{**}(P)\}$ — $\{c(P): P \in \underline{P}^{**}(s, \Gamma)\}$ - Set of all costs of nondominated solution paths from s to Γ.

SG_T - The traversal subgraph defined at any stage in the search by the pointers that A** assigns to the nodes already generated (i.e., visited or searched), with the branches of SG_T directed opposite to the pointers. Note that if $\{n_i\}$ represents the set of all nodes in SG_T at some point in the search, then OPEN $\subseteq \{n_i\}$ at that point in the search also.

SG_e - The explicated subgraph of $\underline{G}$, defined at some point in the search as the union of all the branches in the traversal

subgraphs up to that point in the search. As Pearl points out (Pearl, p. 75), a path P can be in SG_e and not in any of the traversal subgraphs that make up SG_e. Note that if $\{n_i\}$ represents the set of nodes in SG_e, then at any time during the search $\{n_i\}$ = OPEN $\cup$ CLOSED.

$G(n)$ = $\{g(n)\}$ - Set of cost vectors for all paths from s to n that are nondominated in SG_e. Note that $G(s)$ = $\{\underline{0}\}$.

G^{**} = $\{g^{**}\}$ - Set of nondominated cost functions,

$$g^{**}: N \rightarrow R^{M+}, \text{ where } g^{**}(n) = c^{**}(P)$$

$$\forall\ n \in \{\text{nodes on P}\},\ P \in \underline{P}^{**}(s,n).$$

$G^{**}(n)$ = $\{g^{**}(n)\}$ - Set of all costs of nondominated paths between the start node, s, and node n. By this definition,

$$G^{**}(n) = \{c(P): P \in \underline{P}^{**}(s,n)\}.$$

H = $\{h\}$ - A set of heuristic functions estimating the cost to get from any node of the graph to a solution,

$$h: N \rightarrow R^{M+}, \text{ where } h(n) \text{ is an estimate of } c(P)$$

$$\forall\ n \in \{\text{nodes on P}\},\ P \in \underline{P}(n,\Gamma).$$

$H(n)$ = $\{h(n)\}$ - Set of heuristic function values for a node n; i.e., an estimate of the actual nondominated cost-to-go values in $H^{**}_1(n)$. Note that

$$H(\gamma) = \{\underline{0}\}\ \forall\ \gamma \in \Gamma.$$

H^{**}_1 - Set of all actual cost-to-go functions,

$$h^{**}_1: N \rightarrow R^{M+}, \text{ where}$$

$$h^{**}_1(n) = c(P_1)\ \forall\ n \in \{\text{nodes on } P_1\},\ P_1 \in \underline{P}^{**}(n,\Gamma^{**}),$$

$$h^{**}_1(n) = c(P_2)\ \forall\ n \in \{\text{nodes on } P_2\},$$

$$P_2 \in \underline{P}(n,\Gamma) - \underline{P}(n,\Gamma^{**}), \text{ and}$$

$$h^{**}_1(n) = \infty\ \forall\ n \notin \{\text{nodes on P}\},\ P \in \underline{P}(n,\Gamma).$$

$H^{**}_1(n)$ = $\{h^{**}_1(n)\}$ - Set of all actual costs of nondominated paths between n and Γ; i.e., between n and a member of the goal set. In accordance with this definition we assign values as follows:

$$h^{**}_1(n) = c(P)\ \forall\ n \in \{\text{nodes on P}\},\ P \in \underline{P}^{**}(n,\Gamma)$$

$$h^{**}_1(n) = \infty \text{ otherwise.}$$

H^{**}_2 = $\{h^{**}_2\}$ - Set of actual cost-to-go functions corresponding to

the nondominated solution paths in $\underline{G}$. Note that $H^{**}_2 \subseteq H^{**}_1$.

$H^{**}_2(n) - \{h^{**}_2(n)\}$ - Set of actual costs of partial paths from n to Γ along nondominated solution paths. Note that $H^{**}_2(n) \subseteq H^{**}_1(n)$.

$F(n) - \{f(n)\}$ - At some stage of the search, the set of costs of the form

$$f(n) - g(n) + h(n) \text{ for some } n, \text{ with } g(n) \in G(n),$$

$$h(n) \in H(n).$$

$F^{**}(n) - \{f^{**}(n)\}$ - Set of costs of paths that are nondominated among all paths from s through n to the goal set.

D. <u>Definition of Path Cost Dominance.</u> Define a relation

$R_1 \subseteq \underline{P}(n_i, n_j) \times \underline{P}(n_i, n_j) \ \forall n_i, \ n_j \in N$, and $\forall P_\ell, P_k \in \underline{P}(n_i, n_j)$ as follows:

$$(P_\ell, P_k) \in R_1 \Leftrightarrow c(P_\ell) \leq c(P_k).$$

We say that path ℓ <u>dominates</u> path k if and only if:

$$(P_\ell, P_k) \in R_1 \text{ and } (P_k, P_\ell) \notin R_1.$$

Similarly, under these same conditions we say that the cost of path ℓ <u>dominates</u> the cost of path k. A path P and its associated cost $c(P)$ are said to be <u>nondominated</u> (in $\underline{P}(n_i, n_j)$) if and only if there does not exist a path $P' \in \underline{P}(n_i, n_j) \ni (P', P) \in R_1$. Paths and path costs may also be defined as nondominated with respect to specific sets of paths and costs by adapting the above definitions in the obvious way. A node n is said to be <u>nondominated</u> with respect to some set containing nodes and/or complete solution paths if and only if there is at least one partial path to n that has a solution path cost estimate that is nondominated with respect to the solution path costs or solution path cost estimates associated with the elements of the set.

III. <u>FORMAL PROPERTIES OF A**.</u> We first present some preliminary analysis for the multiobjective case. Then we prove some useful results concerning termination and completeness of the algorithm. Finally, we establish some definitions that allow us to show that A** is an admissible algorithm when used with an admissible set of heuristics.

A. <u>Preliminary Results.</u> Using the definition of a nondominated path cost and the other notation defined above, the following is immediate:

$\forall n \in N, P(s, n) \in \underline{P}(s, n), \exists i, j \in \{1, 2, \ldots, M\} \ni$

$$c_i(P(s, n)) \geq g^{**}_i(n) \ \forall g^{**}(n) \in G^{**}(n) \text{ and}$$

$$c_j(P(n, \gamma)) \geq h^{**}_j(n) \ \forall h^{**}_1(n) \in H^{**}_1(n), \ \gamma \in \Gamma.$$

By the definition of $F^{**}(n)$, we have:

694

$$\forall \ f^{**}(n) \in F^{**}(n) \ \exists \ g^{**}(n) \in G^{**}(n) \text{ and } h^{**}_1(n) \in H^{**}_1(n) \ni$$

$$f^{**} - g^{**}(n) + h^{**}_1(n).$$

The notational definitions above also directly imply the following relationship:

$$C^{**} - F^{**}(s) - H^{**}_1(s) - H^{**}_2(s) \subseteq G^{**}(\gamma^{**}) - F^{**}(\gamma^{**}) \ \forall \ \gamma^{**} \in \Gamma^{**}.$$

If n^{**} is a node on any nondominated path from s to Γ; i.e., from s to some $\gamma^{**} \in \Gamma^{**}$, then

$$\exists \ P_1 \in \underline{P}^{**}(s, n^{**}), \ P_2 \in \underline{P}^{**}(n^{**}, \Gamma^{**}) \ni c(P_1) \in G^{**}(n^{**}) \text{ and}$$

$$c(P_2) \in H^{**}_1(n^{**}) \text{ with } f^{**}(n^{**}) - c(P_1) + c(P_2)$$

$$\text{and } f^{**}(n^{**}) \in C^{**}.$$

A somewhat more concise way of expressing this same relation using some different notation is as follows:

$$(\text{eq. } 1) \quad \forall \ P \in \underline{P}^{**}(s, \Gamma^{**}) \ \exists \ f^{**} \in F^{**} \ni f^{**}(n) - c(P)$$

$$\forall \ n \in \{\text{nodes on } P\}, \ P \in \underline{P}^{**}(s, \Gamma^{**}).$$

It is tempting to try to conclude that

$$\forall \ n^{**} \in \{\text{nodes on } P\} \text{ with } P \in \underline{P}^{**}(s, \Gamma^{**}), \text{ we have } F^{**}(n) \subseteq C^{**}.$$

The example in Section IV below provides a counterexample showing that this conclusion would be false. In other words, the example shows that there may exist solution paths P through some node n^{**} such that P is dominated among all solution paths, but nondominated among those solution paths constrained to go through n^{**}.

The importance of F^{**} is found in its ability to identify off-track nodes; i.e., nodes not lying on any nondominated solution path. For any off-track node n

$$f^{**}(n) \notin C^{**} \ \forall \ f^{**}(n) \in F^{**}(n)$$

or, in other words, $F^{**}(n) \cap C^{**} - \emptyset$ for all off-track nodes n.

B. <u>Termination</u> <u>and</u> <u>Completeness.</u> The usefulness of A^{**} will be at least partly determined by its ability to retain the valuable properties of A^{*}. The following results on termination and completeness are fundamental prerequisites that A^{**} must satisfy to have some chance of being useful.

In general, an algorithm is said to be complete if it terminates with a solution whenever one exists. Since A^{**} seeks a set of solutions (the nondominated set), the definition must be adapted to state that A^{**} is complete if it finds at least one (nondominated) solution whenever any solutions exist. Below we show that A^{**} terminates on finite graphs and is complete on infinite graphs. As part of the completeness proof, we show

that A** terminates on infinite graphs as long as some finite length path from the start node to a goal node exists. Recall from the problem statement that we assumed that the problem graph is locally finite and that all arc costs are uniformly bounded away from zero. These two assumptions are critical to the proofs of the results of this section.

Theorem 1. A** always terminates on finite graphs.

Proof. By definition, the number of arcs in a finite graph is finite. The total number of unique combinations and permutations of that finite set of arcs is also finite. The set of acyclic paths in the graph is uniquely described by a subset of the set of all combinations and permutations of the arcs in the graph. Therefore, the number of acyclic paths in a finite graph is finite.

Step 1 of the A** algorithm executes exactly once per problem. The remaining steps form a single loop, with either Step 4 or Step 5 executing on each iteration. Therefore, if we show that both Step 4 and Step 5 can execute only a finite number of times on a given finite problem graph, then we have proved the desired result.

Each time the test in Step 4 is satisfied, at least one new finite length nondominated path from s to Γ has been found. Since there are only a finite number of such paths, this step can execute only a finite number of times.

Step 5 is a node expansion step. Each time it executes, either one or more new arcs are added to the current traversal subgraph, SG_T, or there are no other arcs out of the node. If there are no other arcs out of the node, then it is a stub (only incoming arcs attached to it) and cannot be on any solution path. Furthermore, since there can be only a finite number of incoming arcs, the node will be permanently entered on CLOSED after a finite number of additional visits. Since there can be only a finite number of such nodes in the problem graph, this can occur only a finite number of times.

Each arc that is added to SG_T represents part of at least one new acyclic path in $\underline{G}$ that has been dicovered by A**. Since there are only a finite number of such acyclic paths in $\underline{G}$, this can occur only a finite number of times. Reopened nodes also represent new acyclic paths in $\underline{G}$. This is because A** only reopens a node from CLOSED when it discovers a path to the node with a nondominated cost estimate that is different from any of those already contained in SG_e. □

Theorem 2. A** is complete on infinite graphs.

Proof. As stated above, to show that A** is complete in the general case of an infinite graph, we must show that A** terminates with at least one (nondominated) solution whenever any solution exists. There are only two cases in which A** could fail to terminate with a solution:

Case 1. A** terminates in failure.

Case 2. A** fails to terminate.

We will show that neither of these cases can occur if a solution exists.

Case 1. When we say A** terminates in failure we mean that the algorithm terminates before a solution path is found. Since solutions are collected in the set NONDOMINATED as they are found, A** terminates in failure if and only if NONDOMINATED is empty when A** terminates. There are two conditions under which A** terminates; either OPEN becomes empty or all nodes on OPEN are dominated by costs of solution paths in NONDOMINATED. Note that in the second case, at least one solution has been found since NONDOMINATED is nonempty. Therefore, we only need to show that, if a solution exists, it is not possible for both OPEN and NONDOMINATED to be empty at the same time. To see this, let $P(s,\gamma)$ be a solution path. OPEN cannot become empty before $P(s,\gamma)$ is found. This is because whenever OPEN is nonempty and $P(s,\gamma)$ has not yet been found, OPEN must contain a node on $P(s,\gamma)$. We show this by simple induction as follows:

- At Step 1, OPEN contains s which is on $P(s,\gamma)$.

- Assume that, after k iterations, there remains at least one node from $P(s,\gamma)$ in OPEN. Let n be the deepest such node.

- On iteration k+1, n is either found to be dominated or nondominated in OPEN ∪ NONDOMINATED. Unless n is nondominated and selected for expansion, it is left on OPEN. If selected for expansion, n is either found to be a goal node in Step 4, in which case $n = \gamma$ and $P(s,\gamma)$ has been found, or n is expanded in Step 5. If n has successors that are not already on OPEN or CLOSED, these are added to OPEN in Step 5a. If n has no new successors then either n has no successors or all of its successors are on CLOSED. All nodes on $P(s,\gamma)$ have at least one successor each (except γ and we know at this point that $n \neq \gamma$). Therefore, n cannot have no successors. On the other hand, n must not have successors on CLOSED because by assumption, n is on $P(s,\gamma)$, so n is at the head of a chain of descentent nodes among which is γ. If any of n's immediate successors were on CLOSED, then either 1) all of them would be on CLOSED, a contradiction to the assumption that $P(s,\gamma)$ has not yet been found, or 2) some of n's descentants would be on OPEN, a contradiction to the assumption that n was the deepest node of $P(s,\gamma)$ or OPEN. In any case, at the completion of Step 5, node n will have at least one decendent on OPEN.

  Therefore, on iteration k+1, $P(s,\gamma)$ will either be found or it will have one or more nodes remaining on OPEN. By complete induction, since k was arbitrary, OPEN cannot become empty before a solution path $P(s,\gamma)$ is found, if any such solution paths exist.

Case 2. In any locally finite graph, there is only a finite number of finite-length, acyclic paths from the start node to any node in the graph. This can be shown by induction as follows.

  By the definition of a locally finite graph, there can be at most a

finite number of arcs originating at the start node. Therefore, there are at most a finite number of nodes at path length 1 from the start node. Assume that there are only a finite number of distinct, acyclic paths from s to all nodes at path length k-1 and that the number of such nodes is finite. Since each of these nodes may have only a finite number of immediate successors arrived at through a finite number of arcs, the number of nodes at path length k will also be finite. The set of distinct acyclic paths to these nodes at path length k is uniquely represented by some subset of the possible combinations of the paths to nodes at path length k-1 with the arcs from those nodes to the nodes at path length k. Therefore, by complete induction we have shown that for arbitrary finite k, there are at most a finite number of distinct acyclic paths from the start node to all nodes reachable by traversing k arcs of the graph, where k is any finite number.

By Theorem 1, if A^{**} does not terminate it must be searching an infinite path. Since all arc costs are assumed to be uniformly bounded away from zero in all components ($c_i(a) \geq \delta > 0 \; \forall \; i \in \{1,2,\ldots,M\}, \; a \in A$), an infinite length path must have unbounded costs in all elements of its cost vector. Cost estimates of nodes on such an infinite path will eventually become dominated by cost estimates of any finite length paths, including all (nondominated) solution paths. This will cause A^{**} to eventually terminate on the test for dominance in Step 3. $\square$

Corollary 1. A^{**} is complete on finite graphs.

Proof. This is a subcase of Theorem 2 for which the proof in Theorem 2 is still valid. $\square$

Corollary 2. A^{**} terminates on infinite graphs if there exists any finite length path from s to any goal node.

Proof. This was shown as part of Case 2 in Theorem 2. $\square$

C. Admissibility. One of the most useful properties of A^* is admissibility. This property guarentees that the algorithm will return an optimal solution whenever any solution exists. Using the following definition of an admissible set of multiobjective heuristics, analogous results are derived for A^{**}.

Definition. A multiobjective heuristic function, h, is said to be admissible if

$$\exists \; h^{**}_2 \in H^{**}_2 \ni h(n) \leq h^{**}_2(n) \; \forall \; n \in N.$$

Definition. A set of multiobjective heuristic functions, H, is said to be admissible if

$$\forall \; h^{**}_2 \in H^{**}_2 \; \exists \; h \in H \ni h(n) \leq h^{**}_2(n) \; \forall \; n \in N.$$

In other words, a set of admissible multiobjective heuristic functions contains at least one heuristic function that is admissible with respect to each of the elements in the set H^{**}_2.

In order to prove the admissibility of A^{**}, we first consider the following Lemma and its Corollary$_{**}$ both of which presume the use of an admissible set of heuristics with A^{**}.

Lemma 1. At any time before A^{**} terminates, for every undiscovered nondominated solution path $P \in \underline{P}^{**}(s, \Gamma^{**})$ ∃ $n \in$ OPEN, $n \in$ {nodes on P} ∃ ∃ $f(n) \in F(n)$ ∋ $f(n) \leq c(P)$.

Proof. $_{**}$ Consider any undiscovered nondominated solution path $P \in \underline{P}^{**}(s, \Gamma^{**})$. If no such path exists, then there is nothing to prove. Let $P = s, n_1, n_2, n_3, \ldots, n', \ldots, \gamma^{**}$. There is always a node from P on OPEN. This is shown by induction as follows. At the start of the algorithm, s is in OPEN. Assume that, at some later point in the search, a nongoal node of P is on OPEN. Let n_{k-1} be the deepest such node on P. At the next iteration of A^{**}, n_{k-1} is either chosen for expansion or not. If not, then it remains on OPEN for the next iteration.

If n_{k-1} is chosen for expansion, then it will be found to have at least one successor on P, call it n_k, and this successor will be added to OPEN as n_{k-1} is removed and placed on CLOSED. Therefore, since k was arbitrary, complete induction shows that there will always be a node from P on OPEN until γ^{**} is chosen for expansion and P is discovered to be a nondominated solution path.

Since we have shown that every undiscovered nondominated solution path will have a node on OPEN at all times before A^{**} terminates, let n' be the shallowest such node from P. Since n' is the shallowest node from P on OPEN, all of its ancestors must be on CLOSED. Furthermore, by assumption, the partial path $s, n_1, n_2, n_3, \ldots, n'$ from P is nondominated. Therefore,

(eq. 2) $\forall g(n') \in G(n') \ \exists \ g^{**}(n') \in G^{**}(n') \ \ni g(n') = g^{**}(n')$.

In other notation this equation states that

$$c(s, n_1, n_2, n_3, \ldots, n') = c(P^{**}(s, n')) = g(n')$$

and since G^{**} contains all nondominated costs for paths from s to n', $g(n') \in G^{**}(n')$. Using the admissibility of H, we know that for n'

$$\forall h^{**}_2 \in H^{**}_2 \ \exists \ h \in H \ \ni h(n') \leq h^{**}_2(n').$$

By the definition of H^{**}_2, since $n' \in$ {nodes on P} and $P \in \underline{P}^{**}(s, \Gamma^{**})$, ∃ $h^{**}_2 \in H^{**}_2 \ni h^{**}_2(n') = c(P) - c(P^{**}(s, n'))$

From above we had $g(n') = g^{**}(n') = c(P^{**}(s, n'))$ for some $P^{**}(s, n')$ on P so we can now combine to see that

$$\exists \ h^{**}_2 \in H^{**}_2 \ \ni h^{**}_2 = c(P) - g^{**}(n').$$

Using equation 1 from the preliminary results above, we know

$$\exists \ f^{**} \in F^{**} \ni f^{**}(n') = c(P)$$

so that

$$\exists \; h^{**}_2 \in H^{**}_2 \ni h^{**}_2(n') = f^{**}(n') - g^{**}(n').$$

Combining the results above, we may complete the proof as follows:

$$
\begin{aligned}
f(n') &= g(n') + h(n') && \text{by definition} \\
&= g^{**}(n') + h(n') && \text{substituting from eq. 2} \\
&\leq g^{**}(n') + h^{**}_2(n') \text{ for some } h^{**}_2 \in H^{**}_2 \\
& && \text{by the admissibility of H} \\
&= f^{**}(n') \text{ for some } f^{**}(n') \in F^{**}(n') \\
& && \text{by eq. 1 above} \\
&= c(P) \text{ for some } P \in \underline{P}^{**}(s, \Gamma^{**}) \\
& && \text{by eq. 1 also, since } f^{**}(n') \in C^{**}.
\end{aligned}
$$

In summary, we have completed the proof by demonstrating the existence of an open node n' on an arbitrary nondominated solution path P such that

$$f(n') \leq c(P) \text{ for some } c(P) \in C^{**}. \quad \square$$

<u>Corollary 3.</u>   Let n' be the shallowest open node on a nondominated path $P^{**}(s, n'')$ to an arbitrary node n'', not necessarily a goal node. Then

$$\forall \; g(n') \in G(n') \; \exists \; g^{**}(n') \in G^{**}(n') \ni g(n') = g^{**}(n')$$

so that A^{**} has already found a nondominated path to n' and that path will remain nondominated in $\underline{P}(s, n')$ throughout the search.

<u>Proof.</u>   This proof is the same as that for Lemma 1 where the fact that n' was on a solution path was not used to demonstrate the validity of equation 2. $\square$

<u>Theorem 3.</u> A^{**} using any set of admissible heuristics is admissible.

<u>Proof.</u> By Corollary 2 we know A^{**} terminates whenever a finite length solution path exists so it is sufficient to show that A^{**} will not terminate until all nondominated solution paths have been found; i.e., placed on NONDOMINATED. Assume that A^{**} terminates, but there exists some nondominated solution path P that is as yet undiscovered. By Lemma 2

$$\exists \; n \in \text{OPEN}, \; n \in (\text{nodes on P}) \ni \exists \; f(n) \in F(n) \ni f(n) \leq c(P)$$

$$\text{for some } c(P) \in C^{**}.$$

A^{**} terminates in only 2 cases:

1) OPEN is empty.

2) All current partial path cost estimates for nodes on OPEN are dominated by solution path costs in NONDOMINATED.

Case 1 is ruled out by our assumption that n was in OPEN. According to the condition in case 2, the termination we assumed must have occurred because $f(n)$ was dominated by a member of NONDOMINATED. This is a contradiction since $f(n) \leq c(P)$ implies that $f(n)$ is nondominated in C^{**} and therefore nondominated among all solution paths that could be in NONDOMINATED. Therefore, A^{**} cannot terminate before all nondominated solution paths have been discovered. □

IV. **EXAMPLE.** The following multiobjective shortest path problem illustrates the operation of the A^{**} algorithm on a very simple state-space graph. The purpose of the example is to present an overview of the general behavior of the algorithm, without emphasizing the computational details of the procedure. The computational aspects of the A^{**} algorithm in particular, and of multiobjective search algorithms in general, present an interesting topic for further research.

The problem state-space graph is shown in Figure 1, which also shows the values of the arc costs for the problem. The value of M for this simple example is 2. Based on the information in the figure, we may calculate the derived cost information that is shown in Table 1. Also shown in Table 1 are the heuristic function values used for this illustration. The heuristic function values were simply defined by the nondominated members of the set of costs corresponding to the arcs emminating from a node. For example, the set of costs of arcs for the start node is {(1,2),(3,1),(1,3)}, of which one is dominated so the heuristic function values for the start node are {(1,2),(3,1)} as shown in Table 1. These heuristic function values are clearly admissible since they will always provide lower bounds on the possible costs of paths to the goal set.

Table 1 provides the information necessary to illustrate one of the statements made in Section III. A. Nodes 2, 7, and γ_1 all provide examples of the fact that the set $F^{**}_2(n)$ may not be contained in C^{**}, even if n is on a nondominated solution path. As a specific example,

$$F^{**}_2(2) = \{(7,9),(9,5),(8,8)\},$$

but the path $(s,2,5,7,\gamma_1)$ with cost $(7,9)$ is not a nondominated solution path. In other words, while this path with cost $(7,9)$ is nondominated among those solution paths constrained to go through node 2, it is not nondominated among all solution paths.

Operation of the algorithm with the next-arc-cost heuristic is easily followed using the sequence of 11 graphs that make up Figure 2. Each graph illustrates the state of the search for an iteration of the algorithm. Arcs of the initial problem state-space graph are shown as dashed lines. As the arcs are traversed as part of the search, they are made solid in the figures and the final solution paths are shown as

701

additional numbered dashed lines in the graphs. Nodes of the initial problem state-space graph are shown as dashed circles. OPEN nodes are shown with double circles, the node currently being expanded as part of the search is depicted with three circles, and CLOSED nodes are indicated with solid circles. As shown in the last panel of Figure 2, A^{**} correctly identifies the three nondominated solution paths in this example. The solutions and there associated costs are

$$P^{**}_1 = (s,1,5,7,\gamma_1) \qquad c(P^{**}_1) = (4,11)$$
$$P^{**}_2 = (s,1,5,8,\gamma_3) \qquad c(P^{**}_2) = (6,7)$$
$$P^{**}_3 = (s,2,5,8,\gamma_3) \qquad c(P^{**}_3) = (9,5).$$
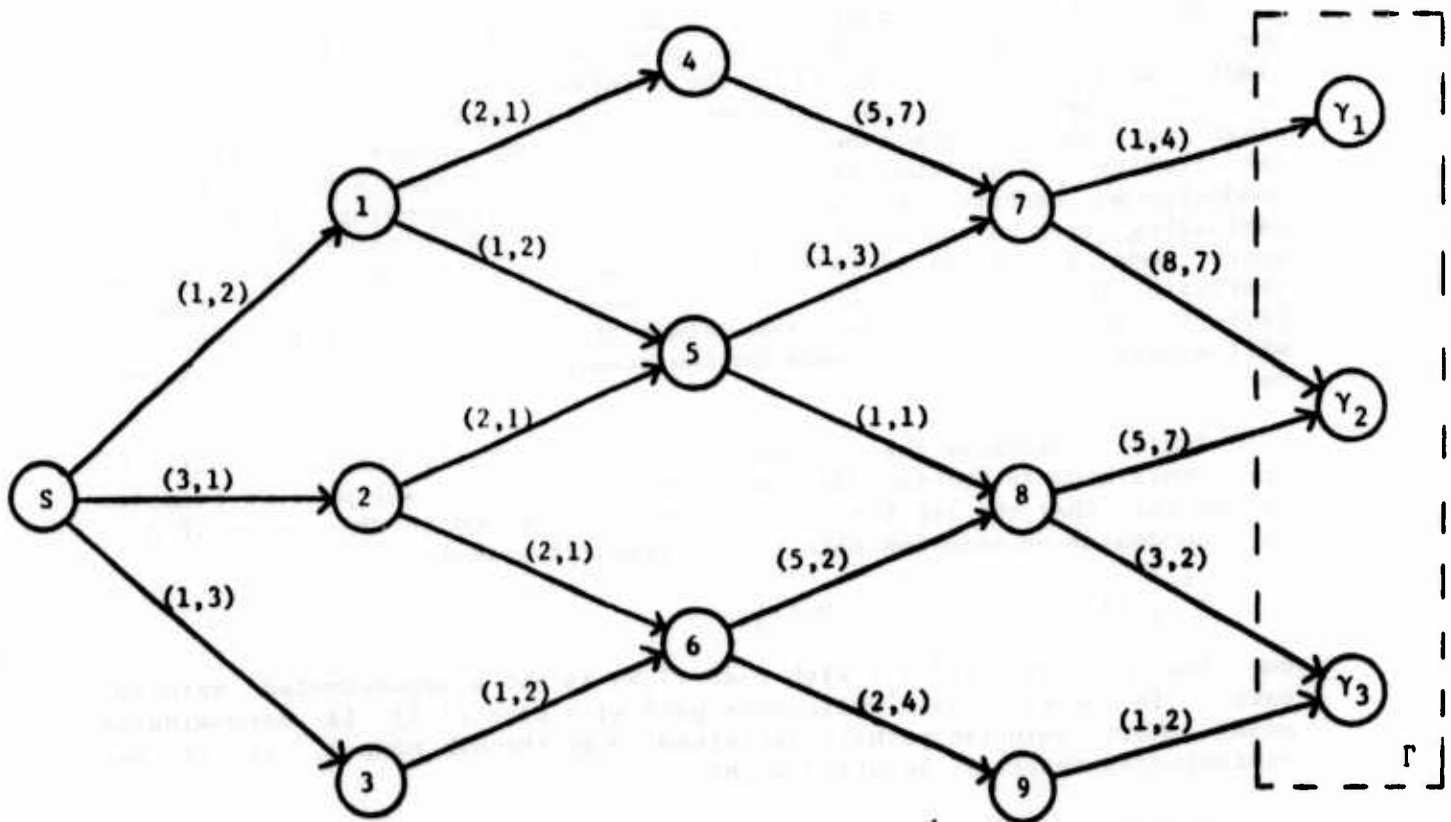
Figure 1. Example Problem Graph.

	Derived Cost Values				Heuristic Values
node	$G^{**}(n)$	$H^{**}_1(n)$	$H^{**}_2(n)$	$F^{**}(n)$	$H(n)$
s	(0,0)	(4,11) (6,7) (9,5)	$H^{**}_1(s)$	$H^{**}_1(s)$	(1,2) (3,1)
1	(1,2)	(3,9) (5,5)	$H^{**}_1(1)$	(4,11) (6,7)	(2,1) (1,2)
2	(3,1)	(4,8) (6,4) (5,7)	(6,4)	(7,9) (9,5) (8,8)	(2,1)
3	(1,3)	(9,6) (4,8)	(∞,∞)	(10,9) (5,11)	(1,2)
4	(3,3)	(6,11)	(∞,∞)	(9,14)	(5,7)
5	(2,4) (5,2)	(2,7) (4,3)	$H^{**}_1(5)$	(4,11) (6,7) (9,5) (7,9)	(1,1)
6	(5,2) (2,5)	(8,4) (3,6)	(∞,∞)	(13,6) (8,8) (5,11) (10,9)	(5,2) (2,4)
7	(3,7) (6,5)	(1,4)	$H^{**}_1(7)$	(4,11) (7,9)	(1,4)
8	(3,5) (6,3)	(3,2)	$H^{**}_1(8)$	(6,7) (9,5)	(3,2)
9	(7,6) (4,9)	(1,2)	(∞,∞)	(8,8) (5,11)	(1,2)
γ_1	(4,11) (7,9)	(0,0)	$H^{**}_1(\gamma_1)$	(4,11) (7,9)	(0,0)
γ_2	(8,12) (11,10)	(0,0)	(∞,∞)	(8,12) (11,10)	(0,0)
γ_3	(6,7) (9,5)	(0,0)	$H^{**}_1(\gamma_3)$	(6,7) (9,5)	(0,0)

Table 1. Derived cost values and next-arc-cost heuristic estimates for example problem.
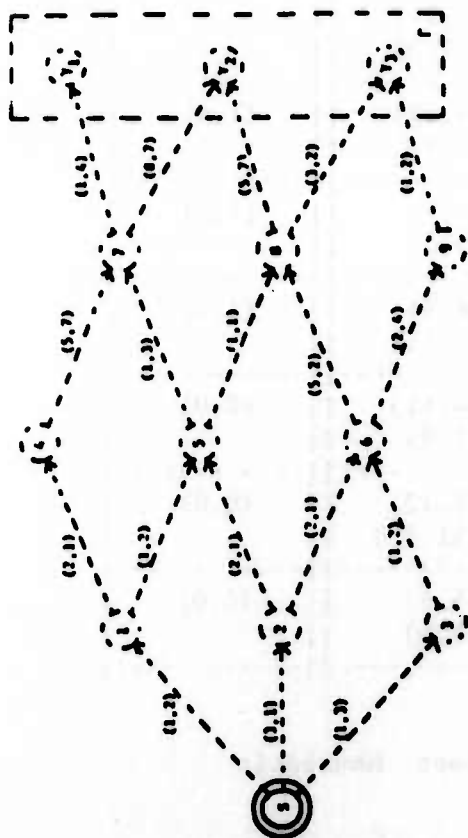
2c. Iteration 2.
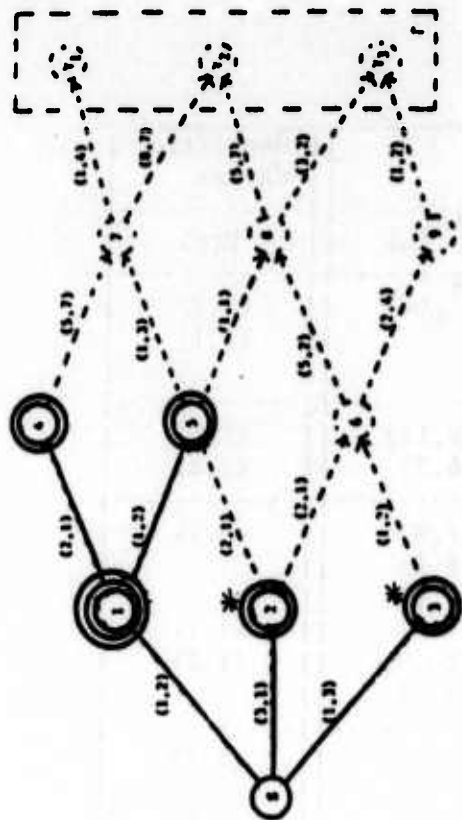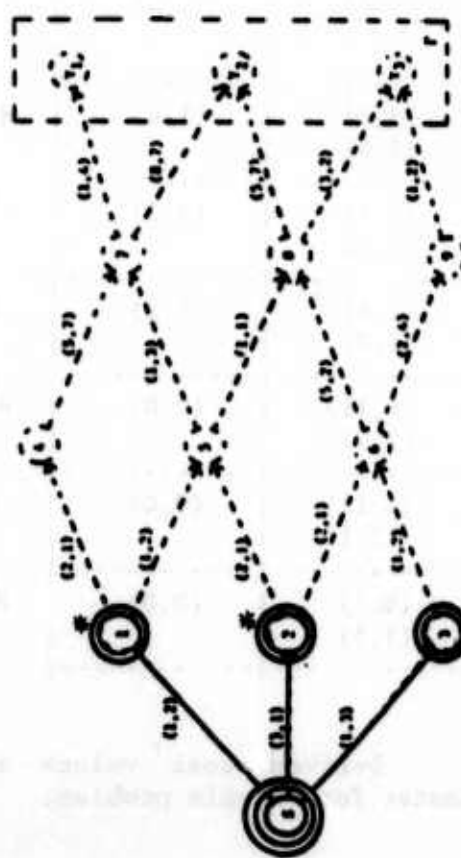
2d. Iteration 3.
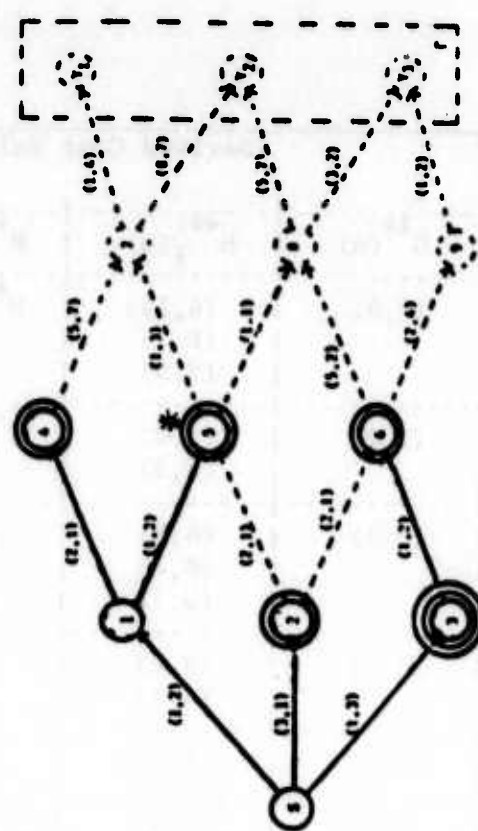
2a. Iteration 0.

2b. Iteration 1.

Figure 2. Operation of A** on example problem.
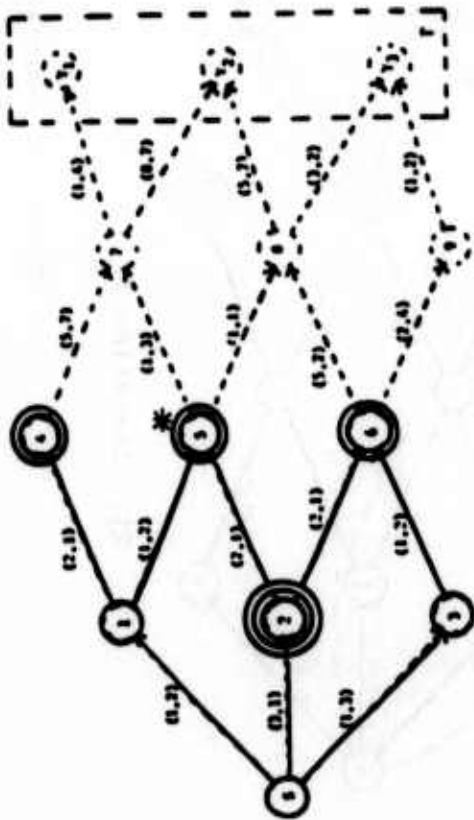
2g. Iteration 6.
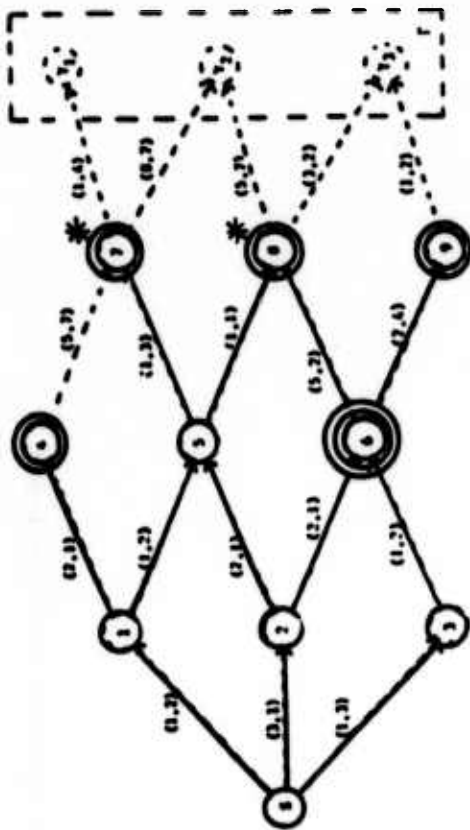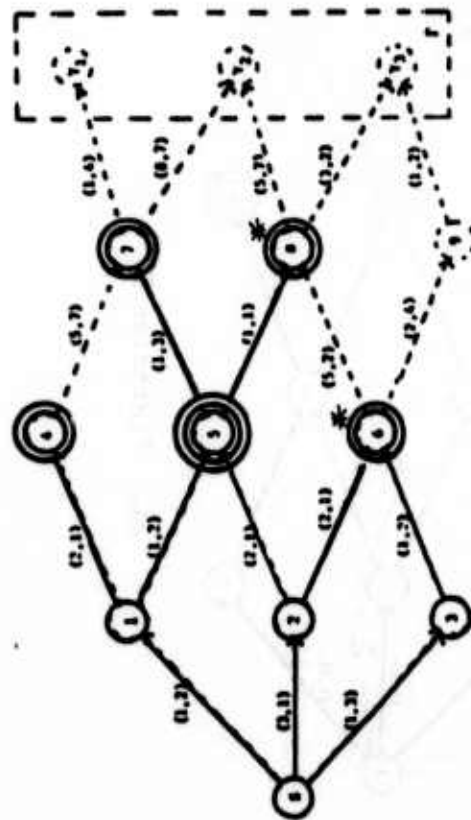
2h. Iteration 7.
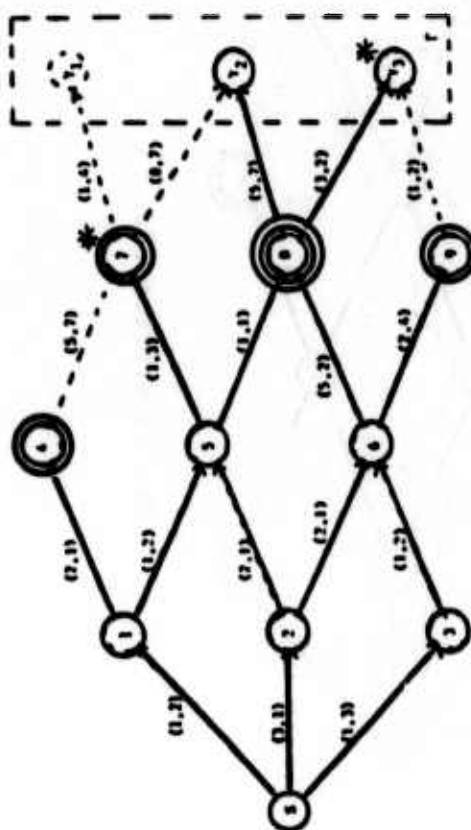
2e. Iteration 4.

2f. Iteration 5.

Figure 2. Operation of A** on example problem (cont.).
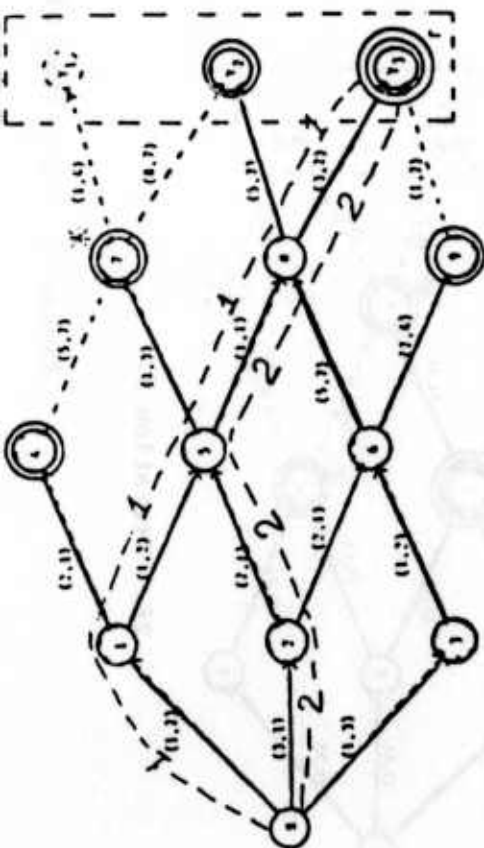
2k. Iteration 10.

2i. Iteration 8.

2j. Iteration 9.
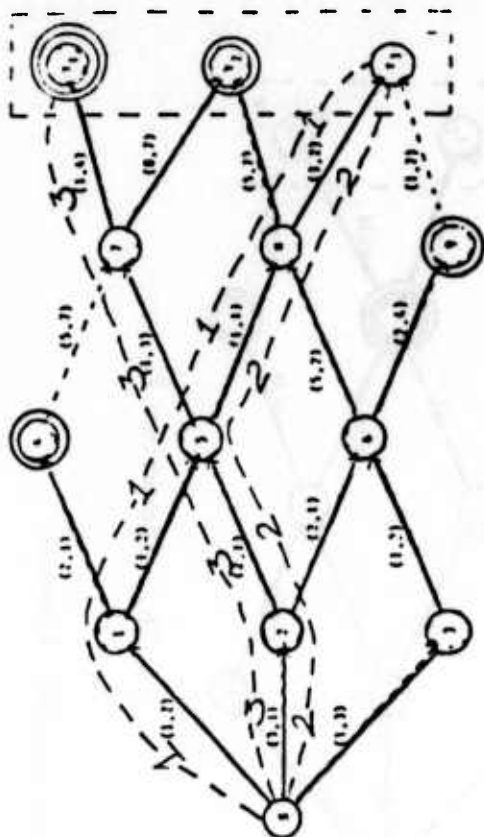
Figure 2. Operation of A^{**} on example problem (cont.).

V. <u>CONCLUSIONS.</u> This paper has presented some early results in an ongoing study of multiobjective search. We defined an abstract multiobjective search problem and outlined an adaptation of the single-objective A^* algorithm, A^{**}, for solving it. By suitably adapting the terminology used in work on A^*, we were able to show that the valuable properties of completeness and admissiblity are inherited by A^{**}. The behavior of the algorithm was briefly illustrated on a simple two objective shortest path problem.

There are two basic directions for future research in the area of multiobjective search. On the theoretical side, much additional work must be done to complete the development of A^{**}. For A^*, heuristic functions may be compared as to there effectiveness in directing search. In this context, the definition of dominance of one heuristic over another becomes useful in identifying the best heuristics. The generalization of this concept of dominance among heuristics to the multiobjective case will be a useful theoretical construct. Under specific assumptions about the information available to aid in the search process, it can be shown that A^* is an optimal search algorithm in some sense. The proof that A^{**} is also optimal in some sense, and the definition of the exact conditions on that optimality, are near-term research objectives. Another item of theoretical interest is the multiobjective generalization of the AO^* algorithm, the counterpart of A^* for use in searching AND/OR graphs.

The other, and perhaps more critical, direction for future research concerns the practical uses of multiobjective search. The preliminary work done so far indicates that these search routines will be very computation intensive. Issues of computational tractibility and efficiency will certainly become critical for any practical applications of multiobjective search. Actual heuristics available for real applications will probably be hard to develop and admissibility will almost certainly be impossible to prove or guaranty. Further work on searching with inadmissible heuristics will therefore be of practical significance.

BIBLIOGRAPHY.

Hart, P. E., N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," IEEE Transactions on Systems Science and Cybernetics, Vol. 4, 1968.

Hart, P. E., N. J. Nilsson, and B. Raphael, "Correction to 'A formal basis for the heuristic determination of minimum cost paths'," SIGART Newsletter, Vol. 37, 1972.

Keeney, R. L., and H. Raiffa, Decisions with Multiple Objectives: Preferences and Value Tradeoffs, Wiley, New York, 1976.

Korhonen, Pekka, Herbert Moskowitz, and Jyrki Wallenius, "A sequential approach to modeling and solving multiple criteria decision problems," Working Paper F-70, Helsinki School of Economics, Finland, 1984.

Pearl, Judea, Heuristics: Intelligent Search Strategies for Computer Problem Solving, Addison-Wesley, Reading, Massachusetts, 1984.

White, Chelsea C., III, Andrew P. Sage, and Shigeru Dozono, "A model of multiattribute decisionmaking and trade-off weight determination under uncertainty," IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-14, No. 2, March/April, 1984, pp. 223-229.

MATHEMATICAL BASIS FOR EXPERT REASONING *

Forouzan Golshani

Department of Computer Science,
Arizona State University
Tempe, AZ 85287

ABSTRACT

We propose an modal system of algebras for knowledge representation. Based on this formal setting, we use the combination of *semantic evaluation* and *proof theoretic* techniques for the design and implementation of Expert Database Systems.

We see expert databases as dynamic objects and use a system of modal logic for the specification of their dynamic properties. The possible worlds of our modal system are instances of the expert database system which are defined as many-sorted algebras. Thus our framework is a modal logic system of algebras where the signature of algebra is the basis for the schema of the expert database. With this setting, in addition to ordinary database operations, many sophisticated expert facilities can be provided.

1. INTRODUCTION

Conventional database systems are mainly concerned with storage and retrieval of data, and the efficiency of these activities. Generally, all of the data must be explicitly stored and there is no mechanism for deriving new facts from the existing information. In expert systems, which are often idealized as systems that can work like human beings, the emphasis is largely on the deduction of new facts, and they are not confined to the data stored. In addition to giving precise and complete answers to questions, expert database systems (or inferential databases) should be able to cope with queries such as: "What would be the consequence if X happens?" (hypothetical queries); "Why would X happen?" (causal queries); "What are the objects that are directly or indirectly related to a certain object?" (transitive closure); "Which objects can be candidates for solutions (in addition to the definite answers obtained from the database)?"

Thus the key issue is to find a suitable setting in which both data and knowledge can be stored, accessed and updated. While theorem proving

---

* A portion of the formal part of this article has been presented in [Gol-86].

techniques can provide the machinery to cope with the above, we note that, at least with the present technology, theorem proving is not efficient enough for answering ordinary database queries in systems with large amounts of data. In the case of simple database queries, we have straightforward computation where we know which actions must take place. Therefore, there is no need for random searches and tentative reasoning which require a great deal of computation time. To deal with the requirements above we propose the combination of semantic evaluation and proof theoretic techniques as tools for the design of inferential databases, whereby ordinary queries are computed straightforwardly, and deduction is used for more sophisticated ones.

Expert database systems (abbreviated to EDS) are seen as dynamic objects, where updates change the state of the database and the states are used for answering queries. For the dynamic part we develop a modal logic system. The domain of interpretation (or the universe) of a modal system for databases is the set of database instances, and the accessibility relation is determined by the update functions.

An EDS instance is seen as a collection of sets together with a collection of functions mapping these sets to each other. As in abstract data type specification methodology, we use the signature of the algebra as the basis for the type checker and the syntax checker of the database language. Queries are expressions which are built up out of the symbols in the signature and which comply with the precise formation rules given by the query language. The semantics of a query is defined to be the value which is assigned to it by the algebra representing a database instance. Integrity constraints are expressed as boolean valued expressions that must hold in all instances.

The power of deduction is provided by allowing inference rules which are activated by programs (queries) or by users. The inference rules are also expressions of type boolean (like the integrity constraints). Although the deduction rules can be invoked by the language processor, it is possible to define operators which explicitly trigger the inferencing mechanism. [Gol-84]

2. RELATED WORK

2.1. The Algebraic Approach:

This line of work has its roots in the research efforts on the specification of abstract data types [Zil-74 , Gut-76 , ADJ-77]. Since a conventional database (at a certain level of abstraction) can be viewed as an abstract data type, it has been proposed that a natural application area for algebraic specification theory is in the formal specification of databases. This view was taken by Ehrig et al [EKW-78] who proposed a hierarchic approach by building tables and sequences. Also taking this

view, Wirsing et al [DMW-82] gave a formal definition of databases and introduced primitives that mimic the generalization and aggregation structures [SmSm-77].

It is pointed out in [Gol-84] that a severe drawback of these works is that they do not cope with the dynamic aspects of databases. We eliminate this shortcoming by introducing our modal system.

Finally, we point out the novel extensions of our approach. We give the following extensions to the ordinary notion of universal algebra:

- we add "variable binding operators" such as the set-construction operator and the quantifiers, to the collection of operators;

- and we allow the operators in the algebra to work on union, cartesian product and powerset of types.

2.2. The Logical Approach:

Three directions can be distinguished here. First and the most prominent appraoch is using Prolog as the language for all purposes (which, as mentioned, is inadequate for large databases) [HaSe-84]. The second approach is interfacing a Prolog system with a relational database system which will have special problems due to the addition of a communication system. (For example, the authors of [VCJ-83] note the limitations of using Prolog directly as a system and interface it with a relational database. Some other variations are discussed in [PCG-86]. The third approach is to consider a richer framework that can cater for a wider variety of tasks. Our approach falls within this category.

Our **modal logic system** is similar to Hoare-style logics [Gold-82]. The motivation for its introduction came from our realization that the NEXT operator of temporal logic cannot adequately reason about the next EDS state because it assumes a fixed sequence of states. Note that from any state, depending on the update to be performed, there are many states that we can go to. (That is, "NEXT instance" will depend on the update.) Our modal operators are like the NEXT operator of temporal logic but are parameterized with respect to the update being performed.

2.3. The Artificial Intelligence Approach

A survey and tutorial on expert systems is presented in [HWL-83]. In this book, the authors describe many concepts which are important in the design and construction of expert systems, and then analyze a number of existing expert systems, including MYCIN [Sho-76] and its derivatives, DENDRAL and Meta-DENDRAL [BuFe-78], SAINT [Sla-61] and its successors, HEARSAY [Erm-80] and its other versions.

A quick study of these systems shows that none have a sound formal foundation and that generally they use ad hoc methods for finding solutions within large search spaces.

3. FORMAL SPECIFICATION OF EXPERT DATABASE SYSTEMS

The framework will be constructed in two steps. First, the static part will be developed and then we design the dynamic part on top of it.

We consider four classes of symbols within a given alphabet. We assume that the appearance of each symbol will determine which class it belongs to. The four classes are: sort symbols, function symbols, operation symbols, and variables. We also assume that the two sorts boolean and integer are given.

3.1. Primitives of EDS

We define inductively *simple-type-expressions* to be:

o   sort symbols

o   or of one of the forms:

    o $\gamma_1 \cup \gamma_2$,

    o $(\gamma_1 {}^* \gamma_2 {}^* \cdots \gamma_n)$

    o   and $P(\gamma_1)$,

where for some n for $1 \leq i \leq n$, γ_i is a simple-type-expression. We will see that $P(A)$ will be interpreted as the powerset of the set A.

Given n to be a natural number, a *function-type-expression* of arity n has the form

$$\gamma_1, \gamma_2, \cdots, \gamma_n \rightarrow \gamma_{n+1}$$

where for $1 \leq i \leq n+1$, γ_i is a simple-type expression. *Operation-type-expressions* are defined in a similar manner. For example, the operation type expression for the operation symbol "+" is

$$int, int \rightarrow int.$$

A *signature* is a function which assigns a function-type-expression to each function symbol and a sort symbol to each variable symbol. Thus, variables, both local and global, are typed (sorted) by the signature and not by the user. There is an unlimited supply of variables of each type. The signature is thus the specification for the type-checker and the syntax-checker of the language.

Given a signature Σ and a function symbol φ in the domain of Σ, the arity of φ in Σ is the arity of $\Sigma(\varphi)$.

Example: A small portion of an EDS for aviation purposes can be specified by giving the necessary sort symbols, function symbols and the function type expressions for each of the function symbols. Some of the sort symbols are 'flights', 'aircrafts', 'bases' and 'staff'. Some function symbols are 'destination_of', 'flight_given_to', 'captain_of' and 'crew_of'. Here we present the unique type-expressions for some of these functions.

captain_of	flights → staff
crew_of	flights → P(staff)
flight_given_to	aircrafts → flights
destination_of	flights → bases
is_suitable_for	aircrafts , flights → boolean

For a signature Σ, we define the set of well-formed expressions on Σ in the usual way. Details of this section were presented in [Gol-83]. Amongst the operations provided, there are a number of variable binding operators such as the logical quantifiers and the set construction operator. Bound and free occurrences of variables in expressions can be detected syntactically in the usual way.

Given a signature Σ, static *integrity constraints* on Σ are defined to be a collection of well-formed expressions of type boolean on Σ. We will use Γ_Σ for a set of integrity constraints on a signature Σ.

Example of a static integrity constraint on our avionic EDS is: (Variables are written in capital letters.)

"Ages of all crew members must be greater than 18"

$$\text{forall C (age\_of(C) GT 18)}$$

Given a signature Σ, *inference rules* on Σ (denoted by Ψ_Σ) are defined as closed expression of type boolean on Σ. Note that the formal definitions of integrity constraints and inference rules are the same and the only distinction is in their designation. An example of an inference rule on the avionic EDS is: "Aircrafts with seating capacity less than 10 fly in an altitude of less than 25000".

$$\text{forall AIRCRAFT (capacity\_of(AIRCRAFT) LT 10) implies}$$
$$\text{(altitude\_of(flight\_given\_to(AIRCRAFT)) LT 25000)}$$

An *EDS schema* is the triple $(\Sigma, \Gamma_\Sigma, \Psi_\Sigma)$ where Σ is a signature, Γ_Σ is a (possibly empty) set of constraints on Σ, and Ψ_Σ is a (possibly empty) set of inference rules on Σ, such that $\Gamma_\Sigma \cup \Psi_\Sigma$ is consistent. It is interesting to note that if Ψ_Σ is empty, then we have an ordinary database system.

Obviously one of the most important features of any system is the language provided by it. We will see in the following section that the query language constructed based on this formalism, despite the mathematical rigour, has a very simple notation. Although we cannot talk about hypothetical queries (because they are expressions of our modal system) here we will present other types queries which demonstrate the power of the proposed language. (Hypothetical queries will be introduced after discussing our modal system.)

For a given signature Σ, a *query* is a closed expression on Σ in which any variable is bound only once. As examples, we construct a few queries

of different types. The first query is an ordinary database query.

1. Who is the captain of the flight to which aircraft AR008522 is assigned?

    captain_of(flight_given_to(AR008522))

In the evaluation of this query, the result returned by the function 'flight_given_to' will be given to 'captain_of'. This is simple function composition.

2. Destinations of all flights that take off from base AAA before the hour 1400.

    { (F,destination_of(F)) |
         (origin_of(F) is AAA) and (departure_time_of(F) LT 1400) } F

While the variable F iterates over the elements of the set 'flights', a set is constructed that contains flight numbers and destinations of all those flights for which the said conditions hold. Obviously, F is a variable of type 'flights'. The appearance of F on the very right indicates the variable which is being bound by the set construction operator.

3. To demonstrate the capability for dealing with queries that involve computation of transitive closure, we formulate a query from a medical expert database system.

    Cures of all those diseases that can be caught as a result of having an ulcer.

This query will require the computation of transitive closure of the set-valued function "results_of_having". (Proofs on least fixed point and termination are presented in [Gol-83]).

    { (D1, cures_of(D1)) | D1 isin S } D1
        where
        S = results_of_having (ulcer) union
            Union { results_of_having(D2) | D2 isin S } D2

"union" is the ordinary set theoretical operator $\cup$. "Union" is the operator which, when given a set of sets, computes the union of all included sets. In the inductively defined part we have results_of_having(ulcer) as the basis.

3.2. Semantics of the language

Having discussed the syntax of the language, we use the notions of algebra to give semantics to the formalisms. Informally speaking, a many

714

sorted algebra is a function which assigns a set (a carrier) to every sort symbol and a function to every function symbol.

Recall that we allowed several forms to be simple-type-expressions. For a simple-type-expression γ, the set of all objects of type γ in an algebra A, denoted by $|A|_\gamma$ is defined as follows:

i-   when γ is a sort symbol then $|A|_\gamma = A(\gamma)$, that is the set that the algebra A assigns to it.

ii-   when γ is $\gamma_1 \cup \gamma_2$ then $|A|_\gamma = |A|_{\gamma_1} \cup |A|_{\gamma_2}$

iii-   when γ is $(\gamma_1 {}^*\gamma_2{}^* \cdots {}^*\gamma_n)$ then $|A|_\gamma = |A|_{\gamma_1}{}^* \cdots {}^* |A|_{\gamma_n}$

iv-   when γ is $P(\gamma_1)$ then $|A|_\gamma = P(|A|_{\gamma_1})$, that is, the powerset.

The evaluation in A of expressions is carried out in the usual way. Given an EDS schema $S = (\Sigma, \Gamma_\Sigma, \Psi_\Sigma)$ where Σ is a signature, and Γ_Σ and Ψ_Σ are as before, an algebra A is an S-algebra iff:

1.   For every function symbol φ in the domain of A, if $\Sigma(\varphi)$ is $\gamma_1, \gamma_2, \ldots, \gamma_n \to \gamma_{n+1}$ then $A(\varphi)$ returns an element of $|A|_{\gamma_{n+1}}$ when given an element of $|A|_{\gamma_1}$, an element of $|A|_{\gamma_2}, \cdots$, and an element of $|A|_{\gamma_n}$.

2.   The evaluation in A of all of the expressions in Γ_Σ results in true.

Let $S = (\Sigma, \Gamma_\Sigma, \Psi_\Sigma)$ be an EDS schema. An *EDS instance* is the ordered pair (S,A) where A is an S-algebra.

Given an expression Ω of type boolean, for an EDS instance i, we write $i \models \Omega$ iff i evaluates Ω as true. We will use I to indicate the collection of EDS instances on a given schema.

3.3. The Dynamic Aspects of EDS

The modal logic system that we will use for reasoning about dynamic characteristics of the EDS is comparable with temporal logic. It was first presented in [GMS-83]. Syntactically, we make a number of extensions. Σ is extended to Σ' by including:

o   update symbols $u_0, u_1, \cdots$,

o   for each of the update symbols $u_0, u_1, \cdots$, we introduce a corresponding modal operator $[u_0], [u_1], \cdots$,

o   global variable symbols $X_0^\gamma, X_1^\gamma, \cdots$ of each sort γ of Σ.

The set of well-formed expressions over Σ is extended to well-formed expressions over Σ' by allowing the construct $[\mu]\Omega$ as an expression of type boolean, where Ω is of type boolean and μ is an update symbol. The expression $[\mu]\Omega$ is read as: "after the update μ is performed Ω will be true". Note that each of the operators $[\mu]$ acts as an operator in a similar way to the more familiar modal operators ALWAYS, NEXT, etc. In fact, one can think of the $[\mu]$ as the $\bigcirc$ operator of temporal logic which is parameterized with respect to the update being made.

Given Ω, Ω_1 and Ω_2 as expressions of type boolean on Σ' and X as a global variable, we can extend our logic for deriving consequences by adding the following axioms and rule:

Ax1. $\;|- \; [\mu](\Omega_1 \rightarrow \Omega_2) \;\equiv\; ([\mu]\Omega_1 \rightarrow [\mu]\Omega_2)$

Ax2. $\;|- \; \neg[\mu]\Omega \;\equiv\; [\mu]\neg\Omega$

Ax3. $\;|- \;$ for all $X\;[\mu]\Omega(X) \;\equiv\; [\mu]$ for all $X\;\Omega(X)$

$$(X \text{ must be a global variable.})$$

Ru1. If $\;|- \;\Omega_1 \rightarrow \Omega_2$ and $\;|- \;\Omega_1,$ then $\;|- \;\Omega_2$.

Ru2. If $\;|- \;\Omega,$ then $\;|- \;[\mu]\Omega$.

The semantics for the modal extensions is defined as follows. For every update symbol μ in Σ' we consider a function $\bar{\mu}$ which when given an EDS instance returns an EDS instance, i.e.

where I is the set of EDS instances. Recall that we used $\;i \;|= \Omega\;$ to indicate that Ω holds in the instance i. We extend our notion of satisfaction to cope with modal expressions. Given an update symbol μ and the corresponding update function $\bar{\mu}$ we have:

$$i \;|= \;[\mu]\Omega \quad \text{iff} \quad \bar{\mu}(i) \;|= \;\Omega.$$

Other modal operators are not essential for our purposes but we can easily capture them if necessary. By adding the "null" update to our system we will get a modal system equivalent to S4.

Transition constraints are boolean type expressions over Σ'. (Transition constraints are statements that guard the system through updates.) For example, the constraint "ages cannot be reduced" is expressed as follows:

forall X forall Y ((age_of(X) is Y) implies ([u] (age_of(X) GE Y)))

This expression reads as follows: for any person X and any age Y, if the age of X is Y then after performing any update u the age of X will be at least Y.

3.4. Hypothetical Queries

Using this type of query, the user asks the expert system to make predictions based on certain assumptions. For example, in a company, the manager may ask the question: "if I increase Jack's salary by 1,000 dollars, would he then earn more than George?" Such a query is expressed as:

[increase_salary(Jack, 1,000)] sal_of(Jack) GT sal_of(George)

In this query the system will not make the update effective but will assume that the update is performed and then answers the question.

Similarly, suppose the management decides to ensure that no employee shall earn less than 20,000 dollars and they want to know whether a uniform 10% pay raise would achieve this. Such a query is formulated as:

forall EMP ([increase_salary(EMP, sal_of(EMP)/10)] sal_of(EMP) GT 20000)

This expression reads as follows: "For every employee, after increasing the salary of that employee by 10%, is it true that her salary will be more than 20,000 dollars?".

4. CONCLUSION

Research into the mathematical foundations of intelligent systems is of increasing importance for a healthy growth of computing technology. Database systems (with different degrees of sophistication) and expert systems are at the heart of a great deal of work on information processing systems, software engineering development environments and on many branches of artificial intelligence. We note that the progress has been slow for systematic software design methodologies. The well-recognized "software crisis" is a symptom of the limitations inherent in the current traditional approach to the specification, design and programming of complex systems. In recent years, these problems have been becoming steadily more apparent. They will be the dominant limiting factor in our ability to apply ever more powerful computing hardware to solve complex problems. Ideally, the current "brute force" methods and ad hoc design will be replaced by sound formally-based techniques.

Here, we have developed a formal setting for the specification of intelligent systems. Based on this setting, we were able to define a functional query language powerful enough to do several novel things. The notation of this language is based on the well-known conventional mathematical notation, similar to SETL and SASL.

5. REFERENCES

[ADJ-77]
  Goguen J A, Thatcher J W, Wagner E G, Wright J B
  Initial algebraic semantics and continuous algebras
  JACM Vol. 24, No. 1, pp 68-95, 1977
[BuFe-78]
  Buchanan B G, Feigenbaum E A
  "DENDRAL and Meta-DENDRAL: Their Applications"

Artificial Intelligence 11, pp 5-24, 1978

[DMW-82]

Dosch W, Mascari G, Wirsing M

On the algebraic specification of databases

Proc. of 8th VLDB Conference, Mexico City, September 1982.

[EKW-78]

Ehrig H, Kreowski H J, Weber H

Algebraic specification schemes for database systems

Proc. of 4th VLDB Conference, 1978.

[Erm-80]

Erman L D, et al

The HEARSAY-II speech understanding system

Computing Surveys, Vol.12, No. 2, pp 213-53, 1980.

[GMS-83]

Golshani F, Maibaum T, Sadler M

"A Modal System of Algebras for Databases Specification and Query/Update Language Support"

Proceedings 9th International Conference on VLDB, Florence, November 1983, pp 331-339

[Gol-83]

Golshani F

"A Mathematically Designed Query Language"

Research Report DoC 83-1, Imperial College, London UK

[Gol-84]

Golshani F

Tools for the Construction of Expert Systems

Proc. of 1st Intl. Workshop on expert database systems, Kiawah Island, SC, October 1984.

[Gol-86]

Golshani F

Specification and Design of Expert Database Systems

In "Expert Database Systems" (Kerschberg, ed.), Benjamin Cummings Publ. Co., 1986, pp 369-381.

[Gold-82]

Goldblatt R

"Axiomatising the Logic of Computer Programming"

Lecture Notes in Computer Science 130, Springer-Verlag, 1982

[Gut-76]

Guttag J V

Abstract data types and the development of data structures

Supplement to the Proc. of Conf. on Data Abstraction, definition and structures, SIGPLAN Notices 8, March 1976.

[HWL-83]

Hayes-Roth F, Waterman D A, Lenat D B

Building Expert Systems
Addison Wesley, MA, 1983.

[HaSe-84]

Hammond P, Sergot M
"APES: Reference Manual"
Logic Based Systems, Ltd., 40 Beaumont Ave., Richmond, Surrey, UK

[PCG-86]

Parker D S, Carey M, Golshani F, Jarke M, Sciore E, Walker A
Logic Programming and Databases
In "Expert Database Systems", (Kerschberg, Ed.), Benjamin Cummings Publ. Co., 1986, pp 35-49.

[Sho-76]

Shortliffe E H,
"Computer-Based Medical Consultation: MYCIN"
New York, Elsevier/North Holland, 1976

[Sla-61]

Slagle J R
A heuristic program that solves symbolic integration problems in freshman calculus
PhD Diss., Report 2G-0001, Lincoln Lab., MIT, 1961.

[SmSm-77]

Smith J M, Smith D C P
Database Abstraction: Aggregation and Generalization
ACM TODS Vol.2, No. 2, 1977.

[VCJ-83]

Vassiliou Y, Clifford J, Jarke M
"How does an expert system get its data?"
Extended Abstract, Proceedings 9th International Conference on VLDB, Florence, November 1983, pp 70-72

[Zil-74]

Zilles S N
Algebraic specification of data types
Project MAC Progress report, CSG Memo 119, MIT, MA, 1974

Toward optimal feature selection:
Past, Present and Future.

Wojciech Siedlecki and Jack Sklansky
University of California, Irvine
Irvine, CA 92717

ABSTRACT. Over the last twenty-five years, extensive research has taken place on the development of efficient and reliable methods for the selection of features in the design of pattern classifiers, where the features constitute the inputs to the classifier. The quality of this design depends on the relevancy, discriminatory power and ease of computation of various features.

Selecting features is an extremely difficult task, charged both with theoretical and computational problems. An effective mathematical theory for feature selection seems achievable only for a very specific aspect of the problem: linear transformations for reducing the dimensionality of the feature space, with the assumption that data are drawn from normal distributions [1,2,3,4]. The theoretical problems are usually associated with two closely related questions:

a) "What does it mean that a feature is good or irrelevant?"
b) "What criteria should be used to evaluate features?".

From the standpoint of Bayesian decision rules there are no bad features. One never can improve the performance (usually understood as an error committed by the classifier) of a Bayes classifier by eliminating a feature (this property is called *monotonicity*). However, in practice the assumptions in the design of Bayes classifiers are (almost) never valid. As a consequence, it is possible to improve the performance of a nonideal classifier by deleting a feature (this phenomenon will be discussed later). Moreover, for a given amount of data, reducing the number of features increases the accuracy of estimates of the classifier's performance. These two facts have tremendous consequences for computational problems associated with feature selection and have led in the past to other methods for evaluating features [5,6,7]. These methods do not evaluate the performance of a classifier associated with a given set of features, but rather tend to approximate the Bayes error for this set of features. The criteria used by these methods (for instance Bhattacharyya distance or Vajda's entropy) satisfy the monotonicity property, which permits the use of efficient computational techniques. However, some evidence [8] indicates that they do not induce over an arbitrary set of features the same preference order as would be obtained by comparing the errors of the Bayes classifier. Thus, it seems that the only promising and legitimate way of evaluating features must be through the error rate of the classifier being designed (this also satisfies our intuitive understanding of the design policy, although it has some theoretical drawbacks [5,9]).

Unfortunately, so far none of the forms of classifiers realisable in practice by known techniques exhibits the monotonicity property. This fact is important when we realise that the problem of feature selection is essentially equivalent to searching a directed graph (at the root node all features are accepted) and could be solved by artificial intelligence or "AI" (e.g. *branch and bound* [6]) techniques. Moreover, the total number of all possible subsets of an n-element set of features totals around 2^n and, therefore, even for small n (say, 10) any brute force method leads to a computational dead end (specially when the evaluation of classifier's error is costly).

Over the last five years, intensive research on feature selection has been carried out at University of California, Irvine [10], leading to a group of suboptimal but efficient and robust methods. This group includes:

1. methods utilising the idea of *approximate monotonicity* [10],
2. other AI methods for graph searching.

Another promising method, currently under consideration, is based on the observation that the monotonicity property of classifier's error rate is highly related to the optimality of this classifier [11]. This method does not require any search, but evaluates all features at the same time in a *fuzzy* decision process that involves the assignment of a weight to each feature. In this report we will discuss the above three classes of methods in more detail.

1. THE PAST: A HISTORICAL NOTE. The pioneering work in the area of feature selection is associated with the names of Sebestyen [12], Lewis [13] and Marill and Green [14], who made their contributions in the early sixties. Since at that time the theoretical framework for evaluating the error rate of classifiers was also in its preliminary stage of development, the original approaches to feature selection were based on the concept of probabilistic class separability measures and entropies. In some cases (e.g. [13]) the independence of features was assumed and the features were selected on the basis of their individual merits. However, even such a simplified model did not guarantee the optimality of a selected feature subset (for instance, two independent features don't have to be the two best, as was pointed out by Cover [15]).

The question of the trade-off between the optimality and efficiency of algorithms for NP-problems (feature selection, by definition, seems to qualify as an NP-problem) was recognized early, and the mainstream of research on feature selection was thus directed toward suboptimal search methods. The invention of sequential backward selection (SBS) in 1963 [14] gave rise to a family of suboptimal stepwise forward and backward methods. The research in this direction was concluded by introducing the generalization of these algorithms proposed by Kitler in 1978 [16]. Another approach to feature selection based on the concept of dynamic programming was proposed by Chang [17], but this approach is burdened by numerous restrictive requirements (e.g. the monotonicity condition and statistical independence of features) and, therefore, has not been heavily pursued by other researchers.

The potential of any suboptimal search algorithm to select the worst possible set of features was indicated by Cover and Campenhout [18]. A breakthrough came in 1977 with the introduction of the *branch and bound algorithm*. The application of this method, proposed by Narendra and Fukunaga [6], guaranteed the selection of an optimal feature subset if *the monotonicity condition* is satisfied. The monotonicity condition requires that a criterion function J used to evaluate feature subsets change (in our case: grow) monotonically over a sequence of nested feature subsets $\{F_1, \ldots, F_k\}$, that is

$$F_1 \subset F_2 \subset \ldots \subset F_k \quad \Rightarrow \quad J(F_1) \geq J(F_2) \geq \ldots \geq J(F_k). \tag{1}$$

Based on this concept Narendra and Fukunaga also clearly defined which subset of features could *not* be considered optimal. Roughly speaking the branch and bound procedure searches in an optimally organized way the feature selection lattice, Fig.1. (In the lattice, nodes represent feature subsets and links represent the relation of subset inclusion. The subsets are coded by sequences of zeros and ones. One means that a feature is present in a subset and zero means that the feature does not belong to it. The percentages next to the nodes denote observed error rates of a hypothetical classifier.)

Since the kind of graph generated in the feature selection problem (each node represents a subset of features) has finite depth, the depth first search technique appeared very effective in this case and has resulted in a very efficient enumeration scheme.

When no restrictions on examining nodes (feature subsets) in the graph have been assumed, the branch and bound leads to exhaustive search. However, if each node is evaluated with the aid of a criterion function
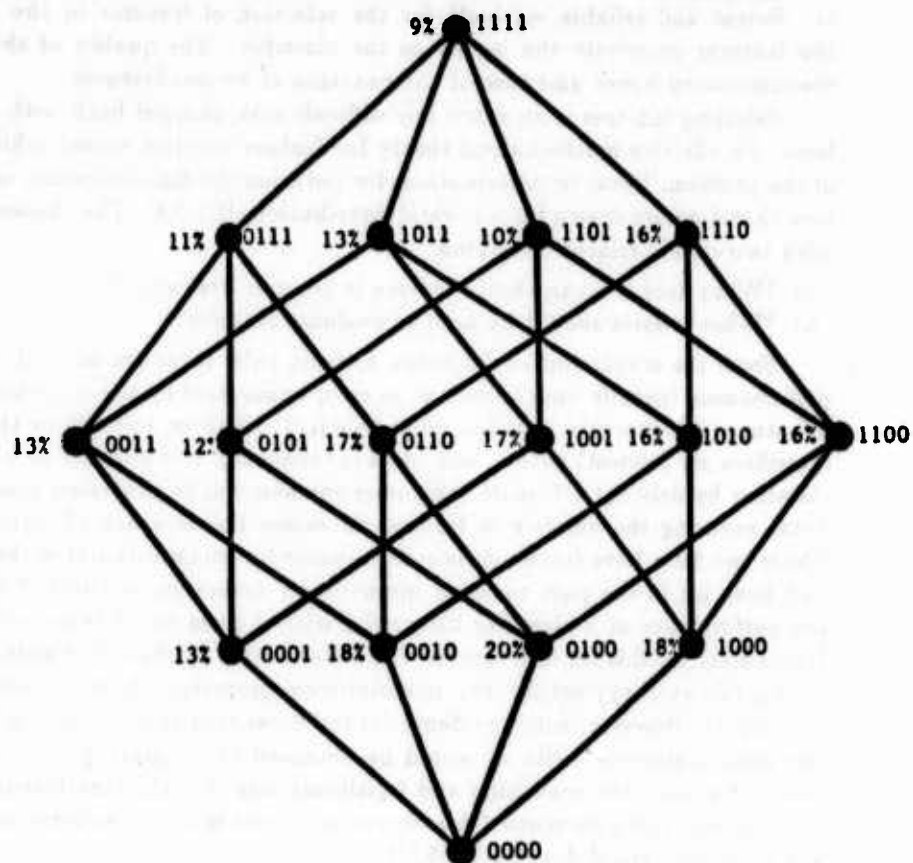
Fig.1.

J_1 and an upper limit (threshold) for its acceptable values is set (that is, some feature subsets are considered *infeasible*), then the algorithm backtracks whenever an infeasible node is discovered. If the criterion function has the monotonicity property (1), no feasible node is omitted as a result of early backtracking and, therefore, the gained savings in the search time do not violate the "optimality" of the selection procedure. Now, among all examined and, therefore, feasible subsets of features one can look for the best group of features according to a second criterion J_2. If J_2 is also monotonic with respect to a sequence of nested feature subsets, but in the direction opposite to that of J_1, then J_1 and J_2 can be interchanged, yielding a search for a feasible node among the best nodes, which is equivalent to a backward branch and bound scheme (in which one takes the empty set of features as a start node).

All the considerations regarding the branch and bound procedure as applied to optimal feature selection are valid only for monotonic evaluation functions. Narendra and Fukunaga originally proposed to use probabilistic separability measures as criterion functions. This approach, however, has a number of disadvantages:

   a) feature selection is done based on finite samples and should refer to any particular classifier's performance rather than to intrinsic discriminant properties of the data, which cannot be reliably uncovered due to the sampling process,

   b) to use any of these criteria one has to estimate them based on the sample, which introduces some error and can turn a monotonic criterion into a nonmonotonic one, and

   c) as some evidence indicates [8] certain criteria in this class can give results which are optimal in the sense outlined above but the selected subset of features need not be optimal, where the optimality refers to expected performance of a classifier which would use this subset of features.

While the second disadvantage can prevent the search procedure from finding an optimal solution, the first and the third ones are much stronger when a selected subset of features is to be used to build a practical classifier.

As many authors pointed out, the only remaining alternative is to use the error rate of a classifier as a design criterion. Unfortunately, due to phenomena similar in origin to those mentioned in the second of the above disadvantages, the error rate of a classifier (if it is not a Bayes classifier) does not satisfy the monotonicity condition. Such a case can be observed in the feature selection lattice presented in Fig.1. The error rate along the path (1111)-(1101)-(1001)-(1000) has a monotonicity defect at the node (1001). So far the lack of monotonicity in the classifier's error rate made it useless for the branch and bound procedure. In 1985 Foroutan and Sklansky [10] introduced the concept of *approximate monotonicity*. Based on the example of a locally trained piecewise linear classifier they showed that the error rate might be used for optimal branch and bound. Although the supporting tests were done only for one data set the idea of approximate monotonicity constitutes another breaktrough in understanding and applying methods for optimal feature selection for classifiers trained on finite samples.

2. THE PRESENT: APPROXIMATE MONOTONICITY AND AI GRAPH SEARCH METHODS.

The concept of approximate monotonicity opened a new chapter in the research on optimal feature selection. It allows the use of branch and bound to obtain with high confidence an optimal subset of features even though the monotonicity condition is in some cases to some extent violated. Below we discuss two ways of coping with the negative effects of the lack of monotonicity in the error rate on the optimal branch and bound search procedure. Also we present other approaches to feature lattice search, originating from artificial intelligence (AI).

2.1. THE BRANCH AND BOUND PROCEDURE FOR NONMONOTONIC CRITERIA.
In their work [10] Foroutan and Sklansky used a tolerance factor imposed on the assumed threshold for branch and bound search. Namely, if the assumed upper limit of the error rate for a subset of features to be considered feasible is e_{max}, then a subset of features, F, in fact is assumed

   a) feasible, if the associated error rate $e(F)$ is less or equal to e_{max},
   b) conditionally feasible, if $e_{max} < e(F) \leq e_{max}(1 + \Delta)$ and
   c) infeasible, if $e(F) > e_{max}(1 + \Delta)$.

In this version the best subset of features is chosen only from the set of feasible nodes in the feature selection lattice, but also conditionally feasible nodes are examined. In Fig.2. a feature subset (00110101) is found

conditionally feasible and search is continued. **This allows the branch and bound algorithm to examine**
the feature subset behind it, which is feasible. In
experiments described in [10], despite the lack of
strict monotonicity, a procedure using the error
tolerance was able to find an optimal subset of
features with over 90% in computational savings
compared with exhaustive search.

Another way of avoiding the negative effects
of using an estimated and, therefore, by defini-
tion nonmonotonic error rate of a classifier is to
estimate the expected value of the error rate for
an examined subset of features. Assuming that
departures from monotonicity are an effect of es-
timation errors and the classifier's true error rate
should increase monotonically over a sequence of
nested feature subsets (1) we can try to estimate
the general trend of error rate changes in the prac-
tical classifier. We consider the error rate a func-
tion of nested feature subsets, which corresponds
to a path in the feature selection graph. Since the
observed error rate may not be monotonic we can
observe that along this path it rises and falls even
though the expected error rate does not decrease.

Fig.2.

As a result, it might happen that the current node is infeasible based on its observed error rate, but that it
ought to be feasible because the expected error associated with a classifier trained on an infinite sample is
below the threshold of acceptability. In Fig.3. the feature subset (00110101) would be considered infeasible
based on the observed error rate associated with it. However, as one can notice the trend along this path
(in this case we use linear prediction) indicates that the value of the expected error rate associated with this
subset should be less than the presumed threshold and, therefore, the subset is treated as if it were feasible.

Thus, if we analyze the trend of changes of
the observed error rate over a sequence of nested
feature subsets and, based on this information, we
use the approximation of the expected error rate
rather than the currently observed error rate, we
may successfully use the branch and bound algo-
rithm to search for the optimal subset of features.

The strategy of enumeration in the branch
and bound method is another important factor in-
fluencing the efficiency and optimality of the fea-
ture selection process. When the monotonicity
condition is satisfied it does not matter in which
order we will examine the descendants of the cur-
rent node — we will always find the optimum solu-
tion and the number of visited nodes in the feature
selection lattice will be about the same in each
case. In this case one usually takes the node with
the highest error rate as the next current node,
for it increases the chance for finding the next in-
feasible node and consequently for pruning some
part of the lattice below it. However, when the
monotonicity condition is not satisfied by using

Fig.3.

this strategy we could prune a part of the lattice including feasible nodes and, which is likely, the best node.
Such a case is observed in the feature selection lattice depicted in the Fig.1. Here, if the threshold is set
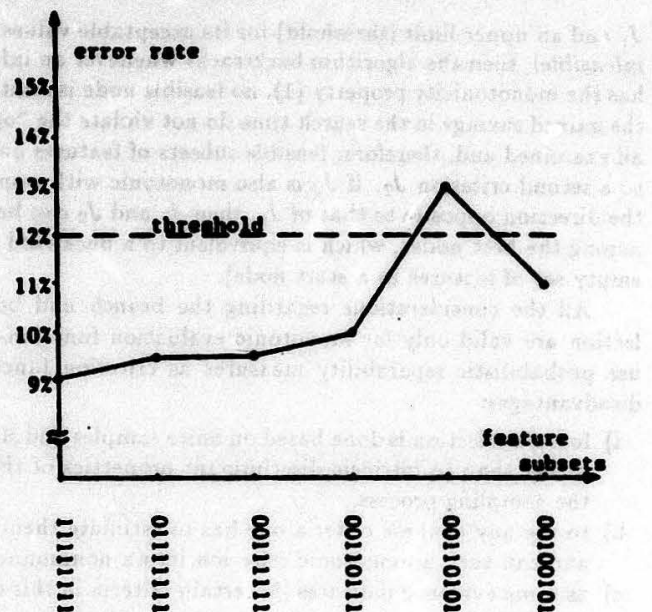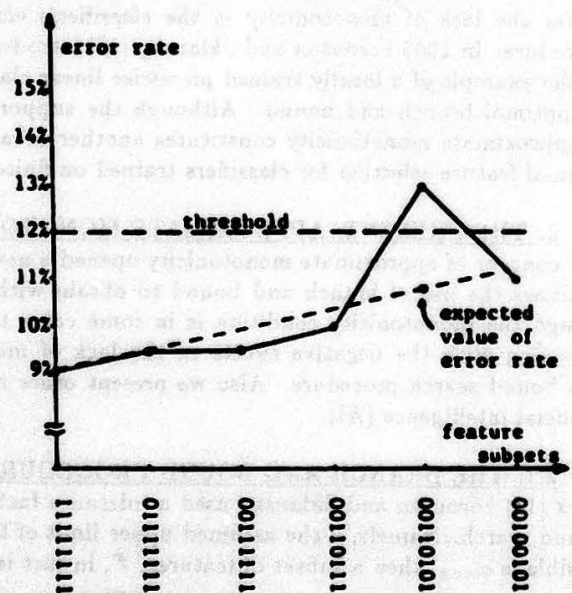
to 15% the optimal node (0001) will be pruned due to the fact that it is a subset of the set (1001), which is infeasible. On the other hand, if we select a subset with the lowest error as the nex node in the lattice this subset (0001) will be discovered in the sequence (1101)-(0101)-(0001). Hence, the strategy in which we choose the node with the lowest error rate is more likely to avoid local nonmonotonicities and continue search in those parts of the feature selection lattice that would be skipped by using the traditional strategy.

An important question in feature selection with the aid of branch and bound is how to choose the threshold that defines the feasibility of subsets of features. We can assume that we do not want a big degradation of the classifier's performance and, therefore, we set the threshold at a low level. However, we don't know what price we will pay for selecting an optimal subset of features. In other words, we do not know in advance if the cost of removing one feature from the selected optimal subset would only minimally increase the value of the error rate or whether by adding one feature we can significantly improve the classifier's performance. This might be important, since by properly setting the threshold we could avoid an examination of a significant number of nodes in the feature selection lattice.

A way to predict the best value of the threshold would be to use a sequential forward or backward method to look for the best path in the feature selection lattice, where the best path is a path along which the error rate increases as slowly as possible. By doing this we can scan the feature se-

Fig.4.

lection lattice and obtain a function, Fig.4., depicting the trade-off between the number of removed features and an expected threshold, which must be set in order to find an optimal subset of features of this size. Or, on the other hand, we can estimate the size of the optimal feature subset given some threshold.
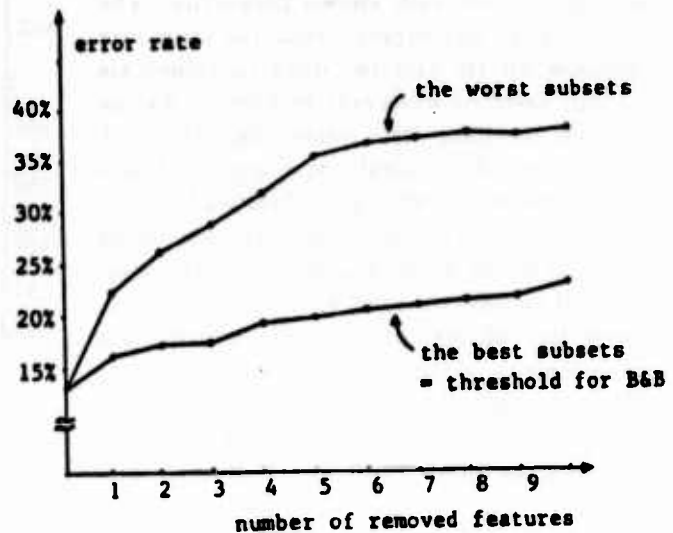
Unfortunately, both forward and backward selection can easily be derailed. For instance, the forward selection algorithm can add two features which are subsequently the best ones but they are bad if used together. This could be to some extent avoided if the sequential forward and backward methods are used at the same time. We call this method a *bidirectional search*. Its concept originates from the MEA (i.e. *means-ends-analysis* method used for problem solving in AI.

In the bidirectional search we conduct the search for the best path from two end nodes (that is, the node representing the full set of features and the node associated with the empty set) at the same time. The feature selection lattice is examined in a DFS (*depth first search*) fashion, in two directions:

a) from the full set node toward the empty set node and
b) from the empty set node toward the full set node.

The search is conducted simultaneously from both terminal nodes and concludes in the middle of the lattice, resulting in a path that goes from the top to the bottom of the feature selection lattice. The path is determined a by local comparison of values of the criterion J associated with the feature subset evaluation. At every step, in the forward as well as in the backward search, for the current feature subset all its successor nodes (its subsets in the forward direction and supersets in the backward direction) are evaluated and the most promising ones are selected. If there is a conflict, then the second best successors are selected. A conflict arises if at a given step the same feature is selected in both directions, that is, is chosen to be both added and discarded. This corresponds to the situation in the sequential forward selection algorithm mentioned above: a feature is considered good by the forward selection method but the backward selection algorithm indicates that it also could be removed with no harm. In other words, the conflict suggests that the information obtained from the two methods is contradictory and should be disregarded. If this conflict were not resolved, it would be impossible to conclude both searches in the same place in the feature

selection lattice for no path connecting the two current nodes contains both nodes determined by adding and removing the same feature at the same time.

An example result of using the bidirectional search procedure is given in Fig.5. This comparison was made for the feature selection lattice obtained from a piecewise linear classifier trained on a synthetic data set with known properties. The analysis of the lattice suggests that the error rate is nonmonotonic (in fact the data contained six deliberately inserted irrelevant features). As one can see, the resulting error rates along the path selected by the bidirectional search seem to follow closely the minimum error rates obtained from exhaustive analysis. This encourages the use of the bidirectional search algorithm to predict the value of threshold for efficient branch and bound search. Moreover, the bidirectional search is insensitive to the monotonicity of the error rate function.

There is one additional benefit of scanning the feature selection lattice for the best path: we can estimate the number of nodes that have to be examined.
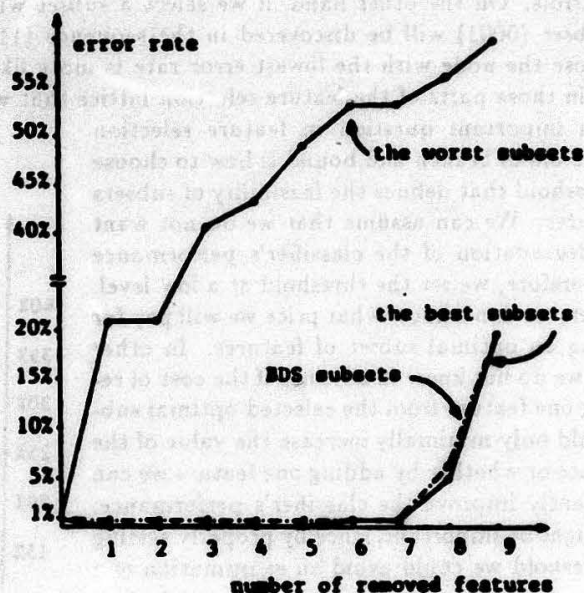
Fig.5.

2.2. OTHER AI METHODS IN FEATURE SELECTION.
Both branch and bound and MEA are commonly recognized as techniques developed by AI researchers. While the branch and bound algorithm can supply the optimum solution, provided that some conditions be satisfied, the other AI techniques are typically heuristic and they generally do not give a guarantee of finding the optimal solution. However, when the original dimensionality d of feature space is large (say $d > 15$) then the optimality must be given up because the complexity of the problem impedes the use of the branch and bound technique. Of course, one could start enumerating nodes in the feature selection lattice in the backward fashion, that is from the node associated with the empty set. However, if the selected threshold allows the algorithm to visit nodes too deep in the lattice, this solution would be as useless as the original version of the branch and bound enumeration scheme. In such a case we have to look for substitute solutions, which are most likely nonoptimal.

The branch and bound technique in application to the search in the feature selection lattice is nothing but a method of enumerating nodes in this graph. Its advantage over other possible enumeration schemes is such that no node is examined more than once and, therefore, by forcing the algorithm to backtrack earlier than at the terminal node corresponding to the empty set, we can eliminate parts of the lattice, which for some reason (in our case the nodes with excessive error rates) are of no interest to us, and increase the efficiency of the search.

Other known AI techniques do not have this property. They guarantee the optimum solution in feature selection only if they are allowed to do exhaustive search. The following are a few examples:

a) DFS, depth first search, which, if terminated without backtracking, turns out to be the sequential forward or backward selection,

b) BFS, breadth first search, which has no equivalent in feature selection literature and

c) best-first search, which also has no equivalent.

In the best-first search method one expands the top node and builds from its descendants a queue according to decreasing values of the so called *heuristic evaluation function* associated with them. Next, the first node in the queue is expanded and the queue updated. This process is repeated until a *goal node* is detected. In the feature selection problem we do not explicitly look for a goal node, because we are unable to detect whether a given node is a goal or not. Instead, we are interested in searching some part of the feature selection lattice, which should contain the node that is optimal with regard to some assumed criterion.

726

The heuristic evaluation function would be another unknown in the definition of the best-first search method. For instance, one could assume that the observed error rate of a classifier is this function.

The major disadvantage of the best-first search algorithm is its space complexity (i.e., the size of the computer memory required to execute an algorithm), which in the worst case, when all nodes from the middle level in the lattice have to be placed in the queue, equals

$$\binom{d}{\lfloor d/2 \rfloor} \le 2^{d-2} \tag{2}$$

where d is the total number of features. This number is prohibitive even for d as small as 20, so the full queue cannot be stored in computer's memory. However, we can assume in advance the maximum size of the queue and this version of the best-first search procedure is refered to as *beam search* technique. For instance, only the feasible subsets can be stored in the queue (although this does not give a full guarantee that the queue will be limited to a reasonable size).

The limitation of the size of the queue has two consequences. First, the space complexity is much less and can be arbitrarily set. Second, some nodes and, as a result, some parts of the feature selection lattice are never visited, which significantly improves the efficiency of the beam search compared to the exhaustive search scheme pursued by the best-first search algorithm. The second observation suggests also that the beam search procedure may not find the optimal solution (unless all feasible nodes are stored in the queue and the error rate has the monotonicity property).

Recently we have conducted experiments with a version of the beam search procedure which incorporates into the heuristic evaluation function not only the error rate associated with the current node but also uses some prediction scheme to speed up the search process. Its algorithm contains the same elements as the original beam search. First the top node is expanded and the priority queue built according to increasing values of the error rate associated with each descendant. Next, from a few levels ahead some assumed number of nodes is drawn at random from the lattice. If there is a node with the error rate lower than the error rate of the first node in the queue, then as the current node we choose the best node in the queue from which there is a path toward the node on the lower level. Otherwise, we select the first node in the queue as the best node. Finally, we expand the best node and the process of generating goal nodes (these nodes are drawn only from levels below the level at which the current best node is placed) and for selection of the next best node is repeated. The natural stop condition is satisfied if all prospective nodes are infeasible with regard to a given threshold.

The drawback to this method is that, for a high dimensionality of the feature space, the number of nodes checked may become very large, and the stop condition may not be reached soon enough. This could be resolved in one of the following ways:

   a) by setting the number of nodes checked to a finite number,
   b) by making the stop condition user interactive or
   c) by using the two options given above.

We have tested this algorithm on the data used for the bidirectional search. The results are very encouraging: for each data set this method performed as well as the branch and bound algorithm, but the number of examined nodes was much less. However, we emphasise that the beam search algorithm is suboptimal, and for this reason it is not competitive with branch and bound wherever the latter method can be used. On the other hand, its usefullness can be appreciated in feature selection problems in which the dimensionality of the feature space prohibits the use of the branch and bound enumeration scheme.

Another interesting aspect of the beam search technique is that it can be viewed as a generalization of the popular sequential selection methods. Namely, if the queue size is assumed to be equal to one and we start searching from the node associated with the full set of features, then this algorithm is equivalent to the sequential backward selection method.

3. **THE FUTURE.** The techniques for feature selection discussed so far assumed a search for the best subset of features among a number of feasible subsets. Such a statement of the problem presents several disadvantages:

   a) It leads to an NP-problem, which for larger tasks must be solved with the aid of suboptimal methods. These methods, by definition, do not guarantee that the selected subset is optimal,

b) Given a selected optimal subset of features we are still unable to determine the usefulness of a particular feature (called sometimes its *discriminatory power*).

c) If a feature selection process uses a criterion function involving an error rate of a classifier, then it must be immediately recognized as a process in which this classifier is optimized and, therefore, trained. Hence, the only error rate that can be computed for this classifier is an apparent error rate, which is known to be very biased.

It is very likely that a panaceum for all these problems does not exist, although we can try to solve each of them independently. For instance, instead of selecting a subset of features we can evaluate a discriminatory power of each feature. A potentially promising approach is based on the concept of classifiers optimized with regard to the use of available features [11]. In this approach we define a classifier as a function $F: X \times P \to \Omega$, where X is a feature space, P is a set of parameters of the classifier and Ω is a set of class labels (decisions). We assume that our classifier is trainable with respect to a criterion function $J: P \to \Re$, where $\Re$ is a set of real numbers, that is, we can find a parameter vector $p^* \in P$ such that $J(p^*)$ is a minimum. In this notation a classifier $F(., p^*)$ is assumed to be an optimally trained classifier. Now suppose we discard the i-th feature, that is for any two feature vectors

$$x = [x_1, \ldots, x_i, \ldots, x_d]^T, \; x' = [x_1, \ldots, x_i', \ldots, x_d]^T \quad \text{and} \quad x_i \neq x_i' \quad \text{but} \quad F(x, p) = F(x', p).$$

We call a classifier a *scalable classifier* if the effect described above can be accomplished by imposing a certain value to the parameter vector, p. In fact, many known classifiers, including linear, piecewise linear and quadratic ones are scalable classifiers. Other types of classifiers like k-NN rule or classifiers based on density function estimation, which involve the use of distance functions, can be transformed to satisfy the definition of scalable classifiers.

In [11] we have shown that if a classifier is a scalable classifier then its error rate satisfies the monotonicity condition provided that we use an optimum training procedure to minimize the error rate of the classifier. This theorem can be rephrased for a linear classifier into the following form: if a linear classifier is trained with the aid of a procedure that guarantees a minimum resubstitution error rate then this error rate is monotonic over a sequence of nested feature subsets. With some additional assumptions a similar theorem was proven for piecewise linear classifiers [10]. Now, if we optimally train a scalable classifier, then we will receive a vector of optimal parameters, p^*. Since some of these parameters are responsible for amplifying or reducing the influence of each feature (for instance, weights in linear classifiers), then by comparing them we can deduce the discriminatory power of each feature (assuming that all features are statistically equivalent, that is, they have the same mean and variance).

The following approach can be called a *fuzzy* formulation of the feature selection problem:

"Given a scalable classifier, train it optimally over a set of statistically equivalent features and compare parameters associated with each feature. These parameters can be viewed as values of a fuzzy membership function computed for associated features and their relatively large values indicate high discriminatory power of these features."

Unfortunately, the optimal training procedures are not known so far.

The approach sketched above may solve the first two problems. However, the problem of biasedness of an error rate of a classifier built for a selected optimal subset of features is more complicated. We could try to use a form of cross-validation for feature selection. Given a finite sample we divide it into two parts: the training set and the test set. Next, we design a classifier and perform feature selection for this classifier based on its observed error rate. Finally, we compute a new error estimate for the classifier whose design is based on the selected optimal subset of features. This procedure is repeated a number of times, and each time we divide the data set into two subsets in a different way. At the end of the process we take a mean of all estimated final error rates. This estimator is known as a *rotation error* estimator, and is less biased than the resubstitution error estimator. However, this solution has one significant drawback: it requires that the feature selection process be repeated a number of times, which further increases the already discouraging computational complexity of the problem.

728

BIBLIOGRAPHY

[1] J. Kittler, "On the discriminat vector method of feature selection," *IEEE Trans. Comput.*, vol. C-26, pp. 604-606, 1977.

[2] H. P. Decell, P. L. Odell and W. A. Coberly, "Linear dimension reduction and Bayes classification," *Pattern Recognition*, vol. 15, pp. 51-54, 1979.

[3] D. M. Young, "A formulation and comparison of two linear feature selection techniques applicable to statistical classification," *Pattern Recognition*, vol. 17, pp. 331-337, 1984.

[4] S. D. Morgera and L. Datta, "Toward a fundamental theory of optimal feature selection: Part I," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. PAMI-6, pp. 601-616, 1984.

[5] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," in P. R. Krishnaiah and L. N. Kanal, eds., *Handbook of Statistics, Vol. 2*, North-Holland Publ. Comp., 1982.

[6] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, pp. 917-922, 1977.

[7] K. S. Fu, "Recent developments in pattern recognition," *IEEE Trans. Comput.*, vol. C-29, pp. 845-854, 1977.

[8] M. Ben-Bassat, "f-entropies, probability of error, and feature selection," *Information and Control*, vol. 39, pp. 227-242, 1978.

[9] M. Ben-Bassat, "On the sensitivity of the probability of error rule for feature selection," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. PAMI-2, pp. 57-60, 1980.

[10] I. Foroutan and J. Sklansky, "Feature selection for piecewise linear classifiers," in IEEE Proc. Computer Vision and Pattern Recognition, San Francisco, 1985.

[11] W. Siedlecki and J. Sklansky, "On the optimality of classifiers and the selection of features," Int. Rep. Pattern Recognition Project, School of Engineering, University of California, Irvine, 1985.

[12] G. Sebestyen, *Decision Making Processes in Pattern Recognition*, New York, 1962.

[13] P. M. Lewis, "The characteristic selection problem in recognition systems", *IRE Trans. Info. Theory*, vol. IT-8, pp. 171-178, 1962.

[14] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems", *IEEE Trans. Info. Theory*, vol IT-9, pp. 11-17, 1963.

[15] T. M. Cover, "The best two independent measurements are not the two best", *IEEE Trans. Systems, Man and Cybern.*, vol. SMC-4, pp. 116-117, 1974.

[16] J. Kittler, "Une généralisation de quelques algorithmes sous-optimaux de recherche d'ensembles d'attributs", *Proc. Congrès Reconneisance des Formes et Traitment des Images*, Paris, 1978.

[17] C. Y. Chang, "Dynamic programming as applied to feature selection in pattern recognition systens", *IEEE Trans. Systems, Man and Cybern.*, vol. SMC-3, pp. 166-171, 1973.

[18] T. M. cover and J. M. Van Campenhout, "On the possible orderings in the measurement selection problem", *IEEE Trans. Systems, Man and Cybern.*, vol. SMC-7, pp. 657-661, 1977.

INTRODUCING TREATMENTS INTO TEST PROCEDURES

D.W. Loveland[1]
Computer Science Department
Duke University
Durham, NC 27706

Abstract. The problem of finding low-cost testing procedures to isolate a faulty, or diseased, object has been extensively studied. We enlarge the problem by allowing treatment of the faulty object before the identity of the object is completely known, a common occurrence in real life. The problem is formalized, a general well-known solution method is mentioned, and a limited subcase is explored, including presentation of exact (optimal) and approximate solution methods. We illustrate the complexity of the extended problem and why analogous results to the simpler binary testing problem is unlikely.

1. Introduction. There is an extensive literature in the binary testing problem, a problem featuring the analysis of optimal and near-optimal test procedures with respect to expected cost. (See [1] for a survey of literature on this general problem.) The solutions are presented as decision trees. While working in this area we were struck by the fact that for computer scientists, physicians and anyone else interested in repairing faults as well as finding faults this was the wrong problem. In many situations one wishes not to isolate the fault as the final solution but to treat the fault, and treatment may often occur before the fault is isolated. Indeed, the treatment can also be in part a test: "Take two aspirin and, if not better, see me in the morning". Surprisingly, no theory of tests and treatments parallel to the theory of binary testing seems to appear in the literature. We outline a model for the test-and-treatment problem and present some initial results for this model. A special case of the binary testing problem (the "complete" test case) has a simple and quickly computed solution, the Huffman coding procedure. We had hoped that there might be an interesting generalization of the Huffman procedure for the analogous case in the test-and-treatment problem but the rich interactions between nodes, which themselves are clusters of treatments, seems to preclude a simple algorithm to solve this special case. We then discuss an approximation algorithm for the simplest case and illustrate the node interaction that makes finding simple optimal algorithms difficult.

2. The Model. We first present the model for the binary testing problem because the model we consider is an extension of this binary testing model. The binary testing problem is presented by n objects, n *a priori* probabilities of fault, and m tests with associated costs. Let $U = \{o_1, \ldots, o_n\}$ denote the n objects, let $\{p_1, p_2, \ldots, p_n\}$ denote the n *a priori* probabilities that report the user's estimate of the likelihood of the corresponding object being faulty, and let $\{T_1, \ldots, T_m\}$ denote the m binary tests with associated costs $\{C_1, \ldots, C_m\}$.

We make various assumptions to simplify the problem analytically. We assume that there is only one faulty object, so $\sum p_i = 1$. The assumption that tests are binary means that they are reliable and unambiguous; in particular we can model a test by a subset of the universe. The test set is defined as follows: an object is placed in the test set if the test gives a positive

---

response when that object is the faulty object. We will let T_i denote the test set as well as the underlying test, because functionally they are equivalent. Previous analytical work has usually assumed that the tests all have the same cost, chosen arbitrarily as a unit cost; i.e., $C_i = 1$, all i.

Although this may seem draconian, much has been learned using this restriction. This knowledge serves as a message regarding the more general case with arbitrary costs. (See [2], [3]). We return to this point later.

The outcome of the problem is a decision tree that instructs one as to how to apply the tests, where the choice of test is a function of the outcome of previous tests. For any particular problem one follows a single path of the tree, branching as determined by test outcome, until the faulty object is isolated. We seek the tree of minimum expected cost, where expected cost is given by

$$EC = \sum_{i=1}^{s} Path_i \cdot p_i \tag{1}$$

with $Path_i$ defined as the sum of the costs of the tests encountered. In the case of uniform cost for tests $Path_i$ becomes the number of tests encountered.

The model for the (binary) test-and-treatment problem that we adopt here extends the binary testing model by the addition of treatments $\{T_{m+1}, ..., T_{m+r}\}$ with associated *a priori* probabilities $\{p_{m+1}, ..., p_{m+r}\}$ and associated costs $\{C_{m+1}, ..., C_{m+r}\}$. Our indexing convention reserves the first m indices for tests and the last r indices for treatments, which allow a uniform notation for both tests and treatments. Like tests, treatments are representable by subsets of U, but of course the meaning is quite different. If a treatment is applied then the unknown object is considered (completely) treated if it is in the treatment set, and not treated (or otherwise altered) if it is not in the treatment set. However, in each case the cost of the treatment is incurred. In the decision tree that represents a given test-and-treatment (TTr) procedure there would be only one arc below a node representing a treatment, the arc that represents the continuing path for non-treatment, i.e., when the unknown object is not in the treatment set. The procedure must treat the unknown object, so every branch of the decision tree will end in a treatment.

Our objective is still the same, to find procedures that minimizes the expected cost. Expected cost still is defined by formula (1) used for the binary testing problem, but the notion of path now changes to include treatment nodes.

Figure 1 is an example of a test-and-treatment problem presentation with two TTr procedures presented. Although we believe that Procedure 2 is optimal, the computation to establish that is sufficiently time consuming that optimality has not been proven. The decision trees have been stylized for easier reading. Although a treatment should have only one arc below it, we have added a second arc, with double lines, to record at the end of that arc the objects treated. Technically the objects treated should label the treatment made itself, the convention we follow when the treatment is at the end of a path and all objects associated with that path are treated. In general, a test or treatment labels a node, and an object with its associated *a priori* probability (alternatively, its *weight*) labels the end of a path. The expected cost value for each procedure is also given.

A Sample Test-and-Treatment Problem

Objects	o_1	o_2	o_3	o_4	o_5
Probabilities	.3	.3	.2	.1	.1

Tests/Treatments

	Name	Set	Cost
Tests	T_1	{2,3}	1
	T_2	{2}	1
	T_3	{3,4}	1
Treatments	T_4	{1,4}	4
	T_5	{2,5}	4
	T_6	{2,3}	5
	T_7	{3}	2
	T_8	{1}	1

Procedure 1

$EC = 1(.3) + 6(.3) + ...$
$\quad = 4.3$

Procedure 2

$EC = 2(.3) + 6(.3) + 4(.2) + 7(.1) + 7(.1)$
$\quad = 3.6$
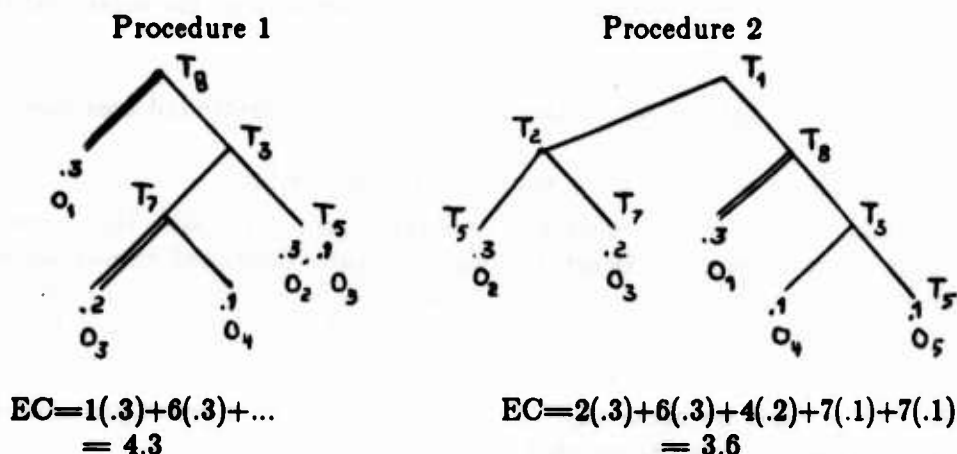
Figure 1

3. Finding Low-Cost Procedures. Using an appropriate dynamic programming or branch-and-bound algorithm, one can determine a minimum cost TTr procedure (decision tree) even with arbitrary test costs. However, such algorithms take at least $m\,2^n$ steps to execute in the general case, where, as before, n is the number of objects and m is the number of tests. This exponential growth in the number of objects means that in practice only small problems can be solved exactly. Thus there has been a great interest in fast algorithms that find low-cost (but not necessarily optimal) test procedures. When algorithms are presented, the obvious questions are: (1) how fast is the algorithm? (2) how close to optimal are the resulting test procedures?

To proceed we need some definitions.

A *complete TTr problem* has a set of tests such that each subset S of U is both a test set and a treatment set. (For tests it actually suffices that either S or $U-S$ be a test, by symmetry for tests.) The *complete testing problem* is a restriction of the complete TTr problem to test

sets.

A complete TTr problem (or testing problem) is an important subcase because it is assured that whenever a test (treatment) is desired, it exists and may be used. The importance of this subcase is documented by some major properties for the binary testing procedure, which we now state. See [2] and [3] for details and further properties.

The following hold for the binary testing problem with unit test costs.

I.   There is a $O(n \log n)$ algorithm to find the optimal test procedure for the (unit cost) complete testing problem.

II.   The incomplete testing problem is NP-hard. (That is, all evidence is that such problems cannot all be solved in $O(n^k)$ steps for any given integer k.)

III. For the incomplete testing problem the most natural fast approximation algorithm has been evaluated regarding its speed (easily seen as $O(mn)$) and its expected cost approximation to optimal. See [2] and [3].

With this background in mind, we undertook the study of the test-and-treatment problem, which as mentioned earlier, seems to be the more correct problem statement for most real-life situations. Our original goals were:

A.   To find a fast optimal algorithm for the complete TTr problem (for a restricted cost case);

B.   For the incomplete TTr problem to find a good approximation algorithm.

After considerable study on the former question we have reason to doubt the interest, perhaps even the feasibility, of our first goal. Besides giving some quite restricted results we will demonstrate why seeking an optimal solution may not be worthwhile except in the restricted case we mention. (We have to date done limited work on the second goal; that work is beyond the scope of this summary.)

By an *equiprobable TTr problem* we mean any *TTr* problem where all *a priori* probabilities have equal value, i.e., $p_i = 1/n$, where n is the number of objects in U.

We now consider briefly a dynamic programming solution to the general TTr problem. For any subset S of U, we define $EC(S)$ by

$$EC(S) = \min\{ \min_{1 \le i \le m} (C_i \cdot |S| + \tag{2}$$

$$EC(S \cap T_i) + EC(S - T_i)),$$

$$\min_{m < i \le m+r} (C_i \cdot |S| + EC(S - T_i))\},$$

where $|S|$ denotes the sum of the weights of the objects in set S, C_i denotes the cost of test or treatment i, $EC|\phi| = 0$, and any term reducing to $EC(S)$ itself on the right is undefined. In general the cost of computing $EC(U)$, the desired answer, is exponential in n because there are 2^n subsets that need consideration. Our colleague Robert Wagner observed that if the partial expected cost $EC(S)$ only depends on the cardinality of S then this minimization can be solved in $O(n^2)$ steps. Basically, this is because one needs to know only $EC(\#S)$ rather than $EC(S)$ for all smaller subsets S. ($\#S$ denotes the cardinality of S). This special case can be realized for the equiprobable complete TTr problem with all test costs the same and treatment costs proportional to the weight of the treatment set, for example. (See [4] for more details. A

method of parallel computation of the general dynamic programming formulation (2) for the TTr problem is presented in [5].)

Before proceeding to discuss an approximation algorithm for a special case we should note that the most obvious special case is not interesting. The complete TTr problem that has all treatments as well as all tests with unit cost is clearly easy to solve: simply invoke the universal treatment (that treatment with treatment set U) so that every object is treated by that one treatment. That gives an expected cost of 1. Clearly, any other procedure is more costly. For uniform treatment costs other than unit cost (the cost of each test) the answer remains the same because it costs as much to treat a subset of U as to treat all of U.

Some experimental work has shown us that we often do quite well in an arbitrary TTr problem (including the incomplete TTr problem case) if we choose the treatment with the lowest cost/power ratio and invoke that treatment, and recurse on that strategyy. By power we simply mean the weight of the treatment set. This simple rule fails if a number of treatments have nearly the same cost/power ratio. The example in Figure 1 has all the treatments in the problem except the last with a cost/power ratio of 10; and this makes the outcome more difficult to ascertain. Treatment T_8 has cost/power ratio less than 4, so by our guideline should be favored over the other treatments. Both the procedure illustrated have T_8 near the top of the decision tree, and the reader may wish to verify that placing other treatments before T_8 yields worse procedures. The example procedure does illustrate that one may want to use tests before invoking the most effective treatment for best expected cost.

Because the single most important determiner of value for treatments seems to be the cost/power ratio, our special case investigations have focused first on the subcase where all treatments have the same cost/power ratio. We hereafter denote that ratio by k. Test costs will be fixed at unit cost. Again, we demand that all tests and treatments be present.

We have determined certain properties of optimal procedures for this special case.

Lemma 1. No non-singleton treatment appears in an optimal procedure for this subcase.

Lemma 2. All tests occur prior to treatments.

We omit all proofs although the proof of Lemma 1 follows very quickly from the nature of expected cost computations in this special subcase. Any multiobject treatment can be replaced by a sequence of singleton treatments for the same objects with lower expected cost, which we call a *cascade*. For example, the single treatment given by treatment set $\{o_1, o_2, o_3\}$ can be replaced by the cascade of Figure 2 with the expected cost then reduced.
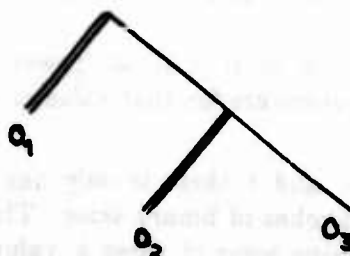
A cascade
Figure 2

We now give an approximation algorithm for a yet more restricted subcase, namely, the subcase under consideration (complete TTr problem, unit cost tests, treatments of cost kw where w is the weight of the treatment set) plus the equiprobable requirement. We noted that for the special case the dynamic programming method allowed computation of the optimal decision tree in $O(n^2)$ steps. The approximation algorithm gives a decision tree in time essentially independent of n and k (constant time). We present a bound on the relative error, a result that is non-trivial, requiring considerable understanding of the nature of low-cost decision trees for this problem. We will close the paper with an attempt to illustrate why a fast optimal algorithm (beyond the special case dynamic programming algorithm) is likely to be complex and not likely to be found soon, if it exists. Also the problem of finding approximation algorithms for less restricted classes must take these concerns into consideration.

The decision tree class to be used for this special case is a simple class. The trees have the following properties:
a) the superleaves of the tree are all at the same level, where each superleaf is a cascade (so the position of the superleaf locates the root of the cascade);
b) the cascades differ by at most one in the number of objects treated;
c) the level of occurrance of the superleaves is level l (the rest is level 0) where l is the least integer z such that

$$k \leq 4 \cdot 2^z.$$

Thus all tests, and only tests, occur above level l with nearly identically formed cascades beginning at level l. The cost/power ratio k determines the transition level.

We shall call the class just defined the class of *level l* procedures. See Figure 3 for examples of level l procedures.

An upper bound that holds for all k is given by

$$EC_l - \frac{opt}{EC_l} < 1/l \tag{3}$$

where EC_l is the expected cost of the above mentioned decision tree and opt is the minimal expected cost possible. Experimentation shows that for small n the approximation is actually much better than the upper bound suggests. The actual relative error does not seem to improve with n for a fixed k, and also varies considerably with k even for small values of k. The upper bound is as weak as it is partly because it represents all values of k. The theorem statement below helps explain the varying relative error for small k.

The relative error result follows from a key theorem that holds for this special case.

Theorem. For every $n > 0$ and every $l > 0$, there is a cost/power ratio value k such that the n-object level l procedure is an optimal procedure for that value of k.

One should note that for each n and l there is only one n-object level l decision tree modulo the left-right orientation of branches of binary trees. The theorem states that this tree is optimal for some k. We can determine some of these k values but the expression is messy. At intermediate k values the approximation seems quite good but is hard to characterize analytically. Thus our relatively modest upper bound.

What is at least as interesting as this upper bound is a characteristic of optimal trees that we can hint at by example. Although there is symmetry to the problem presentation (equiprobable weights, costs uniform for tests and dependence only on the number of objects in treatments) the optimal tree is not necessarily fully symmetric, due to what we call "migration of objects" from cascade to cascade as k changes. Without going into the specific analysis, we demonstrate the effect in Figure 3 where we present three decision trees for a specific TTr problem.

For Figure 3 we choose $n = 64$ and $k = 16$, which by our formula for determining the level l, $l = \min z (k \leq 4 \cdot 2^z)$, puts $l = 2$ by virtue of the equal sign; had $k = 16.001$ then $l = 3$ would be needed. That is reflected by the same expected cost value for the level 2 tree and the level 3 tree. (The circle with enclosed number represents the number of elements in a cascade; we chose n so that all cascades are equally populated to remove the "excess objects" effect. The branching above the superleaves represents tests that split the relevant sets of objects exactly in half.)

Level l procedures are not always optimal

Example: $n = 64$ $k = 16$

level 2 procedure: 16 obj/cascade

$$EC_2 = \frac{1}{64}[l \cdot 64 + \frac{k}{64}(4 \cdot \frac{s(s+1)}{2})]$$
$$= \frac{1}{64}[2 \cdot 64 + \frac{1}{4}(4 \cdot 136)] = \frac{1}{64}(264)$$
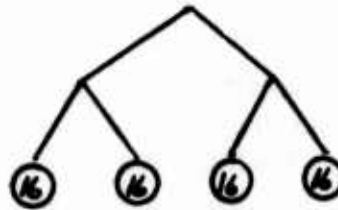
level 3 procedure: 8 obj/cascade

$$EC_3 = \frac{1}{64}[3 \cdot 64 + \frac{1}{4}(8 \cdot 36)] = \frac{1}{64}(264)$$
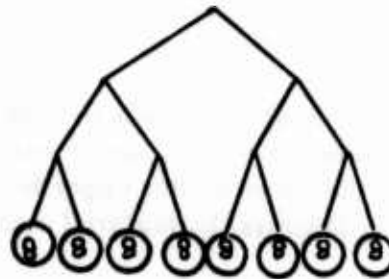
level 2,3 tree:

$$EC_{2,3} = \frac{1}{64}[2 \cdot 28 + 3 \cdot 36$$
$$+ \frac{16}{64}(4 \cdot 45 + 2 \cdot 105)]$$
$$= \frac{1}{64}[56 + 108 + 45 + 52.5]$$
$$= \frac{1}{64}[261.5]$$

Figure 3

737
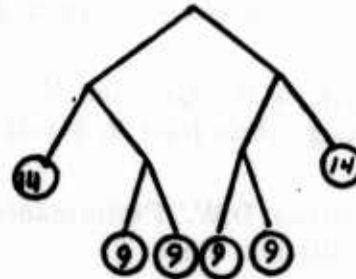
The third tree of Figure 3 has cascades with roots at levels 2 and 3 so is not a level l procedure. Yet its expected cost is lower than the two level l trees. One might have expected, even hoped, at the transition value of k (viewing k as increasing) where a level 2 tree passes to a level 3 tree, that an even splitting of each cascade from 16 members to two cascades of 8 members each would yield optimal trees. We see by example that instead the 8-member cascades absorb an extra member and allow a 14-member cascade which is not "big enough" to split at that k value. This is what we term "migration". This migration can be characterized but the resulting algebra makes computation very messy and the determination of an algorithm for finding optimal trees very unpleasant, if doable, even for this simple case. It is better to use the dynamic programming formulation in this special case if optimal trees are needed, and in general we surely will settle for approximate solutions, even in the complete TTr problem case. (Recall that the incomplete TTr problem is NP-hard anyway, since the simpler incomplete testing problem is NP-hard.)

We now state briefly how the terms in the expected cost are computed in Figure 3. The first line in the computation of EC_l, the expected cost for the level 2 procedure outlines the computation symbolically. We do not sum the terms $Path_i \cdot p_i$ directly but aggregate components. First we factor out the common weight p_i (i.e. 1/64) so we need only determine $Path_i$. The tests, of unit cost, cost l units for each of the 64 objects. The cascade is composed of unit treatment costs of k/64, and there is one less object subject to each sequential treatment so for s objects in a cascade there are $\sum_{i=1}^{s} i$ object-treatments. (Compare with $(ks/64) \cdot s$ for a single treatment for all s objects.)

We presently have a model for reasonable approximate procedures in the arbitrary weight case but no upper bound or relative error yet. Such models may serve as well for the incomplete case. Finding bounds on their relative error is another matter however.

The integrated theory of test-and-treatment procedure design is clearly of interest and it is our hope eventually to better understand how to find, with reasonable effort, good low-cost procedures for accomplishing this task. We are hopeful that at least for the complete TTr we can do well for the cost structure outlined here.

References

[1] Payne, R.W. and D.A. Pierce. Identification keys and diagnostic tables: a review. *J. Royal Stat. Soc. (Series A) 143*, 253-292, 1980.

[2] Garey, M.R., Graham, R.L. Performance bounds on the splitting algorithm for binary testing. *Acta Infor 3*, 347-355, 1976.

[3] Loveland, D.W. Performance bounds for binary testing with arbitrary weights. *Acta Infor 22*, 101-114, 1985.

[4] Wagner, R.A. A polynomial time algorithm for the complete test-and-treatment decision tree problem. C.S. Report, Comp. Sci. Dept., Duke University, 1986.

[5] Duval, L.D., R.A. Wagner, Y. Han and D.W. Loveland. Finding test-and-treatment procedures using parallel computation. *Proc. of the 1986 Int'l Conf. on Parallel Processing*, St. Charles IL, August, 1986.

ON THE ERRORS THAT LEARNING MACHINES WILL MAKE

A.W. Biermann, K.C. Gilbert, A. Fahmy, B. Koster
Department of Computer Science
Duke University
Durham, N. C. 27706

A learning model has been studied where binary function f of p binary variables $x_1, x_2, ..., x_p$ is to be learned from example input-output behaviors. At each time $t = 1,2,3,...$, the learning machine receives a sample input-output pair for the target function and guesses what that target function is.

Most learning machines are capable of guessing (or learning) only L functions where L is less than the set of all possible functions 2^{2^p}. There are advantages to having L large: more functions can be learned, and in general, when the target function is not precisely learnable there will be a learnable function not too distant from that target function. Thus the learning machine will be able to choose a function which agrees with the target function on most inputs even though it will be in error on some inputs. There are also advantages to having L small if the problems with error are not too severe: learning can occur much more quickly if there are fewer learnable functions to choose from. This paper is concerned with a number of different learning machines, their associated values for L and the nature of the trade-off between having large L and little expected error versus having small L and short expected learning time.

For example, the signature table learning model of Arthur Samuel was studied. A characterization of the class of learnable functions was found which gives insight into how the mechanism works and what types of functions it can acquire. The characterization specifies the form that certain matrices of function values must have in order for the function to be realized. This leads to a methodology for computing L and estimating the expected error that signature table systems will have in attempting to learn a randomly selected target function from the set of all possible functions.

Other learning models have similarly been studied such as the linear evaluation systems, the Boolean conjunctive normal form learning methodology of Valiant, and "truncation machines" which simply memorize the outputs with the assumption that they are determined by a specified subset of the inputs.

The L learnable functions for a given machine may be widely spread across the space of all possible functions so that every possible function is near, using Hamming distance as a measure, some learnable function. They may also be very poorly scattered so that some possible target functions are very far from any learnable function. In order to gather information regarding the quality of these learnable function distributions, a new learning machine was invented, the "G-machine", which spreads its learnable behaviors in a near optimal fashion. The G-machine thus can learn with very low expected error for a given value of L and serves as a standard for comparison with other learning machines. The general result in some simulations was that most learning machines achieved expected errors which were surprisingly close to the best known values.

A MODEL OF DECISION MAKING
WITH SEQUENTIAL INFORMATION-ACQUISITION
WITH APPLICATION TO THE FILE SEARCH PROBLEM

James C. Moore, William Richmond and Andrew B. Whinston
Department of Computer Science
and
Management Information Research Center
Krannert Graduate School of Management
Purdue University
West Lafayette, IN 47907

ABSTRACT. We present the file search problem in a Decision Theoretic Framework, and discuss a variation of it which we call the common index problem. The goal of the common index problem is to return the best available record in the file where "best" is in terms of a class of user preferences. We use dynamic programming to construct an optimal algorithm using two different optimality criteria, and we develop sufficient conditions for obtaining complete information.

1. Introduction

Many areas of computer science are benefiting from the application of economic decision theory. The literature in data base design [Mendelson and Saharia, 1986] and expert systems [Hall et al., 1985] is starting to draw on theories of rational decision making. Mendelson and Saharia use a decision theoretic framework for defining an optimal data base design. They begin by defining a minimum cost answer to an information request, and then they balance the trade-offs between incomplete information costs and data-related costs over the set of possible queries.

Hall, Moore and Whinston [1985] develop a decision model that characterizes the ideal (economically rational) behavior of an expert and present the model as a theoretical basis for expert systems. The decision model combines sequential information acquisition with classical decision theory. Decision making is viewed as a dynamic process where knowledge is obtained by a sequence of actions followed by a final decision. The information gathering actions taken are determined by the trade-off between the cost of collecting more information and the payoff from a better decision. The construction of an expert system is then shown to correspond to determining a feasible information gathering and decision strategy.

Moore and Whinston [1986] expand the decision model developed by Hall et al., and show that the file search problem can be interpreted as a special case of a class of decision problems called the categorization problem. They show that there are essentially two mathematically equivalent ways of looking at file search - from a decision theoretic perspective and from a computer science viewpoint. The decision theoretic approach treats file search as a decision problem with a payoff based on the value of a correct answer and the cost of retrieval, and then develops an algorithm that maximizes

the expected return. The computer science approach views the file search problem as finding the algorithm that minimizes the expected cost (search time). Under certain criteria, the two approaches are shown to be equivalent, and the decision theoretic approach is used to construct an optimal algorithm that is equivalent to the optimal binary search tree. In this paper, we continue to use the decision model developed by Moore and Whinston, and the associated information economics terminology. Since this terminology is not universally understood, we define three of the key terms. The idea of an information and decision strategy corresponds to an algorithm in computer science; an individual utility function represents a person's preferences over a given set; and a social welfare function is an aggregate of the individual utility functions and represents the well-being of the users as a whole. At the end of this paper there is a Glossary of our notation.

We generalize the file search using the decision theoretic approach [Moore and Whinston, 1986] of treating the selection of an optimal algorithm as an optimization problem. The decision theoretic approach provides several advantages over the typical approach used in computer science. Instead of constructing an algorithm and then evaluating its complexity, we are able to develop an algorithm that we know to be optimal with respect to predefined criteria. In addition, these optimality criteria are not limited to those usually used in computer science (worst case time and expected time), but include functions of both the value of the information and the cost of the retrieval. Thirdly, if there is a binary relation, ≥, on the set of experiments that defines a total ordering over the experiments and induces a binary or trinary information structure on the state space then we can use dynamic programming to calculate optimal algorithms. These conditions are general enough that we expect dynamic programming can be used to create optimal algorithms for a broad range of decision problems. Finally, by using the decision theoretic framework we are not restricted to obtaining complete information, and therefore finding the requested element in the file. In general, if the file is very large, it will not be optimal to find the best record in the file. A person concerned with finding "the best job" in the U.S. for him/her would generally not find it worthwhile to search the whole file, even if it were available. At some point the expected cost of further search would outweigh the expected benefits of further search even if the only costs involved were the time required to continue the search.

We use the decision theoretic framework to study a generalization of the file search problem that we term the common index problem. To date, the primary role for file search and query processing has been to aid a decision maker by supplying a specific record or request. The user specifies some specific record and if the record exists, it is returned. In many decision problems, however, the need is not for a specific record, but for the best record available. We propose that for these goal oriented queries, the choice process should be embedded within the query system [in a sense, making the query processor a generalized expert system]. The common index problem is defined as: Given a random individual, find the element in a set that maximizes the individual's utility, where the utility function is a function of the index. We give a formal definition and discussion of the common index problem in Section 3.

Motro [1986] has taken a step in the direction of goal-oriented queries by proposing a distance measure for data base queries to determine the best available records. If the ideal record for a query is not in the data base, then the "closest" available record is

returned. In Motro's formulation, however, a record will be returned only if it is within some minimal distance of the ideal record, so null responses to a query are possible. The approach that Motro takes is ad hoc; presenting no theoretical basis for the use of a distance function, nor presenting a method for generating an optimal search algorithm.

In this paper, we confine ourselves to finding the best available record in a file. We restrict ourselves to searching a file in part to investigate the effect of the optimality criteria on the information and decision strategy. We are assuming that the search process will be repeated many times, and that our goal is to provide an algorithm for searching the file that will maximize a social welfare function. Although computer science has generally limited itself to analyzing an algorithm's worst case time and occasionally an algorithm's expected time, we shall analyze the algorithms using two different payoff functions.

In Section 2 of the paper we present a formal characterization of the general decision model. In Section 3 we define the common index problem, develop a dynamic programming solution for our characterization of it and work out two examples, one with each optimality criteria.

2. Decision Model

Our decision problem is defined by eight elements:
$$D = <X, \phi, D, \omega^*, A, \{M_a | a \in A\}, c, r>,$$
where:

X = the set of possible (mutually exclusive) states. We use the generic notation "x" to denote elements of X.

$\phi: X \rightarrow [0,1]$ the probability density function. ϕ defines the probability distribution function $\pi: P(X) \rightarrow [0,1]$ by:
$$\pi(Y) = \sum_{x \in Y} \phi(x) \quad \text{for } Y \subseteq X,$$
where "P(X)" denotes the power set of X.

D = the set of available (final) decisions.

$\omega: X \times D \rightarrow \mathbb{R}$ is the gross payoff function.

A is the set of "initial" (information-gathering) actions, or experiments, available.

M_a is the information structure associated with action $a \in A$. (Each M_a is a partition of X, as will be explained in more detail below.)

$c: A \rightarrow \mathbb{R}_+$ is the cost function; $c(a)$ is the cost of utilizing action $a \in A$.

r is a positive integer representing the number of information-gathering actions which can be taken before a final decision is made.

<u>Assumptions:</u> X, D, and A are all finite, and:
$$(\forall x \in X): \phi(x) > 0.$$
In particular, we shall assume that A has $n+1$ elements, where $n \geq 1$, and write
$$A = \{0,1,...,n\}.$$

The decision-maker is assumed to have a finite set of feasible (final) decisions, D, and to receive a (net) payoff which depends upon the state of the environment, $x \in X$, the decision chosen, $d \in D$, and the cost of information-gathering, c. One may suppose (see, e.g., Marschak and Radner [1972]), that there is a deterministic relationship between decisions, states of the environment, costs and a set of outcomes (or effects), E, such that there exists an outcome function, $\rho(x,d)$ mapping the set $X \times D$ into the set of outcomes. If the decision-maker's preferences over the outcomes may be represented by a real valued utility function, $u(e,d)$, for all $e \in E$ and $d \in D$, then the (gross) payoff function may be defined by

$$\omega(x,d) = u[\rho(x,d), d]. \tag{1}$$

For the remainder of our discussion we will take the payoff function, $\omega(\cdot)$, as given, and will identify the net payoff of a strategy with the difference between the gross payoff obtained and the cost of the information-gathering actions undertaken.[1]

The remaining elements of our decision problem revolve around the construction of an information structure, and the costs of obtaining information. The result of an information acquisition strategy, α, is a partition,

$$\mathbf{B} = \{B_1,...,B_q\}$$

on X such that $B_i \cap B_j = \emptyset$ for $i \neq j$, and $\bigcup_{i=1}^{q} B_i = X$. With each set $B \in \mathbf{B}$, there will be associated a cost of information-gathering, $C(B)$. Thus if the decision-maker follows the decision function $\delta: \mathbf{B} \rightarrow D$, the expected net payoff for the joint strategy $(\alpha, \mathbf{B}, \delta)$ will be given by:

$$\Omega^*(\alpha,\mathbf{B},\delta) = \sum_{B \in \mathbf{B}} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{B \in \mathbf{B}} \pi(B)C(B) \tag{2}$$

In a summary statement, we can roughly describe the goal of the decision problem being analyzed as:

Choose an information strategy α and a decision function $\delta: \mathbf{B} \rightarrow D$ in such a way as to maximize (2) over all α' and $\delta': \mathbf{B} \rightarrow D$.

Associated with each $a \in A$ is a set of information signals, Y_a, and a function $\eta_a: X \rightarrow Y_a$. We shall assume that each Y_a contains a finite number, $n(a)$, of different signals, so that, without loss of generality, we can write:

$$Y_a = \{1,2,...,n(a)\}.$$

We shall also assume that:

i.   for each $a \in A$, η_a is onto Y_a, and

ii. $n(0) = 1$ (so that the $a = 0$ action is the null information action).

For a given element of the set of states, $x \in X$, there is a single signal receivable from each of the n information signal sets. We shall only consider the case where information is obtained deterministically ("noiseless information"); however, it can be shown that noisy information can be incorporated within the present model by including the signals as a part of the specification of the state space. (See Marschak and

---

1. For further discussion of these points, see e.g. Marschak and Radner [1972], pp. 41-44, or DeGroot [1970], pp. 86-115.

Radner [1972].)

We define
$$M_{ay} = \{x \in X | \eta_a(x) = y\} = \eta_a^{-1}(\{y\}) \quad \text{for } a = 0,1,...,n; \; y = 1,...,n(a);$$

and
$$M_a = \{M_{a1},...,M_{a,n(a)}\} \quad \text{for } a = 0,1,...,n.$$

2.1. Definition. Let $B \subseteq X$ be non-empty. We shall say that a family of subsets of X, **B**, is an information structure on **B** iff:

i. **B** is a partition of B (that is, the sets in **B** are pairwise disjoint, and their union equals B).

ii. $(\forall \, B' \in B): B' \neq \varnothing$.

Notice, that for $a \in A$, M_a is an information structure on X (by Definition 1). We shall refer to M_a as the information structure associated with (or induced by) a.

2.2. Definition. Let $B \subseteq X$ be non-empty, and let $a \in A$. We define the information structure induced on **B** by **a**, $\iota(B,a)$, as:

$$\iota(B,a) = \{B \cap M_{a1}, B \cap M_{a2},..., B \cap M_{a,n(a)}\} \setminus \{\varnothing\}.$$

Notice that if $B \subseteq X$ is non-empty, and $a \in A$, then $\iota(B,a)$ is an information structure on B.

Assumption: The decision-maker can take up to r information-gathering actions, where $1 \leq r \leq n$. Since we include the null information action in A (and its associated cost will be assumed to be zero), we can, without loss of generality, assume that the decision-maker takes exactly r information-gathering actions. We also assume that there are no duplicate information structures, i.e.,
$$(\forall \, a, a' \in A): M_a = M_{a'} \Rightarrow a = a'.$$

2.3. Definition. A feasible strategy for **D**, σ, is a sequence of r+1 pairs:
$$\sigma = \langle (B_1,\alpha_1),(B_2,\alpha_2),...,(B_r,\alpha_r),(B_{r+1},\delta) \rangle$$
satisfying:

1. $B_1 = \{X\}$

2. a. $\alpha_t : B_t \to A$ for $t = 1,2,...,r$.

   b. $B_{t+1} = R(B_t,\alpha_t)$ for $t = 1,2,...,r$.

3. $\delta: B_{r+1} \to D$.

We shall denote the set of all feasible strategies for D by "$\sum(D)$".

We shall often find it convenient to regard a feasible strategy,
$\sigma = \langle (B_1,\alpha_1),...,(B_r,\alpha_r),(B_{r+1},\delta) \rangle$ as being composed of two parts:

i. the information-gathering strategy:

$$\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>,$$

ii.   the <u>decision</u> <u>strategy,</u> (B_{r+1},δ).

Accordingly, we define the following:

    <u>2.4. Definition.</u> A <u>feasible information-gathering strategy for D,</u> α is a sequence of r pairs:

$$\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>$$

satisfying 1 and 2 of Definition 2.3; and a <u>feasible decision strategy for D</u> is a pair (B,δ), where

1.   there exists a feasible information-gathering strategy for **D**,
$$\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>$$
    such that
$$R(B_r,\alpha_r) \geq B,$$

2. $\delta: B \rightarrow D$.

2.1. Costs and Payoffs of Strategies

    The following two definitions (and the preceding results) will enable us to provide a convenient characterization of the expected cost of a feasible strategy.

    <u>2.2.1. Definition.</u> Let $\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>$ be a feasible information-gathering strategy for **D**, and let $B_{r+1} = R(B_r,\alpha_r)$. For each $q \in \{1,...,r+1\}$, and each $B \in B_q$, we define the sequence $<\beta_t(B)>_{t=1}^{q}$ by:

$$\beta_t(B) = \text{that } B' \in B_t \text{ such that } B \cap B' \neq \emptyset. \tag{18}$$

(It is shown in Moore and Whinston [1986] that $\beta_t(B)>_{t=1}^{q}$ is well-defined.) We shall refer to $\beta_t(B)$ as the <u>predecessor</u> of **B** at <u>t</u>.

    <u>2.1.2. Definition.</u> Let $\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>$ be a feasible information-gathering strategy for **D**, and let $B_{r+1} = R(B_r,\alpha_r)$. For each $q \in \{1,...,r+1\}$, and each $B \in B_q$, we define a(B) as the sequence (of length q-1) of actions taken by the strategy α along the path that yields B; that is,

$$a(B) = <a(1,B),...,a(q-1,B)>,$$

where we define [2]

$$a(t,B) = \alpha_t[\beta_t(B)] \text{ for } t = 1, ...,q-1 .$$

_____

2. That is, a(t,B) is the action taken at step t (t = 1, ...,q-1) along the path that yields B.

Assumption: We suppose that, with each $a \in A$ is associated a nonnegative cost, $c(a)$, the cost of employing action a. Further, we assume that $c(0) = 0$.

In a given realization of the type of decision problem under study, the application of a feasible information-gathering strategy,

$$\alpha = <(B_1,\alpha_1),...,(B_r,\alpha_r)>,$$

will result in the determination that $\hat{x}$, the true state, is an element of some $B \in B_{r+1} \equiv R(B_r,\alpha_r)$. The cost of determining that $\hat{x} \in B$ will be the sum of the costs of all the actions taken along the path yielding (ending in) B, and will therefore be given by:

$$C(B) = \sum_{t=1}^{r} c[a(t,B)]. \tag{1}$$

The expected cost of the information-gathering strategy, α, will then be given by

$$\gamma(\alpha) = \sum_{B \in B_{r+1}} \pi(B)C(B); \tag{2}$$

and of a feasible strategy, $\sigma = <\alpha,B_{r+1},\delta>$ will be given by

$$\Gamma(\sigma) = \gamma(\alpha) = \sum_{B \in B_{r+1}} \pi(B)C(B). \tag{3}$$

As noted earlier, the function $\omega:X \times D \to \mathbb{R}$ yields the (gross) payoff associated with a given state $x \in X$, given a final decision $d \in D$. Thus if $\sigma = <\alpha,B_{r+1},\delta>$ is a feasible strategy for D, its expected gross payoff is given by

$$\Omega(\sigma) = \sum_{B \in B_{r+1}} \sum_{x \in B} \phi(x)\omega[x,\delta(B)]; \tag{4}$$

and the expected net payoff of the strategy will be given by

$$\Omega^*(\sigma) = \sum_{B \in B_{r+1}} [\sum_{x \in B} \phi(x)\omega(x,\delta[B]) - \pi(B)C(B)] \tag{5}$$

$$= \sum_{B \in B_{r+1}} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{B \in B_{r+1}} \pi(B)C(B)$$

$$= \Omega(\sigma) - \Gamma(\sigma).$$

We can now more formally state the goal of our decision problem as:

$$\text{choose } \sigma^* \in \Sigma(D) \text{ such that for all } \sigma \in (D); \tag{6}$$
$$\Omega(\sigma^*) - \Gamma(\sigma^*) \geq \Omega(\sigma) - \Gamma(\sigma).$$

The following definitions and results will be useful in our investigations in Section 3.

2.1.3. Definition. We define the finest information structure obtainable from $\underline{A}$, $\underline{B^A}$, by:

$$B^A = \{\bigcap_{a=1}^{n} M_{a1}, \bigcap_{a=1}^{n-1} M_{a1} \cap M_{n2}, ..., \bigcap_{a=1}^{n-1} M_{a1} \cap M_{n,n(a))}$$

$$\bigcap_{a=1}^{n-2} M_{a1} \cap M_{n-1,2} \cap M_{n1}, ..., \bigcap_{a=1}^{n} M_{a,n(a)}\} \setminus \{\varnothing\}.$$

It can be shown that $B^A \geq B_{r+1}$, for any feasible final information structure, B_{r+1}; which justifies our terminology in the above definition. This gives us a simple necessary condition which must be satisfied by any feasible final information structure.

Notice that in a given realization of the decision model, B^A represents the best information that can be obtained even if $r \geq n$. (Thus, for example, the finite memory of even the largest computer available puts an upper limit on the number of decimal places one can use in representing a real number.) From the standpoint of the decision problem, therefore, any data concerning individual elements of members of B^A is, in a sense, irrelevant to the decision at hand.

2.1.4. Definitions. If $B \subseteq X$ is non-empty, we define the potential gross payoff associated with B, $v(B)$, and the conditionally optimal decision set for B, $D^*(B)$, by

$$v(B) = \underset{d \in D}{\text{Max}} \sum_{x \in B} \phi(x|B)\omega(x,d), \tag{11}$$

and

$$D^*(B) = \{d \in D \mid \sum_{x \in B} \phi(x|B)\omega(x,d) = v(B)\}, \tag{12}$$

respectively.

Notice that we can equally well define $D^*(B)$, the conditionally optimal set for B, as:

$$D^*(B) = \{d \in D \mid \sum_{x \in B} \phi(x)\omega(x,d) = \pi(B)v(B)\} \tag{13}$$

Given this consideration, the following result becomes more or less immediate.

2.1.5. Proposition. If $\sigma = <\alpha, B_{r+1}, \delta>$ is optimal for D, then for each $B \in B_{r+1}$ we must have $\delta(B) \in D^*(B)$. Furthermore, the expected gross payoff for σ, $\Omega(\sigma)$, will be given by:

748

$$\Omega(\sigma) = \sum_{B \in B_{i+1}} \pi(B)\nu(B).$$

3. Common Index Problem

The common index problem was briefly described in Section 1 as a generalization of the file search problem. We shall formalize the common index problem using the decision theoretic framework just presented, but we first present the formalization of the file search problem in both the decision theoretic framework and the computer science framework.

Aho, Hopcroft and Ullman present the file search problem as:

> ... we were given a set $S = \{a_1, a_2, ..., a_n\}$, that is, a subset of some large universal set U [which is linearly ordered by a relation $\leq$], and we were asked to design a data structure that would allow us to process efficiently a sequence σ consisting only of MEMBER instructions. Let us reconsider this problem, but this time let us assume that, in addition to being given the set S, we are given the probability that the instruction MEMBER(a,S) will appear in σ for all elements a in the universal set U. We would now like to design a binary search tree for S such that a sequence σ of MEMBER instructions can be processed on-line with the smallest expected number of comparisons. [Aho, Hopcroft and Ullman, 1974, p. 119].

3.1. The Computer File Search Problem - Decision Theoretic Framework

This section presents the decision theoretic approach to file search developed in Moore and Whinston [1987, Sec. 4.2].

We suppose that there is some universal set, U, which is finite and linearly ordered, and that we are dealing with a non-empty subset,

$$S = \{b_1, b_2, ..., b_n\} \subseteq U, \tag{1}$$

with

$$b_i < b_{i+1} \quad \text{for} \quad i = 1, ..., n-1. \tag{2}$$

We suppose that there is a probability measure on U, so that the probabilities

$$Pr(b < b_1), \quad Pr(b > b_n),$$

$$Pr(b_i < b < b_{i+1}) \quad \text{for} \quad i = 1, ..., n-1,$$

and

$$Pr(b = b_i) \quad \text{for} \quad i = 1, ..., n,$$

are well-defined, for b a random element of U.

The basic idea is that the elements of S correspond to a stored data set drawn from U. We consider the problem of searching the set in order to determine whether a

randomly drawn element from U, b, is in the set S or not; and if it is in S, to determine its location (i.e., for which i we have $b=b_i$). The available experiments can be denoted by:

$$A = \{0,1,...,n\},$$

where for $a = 1,...,n$, the experiment a is interpreted as:

"compare b with b_a".

(and a=0 represents the null information experiment). Thus, for $a \in A_1$, the possible outcomes of the experiment are:

$$b<b_a, \quad b=b_a, \quad \text{or} \quad b>b_a. \tag{3}$$

For notational convenience, we shall represent the state space as:

$$X = Y \cup Z,$$

where

$$Y = \{y_1,...,y_n\},$$

with the interpretation:

$$x = y_i \iff b = b_i \quad \text{for} \quad i = 1,...,n;$$

and

$$Z = \{z_0,z_1,...,z_n\},$$

with the interpretation:

$$x = z_j \iff \begin{cases} b<b_1 & \text{if } j = 0 \\ b_j < b < b_{j+1} & \text{for } j = 1,...,n-1 \\ b > b_n & \text{for } j = n. \end{cases}$$

We then define

$$p_i = \Pr(x = y_i) = \Pr(b = b_i) \quad \text{for} \quad i = 1,...,n,$$

$$q_0 = \Pr(x = z_0) = \Pr(b < b_1),$$

$$q_j = \Pr(x = z_j) = \Pr(b_j < b < b_{j+1}) \quad \text{for} \quad j = 1,...,n-1,$$

and

$$q_n = \Pr(x = z_n) = \Pr(b > b_n).$$

From (3) and our specification of X, we see that for each $a \in A_1$, the information structure for a, M_a, can be written as:

$$M_a = \{M_{a1},M_{a2},M_{a3}\},$$

where

750

$$M_{a1} = \{y_1,...,y_{a-1}\} \cup \{z_0,...,z_{a-1}\},$$

$$M_{a2} = \{y_a\},$$

and

$$M_{a3} = \{y_{a+1},...,y_n\} \cup \{z_a,...,z_n\}.$$

To complete our specification of the problem, we note that we can specify **D** as

$$\mathbf{D} = \{0,1,...,n\},$$

with the interpretation:

$d = 0$ corresponds to the decision $b \notin S$ (i.e., $x \in Z$),

$d = j$ corresponds to the decision $b = b_j$ (i.e., $x = y$) for $j = 1,...,n$.

It also seems appropriate here to specify our gross payoff function $\omega: X \times D \to \mathbb{R}$ as

$$\omega(x,d) = \begin{cases} \bar{\omega} > 0 & \text{if } d=0 \text{ and } x \in Z, \text{ or if } d \in \{1,...,n\} \text{ and } x=y_d, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

The user is thus indifferent to whether the requested record is in a gap or an element of the file, as long as he/she is correctly informed of the status of the requested record.

We also suppose that there exists some constant $c > 0$ such that:

$$c(a) = c \quad \text{for} \quad a = 1,...,n.$$

3.2. The Common Index Problem - Decision Theoretic Framework

In the common index problem, a random individual wants to determine if a given element, the ideal element, from a universal set is present in a known subset (the file). In the common index problem, however, if the ideal element is not in the file then the best available element is to be returned. To determine the best available element in the file we must make a couple of assumptions. The first assumption is that the index associated with each record has an intrinsic meaning. As an example, if the index for a file containing data on a set of cars is the price of the car, then the index has a universally agreed upon meaning. The second assumption necessary to determine a best available record is that there exists an individual utility function associated with each potential user and that the utility function is a function of the index. The utility functions measure the individuals' preferences over the elements in the file, and can be interpreted as measuring the amount or quality of information in a record, or merely the desirability of the object that the record represents. We are assuming that all of the relevant information for choosing between records is contained in the index. In a future paper, we plan to extend the ideas developed here to multiattribute data base queries where the utility functions are functions of the attributes.

As an example of the common index problem, consider a person who wants to vote in an upcoming election. The only information available about the candidates is a rating from 0 to 20 where 0 implies the candidate is left-wing and 20 implies the candidate is right-wing. The election is for the U.S. economic advisor, and the set of

751

possible candidates is the set of all people, both historical and modern. As of election day, the set of candidates include: Karl Marx, Fredrick Engels, John Maynard Keynes, J.R. Hicks and Adam Smith. An omnipotent panel has assigned index values to each of the candidates as shown in Table 1. The prospective voter must find the one candidate that maximizes his/her utility by searching the index. The index value returned will correspond to the best available candidate the he/she can vote for. The search process is similar to the computer file search, except that the user is not looking for a particular record, but for the best record available.

Table 1

Candidate	index value
Karl Marx	1
Fredrick Engels	2
John Maynard Keynes	6
J.R. Hicks	19.5
Adam Smith	20

To formalize the common index problem, we assume that there exists some universal set T. The set S is the set of available alternatives from which a best element can be chosen. $S = \{s_1,...,s_m\}$ is a subset of T, and is assumed to be non-empty.

Define W as the common index on T where $W = [0,\overline{w}] \in \mathbb{R}_+$. The function γ, $\gamma:T \to W$, assigns an index value to each element in the set T. γ may or may not be known, but we assume $\gamma(s)$ is known for all $s \in S$. We define $\hat{W} = \{w \in W:(\exists s \in S):w = \gamma(s),\}$. $\hat{W} = \{\hat{w}_1,...,\hat{w}_n\}$, is the set of indices of the available elements $s \in S$, and it is over $\hat{W}$ that the search is conducted.

The set of potential users is designated by U, and each user is identified by his/her utility function $u \in U$. The set U does not necessarily correspond to the set of people who may search the file, but rather to the set of possible requests they may make. Thus, if a person searches the file multiple times, each time with a different set of preferences over the index, then the person is considered to be multiple users with different utility functions.

The probability distribution, $h(\cdot)$, is a distribution over U and is assumed to be known. $h(u)$ denotes the probability that user $u \in U$ will query the system.

$$h:U \to [0,1]$$
$$u:W \to \mathbb{R}; \quad \forall u \in U.$$

To make the problem more tractable we will restrict the set of utility functions to single-peaked, symmetric functions that are strictly decreasing away from the maximum. Specifically we require that for each $u \in U$

$$(\exists v_u \in W \text{ and } \hat{u}\mathbb{R}_+ \to \mathbb{R}:u(w) = \hat{u}(|w - v_u|))$$

where $\hat{u}$ is strictly decreasing, and v_u is the ideal element in W for the user represented

by the utility function u. [3]

These restrictions allow the individual utility functions to be completely characterized by v_u. All individual utility functions maximized at v_u will have the same best available element; thus a user need merely specify his/her ideal point without having to express the complete functional form of his/her utility function.

Due to the restriction on the individual utility functions, the best available element will always be the element $\hat{w} \in \hat{W}$ closest to v_u. The problem therefore becomes: given a randomly drawn utility function $u \in U$, find the $k \in \{1,...,n\}$ satisfying $|\hat{w}_k - v_u| \leq |\hat{w}_j - v_u|$; $j = 1,...,n$. We now want to develop a search strategy that will maximize the social welfare of the users, where the social welfare function is defined by:

$$\sum_{B \in B_{t+1}} \sum_{x \in B} \phi(x)\omega(x,\delta(B)) - \sum_{B \in B_{t+1}} \pi(B)C(B)$$

To compute the optimal information structure for the search we must first define the decision problem's state space. We define

$$X = \{\chi_0, \chi_1,...,\chi_n, \chi_{n+1}\}$$

where

$$\chi = \chi_j \leftrightarrow \begin{cases} v_u < 0 & j = 0 \\ v_u \in [v_{j-1},v_j) & j = 1,...,n \\ v_u \in [v_{n-1},\overline{w}] & j = n \\ v_j > \overline{w} & j = n+1 \end{cases}$$

and

$$v_j = \begin{cases} 0 & j = 0 \\ (\hat{w}_j + \hat{w}_{j+1})/2 & j = 1,...,n-1 \\ \overline{w} & j = n \end{cases}$$

The elements of the state space partition W into half open intervals. The endpoints of the elements $\chi \in X$ bisect the intervals between the elements of $\hat{W}$. For example, if $W = [0,10]$ and $\hat{W} = \{3,7\}$, then $X = \{(-\infty,0),[0,5),[5,10],(10,\infty)\}$.

We now define the probability that $\chi \in X$ is the true state (i.e., $v_u \in \chi_i$). First note that the probability measure $h(\cdot)$ induces a probability measure $h^*(\cdot)$ on W, where $h^*(w)$ is the probability that $v_u = w$. The only restriction needed here on $h(\cdot)$ and $h^*(\cdot)$ is that they can be used to generate a probability distribution for the state space, X. If $h(\cdot)$ is a continuous distribution on U, the probability that χ is the true state is $\phi(\chi)$ where:

---

3. This is essentially the assumption basic to Coombs' "unfolding technique". See Coombs [1950] and, for a discussion of empirical tests and extensions, Coombs, Dawes, and Tversky [1970], pp. 55-66.

$$\phi(\chi_j) = \begin{cases} 0 & j \in \{0,n+1\} \\ p_j & j \in \{1,...,n\} \end{cases}$$

and

$$p_j = \int_{v_{j-1}}^{v_j} h^*(w)dw$$

The set of available experiments is defined by $A = \{0,1,...,n\}$ where $a = 0$ is the null experiment and $a = 1, ...,n$ can be interpreted as:

$$\text{compare } v_u \text{ with } v_{a-1} \text{ and } v_a$$

With possible outcomes

$$v_u < v_{a-1} \quad \text{or} \quad v_{a-1} \le v_u < v_a \quad \text{or} \quad v_u \ge v_a .$$

For each $a \in A_1 = A\backslash 0$ the information structure induced by a, M_a, is trinary.

$M_a = (M_{a1}, M_{a2}, M_{a3}\}$ where

$M_{a1} = \{\chi_0,...,\chi_{a-1}\}$

$M_{a2} = \{\chi_a\}$

$M_{a3} = \{\chi_{a+1},...,\chi_{n+1}\}$.

and

$M_{01} = M_{02} = \phi, \quad M_{03} = X$

$M_{n+1,1} = X, \quad M_{n+1,2} = M_{n+1,3} = \phi$

To complete the specification we must define D, c(a), and $\omega(x,d)$. The decision set is $D = \{1,...,n\}$ where $d = j$ corresponds to the decision that $v_u \in \chi_j$, and that the best available element is $\hat{w}_j$.

The cost function is $c(a) = c$ for $a = 1, ...,n$ where c is a strictly positive constant and $c(0) = 0$.

We shall look at two payoff functions:

$$\bar{\omega}(x,d) = \begin{cases} \bar{\omega} > 0 & \text{if } v_u \in \chi_d \\ 0 & \text{otherwise} \end{cases}$$

and

$$\omega(x,d) = -|\hat{w}_u - \hat{w}_d|$$

where

$\hat{w}_d \in \chi_d$ and $\hat{w}_u$ is such that $v_u \in \chi_u$ ($\hat{w}_d$ is the element $\hat{w} \in \hat{W}$ chosen by the search process and $\hat{w}_u$ is the element $\hat{w} \in \hat{W}$ that maximizes $\hat{u}(|v_u - \hat{w}_i|)$ $\hat{w}_i \in \hat{W}$).

The first payoff function identifies the social welfare function with the utility of the designer or operator of the system; the idea basically being that if the programmer (or operator) finds the best element in the file for a given user, then he/she gets a pat on the head, and gets trouble otherwise. The second payoff function identifies the social welfare more closely with the utility of the user, and is used as an approximation to the

754

function $\hat{\omega}(x,v_u,d)$ given by:

$$\hat{\omega}(x,v_u,d) = -(|\hat{w}_d - v_u| - |\hat{w}_u - v_u|) = |\hat{w}_u - v_u| - |\hat{w}_d - v_u|,$$

where, again,

$$\hat{w}_u \text{ is that } w_j \text{ such that } v_u \in \chi_j,$$

and

$$\hat{w}_d \text{ is the element chosen by the search process.}$$

Note that we always have

$$\omega(x,d) \equiv -|\hat{w}_u - \hat{w}_d| \le |\hat{w}_u - v_u| - |\hat{w}_d - v_u| \equiv \hat{\omega}(x,v_u,d).$$

Moreover, if

$$\hat{w}_d \ge \hat{w}_u \ge v_u, \quad \text{or} \quad \hat{w}_d \le \hat{w}_u \le v_u,$$

then

$$\omega(x,d) = \hat{w}(x,v_u,d).$$

If, however,

$$\hat{w}_u \ge v_u > \hat{w}_d \quad \text{or} \quad \hat{w}_d > v_u \ge \hat{w}_u,$$

the two payoff functions will differ. [4] On the other hand, since

$$\hat{\omega}(x,v_u,d) \equiv |\hat{w}_j - v_u| - |\hat{w}_d - v_u|$$

is a function of v_u as well as x, we cannot use it as the payoff function in our problem (that is, its use would necessitate a different specification of the state space). [5]

The dynamic programming solution follows in a manner almost identical to the one developed in Moore and Whinston [1987, Sec. 6.3]. [6] We present it here for

4. Notice that we cannot have $\hat{w}_u > \hat{w}_d \ge v_u$ or $v_u \ge \hat{w}_d > \hat{w}_u$; given the definition of $\hat{w}_u$.
5. If it were the case that for each $u \in U$, v_u were an element of $\hat{W}$, then this difficulty would disappear; i.e., $\omega(x,d)$ would always equal $\omega(x,v_u,d)$. However, this is a condition which we would not expect to be realized in practice.
6. The key condition needed to develop this sort of dynamic programming solution is that the relation $>^*$ defined on A by
$$a >^* a' \leftrightarrow M_{a'1} \cup M_{a'2} \subseteq M_{a1}$$
is a (strict) <u>linear order</u>; that is, that it is total, asymmetric, and transitive. Notice that this condition is satisfied here.

completeness. The purpose is to calculate the information and decision strategy, σ^*, such that $\Omega^*(\sigma^*) \geq \Omega^*(\sigma)$, $\forall\ \sigma \in \Sigma(D)$. We assume $r \geq n$ and we define for $i \in \{0,1,...,n\}$ and $j \in \{i+1,...,n+1\}$, $B_{ij} \subseteq X$ where

$$B_{ij} = M_{j1} \cap M_{i3}\ (= \bigcup_{k=i+1}^{j-1} M_{k2}\ \text{for our given state space})$$

1. For $k \in \{1,2,...,n\}$ we define $\Delta(M_{k2})$ by

$$\Delta(M_{k2}) = \pi(M_{k2})v(M_{k2}),$$

and for $i \in \{0,1,...,n\}$, we define

$$\Delta(B_{i,i+1} = 0.$$

For $i \in \{0,1,...,n-1\}$ and $j \in \{i+2,...,n+1\}$ we define $\Delta(B_{ij})$ by

$$\Delta(B_{ij}) = \pi(B_{ij})v(B_{ij}) = \pi(M_{i+1,2})v(M_{i+2,2}) = \Delta M_{i+2,2}$$

2. For each $i \in \{0,1,...,(n+1)-3\}$ we calculate

$$f(j) = \Delta(B_{i,j}) + \Delta(M_{j2}) + \Delta(B_{j,i+3}) - \pi(B_{i,i+3})C(j);\quad j = i+1,i+2$$

and

$$f(0) = \pi(B_{i,i+3})v(B_{i,i+3})$$

We then define [7]

$$\Delta(B_{i,i+3}) = \text{Max}\{f(0),f(i+1),f(i+2)\}$$

and

$$â(i,i+3) = \text{Min}\{j \in \{0,i+1,i+2\}\ |\ f(j) = \Delta(B_{i,i+3})\}$$

3. For each $i \in \{0,1,...,(n+1)-4\}$ we calculate

$$f(j) = \Delta(B_{ij}) + \Delta(M_{j2}) + \Delta(B_{j,i+4}) - \pi(B_{i,i+4})c(j);\quad j = i+1,i+2,i+3$$

$$f(0) = \pi(B_{i,i+4})v(B_{i,i+4})$$

$$\Delta(B_{i,i+4}) = \text{Max}\{f(0),f(i+1),f(i+2),f(i+3)\}$$

$$â(i,i+4) = \text{Min}\{j \in \{0,i+1,i+2,i+3\}\ |\ f(j) = \Delta(B_{i,i+4})\}$$

4. Having found $\Delta(B_{i,i+(l-1)})$ $i = 0,1,...,n+1-(l-1), l \geq 3$

---

7. Notice that if $C(i+1) = C(i+2)$, then $f(i+1) = f(i+2)$.

We compute for $i \in \{0,1,...,n+1-l\}$

$$f(j) = \Delta(B_{ij}) + \Delta(M_{j2}) + \Delta(B_{j,i+l}) - \pi(B_{i,i+l})c(j) \text{ for } j = i+1,...,i+l-1$$

$$f(0) = \pi(B_{i,i+l})v(B_{i,i+l})$$

$$\Delta(B_{i,i+l}) = \text{Max}\{f(j) \mid j \in \{0,i+1,...,i+l-1\}\}, \text{ and}$$

$$\hat{a}(i,i+l) = \text{Min}\{j \in \{0,i+1,...,i+l-1\} \mid f(j) = \Delta(B_{i,i+l})\}$$

5. Proceeding as above, we eventually obtain
$$\Delta(B_{0,n+1}) = \Delta(X) \text{ and } \hat{a}(0,n+1).$$

We then define the strategy $\sigma^* = <(B^*_1,\alpha^*_1),...,(B^*_r,\alpha^*_r),(B^*_{r+1},\delta^*)>$ by

$$\alpha^*_1(X) = \hat{a}(0,n+1) \equiv a^*. \tag{16}$$

We obtain one of two cases.

    a. $\alpha^*_1(X) = 0$ and $B^*_2 = M_{a^*} = \{X\}$; in which case, we complete the definition of σ^* by defining

$$\alpha^*_t(X) = 0 \quad \text{for } t = 2,...,r, \tag{17}$$

and let

$$d \in D^*(X). \tag{18}$$

    b. $\alpha^*_1 \in \{1,...,n\}$ and
$$B^*_2 = M_{a^*} = \{B_{0a^*}, M_{a^*2}, B_{a^*,n+1}\}.$$

Here we define α^*_2 by

$$\alpha^*_2(B_{0,a^*}) = \hat{a}(0,a^*), \quad \alpha^*_2(M_{a^*2}) = 0, \quad \alpha^*_2(B_{a^*,n+1}) = \hat{a}(a^*,n+1). \tag{19}$$

Having obtained α^*_{t-1} and B^*_t, for $t \in \{3,...,r\}$, we have that each $B \in B^*_t$ is either of the form

$$B = B_{ij} \quad \text{for some } i \in \{0,1,...,n\}, j \in \{i+2,...,n+1\}, \tag{20}$$

or is of the form

$$B = M_{k2} \quad \text{for some } k \in \{1,...,n\}. \tag{21}$$

We then define α^*_t on B^*_t by:

$$\alpha^*_t(B) = \begin{cases} \hat{a}(i,j) & \text{if B is of the form (20) with } j > i+2. \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

757

Proceeding in this fashion we eventually obtain B^*_{r+1}, and let $\delta^*(B)$ be an element of $D^*(B)$, for each $B \in B^*_{r+1}$.

$\square$

It will be an easy consequence of the following result that α^*, as defined in (17)-(22), above, is optimal for D.

3.2.1. Theorem. If $\sigma = <(B_1,\alpha_1),...,(B_r,\alpha_r),(B_{r+1},\delta)>$ is a feasible strategy for D, $q \in \{1,...,r+1\}$, and $i \in \{0,1,...,n\}$ and $j \in \{i+2,...,n+1\}$ are such that $B_{ij} \in B_q$, then [8]

$$\sum_{B \in B_{r+1}(B_{ij})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{t=q}^{r} \sum_{B' \in B_t(B_{ij})} \pi(B')c[\alpha_t(B')] \leq \Delta(B_{ij}), \qquad (23)$$

where for $q = r+1$, we define

$$\sum_{t=q}^{r} \sum_{B' \in B_t(B_{ij})} \pi(B')c[\alpha_t(B')] = 0.$$

Proof. We distinguish two cases, based on the value of q.

a. $q = r+1$. Here the left-hand-side of inequality (23) becomes:

$$\sum_{x \in B_{ij}} \phi(x)\omega[x,\delta(B)]; \qquad (24)$$

and, since (24) is less than or equal to
$$\pi(B_{ij})v(B_{ij}) \leq \Delta(B_{ij}),$$
the desired inequality follows at once.

b. $q \in \{1,...,r\}$. Here we establish our result for arbitrary $i \in \{0,1,...,n-1\}$ by induction on $l = j-i$, as follows.

i. $l = 2$ (and $j = i+2$). Here we have $B_{ij} \in B^\Lambda$ and thus

$$B_{r+1}(B_{ij}) = \{B_{ij}\},$$

so that the left-hand-side of (23) becomes:

$$\sum_{x \in B_{ij}} \phi(x)\omega[x,\delta(B)] - \sum_{t=q}^{r} \pi(B_{ij})c[\alpha_t(B_{ij})]$$

$$\leq \sum_{x \in B_{ij}} \phi(x)\omega[x,\delta(B_{ij})] \leq \pi(B_{ij})v(B_{ij}) = \Delta(B_{ij}).$$

---

8. Notice that if $B \in B_q$, then there must exist $i \in \{0,1,...,n\}$ and $j \in \{i+2,...,n+1\}$ such that $B = B_{ij}$. This is intuitively fairly apparent and can be proved rigorously by an argument essentially identical to that in Moore and Whinston [1987, Corollary 6.2.6.].

758

ii. Suppose the desired inequality holds for $j = i+l$, where $l \in \{2,...,n-i\}$. Then for $j = i+l+1$, we have three possible cases.

Case 1. $\alpha_q(B_{ij}) \equiv k \in \{i+1,...,j-1\}$. Here it follows from Theorem 6.2.5 of Moore and Whinston [1987] that we can write the left-hand-side of (23) as:

$$\sum_{B \in B_{r+1}(B_{ik})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)j - \sum_{t=q+1}^{r} \sum_{B' \in B_t(B_{ik})} \pi(B')c[\alpha_t(B')] \qquad (25)$$

$$+ \sum_{B \in B_{r+1}(B_{kj})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{t=q+1}^{r} \sum_{B' \in B_t(B_{kj})} \pi(B')c[\alpha_t(B')]$$

$$+ \sum_{B \in B_{r+1}(M_{k2})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{t=q+1}^{r} \sum_{B' \in B_t(M_{k2})} \pi(B')c[\alpha_t(B')]$$

$$- \pi(B_{ij})c(k).$$

However, since $k \in \{i+1,...,j-1\}$, and $j = l+i$, it follows that $k-i \leq l$ and $j-k \leq l$. Consequently, it follows from our inductive hypothesis that (25) is less than or equal to

$$\Delta(B_{ik}) + \Delta(B_{kj}) + \Delta(M_{k2}) - \pi(B_{ij})c(k) \leq \Delta(B_{ij}) \qquad (26)$$

Case 2. $\alpha_q(B_{ij}) \equiv k \notin \{i+1,...,j-1\}$ and
$$B_{r+1}(B_{ij}) = \{B_{ij}\}.$$

Here it is immediate that the left-hand side of (23) is less than or equal to
$$\pi(B_{ij})v(B_{ij}) \leq \Delta(B_{ij})$$

Case 3. $\alpha_q(B_{ij}) \equiv k \notin \{i+1,...,j-1\}$ and
$$B_{r+1}(B_{ij}) \neq \{B_{ij}\}.$$

In this case, it follows that for some $t \in \{q+1,...,r\}$ we have

$$\alpha_t(B_{ij}) = k' \in \{i+1,...,j-1\} \qquad (27)$$

and

$$\alpha_s(B_{ij}) \notin \{i+1,...,j-1\} \quad \text{for } s = q,...,t-1; \qquad (28)$$

and thus

$$\sum_{B \in B_{r+1}(B_{ij})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{t=q}^{r} \sum_{B' \in B_t(B_{ij})} \pi(B')c[\alpha_t(B')]$$

759

$$= \sum_{B \in B_{r+1}(B_{ij})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{s=q}^{t-1} \pi(B_{ij})c[\alpha_s(B_{ij})]$$

$$- \sum_{s=t}^{r} \sum_{B' \in B_s(B_{ij})} \pi(B')c[\alpha_s(B')]$$

$$\leq \sum_{B \in B_{r+1}(B_{ij})} \sum_{x \in B} \phi(x)\omega[x,\delta(B)] - \sum_{s=t}^{r} \sum_{B' \in B_s(B_{ij})} \pi(B')c[\alpha_s(B)];$$

and it follows from our analysis of the preceding cases that the right-hand side of (28) is less than or equal to $\Delta(B_{ij})$.

□

3.2.2. <u>Corollary</u>. The strategy $\sigma^* = <(B^*_1,\alpha^*_1),...,(B^*_r,\alpha^*_r),(B^*_{r+1},\delta^*)>$, defined in (17)-(22), above, is optimal for **D**.

<u>Proof</u>. It is an immediate consequence of our definition of σ^* (in particular of our definition of $\alpha^*_1(X)$ and δ^*) that:
$$\Omega(\sigma^*) - \Gamma(\sigma^*) = \Delta(B_{0,n+1}) = \Delta(X).$$
Consequently, it follows from Theorem 3.2.1. that σ^* is optimal for **D**.

□

We now continue the example discussed at the beginning of this section, which illustrates how the payoff function affects the optimal information structure. Since the utility functions can be represented by the index value where they are maximized, we will define $U = W = [0,20]$. Since $U = W$, $h(\cdot) = h^*(\cdot)$, and we define $h^*(\cdot)$ by

$$h^*(\cdot) = \begin{cases} 1/30 & 0 \leq w < 1.5 \\ 3/50 & 1.5 \leq w < 4 \\ 2/145 & 4 \leq w < 11.25 \\ 1/17 & 11.25 \leq w < 19.75 \\ 4/5 & 19.75 \leq w \leq 20 \end{cases}$$

and let $\hat{W} = [1,2,6,19.5,20]$.

We must now generate an information and decision strategy that is optimal with respect to our payoff function. The decision problem can be defined as

$$D = <X,\phi,D,\omega^*,A,\{M_a \mid a \in A\},c,r>$$

where:

$X = \{(-\infty,0),[0,1.5),[1.5,4),[4,11.25),[11.25,19.75),[19.75,20],(20,\infty)\}$

$\phi(\chi_0) = 0, \quad \phi(\chi_1) = .05, \quad \phi(\chi_2) = .15, \quad \phi(\chi_3) = .1$

$\phi(\chi_4) = .5, \quad \phi(\chi_5) = .2, \quad \phi(\chi_6) = 0$

$D = \{1,2,3,4,5\}$

$A = \{1,2,3,4,5\}$; i.e., experiment j consists of comparing v_u with v_{j-1} and v_j

$c = 1$ and $r = 5$.

When we use the payoff function,

$$\bar{\omega}(x,d) = \begin{cases} \bar{\omega} > 0 & \text{if } v_u \in \chi_d \\ 0 & \text{otherwise} \end{cases}$$

the resulting information structure is a binary tree (Figure 1). When $\omega(x,d) = -|\hat{w}_d - \hat{w}_j|$ is used, the resulting information structure consists of taking a single experiment at $\chi_3 = [4,11.25)$. If the optimal value, v_u, is an element of χ_3, then we choose $\hat{w}_3 = 6$ as the best available element $(d = d_3)$. If v_u is less than χ_3, we choose $\hat{w}_2 = 2$ as the best available element, and if v_u is greater than χ_3 we choose $w_4 = 19.5$ as the best available element (see Appendix B for the detailed calculations). In this case the expected value of distinguishing between χ_1 and χ_2 and between χ_4 and χ_5 is not worth the expected cost of doing so. It is important to note that complete information was not obtained. Computer science generally assumes that if the requested record is in the file, then it will be found and returned. Using the decision theoretic framework, however, that is not the case for either the common index problem, or the file search problem. Complete information will be obtained only if the expected payoff exceeds the expected cost of doing so. It is also not necessarily the case that the optimal algorithm will be the same for the file search problem as the common index problem, even if similar payoff functions are used.

We want now to determine the conditions under which complete information will be obtained for each payoff function.

3.2.3. Definition: We say that a strategy σ^* strictly dominates σ if $\Omega(\sigma^*) - \Gamma(\sigma^*) > \Omega(\sigma) - \Gamma(\sigma)$.

3.2.4. Proposition: Suppose D is a common index problem with the payoff function

$$\omega(x,d) = \begin{cases} \bar{\omega} & \text{if } v_u \in \chi_d \\ 0 & \text{otherwise,} \end{cases}$$

and that $r \geq n$ and $\theta\bar{\omega} > \bar{c}$, where: we define $\theta > 0$ and $\bar{c}$ by

$$\theta = \min\{\pi(B) \mid B \in B^A\} \quad \text{and} \quad \bar{c} = \max_{a \in A} c(a) = c.$$
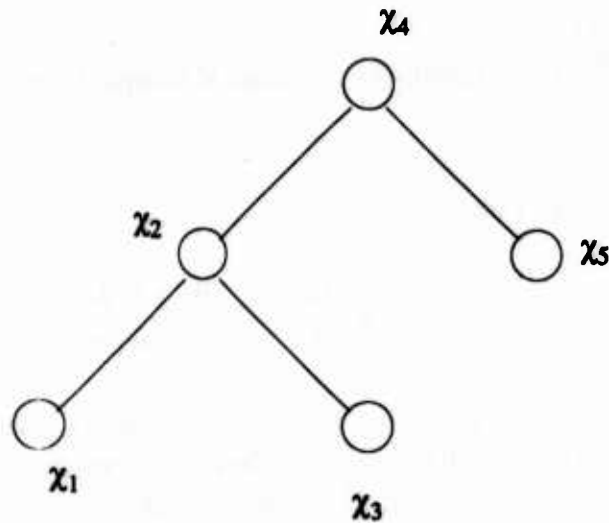
Figure 1. Optimal Binary Search Tree for the Common Index Problem

$$\text{with Payoff Function } \omega(x,d) = \begin{cases} \overline{w} > 0 & \text{if } v_u \in x_d \\ 0 \end{cases}$$

If $\sigma = <(B_1,\alpha_1),...,(B_r,\alpha_r),(B_{r+1},\delta)>$ is a feasible strategy for D such that for some $B^* \in B_{r+1}$, $B^* \notin B^A$, then σ is strictly dominated.

Proof. The proof of this can proceed by an argument essentially identical to the proof of Corollary 6.5.2 and Proposition 6.5.3. in Moore and Whinston [1987].

We shall prove a similar result for the common index problem with the alternative payoff function

$$\omega(x,d) = -|w_d - w_u|;$$

but it will be convenient to first establish the following lemma.

3.2.5. Lemma. Suppose D is a common index problem with the payoff function

$$\omega(x,d) = -|w_d - w_u|$$

that

$$p_j = \phi(\chi_j) > 0 \quad \text{for} \quad i = 1,...,n,$$

and that

$$\sigma = <(B_1,\alpha_1),...,(B_r,\alpha_r),(B_{r+1},\delta)>$$

is a feasible strategy for D, then given any $B \in B_{r+1}$,

i.   there exists $i \in \{1,...,n-1\}$ and $l \in \{1,...,n-i\}$ such that $B^* = \{\chi_i,...,\chi_{i+l}\}$, and

ii.   if $\delta(B^*) \notin \{i,...,i+l\}$, then σ is strictly dominated.

Proof. Part (i) of our conclusion follows from Theorem 6.2.3. of Moore and Whinston [1987]. To prove part (ii), we distinguish two cases.

a. $\delta(B^*) \leq i-1$. Here if we define a new strategy σ^* which is identical to σ except that we take

$$\delta^*(B^*) = i,$$

we will have; writing $d = \delta(B^*)$:

$$\Omega(\sigma^*) - \Gamma(\sigma^*) - [\Omega(\sigma) - \Gamma(\sigma)]$$

$$= \sum_{x \in B^*} \phi(x)\omega[x,\delta^*(B^*)] - \sum_{x \in B^*} \phi(x)\omega[x,\delta(B^*)]$$

$$= \sum_{x \in B^*} \phi(x)[\omega(x,i) - \omega(x,d)]$$

$$= -\sum_{k=0}^{l} p_{i+k}(|w_{i+k} - w_i| - |w_{i+k} - w_d|)$$

$$= -\sum_{k=0}^{l} p_{i+k}(w_{i+k} - w_i - w_{i+k} + w_d)$$

$$= -\sum_{k=0}^{l} p_{i+k}(w_d - w_i) = -(w_d - w_i)\sum_{k=0}^{l} p_{i+k}$$

$$= |w_i - w_d|\sum_{k=0}^{l} p_{i+k} > 0.$$

b. $\delta(B^*) \geq i+l+1$. A similar argument will suffice for this case, except that we here obtain σ^* from σ by changing $\delta(B^*)$ to

$$\delta^*(B^*) = i + l.$$

□

Using Lemma 3.2.5, we can now prove the following.

<u>3.2.6. Proposition:</u> Suppose D is a common index problem with the payoff function

$$\omega(x,d) = -|w_d - w_u|$$

that $r \geq n$, and

$$\theta \underset{j \in 1,...,n-1}{\text{Min}} |\hat{w}_{j+1} - \hat{w}_j| > \bar{c}, \tag{29}$$

where $\theta > 0$ and $\bar{c}$ are defined by

$$\theta = \min\{\pi(B) \mid B \in B^A\} = \min\{\delta_j \mid j \in \{1,...,n\} \tag{30}$$
$$\text{and} \quad \bar{c} = \max_{a \in A} c(a) = c.$$

If $\sigma = <(B_1,\alpha_1),...,(B_r,\alpha_r),(B_{r+1},\delta)>$ is a feasible strategy for D such that for some $B^* \in B_{r+1}$, $B^* \notin B^A$, then σ is strictly dominated.

<u>Proof.</u> As noted in Lemma 3.2.5, it follows from the fact that $B^* \in B_{r+1}$ that there exist $i \in \{1,...,n-1\}$ and $l \in \{1,...,i-n\}$ such that

$$B^* = \{\chi_i,...,\chi_{i+l}\}; \tag{31}$$

and, since we also have $B^* \notin B^A$, it follows that
$$l \geq 2 \tag{32}$$

Furthermore, since $r \geq n$ and $B^* \notin B^A$, it is easy to see that there must exist some $q \in \{1,...,r\}$ such that
$$\iota[\beta_q(B^*),\alpha_q(\beta_q[B^*])] = \{\beta_q(B^*)\}.$$

If $q < r$, and we define a new information-gathering strategy
$$\alpha' = <(B_1',\alpha_1'),...,(B_r',\alpha_r')>$$

by letting:
$$(B_t',\alpha_t') = (B_t,\alpha_t) \quad \text{for} \quad t = 1,...,q-1,$$

and for $t \geq q$:

$$\alpha_t'(B) = \begin{cases} \alpha_t(B) & \text{for } B \in B_t \backslash \beta_t(B^*) \\ \alpha_{t+1}(B) & \text{for } B = \beta_{t+1}(B^*) \text{ and } t < r^\dagger \\ 0 & \text{for } B = \beta_{r+1}(B^*) = B^* \text{ and } t = r; \end{cases}$$

---

† Notice that $\beta_{q+1}(B^*) = \beta_q(B^*)$, so that this function is well-defined.

it is clear that

$$B_{r+1}' = B_{r+1},$$

and

$$\Gamma(\alpha') \leq \Gamma(\alpha).$$

From these considerations, and from part (ii) of Lemma 3.2.5, we can see that it suffices to prove our result for the case where

$$B^* \in B_r, \quad \alpha_r(B^*) = 0, \tag{33}$$

and

$$\delta(B^*) \in \{i,...,i+l\} \tag{34}$$

[see (31), above].

Now, we are going to prove that σ is strictly dominated by defining a new strategy σ^* which differs from σ only in that we will set

$$\alpha_r^*(B^*) \in \{i,...,i+l\},$$

and we will re-define $\delta(\cdot)$ on $\iota[B^*,\alpha_r(B^*)]$. Notice that for any such altered strategy, σ^*, we will have

$$\Omega(\sigma^*) - \Gamma(\sigma^*) - [\Omega(\sigma) - \Gamma(\sigma)] \tag{35}$$

$$= \sum_{B \in \iota[B^*,\alpha_r^*(B^*)]} \sum_{x \in B} \phi(x)\omega[x,\delta^*(B)] - \pi(B^*)c[\alpha_r^*(B^*)]$$

$$\sum_{x \in B^*} \phi(x)\omega[x,\delta(B^*)].$$

$$= \sum_{B \in \iota[B^*,\alpha_r^*(B^*)]} \sum_{x \in B} \phi(x)(\omega[x,\delta^*(B)] - \omega(x,\delta(B^*))) - \pi(B^*)c[\alpha_r^*(B^*)]$$

Defining

$$i^* = \delta(B^*),$$

we now distinguish two cases.

    1. $i^* = 1$. Here if we set

$$\alpha_r^*(B^*) = i + l - 1 = i^* + l - 1,$$

we will have

$$\iota[B^*,\alpha_r^*(B^*)] = \{\{\chi_i,...,\chi_{i+l-2}\},\{\chi_{i+l-1}\},\{\chi_{i+l}\}\},$$

where we follow the convention that

$$\{\chi_i,...,\chi_{i+l-2}\} = \varnothing \quad \text{if } l = 1.$$

If we then define $\delta^*(\cdot)$ on $\iota[B^*,\alpha^*_r(B^*)]$ by:

$$\delta^*(B) = \begin{cases} i = i^* & \text{if } B \in \{\{\chi_i,...,\chi_{i+l-2}\},\{\chi_{i+l-1}\}\} \\ i+l & \text{if } B = \{\chi_{i+l}\}, \end{cases}$$

equation (35) becomes:

$$\Omega(\sigma^*) - \Gamma(\sigma^*) - [\Omega(\sigma) - \Gamma(\sigma)]$$

$$= -p_{i+l}(0 - |w_{i+l} - w_i|) - \pi(B^*)c(i + l)$$

$$= p_{i+l}(w_{i+l} - w_i) - \pi(B^*)c(i + l)$$

$$\geq p_{i+l}(w_{i+1} - w_i) - \pi(B^*)c(i + l)$$

$$\geq \theta \mathrm{Min}|w_{j+1} - w_j| - c(i + l) > 0,$$

where the last two inequalities are by (30) and (29), respectively. Thus σ is strictly dominated.

2. $i^* \geq i + l$. Here if we set
$$\alpha^*_r(B^*) = i + 1,$$

and define $\delta^*(\cdot)$ on $\iota[B^*,\alpha_r^*(B^*)]$ by:

$$\delta^*(B) = \begin{cases} i & \text{for } B = \{\chi_i\} \\ i^* & \text{for } B \in \{\{\chi_{i+1}\},\{\chi_{i+2},...,\chi_{i+l}\}\} \end{cases}$$

(where again we follow the convention of letting $\{\chi_{i+2},...,\chi_{i+l} = \varnothing$ for $l = 1$), we can show by an argument similar to that used in case 1 that
$$\Omega(\sigma^*) - \Gamma(\sigma^*) - [\Omega(\sigma) - \Gamma(\sigma)] > 0.$$

$\square$

For both payoff functions, the condition for complete information presented in Propositions 3.2.4 and 3.2.5 is sufficient, but not necessary. In the future, we hope to find either necessary conditions or at least tighter sufficient conditions.

4. Conclusion

We have found and characterized a solution for a version of the common index problem. There are other ways in which one might formalize the problem, and we intend to explore some of these in later work. Our present view, however, is that the principle weakness in the applicability of the solution presented here lies in our assumption that the preferences of the ith user can be represented as a composite function

$$u^i(s) = \hat{u}^i(|\gamma(\sigma) - v_i|),$$

where $\gamma:S \to W$ is the same index for all users. On the other hand, psychological researchers have appeared to obtain good results with this sort of model in a fairly wide variety of contexts, [9] but at the same time it is clear that this is a restrictive assumption; particularly in that it is necessary for applications that the $\gamma(\cdot)$ function be known (up to similarity transform), or at least that its values at the points in the file be known. [10] We are currently investigating the question of whether and how this assumption can be weakened.

References

Aho, A.V., Hopcroft, J.E., and Ullman, J.D. (1974). *The Design and Analysis of Compter Algorithms*, Addison-Wesley Publishing Co., Reading, Mass., pp. 115-119.

Bennett, Joseph F. and Hayo, W.L. (1960). "Multidimensional Unfolding: Determining the Dimensionality of Ranked Preference Data," *Psychometrika*, Vol. 25, pp. 27-43.

Bonczek, R.H., C.W. Holsapple and A.B. Whinston (1981). *Foundations of Decision Support Systems*, Academic Press, 1981.

Coombs, Clyde H. (1950). "Psychological Scaling Without a Unit of Measure," *Psychological Review*, Vol. 57, pp. 145-158.

Coombs, Clyde H., Dawes, R.M. and Tversky, A. (1970). *Mathematical Psychology*, Prentice-Hall.

Coombs, Clyde H. (1965). *A Theory of Data*, Wiley.

DeGroot, H.M. (1970). *Optimal Statistical Decisions*, McGraw-Hill, Inc.

---

9. In addition to the reference already cited on the subject, we should mention Coombs [1965], Torgerson [1965], Bennett and Hayo [1960].

10. The literature on "unfolding" cited earlier does, however, discuss methods of constructing such a scale for a given set of subjects (users).

Hall, H.K., Moore, J.C. and Whinston, A.B. (1985). "An Economic Basis for Expert Systems," Working Paper, Krannert Graduate School of Management, Purdue University, West Lafayette, IN.

Marschak, J. and Radner, R. (1972). *Economic Theory of Teams*, Cowles Foundation.

Mendelson, H. and Samaria, A.N. (1986). "Incomplete Information Costs and Database Design," *ACM Transactions on Database Systems*, Vol. II, No. 2, June, pp. 159-185.

Moore, J.C. and Whinston, A.B. (1986). "A Model for Decision-Making with Sequential Information-Acquisition - Part 1," forthcoming in *Decision Support Systems*.

Moore, J.C. and Whinston, A.B. (1987). "A Model for Decision-Making with Sequential Information-Acquisition - Part 2," forthcoming in *Decision Support Systems*.

Motro, A. (1986). "Supporting Goal Queries in Relational Databases," *Proceedings of the Conference on Expert Systems and Data Base*, University of South Carolina, 1986.

Sleator, D.D. and Tarjan, R.E. (1985). "Self-Adjusting Binary Search Trees," *Journal of the Association for Computing Machinery*, Vol. 32, No. 3, July, pp. 652-686.

Torgerson, Warren S. (1965). "Multidimensional Scaling of Similarity," *Psychometrica*, Vol. 30, pp. 379-393.

Appendix A
Notation

T	Universal set of possible elements from which to choose
S	$S \subseteq T$. Set of available alternatives.
U	Set of possible users.
u	$u \in U$ utility function representing a specific user.
v_u	The index of ideal elements for user u.
$h(\cdot)$	Probability distribution over U.
W	Common index over T.
γ	$\gamma: T \rightarrow Y$ Function that assigns an index to each element of T.
$\hat{W}$	$\{w \in W \mid w = \gamma(S), s \in S\}$ set of indices of the available elements.
$h^*(\cdot)$	Probability distribution on W. Induced by $h(\cdot)$.
X	Set of possible mutually exclusive states.
ϕ	Probability density function over X.
π	Probability distribution over the power set of X.
D	Set of final decisions.
ω^*	Payoff function.
A	Set of initial experiments.
M_a	Information structure associated with $a \in A$.
$c(a)$	Cost of using action a.
r	Number of experiments that can be taken.
B	Partition of X.
B	element of **B**.
$t(B,A)$	Partition of B induced by experiment a.
$a(t,B)$	Action taken on the set B of time t
Ω^*	Expected payoff for a search strategy (social welfare function).
$\Omega(\sigma)$	Expected gross payoff for the strategy σ.
$\Gamma(\sigma)$	Expected cost of the strategy σ.
$C(B)$	Cost of a path resulting in B.
$V(B)$	Potential gross payoff associated with B.
$D^*(B)$	Conditionally optimal decision set for B.
$B^A(d)$	Set of elements in B^A for which d is the optimal decision.
X_d	Union of the elements in $B^A(d)$ for a given d.
$\psi(B)$	$\text{Max}\{\pi(B \cap X_d) \mid d \in D\}$ for $B \subseteq X$.
B^A	The finest possible partition of X.

$$\omega(x,d) = -|\hat{w}_j - \hat{w}_d|$$

$$
\begin{array}{ll}
\Delta M_{02} = 0 & \Delta B_{02} = 0 \\
\Delta M_{12} = 0 & \Delta B_{13} = 0 \\
\Delta M_{22} = 0 & \Delta B_{24} = 0 \\
\Delta M_{32} = 0 & \Delta B_{35} = 0 \\
\Delta M_{42} = 0 & \Delta B_{46} = 0 \\
\Delta M_{52} = 0 & \\
\Delta M_{62} = 0 &
\end{array}
$$

	f(0)	f(1)	f(2)	f(3)	f(4)	f(5)
$\Delta B_{03} = -0.05$	-0.05	-0.2	-0.2			
$\Delta B_{14} = -0.2$	-0.2		-0.25	-0.25		
$\Delta B_{25} = -0.6$	-0.675			-0.6	-0.6	
$\Delta B_{36} = -0.025$	-0.025				-0.7	-0.7
$\Delta B_{04} = -0.3$	-0.45	-0.5	-0.3	-0.35		
$\Delta B_{15} = -0.75$	-1.55		-1.35	-0.75	-0.95	
$\Delta B_{26} = -0.725$	-0.725			-0.825	-0.8	-1.4
$\Delta B_{05} = -0.85$	-4.9	-1.55	-1.4	-0.85	-1.1	
$\Delta B_{16} = -0.975$	-3.05		-1.675	-0.975	-1.15	-1.7
$\Delta B_{06} = -1.075$	-5	-1.975	-1.725	-1.075	-1.3	-1.85

Figure 2. Details of the Dynamic Programming Solution
for $\omega(x,d) = -|\hat{w}_j - \hat{w}_d|$

$$\omega(x,d) = \begin{cases} \bar{\omega} & \text{if } v_u \in \chi_d \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta M_{02} = 0 \qquad \Delta B_{02} = 5$$
$$\Delta M_{12} = 5 \qquad \Delta B_{13} = 15$$
$$\Delta M_{22} = 15 \qquad \Delta B_{24} = 10$$
$$\Delta M_{32} = 10 \qquad \Delta B_{35} = 50$$
$$\Delta M_{42} = 50 \qquad \Delta B_{46} = 20$$
$$\Delta M_{52} = 20$$
$$\Delta M_{n+12} = 0$$

	f(0)	f(1)	f(2)	f(3)	f(4)	f(5)
$\Delta B_{03} = 19.8$	3	19.8	19.8			
$\Delta B_{14} = 24.75$	3.75		24.75	24.75		
$\Delta B_{25} = 59.4$	30			59.4	59.4	
$\Delta B_{36} = 69.3$	35				69.3	69.3
$\Delta B_{04} = 29.7$	4.5	25.45	29.7	29.5		
$\Delta B_{15} = 74.25$	37.5		73.65	74.25	74	
$\Delta B_{26} = 79.2$	40			78.5	79.2	78.6
$\Delta B_{05} = 79$	40	78.45	78.6	79	78.9	
$\Delta B_{16} = 93.8$	47.5		93.25	93.35	93.8	93.3
$\Delta B_{06} = 98.7$	50	97.8	98.2	98.1	98.7	98

Figure 3. Details of the Dynamic Programming Solution for

$$\omega(x,d) = \begin{cases} \hat{\omega} & \text{if } v_u \in \chi_d \\ 0 & \text{otherwise} \end{cases}$$

Comments on Multiple Bifurcations

John Guckenheimer
Cornell University

This lecture is a summary of several problems involving bifurcations in systems of ordinary differential equations that depend upon several parameters. All mathematical details of the problems discussed here appear in papers listed in the bibliography. The discussion here is concerned with the philosophical viewpoint and practical significance of this type of study.

Many applications of mathematics involve solving systems of ordinary differential equations that depend upon several parameters. Although there are many well established algorithms for numerically integrating individual solutions of such systems, achieving a qualitative understanding of the dynamics of systems of moderate size and how these dynamics vary with parameters is still a formidable task. Even for simple models such as the Henon mapping and forced oscillators with one degree freedom [11]. the details associated with chaotic behavior in systems with nonhyperbolic nonwandering sets still eludes us. Going further to describe the details of bifurcations in these systems is still more complex, and it is presumptuous to expect answers that are mathematically complete. The recent work of Chenciner [1] is a good benchmark for the level of detail that can be achieved at this time in studies of bifurcations which involve chaotic and quasiperiodic behavior.

By definition, bifurcations are qualitative changes in the dynamics of a system which occur as parameters are varied. A simple example is the Hopf bifurcation which occurs as a family of periodic orbits emerge from an equilibrium which has a simple pair of purely imaginary eigenvalues. The approach to bifurcation theory adopted here relies upon a progressive classification of more and more complicated bifurcations which occur robustly in families of increasing numbers of parameters. The viewpoint is one of looking for generic or typical cases which are expected to occur in a wide variety of examples. The mathematical

techniques trace their origins to transversality theorems. The search for simplicity in the typical bifurcations and their classification is complicated by the presence of symmetry. Many physical systems are at least approximately symmetric with respect to various symmetry groups, and this approach to bifurcation theory requires that careful, explicit consideration of the symmetry be built in from the beginning. The complications inherent in the presence of symmetry increase the mathematical richness of the subject.

From a practical point of view, bifurcation problems involving degenerate equilibria are approached in the following manner. One has a system of differential equations with an equilibrium point at which the linearization has zero or purely imaginary eigenvalues. In the presence of a symmetry group for the system of equations, the symmetry may force the eigenvalues to have high multiplicity. An initial classification of bifurcations is made, based upon the structure of the generalized eigenspaces associated to zero and purely imaginary eigenvalues. Having determined this information, one proceeds to calculate <u>normal forms</u> for biifurcations with the specified type of linear part. This entails an algebraic calculation in which the Taylor expansion of the equilibrium is transformed by near identity coordinate transformations into one for which there are as few nonzero terms in the Taylor expansion as possible. In the case of a system with a symmetry, one can insist that the coordinate transformations respect the symmetry. Different bifurcations correspond to the vanishing of terms or expressions in the coefficients of the normal forms which play an essential role in determining qualitative features of the dynamics. Finally, parameters are introduced to the normal forms in a way that hopefully produces a family of equations whose main dynamical features remain qualitatively unchanged if the family is perturbed.

The normal form equations are closely related to the <u>reduced</u> <u>bifurcation equations</u> that are often obtained by introducing asymptotic expansions at a bifurcation. In either approach, one is faced with the problem of solving systems of differential equations with polynomial right hand sides. This is a notoriously difficult mathematical problem in

general, and each example requires individual treatment at this time. Beyond the calculation of the location and stability of the equilibria, further analysis of the dynamics has relied upon either numerical data or perturbation arguments. In many multiple bifurcation problems, it has been possible to analyze the occurrence of periodic orbits and invariant tori via perturbation methods that begin with systems which have explicit analytic integrals. By appropriately scaling the variables, coefficients and parameters of the normal form, one can seek to obtain a limit in which the system does have explicit integrals. For two dimensional systems (or ones which can be reduced to an analysis of two dimensional systems) with trajectories lying in algebraic curves, the perturbation arguments involve the study of abelian integrals and how these depend upon the parameters of the system. These arguments establish a connnection between these problems and algebraic geometry.

Following the analysis of the dynamics of the normal form equations, one still has an additional step that is needed to complete the mathematical study of particular bifurcations. This final step involves defining an appropriate equivalence relation between families of vector fields and showing that the normal form is stable with respect to this equivalence relation; i.e., perturbations of the family lie in the same equivalence class. This aspect of the work involves verifying that the computation of the normal forms to higher degree produces systems which do not differ qualitatively from the normal form which has been analyzed in detail. There is a subtle and difficult point frequently encountered here in that "flat" terms with zero Taylor expansions may play an important role in determining the dynamics of a system which has been formally reduced to its normal form. Perturbations methods that successfully cope with these flat terms are lacking in the theory. The issue of choosing an equivalence relation in terms of which stability of a family will be determined is also an awkward and messy issue to deal with in many of the examples.

The analysis of codimension two bifurcations by the methods outlined above has been quite successful. Corresponding to the three cases of a two dimensional generalized eigenspace for zero, a zero and a

pair of purely imaginary eigenvalues, and two pairs of purely imaginary eigenvalues with an irrational ratio, there are analyses of two parameter families of vector fields which are reasonably complete apart from the difficulties mentioned in the previous paragraph [11]. Recent analytic work has sought to extend these analyses, to higher codimension bifurcations which involve degeneracies due to the vanishing of certain nonlinear terms in the normal forms [4, 6]. Perhaps the most intricate problem of this type which has been examined thus far is a four parameter family of two dimensional vector fields studied by Dangelmayr and Guckenheimer [4]. The system they studied is

$$\dot{x} = y$$
$$\dot{y} = -(x^3 + \lambda_1 x^2 + \lambda_2 x + \lambda_3) + y(\lambda_4 - x^2) .$$

One motivation for studying this system is that it represents the effect of symmetry destroying perturbations on a codimension two bifurcation within the class of systems that are symmetric with respect to rotation by π in the plane. In this work as well as that of other groups who have studied codimension three bifurcations of two dimensional vector fields, it is necessary to examine "global" codimension two bifurcations which entail the exitence of homoclinic or heteroclinic orbits as part of their degeneracy.

A second area in which there has been extensive recent activity extending the analysis of codimension bifurcations involves degenerate bifurcations in the presence of $O(2)$ as symmetry group. Here $O(2)$ is the group of 2×2 orthogonal matrices, including the reflections that reverse orientation. There are many different cases of codimension 2 bifurcation of equilibria in the context of systems with $O(2)$ symmetry. This is due to the variety of representations of $O(2)$ which produce different combinations of purely imaginary and zero eigenvalues. A vigorous effort which includes studies of Chossat et al [2,3], Dangelmayr and Knobloch [5], Golubitsky et al. [7,8], Guckenheimer [10], Keyfitz et al. [12], and Knobloch [13] has extended the initial analysis by Ruelle of Hopf bifurcation in the presence of $O(2)$ symmetry to many of the codimension 2 cases which occur. There are applications of these results to a number of systems, perhaps the most notable being Taylor-Couette flow and the study of oscillatory behavior in systems

776

which are homogeneous in one spatial dimension with periodic boundary conditions. The volume edited by Golubitsky and Guckenheimer [7] provides a thorough overview of work in this area through the summer of 1985, but additional work on problems of this type has been appearing since then at a rapid rate.

The complexity of recent studies of multiple bifurcation phenomena prompts a reconsideration of the utility of the results for applications. One motivation for the study of multiple bifurcations is that they represent an entree into parameter space regimes in which complicated dynamical behavior occurs. Without integration of a system of equations, the location of parameter values at which multiple bifurcations occur identifies regions in which complicated dynamics appears. This is thoroughly nontrivial for applications like the dynamics of chemical reactors in which the parameter space of a system has a high dimension that cannot be searched with fine resolution in numerical studies. It is also important for problems of fluid dynamics in which numerical integration of the underlying equations requires enormous computational resources and is therefore expensive when it is feasible at all. A difficulty which is encountered in such problems is that more degenerate bifurcations and larger symmetry groups entail more nonzero coefficients in normal forms. For application to problems in which one is confident of the validity of equations describing a mathematical model, one finds extensive algebraic manipulations are required to compute the coefficients of the normal forms associated with the model. These can be performed in some cases with symbolic computational systems such as MACSYMA, SMP, MAPLE, SCRATCHPAD or REDUCE, but the problems quickly strain the limits of this technology. In other cases, a valid mathematical model for experimental data is not readily available and one must try to estimate or fit the values of normal form coefficients to data. When there are many coefficients, this is a formidable task.

A second difficulty which arises in applying the results of a multiple bifurcation analysis of a high codimension bifurcation to experimental data is that there are qualitative features that seem to

occur in very small parameter regions and are difficult to detect quantitatively. This observation is based upon numerical computations in which it is difficult to resolve subsidiary bifurcations that occur close to one another in a multidimensional parameter space. Experience seems to indicate that some features are easily missed unless one works systematically, piecing together complicated stability diagrams from a knowledge of the stability diagrams of lower codimension bifurcations which occur in examples. The incidence of errors in drawing stability diagrams on the basis of numerical computations with standard floating point calculations has been quite high. Perhaps one should not be overly concerned with these difficulties when dealing with applications, but the goal of the theory is to develop a comprehensive explanation of complicated dynamical phenomena found in the mode interactions associated with multiple bifurcations. Ignoring aspects of the problem that are quantitatively insignificant can easily lead to a mathematically inconsistent picture. No good resolution of this dilemna has appeared. It suggests that considerable caution should be used in trying to use numerical software for such tasks as tracking periodic orbits of a dynamical system in the close vicinity of highly degenerate bifurcations.

References:
1. A. Chenciner, Bifurcation de points fixes elliptiques,
     I. Publicationes Math. de l'I.H.E.S. **61**, 67-127, 1985.
     II. Inventiones Math. **80**, 81-106, 1985.
     III. Preprint, Université Paris VII

2. P. Chossat, Y. Demay and G. Iooss, Interaction de modes azimutaux dans le probleme de Couette-Taylor, Arch. Rat. Mech . and Anal., to appear.

3. P. Chossat, M. Golubitsky, and B. Keyfitz, Hopf-Hopf mode interactions with O(2) Symmetry, University of Houston, preprint.

4. G. Dangelmayr and J. Guckenheimer, On a four parameter family of vector fields, Arch. Rat. Mech . and Anal., to appear.

5. G. Dangelmayr and E. Knobloch, The Takens-Bogdanov bifurcation with O(2) symmetry, University of Tubingen, preprint.

6. F. Dumortier, R. Roussarie, and J. Sotomayor, Generic 3-parameter families of vector fields in the plane, unfolding a singularity with nilpotent linear part. The cusp case of codimension 3, preprint.

7. M. Golubitsky and J. Guckenheimer, eds., Multiparameter Bifurcation Theory, Contemporary Mathematics, **56**, Am. Math. Soc., 1986.

8. M. Golubitsky and M. Roberts, A classification of degenerate Hopf bifurcations with O(2) symmetry, University of Houston, preprint.

9. M. Golubitsky and I. Stewart, Symmetry and stability in Taylor-Couette flow, SIAM J. Math. Anal., **17**, 249-288, 1986.

10. J. Guckenheimer, A codimension two bifurcation with O(2) symmetry, Contemporary Mathematics, **56**,175-184, 1986.

11. J. Guckenheimer and P. Holmes, Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields, Springer-Verlag, 1983.

12. B. Keyfitz, M. Golubitsky, M. Gorman, and P. Chossat, The use of symmetry and bifurcation techniques in studying flame stability, Lectures in Appl. Math., **24, part 2**, 293-315, 1986.

13. E. Knobloch, On the degenerate Hopf bifurcation with O(2) symmetry, Contemporary Mathematics, **56**,193-202, 1986.

MEASURES OF BLOCK DESIGN EFFICIENCY RECOVERING INTERBLOCK INFORMATION

Walter T. Federer
337 Warren Hall, Biometrics Unit
Cornell University
Ithaca, NY 14853
and
Terry P. Speed
Division of Mathematics & Statistics, CSIRO
G.P.O. Box 1965
Canberra, ACT 2601, Australia

ABSTRACT.   In evaluating goodness of a class of designs, researchers have used a measure of design efficiency proposed by F. Yates in the thirties. This measure considers only intrablock information and does not make use of the information contained in the interblock variance. The measures of efficiency proposed here are dependent upon the ratio of the interblock and intrablock components of variance,   i.e., $\sigma_\beta^2/\sigma_\epsilon^2 = \gamma$. The efficiency of one block design to a second may not remain invariant with respect to this ratio. Incomplete block designs which were inefficient under the intrablock measure, now become quite efficient for some ratios of γ. Likewise, the indications are that interblock information should always be recovered when analyzing data from experiments arranged in an incomplete block design.

I. INTRODUCTION. In the mid-thirties Yates (e.g. in 1937) introduced an efficiency factor for partially confounded factorials and for incomplete block designs. The factor is computed as the ratio of the average variance of a difference between two adjusted means (or for factorial effects) to the variance of a difference of two means from an orthogonal design such as a completely randomized or randomized complete block design assuming no change in the error variance σ_ϵ^2 for the two designs. It is common practice in statistical literature to present this efficiency factor for designs and to discuss optimality of classes of designs in terms of the Yates efficiency factor, which only makes use of the intrablock error variance. No use is made of the information contained in the interblock variance obtained from the incomplete blocks (eliminating treatment effects) sum of squares. A more proper efficiency factor should make use of the information contained in both the intrablock and the interblock variances. Some measures accomplishing this are presented in this paper.

II. BALANCED INCOMPLETE BLOCK DESIGN.

A classical balanced incomplete block design (BIBD) consists of v treatments arranged in b incomplete blocks of size k, k<v, with r repetitions of each treatment, and with each and every pair of treatments occurring together in an incomplete block λ times. The standard relations are bk = rv, $\lambda = r(k-1)/(v-1)$, and $e = (1-k^1)/(1-v^1) = v(k-1)/k(v-1)$. The factor e is the Yates efficiency factor. The usual response model assumed for a BIBD is

$$Y_{hij} = n_{hij} (\mu + \rho_h + \beta_{hi} + \tau_j + \epsilon_{hij}), \tag{1}$$

where Y_{hij} is the response for the jth treatment in the ith incomplete block in the hth complete block, h=1,...,r; i=1,...b/r; j=1,...,v; n_{hij} is one if the jth treatment occurs in the hith incomplete block and zero otherwise; μ is a general mean effect; ρ_h is the hth complete block effect; β_{hi} is the hith random incomplete block effect distributed with mean zero and common variance σ_β^2, τ_j is the jth treatment effect, and ϵ_{hij} are random error effects which were distributed with mean zero and variance σ_ϵ^2. An analysis of variance is given in Table 1.

TABLE 1. Analysis of variance for a resolvable BIBD.

Source of variation	Degree of freedom	Expected value of mean square
Total	bk = vr	–
Correction for mean	1	–
Treatment (ignoring incomplete block effects)	v-1	–
Within treatments	v(r-1)	–
Blocks (eliminating treatment effects)	b-1	$\sigma_\epsilon^2 + \dfrac{bk-v}{b-1} \sigma_\beta^2$ [1]
Complete blocks	r-1	
Incomplete blocks (elim. tr.)	b-r	$\sigma_\epsilon^2 + \dfrac{r-1}{r} k\sigma_\beta^2$
Intrablock error	vr-v-b+1	σ_ϵ^2

[1] Expected mean square for $\rho_h = 0$, i.e., no complete block effects.

Intrablock information $\omega = 1/\sigma_\epsilon^2$

Interblock information $\omega' = 1/(\sigma_\epsilon^2 + k\sigma_\beta^2)$

For intrablock contrasts the variance of a difference between two estimated treatment effects is $2\sigma_\varepsilon^2/re$, when $e=(1-1/k)/(1-1/v)$. For interblock contrasts the variance of a difference between two treatments effects is

$$\frac{2(\sigma_\varepsilon^2 + k\sigma_\beta^2)}{r(1-e)} . \tag{2}$$

For the combined estimator of a difference between two treatment effects the variance is

$$\frac{2}{r} \left\{ \frac{e}{\sigma_\varepsilon^2} + \frac{1-e}{\sigma_\varepsilon^2 + k\sigma_\beta^2} \right\}^{-1}$$

$$= \frac{2\sigma_\varepsilon^2}{r} \left\{ \frac{1 + k\gamma}{1 + ke\gamma} \right\} , \tag{3}$$

where $\gamma = \sigma_\beta^2/\sigma_\varepsilon^2$.

Since the intrablock contrast variance is of the form $2\sigma_\varepsilon^2/re$, it would be logical to have the combined estimator variance in the same form i.e., $2\sigma_\varepsilon^2/re_1^*$, where

$$e_1^* = \frac{1 + ke\gamma}{1 + k\gamma} = 1 - \frac{k(1-e)}{1 + k\gamma} . \tag{4}$$

A second measure not involving e is

$$e_2^* = \frac{1 + k\gamma}{1 + (k+1)\gamma} = 1 - \frac{1}{k + 1 + 1/\gamma} \tag{5}$$

The latter measure of efficiency depends only upon $\gamma = \sigma_\beta^2/\sigma_\varepsilon^2$ and k; note that (4) is also a function of k and γ since $e=k/(k+1)$ for $v=k^2$ and $r=k+1$. A comparison of the two measures is given in Table 2 for k=3,7, and 11. There is little to choose between e_1^* and e_2^* and it is suggested that e_1^* be used as a measure of efficiency rather than e_2^*. Note that as γ approaches zero the efficiency for all k approaches unity. When γ approaches infinity, the Yates efficiency factor e is approached for all k. For small k, e is relatively low indicating an inefficient design. However, e_1^* indicates that designs with small k are quite efficient if γ is 1/4 to 1/16, say. From these results, it is suggested that interblock information should always be recovered and that inefficiency of incomplete block design is not a problem.

TABLE 2. Intrablock - interblock efficiencies for various values of γ for k = 3, 7, and 11.

| | | k | | | | | |
| | | 3 | | 7 | | 11 | |
γ		e_1^*	e_2^*	e_1^*	e_2^*	e_1^*	e_2^*
0		1	1	1	1	1	1
1/32		.98	.97	.98	.98	.98	.98
1/16		.96	.95	.96	.96	.97	.96
1/4		.89	.88	.92	.92	.94	.94
1/2		.85	.83	.90	.90	.93	.93
1		.81	.80	.89	.89	.92	.92
2		.79	.78	.88	.88	.92	.92
4		.77	.76	.88	.88	.92	.92
∞		.75	.75	.875	.875	.917	.917

$$e_1^* = \frac{1 + ke\,\gamma}{1 + k\,\gamma} \, . \qquad e_2^* = \frac{1 + k\,\gamma}{1 + (k+1)\gamma} \, .$$

For a randomized complete block design, the variance of a difference between two arithmetic means is

$$\frac{2\sigma_e^2}{r} \left\{ 1 + \frac{v - k}{v - 1}\, \gamma \right\} , \qquad (6)$$

whereas the variance of a difference between two adjusted treatment means from a BIBD is

$$\frac{2\sigma_e^2}{re_1^*} \qquad (7)$$

784

Now (6) $\geq$ (7), their difference being

$$1 + \frac{v - k}{v - 1} \gamma - \frac{1 + k \gamma}{1 + ke \gamma}$$

$$= \frac{v(v-k)(k-1) \gamma^2}{(v-1)[v-1 + v(k-1) \gamma]} .$$

(8)

(8) is zero when $\gamma = 0$ and/or $v = k$. Equation (8) could be another measure of intrablock - interblock efficiency. Perhaps a more appropriate measure would be a ratio rather a difference to obtain

$$e_3^* = 1/e_1^* \left(1 + \frac{(v-k)\gamma}{v-1} \right).$$

(9)

The measure e_3^* would conform more to the definition of efficiency origin-ally presented by Yates but would include both intrablock and interblock information.

III. OTHER BLOCK DESIGNS.

A class of generalized N-ary designs were discussed by Shafiq and Federer (1979, 1983). For these designs the response model equation (1) is replaced by

$$Y_{ghij} = \mu + \rho_h + \beta_{hi} + \tau_j + \varepsilon_{ghij},$$

(10)

where $g = 0, .., n_{hij}$ and when $n_{hij} = 0$ there is no response Y_{ghij}. The other symbols are as defined in (1). The above authors generalized the Yates efficiency factor for this class of designs and they also proved that the Fisher inequality $v \leq b$ holds for this general class of balanced block designs. The efficiency factor e_1^* for the generalized balanced block design is

$$e_1^* = 1 - \frac{c - \lambda}{r(k + \gamma^{-1})} ,$$

(11)

where $c = \sum_h \sum_i n_{hij}^2$, n_{hij} is the number of times treatment j occurs in block hi, and $r = \sum_h \sum_i n_{hij}$ is the number of replicates for treatment j.

For the class of resolvable incomplete block designs known as lat-tices, the average variance of a difference between two adjusted treatment means is

$$\frac{2}{k+1} \left\{ \frac{r}{(r-1)\omega + \omega'} + \frac{k-r+1}{r\omega} \right\}$$

785

$$
= \frac{2\sigma_e^2}{r} \left(\frac{r}{k+1}\right) \left(\frac{r}{r-1 + (1+k\gamma)^{-1}} + \frac{k-r+1}{r}\right) . \tag{12}
$$

As before, we may take

$$
e_1^* = \frac{k+1}{r} \left(\frac{r}{r-1 + (1+k\gamma)^{-1}} + \frac{k-r+1}{r}\right)^{-1} . \tag{13}
$$

Note that r is the number of confounding arrangements and r=2 for the simple (double) lattice, r=3 for the triple lattice, etc.

An intrablock - interblock measure of efficiency like e_3^* would be

$$
e_3^* = \frac{\left[1 + \dfrac{r}{k+1} \left(\dfrac{1 - (1+k\gamma)^{-1}}{r-1 + (1+k\gamma)^{-1}}\right)\right]}{\left[1 + \dfrac{k}{k+1}\gamma\right]} \tag{14}
$$

Note that $v=k^2$ and r is the number of geometrical factorial effects confounded.

Using the average variance of a difference between two adjusted treatment effects, we could construct e_1^* and e_3^* for any class of incomplete block designs. The ideas in this paper may be used to construct efficiency factors similar to e_1^* and e_3^* for cubic lattices, for lattice squares, and for other designs.

IV. LITERATURE CITED.

Shafiq, M. and W. T. Federer (1979). "Generalized N-ary balanced block designs." *Biometrika* 66, pp. 115-123.

Shafiq, M. and W. T. Federer (1983). "General binary partially balanced designs." *Indian J. Agric. Statist.*, XXXV(2).

Yates, F. (1937). "The design and analysis of factorial experiments." *Imperial Bur. Soil Sci., Tech. Comm.* 35, pp. 1-95.

COMPUTING ASYMPTOTIC CONFIDENCE BANDS
FOR NONLINEAR REGRESSION MODELS

John J. Peterson

Department of Mathematics
Syracuse University
Syracuse, New York 13244-1150

ABSTRACT. This paper introduces an easy-to-implement, quadratically constrained, optimization algorithm that is particularly suited to computing asymptotic confidence bands for many nonlinear models. Convergence of this algorithm is proved and several applications are discussed. These applications include: nonlinear regression, multinomial logistic regression, and covariate-dependent reliability models.

1. INTRODUCTION. In nonlinear regression, investigators are often interested in the underlying form of the response curve. As such, simultaneous confidence bands are particularly useful in nonlinear regression. If a closed confidence region exists for the regression parameters, then in theory an approximate confidence band can be obtained for the nonlinear response function. This is achieved by minimizing and maximizing the response function over the confidence region for various values of the predictor variables. Such confidence bands are generally conservative, but close to nominal value if the predictor variable can take on a considerable range of values. Procedures for obtaining approximate confidence regions that are closed and have a quadratic form can be found in Bates and Watts (1981) and Hamilton, Bates and Watts (1982). However, even with this simplification, construction of the confidence band generally requires solving many pairs of nonlinear programming problems. In fact, in one of the very few papers on confidence bands in nonlinear regression, Khorasani and Milliken (1982) only describe the general process as a "computational tedium" and go on to discuss two special cases where the confidence bands have a closed functional form.

Indeed, consideration of general, nonlinear optimization techniques (for nonlinear constraints) may discourage practioners form attempting to compute confidence bands for nonlinear regression models. Nonlinearly constrained optimization procedures that seem to be available in the literature can be classified into Lagrangian-penalty function approaches (Bertsekas, 1982) or feasible direction approaches (Ben-Isreal, Ben-Tel, Zlobe, 1981). These procedures are for quite general problems, and may require substantial calculations to complete one iteration. This can be undesirable since the confidence band requires the solution of many pairs of nonlinear optimization problems. Furthermore, these algorithms are considerably more difficult to implement in practice than the widely available procedures used to estimate the parameters in nonlinear regression. However, by exploiting two properties special to most regression inference situations, it is possible to compute approximate confidence bands for nonlinear regression models in a relatively easy and efficient manner.

787

In this paper an algorithm is presented for computing confidence bands for nonlinear regression models using a quadratic confidence region for the regression coefficients. This algorithm is especially easy to implement if matrix oriented software is used such as PROC MATRIX (SAS Institute, Inc, 1985a), IML (SAS Institute, Inc. 1985b), or APL.

In the following sections, an algorithm for computing asymptotic confidence bands is presented and investigated. In section 2, the problem is formulated. In section 3, the algorithm is derived and a stepsize modification is briefly discussed. A convergence theorem is proved in section 4. Also in section 4, a geometric justification is given for the algorithm's convergence properties. Finally, in section 5, applications of the algorithm to other nonlinear simultaneous estimation problems are discussed.

2. PROBLEM STATEMENT. In this paper we define the nonlinear regression model as,

$$y_t = f(x_t, \theta) + \varepsilon_t,$$

where the ε_t's are independent, normally distributed random errors with zero mean and variance σ^2 for all x_t values $(t = 1, \ldots, n)$. The x_t's are (possibly vector valued) predictor variables, and θ is a $p \times 1$ vector of unknown model parameters. Here, $f(x, \theta)$ is a known function of θ for each $x \varepsilon X$, where X is the set of all possible predictor variable values. For notational compactness $f(x, \theta)$ will sometimes be denoted as simply $f(\theta)$.

It is assumed that an approximate $100(1-\alpha)\%$ confidence region for θ can be written in the form

$$\theta_\alpha = \{\theta: (\theta - \hat{\theta})' \, C \, (\theta - \hat{\theta}) \leq b_\alpha\},$$

where $\hat{\theta}$ is the least squares estimate of θ, C is a $p \times p$ positive-definite, symmetric matrix, and b_α is a (boundary) value that is determined by α. Typically, b_α has the form $ps^2 F(p, n-p; \alpha)$, where s^2 is an estimate of σ^2, and C has the form $P'P$ where P is the $n \times p$ matrix of partial derivatives

$$\frac{\partial}{\partial \theta_j} f(x_t, \theta), \quad t=1, \ldots, n, \ j=1, \ldots, p.$$

See Bates and Watts (1981) and Hamilton, Watts, and Bates (1982) for other possible formulations of b_α and C, and for possible reparameterizations to improve the coverage probability of the $100(1-\alpha)\%$ confidence region.

A further assumption imposed upon f is that for each $x \varepsilon X$, and each $\theta \varepsilon \theta_\alpha$, f is a continuously differentiable function of θ, with $\nabla f(\theta) \neq 0$ on θ_α. This condition is satisfied quite often in practice. In fact, most $f(\theta)$ will be monotonic in each θ_j on θ_α for all $x \varepsilon X$.

Approximate $100(1-\alpha)\%$ simultaneous confidence bounds for the set of functions $\{f(x,\theta): x\epsilon X\}$ can be obtained by computing

$$(2.1) \qquad [\min_{\theta_\alpha} f(\theta), \max_{\theta_\alpha} f(\theta)]$$

for each $x\epsilon X$ (Rao, 1973). Computation of the pairs (2.1) for various x's in X requires the repeated use of nonlinear programming. However, the optimization problems in (2.1) have two important characteristics which will be exploited in deriving the algorithm used to construct the confidence band for $f(\theta)$. Firstly, θ_α is a closed, quadratic region. This property is not only useful in deriving the iterative form of the algorithm, but it also proves useful in establishing global and local convergence properties. Secondly, the assumption that $f(\theta)$ is continuously differentiable with $\nabla f(\theta) \neq 0$ on θ_α for each $x\epsilon X$, implies that (2.1) can be simplified to

$$(2.2) \qquad [\min_{\theta_\alpha^b} f(\theta), \max_{\theta_\alpha^b} f(\theta)]$$

where θ_α^b is the boundary of θ_α.

3. THE ALGORITHM. In this section the algorithmic maps for seeking the solution to (2.2) are derived. If $f(\theta)$ was linear in θ then (2.2) could be obtained in closed form by using the method of Lagrange multipliers. The algorithms to be presented attempt to solve (2.2) by obtaining successive linear approximations to $f(\theta)$. In theory, stepsize modifications may be necessary to obtain convergence.

To begin, approximate $f(\theta)$ by

$$(3.1) \qquad f(\theta_0) + \nabla f(\theta_0)'(\theta-\theta_0),$$

where θ_0 is the starting value. The optimal values to minimize and maximize (3.1) over θ_α^b can be obtained by solving

$$(3.2) \qquad [\min_{\theta_\alpha^b} \nabla f(\theta_0)'\theta, \max_{\theta_\alpha^b} \nabla f(\theta_0)'\theta],$$

for a fixed θ_0. Finding the θ-values to solve (3.2) by Lagrange multipliers involves solving

$$(3.3) \qquad \nabla f(\theta_0) = \lambda \nabla g(\theta),$$

where λ is the Lagrange multiplier and $g(\theta) = (\theta-\hat{\theta})'C(\theta-\hat{\theta}) - b_\alpha$. Using (3.3) to solve for $(\theta-\hat{\theta})$ yields

$$(3.4) \qquad (\theta-\hat{\theta}) = 1/2 \; \lambda^{-1} \; C^{-1} \nabla f(\theta_0).$$

789

Substituting the right hand side of (3.4) into the equation $g(\theta) = 0$ and solving for λ yields

(3.5) $\lambda = \pm 1/2 \{b_\alpha^{-1} \nabla f(\theta_0)' C^{-1} \nabla f(\theta_0)\}^{1/2}$.

Since $g(\theta)$ is convex, all gradient vectors, $\nabla g(\theta)$, for $\theta \varepsilon \theta_\alpha^b$, are normal to the outside of θ_α^b, the direction of steepest ascent. It follows, then, that the λ-value associated with the minimum in (3.2) corresponds to the minus sign in (3.5), whereas the λ-value associated with the maximum corresponds to the plus sign.

Substituting the expression for λ in (3.5) into (3.4), and adding $\hat{\theta}$, yields the following iterative expressions,

(3.6) $\theta_{k+1} = \hat{\theta} \pm b_\alpha^{1/2} C^{-1} \nabla f(\theta_k) \{\nabla f(\theta_k)' C^{-1} \nabla f(\theta_k)\}^{-1/2}$,

for $k = 0,1,\dots$ The plus and minus signs correspond to the iterations for computing the maximum and minimum respectively. In practice, θ_0 can be taken as $\hat{\theta}$ since the first iteration will put θ_1 on θ_α^b. The (gradient-based) algorithmic maps corresponding to (3.6) will be denoted by $G_*(\theta)$ for the minimum and by $G^*(\theta)$ for the maximum. To refer to G_* and G^* simultaneously the generic symbol G will be used.

In practice, this author has never encountered a situation where G failed to converge. In theory, there can exist $f(\theta)$'s, as described in section 2, for which G_* (G^*) may produce an increased (decreased) value. However, it will be shown in the next section that by modifying G with stepsize adjustment, algorithmic convergence can be proved. In more standard iterative notation (3.6) can be written as

(3.7) $\theta_{k+1} = \theta_k + d_k$,

where $d_k = (\hat{\theta} - \theta_k) \pm b_\alpha^{1/2} C^{-1} \nabla f(\theta_k) \{\nabla f(\theta_k)' C^{-1} \nabla f(\theta_k)\}^{-1/2}$.
Here, d_k is interpreted as a direction vector. It takes θ_k a distance $\|d_k\|$ in the direction induced by d_k. To obtain a stepsize adjustment, (3.7) is modified to

(3.8) $\theta_{k+1} = \theta_k + sd_k$,

where $s \varepsilon (0,1]$. The stepsize modifications of G_*, G^*, and G are SG_*, SG^*, and SG respectively. It will be shown that there exists a $\delta \varepsilon (0,1]$ such that for any $s \varepsilon (0,\delta)$, the resulting $\theta_{k+1} = SG(\theta_k)$ will be an improved value on θ_α.

It should be noted however that SG may converge to a suboptimal, fixed point of the SG-map. This is not surprising in that no algorithm can guarantee convergence to a global, or even local, optimum unless strong conditions are imposed upon the objective function. It is

therefore recommended that if θ_α is not small, different starting values on θ_α^b should be tried other than $G(\hat{\theta})$. This can be achieved by replacing the initial gradient-direction vector, $\nabla f(\theta)$, by some other direction vector, v_0, at the first iteration. The resulting θ_1 value will fall different parts of θ_α^b depending on the relative orientation of θ_α^b and the hyperplane $v_0'\theta$.

4. CONVERGENCE. In this section a convergence theorem for SG∗ is proven. The proof SG* follows similarly. Following this a geometric interpretation of the convergence properties of G is presented.

The convergence proof in this section employs three important results in nonlinear optimization theory. We refer to these three results as: The alogrithmic map convergence theorem, the descent direction theorem, and the composite map theorem. Proofs for these theorems can be found in Bazaraa and Shetty (1979). Before presenting these results, we need to elaborate on the notion of an algorithmic map. An *algorithmic map* is a point-to-set map, A, that assigns to each point in the domain θ, a subset of points in θ. The map A is *closed* at a point $\theta \in \theta$ if $\theta_k \to \theta$ and $a_k \to a$, for $a_k \in A(\theta_k)$, imply $a \in A(\theta)$. If A is a point-to-point map, then continuity of A implies that A is closed. A is said to be closed on some $\theta_0 \subset \theta$ if it is closed at each point in θ_0.

ALOGRITHMIC MAP CONVERGENCE THEOREM. Let θ be a nonempty, closed set in $\mathbf{R}^p$, and let $\Omega \subset \theta$ be some solution set. Let $A:\theta \to \theta$ be an algorithmic map. Given $\theta_0 \in \theta$, suppose that A generates the sequence $\{\theta_k\}_{k=0}^\infty$ which is contained in a compact subset of θ. Suppose also that there is a continuous function $f(\theta)$ such that $f(\tilde{\theta}) < f(\theta)$ if $\theta \notin \Omega$ and $\tilde{\theta} \in A(\theta)$. If A is closed over the complement of Ω, then all the limit points of $\{\theta_k\}_{k=0}^\infty$ are in Ω and $f(\theta_k) \to f(\theta)$ for some $\theta \in \Omega$.

The solution set considered for SG∗ and G∗ in this paper is of the type $\Omega_* = \{\theta: \theta \in \theta_\alpha, \|\theta - G_*(\theta)\| \leq e\}$, where e is some small, positive, error tolerance. The solution set for SG* and G* is $\Omega^* = \{\theta: \theta \in \theta_\alpha, \|\theta - G^*(\theta)\| \leq e\}$. Both Ω_* and Ω^* contain the fixed points, θ_* and θ^*, of G∗ and G* respectively.

DESCENT DIRECTION THEOREM. If there is a vector d such that $\nabla f(\theta)'d < 0$ then there is a $\delta > 0$ such that $f(\theta + sd) < f(\theta)$ for all $s \in (0,\delta)$.

COMPOSITE MAP THEOREM. Suppose M_1 is a point-to-point map continuous at θ, and M_2 is a point-to-set map closed at $M_1(\theta)$. Then the composite map M_2M_1 is closed at θ.

THEOREM 4.1. The composite algorithmic map SG∗ (SG*) generates a sequence, $\{\theta_k\}_{k=0}^\infty$, with its limit point in the solution set Ω_* (Ω^*) containing θ_* (θ^*).

PROOF. First we show that SG∗ always generates points in the compact set θ_α. Then we show that this sequence of points on θ_α is strictly

791

decreasing. Given a point $\theta \in \Theta_\alpha$, the point-to-point map, G_*, generates a point $G_*(\theta)$ on Θ_α^b. Since Θ_α is a convex set, any linear stepsize interpolation between θ and $G_*(\theta)$ must lie in Θ_α. Hence, the composite map SG_*, always generates points in Θ_α.

Next, define the function,

$$h_k(\theta) = \nabla f(\theta_k)'(\theta - \theta_k).$$

Note that $\theta_{k+1} = G_*(\theta_k)$ minimizes $h_k(\theta)$ on Θ_α. Since $\nabla f(\theta) \neq 0$ on Θ_α, $h_k(\theta_{k+1}) < h_k(\theta)$ for all $\theta \in \Theta_\alpha$, not equal to θ_{k+1}. Since all θ_k's are in Θ_α, $h_k(\theta_{k+1}) < h_k(\theta_k) = 0$. But $h_k(\theta_{k+1}) < 0$ implies

$$\nabla f(\theta)'d_k < 0$$

since $d_k = \theta_{k+1} - \theta_k$ by (3.7). Thus by the Descent Direction Theorem for each k, there is a $\delta < 0$ such that

(4.1) $\qquad f(\theta_k + sd_k) < f(\theta_k)$

for some $s \in (0, \delta)$. Let S denote a closed line search map that generates at least one s satisfying (4.1) at each iteration for some s $\epsilon[\eta, 1]$. (S is a point-to-set map, where S is the set of s-values in $[\eta, 1]$ satisfying 4.1). Here η is some small, positive value less than δ. Any stepsize adjustment based on a linesearch, or simple stepsize halving, forms a closed algorithmic map if the search interval is closed and f is continuous (Bazaraa and Shetty, 1979, section 8.3). Since G_* is a continuous point-to-point map, it follows by the Composite Map Theorem that SG_* is a closed algorithmic map. Since SG_* generates a sequence of improving, points on the compact set, Θ_α, the Algorithmic Map Convergence Theorem implies that SG_* converges as stated. The proof for SG^* follows similarly.

Theorem 4.1 shows that G will converge if a stepsize adjustment is made. However, with examples used in practice G has never failed to converge quickly (3 to 4 iterations). At first this may seem curious since G is a constrained, steepest descent (ascent) mapping. The steepest descent algorithm is not generally considered to have good convergence properties. This consideration has arisen out of the tendency for steepest descent procedures to zig-zag back and forth along long, narrow valleys (or ridges) and thereby make slow progress toward the optimal point. (See Bazaraa and Shetty, 1979, Figure 8.14.) However, for most nonlinear regression models used in practice, $f(\theta)$ does not have such valleys or ridges on Θ_α. To see this note that for most regression models, $f(\theta)$ is monotonic in each θ_j on Θ_α. Without loss of generality, we may assume that $f(\theta)$ is increasing in each θ_j if it is monotonic in each θ_j. But if $f(\theta)$ is increasing in each θ_j on Θ_α, then $\nabla f(\theta)$ must lie in the nonegative orthant for all $\theta \in \Theta_\alpha$. Hence the angles between any two such gradient vectors cannot exceed $90°$. So, for most regression models, the contours of f cannot form long narrow valleys or ridges.

It is also reasonable to assume that G will converge quickly for a sufficiently large sample size. This is due simply to the fact that as the sample size increases the size of θ_α decreases, thus making the overall linear approximation to $f(\theta)$ on θ_α more accurate. The overall low curvature of the contours of f make it unlikely that a fixed point of G_* (G^*) will not be a global minimum (maximum) of $f(\theta)$ on θ_α, especially if θ_α is relatively small.

5. OTHER APPLICATIONS. Besides the obvious extensions of SG to weighted least squares regression and to Scheffe'-type simultaneous confidence intervals for several $f_i(\theta)$, there are other applications to confidence band estimation. Outside of the (nonlinear) mean response setting, there are some useful applications to simultaneous probability estimation. Such situations often arise from covariate dependent probability models. As discussed below, applications include multinomial logistic regression and various forms of covariate dependent reliability models. All of these models satisfy the original assumptions on $f(\theta)$ given in section 2.

Large sample confidence bands for the univariate-response, logistic regression probability model exist in closed form (Hauck, 1983). However, this is not the case for multinomial logistic regression. The probability model has the form,

(5.1) $$\exp(x'\beta_i) / \sum_{s=0}^{m} \exp(x'\beta_s)$$

for the ith out of (m+1) possible outcomes, conditional on the predictor variable x. Here each β_i is a vector, and $\beta_0 = 0$.

Heteroskedastic regression models are often used in econometrics. Some of these models allow the standard deviation of the error variable to be a linear (Rutemiller and Bowers, 1968) or log-linear (Harvey, 1976) function of predictor variables. To form a confidence band about the reliability, $Pr(Y > y^*/x)$ (with y^* fixed), we must work with $p(x) = 1 - F\{(y^*-x'\beta)/g(x'\gamma)\}$, where F is the distribution function of the standardized residuals and $g(\cdot)$ is the identity function or $\exp(\cdot)$. (In the former case, the estimated $x_j'\gamma$'s should, of course, not be near zero.) Since F is strictly increasing in applications, a confidence band about

(5.2) $$(y^*-x'\beta)/g(x'\gamma)$$

can be directly converted into a confidence band for p(x). The large sample confidence band for (5.2) will not generally exist in closed form. In this case SG can be used to search for the optimal (β, γ)-points to form the confidence bounds for (5.2) and convert them to bounds for p(x).

For parametric, or semi-parametric, proportional hazard models (Elandt-Johnson and Johnson, 1980), it is possible to get a relatively tractable expression for the conditional probability,

(5.3) $\Pr(\max\{T_0,\ldots,T_j\} < \min \{T_{j+1},\ldots,T_k\}/x)$,

provided j is small (Peterson, 1986). Here each T_i is a response time with hazard function

$$h_i(t) = h_B(t)\exp(x'\theta_i),$$

where $h_B(t)$ is the common baseline hazard and $\theta_0 = 0$. For j=0, (5.3) has the logistic form in (5.1). As in the previous two examples, SG can be used to search for optimal points on the corresponding θ_i^b in order to compute a confidence band for (5.3).

REFERENCES

Bates, D.M. and Watts, D.G. (1981), "Parameter Transformations for Improved Approximate Confidence Regions in Nonlinear Least Squares", Annals of Statistics, 9, 1152-1167.

Bazaraa, M.S. and Shetty, C.M. (1979), Nonlinear Programming: Theory and Algorithms, Wiley, New York.

Ben-Israel, A., Ben-Tal, A., Zlobec, S. (1981), Optimality in Nonlinear Programming: A Feasible Directions Approach, Marcel-Dekker, New York.

Bertsekas, D.P. (1982) Constrained Optimization and Lagrange Multipliers, Academic Press, New York.

Elandt-Johnson, R.C. and Johnson, N.L. (1980) Survival Models and Data Analysis, Wiley, New York.

Hamilton, D.C., Watts, D.G., and Bates, D.M. (1982) "Accounting for Intrinsic Nonlinearity in Nonlinear Regression Parameter Inference Regions", Annals of Statistics, 10, 386-393.

Hauck, W.W. (1983), "A Note on Confidence Bands for the Logistic Response Curve", The American Statistician, 37, 158-160.

Harvey, A.C. (1976), "Estimating Regression Models with Multiplicative Heteroscedasticity", Econometrica, 44, 461-465.

Khorasani, F. and Milliken, G. (1982), "Simultaneous Confidence Bands for Nonlinear Regression Models", Communications in Statistics-Theory and Methods, 11, 1241-1253.

Peterson, J.J. (1986), "Confidence Estimation of Some Treatment Comparison Probabilities for Response-Time Studies", Technical Report S-37, Department of Mathematics, Syracuse University.

Rao, C.R. (1973), Linear Statistical Inference and Its Applications, Wiley, New York.

Rutemiller, H.C. and Bowers, D.A. (1968), "Estimation in a Heteroscedastic Regression Model", Journal of the American Statistical Association, 68, 552-557.

SAS Institute Inc. (1985a), "MATRIX Procedure", Technical Report: P-135, Cary, North Carolina.

_____ (1985b), SAS/IML™ User's Guide (Version 5 ed.), Cary, North Carolina.

TESTING CURVE FIT

Royce Soanes
U.S. Army Armament, Munitions, and Chemical Command
Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

   ABSTRACT. A test is derived for the hypothesis of residual randomness
in a curve fit. The test is based on a binomial process where the main
problem involved is equivalent to finding the distribution of the number of
runs in a sequence of coin tosses. Although this distribution turns out to be
quite simple in form and easy to apply, its use in testing curve fit seems to
have been neglected in the statistical literature. We also obtain the more
general distribution of the number of runs in a sequence of throws of a multi-
faceted die, which is used to test sequences for randomness per se.

   I. INTRODUCTION. One question we wish to answer in this paper is:
"Will it do any good to add more parameters to my model in order to get a
better fit of the curve to the data?" The answer to this question is "yes" if
we can reject the hypothesis of residual randomness, and "no" if we cannot.
We will be concerned only with the signs of the residuals and not their magni-
tudes and we will examine the residual signs in a left to right manner. For
example, the residual sign sequence ++++-+++-----+ contains exactly five runs
of lengths 4, 1, 3, 5, and 1. The number of runs is equal to the number of
changes in sign plus one.

   If we assume (as our hypothesis of randomness) that the probability of
obtaining a positive (or negative) residual at any point is one-half, the
analysis of sign runs is equivalent to the analysis of runs of heads and tails
in coin flipping.

   Suppose, for example, that someone were to flip a coin for us one hundred
times. Our hypothesis is that this is done with a fair coin. Suppose also
that the first fifty flips come up heads and the second fifty flips come up
tails. What could we conclude from this? We could either conclude (1) that
an extremely unlikely event has taken place (why?), or (2) that the flipper
has flipped a two headed coin fifty times and then exchanged it for a two
tailed coin and flipped that fifty times. Statistics is based on the idea
that conclusion (2) is the shrewder of the two. In the context of residual
signs, we would just say that "the fit is obviously very poor." Saying this
another way, we reject the possibility that an extremely unlikely event has
taken place in favor of the possibility that our original hypothesis (on the
basis of which the probability of the unlikely event was computed) is false,
i.e., we reject the hypothesis.

II. RUN DISTRIBUTION.

First we need to answer the following question: "What is the probability of obtaining i runs in n flips of a coin?" We will call this probability r_i. We know from basic probability theory and the binomial distribution that the probability of obtaining i **heads** in n flips is given by

$$b_i = \binom{n}{i}/2^n \qquad (0 \leqslant i \leqslant n)$$

The interesting thing is that the probability of obtaining i **runs** in n flips may be obtained mnemonically by merely subtracting one from each variable parameter in b_i:

$$r_i = \binom{n-1}{i-1}/2^{n-1} \qquad (1 \leqslant i \leqslant n)$$

One feels that a result as simple and practical as this should be common knowledge. Is it? As is often the case, our result can be easily guessed by examining a special case. Take n = 5, and draw the tree of all possible outcomes.

1 2 3 2 3 4 3 2 3 4 5 4 3 4 3 2 2 3 4 3 4 5 4 3 2 3 4 3 2 3 2 1
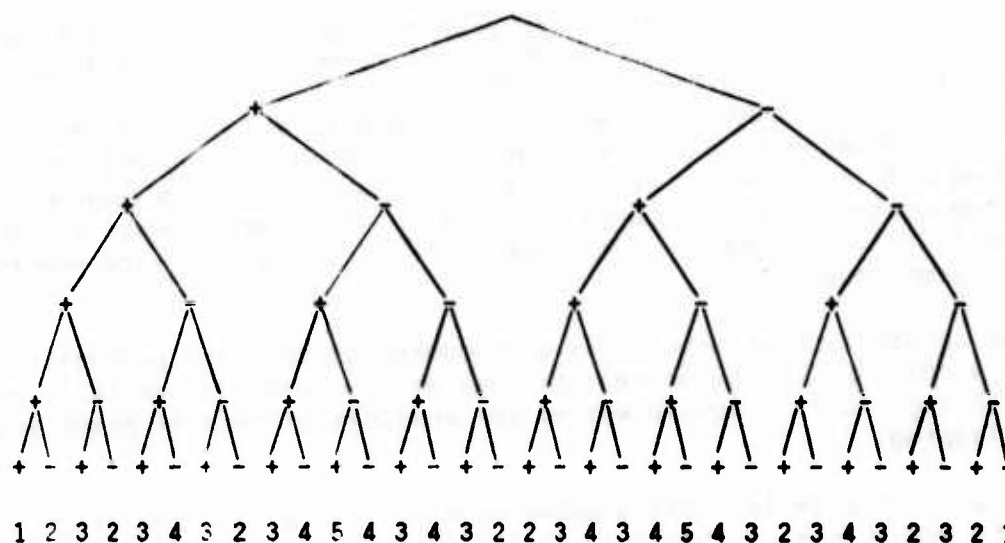
Also, label each leaf of the tree with the number of runs in the path leading to it. The probability weight of each path is 1/32. A leaf labeled k will be called a k leaf. We then simply count the number of k leaves (1 ≤ k ≤ 5) to obtain the entire probability distribution:

$$(n = 5)$$

i	r_i				
1	2/32	=	1/16	=	$\binom{4}{0}/2^4$
2	8/32	=	4/16	=	$\binom{4}{1}/2^4$
3	12/32	=	6/16	=	$\binom{4}{2}/2^4$
4	8/32	=	4/16	=	$\binom{4}{3}/2^4$
5	2/32	=	1/16	=	$\binom{4}{4}/2^4$

We therefore guess that:

$$r_i = \binom{n-1}{i-1}/2^{n-1}$$

in general.

Reasoning more rigorously (and recursively):

Suppose we have labeled the leaves of an n tree. We then want to extend the n tree to an n+1 tree by adding two leaves onto each previous leaf and labeling them. If a leaf labeled k (k leaf) has two more leaves added to it, one leaf will also be labeled k and the other will be labeled k+1. Note that each k leaf "grows" a k leaf and each k-1 leaf also grows a k leaf, but no other leaves grow k leaves. Therefore, if there are R_k^n k leaves in the n tree, there will be $R_k^n + R_{k-1}^n$ k leaves in the n+1 tree. We therefore have the recursion:

$$R_k^{n+1} = R_{k-1}^n + R_k^n$$

This is the same as the recursion for the binomial coefficients:

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$$

Again, we can easily guess the solution to the recursion as:

$$R_k^n = 2\binom{n-1}{k-1}$$

This checks in the recursion since

$$R_{k-1}^n + R_k^n = 2\binom{n-1}{k-2} + 2\binom{n-1}{k-1}$$

$$= 2\left(\binom{n-1}{k-2} + \binom{n-1}{k-1}\right)$$

$$= 2\binom{n}{k-1} = R_k^{n+1}$$

We also know that $R_1^n = 2$ and this checks also.

We therefore have:

$$r_i = R_i^n \cdot \frac{1}{2^n} = 2\binom{n-1}{i-1} \cdot \frac{1}{2^n}$$

$$= \binom{n-1}{i-1}/2^{n-1} \text{ for any } n \text{ and } 1 \leqslant i \leqslant n$$

III. THE TEST. The main idea behind the run test in curve fitting is that regardless of the fit, the number of positive residuals will usually be roughly the same as the number of negative residuals, while the number of runs will be _small_ for a bad fit and larger for a good fit. Technically speaking, a large number of runs violates randomness to the same extent that a _small_ number of runs does, but we do not apply the term "bad fit" to a case where the number of runs is unusually large. Our test is therefore a one-sided test. To answer the question of how small _small_ is, we must arbitrarily assign a significance level to the test. The significance level is simply the probability of the "unlikely event" (that the number of runs is _small_) based on the hypothesis of good fit. A high significance level (say 1/10) might be used if we want to be fairly liberal in adding more parameters to obtain a better fit. Lower significance levels can be used if we wish to be more conservative about adding more parameters.

The criterion for rejecting the fit is given by:

$$c_i \leqslant \alpha = \text{significance level}$$

where i is the observed number of runs and

$$c_i = \sum_{j=1}^{i} r_j$$

is the cumulative probability that the observed number of runs is at most i.

If we wanted to, we could compute tables relating various values of n, i, c_i, and α, but in this day of personal computers it hardly seems worth it - especially since the computation to test for good fit is so straightforward.

IV. ALGORITHM FOR TEST. We note in passing that

$$r_{i+1}/r_i = \frac{\binom{n-1}{i}}{2^{n-1}} \cdot \frac{2^{n-1}}{\binom{n-1}{i-1}}$$

$$= \frac{(n-1)!}{i!(n-1-i)!} \cdot \frac{(i-1)!(n-i)!}{(n-1)!}$$

$$= (n-i)/i$$

Therefore,

$$r_{i+1} = (n-i)r_i/i \quad , \quad r_1 = 1/2^{n-1}$$

This recursion gives us an efficient means of computing c_i.

Given n (the number of residuals), i (the number of runs), and α (the significance level), the test algorithm will end in one of two ways (as j is increased):

1. $c_j > \alpha$ and $j < i$ with no rejection of good fit hypothesis, or
2. $c_j < \alpha$ and $j = i$ with rejection of good fit hypothesis.

We may then just write down the test algorithm:

```
input: n,i,α

    c:=0
     ↓
    j:=1
     ↓
    r:=1.0/2.0**(n-1)
     ↓
→→c:=c+r
↑    ↓
↑  c>α and j<i?  →yes→  good fit → end
↑    ↓no
↑  c<α and j=i?  →yes→  bad fit → end
↑    ↓no
↑  r:=(n-j)*r/j
↑    ↓
←←j:=j+1
```

<u>**V. MOMENTS OF THE RUN DISTRIBUTION.**</u> We first obtain the moment generating function of the run random variable:

$$r_i = \binom{n-1}{i-1}/2^{n-1}$$

$$M(t) = E(e^{ti}) = \sum_{j=1}^{n} r_j e^{tj}$$

$$= \sum_{j=1}^{n} \frac{\binom{n-1}{j-1}}{2^{n-1}} e^{tj} = \sum_{j=0}^{n-1} \frac{\binom{n-1}{j}}{2^{n-1}} e^{t(j+1)}$$

$$= \frac{e^t}{2^{n-1}} \sum_{j=0}^{n-1} \binom{n-1}{j}(e^t)^j \cdot 1^{n-1-j}$$

Therefore,

$$M(t) = \frac{e^t}{2^{n-1}}(1+e^t)^{n-1}$$

$$M(0) = 1$$

$$M'(t) = \frac{e^t}{2^{n-1}}(1+e^t)^{n-1} + \frac{e^t}{2^{n-1}} \cdot (n-1)(1+e^t)^{n-2} \cdot e^t$$

$$= \frac{e^t}{2^{n-1}}(1+e^t)^{n-1}\left(1 + \frac{(n-1)e^t}{1+e^t}\right)$$

$$= M(t)\left(\frac{1+ne^t}{1+e^t}\right)$$

$$M'(0) = M(0)\left(\frac{1+n}{2}\right) = E(i)$$

Therefore

$$\mu = \frac{n+1}{2}$$

$$M''(t) = M'(t)\left(\frac{1+ne^t}{1+e^t}\right) + M(t)\left(\frac{ne^t(1+e^t) - e^t(1+ne^t)}{(1+e^t)^2}\right)$$

$$= M'(t)\left(\frac{1+ne^t}{1+e^t}\right) + \frac{M(t)e^t(n-1)}{(1+e^t)^2}$$

$$M''(0) = M'(0) \cdot \frac{n+1}{2} + \frac{n-1}{4}$$

$$= \mu^2 + \frac{n-1}{4} = E(i^2)$$

Therefore,

$$\sigma^2 = \frac{n-1}{4}$$

We therefore have the run mean and standard deviation:

$$\mu = \frac{n+1}{2}$$

$$\sigma = \frac{\sqrt{n-1}}{2}$$

   VI. ASYMPTOTIC RESULTS. For large n, we may further simplify our test by using approximate asymptotic formulas based on the normal approximation to our run distribution. If we think of our run random variable i as being continuous such that $f(x) = r_i$ for $i - \frac{1}{2} < x < i + \frac{1}{2}$, and we further approximate f by the normal density g, we may write

$$c_i = \sum_{j=1}^{i} r_j \approx \int_{-\infty}^{i+\frac{1}{2}} g(x)dx = G(i+\frac{1}{2})$$

$$= \Phi\left(\frac{i+\frac{1}{2}-\mu}{\sigma}\right)$$

where Φ is the standard normal distribution function:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}dt$$

803

Therefore, we have, approximately

$$c_i \simeq \phi\left(\frac{i+\frac{1}{2}-\mu}{\sigma}\right)$$

but

$$\mu = \frac{n+1}{2} \quad \text{and} \quad \sigma = \frac{\sqrt{n-1}}{2}$$

Therefore

$$c_i \simeq \phi\left(\frac{i+\frac{1}{2} - \frac{n+1}{2}}{\frac{\sqrt{n-1}}{2}}\right)$$

$$= \phi\left(\frac{2i+1 - (n+1)}{\sqrt{n-1}}\right)$$

$$= \phi\left(\frac{2i-n}{\sqrt{n-1}}\right)$$

Equivalently,

$$\frac{2i-n}{\sqrt{n-1}} = \phi^{-1}(c_i)$$

Therefore,

$$i = \frac{1}{2}(n + \phi^{-1}(c_i)\sqrt{n-1})$$

but

$$\phi^{-1}(c_i) = -\phi^{-1}(1-c_i)$$

Therefore, we have the simple asymptotic percentile

$$i_\alpha = \frac{1}{2}(n-\phi^{-1}(1-\alpha)\sqrt{n-1})$$

where we reject good fit on the α significance level if $i \leq i_\alpha$. For significance levels of 0.1, 0.01, and 0.001 we have from any table of the normal distribution:

$$\phi^{-1}(0.9) = 1.2816 \quad , \quad \phi^{-1}(0.99) = 2.3263 \quad \text{and} \quad \phi^{-1}(0.999) = 3.0902$$

804

Therefore,

$$i_{0.1} = \frac{1}{2}(n-1.2816\sqrt{n-1})$$

$$i_{0.01} = \frac{1}{2}(n-2.3263\sqrt{n-1})$$

and

$$i_{0.001} = \frac{1}{2}(n-3.0902\sqrt{n-1})$$

VII. GENERALIZATIONS. The previous formulas can be used to test for residual randomness in a curve fit if we are only concerned about the number of runs being too small. If, on the other hand, we are concerned about randomness per se, we must not only be concerned about the number of runs being too large, but we must go even further in our testing. Take the following sequence as an example:

HHTTHHTT...HHTT
(alternate pairs of heads and tails)

For a sequence such as this of even modest length, it becomes intuitively evident that the sequence is nonrandom; yet, our basic run test would declare it perfectly random because the actual number of runs is (almost) exactly equal to the expected number of runs.

How do we mathematically reconcile this fact with our intuition? We create a new sequence of composite symbols. In our example, we have used the symbol alphabet {H,T}. We now step up to the symbol alphabet {HH, HT, TH, TT}. We could say, for instance, that

a = HH

b = HT

c = TH

d = TT

Our example sequence would then be:

adadad...ad

a sequence whose randomness could easily be rejected by our basic run test (for too many runs as opposed to too few runs). We could also take triples or quadruples as our composite symbols. In the case of quadruples, our example sequence would obviously consist of only one symbol (too few runs).

It should now be clear that the next basic question we want to answer is: "What is the probability of obtaining i runs in n trials where each trial has m equally likely outcomes?" Drawing a multiway tree and counting will not suffice to answer this question, but reasoning recursively will. Imagine an m way tree representing n trials with its leaves labeled with the number of runs in the path leading to each leaf. We now want to grow m leaves on each leaf of the n tree in order to obtain the n+1 tree. Consider a leaf labeled i (i leaf) in the n tree. Regardless of what symbol this leaf represents, it will grow exactly one i leaf and it will grow exactly m-1 i+1 leaves. By the same token, a leaf labeled i-1 in the n tree will grow one i-1 leaf and m-1 i leaves. No other leaves of the n tree grow i leaves. Since there are R_i^n i leaves and R_{i-1}^n i-1 leaves in the n tree, the number of i leaves in the n+1 tree is given by the recursion:

$$R_i^{n+1} = R_i^n + (m-1)R_{i-1}^n$$

It is clear that we want

$$R_1^n = m$$

because there are m sequences with only one run.

Examining the recursion for various values of n and i enables us to once again guess the solution:

$$R_i^n = m(m-1)^{i-1}\binom{n-1}{i-1}$$

Checking this solution in the recursion:

$$R_i^n + (m-1)R_{i-1}^n = m(m-1)^{i-1}\binom{n-1}{i-1}$$

$$+ (m-1)m(m-1)^{i-2}\binom{n-1}{i-2}$$

$$= m(m-1)^{i-1}(\binom{n-1}{i-1} + \binom{n-1}{i-2})$$

$$= m(m-1)^{i-1}\binom{n}{i-1} = R_i^{n+1}$$

Also,

$$R_1^n = m(m-1)^0\binom{n-1}{0} = m$$

and we see that our guessed solution is indeed correct.

806

Since the total number of paths in the n tree is m^n, we have:

$$r_i = R_i^n/m^n = \binom{n-1}{i-1}(m-1)^{i-1}/m^{n-1}$$

= Prob (i runs in n throws of an m faceted die)

As before, it is a trivial matter to obtain the recursion for r_i to be used in computing the cumulative distribution function c_i.

$$r_{i+1} = (m-1)(n-i)r_i/i$$

$$r_1 = 1/m^{n-1}$$

$$c_{i+1} = c_i + r_{i+1}$$

$$c_1 = r_1$$

The moment generating function may be obtained in a manner similar to the m = 2 case:

$$M(t) = E(e^{ti}) = \sum_{j=1}^{n} e^{tj}\binom{n-1}{j-1}(m-1)^{j-1}/m^{n-1}$$

$$= \sum_{j=0}^{n-1} e^{t(j+1)}\binom{n-1}{j}(m-1)^{j}/m^{n-1}$$

$$= \frac{e^t}{m^{n-1}} \sum_{j=0}^{n-1} \binom{n-1}{j}((m-1)e^t)^j \cdot 1^{n-1-j}$$

$$= \frac{e^t}{m^{n-1}} (1 + (m-1)e^t)^{n-1}$$

Using the first and second derivatives of M at t = 0 gives us the mean and standard deviation of the number of runs.

$$\mu = \frac{n(m-1) + 1}{m}$$

$$\sigma = \frac{\sqrt{(n-1)(m-1)}}{m}$$

807

An asymptotic normal approximation to c_i is therefore given by:

$$c_i \approx \phi \left(\frac{i+\frac{1}{2}-\mu}{\sigma} \right)$$

$$= \phi \left(\frac{(m-1)(i-n+\frac{1}{2}) + i-\frac{1}{2}}{\sqrt{(n-1)(m-1)}} \right)$$

Another question one might ask is: "What is the most probable number of runs (the mode) in n throws of an m faceted die?" Using the recursion formula for r_i, it is easily concluded that: if m divides n(m-1) exactly, we have two modes, i and i+1, where i = n(m-1)/m; and if m does not divide n(m-1) exactly, we have one mode i, where i = the greatest integer less than n(m-1)/m+1.

We can use these modal values of the run distribution to define the idea of "run perfect" superrandom sequences. That is, if the basic symbol sequence and all composite symbol sequences up to a given composite symbol length l each have the modal number of runs, we say that the original sequence is run perfect up to this l value.

In [1], the authors display a page of H's and T's and ask whether this sequence represents the flips of a fair coin. They do not answer the question, but the answer is: very probably not, because the probability of obtaining the given number of runs or fewer is less than three chances in one thousand. As a complement to [1], we display the following sequence of 1200 flips which is demonstrably more random (after the fact) than most such sequences produced by a real coin, because the sequence is run perfect for composite symbols of up to length ten.

REFERENCES

1. P. J. Davis and R. Hersh, The Mathematical Experience, Houghton Mifflin Company, Boston, 1982.

808

ON THE ESTIMATION OF SOME NETWORK PARAMETERS IN

THE PERT MODEL OF ACTIVITY NETWORKS

Salah E. Elmaghraby

Graduate Program in Operations Research
North Carolina State University, Raleigh, NC

The PERT model of activity networks[3] (ANs), which dates to the later half of 1957, represents a landmark development in the theory and practice of operations research for several reasons, not the least important of which is that it represented the first attempt at the explicit recognition of randomness in the duration of activities within the context of project planning and control. The model's almost instantaneous popularity, which was bolstered by requiring all tenders to the DoD to be couched in the vernacular of PERT, invited theoreticians to take a closer look at the model's constructs, and they found them lacking.

The critique of the assumptions and derivations of the original PERT model are presented in Chapter 4 of Elmaghraby's book [1]. There one can also find description of some early attempts at rectification relative to the estimation of the expected duration of the project, and to the estimation of its probability distribution function (pdf).

Briefly, the main line of criticism to the estimation fo the pdf of the completion time of the project runs as follows. The duration of the project is the time of realization of its last event. Now, assuming the nodes of the network to be numbered sequentially from 1 to n in the 'acitivity-on-arc' mode of representation, the last event is node n and its time of realization is

denoted by τ_n. Evidently, it is a random variable (rv) equal to the maximum of a finite number of rv's, each representing the duration of a path from the start node (node 1) to the terminal node (node n). These paths are not independent because they usually share activities. Even if they were considered 'approximately' independent, their durations are only approximately normally distributed. But the time of realization of node n is definitely not normally distributed. In fact, under the assumption of independence, the pdf of τ_n is the product of the individual path pdf's, which is known to converge to the step function as the number of paths grows without bound. Finally, even if we are willing to 'approximate' the pdf of τ_n by a normal distribution, it should be with a different mean and different variance from that suggested by PERT!

This talk is based on the paper with the same title by Elmaghraby, ref.[2], which should be consulted for a more detailed exposition of the concepts outlined here.

The introduction of uncertainty in the duration of the activities has enriched the field with new concepts which came into being in response to a variety of questions. These latter may be viewed as the probabilistic counterparts to their deterministic equivalents. For instance, since (almost) any path may be the critical path (cp), in the sense of being the longest path in a realization of the project, it is meaningless to inquire of the cp, but it is meaningful to inquire of the probability that a particular path is the cp. This gave rise to the concept of the 'criticality index' of a path, and subsequently to the concept of the criticality index of an activity. The issue of the duration o the project is now re-cast into the determination of the pdf of τ_n or, for that matter, the pdf of τ_j for any node j in the network. Interestingly enough, research in the approximation of the pdf gave rise to the

810

concept of 'network reducibility', which is of both theoretical as well as practical significance. Finally, the probabilistic counterpart of the critical subnetwork (i.e., the set of cp's) in the CPM model is the set of minimum number of paths whose criticality index is at least α, $0 < \alpha < 1$. Alternatively, a set of K paths are referred to as the 'topmost K-cp's' when their criticality index is the maximum among all sets of K paths in the network.

The approaches to the solution of the problems posed adopt one of the following four avenues: the analytical determination of the exact value(s), analytical approximation, analytical bounding, and estimation *via* Monte Carlo sampling (MCS). The difference among the various approaches is almost evident from their names, but a word of clarification is still in order.

There is no substitute for the analytical determination of the exact value(s) that is sought. However, it is not always possible to achieve that lofty objective, or it is possible but not economical in effort. Then approximation is admissible. There are two analytical approches to achieve such approximation, in addition to the approach *via* MCS. The choice among these options is a matter of taste; it is also a matter of the requirements of the analysis. For, sometimes analytical approximation may give no clue to the error committed. If the magnitude of the error must be controlled, then analytical bounding may be the only route open to the analyst. Finally, recall that MCS does not bound the error, but only the probability of committing an error of specified magnitude.

The accompanying diagram gives a global view of the methodologies adopted in the resolution of this problem, and represents a synopsis of the talk presented at the conference.

811

REFERENCES

[1] Elmaghraby, S.E. (1977). *Activity Networks: Project Planning and Control by Network Models*, J. Wiley & Sons, N.Y., NY.

[2] _____, (1985). "The estimation of some network parameters in the PERT model of activity networks: Review and critique", OR Report No. 207, NCSU, Raleigh, NC 27695-7913.

[3] Malcolm, D.G., J.H. Roseboom, C.E. Clark, and H. Fazar (1958). "Application of a technique for research and development evaluation", *Oper. Res. 7*, 646-669.
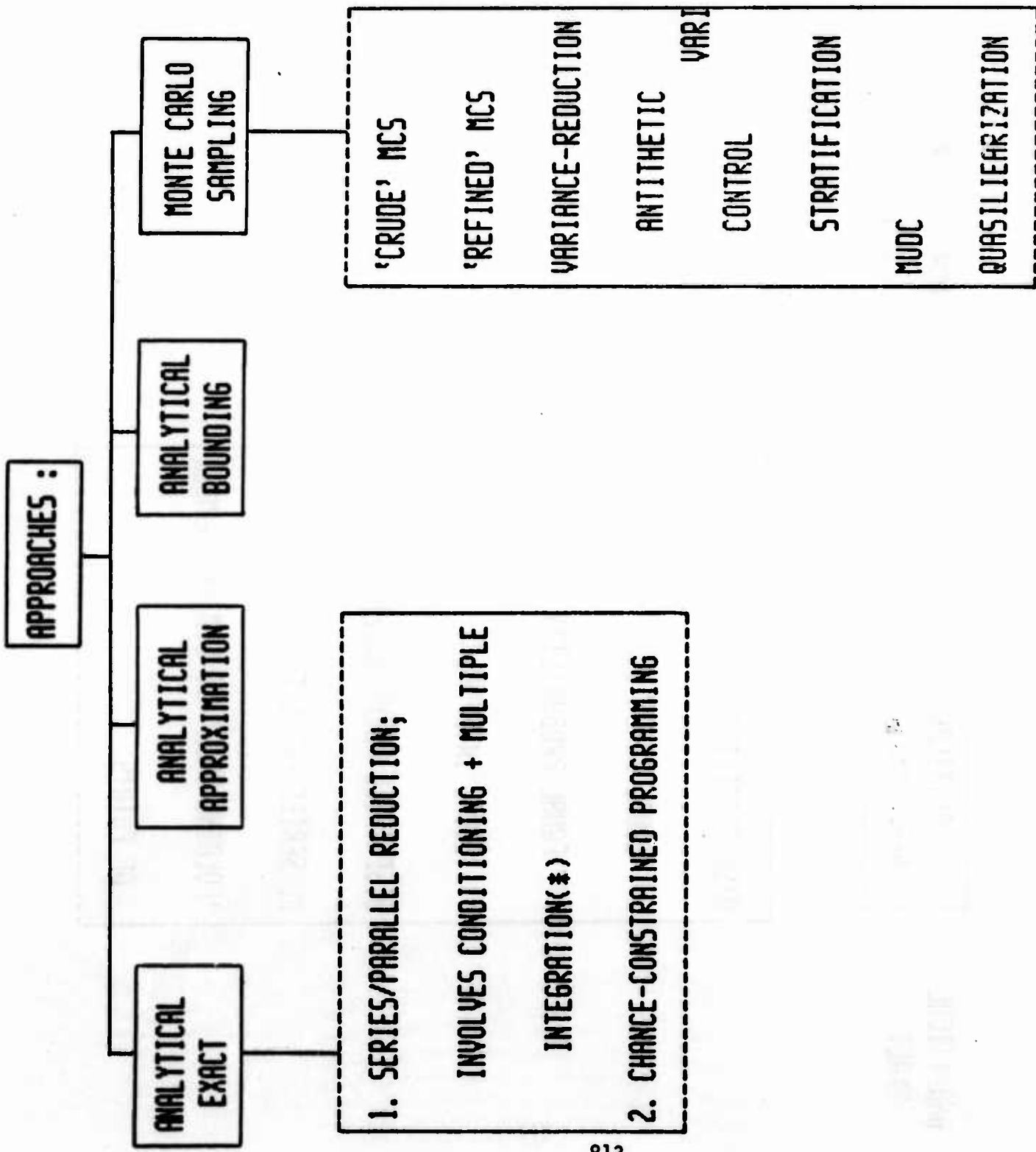
ANALYTICAL
EXACT

ANALYTICAL
APPROXIMATION

ANALYTICAL
BOUNDING

MONTE CARLO
SAMPLING

DISCRETIZE THE PDF:

EQUAL MOMENTS

EQUAL PROBABILITY

EQUAL INTERVALS

DISENTANGLE PATHS DEPENDENCY(*)

DO SERIES/PARALLEL REDUCTION,

'FOLDING BACK' TO LIMITED NUMBER

OF POINTS

814

ANALYTICAL
EXACT

ANALYTICAL
APPROXIMATION

ANALYTICAL
BOUNDING

MONTE CARLO
SAMPLING

PDF: BASED ON IGNORING DEPENDENCE

AMONG PATHS

BOUNDS ON $E(\Upsilon_n)$:

IGNORING DEPENDENCE

USING SURROGATE PDF

SOLIDIFICATION AND MELTING WITH
INTERFACIAL ENERGY AND ENTROPY

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

ABSTRACT. The classical theory of Stefan for solidification and melting is too simplistic to describe phenomena such as super-cooling, superheating, and the formation of dendrites. Recently, a general theory was developed for phenomena of this type; this paper describes that theory, a theory based on general thermodynamical laws which are appropriate to a continuum and which include contributions of energy and entropy for the interface between phases.

1. <u>INTRODUCTION</u>. A classical problem of mathematical physics is the Stefan problem for the melting of a solid or the freezing of a liquid. The underlying theory, however, is too simplistic to describe phenomena such as supercooling, in which a liquid supports temperatures below its freezing point, or superheating, the analogous phenomenon for solids, or dendrite formation, in which simple shapes, such as spheres, evolve to complicated tree-like structures.[1] The past two decades have seen the development of more general theories[2] for phenomena of this type, a critical ingredient being a free-boundary condition at the solid-liquid interface $I = I(t)$ in which the temperature depends on the curvature of I. In these theories questions arise as to what are the interface conditions;[3] in fact, it is not clear which of the interface conditions are constitutive assumptions and which follow directly from the underlying balance laws.

Here we shall discuss a recent paper of Gurtin [1986]. That paper develops a theoretical framework for theories of the above type starting from general thermodynamical laws which are appropriate to a continuum and which include interfacial contributions for both energy and entropy.

2. <u>GENERAL RESULTS</u>. The chief assumptions - apart from general equations of state for the bulk and interfacial quantities - are that the interface I produce no entropy and that the temperature be continuous across I. Among the main results are the interface conditions[4]

---

[1] Cf. Chalmers [1964] and Delves [1974] for discussions of these phenomena.

[2] Mullins [1960], Mullins and Sekerka [1963,1964], Voronkov [1965]. See also the review articles by Sekerka [1968,1973,1984], Chernov [1972], Delves [1974], and Langer [1980].

[3] Cf. Rogers [1983] for a discussion of some of the inconsistencies in the literature.

[4] Cf. Moeckel [1975], Fernandez-Diaz and Williams [1979], and Wollkind [1979] for the first relation in (1).

$$[q] \cdot m = v[E] - v\kappa e - e^{\Delta} \quad \text{on} \quad I,$$

$$T = ([E] - \kappa e)/([S] - \kappa s) \quad \text{on} \quad I, \qquad (1)$$

$$vm \cdot n = 0 \quad \text{on} \quad \partial I,$$

in which T is the temperature; $[E]$, $[S]$, and $[q]$ are the jumps[5] in energy, entropy, and heat flux across the interface; e and s are the interfacial values of energy and entropy; κ, v, and m, respectively, are, for the interface, the sum of principal curvatures, the normal velocity, and the unit normal vector (outward relative to the region occupied by phase 1); e^{Δ} is the time derivative of e following the interface; n is the outward unit normal on the boundary of the region B occupied by the body.

The first of (1) is essentially the first law of thermodynamics at the interface. The second - derived within the fully dynamical theory - is a condition of local equilibrium expressing balance of free-energy across the interface. The third is a contact condition for that portion of the interface which intersects the boundary of B; it asserts that - where the interface meets ∂B - it is orthogonal to ∂B or stationary.

Two types of boundary conditions are discussed: an isolated boundary on which $q \cdot n = 0$; an isothermal boundary on which T is constant. It is shown that, for either of these boundary conditions,

$$\underline{\text{interfacial}} \ \underline{\text{area}} \ \text{is} \ \underline{\text{uniformly}} \ \underline{\text{bounded}} \ \underline{\text{in}} \ \underline{\text{time}}, \qquad (2)$$

at least when B is bounded.

3. **EQUILIBRIUM THEORY.** For isothermal boundary conditions stable states are defined as minimizers of a global free-energy. It is assumed that the bulk free-energies cross at a single temperature T_M; it follows that - for bounded B - stable states are always single phase,[6] the stable phase being the phase with lower free-energy. T_M therefore represents the temperature at which a change in stable phase occurs, and, for that reason, is referred to as the transition temperature.

---

[5] Our convention for jumps and for the latent heat L is "phase 2 minus phase 1", with phases labeled so that $L \geq 0$. Thus for a solid-liquid system phase 2 would denote the liquid.

[6] Here it is important to emphasize that the boundary is held at constant temperature; two-phase solutions are possible when, for example, the body is isolated and the total energy constrained.

The question of stability for unbounded B is far more interesting There the results, expressed in terms of a solid-liquid system in isothermal equilibrium, assert that:[7]

(i) There are no stable states in which the bounded phase is solid and the unbounded phase supercooled liquid.

(ii) Under the conditions of (i), minimizing sequences of the free-energy are consistent with interfacial instabilities such as the formation of complicated arrays of thin spikes, behavior indicative of dendritic growth.

4. QUASI-STATIC THEORY. A quasi-static model is developed for situations in which the interface moves slowly compared with the time scale for heat conduction. The chief constitutive hypothesis underlying this model is that - in each of the phases - both the bulk energy and the bulk entropy are constant. It is also assumed that the conductivities k_i, the interfacial energy e, and the interfacial entropy s are constant. Let $B_i(t)$ denote the subregion of B occupied by phase i $(i = 1,2)$, and let

$$u = T - T_M.$$

Then these assumptions lead to the system[8]

$$\Delta u = 0, \qquad q = -k_i \operatorname{grad} u \qquad \text{in } B_i,$$

$$u = -h\kappa/(1-a\kappa), \quad [q] \cdot m = (L-\kappa e)v \qquad \text{on } I, \qquad (3)$$

$$vm \cdot n = 0 \qquad \text{on } \partial I,$$

where

$$h = T_M f(T_M)/L, \quad a = T_M s/L \qquad (4)$$

with $f(\cdot)$ the interfacial free-energy and L the latent heat. Global growth-conditions are established for (3) under the two types of boundary conditions discussed previously. In particular, letting

$$\mu = T_M f(T_0)/L, \quad \beta = e/L,$$

---

[7]In their paper of [1963], Mullins and Sekerka, working with the dynamical theory described by (13), established the instability of the interface or infinitesimal perturbations of a sphere solidifying in a supercooled melt. The assertions (i) and (ii) are analogs, within the equilibrium theory, of the Mullins-Sekerka instability.

[8]grad, div, and Δ are the gradient, divergence, and Laplacian operators; for $F = F(t)$ and $f = f(x,t)$, $F' = dF/dt$ and $f' = \partial f/\partial t$; vol($\cdot$) and area($\cdot$) denote the volume and area measures.

819

it is shown that:

(iii)   for an isolated boundary,

$$\text{vol}(B_2)^{\cdot} + \beta\,\text{area}(I)^{\cdot} = 0, \qquad \text{area}(I)^{\cdot} \leq 0; \tag{5}$$

(iv)   for an isothermal boundary ($u = u_0 = \text{constant on } \partial B$),

$$u_0\,\text{vol}(B_2)^{\cdot} + \mu\,\text{area}(I)^{\cdot} \leq 0. \tag{6}$$

The results (5) and (6) seem to indicate (asymptotically as $t \to \infty$) interfacial instabilities such as those described in (i) and (ii), as well as an instability characterized by a solid phase whose volume tends to zero, but whose interfacial area does not. This phenomenon is referred to as the formation of a dendrite with null volume.

5.   THEORIES BASED ON THE CAPILLARITY RELATION. Thus far no assumptions have been made concerning the size of the interfacial quantities; even though the hypotheses underlying (3) are strong, the theory is exact in the sense that the underlying equations are fully compatible with the first two laws of thermodynamics.

Consider now the general relations (1), but in situations for which interfacial energy and entropy are small. Then, to within terms of higher order in these quantities,

$$[q] \cdot m \approx Lv,$$
$$u \approx -h\kappa, \tag{7}$$

with h as defined in (4). The relations (7) are central to the modern work on solidification.[9]

A model is developed based on the interface conditions (7) in conjunction with assumptions of constant specific heats and constant conductivities. These assumptions lead to the equations

$$C_i u^{\cdot} = -\text{div}\,q, \qquad q = -k_i\,\text{grad}\,u \qquad \text{in } B_i$$
$$u = -h\kappa, \qquad\qquad [q] \cdot m = Lv \qquad \text{on } I, \tag{8}$$
$$vm \cdot n = 0 \qquad\qquad\qquad\qquad \text{on } \partial I.$$

The assumption

$$C_1 = C_2 \tag{9}$$

is common in the literature; granted (9), the following growth conditions follow from (8):

---

[9] Cf. the references cited in Footnote 2.

820

(v) for an isolated boundary,

$$\{vol(B_2) + CVu_m\}^{\cdot} = 0,$$

(10)

$$\{h\, area(I) - u_m\, vol(B_2) + (C/2) \int_B (u-u_m)^2\}^{\cdot} \leq 0;$$

(vi) for an isothermal boundary,

$$\{h\, area(I) - u_0\, vol(B_2) + (C/2) \int_B (u-u_0)^2\}^{\cdot} \leq 0.$$

(11)

Here $V = vol(B)$, $C = C_i/L$, and u_m is the mean value of u.

A standard model for solidification follows from (8) when the terms $C_i u^{\cdot}$ are neglected:

$$\Delta u = 0, \quad q = -k_i\, grad\, u \quad in\ B_i,$$

$$u = -h\kappa, \quad [q]\cdot m = Lv \quad on\ I,$$

(12)

$$vm \cdot n = 0 \quad on\ \partial I.$$

Here (10) and (11) are replaced by:

(vii) for an isolated boundary,

$$vol(B_i)^{\cdot} = 0, \quad area(I)^{\cdot} \leq 0;$$

(13)

(viii) for an isothermal boundary,

$$u_0\, vol(B_1)^{\cdot} + h\, area(I)^{\cdot} \leq 0.$$

(14)

REFERENCES.

[1960] Mullins, W. W., Grain boundary grooving by volume diffusion, Trans. Metall. Soc. AIME 218, 354-361.

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

[1964] 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] Chalmers, B., Principles of Solidification, Wiley, New York.

[1965] Voronkov, V. V., Conditions for formation of mosaic structure on a crystallization front, Sov. Phys. Solid State $\underset{\sim}{6}$, 2378-2381.

[1968] Sekerka, R. F., Morphological stability, J. Crystal Growth $\underset{\sim}{3},\underset{\sim}{4}$, 71-81.

[1972] Chernov, A. A., Theory of the stability of face forms of crystals, Sov. Phys. Crystallog. $\underset{\sim}{16}$, 734-753.

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

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

[1975] Moeckel, G. P., Thermodynamics of an interface, Arch. Rational Mech. Anal. $\underset{\sim}{57}$, 255-280.

[1979] Fernandez-Diaz, J. and W. O. Williams, A generalized Stefan condition, Zeit. Angew. Math. Phys. $\underset{\sim}{30}$, 749-755.

[1979] Wollkind, D. J., A deterministic continuum mechanical approach to morphological stability of the solid-liquid interface, Preparation of Properties of Solid State Materials, (ed. C. R. Wilcox), Dekker, New York.

[1980] Langer, J. S., Instabilities and pattern formation in crystal growth, Rev. Mod. Phys. $\underset{\sim}{52}$, 1-27.

[1983] Rogers, J.C.W., The Stefan problem with surface tension, Free Boundary Problems: Theory and Applications, (eds. A. Fasano, M. Primicerio), Pitman, Boston.

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

[1986] Gurtin, M. E., On the two-phase Stefan problem with interfacial energy and entropy, Arch. Rational Mech. Anal. Forthcoming.

NUMERICAL COMPUTATION OF THE APPROXIMATE ANALYTICAL
SOLUTION OF A STEFAN'S PROBLEM IN A FINITE DOMAIN

Shunsuke Takagi
U.S. Army Corps of Engineers
Cold Regions Research and Engineering Laboratory
Hanover, NH 03755-1290

ABSTRACT. The approximate analytical solution of Stefan's problem in
a finite domain with constant boundary and initial conditions was found and
reported last year. This year we report the numerical computation of the
analytical solution. We start with the presentation of the temperature
solutions, which are easily verifiable. The interfacial position is deter-
mined by solving a complicated nonlinear equation composed of a summation
of transcendental functions, which we describe in detail.

I. THE ANALYTICAL SOLUTION. We consider the simplest freezing prob-
lem in a finite domain $0 \le x \le \ell$. The boundary temperature T_A at $x = 0$
and T_B at $x = \ell$ are constant, the latter being also the initial tempera-
ture. The freezing temperature is T_F.

At $t = 0$, a new phase emerges at $x = 0$, whose temperature we express
by $T_I(x, \kappa_I t)$, where κ_I is the thermal diffusivity of the new phase. The
domain of the new phase is $0 \le x \le s(t)$, where initially $s(t) = s_0\sqrt{t}$ and
finally $s(\infty) = $ constant, s_0 being a constant. We express the temperature
of the old phase by $T_{II}(x, \kappa_{II} t)$, where κ_{II} is the thermal diffusivity of
the old phase. The domain of the old phase is $s(t) \le x \le \ell$. The quanti-
ties of the new and old phases are designated by the roman numerals I and
II, respectively, used as a sub- or superindex. The conditions to be
satisfied are:

$$T_I(0, \kappa_I t) = T_A \ , \tag{1}$$

$$T_I\big(s(t), \kappa_I t\big) = T_{II}\big(s(t), \kappa_{II} t\big) = T_F \ , \tag{2}$$

$$K_I \frac{\partial T_I}{\partial x}\bigg|_{s(t)} - K_{II} \frac{\partial T_{II}}{\partial x}\bigg|_{s(t)} = L\rho \frac{ds}{dt} \ , \tag{3}$$

$$T_{II}(\ell, \kappa_{II} t) = T_B \ , \tag{4}$$

and

$$T_{II}(x,0) = T_B \quad \text{for} \quad 0 < x < \ell \ . \tag{5}$$

The approximate analytical solution we have found [1] is as follows:

The new and old phase temperatures are given by

$$T_I(x, \kappa_I t) = T_A + (T_F - T_A) \text{ Erf } \frac{x}{\sqrt{4\kappa_I t}} / \text{ Erf } \frac{s(t)}{\sqrt{4\kappa_I t}} \qquad (6)$$

for $0 \leq x \leq s(t)$, and

$$T_{II}(x, \kappa_{II} t) = T_B - \frac{T_B - T_F}{R_N} \sum_{n=0}^{N} \left[\text{erfc} \frac{2n\ell + x}{\sqrt{4\kappa_{II} t}} - \text{erfc} \frac{(2n+2)\ell - x}{\sqrt{4\kappa_{II} t}} \right] , \qquad (7)$$

for $s(t) \leq x \leq \ell$, respectively, where

$$R_N = \sum_{n=0}^{N} \left[\text{erfc} \frac{2n\ell + s(t)}{\sqrt{4\kappa_{II} t}} - \text{erfc} \frac{(2n+2)\ell - s(t)}{\sqrt{4\kappa_{II} t}} \right] . \qquad (8)$$

Substituting (6) and (7) into (3), we find the equation for the determination of $s(t)$,

$$\frac{K_I(T_F - T_A)}{\sqrt{4\kappa_I}} \frac{i^{-1}\text{erfc}\left(s(t)/\sqrt{4\kappa_I t}\right)}{\text{Erf}\left(s(t)/\sqrt{4\kappa_I t}\right)} - \frac{K_{II}(T_B - T_F)}{\sqrt{4\kappa_{II}}} \frac{Q_N}{R_N} = \frac{L\rho \cdot s(t)}{\sqrt{4t}} , \qquad (9)$$

where

$$Q_N = \sum_{n=0}^{N} \left[i^{-1}\text{erfc} \frac{2n\ell + s(t)}{\sqrt{4\kappa_{II} t}} + i^{-1}\text{erfc} \frac{(2n+2)\ell - s(t)}{\sqrt{4\kappa_{II} t}} \right] . \qquad (10)$$

It is obvious that the temperature solutions (6) and (7) satisfy all the assigned conditions but not the differential equation of the heat conduction, unless $N = 0$, $s(t) = s_0\sqrt{t}$ and the second summand in R_N is negligible. In other words, the solution is exact at the initial stage, where the semi-infinite domain solution is applicable, but is approximate beyond the initial stage. The approximate solution, however, approaches the exact one as the limit as t and N increase indefinitely together, as proved in [1]. We may remark that if we adopt a hypothetical procedure that t in $s(t)$ and R_N are an extraordinary parameter that may be kept constant during the time differentiation, the solution is always exact.

The summation in (7), (8) and (10) follow a convention. Because the magnitudes of erfc x and i^{-1}erfcx are governed by $\exp(-x^2)$, we apply the same convention to all the summations. We explain the motivation for the convention here. It is formalized later.

Initially the second summand in the 0th (i.e. n=0) bracket and the dual summands in all the subsequent brackets are less than some small number 10^{-m}, where m is a positive number. These summands are in this case regarded negligible. The temperature distributions then reduce to that of a semi-infinite domain. We call 10^{-m} the threshold number, and m the threshold power.

At an appropriate time, called the first lead time, the second summand in the 0th bracket exceeds the threshold value. We then add it, completing the 0th bracket. At the second lead time, the first summand in the 1st (i.e. n=1) bracket exceeds the threshold value. Summands appear successively in this way and the brackets are completed successively. We let mainly m = 10 in our numerical computation.

To describe the successive emergence of latent summands, we use a sequence of lead times as a parameter. The parametric lead time grows to infinity together with N.

II. INTERFACIAL COORDINATES. Introducing the nondimensional interfacial coordinates ξ and η by

$$\xi = s(t)/\ell \tag{11}$$

and

$$\eta = (4\kappa_{II}t)^{1/2}/\ell \tag{12}$$

and defining

$$U_k(\xi, \eta) = \sum_{n=0}^{N} \left[i^{-k}\text{erfc}\,\frac{2n+\xi}{\eta} + (-1)^{k-1} \cdot i^{-k}\text{erfc}\,\frac{2n+2-\xi}{\eta} \right] \tag{13}$$

for integers $k \geq 0$, we rewrite the transcendental equation (9) to

$$W(\xi, \eta) = \text{Erf}(\beta\xi/\eta) - (2a/\sqrt{\pi})\exp\left(-(\beta\xi/\eta)^2\right) \cdot \tag{14}$$

$$\cdot\, U_0(\xi, \eta)/\left\{(b\xi/\eta)U_0(\xi, \eta) + U_1(\xi, \eta)\right\} = 0 \,,$$

where

$$\beta = (\kappa_{II}/\kappa_I)^{1/2} \tag{15}$$

$$a = \beta(K_I/K_{II})(T_F - T_A)/(T_B - T_F) \tag{16}$$

and

$$b = 2L\rho\kappa_{II}/\left(K_{II}(T_B - T_F)\right) \,. \tag{17}$$

825

The domains of ξ and η in $W(\xi, \eta)$ are:

$$0 \leq \xi < \xi_\infty \tag{18}$$

and

$$0 \leq \eta < \infty , \tag{19}$$

where

$$\xi_\infty = s(\infty)/\ell , \tag{20}$$

at which η becomes infinite. $s(\infty)$ is found by solving

$$\frac{K_I(T_F - T_A)}{s(\infty)} - \frac{K_{II}(T_B - T_F)}{\ell - s(\infty)} = 0 , \tag{21}$$

an expression of the linearity of the final temperature distribution. The derivatives of $U_k(\xi, \eta)$ are given by

$$\eta \cdot \partial U_k/\partial \eta = -U_{k+1} , \tag{22}$$

$$\eta \cdot \partial U_k/\partial \eta = k \, U_k + \frac{1}{2} U_{k+2} . \tag{23}$$

The derivatives W_ξ and W_η are given by

$$\frac{\sqrt{\pi}}{2} \frac{\eta}{\beta} \, e^{(\beta\xi/\eta)^2} \cdot W_\xi(\xi, \eta)$$

$$= 1 + \frac{2a\xi}{\eta} U_0/\left(\frac{b\xi}{\eta} U_0 + U_1\right) +$$

$$+ \frac{a}{\beta} (b \, U_0^2 + 2 \, U_1^2 - U_0 U_2)/\left(\frac{b\xi}{\eta} U_0 + U_1\right)^2 \tag{24}$$

and

$$\frac{\sqrt{\pi}}{2} \eta \, e^{(\beta\xi/\eta)^2} \cdot W_\eta(\xi, \eta)$$

$$= - \frac{\beta\xi}{\eta} - a\left\{U_0^2 \cdot \frac{b\xi}{\eta} \left(1 + \frac{2\beta^2\xi^2}{\eta^2}\right) - \right.$$

$$\left. - U_0 U_1 \cdot \left(1 - \frac{2\beta^2\xi^2}{\eta^2}\right) - \frac{1}{2} U_0 U_3 + \frac{1}{2} U_1 U_2\right\}/\left(\frac{b\xi}{\eta} + U_1\right)^2 . \tag{25}$$

826

The first summand in the Nth bracket of $U_0(\xi, \eta)$ becomes effective when the inequality

$$\text{erfc}\big((2N+\xi)/\eta\big) \geq 10^{-m} \tag{26}$$

is satisfied, or, when

$$(2N+\xi)/\eta \leq z , \tag{27}$$

where erfc $z = 10^{-m}$. We call z the threshold root, whose values are shown in Table 1 for several values of threshold powers m. The second summand in the Nth bracket of $U_0(\xi,\eta)$ becomes effective when

$$(2N+2-\xi)/\eta \leq z. \tag{28}$$

Given ξ and η that satisfy $\eta z - \xi > 0$, we find the maximum N among the non-negative integers that satisfy (27), and sum up (13) to obtain $U_0(\xi,\eta)$, ..., $U_3(\xi,\eta)$. The second summand in the Nth bracket in (13) is simply added if it does not underflow.

The graph of $W(\xi, \text{const.})$ runs as shown in Figure 1. It cuts the ξ axis from below. Numerical computation shows that the graph is steadily increasing in the domain $0 \leq \xi \leq 1$. If the tangent at a point on the curve in this domain cuts the ξ axis to the right of the origin, we may apply the Newton iteration to find the root ξ.

The graph of $W(\text{const.}, \eta)$ runs as shown in Figure 2. As η tends to ∞, the graph approaches to the W axis from below. To find the root η we first discover, as shown in the figure, such points P and Q on the η axis that satisfy $W(P) \cdot W(Q) < 0$, where Q needs to be located to the left of the minimum B. Let S be the regula falsi [2, 3], i.e. the intersection of the η axis with the straight line connecting points $W(P)$ and $W(Q)$. Interval PS is narrower in this figure than interval PQ; therefore, we use the former to locate the root R in this case. If the tangent drawn at point $W(S)$ falls in the interval PS, we apply Newton iteration at S. Otherwise we subdivide the range PS by a new regula falsi, and repeat the procedure.

In the 0th bracket, ξ/η, or, equivalently, s_0, is constant prior to the entrance of the second summand. The constant zone satisfies the condition

$$(2-\xi)/\eta \geq z , \tag{29}$$

which defines a domain contiguous to the one found by letting $N = 0$ in (28). The end of the constant coefficient zone may therefore be defined by the solution of the simultaneous equations $W(\xi, \eta) = 0$ and $\xi + \eta z - 2 = 0$. Table 1 shows the end of the constant coefficient zone for various threshold powers in the case of the freezing of water, $T_A = -5°C$, $T_F = 0°C$ and $\xi + \eta z - 2 = 0$. The material constants used are: $K_I = 2.2180$ J/(m s C), $K_{II} = 0.5688$ J/(m s C), $\kappa_I = 1.15 \times 10^{-6}$ m²/s, $\kappa_{II} = 1.44 \times 10^{-7}$ m²/s, $L = 3.35176 \times 10^5$ J/kg. These values yield $\xi_\infty = 0.7958985952$.

On a renewed assumption that the domain is semi-infinite whose temperature at $x = \infty$ is 5°C, we have computed, on the basis of the threshold powers shown in Table 1, temperatures at $x = \ell$ at the times when the interface reaches the end of the constant coefficient zone, which are included in the table. We accept that the temperature in a semi-infinite domain applies prior to a time at which one of the computed temperatures is deemed close enough to 5°C.

Table 2 shows the nondimensional interfacial coordinates for $m = 10$. The table also shows the differences of the temperature gradients at the terminals of both phases, a quantification for demonstrating the approach to the final steady temperature distribution.

The minimum absolute value of $W(\xi, \eta)$ chosen in this computation for defining the root of the equation (14) is 10^{-10}. If the power is higher than 10, the solution process described above does not necessarily converge because of the error bound in the subroutine for evaluating $E_0(x)$, which we have defined through the formula

$$\text{erfcx} = e^{-x^2} \cdot E_0(x) . \tag{30}$$

The subroutine $E_0(x)$ produces 12 effective digits for any nonnegative x. The program uses Erfx continued fraction [4] from $x = 0$ to 0.5, ERFC 5707 RATIONAL APPROXIMATION [5] from $x = 0.5$ to 8.0, and erfcx continued fraction [4] from $x = 8.0$ to ∞.

The computer programs will be made available, which are written in Fortran 77 and run on the PRIME 9750.

828

REFERENCES

1. S. Takagi, Analytical approximate solution of a Stefan's problem in a finite domain, to be published.

   A preliminary report was published:
   S. Takagi, Stefan's problem in a finite domain with constant boundary and initial conditions, Special Report 85-8, U.S. Army Corps of Engineers Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire 03755, June 1985.

2. A.M. Ostrowski, Solutions of equations and systems of equations (Second Edition), Academic Press, 1966.

3. J.R. Rice, Numerical methods, softwares, and analysis: IMSM reference edition, McGraw-Hill, New York, 1983.

4. W.B. Jones and W.J. Thron, Continued fractions, analytic theory and applications, Addison-Wesley Pub. Co., Reeding, Mass., 1980.

5. J.F. Hart, et al., Computer approximations, John Wiley & Sons, Inc., New York, 1968.

Threshold power	Threshold root	ξ_{end}	η_{end}	ξ_{end}/η_{end}	T at $x = \ell$ in a semi-infinite domain
3	2.326753766	0.2386364804	0.7570046928	0.3152377821	4.529240866
4	2.751063906	0.2056840653	0.6522261918	0.3153569542	4.770161762
6	3.458910737	0.1671161673	0.5299020333	0.3153718175	4.941947352
8	4.052237244	0.1444140153	0.4579164232	0.3153719935	4.984650553
10	4.572824967	0.1290340786	0.4091488161	0.3153719955	4.995826039
12	5.042029746	0.1177331889	0.3733152929	0.3153719956	4.998842966
15	5.675846347	0.1052780845	0.3338219183	0.3153719955	4.999826798

Table 1. End of the Constant Coefficient Zone, ξ_{end} and η_{end}, for Various Threshold Powers.

$s_C^{(0)} = 0.3153719956$, which is identical to ξ/η in the constant coef. zone.

ξ/ξ_∞	η	ξ/η	N	$\left.\dfrac{\partial T_I}{\partial x_\ell}\right\|_A - \left.\dfrac{\partial T_I}{\partial x_\ell}\right\|_S$	$\left.\dfrac{\partial T_{II}}{\partial x_\ell}\right\|_S - \left.\dfrac{\partial T_{II}}{\partial x_\ell}\right\|_B$
0.2000000000E-02	0.5047363787E-02	0.3153719956	0	-19.51917859	-1543.583865
0.2000000000E-01	0.5047363787E-01	0.3153719957	0	-1.951917859	-154.3583865
0.1621237672*	0.4091488161	0.3153719955	0	-0.2407935481	-18.93498242
0.2000000000	0.5047364533	0.3153719491	1	-0.1951917284	-14.76285098
0.3000000000	0.7574358477	0.3152340615	1	-0.1300142903	-6.371580060
0.4000000000	1.016674337	0.3131380683	2	-0.962097573E-01	-1.878077954
0.5000000000	1.297904755	0.3066090143	2	-0.7380648687E-01	-0.3476011954
0.6000000000	1.618361976	0.2950756160	3	-0.5697351501E-01	-0.3204738620E-01
0.7000000000	2.006635332	0.2776433803	4	-0.4324388911E-01	-0.8985652141E-03
0.8000000000	2.543809268	0.2503013430	5	-0.3076205766E-01	-0.1869328571E-05
0.9000000000	3.576474772	0.2002834582	7	-0.1751584018E-01	0.4986304702E-09#
0.9900000000	10.75615472	0.7325476716E-01	24	-0.2131743132E-02	Less than 1.E-10
0.9990000000	33.73851528	0.2356661785E-01	76	-0.2186607922E-03	Less than 1.E-10
0.9999000000	106.6011158	0.7465390953E-02	243	-0.2192274630E-04	Less than 1.E-10
0.9999900000	337.1043118	0.2360962492E-02	770	-0.2192448449E-05	Less than 1.E-10

Table 2. An Example of the Interfacial Coordinates and the Approach to the Steady State.

1. * shows the ξ/ξ_∞ at the end of the constant coef. zone for the threshold power 10.
2. N+1 is the number of brackets used to compute $U_0(\xi,n)$.
3. The positive value marked by # shows that this is in the error bound.
4. Suffixes A, S, and B of the temperature gradients mean the cold side, interface, and warm side, respectively.
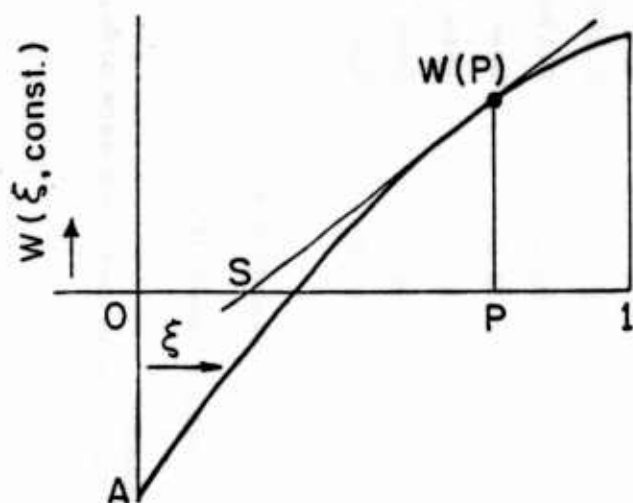5. x_ℓ is the nondimensional space coordinate x/ℓ.

831

Fig. 1 W(ξ, const.) as function ξ.

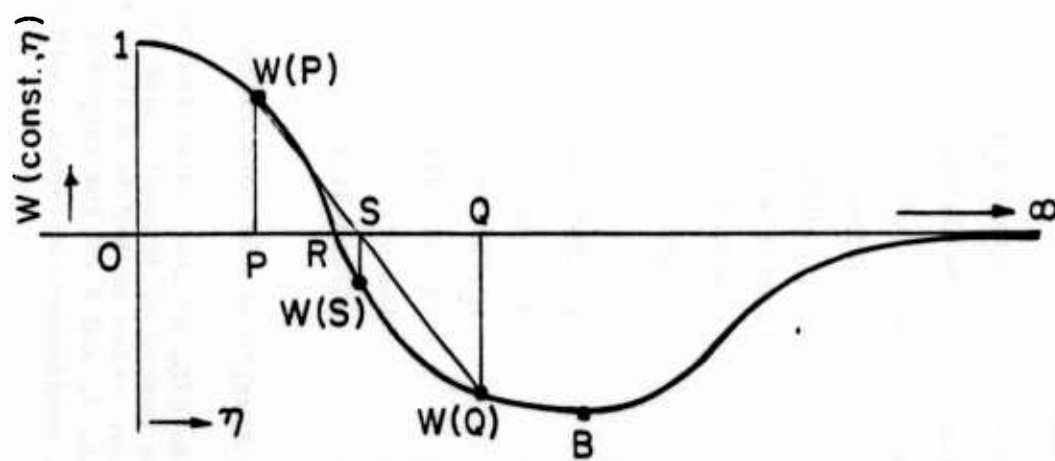Fig. 2 W(const., η) as function of η.

THIN FILM CONDUCTIVE COATINGS FOR SURFACE HEATING AND DECONTAMINATION

S.S. Sadhal
Department of Mechanical Engineering
University of Southern California
Los Angeles, CA 90089-1453

P.S. Ayyaswamy
Department of Mechanical Engineering & Applied Mechanics
University of Pennsylvania
Philadelphia, PA 19104

and

Arthur K. Stuempfle
Chief, Physics Division
U.S. Army Chemical Research & Development Center
Aberdeen Proving Ground, MD 21010-5423

Abstract:

Solid substrates, when exposed to undesirable vapors, can experience temporary or permanent material damage by the adsorption of these vapors. However, the supply of thermal energy to the contaminant can put them back into the gaseous phase. In this analysis the effect of heating as a means of decontamination of substrates is examined in detail. In particular, the use of thin electrically conducting films imbedded in the substrate is considered as a heat source.

In the current development a one-dimensional model for infinitely large substrates is adopted. The thermal transport mechanisms include heat generation in the conductive layer, unsteady heat conduction through the solid substrate, heat of desorption in the adsorbed layer and thermal convection in the gaseous exterior. The mass transport of the contaminant takes place by diffusion in the substrate and desorption at the surface. These coupled phenomena are mathematically modeled by a set of governing differential equations and boundary conditions. This set of equations is solved numerically by finite-difference methods. The analysis predicts the temperature and contaminant concentration of a solid undergoing heating. Results for some special cases have been included to 'exhibit the typical behavior of such systems. For most cases the heating process decontaminates the solid. However, for some cases the heating increases the solubility of the contaminant and may increase its concentration after long periods.

c_v -   concentration of absorbed vapor within the solid

c_o -   maximum vapor concentration that is dissolvable in the solid (This is related to the available absorption sites per unit volume)

c_p -   specific heat

D -   mass diffusion of vapor (in ambient gas or solid)

h -   heater depth from outside surface

H -   dimensionless adsorption depth [equation (3.28)]

k -   thermal conductivity

L -   thickness of the slab

Le -   Lewis number ($= \alpha/D$)

m -   mole friction of the contaminant vapor molecules

Nu -   Nusselt number

p_v -   partial pressure of vapor

$\dot{q}$ -   volumetric heating rate

Q_a -   molar heat of adsorption for the first layer of adsorbed molecules

Q_o -   molar heat of condensation

Q_s -   molar heat of solution

R -   universal gas constant

t -   time

T -   temperature

T_s -   surface temperature

T_∞ -   ambient temperature

x -   dimensionless partial pressure (also equal to mole fraction)

y -   coordinate normal to the slab

Greek Letters

α - thermal diffusivity

δ - thickness of heating element

Δ - coupling parameter governing the thermal transients of the slab surface [Equation (3.27)]

ϵ - solubility parameter which links surface coverage and bulk concentration at the surface [$c = \epsilon\theta/(1 + \epsilon\theta)$].

ϕ - property ratio

δ - density

σ - dimensional surface coverage

σ_o - available adsorption sites per unit area

θ - dimensionless surface coverage ($= \sigma/\sigma_o$)

Subscripts

0 - upper surface

1 - substrate material (plexi-glass)

2 - heater material (Indium-Tin Oxide)

∞ - Ambient far-stream quantity

D - mass diffusivity ratio

g - for property of gas-vapor mixture

k - thermal condictivity ratio

L - lower surface

m - mass transfer

s - at the slab surface; property ratio between solids 2 and 1

t - thermal

v - contaminant vapor

α - thermal diffusivity ratio

835

INTRODUCTION

The problem of contamination of solids by chemical vapors is one of important consideration in many industrial applications. Typically, industrial equipment exposed to undesirable vapors will undergo contamination by the vapors being adsorbed on solid surfaces. After long periods of exposure, absorption of the vapor into the solid will take place. This can cause deterioration of materials such as plexiglass windows and may result in the entire unit being temporarily non-functional.

In the present analysis we examine the process of decontamination by heating the plexiglass substrates with imbedded electrically conducting layers. The application of such heating elements for the purpose of deicing is well known. However, little is known about its overall effectiveness for the removal of adsorbed and absorbed contaminants. While the heat input supplies energy to the contaminant molecules and sets them into a free state (gaseous state), it also increases the solubility of the contaminant in the solid substrate. It is therefore necessary to carry out a detailed mathematical analysis to determine the effect of heating on adsorbed and absorbed contaminants.

In the current study we make some fundamental assumptions relating to the adsorption/desorption kinetics and then develop a one-dimensional model for the removal of physically adsorbed contaminants. The mathematical analysis provides information as to what data are needed to predict the performance of such a decontamination system. At the same time some typical cases have been run to simulate such predictions. Also included are cases for which the increased solubility due to heating may cause the contaminant level to increase.

Kinetic Theory of Adsorption and Desorption

According to Langmuir's [1] monolayer model, adsorption takes place at a constant

836

heat of adsorption Q_a. The ratio of the occupied adsorption sites, σ, to the maximum available, σ_o, on a unit area is given by

$$\theta = \frac{\sigma}{\sigma_o} = \frac{k_3 p}{1+k_3 p}, \tag{2.1}$$

where k_3 is a constant and p is the partial pressure of the vapor. This model for adsorption is suitable for low pressures. A model for multilayer adsorption was suggested by Brunauer, Emmett & Teller [2] in 1938. By simple accounting of the number of molecules in each layer, they gave

$$\theta = \frac{kx}{(1-x)(1-x + kx)} \tag{2.2}$$

Here $x = p/p_o$ where $p_o(T)$ is the equilibrium vapor pressure at temperature T, and $k = e^{(Q_a - Q_o)/RT}$. The heats of adsorption Q_a and Q_o correspond to monolayer and multilayer states, respectively. The total heat of adsorption is given by

$$Q_t(\theta) = [Q_a - (Q_a - Q_o)x]\theta\sigma_o/N = Q\theta\sigma_o/N, \tag{2.3}$$

where N is the Avogadro number and Q is defined as

$$Q = Q_a - (Q_a - Q_o)x, \tag{2.4}$$

The fundamental principle behind the decontamination lies in behavior of θ with temperature. From (2.2) it is clear that for small x, θ is nearly proportional to x. However $p_o(T)$ increases with T and hence $x = p/p_o$ decreases with T. Therefore θ decreases with T, indicating a reduction in the amount of contaminant adsorbed in the surface of the solid. It is clear that by raising the temperature a surface may be decontaminated.

In the next section we formulate the governing differential equations for substrates electrically heated by imbedded conductive layers.

3.DECONTAMINATION OF SUBSTRATES BY ELECTRICAL HEATING: FORMULATION

Description of Problem

Since with a rise in the temperature of the substrate, decontamination of the surface takes place, the possibility of heating by imbedded elements (such as in an automobile windshield) is considered here. The application to defogging and deicing is well known [1,4,8-9]. We adopt here a one-dimensional model in which the heat and mass flow in the direction parellel to the plane of the substrate are considered negligible. At this point we can define a specific one-dimensional time-dependent problem.

Let us consider a long slab of thickness L which is exposed to an environment containing a chemical vapor and some inert gases. An electrically heated layer of thickness δ is imbedded in the slab to a depth h (see Fig. 3.1). The substrate is referred to as phase 1 and the heating element as phase 2. The chemical vapor deposits itself on the surface of the substrate by adsorption and then diffuses into the bulk of the substrate.

The physico-chemical processes involved are as follows:

Mass Transfer

1. Diffusion and convection of contaminant vapor in the environment takes place due to the wind pattern, or to the motion of the substrate. This sets up a velocity profile near the surface of the substrate. As a result, convective transport of the contaminant to or from the surface takes place.

2. Adsorption/desorption of the vapor at the solid surfaces.

3. Diffusion of vapor into the solid phases (1 and 2). Chemical reactions within the solid are not being considered.

Heat Transfer

1. Heat release by electrical heating within the conductive layer.

2. Heat conduction in the solid phases 1 and 2.

3. Heat associated with adsorption/desorption at the surface.

4. Thermal diffusion and convection in the gaseous environment.

Assumptions

1. Fluid flow processes are considered only for their effects on heat/mass transfer. Simple lumped parameter models (i.e., heat and mass transfer coefficients) are to be used for these processes.

2. The adsorbate (solid) is infinite in length. A one-dimensional formulation is used for the solid phase.

3. A uniform volumetric heat generation rate $\overset{\circ}{q}''$ is considered in the conductive layer of thickness δ. Any chemical reactions within the solid phase and the latent heat release thereof is not considered.

4. The concentration of the sorbed species within the solid is linearly related to the surface concentration of the adsorbed layer within certain limits.

5. The solid surfaces are taken to have reached an adsorption equilibrium with the surrounding gas-vapor mixture. At this equilibrium, the "fraction" of the surface covered (θ) depends on the surface temperature T_s and the partial pressure $p_{v,s}$ of the vapor adjacent to the surface. Furthermore, at equilibrium, $p_{v,s}$ is purely a function of the surface temperature and the heat of adsorption.

   If $p_{v,s}$ is different from $p_{s,\infty}$, (the far-field partial pressure) then vapor transport takes place in the gaseous phase as well.

6. For the present model, only physical adsorption is treated for a multi-molecular adsorbed layer. The heat of adsorption for the first layer, Q_a, is different from the subsequent layers, for which it is taken to be the latent heat of condensation, Q_o. For the first layer Q_a actually varies with the fraction covered but here we take it to be the average value.

7. The dependence of the partial pressure with temperature is given by

$$p_{v,s} = xp_{o,s} = xp_{o,\infty}e^{-\frac{Q_o}{R}\left(\frac{1}{T_s} - \frac{1}{T_\infty}\right)} \tag{3.1}$$

where $x = m_{v,s}$ is the mole fraction of the vapor in the air adjacent to the surface.

Governing Equations

At this point the problem may be precisely cast into a mathematical form as a closed set of partial differential equations. With the assumptions made in §3.1, the equations for heat and mass balance are as follows:

Gas Phase

Heat transfer between solid surface and air:

$$q_g = h_{gt}(T_s - T_\infty) \tag{3.2}$$

Vapor mass flux between solid surface and air.

$$j_{g,v} = h_{gm}\rho_g(m_{v,s} - m_{v,\infty}) \tag{3.3}$$

where

$$h_{gt} = Nu_t k_g/L \tag{3.4}$$

$$h_{gm} = Nu_m D_{gv}/L \tag{3.5}$$

with

Nu   = Nusselt number
k_g = thermal conductivity of gas
m_v = mole fraction of the vapor
D_{gv} = binary diffusion coefficient between vapor and air.
ρ = density.

Equation (3.2) and (3.3) are based on lumped parameter modeling of the convective transport. The Nusselt numbers, Nu_t and Nu_m depend on the external flow conditions. The general characteristics of these Nusselt numbers are available in most heat transfer textbooks (see, e.g., Burmeister [2]).

Solid-Gas Interface

Heat transfer at the interface:

$$q_g - q_a = q_s \tag{3.6}$$

where

$$q_s = -k\frac{\partial T_1}{\partial y}\Big|_{y=0}$$

or

$$q_s = +k\frac{\partial T_1}{\partial y}\Big|_{y=-L} \tag{3.7}$$

and q_a is the heat release by adsorption.

Mass transfer at the interface:

$$j_{g,v} + j_a = j_{s,v} \tag{3.8}$$

where

j_a = mass rate of adsorption (g/cm^2–sec)
$j_{s,v}$ = mass flux from the solid

$$j_{s,v} = -D_{1v}\frac{\partial c_{1v}}{\partial y}\Big|_{y=0}$$

or   (3.9)

$$j_{s,v} = +D_{iv}\frac{\partial c_{1v}}{\partial y}\Big|_{y=-L}$$

and $j_{g,v}$ is given by (3.3)

<u>Adsorption Equilibrium</u>

From (2.2)

$$\theta = \frac{kx}{(1-x)(1-x+kx)} \tag{3.10}$$

where

$$k = e^{(Q_a-Q_o)/RT} \tag{3.11}$$

and following (2.4) it can be seen that

$$Q - Q_o = (Q_a - Q_o)(1-x) \tag{3.12}$$

By employing the quasi-equilibrium assumption we may write the adsorption heat flux as

841

$$q_a = \frac{\sigma_o}{N}\dot{\theta}Q \tag{3.13}$$

and

$$j_a = \frac{\sigma_o}{N}\dot{\theta} \times M_v = m\sigma_o\dot{\theta} \tag{3.14}$$

where N is the Avogadro number, σ_o is the maximum number of available sites per unit area, M_v is the molecular weight of vapor, and m is the mass of one molecule of the vapor.

<u>Solid Phase 1</u>

Heat transfer:

$$\frac{1}{\alpha_1}\frac{\partial T_1}{\partial t} = \frac{\partial^2 T_1}{\partial y^2} \tag{3.15}$$

where α_1 is the thermal diffusivity and T_1 is the temperature.

Mass transfer:

$$\frac{1}{D_{1v}}\frac{\partial c_{1v}}{\partial t} = \frac{\partial^2 c_{1v}}{\partial y^2} \tag{3.16}$$

where D_{1v} is the diffusion coefficient, and c_{1v} is the mass concentration in g/cm^3.

We assume a linear relatioship between the superficial mass concentration that is adsorbed at the surface and the volumetric concentration adjacent to the surface. This is the basis of assumption 4 in § 3.1. Thus at $y = 0$ and $y = -L$,

$$m\sigma_o\theta = \phi c_{1v} \tag{3.17}$$

where ϕ is a constant with dimensions of length. This is called the penetration depth. The relationship (3.17) is used for cases when

$$\frac{m\sigma_o}{\phi} << c_o \tag{3.18}$$

where c_o is the maximum possible concentration.

842

<u>Solid Phase 2</u>

Heat Transfer

$$\frac{1}{\alpha_2}\frac{\partial T_2}{\partial t} = \frac{\partial^2 T_2}{\partial y^2} + \frac{\dot{q}'''}{k_2} \tag{3.19}$$

where $\dot{q}'''$ represents the volumetric heat generation rate in cals/cm^3-sec.

Mass Transfer

$$\frac{1}{D_{2v}}\frac{\partial c_{2v}}{\partial t} = \frac{\partial c_{2v}}{\partial y^2} \tag{3.20}$$

The boundary condition between solid 1 and solid 2 at $y = -h$ and $y = -(h+\delta)$ are

$$T_1 = T_2 \tag{3.21}$$

$$k_1\frac{\partial T_1}{\partial y} = k_2\frac{\partial T_2}{\partial y}$$

and

$$c_{1v} = c_{2v}$$

$$\tag{3.22}$$

$$D_{1v}\frac{\partial c_{1v}}{\partial y} = D_{2v}\frac{\partial c_{2v}}{\partial y}$$

The initial conditions are

$$T_1 = T_2 = T_\infty \quad \text{at } t = 0$$

$$\tag{3.23}$$

$$c_{1v} = c_{2v} = 0 \quad \text{at } t = 0 \qquad \qquad .$$

The equations (3.1-3.23) form a closed set. However, it is convenient to identify dimensionless groups and make these equations non-dimensional.

Dimensionless Grouping

The dimensionless parameter are selected as follows:

$$Nu_t = \frac{h_t L}{k_g} \tag{3.24}$$

$$Nu_m = \frac{h_m L}{D_{gv}} \tag{3.25}$$

$$\varepsilon = \frac{m\sigma_o}{\phi c_o} \tag{3.26}$$

$$\Delta = \frac{m\sigma_o}{Lc_o} \cdot \frac{c_o}{\rho_1} \cdot \frac{R}{c_{p1}} \tag{3.27}$$

$$H = \frac{m\sigma_o}{Lc_o} \text{ (dimensionless adsorption depth)} \tag{3.28}$$

$$Q_o^* = \frac{Q_o}{RT_\infty} \tag{3.29}$$

$$Q_a^* = \frac{Q_a}{RT_\infty} \tag{3.30}$$

$$\tilde{\rho}_o = \frac{\rho_\infty}{c_o} \text{ (ratio of saturation density at infinity to maximum solubility in solid 1)}$$

$$\tag{3.31}$$

$$r_\infty = \text{vapor mole fraction at infinity} \tag{3.32}$$

$$\dot{q}^* = \frac{\dot{q}''' L^2}{kT_\infty} \text{(dimensionless volumetric heating rate)} \tag{3.33}$$

$$\varepsilon = \frac{m\sigma_o}{\phi c_o} \text{ (dimensionless penetration depth)} \tag{3.34}$$

$$\phi_{kg} = k_g/k_1 \tag{3.35}$$

$$\phi_{Dg} = D_{gv}/D_{1v} \tag{3.36}$$

$$Le_1 = \alpha_1/D_{1v} \tag{3.37}$$

$$\phi_{\alpha s} = \alpha_2/\alpha_1 \tag{3.38}$$

844

$$\phi_{Ds} = D_{2v}/D_{1v} \tag{3.39}$$

$$\phi_{ks} = k_2/k_1 \tag{3.40}$$

$$T^* = T/T_\infty \tag{3.41}$$

$$c_{1v}^* = c_{1v}/c_o \tag{3.42}$$

$$y^* = y/L, \ h^* = h/L, \ \delta^* = \delta/L, \ t^* = \alpha_1 t/L^2 \tag{3.43}$$

By combining (3.2), (3.6), (3.7) and (3.13) we obtain

$$Nu_t \ \phi_g(T_s^* - 1) - \Delta \frac{d\theta}{dt^*}[Q_o^* + (Q_a^* - Q_o^*)(1-x)] = \pm \frac{\partial T_1}{\partial y} \ \text{at} \ y = 0 \ \text{or} \ y = -L. \tag{3.44}$$

Similarly, by combining (3.1), (3.3), (3.8), (3.9) and (3.14) we find

$$Nu_m \phi_{Dg}\tilde{\delta}_o \ r_\infty \ \frac{x}{T_s^*}e^{Q_o^*(1 - 1/T_s^*)} - Le_1 H\dot{\theta} = \pm \frac{\partial c_{1v}^*}{\partial y} \quad \text{at} \ y = 0 \ \text{or} \ y = -L \tag{3.45}$$

The adsorption equilibrium given by (3.10) and (3.12), when non-dimensionalized, gives

$$\theta = \frac{kx}{(1-x)(1-x+kx)} \tag{3.46}$$

with

$$k = e^{(Q_a^* - Q_o^*)/T_s^*} \tag{3.47}$$

The remaining equations may be written as

$$\frac{\partial T_1^*}{\partial t^*} = \frac{\partial T_1^*}{\partial y^{*2}} \tag{3.48}$$

$$\frac{\partial c_{1v}^*}{\partial t^*} = \frac{1}{Le_1} \frac{\partial^2 c_{1v}^*}{\partial y^{*2}} \tag{3.49}$$

$$\frac{\partial T_2^*}{\partial t^*} = \phi_{\alpha s} \frac{\partial^2 T_2^*}{\partial y^{*2}} + \dot{q}^* \qquad (3.50)$$

$$\frac{\partial c_{2v}^*}{\partial t^*} = \phi_{Ds} \frac{1}{Le_1} \frac{\partial^2 c_{2v}^*}{\partial y^{*2}} \qquad (3.51)$$

with boundary conditions

$$c_{1v}^* = \varepsilon \theta \quad \text{at} \quad y = 0, -1 \qquad (3.52)$$

$$T_1^* = T_2^*; \frac{\partial T_1^*}{\partial y^*} = \phi_{ks} \frac{\partial T_2^*}{\partial y^*} \quad \text{at } y = -h^*, -(h^* + \delta^*) \qquad (3.53)$$

and

$$c_{1v}^* = c_{2v}^*; \frac{\partial c_{1v}^*}{\partial y^*} = \phi_{Ds} \frac{\partial c_{2v}^*}{\partial y^*} \quad \text{at } y^* = -h^*, -(h^* + \delta^*) \qquad (3.54)$$

and initial conditions

$$T_1^* = T_2^* = 1 \quad \text{at} \quad t^* = 0$$

$$\qquad (3.55)$$

$$c_{1v}^* = c_{2v}^* = 0 \quad \text{at} \quad t^* = 0$$

We next examine the magnitudes of the various dimensionless parameters so as to identify the relatively important transport mechanisms.

2. IDENTIFICATION OF THE IMPORTANT TRANSPORT MECHANISMS

In this section we first state the approximate range of physical properties and external conditions for typical situations of practical interest. This is followed by an estimate of the magnitudes of the dimensionless groups.

Physical Properties

Plexiglass substrate (Poly Methyl Methacrylate)

Property		Value at 378K	Range	
Density,	ρ_1	1.19	1.0	g/cc
Specific heat,	c_{p1}	0.37	.5	cal./g-K
Thermal conductivity,	k_1	5×10^{-4}	$1-10 \times 10^{-4}$	cal/cm-sec-K
Thermal conductivity,	α_1	1.14×10^{-3}	$1-10 \times 10^{-3}$	cm^2/sec
Mass diffusivity (O_2),	D_{1v}	1×10^{-8}	$10^{-6}-10^{-8}$	cm^2/sec

Air at 20°C

Thermal conductivity:	k_g	0.6267×10^{-4}	cal/cm-sec-K
Thermal diffusivity:	α_g	0.2216	cm^2/sec
Mass Diffusivity (O_2, N_2)	D_g	0.2	cm^2/sec

Electrically Conducting Layer (Indium Oxide + Stannic Oxide)

Density:	ρ_2	6.3	g/cc
Specific heat:	c_{p2}	0.2	cal/g-K
Thermal conductivity:	k_2	0.0136	cal/cm-sec-K
Thermal diffusivity:	α_2	1.08×10^{-2} .	cm^2sec

Mass diffusivity: D_2 Not Available (consider: $D_{2v}/D_{1v} = 10^{-1}$, 1 and 10)

Thickness of layer: δ typically 1000 $\overset{o}{A} = 10^3 cm$, (consider: $1 - 10 \times 10^{-3} cm$)

847

Heats of Adsorption

The ambient temperature is taken to be $T_\infty = 20°C$ (293 K). Assuming $Q_a \simeq 3$ kcal/g-mole, we obtain:

$$Q_a^* = Q_a/RT_\infty \simeq 5.2.$$

With this as an approximate estimate, a range $Q_a^* = 1-25$ is considered.

Similarly, with $Q_o \simeq 1$ kcal/g-mole, $Q_o^* = 1-5$ is considered.

Heating Levels

The maximum heating levels quoted in the literature are of the order of 4 cal/cm²-sec. Therefore, the volumetric heating rating (for $\delta = 10^{-3}$cm) is

$$\dot{q}^* = \frac{\dot{q}L^2}{k_2 T_\infty} \simeq 60.$$

We consider $\dot{q}^* = 1-100$.

Solubility Parameter

The solubility parameter ϵ is estimated as follows:

At the surface we have $c_{1v}^* = \epsilon\theta$. As θ takes on large values (say, 10) the solid phase will also approach saturation ($c_{1v}^* \to 1$). Therefore, we take $\epsilon = 1/10 = 0.1$.

Adsorption Depth

The dimensionless adsorption depth, $H = m\sigma_o/Lc_o$ is estimated by examining the solubility of N_2 in Poly Ethyl Methaclylate. At 25°C the solubility is

848

$$s = 7.5 \times 10^{-2} c.c. gasSTP/c.c. substrate.$$

where s is defined by $c_{1v} = sp$. Assuming a partial pressure of $p = 0.1$ atm, we find the volumetric concentration c_{1v} to be

$$c_{1v} = 7.5 \times 10^{-3} c.c.(STP)/c.c..$$

In units of mass concentration this is

$$c_{1v} = 9 \times 10^{-6} g/c.c. \simeq 10^{-5} g/c.c.$$

We therefore take $c_o \simeq 10^{-5}$ g/c.c.. If the gas is more soluble, c_o may be as large as 10^{-3} g/c.c.. For very low solubilities it may be 10^{-7} or 10^{-8} g/c.c..

Dimensionless Groups

The maximum number of available sites is of the order of $\sigma \sim 10^{14}/cm^2$. The parameter H is therefore given by

$$H = \frac{m\sigma_o}{Lc_o} \sim 10^{-4}$$

where L is taken to be 1 cm, and $m = 4.67 \times 10^{-23} g$ for N_2.

The dimensionless parameter Δ is given by

$$\Delta = \frac{m\sigma_o R}{L\rho_1 c_{p1}} = H \frac{c_o R}{\rho_1 c_{p1}} \simeq 10^{-9}$$

From these estimates it is clearly seen that surface transients (θ terms) would be negligibly small. Furthermore, due to the very small mass diffusivity of the solid ($10^{-6} - 10^{-7}$ cm^2/sec) compared to that in the gaseous phase (0.1 cm^2/sec), the mass transfer relationship betwen the surface and the ambient reduces to the quasisteady relation $c_{1v} = \epsilon\theta$ where θ is the surface coverage at the adsorption quasi-equilibrium.

The above linear relationship can be generalized by considering saturation phenomena within the solid and a Langmuir type absorption isotherm, $c = \epsilon\theta/1 + \epsilon\theta$ can be employed. The results of the estimates of these dimensionless groups are summarized in Table 1.

Dimensionless Parameter	Range	Comments
Nu_t	O(100)	Included in the model
Nu_m	O(100)	This is large, but irrelevant since gas-phase mass transfer does not affect problem
Δ	10^{-9}	Negligible
Q_o^*	1–5	Included
Q_a^*	1–25	Included
$\dot{q}^*$	1–100	Included
H	$10^{-6} - 10^{-1}$	Irrelevant
ϵ	< 1	Strongly temperature dependent. Consider values 10^{-4} – 1
δ^*	0.001	Consider 0.001, 0.01, 0.1
h^*	$0 < h^* < 1$	
ϕ_{k_g}	0.05 – 0.5	
ϕ_{D_g}	$10^5 - 10^8$	
Le_1	$10^2 - 10^5$	
ϕ_{α_s}	1 – 10	
ϕ_{k_s}	10 – 100	
ϕ_{D_S}	---	No data available, try 0.1, 1, 10

Table 1: Summary of estimate of magnitudes of dimensionless parameter

5.RESULTS AND DISCUSSION

The non-dimensional governing differential equations together with the simplifications discussed in §4 have been programmed for a finite-difference solution. The results from the various sets of data have been plotted in Figs. 5.1-5.5. Here we discuss each case in detail.

In Fig. 5.1, the temperature profile at various times is shown. The heater is placed at $y^* = -0.25$. The thermal parameters for this plot are $\dot{q}^* = 50.0$, $\delta^* = 0.001$, $\phi_{ks} = 50.0$, $Nu_{t0} = 200.0$ and $Nu_{tL} = 10.0$. The plot shows the following important features.

1. The region near $y^* = 0$ reaches a steady state faster than the rest of the substrate. This is because of the higher Nusselt number and the shorter distance from the heating element.

2. More heat leaves through the surface $y^* = 0$ than $y^* = -1$. This is owing to the relatively lower thermal resistance of the region $-h^* \leq y^* \leq 0$ than $-1 \leq y^* \leq -h^*$.

3. A large Nusselt number causes the corresponding surface to be cooler and, as a result, leads to higher adsorption. The heating is therefore wasteful. If the Nu_t is controllable, then it should be minimized so that very high heating levels are not needed.

4. The maximum steady temperature in this case is $\bar{T}_{max}/T_\infty \simeq 1.5$. Assuming $T_\infty = 300K$, we have $\bar{T}_{max} = 450K$. At such high temperatures the plexiglass would deteriorate.

In Fig. 5.2 the dimensionless surface coverage of the contaminant on the outside, (θ_o) and on the inside (θ_L) are shown as functions of time. The parameters for the graph are:

	H	δ^*	ϕ_{ks}	ϕ_{kg}	$\dot{q}^*$	Nu_{to}	Nu_{tL}	$\phi_{\alpha s}$
Run I	0.5	0.0001	10	0.05	50	200	10	1
Run II	0.25	0.001	50	0.25	50	200	10	5
Run III	0.5	0.001	50	0.25	50	10	10	5

851

Run IV   0.5   0.001   50   0.25   10   2   2   5

In addition we use $Q_a^* = 5.0$, $Q_o^* = 1.0$, $p_{vo} = 0.5$ and $p_{vL} = 0.01$. The following features are observed.

1. Curves I and II, between which there is no systematic change in the thermal parameters, both reach the same steady state value for θ_o and θ_L. This is because the groups

$$Q_1 = \frac{1 + Nu_{to}\phi_{kg}h^*}{1 + Nu_{tL}\phi_{kg}(1-h^*)} \tag{5.1}$$

and

$$Q_2 = \frac{\dot{q}^* \delta^* \phi_{ks}}{\phi_{kg}(Nu_{to} + Nu_{tL}Q_1)} \tag{5.2}$$

have approximately the same values for these curves.

2. The curve II corresponds to a larger thermal diffusivity than curve I. It therefore approaches a steady state faster. For curve IV the Nusselt number is lower that curve III. The eigenvalues determining the rate of thermal transport are smaller for curve IV and the transport process lasts longer in this case.

3. The higher level of adsorption on the outside is due to larger partial pressure of the vapor on the outside.

In Fig. 5.3, the effects of partial pressure of the vapor and the heats of adsorption on the fraction covered are shown. The plots correspond to fixed values of the parameters Q_1 and Q_2 (Q_1 = 0.0317; Q_2 = 11.5556) or fixed values of the parameters

$$T_o = 1 + Q_2 \tag{5.3}$$

and

$$T_L = 1 + Q_1Q_2. \tag{5.4}$$

1. We find that $\theta_L < \theta_o$ because $T_L > T_o$. This is due to the convective cooling which occurs in the outside.

852

2. Each curve shows a steep linear portion for small pressures, a flat portion for moderate pressures, and saturation ($\theta \to \infty$) as the partial pressure approaches saturation.

3. At large Q_a^*, the linear and the flat portions are separated around $\theta \sim 1$. this implies that for large Q_a^* (such as in chemisorption) a monolayer is formed first until $\theta \sim 1$. Subsequently, more layers build up. For $Q_a^* \sim Q_o^*$, all layers may be formed simultaneously.

4. At large Q_o^*, the saturation phenomenon is delayed. This happens because the parameter $k = e^{(Q_a^* - Q_o^*)/T_s^*}$, which signifies the ratio of adsorption times between the first and the subsequent layers, decreases.

In Fig. 5.4, the concentration profile within the solid is plotted. The solubility has been assumed to be the same in both the substrate and the heating element. Also the ratio of the two diffusivities is taken to be unity.

The parameters are: $L^* = 0.5$, $\delta^* = 0.001$, $\phi_{\alpha s} = 5.0$, $\phi_{Ds} = 1.0$, $\phi_{ko} = 50.0$, $Le_1 = 500$, $\phi_{kg} = 0.25$, $\phi_{Dg} = 5.0 \times 10^5$, $Q_a^* = 5.0$, $Q_o^* = 1.0$, $\dot{q}^* = 50.0$, $Nu_{to} = 10.0$, $Nu_{tL} = 200.0$, $\varepsilon = 0.1$, $p_{vo}^* = 0.5$ and $q_{vL}^* = 0.01$.

We observe from the plot that:

1. The time taken to reach mass transfer steady state is approximately equal to $Le_1 \times$ (time for thermal steady state).

2. For short times, diffusion occurs from both ends and the effects from each end grow independently until they interact.

3. As time increases, the concentration profile becomes monotonic and a straight line is obtained for constant mass diffusivity. At steady state, a steady stream of vapor diffuses from the higher concentration side to the lower concentration side.

853

4. The surface concentrations are given by

$$c_o = \varepsilon\theta_o/(1 + \varepsilon\theta_o) \qquad \text{at } y^* = 0 \tag{5.5}$$

and

$$c_L = \varepsilon\theta_L/(1 + \varepsilon\theta_L) \qquad \text{at } y^* = 1 \tag{5.6}$$

Here θ_o and θ_L are functions of the thermal parameters Q_1 and Q_2, the heats of adsorption Q_a and Q_o, and the partial pressures p_{vo} and p_{vL}. Thus after the thermal steady state is reached, c_o and c_L remain constant due to thermal equilibrium.

In Fig. 5.5, we have plotted the variation in the steady state bulk concentration ($c_{bulk} = \int_{-1}^{o} c^* dy^*$) as well as $T_{max}^* = T_{max}/T_\infty$ as a function of $\dot{q}^*$. The important feature incorporated here is that the solubility parameter ε is taken to be temperature dependent. It is given by

$$\varepsilon = \varepsilon_o e^{Q_s^*(1 - 1/T_s^*)} \tag{5.7}$$

where $Q_s^* = Q_s/RT_\infty$ is the heat of solution. Since the solubility changes with changing temperature, the heat of solution plays a role. The program was modified to include this addition parameter. The plot corresponds to the following values of the parameter:

$L^* = 0.8,$ $\delta^* = 0.001,$ $\phi_{ks} = 50.0,$ $\phi_{kg} = 0.25,$

$Nu_{to} = 200.0,$ $Nu_{tL} = 10.0,$ $\varepsilon_o = 0.1,$ $Q_a^* = 10.0,$

$Q_o^* = 1.0,$ $p_{vo}^* = 0.5,$ $p_{vL}^* = 0.01,$ $Q_s^* = 0.0-1.0.$

The results exhibit the following features:

1. For very low heat of solution, Q_s^*, the bulk concentration decreases monotonically with the heating level $\dot{q}^*$. It may be noted that the duration of heating does not determine the level of contamination after the attainment of thermal and mass transfer steady states. For this reason the level of initial contamination or the initial temperature do not affect c_{bulk}.

2. For a reasonably large heat of solution, while the surface coverage (θ_o and θ_L)

854

decreases with increasing temperature, the dissolved contaminant in the bulk increases. this usually happens when the heating levels are low and the change of θ_0 and θ_L are not as rapid as that of the solubility. If the outside environment is very cold and if the convective cooling is strong, the interior of the substrate may be very hot, but the surface will remain fairly cool. As a result, the heating may not substantially remove the surface contaminant and at the same time it will increase the solubility. This will lead to increased contamination if heating is sustained for long periods.

3. An important consideration for design would be the maximum temperature reached:

$$T_{max} = \frac{T_o + T_L h^*/(1-h^*) + \dot{q}^* \delta^* h^* \phi_{ks}}{1 + h^*/(1-h^*)}. \tag{5.7}$$

If $T_\infty \simeq 300K$, then for a material such as plexiglass, we would require $T_{max} \leq 1.5$ to avoiding thermal damage.

Acknowledgement

References

1. LANGMUIR, I., "Adsorption of Gases on Glass, Mica and Platinum", _Journal Amer. Chem. Soc._ 40, 1361 (1918).

2. BRUNAUER, S., EMMETT, T. H. & TELLER, E., "Adsorption of Gases in Multimolecular Layer", _Journal Amer. Chem. Soc._ 60, 309 (1938).

3. SADHAL, S. S. & AYYASWAMY, P.S., "Thin Film Conductive Coatings for Surface Heating and Decontamination", U. S. Army Technical Report CRDC-CR-85028 (1985).

4. BOAZ, P.T. & YOUNGS, J.D. Ford Motor Company, "Electrically Heatable Windshield and Bakelite System", _Trans. Soc. Automotive Engnrs._ #740157, March 1974.

5. BURMEISTER, L. C., _Convective Heat Transfer_, Wiley (1983).

6. KING, R. D., Triplex Safety Glass co., Ltd., "Defrosting of Automobile Windshields Using High Light Transmitting Electroconducting Films", _Trans. Soc. Automotive Engnrs._ #740158, March 1974.

7. MIZUHASHI, M., "Electrical Properties of In_2O_3 and In_2O_3:Sn Films", _Thin Solid films_ 70, 91-100 (1980).

8. PRESSON, E. W., PPG Industries, Inc., "Progress Update – Electrically Heated Windshields", _Trans. Soc. Automotive engnrs._ #790600, August 1980.

9. ROTHE, W., PPG Industries, Inc., "AIRCON Electrically Heated Acrylic", _Trans. Soc. Automotive Engnrs._ #790600, April 1979.

10. MORRIS, J. E., BISHOP, C. A. RIDGE, M. I. & HOWSON R. P., "Structural Determination of In_2O_3 Thin Films on Polyester Substrates by Transmission Electron Microscopy", _Thin Solid Films_ 62, 19-23 (1979).
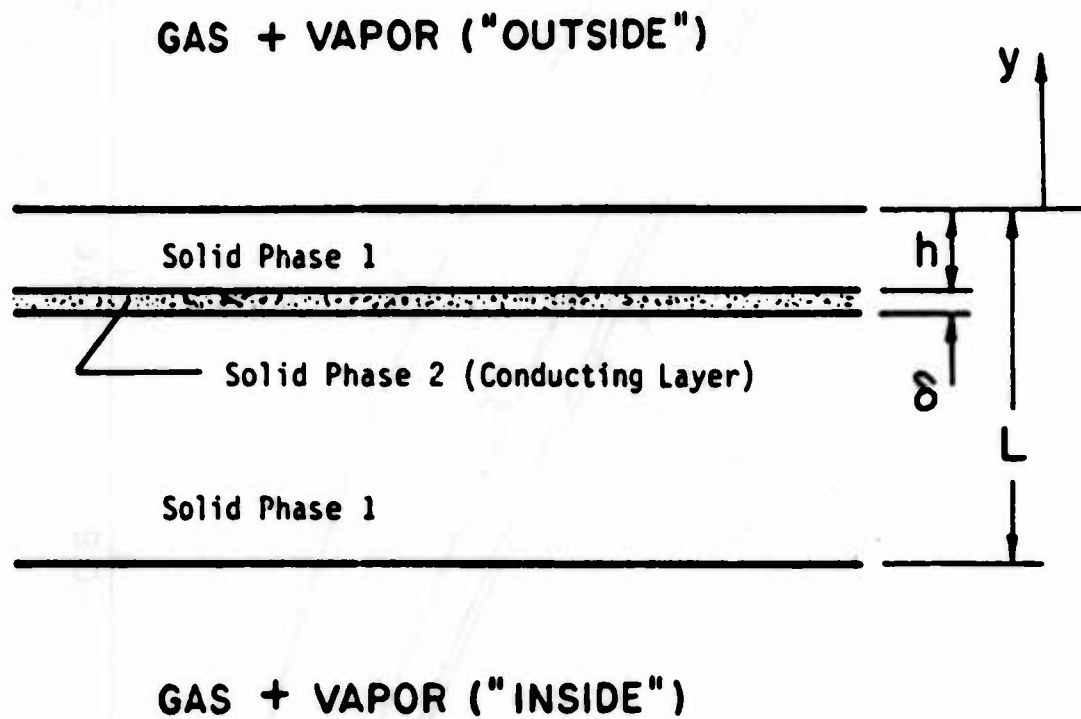
GAS + VAPOR ("OUTSIDE")

GAS + VAPOR ("INSIDE")

Fig. 3.1
One-dimensional model of substrate with
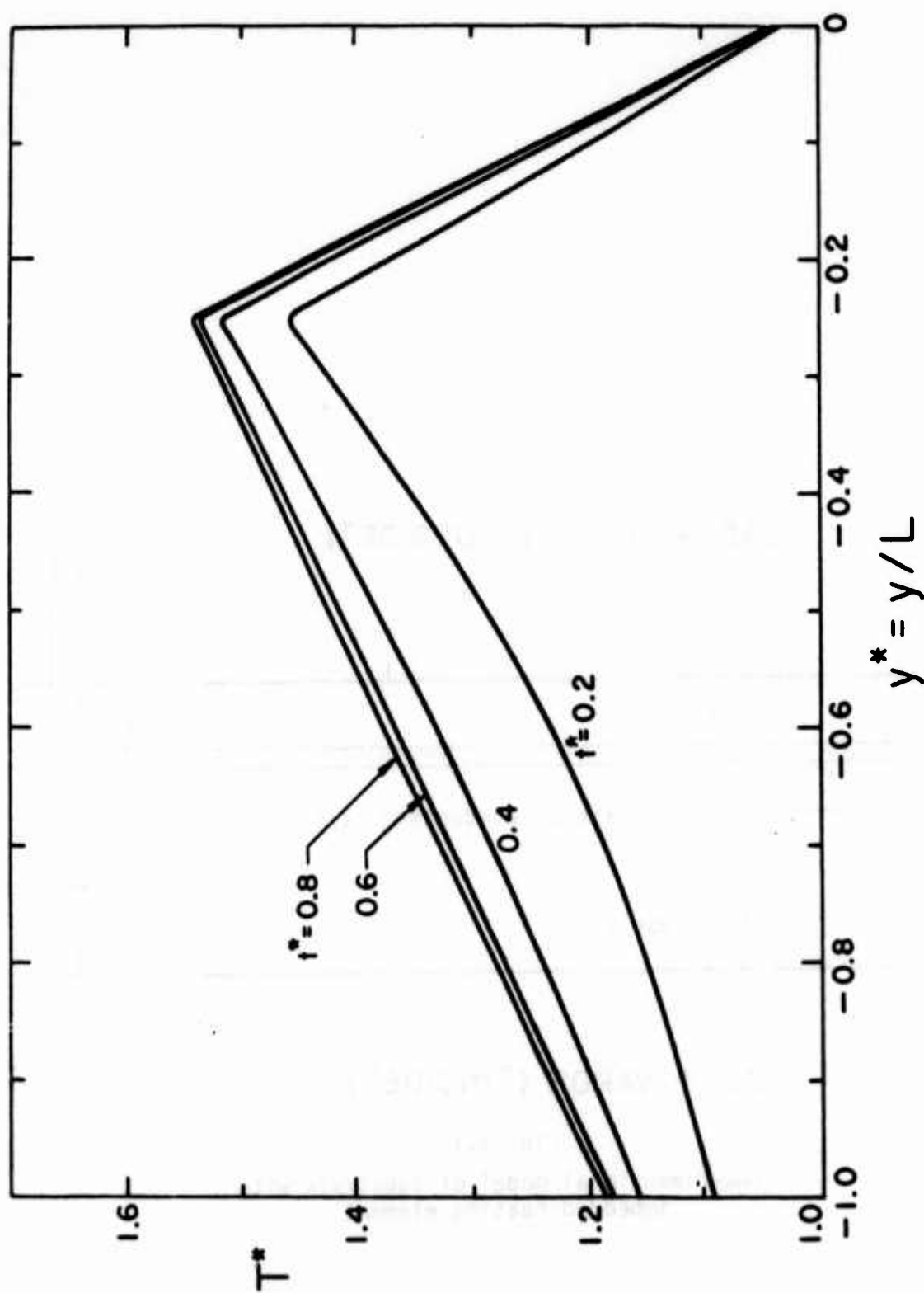imbedded heating element

Fig. 5.1

Temperature distribution within the substrate at various times.
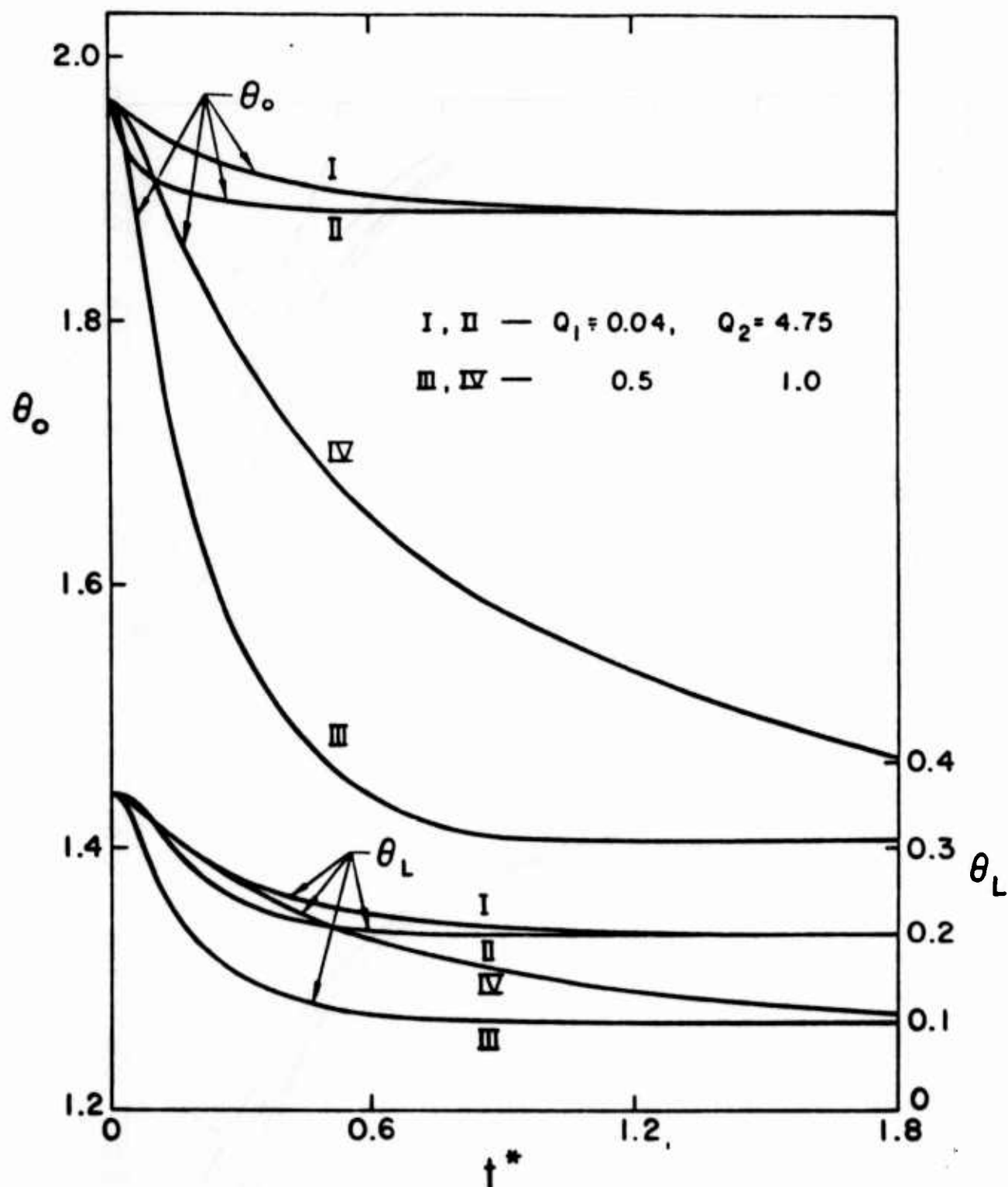A sharp variation of T^* takes place at the heater location.

Fig. 5.2

Dimensionless surface coverage at various times.
The upper surface is denoted by θ_o and the lower one by θ_L.
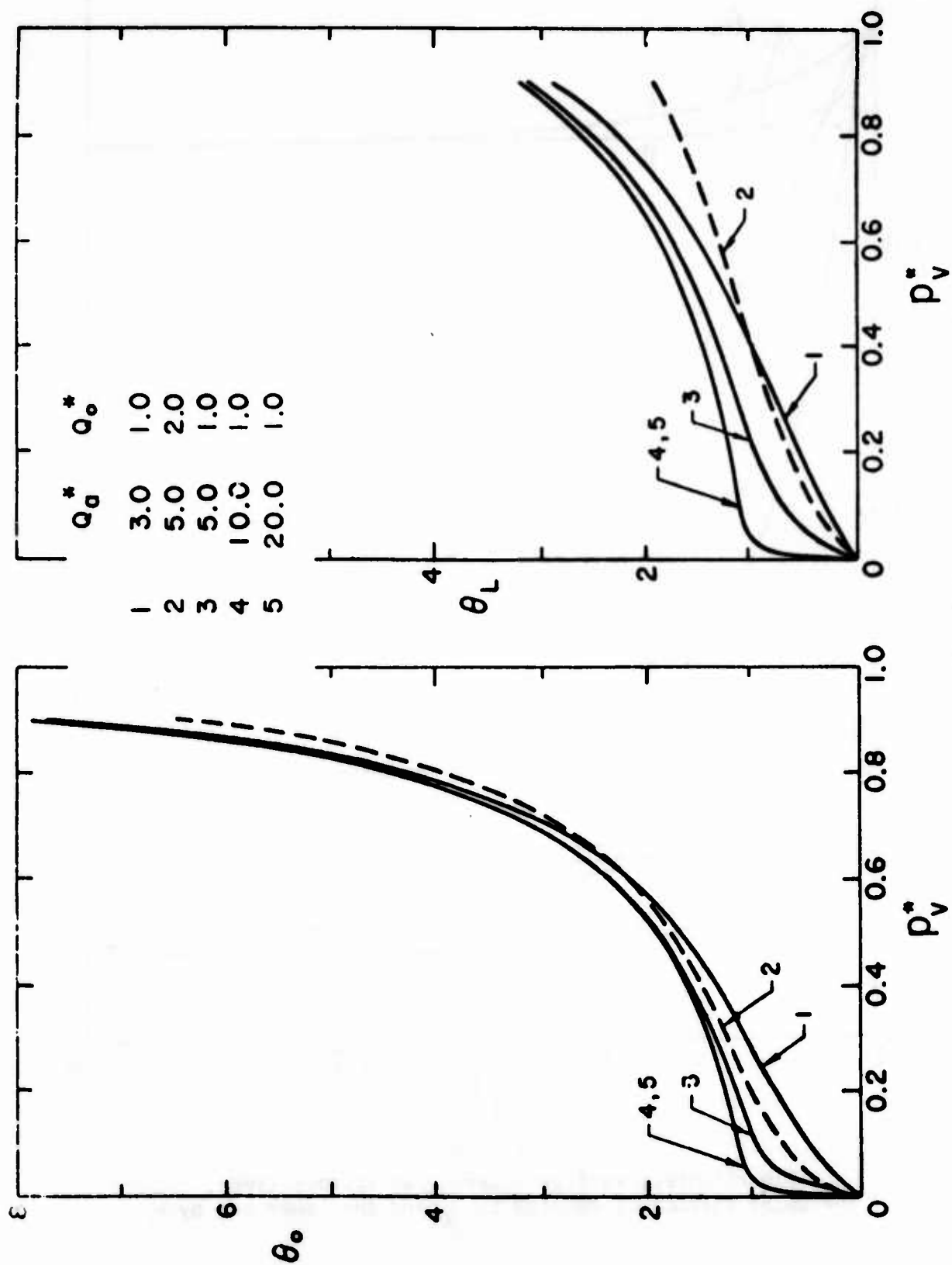
Fig. 5.3

Dimensionless surface coverage as a function of partial pressure of the vapor

860

Fig. 5.4

Volumetric concentration profile at various times

Fig. 5.5

Steady state average volumetric concentration c_{bulk}^* and maximum temperature, T_{max}^*, as function of heating level $\dot{q}^*$. Note that for $Q_s^* = 0.1$–0.5 an increase in heating level $\dot{q}^*$ from 0 to about 25 actually increases in the contaminant level in the steady state.

862

THE POISEUILLE FLOW OF A PARTICLE-FLUID MIXTURE — EFFECTIVE VISCOSITY

Donald A. Drew
Department of Mathematical Sciences
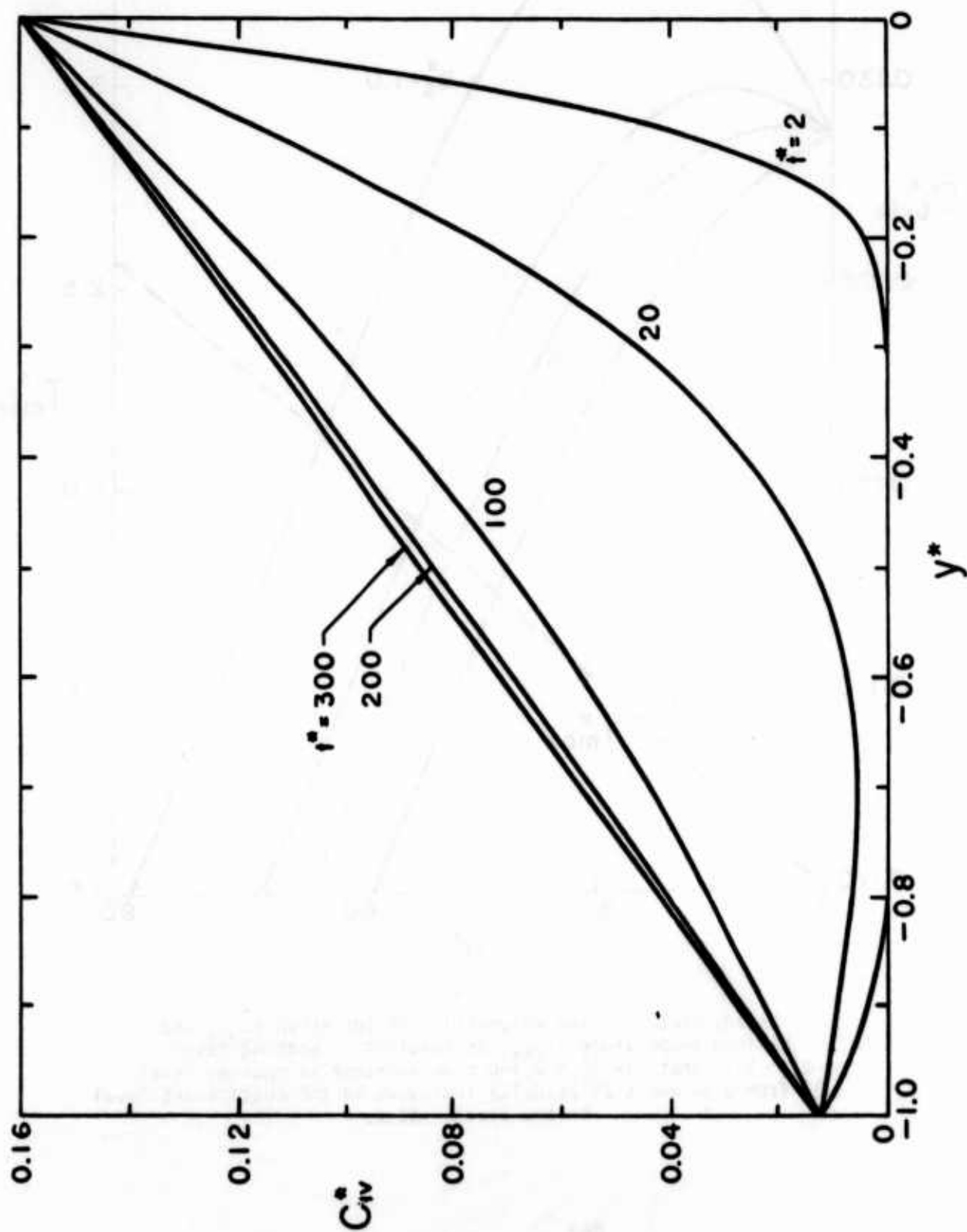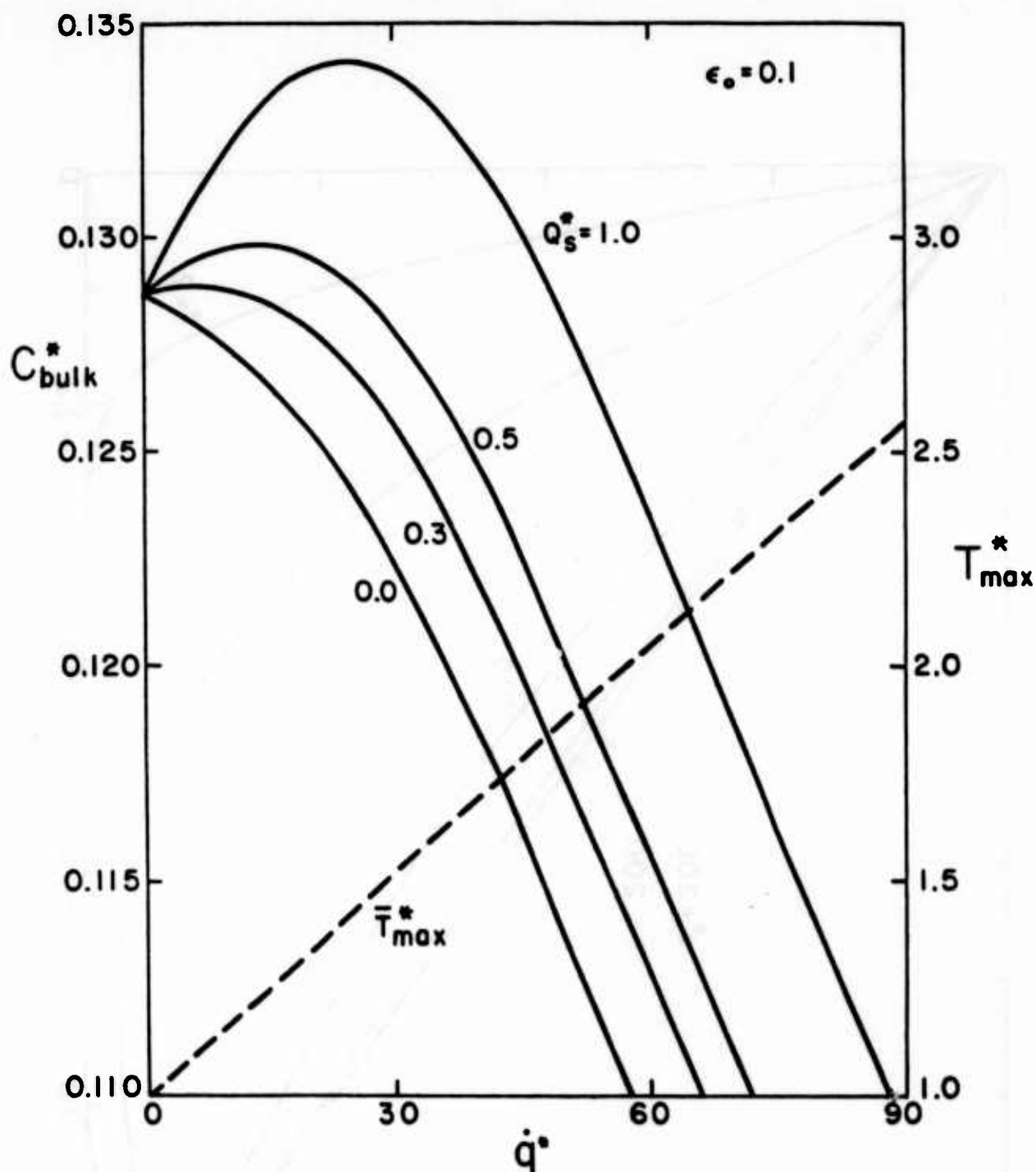Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Abstract. The equations of motion for the Poiseuille flow of a particle-laden fluid mixture are solved for a symmetric flow in a channel. We assume a formula for mixture viscosity as derived by Graham. For mixtures with particle concentrations below the effective maximum packing, this results in a flow with the particles concentrated at the core and a clear fluid layer at the walls separated by a transition layer where the particle concentration varies between the packed value in the core, and zero in the clear layer. This flow structure leads to a flow-pressure drop relation which can be interpreted as an "effective viscosity" for the mixture in Poiseuille flow. This points out the difficulty in measuring the mixture viscosity using such a flow. If the walls are porous, the flow leaking through the the walls leads to a force opposing the formation of the transition layer. If this force is sufficiently large, the clear layer does not form and particles can foul the wall, reducing the effectiveness of this device as a filter.

Introduction. The two-fluid model for the flow of a dispersed mixture is based on equations of conservation of mass and momentum for each material. These equations are assumed to govern the multidimensional motion of such mixtures, provided the correct constitutive equations are supplied. One test of the constitutive equations for a multidimensional flow is plane parallel flow, where the equations should be able to predict the velocity profiles and the concentration across the channel.

Equations of Motion. The mass and momentum equations for the flow of a particle-laden fluid mixture are:

$$(1) \quad \frac{\partial \alpha}{\partial t} + \nabla \cdot \alpha v_p = 0$$

$$(2) \quad \frac{\partial (1 - \alpha)}{\partial t} + \nabla \cdot (1 - \alpha) v_f = 0$$

$$(3) \quad \alpha \rho_p \left[\frac{\partial v_p}{\partial t} + v_p \cdot \nabla v_p \right] = -\alpha \nabla p_p + (p_{pi} - p_p) \nabla \cdot \alpha$$

$$+ \alpha S (v_f - v_p)$$

$$+ \alpha \rho_f C_{vm} \left[\left(\frac{\partial v_f}{\partial t} + v_f \cdot \nabla v_f \right) - \left(\frac{\partial v_p}{\partial t} + v_p \cdot \nabla v_p \right) \right]$$

863

$$+ \, \alpha \, \rho_f \, L \, (v_p - v_f) \, \rightarrow \, (\nabla \rightarrow v_f) + F_F + \nabla \cdot \alpha \, \tau_p$$

$$- \, \nabla \, \alpha \cdot \tau_{pi} + \nabla \cdot \alpha \, \tau_p^T - \alpha \, \rho_p \, g$$

$$(4) \quad (1 - \alpha) \, \rho_f \left[\frac{\partial v_f}{\partial t} + v_f \cdot \nabla \, v_f \right] = - \, (1 - \alpha) \, \nabla \, p_f$$

$$- (p_{fi} - p_f) \, \nabla \, \alpha \; + \, \alpha \, S \, (v_p - v_f)$$

$$+ \, \alpha \, \rho_f \, C_{vm} \left[\left[\frac{\partial v_p}{\partial t} + v_p \cdot \nabla \, v_p \right] - \left[\frac{\partial v_f}{\partial t} + v_f \cdot \nabla \, v_f \right] \right]$$

$$+ \, \alpha \, \rho_f \, L \, (v_f - v_p) \, \rightarrow \, (\nabla \rightarrow v_f) - F_F$$

$$+ \, \nabla \cdot (1 - \alpha) \, \tau_f + \nabla \, \alpha \cdot \tau_{fi}$$

$$+ \, \nabla \cdot (1 - \alpha) \, \tau_f^T - (1 - \alpha) \, \rho_f \, g$$

Here α is the volume fraction of particles, ρ_p is the particle density, ρ_f is the fluid density, v_p is the particle velocity, v_f is the fluid velocity, p_p is the particle pressure, p_{pi} is the pressure at the particle interface, p_f is the fluid pressure, p_{fi} is the fluid pressure at the interface, S is the drag coefficient, C_{vm} is the virtual mass coefficient, L is the lift coefficient, F_F is the Faxen force, g is the acceleration due to gravity, τ_p is the particle shear stress, τ_{pi} is the particle shear stress at the interface, τ_f is the fluid shear stress, τ_{fi} is the fluid shear stress at the interface, τ_p^T is the particle turbulent shear stress, and τ_f^T is the fluid turbulent shear stress.

The virtual mass and lift force is calculated by Drew and Lahey (1986) by considering the force on a single sphere accelerating relative to an inviscid fluid which is undergoing a pure shear plus a rotation. This results in $C_{vm} = 1/2$ and $L = 1/2$. For our present purposes, we shall assume that the virtual mass-lift combination is objective. This forces the choice $L = C_{vm}$.

The Faxen force is comparable to the viscous forces in the fluid phase. This term is not usually included in two-phase flow models, and consequently its form is not common knowledge. We assume

$$(5) \quad F_F = k \, \alpha \, \mu_f \, \nabla^2 \, v_f \, .$$

864

The pressure terms are very important for wave propagation in bubbly flows. Single-sphere calculations such as the ones used to determine the forms of the virtual mass and lift suggest;

(6) $p_p = p_{pi} = p_{fi} = p_f - \xi \, \rho_f \, |v_f - v_p|^2$.

Stuhmiller (1977) gives $\xi = 1/4$ for inviscid flow.

The stress terms are most difficult to model. If the particles are small and rigid, they are essentially stress transmitters on the microscale. Thus, we assume that the particle stress and the interfacial stresses are the same. Therefore,

(7) $\underline{T}_p = \underline{T}_{pi} = \underline{T}_{fi}$.

The stress which the particles are transmitting corresponds to the extra needed to make the mixture more viscous. To account for this effect, we take

(8) $\underline{T}_{pi} = \beta \, (\alpha) \, \underline{T}_f$.

For the viscous stress, we take Ishii's (1975) form, which is derived from averaging the microscopic viscous stress tensor. This gives:

(9) $\underline{T}_f = \mu \left\{ \left[\nabla \, v_f + (\nabla \, v_f)^T \right] + \nabla \alpha \, (v_f - v_p) + (v_f - v_p) \, \nabla \alpha \right\}$.

The Reynolds stresses are responsible for diffusive effects in the momentum balance. For the Reynolds stress in the fluid, we take a simple form of a model proposed by Drew and Lahey (1979) which has coefficients which can be calculated from inviscid flow around a single sphere Lamb(1933). This gives:

(10a) $\underline{T}_f^T = \alpha \, \rho_f \, a \, |v_f - v_p|^2 \, \underline{I} + \alpha \, \rho_f \, b \, (v_f - v_p)(v_f - v_p)$

For the Reynolds stress in the particle phase, we assume that the particles follow closely the motions of the fluid. This leads to:

(10b) $\underline{T}_p^T = \rho_p \, c \, |v_f - v_p|^2 \, \underline{I} + \rho_p \, d \, (v_f - v_p)(v_f - v_p)$

Plane Poiseuille Flow. Let us specify the flow conditions. We wish to examine the symmetric flow in a channel of width 2h. We shall neglect gravitational forces. Then the no slip condition on the fluid gives $v_f = 0$ and the condition of impenetrability of the wall to particles gives $n \cdot v_p = 0$ at $y = \pm h$. We also impose conditions that the total flow is given, and the concentration of the incoming fluid is known.

(11a) $\quad \int_{-h}^{h} \mathbf{i} \cdot v_f \, dy = 2 \, h \, V_0$

(11b) $\quad \int_{-h}^{h} \mathbf{i} \cdot v_p \, dy = 2 \, h \, U_0$

(11c) $\quad \int_{-h}^{h} \alpha (y) \, dy = 2 \, h \, \langle \alpha \rangle.$

For plane parallel flow, let

(12a) $\quad v_p = U(y) \, \mathbf{i}$

(12b) $\quad v_f = V(y) \, \mathbf{i}$

(12c) $\quad \alpha = \alpha(y).$

Then the continuity equations are satisfied automatically, and the momentum equations yield

(13) $\quad 0 = - \alpha \dfrac{\partial p_f}{\partial x} + \alpha \, S \, (V - U) + \alpha \, \mu \, \dfrac{d}{dy} \, \beta \, \dfrac{dV}{dy}$

$\quad\quad\quad\quad + \alpha \, k \, \mu \, \dfrac{d^2 V}{dy^2}$

(14) $\quad 0 = - (1 - \alpha) \dfrac{\partial p_f}{\partial x} + \alpha \, S \, (U - V) + \mu \, \dfrac{d}{dy} \left[(1 - \alpha) \, \dfrac{dV}{dy} \right]$

$\quad\quad\quad\quad \alpha \, k \, \mu \, \dfrac{d^2 V}{dy^2} + \dfrac{d\alpha}{dy} \, \beta \, \mu \, \dfrac{dV}{dy}$

(15) $\quad 0 = - \alpha \dfrac{\partial p_f}{\partial y} + 2 \, \alpha \, \xi \, \rho_f \, (V - U) \, \dfrac{d}{dy} \, (U - V)$

$\quad\quad\quad\quad + \alpha \, \rho_f \, L \, (U - V) \, \dfrac{dV}{dy} + \dfrac{d}{dy} \, \alpha \, \rho_p \, c \, (U - V)^2$

(16) $\quad 0 = - (1 - \alpha) \dfrac{\partial p_f}{\partial y} + \xi \, \rho_f \, (V - U)^2 \, \dfrac{d\alpha}{dy} + \alpha \, \rho_f \, L \, (V - U) \, \dfrac{dV}{dy}$

$\quad\quad\quad\quad + \dfrac{d}{dy} \, \alpha \, \rho_f \, a \, (U - V)^2$

It is straightforward to show that

(17) $\quad p_f = - \dfrac{\Delta p}{\Delta x} x + P(y)$,

where Δp is the imposed pressure drop on the channel and Δx is its length. The interfacial force terms can be eliminated by adding (13) and (14). This gives

(18) $\quad - \dfrac{\Delta p}{\Delta x} = \dfrac{d}{dy} \left[(1 - \epsilon) \mu \dfrac{dV}{dy} + \epsilon \beta \mu \dfrac{dV}{dy} \right]$.

If $\beta = 7/2$, eq. (18) is the Einstein (1906) formula for effective viscosity. We shall use a form given by Graham [8] as

(19) $\quad \mu_{eff}/\mu = \left(1 + \dfrac{5}{2} \epsilon \right) +$

$$\dfrac{9}{4} \left[1 + (h/2a) \right]^{-1} \left[\dfrac{1}{(h/a)} - \dfrac{1}{\left[2 + (h/a) \right]} - \dfrac{1}{\left[1 + (h/a) \right]^2} \right]$$

where, for a simple cubic packing,

$h/a = \quad 2 \left[(1 - (\epsilon/\epsilon_m)^{1/3})/(\epsilon/\epsilon_m)^{1/3} \right]$,

where ϵ_m is the experimentally determined maximum packing of spheres. This form agrees with Einstein's for small ϵ and with that derived by Frankel and Acrivos (1967) which agrees with data for larger concentrations.

The relative motion between the particles and the fluid can be obtained from either of the remaining momentum equations in the x-direction. If we divide equation (13) by ϵ S, we have

(20) $\quad U - V = \dfrac{1}{S} \dfrac{\Delta p}{\Delta x} + \dfrac{\mu}{S} \dfrac{d}{dy} (\beta + k) \dfrac{dV}{dy}$.

The momentum equations in the y-direction are instrumental in determining the distribution of particles across the channel. These equations involve the transverse pressure gradient dP/dy. The transverse pressure gradient can be eliminated from (15) and (16) by subtracting $\epsilon/(1 - \epsilon)$ times eq (16) from eq (15). This gives

(21) $\quad 0 = \left[\zeta + c - a \dfrac{\epsilon}{1 - \epsilon} \right] 2 \epsilon \rho_f (U - V) \dfrac{d(U - V)}{dy}$

$\qquad + \dfrac{\epsilon}{1 - \epsilon} \rho_f L (U - V) \dfrac{dV}{dy} +$

$\qquad + \left[- (\zeta + a) \dfrac{\epsilon}{1 - \epsilon} + c \right] \rho_f (U - V)^2$

We can further eliminate $U - V$ by using equation (20). This gives

867

(22) $$0 = \epsilon \left[\zeta + c - a \frac{\epsilon}{1 - \epsilon} \right] \frac{2}{S} \frac{\mu}{dy^2} (\beta + k) \frac{dV}{dy} - \frac{\epsilon}{1 - \epsilon} \frac{dV}{dy}$$

$$+ \left[- (\zeta + a) \frac{\epsilon}{1 - \epsilon} + c \right] \left[\frac{1}{S} \frac{\Delta p}{\Delta x} + \frac{\mu}{S} \frac{d}{dy} (\beta + k) \frac{dV}{dy} \right] \frac{d\epsilon}{dy}$$

Equations (18) and (22) govern the fluid velocity profile and the concentration of particles in the channel. The appropriate boundary conditions are

(23a) $$V(h) = 0$$

(23b) $$\int_{-h}^{h} V(y) \, dy = 2 \, h \, V_0$$

(23c) $$\int_{-h}^{h} \epsilon (y) \, dy = 2 \, h \, \langle \epsilon \rangle .$$

Let us nondimensionalize the problem by

(24a) $$w = V(y)/V(0)$$

(24b) $$\zeta = y/h .$$

The equations become

(25) $$\left(1 + (\beta - 1) \epsilon \right) \frac{dw}{d\zeta} = - R \zeta$$

(26) $$2 \epsilon \left[\zeta + c - a \frac{\epsilon}{1 - \epsilon} \right] \frac{d^2}{d\zeta^2} (\beta + k) \frac{dw}{d\zeta} + \epsilon \, D \frac{\epsilon}{1 + \epsilon} \frac{dw}{d\zeta}$$

$$+ \left[- (\zeta + a) \frac{\epsilon}{1 - \epsilon} + c \right] \left[R + \frac{d}{d\zeta} (\beta + k) \frac{dw}{d\zeta} \right] \frac{d\epsilon}{d\zeta}$$

where

(27a) $$R = \frac{\Delta p \, h^2}{\mu \, \Delta x \, V(0)}$$

is the channel Reynolds number, and

(27b) $$D = \frac{S \, h^2}{\mu}$$

is the dimensionless drag per unit velocity.

The boundary conditions are

(28a) $w(0) = 1$

(28b) $\dfrac{dw}{d\zeta}(0) = 0$

(28c) $w(1) = 0$.

(28d) $\displaystyle\int_0^1 \alpha(\zeta)\, d\zeta = \langle\alpha\rangle$.

Approximate Solution.   Let us seek a solution for D large.   The outer solution assumes $\zeta = O(1)$.   With D large, we must have β large, or dw/dζ small, or α small. From eq. (25), we see that if $R = O(1)$ and dw/dζ is small, β must be large.   With dw/d$\zeta = D^{-q}$ g, and $\beta = D^q$ b', we have $\alpha = \alpha_m + D^{-q} \alpha'$, and $b' = -(9/8)(\alpha_m/\alpha')$.   In order to obtain a balance in eq. (26), we must have q = 1.

Let us assume that the region with α small is near the wall, and that the region with dw/dζ small is in a region around the center of the channel, which we shall call the core.   Then in the core we have $\alpha = \alpha_m$.   Furthermore, the approximate particle concentration is given by

(29) $\alpha \cong \begin{cases} \alpha_m\ , & 0 < \zeta < \zeta^* \\ 0\ , & \zeta^* < \zeta < 1\ , \end{cases}$

where ζ^* is the location of the edge of the core.   In the clear fluid region, the fluid velocity must satisfy

(30) $\dfrac{dw}{d\zeta} = -R\,\zeta$

so that

(31) $w = -\dfrac{1}{2} R \left(\zeta^2 - 1\right)$

At ζ^*, we have $w(\zeta^*) = -(R/2)(\zeta^{*2} - 1) = 1$.   Thus, we see that $R = 2/(1 - \zeta^{*2})$.   If $1 - \zeta^* = O(1)$, we see that $R = O(1)$. Since $\langle\alpha\rangle = \alpha_m \zeta^*$, we have

(32) $R = -\dfrac{2}{\left(\left[\dfrac{\langle\alpha\rangle}{\alpha_m}\right]^2 - 1\right)}$

Note that R is not small when $\langle\alpha\rangle$ is near α_m .

The Transition Layer.   The crux of the argument is whether a layer exists at $\zeta = \zeta^*$ where α makes a transition from α_m to 0, while dw/dζ goes from 0 to $-R\,\zeta^*$.

We let $\zeta = \zeta^* + D^{-p} \zeta'$.   The right hand side of eq (25) becomes $-R\,\zeta^*$ to first approximation.   Using this, we obtain a balance in eq (26) for p = 1/2, and dw/dζ' can be eliminated to give

(33) $\quad \dfrac{d^2\alpha}{d\zeta^{\cdot 2}} + f_1(\alpha)\left[\dfrac{d\alpha}{d\zeta^{\cdot}}\right]^2 + f_2(\alpha) = 0$

where

(34a) $\quad f_1(\alpha) = \dfrac{f''(\alpha)}{f'(\alpha)} + \dfrac{\left[-(\zeta + a)\dfrac{\alpha}{1-\alpha} + c\right]}{\left[\zeta + c - a\dfrac{\alpha}{1-\alpha}\right]}\dfrac{1}{2}\alpha$

(34b) $\quad f_2(\alpha) = \dfrac{1}{2\left[\zeta + c - a\dfrac{\alpha}{1-\alpha}\right]}\dfrac{1}{1-\alpha}\dfrac{1}{1+(\beta-1)\alpha}\dfrac{1}{f'(\alpha)}$

(34c) $\quad f(\alpha) = \dfrac{\beta + k}{1 + (\beta - 1)\alpha}$

It is not known whether the solutions to eq (33) exhibit the proper behavior. A numerical solution to equation (33) with a = - 1/5, c = - 1/5, and ζ = 1/4 is shown in Fig 1.

When $\zeta^* = 1 - o(1)$, the clear-fluid layer no longer exists, and the wall lies in the transition layer. In this case, R need not be O(1). The analysis in this case is straight-forward, and is given in Drew (1986). Figure 2 is a plot of R versus <α>. The equations for the Poiseuille flow of a particle-fluid mixture in a channel results in a flow with a strong structure. The structure which occurs consists of a core of particles which are sufficiently "packed" so that they cannot shear, surrounded by a clear-fluid layer where all the shear occurs. The presence of this structure has several implications. First, if such a structure occurs in a flow, one must be careful in interpreting measurements of fundamental quantities such as mixture viscosity. Measurements of properties such as viscosity usually assume uniform conditions. Clearly, a flow with such a structure is not uniform. The quantity measured may be strongly dependent on the structure. Second, this flow is not useful for measuring any of the terms in the equations of motion that are responsible for the separation of particles and fluid, because the flow is so degenerate that none of these terms are really acting during this motion.

Finally, we note that the presence of such a structure in this flow may allow the efficient filtration of such a mixture using a flow-through membrane device. The situation is essentially plane Poiseuille flow, with a small amount of fluid drawn through the walls, which are assumed to be porous. Since the flow pushes particles away from the walls, if fluid can be drawn through the walls slowly enough that the transition layer is not disrupted, then no particles will stick to the wall to impede the further flow of fluid.

Figure 1.

Distribution of α in the transition layer.

Figure 2.

Reynolds number R versus $\langle\alpha\rangle$.

$\langle\alpha\rangle$

871
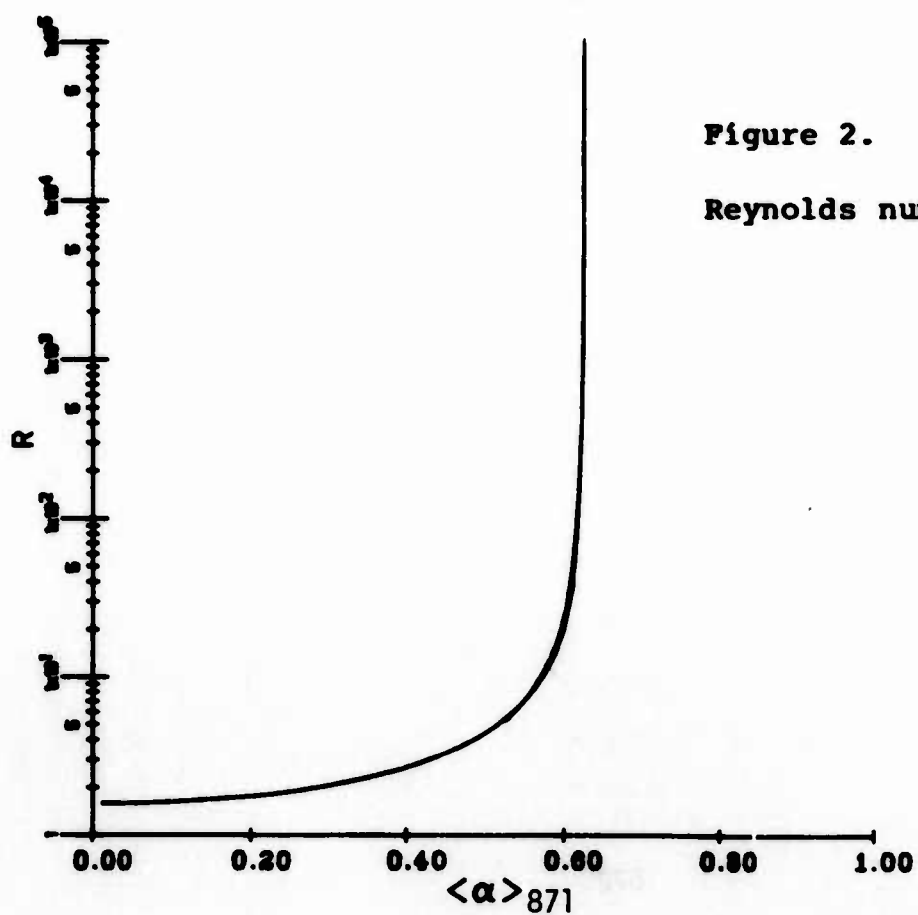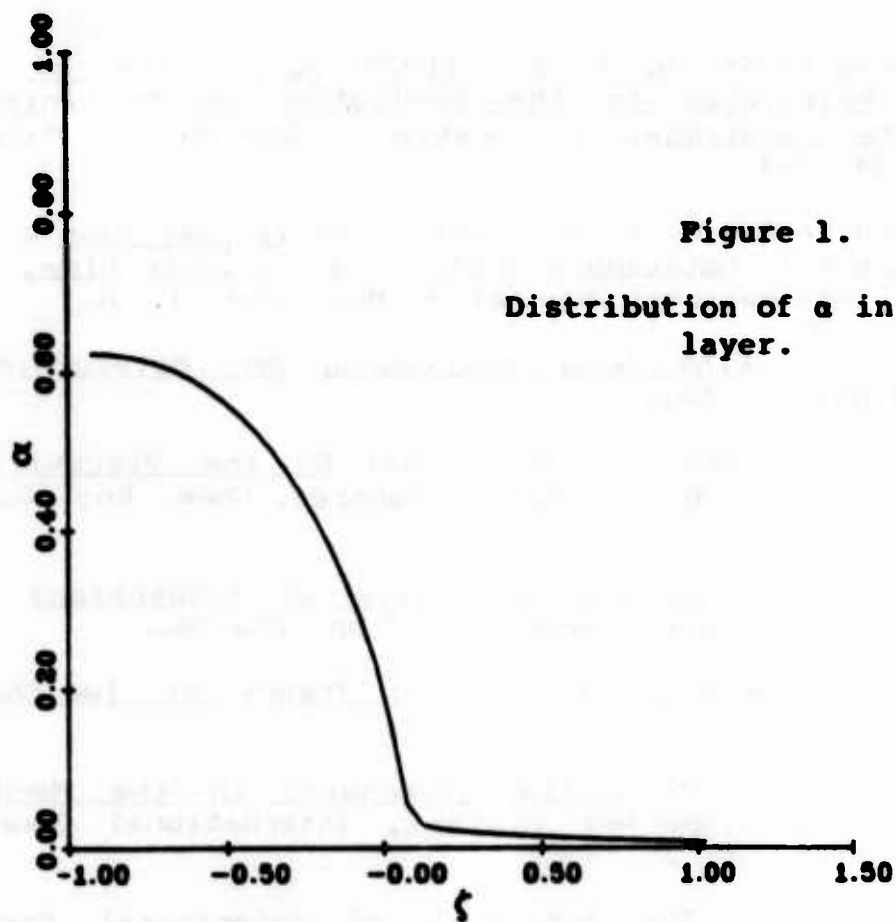
References

DREW, D. A. and LAHEY, R. T. Jr. (1979) <u>Application of General Constitutive Principles to the Derivation of Multidimensional Two-Phase Flow Equations</u>, International Journal of Multiphase Flow, 5 pp. 243-264.

DREW, D. A. and LAHEY, R. T. Jr. (1986) <u>The Virtual Mass and Lift Force on a Sphere in Rotating and Straining Inviscid Flow,</u> to appear in International Journal of Multiphase Flow.

EINSTEIN, A. (1906) <u>Eine Neue Bestimmung der Molekuldimensionen</u>, Ann. Phys. 19 pp. 289-306.

FRANKEL, N. A. and ACRIVOS, A. (1967) <u>On the Viscosity of a Concentrated Suspension of Solid Spheres</u>, Chem. Eng. Sci. 22 pp. 847-853.

GRAHAM, A. L. (1981) <u>On the Viscosity of Suspensions of Solid Spheres</u>, Applied Scientific Research 37 pp. 275-286.

ISHII, M. (1975) <u>Thermo-fluid Dynamic Theory of Two-Phase Flow</u>, Eyrolles, Paris.

NIGMATULIN, R. I. (1979) <u>Spatial Averaging in the Mechanics of Heterogeneous and Dispersed Systems</u>, International Journal of Multiphase Flow, 5 pp. 353-385.

STUHMILLER, J. (1977) <u>The Influence of Interfacial Pressure on the Character of Two-Phase Flow Model Equations</u>, International Journal of Multiphase Flow, 3 pp. 551-560.

SOME REMARKS ON BLOW-UP IN THE STEFAN MODEL FOR PHASE TRANSITIONS AND THE HELE-SHAW PROBLEM

S.D. Howison
Oxford University

Abstract

The classical Stefan model for melting or solidification and the closely related Hele-Shaw model can in certain circumstances exhibit irregular or ill-posed behaviour. Examples of this behaviour are given, and the effectiveness of several smoothing modifications of the models is discussed.

1.1 Introduction

The simplest model[1] for the solidification or melting of a pure substance is the classical 2-phase Stefan model which in suitable dimensionless variables is described by the equations

$$\frac{\partial u}{\partial t} = \nabla^2 u \tag{1.1}$$

in the phase regions $S(t)$ (solid) and $L(t)$ (liquid), with

$$u = u_m \tag{1.2}$$

and

$$\left[\frac{\partial u}{\partial n}\right]_S^L = -\lambda V_n \tag{1.3}$$

on the phase-change boundary $\Gamma(t)$ separating S from L. Here $u(\underset{\sim}{x},t)$ represents the material temperature, u_m is the fixed melting temperature, λ is the dimensionless latent heat and $\partial/\partial n$ is the derivative normal to Γ from S to L at a point whose speed in that direction is V_n; it is assumed that the material properties of S and L are the same. The formulation is completed by appropriate initial conditions $u_0(\underset{\sim}{x})$ and boundary conditions on the edge of the region in consideration or at infinity.

We note here two special cases of this general problem. Firstly, if the diffusion coefficient in the solid is negligible we obtain a 1-phase problem with $u \equiv u_m$ in $S(t)$; here[2] u may more realistically be thought of as the concentration of a dissolved substance diffusing through $L(t)$ with solidification or melting on Γ at an equilibrium concentration u_m. Secondly, if in addition the diffusion is fast compared to the timescale imposed by λ, we may ignore the $\partial/\partial t$ term in (1.1) and replace (1.1) by Laplace's equation. This is known as the Hele-Shaw problem (in two dimensions) and is also equivalent to flow of a viscous liquid through a porous medium[3]; u here represents the pressure in the liquid.

With the additional assumption that $u > u_m$ in the liquid and $u < u_m$ in the solid, the basic Stefan model is known to be well-posed at least for small times[4]. Nonetheless, there are some circumstances in which related problems exhibit irregular

(ill-posed) behaviour, and three of the more important of these are:

      (1) In the presence of superheating or supercooling,
      (2) When there is volumetric heating,
      (3) When impurities are present,

and we examine these in turn. The one-phase Stefan problem is correspondingly ill-posed if the phase in which u is not constant is superheated or supercooled, while the Hele-Shaw problem is well-posed if the fluid region is expanding, and ill-posed if it is contracting.

1.2 Supercooling and superheating

A liquid is supercooled if its temperature is less than u_m; we deal only with supercooling here since it is more common and since the corresponding results for superheating can be obtained by reversing the sign of $u - u_m$.

The solution to the Stefan problem with supercooling can blow up in finite time in two ways.

(a) Sherman blow-up[5]

Under certain conditions it is possible to show that the whole phase boundary Γ may move with infinite speed at a finite time $t^* < \infty$, and that there is no solution to the problem for $t > t^*$. This form of blow-up is known to occur in one-dimensional and radially symmetric geometries, and its cause is that the total energy stored in the system in the form of latent heat is insufficient to raise the supercooled liquid to its melting point. Consider for example a finite solid region in which $u \equiv u_m$ immersed in liquid whose temperature at infinity is u_∞: blow-up in finite time can be shown to occur whenever $u_\infty - u_m < -\lambda$ (undercoolings of up to -2λ can be achieved with certain materials[6],[7]). To show this, let $Q(t) = \int_{L(t)} (u - u_\infty)$ which is positive for suitable initial data $u_0(\underset{\sim}{x}) \geq u_\infty$; on the other hand (using (1.1)-(1.3), $dQ/dt = $ (the rate at which the area of $S(t)$ increases). $(u_\infty - u_m + \lambda)$ so that if the solution exists for all t and $u_\infty - u_m < -\lambda$, $Q \to -\infty$. This contradiction shows that blow-up must occur at a finite time $t^* < \infty$. If in addition the problem has planar, cylindrical or spherical symmetry we see that the velocity of Γ must become infinite at $t = t^*$. This kind of blow-up is not, however, possible for the Hele-Shaw problem since it relies on the $\partial/\partial t$ term in (1.1).

(b) Cuspidal blow-up

The Sherman blow-up with $V_n \to \infty$ on all Γ has hitherto only been shown to occur in symmetric geometries. Nevertheless, the argument leading to blow-up given above does not depend upon symmetry (until the last line which argues only that $V_n \to \infty$).

The form taken by blow-up when Γ is not symmetric is thought to be via a cusp in Γ with infinite V_n at its tip; this is known to be the case for the Hele-Shaw problem.[4] Although this may seem like a pointwise version of Sherman blow-up, it in fact

874

occurs even if the conditions leading to Sherman blow-up do not hold, that is even if there is no deficiency of latent heat in the system. Indeed, it is likely that the set of initial value problems for supercooled liquids or receding Hele-Shaw flows which do not blowup in this way is small.[6],[10] Complex variable methods can be used to investigate blow-up in Hele-Shaw flows in some detail;[10] typically a 3/2-power cusp forms in the moving boundary and the solution ceases to exist at that time, although in certain special circumstances other kinds of cusp can appear momentarily without this non-existence. All, however, have infinite fluid velocities at their tips, and will thus be prevented in practice by surface tension and inertial effects.

Even if blow-up does not occur, the moving boundary is unstable to small perturbations which for large wavenumber n grow as e^{int}; consequently the morphology of the moving boundary may be complicated. A possible situation here both for Stefan and Hele-Shaw is an array of parallel 'fingers' as shown in Fig. 1; we will return to this point later.

S   L

Fig. 1

1.3 Volumetric Heating

If we impose a volumetric heating Q, so that $\partial u/\partial t = \nabla^2 u + Q$ in each phase, the liquid particle whose temperature first reaches u_m must either become superheated or remain at that temperature for a time λ/Q until it has acquired enough energy to change phase [11],[12]. In the first of these cases the classical Stefan model has the difficulties described above; the second is not even possible within the classical framework. This situation has no direct analogy with any Hele-Shaw type flow since it is inherently a two-phase problem and the term $\frac{\partial u}{\partial t}$ in (1-1) is essential to the argument just given; nevertheless there is a remote similarity with the squeeze film problem described in ref. 6., but we do not pursue this point here.

1.4 Impurities

The simplest model for solidification of a dilute binary alloy consists of equations (1.1), (1.3) for the heat flow, together with diffusion of the impurity in solid and liquid phases. The diffusion problems are coupled through the conditions on the phase boundary, and (1.2) is replaced by a relationship between the melting temperature on the interface and the concentration there. We do not go into the details save to remark that in most practical situations the model predicts 'constitutional supercooling,' and that in the absence of

surface tension the interface may be linearly unstable[13] in the same way as a supercooled liquid, with a linear growth rate approximately e^{int}. We therefore conjecture that both cuspidal and Sherman blow-up are possible although this remains unproven. Again there is no analogy with a Hele-Shaw type flow.

2. Regularisations of the models

The two kinds of blow-up described above are physically unrealistic. The simple model (1.1)-(1.3) (or its amended versions) is plainly inadequate, and we seek to modify it in such a way as to incorporate hitherto neglected physical effects and/ or to render it mathematically better behaved. We describe three such modifications, two based on extra physics and one more mathematically motivated.

2.1 Surface tension

Surface energy effects may be incorporated into (1.2) via the Gibbs-Thompson condition

$$u = u_m(1-\gamma k) \text{ on } \Gamma, \tag{2.1}$$

where k is the appropriately signed curvature of Γ and γ is a dimensionless surface tension. This has a dramatic effect on the linear stability of Γ in that only a finite band of larger wavelengths is now unstable, and it almost certainly prevents cuspidal blow-up both for Stefan and Hele-Shaw, although this has not yet been rigorously shown for either problem. On the other hand, examples can be given where the Sherman blow-up is <u>not</u> prevented by surface tension. This can be demonstrated in, for example, a spherical geometry using a version of the argument given in section 1.2; the physical interpretation is that the energy stored in Γ is not sufficient to materially alter the energy imbalance which is the reason for this form of blow-up. A version of this argument can be carried through for regions without symmetry, and we conjecture that, for the 1-phase Stefan problem at least, the only way to avoid Sherman blow-up is for S(t) to split into infinitely many disjoint components, each a sphere of radius R, where $2\gamma/R = |u_\infty|$. This seems the only plausible equilibrium configuration; it bears some resemblance to the ripening process described by Glicksman[15].

2.2 Kinetic Undercooling

A second approach is to modify (1.2) by introducing a kinetic undercooling on Γ, so that the melting temperature is now

$$u = u_m - \frac{1}{\mu}V_n;$$ (2.2)

this represents the fact that the interface departs slightly from thermodynamic equilibrium. It is not, however, used for Hele-Shaw flows as its physical basis there has not been established.

With this condition on Γ the Stefan model avoids both Sherman and cuspidal blow-up[14]; it works because the kinetic term $-V_n/\mu$ allows a greater energy transfer across Γ when V_n is large, which is a stabilizing process. Its only practical limitation is that μ is usually so large that V_n must be about 10m/sec before V_n/μ is significant. Kinetic undercooling was doubtless significant in the experiments of Glicksman[8,9] with $u_\infty - u_m < - \lambda$. Both (2.1) and (2.2) are special cases of the phase field model of Caginalp[15].

2.3 Weak Solutions

A modification which is more purely mathematical in its approach is the idea of a weak solution; it works particularly well for Stefan problems involving volumetric heating. We rewrite (1.1)-(1.3) (with heating) in the form

$$\frac{\partial h}{\partial t} = \nabla^2 u + Q$$ (2.3)

where h is the enthalpy, defined by $h = u + \lambda H(u-u_m)$, H being the Heaviside function. Equation (2.3) is to be interpreted in the sense of distributions, and this can lead to solutions which are not consistent with the classical formulation (1.1)-(1.3). Thus for instance the solid particle whose temperature first reaches u_m remains at that temperature while its enthalpy increases continuously from u_m to $u_m + \lambda$. Neighbouring particles also have this behaviour, and the result is a 'mushy region' in which $u \equiv u_m$ but h varies. This formulation is well suited to numerical solutions since no special treatment is necessary to follow the free boundaries (solid-mush and mush-liquid, or solid-liquid). Its physical interpretation can, on the other hand, be a difficulty, one approach[12] being to regard the mush as a mixture of liquid and superheated regions, each of the latter being small enough to be stabilized by surface tension. The enthalpy method does not, however, explicitly incorporate the effects of the surface energy stored in Γ into the definition of h, and a more realistic definition would take account of the variation of this energy as the solid volume fraction changes; this might involve a model of a ripening process similar to that described by Glicksman[15].

Finally we note that attempts to find a weak formulation for the binary alloy problem (a worthwile goal in view of its potential numerical effectiveness) have not hitherto succeeded.

3. Conclusion

The basic Stefan model (1.1)-(1.3) is mathematically well understood, but it is physically unrealistic in that it predicts finite time blow-up of two kinds. The corresponding Hele-Shaw model suffers from only one of these.

Surface tension is probably an effective regularisation in all but the extreme situation of Sherman blow-up. Nevertheless, there is little rigorous mathematics on this version of the problem, and some subtle and interesting questions remain to be answered. Among these are to explain the mechanism by which cusps are prevented and the question of the selection of the width of dendrites in an array such as that of fig. 1.

Kinetic undercooling is also an effective regularisation for Stefan problems but only comes into play at high interface speeds.

Probably the safest condition for Stefan problems to take is the combination of surface tension and kinetic undercooling

$$u = u_m(1 - \gamma k) - V_n/\mu,$$

and this condition is discussed by Caginalp[15]. For Hele-Shaw problems the term V_n/μ should be ignored.

Weak solutions work well for volumetric heating but their extension to include surface energy and impurity effects has yet to be accomplished.

Acknowledgements

References

1. L. Rubinstein, "The Stefan Problem', Transl. Math. Monogrpahs 27 Am. Math. Soc., Providence, R.I. 1971.

2. W.W. Mullins & R.F. Sekerka, 'Morphological stability of a particle growing by diffusion or heat flow', J. Appl. Phys. 34, 323-328, 1963.

3. P.G. Saffman & G.I. Taylor, 'The penetration of a fluid into a porous medium or Hele-Shaw cell containing a more viscous liquid', Proc. R. Soc. Lond. A 245, 312-329, 1958.

4. A.M. Meirmanov, 'On a free boundary problem for a parabolic equation' (in Russian), Matem. Sb. 115, 532-543, 1981.

5. B. Sherman, 'A general one-phase Stefan problem', Quart. Appl. Math. 27, 427-439, 1970.

6. S.D. Howison, A.A. Lacey & J.R. Ockendon, "Singularity development in moving boundary problems', Q.J. Mech. Appl. Math. 38, 343-360, 1985.

7. S.D. Howison & J. Chadam, 'Existence and stability results for spherical crystals growing in a supersaturated solution', Preprint, 1986.

8. M.E. Glicksman & R.J. Schafer, J. Chem. Phys. 45, 2367. 1966.

9. _____, _____, J. Crystal Growth 1, 297, 1967.

10. E. DiBenedetto & A Friedman, "The ill-posed Hele-Shaw model and the Stefan problem for super-cooled water', Trans. Am. Math. Soc. 282, 183-204, 1984.

11. D.R. Atthey, 'A finite difference scheme for melting problems', J. Inst. Maths. Applcs. 43, 143-158, 1985

12. A.A. Lacey & A.B. Tayler, 'A mushy region in a Stefan problem, I.M.A. J. Appl. Math. 30, 303-314, 1983.

13. W.W. Mullins & R.F Sekerka, 'Stability of a planar interface during solidification of a dilute binary alloy', J. App. Phys. 35, 444-451, 1964.

14. A. Visintin, 'Stefan problem with a kinetic condition at the interface', I.N.A. of C.N.R. Pavia preprint, 1985.

15. Paper to appear in Proceedings of the NATO Workshop on Structure and Dynamics of Partially Solidifed Systems, Lake Tahoe, 1986.

16. J.R. Ockendon & A.B. Crowley, 'Modelling Mushy regions', to appear in proceedings of meeting on Mixing, Stirring and and Solidification in Metallurgical Processes, Cambridge 1985.

17. C.M. Elliott and V. Janovsky, A variational inequality approach to the Hele-Shaw flow with a moving boundary, Proc. Roy. Soc. Edinburgh A 88, 93-107, 1981.

18. S.D. Howison, Cusp development ir Hele-Shaw flow with a free surface, SIAM J. Appl. Math. 46, 20-26, 1986.

GLOBAL OPTIMIZATION USING AUTOMATIC DIFFERENTIATION AND INTERVAL ITERATION

L. B. Rall

Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53705

Abstract. Algorithms are presented which find one or all of the critical points of a smooth function in a rectangular region, or the critical points at which the function has a maximum or minimum value. If no critical points of the function exist in the given region, then the algorithm verifies this fact. The computation is self-validating, in that the existence or nonexistence of critical points is established conclusively, and guaranteed upper and lower bounds are computed for all quantities of interest, including the values of the gradient vector and Hessian matrix of the function. The algorithms make use of an existing implementation of automatic differentiation and interval computation. Numerical results are given.

1. Preliminaries. This paper presents an algorithm for global, unconstrained optimization of a smooth (at least twice differentiable) function $f : \mathbf{R}^n \to \mathbf{R}$, that is,

$$(1.1) \qquad f(x) = f(x_1, x_2, \ldots, x_n).$$

As is well-understood, this also includes the case of optimization of a function $\phi : \mathbf{R}^m \to \mathbf{R}$ subject to $n - m$ smooth constraints

$$(1.2) \qquad g_i(x_1, x_2, \ldots, x_m) = 0, \qquad i = 1, 2, \ldots, n - m,$$

by formation of the function

$$(1.3) \qquad f(x) = \phi(x_1, \ldots, x_m) + \sum_{i=1}^{n-m} x_{m+i} \cdot g_i(x_1, \ldots, x_m),$$

where the new variables $x_{m+1}, \ldots, x_n$ are simply the Lagrange multipliers for the problem. No special properties of f, such as convexity, are assumed.

The method to be used is a *critical point* method, which will find one or all solutions of the system of equations

$$(1.4) \qquad \nabla f(x) = 0$$

in a rectangular region $X \subset \mathbf{R}^n$, where $\nabla f(x)$ denotes the *gradient vector*

$$(1.5) \qquad \nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)^T,$$

or just the critical points at which the value of f is a maximum or minimum in X. Such points will be called *critical extremal points* to distinguish them, if necessary, from non-critical points on the boundary ∂X of X at which f might attain a maximum or minimum value.

The algorithm will make use of automatic differentiation [11] to compute the gradient vector $\nabla f(x)$ of f at $x = (x_1, x_2, \ldots, x_n)$, and also its *Hessian matrix*

$$(1.6) \qquad Hf(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right).$$

This technique will be combined with the use of interval arithmetic and interval evaluation of library functions [8] in order to compute guaranteed bounds for values of functions and their derivatives over the region of interest. The result will be an automatic, self-validating optimization algorithm.

Automatic differentiation has been used, at least in a restricted form, by McCormick [6] for optimization problems. Interval methods have been applied by Hansen [2], [3] and Hansen and Sengupta [4] to global optimization problems, including constrained problems. Although the basic algorithm given below is for unconstrained problems, the ideas presented by Hansen indicate the possibility of introducing constraints into the calculations.

2. Automatic Differentiation. The basic idea behind automatic differentiation is to use the formula or subroutine for the evaluation of the function f at x to obtain also values of its derivatives at the same point. This is done by the introduction of a new representation of variables, and arithmetic operations which include the rules for differentiation. The resulting computational scheme is simple to program for computers [11], [13], and avoids both the complexity of symbolic differentiation and the inaccuracy of numerical differentiation. The new variables are *triples*

$$(2.1) \qquad U = (u, u', u''),$$

where $u \in \mathbf{R}$ is a real number, $u' \in \mathbf{R}^n$ is an n-dimensional real (column) vector, and u'' is a symmetric real $n \times n$ matrix. The set of these elements will be denoted by $\mathbf{H}^n$, and each $u \in \mathbf{H}^n$ is said to be of type HESSIAN. A variable U of type HESSIAN will be interpreted in the following way: Its first component u will represent the value of a real-valued function at some point $x \in \mathbf{R}^n$, and u' and u'' the values of its gradient vector and Hessian matrix, respectively, at the same point.

It is obvious that $\mathbf{H}^n$ forms a linear space. More importantly, *all* the standard arithmetic operations can be defined in $\mathbf{H}^n$:

$$(2.2) \qquad U + V = (u, u', u'') + (v, v', v'') = (u + v, u' + v', u'' + v''),$$

$$(2.3) \qquad U - V = (u, u', u'') - (v, v', v'') = (u - v, u' - v', u'' - v''),$$

$$(2.4) \qquad \begin{aligned} U \cdot V &= (u, u', u'') \cdot (v, v', v'') \\ &= (u \cdot v, u \cdot v' + v \cdot u', u \cdot v'' + u'v'^T + v'u'^T + v \cdot u''), \end{aligned}$$

$$(2.5) \qquad \begin{aligned} U/V &= (u, u', u'')/(v, v', v'') \\ &= \left(\frac{u}{v}, \frac{v \cdot u' - u \cdot v'}{v^2}, \frac{v^2 \cdot u'' - v \cdot (v'u'^T + u'v'^T) + 2u \cdot v'v'^T - uv \cdot v''}{v^3} \right), \\ v &\neq 0. \end{aligned}$$

The above definitions implement the rules for evaluation and differentiation of sums, differences, products, and quotients of functions with known values and derivatives. In order to use an algorithm for evaluation of a real function to obtain the corresponding values in $\mathbf{H}^n$, it is necessary to be able to represent the independent variables x_i, $i = 1, 2, \ldots, n$ and constants c as elements of $\mathbf{H}^n$. This is done by the mapping

$$(2.6) \qquad x_i \mapsto (x_i, e_i, 0),$$

for the ith independent variable x_i, where e_i denotes the ith unit vector, and 0 the $n \times n$ zero matrix. (0 will be used to denote zero vectors and matrices, as well as the real number zero.) Similary, constants c are represented by

$$(2.7) \qquad c \mapsto (c, 0, 0).$$

It follows that calculation of the value, gradient vector, and Hessian matrix of a rational function can be done simply by making the substitutions (2.6) and (2.7), and applying the rules (2.2)–(2.5). The results are exact, not numerical approximations, and are obtained without symbolics.

In actual practice, instead of using the representation (2.7) for constants, it is simpler to define a *mixed arithmetic* between elements $c \in \mathbf{R}$ and $U = (u, u', u'') \in \mathbf{H}^n$ [13]:

$$(2.8) \qquad c + U = U + c = (c + u, u', u''),$$

$$(2.9) \qquad c - U = (c - u, -u', -u''),$$

$$(2.10) \qquad U - c = (u - c, u', u''),$$

$$(2.11) \qquad c \cdot U = U \cdot c = (c \cdot u, c \cdot u', c \cdot u''),$$

$$(2.12) \qquad c/U = \left(\frac{c}{u}, -\frac{c \cdot u'}{u^2}, \frac{2c \cdot u'u'^T - cu \cdot u''}{u^3} \right), \quad u \neq 0,$$

$$(2.13) \qquad U/c = \left(\frac{u}{c}, \frac{u'}{c}, \frac{u''}{c} \right), \quad c \neq 0.$$

For example, consider the two-dimensional Rosenbrock function ([1], p. 95):

$$(2.14) \qquad f(x) = 100(x_2 - x_1)^2 + (1 - x_1)^2.$$

In order to evaluate this function together with its gradient vector and Hessian matrix at the point $x = (-1.2, 1.0)$, one sets

$$(2.15) \qquad x_1 = \left(-1.2, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right), \qquad x_2 = \left(1.0, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right),$$

and evaluates (2.14) using the above rules. The result is

$$(2.16) \qquad \left(f(x), \nabla f(x), Hf(x) \right) = \left(24.2, \begin{pmatrix} -215.6 \\ -88.0 \end{pmatrix}, \begin{pmatrix} 1330.0 & 480.0 \\ 480.0 & 200.0 \end{pmatrix} \right),$$

which is exactly what one would get by differentiating (2.14) symbolically and then evaluating the results for $x_1 = -1.2$, $x_2 = 1.0$ in real arithmetic.

In addition to rational functions of several variables, other standard functions can be defined readily on $\mathbf{H}^n$. For example,

$$(2.17) \qquad \sin U = \sin(u, u', u'') = \left(\sin u, \cos u \cdot u', \cos u \cdot u'' - \sin u \cdot u'u'^T \right).$$

In general, if $g : \mathbf{R} \rightarrow \mathbf{R}$ is twice differentiable, then it can be extended immediately to the mapping $g : \mathbf{H}^n \rightarrow \mathbf{H}^n$ by use of the chain rule:

$$(2.18) \qquad g(U) = g((u, u', u'')) = \left(g(u), g'(u) \cdot u', g'(u) \cdot u'' + g''(u) \cdot u'u'^T \right),$$

[11], [13].

It is easy to program automatic differentiation in languages such as Ada and Pascal-SC [13], which permit introduction of data types and additional definitions of the standard operator symbols to manipulate such types. (This is sometimes called "overloading" the standard operator symbols.) In these languages, the variables x_1 and x_2 in (2.14) would be declared to be of type HESSIAN, along with the result f, and the evaluation would be carried out on the basis of an expression of the same form as (2.14). In ordinary Pascal

or FORTRAN, (2.14) would have to be rewritten as a sequence of calls to subroutines for addition, exponentiation, etc. [11]. The algorithms described in this paper have been programmed in Pascal-SC, and some results are given in the final section.

3. Interval Computation. In ordinary optimization algorithms, the function to be optimized is sampled only at a discrete set of points. This can result in the loss of valuable information about the function. The algorithms presented in this paper, on the other hand, use interval computation, which produces guaranteed bounds for the values of functions and their derivatives over entire regions [8]. This prevents the process from being misled by incomplete information.

The basic component of interval computation is *interval arithmetic* [8]. Let **IR** denote the set of bounded, closed intervals on the real line **R**. For $I = [a, b] \in \mathbf{IR}$, $J = [c, d] \in \mathbf{IR}$, the arithmetic operations are defined by

$$(3.1) \qquad I \star J = [a, b] \star [c, d] = \{x \star y \mid x \in I,\ y \in J\} = [r, s],$$

where $\star \in \{+, -, \cdot, /\}$, and division by an interval containing 0 is excluded. In actual implementation on computers, directed rounding is used (downward for lower endpoints, upward for upper endpoints), so the actual result computed is $[\nabla r, \Delta s]$, which always contains the exact result $[r, s]$ of the interval operation.

Evaluation of a real rational function $f : \mathbf{R} \to \mathbf{R}$ in interval arithmetic results in an *interval inclusion* $F : \mathbf{IR} \to \mathbf{IR}$ of f [8], which has the property

$$(3.2) \qquad f(X) = \{f(x) \mid x \in X\} \subseteq F(X), \qquad X \in \mathbf{IR}.$$

Denoting the endpoints of an interval $I = [a, b]$ by $\inf I = a$, $\sup I = b$, respectively, (3.2) means that

$$(3.3) \qquad \inf F(X) \leq f(x) \leq \sup F(X), \qquad x \in X.$$

These bounds for the range of $f(x)$ over X are obtained *automatically*, without investigation of the minimum and maximum values of $f(x)$ on X, and are furthermore *guaranteed* (although they may be somewhat crude) [8]. This is the basis of the self-validating character of interval computation. Furthermore, interval extensions obtained by using interval arithmetic are *monotone* in the sense that $X \subseteq Y$ implies that $F(X) \subseteq F(Y)$. In exact arithmetic, F is an *extension* of f in the sense that $F([x, x]) = f(x)$ for $x \in \mathbf{R}$ [8]. In what follows, x will be used to denote the degenerate interval $[x, x] \in \mathbf{IR}$ as well as the real number x. Other handy notations to be used from time to time are

$$(3.4) \qquad w(I) = w([a, b]) = b - a, \qquad m(I) = m([a, b]) = \frac{a + b}{2},$$

for the width and midpoint, respectively, of an interval $I \in \mathbf{IR}$.

Just as in the case of differentiation arithmetic, interval arithmetic can be extended to include various standard functions encountered in applications. Efficient implementations of interval arithmetic and interval inclusions of standard functions are now available in a

885

number of computational environments, for example, Pascal-SC for microcomputers and the ACRITH package for IBM 370 computers.

The space $\mathbf{IR}^n$ of interval vectors $X = (X_1, X_2, \ldots, X_n)$ is defined in the same way as $\mathbf{R}^n$, and the notions of interval matrices and vector and matrix-vector interval arithmetic arise in a natural way. The *interval scalar product* of interval vectors X, Y is defined to be

$$(3.5) \qquad X \cdot Y = \left\{ \sum_{i=1}^{n} x_i y_i \;\middle|\; x_i \in X_i, y_i \in Y_i \right\},$$

and the notation $m(X)$ will be used for the *midpoint*

$$(3.6) \qquad m(X) = \big(m(X_1), m(X_2), \ldots, m(X_n)\big)$$

of the interval vector X.

Now, if the evaluation of the function (2.14) is performed in interval arithmetic with $x_1 = [0.9, 1.2]$, $x_2 = [0.8, 1.1]$, then the result is

$$(3.7) \qquad F(X) = [0.0, 41.0],$$

where X denotes the interval vector $X = ([0.9, 1.2], [0.8, 1.1])$. This means that

$$(3.8) \qquad 0 \le f(x) \le 41$$

for $0.9 \le x_1 \le 1.2$, $0.8 \le x_2 \le 1.1$. Thus, the bounds (3.8) are obtained automatically, simply by evaluation of (2.14) in interval arithmetic, in much the same way that values of the gradient vector and Hessian matrix of (2.14) were obtained in §2 by the use of differentiation arithmetic. Furthermore, as stated above, the bounds given by interval arithmetic are guaranteed to be valid.

The next step is to combine the differentiation arithmetic in §2 with interval arithmetic. An element Υ of type IHESSIAN will be a triple

$$(3.9) \qquad \Upsilon = (U, U', U''),$$

where $U \in \mathbf{IR}$ is an interval, $U' \in \mathbf{IR}^n$ is an interval (column) vector, and U'' is a symmetric interval $n \times n$ matrix. The resulting set of elements will be denoted by $\mathbf{IH}^n$. Arithmetic operations in $\mathbf{IH}^n$ are defined by (2.2)–(2.5), with the operations inside the parentheses replaced by the interval operations (3.1). Similarly, operations between constants $c \in \mathbf{IR}$ and elements of $\mathbf{IH}^n$ are defined by (2.8)–(2.13) and the corresponding interval operations. Real constants c are mapped into $\mathbf{IR}$ by $c \mapsto [c, c]$, as before.

For example, the evaluation of (2.14) as type IHESSIAN can be carried out over the intervals $0.9 \le x_1 \le 1.2$, $0.8 \le x_2 \le 1.1$ by setting

$$
\begin{aligned}
x_1 &= \left([0.9, 1.2], \begin{pmatrix} [1,1] \\ [0,0] \end{pmatrix}, \begin{pmatrix} [0,0] & [0,0] \\ [0,0] & [0,0] \end{pmatrix} \right), \\[2mm]
x_2 &= \left([0.8, 1.1], \begin{pmatrix} [0,0] \\ [1,1] \end{pmatrix}, \begin{pmatrix} [0,0] & [0,0] \\ [0,0] & [0,0] \end{pmatrix} \right).
\end{aligned}
$$

(3.10)

The result is

$$\Phi(X) = (F(X), F'(X), F''(X)) =$$

$$(3.11) \quad = \left([0.0, 41.0], \begin{pmatrix} [-139.4, 307.6] \\ [-128.0, 58.0] \end{pmatrix}, \begin{pmatrix} [534.0, 1410.0] & [-480.0, -360.0] \\ [-480.0, -360.0] & [200.0, 200.0] \end{pmatrix} \right),$$

where X denotes the interval vector $X = ([0.9, 1.2], [0.8, 1.1])$. This gives not only the bounds (3.8) for $f(x)$ over X, but also the bounds

$$(3.12) \quad \nabla f(x) \in F'(X) = \begin{pmatrix} [-139.4, 307.6] \\ [-128.0, 58.0] \end{pmatrix},$$

for the gradient vector of f, and

$$(3.13) \quad Hf(x) \in F''(X) = \begin{pmatrix} [534.0, 1410.0] & [-480.0, -360.0] \\ [-480.0, -360.0] & [200.0, 200.0] \end{pmatrix},$$

for the Hessian matrix of f over X. These bounds, obtained automatically by the use of IHESSIAN arithmetic, are guaranteed.

In addition to bounds for the values of f and its derivatives on X, the IHESSIAN computation (3.11) provides information about the continuity of f and ∇f on X. For an interval $I = [a, b] \in \mathbf{IR}$, let

$$(3.14) \quad |I| = \|[a, b]\| = \max\{|a|, |b|\}.$$

If $X = (X_1, X_2, \ldots, X_n)$ is an interval vector, then $\|X\|$ will denote the quantity

$$(3.15) \quad \|X\| = \max_i |X_i|,$$

and for an $n \times n$ interval matrix $M = (M_{ij})$, let

$$(3.16) \quad \|M\| = \max_i \sum_{j=1}^{n} |M_{ij}|,$$

analogously to the ∞-norm in $\mathbf{R}^n$ [8]. If the IHESSIAN value of a function $f : \mathbf{R}^n \to \mathbf{R}$ over $X \in \mathbf{IR}^n$ is denoted by $\Phi(X) = (F(X), F'(X), F''(X))$, then the existence of $F'(X)$ implies that f is Lipschitz continuous on X, and $L = \|F'(X)\|$ is a Lipschitz constant for f on X. Similarly, the existence of $F''(X)$ implies that ∇f is Lipschitz continuous on X, and $\|F''(X)\|$ is a Lipschitz constant for ∇f on X. Thus, for the function (2.14), it follows from (3.11) that

$$(3.17) \quad |f(x) - f(y)| \leq 307.6 \cdot \|x - y\|_\infty,$$

and

(3.18)
$$\|\nabla f(x) - \nabla f(y)\|_\infty \le 1890.0 \cdot \|x - y\|_\infty$$

for $x, y \in X = ([0.9, 1.2], [0.8, 1.1])$ [14]. The Lipschitz continuity of ∇f will enter into the discussion later.

In actual practice, the arithmetic operators and standard library functions for type IHESSIAN can be programmed once and for all, and stored in a small subroutine library, as has been done in Pascal-SC [13].

4. Tests for Existence or Nonexistence of Critical Points.

The key issue in the algorithm described in this paper is to determine if a region in $\mathbf{R}^n$ defined by an interval vector $X \in \mathbf{IR}^n$ contains a critical point of the function $f : \mathbf{R}^n \to \mathbf{R}$ or not. First of all, if

(4.1)
$$0 \notin F'(X),$$

then it is impossible that $\nabla f(x) = 0$ for $x \in X$, and X can be *rejected*, since it does not contain a critical point of f [7]. On the other hand, $0 \in F'(X)$ does *not* necessarily mean that X contains a critical point of f, because $F'(X)$ overestimates $\nabla f(X)$ in general. The intersection of all interval vectors containing $f(X)$ is called the *interval hull* of $f(X)$. In several dimensions, the interval hull of $f(X)$ can contain points outside of $f(X)$ in general [8].

In addition to the rejection criterion (4.1), a test which is capable of establishing the existence of a critical point x^* in X is necessary. For this purpose, the test given by Moore [m3] will be used. This test is based on the application of the *Krawczyk transformation* K [5] to X:

(4.2)
$$K(X) = x - \left(Hf(x)\right)^{-1}\nabla f(x) + \left\{I - \left(Hf(x)\right)^{-1}F''(X)\right\}(X - x),$$

where I denotes here the $n \times n$ identity matrix, and $x = m(X)$ the midpoint vector of the interval region X. The real-valued vectors and matrices in (4.2) are of course interpreted as degenerate interval-valued objects.

In actual practice, $K(X)$ is computed by solving the linear system

(4.3)
$$\left(Hf(x)\right)\Xi = -\nabla f(x) + \left\{Hf(x) - F''(X)\right\}(X - x)$$

for Ξ, from which

(4.4)
$$K(X) = x + \Xi.$$

Once $K(X)$ has been computed, one of the following alternatives holds. If

(4.5)
$$K(X) \subseteq X,$$

then there exists a critical point $x^* \in K(X)$ of f; if

(4.6)
$$K(X) \cap X = \emptyset,$$

888

is empty, then X does *not* contain a critical point of f (another rejection criterion); otherwise, the test is inconclusive [7].

With regard to (4.6), the intersection of interval vectors X, Y is said to be empty if for some i, X_i and Y_i are disjoint intervals. It also follows from (4.2) that if $x^* \in X$ is a critical point of f, then $x^* \in K(X)$. Thus, in the inconclusive case, the region

$$Z = X \cap K(X) \subseteq X \tag{4.7}$$

will also contain any critical points of f which lie in X. This suggests decomposing Z (which may be equal to X) into several subregions, and applying the above tests to each. The resulting algorithms, described in more detail below, are essentially modifications of the one given by Moore and Jones [9] for locating solutions of systems of nonlinear equations in several variables. These algorithms differ from the Moore-Jones method in that they make use of the fact that an optimization problem underlies the system of equations being solved, which provides information additional to that inherent in an arbitrary system of nonlinear equations. Furthermore, the algorithms given here differ by bisecting intersected intervals in the inconclusive case, which results in a certain amount of increase in speed.

The use of subregions has the advantage that the tests (4.1) and (4.5)–(4.6) become more sensitive as the size of the region decreases, that is, as $\|w(X)\|_\infty \to 0$. In fact, if x^* is a *regular* critical point of f, that is, if $(Hf(x^*))^{-1}$ exists, then (4.5) will hold for sufficiently small X such that $x^* \in X$ if ∇f is Lipschitz continuous [10]. The disadvantages are the extra bookkeeping and storage required for pending subregions. However, these are not overwhelming on modern computers.

5. Implementation of the Krawczyk Transformation. The system of equations (4.3), as stated, has the real coefficient matrix $Hf(x)$, interpreted as a degenerate interval matrix, and an interval right side. In actual computation, instead of solving (4.3), one obtains an inclusion $\tilde{\Xi}$ of the solution Y of the system

$$F''(x)Y = -F'(x) + \{F''(x) - F''(X)\}(X - x), \tag{5.1}$$

which has an interval coefficient matrix and an interval right side. The solution of such a system

$$AY = B \tag{5.2}$$

is defined to be

$$Y = \{y \mid ay = b, \ a \in A, \ b \in B\}, \tag{5.3}$$

where a is a real matrix, and $b, y \in \mathbf{R}^n$, provided all the indicated real systems are solvable. In this case, it follows that $\Xi \subseteq Y \subseteq \tilde{\Xi}$, where Ξ is the solution of (4.3), and thus

$$K(X) = x + \Xi \subseteq x + \tilde{\Xi} = \tilde{K}(X). \tag{5.4}$$

Furthermore,

$$\tilde{K}(X) \subseteq X \tag{5.5}$$

889

or

(5.6)
$$\tilde{K}(X) \cap X = \emptyset$$

thus imply (4.5) or (4.6), respectively. In this way, existence or nonexistence of a critical point of f in X can be established conclusively by a computation done in floating-point interval arithmetic by a computer, providing it is possible to obtain an inclusion for the solution of the system (5.1). In actual practice, the widths of the components of $F''(x)$ will be small, and good inclusions of Y can be obtained by a process of floating-point approximation followed by interval iterative refinement [15], which will fail only if $F''(x)$ contains a singular or very badly conditioned matrix. Interval linear system solvers of this type are available in Pascal-SC and the ACRITH package.

6. The Basic Algorithm. The algorithm described in this section will find one or all the critical points of a function f in a given initial region X, or show that X contains no critical points, provided no exceptions arise. Exceptions will be discussed in a later section. The computer program implementing this algorithm handles exceptions in such a way that the computation always terminates in a finite number of steps. Validated upper and lower bounds are given for all critical points and values found. The algorithm will be presented first for the case of a single processor, in the way it has actually been implemented. Adaptation to a multiprocessor environment will be discussed at the end of the section.

The basic steps of the algorithm are:

$1°$. Compute $\Phi(X) = (F(X), F'(X), F''(X))$ in IHESSIAN arithmetic. If $0 \notin F'(X)$, then X is rejected.

$2°$. Compute $\Phi(x) = (F(x), F'(x), F''(x))$ in IHESSIAN arithmetic. Compute an inclusion $\tilde{K}(X)$ of the Krawczyk transformation of X by (5.1).

$3°$. If $\tilde{K}(X) \subseteq X$, then the interval iteration

(6.1)
$$X^0 = X, \qquad X^{n+1} = \tilde{K}(X') \cap X^n$$

is performed until it converges to

(6.2)
$$X^* = X^N \subseteq X^{N+1},$$

in a finite number of steps [12]. The values

(6.3)
$$X^*, \qquad \Phi(X^*) = (F(X^*), F'(X^*), F''(X^*)),$$

are output. The existence of a critical point x^* of f in X is guaranteed, and furthermore the bounds

(6.4)
$$x^* \in X^*, \qquad (f(x^*), \nabla f(x^*), Hf(x^*)) \in (F(X^*), F'(X^*), F''(X^*)),$$

for x^* and the values of the function f, its gradient vector ∇f, and its Hessian matrix Hf at x^*. These bounds are usually as good as can be obtained by floating-point computation, and $F''(X^*)$ can be used to determine the nature of the critical point x^*, if necessary.

890

4°. If

$$\tilde{K}(X) \cap X = \emptyset,$$ (6.5)

then X is rejected.

5°. In the indeterminate case, the region

$$Z = (Z_1, Z_2, \ldots, Z_n) = \tilde{K}(X) \cap X$$ (6.6)

is *bisected* in the direction of its widest component. An index j is determined such that

$$w(Z_j) \geq w(Z_i), \quad i = 1, 2, \ldots, n,$$ (6.7)

and one takes

$$Z^l = (Z_1, \ldots, Z_{j-1}, [\inf Z_j, m(Z_j)], Z_{j+1}, \ldots, Z_n),$$

(6.8)

$$Z^r = (Z_1, \ldots, Z_{j-1}, [m(Z_j), \sup(Z_j)], Z_{j+1}, \ldots, Z_n).$$

6°. (Single processor) Compute $\Phi(Z^l)$, $\Phi(Z^r)$. The test (4.1) is applied to $F'(Z^l)$ and $F'(Z^r)$. There are four cases:

(i) If $0 \notin F'(Z^l)$ and $0 \notin F'(Z^r)$, then X is rejected;

(ii) If $0 \in F'(Z^l)$ but $0 \notin F'(Z^r)$, then return to step 2° with $X = Z^l$;

(iii) If $0 \notin F'(Z^l)$ but $0 \in F'(Z^r)$, then return to step 2° with $X = Z^r$;

(iv) If $0 \in F'(Z^l)$ and $0 \in F'(Z^r)$, then one of the interval vectors is to be placed on a push-down (last in, first out) stack in storage, while the other replaces X for continued processing at step 2°. In this algorithm, the choice is made by the following heuristic: If $w(F(Z^l)) \leq w(F(Z^r))$, then one takes $X = Z^l$ and stacks Z^r for processing later; otherwise, one takes $X = Z^r$ and stacks Z^l.

The region selected for X is considered to be "more promising" than the one stacked because the variation of a function in the neighborhood of a critical point is asymptotically less than it is elsewhere. The goal is to find critical points as quickly as possible, particularly if only one is desired.

6°. (Multiprocessors) Z^l is sent to another processor following the bisection (6.8). Return to step 1° with the current processor taking $X = Z^r$, while the other takes $X = Z^l$. If no processors happen to be free, then Z^l circulates or is put on a common stack to await the first available processor. If a number of processors are free, it could be expeditious to decompose X into more than two subregions, and then send one to each processor. The choices here will depend to a great extent on the multiprocessor configuration actually used.

In the case of a single processor, the choice of which interval to stack in step 6°(*iv*) will be modified in the case global critical maxima or minima are sought. If only one critical point is sought, the algorithm is terminated at the end of step 3°. Otherwise, the

891

algorithm can continue until all critical points of f in X are found, and no regions remain on the stack to process. Complete processing of X without finding critical points proves that it contained none.

Because the processes of intersection and bisection can result in subregions of a wide range of sizes, a simple count of the number processed at any given time does not give a good indication of the progress being made by the algorithm. For this reason, it has been found convenient to compute the initial volume

$$(6.9) \qquad\qquad V_0 = \prod_{i=1}^{n} w(X_i)$$

of the region $X = (X_1, X_2, \ldots, X_n)$ to be searched. The unexplored part of the initial region has volume $V_u(t)$ at time t, where $V(0) = V_0$. $V_u(t)$ can be computed simply as the sum of the volumes of the intervals being processed and those on the stack awaiting processing at time t. $V_u(t)$ is a monotone decreasing function of t, and the algorithm terminates when the stack is empty and $V_u(t) = 0$, if an exhaustive search is desired.

7. Exceptions. Several exceptions can arise in the execution of the algorithm in §6 which could terminate the computation prematurely, or cause it to run indefinitely. These and the way they are handled will be discussed now, because they may also occur in the search for global critical extrema.

$1°$. If $F''(x)$ contains a singular or badly conditioned matrix, then the attempt to perform the Krawczyk transformation by solving (5.1) will fail. One solution is to replace $F''(x)$ by some nonsingular matrix, for example, $m(F''(X))$ could work [9]. The implementation used for the examples given below simply outputs X to a file for later examination, with an appropriate message, and then selects the next region to be processed from the stack.

$2°$. The intersection-bisection process can lead to regions which do not differ from the previous ones, because of outward rounding, or which are so small that total time to explore the entire region is prohibitive. This can happen, in particular, if a critical point lies exactly on a bisection coordinate. For this reason, the user is provided with a parameter ϵ such that if the volume V of the region to be processed satisfies

$$(7.1) \qquad\qquad V \leq \epsilon \cdot V_0,$$

then the region will be output to a file for later examination, with an appropriate message.

The choice $\epsilon = 0$ is permitted; this allows the processing of smaller and smaller regions until their volume (6.9) underflows to 0 or some coordinate becomes degenerate.

$3°$. If the storage space allotted to the stack is full, then additional regions will be output to a file.

$4°$. Numerical exceptions, such as division by zero and overflow, are allowed to terminate the present program. However, they could be used as signals to output the offending region to a file with an appropriate message.

Successful termination of the computer program will be accompanied with a list of the number of regions processed (given by the number of times the Krawczyk transformation was performed), the number rejected, and the number of critical points found, and the number output to the exception file. Given all its critical points, the global critical maximum and minimum of the function can be found simply by sorting the function values.

8. Global Critical Extrema. The algorithm of §6 can be speeded up if only the global critical maximum or mininum of f on X is desired. The modification of the algorithm to find the global critical minimum will be described; finding the global critical maximum follows exactly the same pattern.

Suppose first that the global critical minimum of f on X is actually its global minimum on X, that is, $f(x) \geq m = f(x^*)$ for $x \in X$, where x^* is the global critical minimum point. The algorithm will compute a decreasing sequence of upper bounds $m(t)$ for m, and reject subregions Z such that $\inf F(Z) > m(t)$. The modifications of the corresponding steps for the single-processor algorithm are:

$1°$. Set $m(0) = $ MAXREAL, the largest floating-point number (for example, in Pascal-SC, MAXREAL$=9.99999999999 \times 10^{99}$).

$2°$. If $\sup F(x) < m(t-1)$, then set $m(t) = \sup F(x)$, otherwise, $m(t) = m(t-1)$.

$6°$. Reject Z^l if $0 \notin F'(Z^l)$ or $\inf F(Z^l) > m(t)$; reject Z^r if $0 \notin F'(Z^r)$ or $\inf F(Z^r) > m(t)$. If neither Z^l nor Z^r can be rejected, then Z^l is considered to be more promising if $\inf F(Z^l) < \inf F(Z^r)$, and Z^r is stacked, or conversely. In case $\inf F(Z^l) = \inf F(Z^r)$, then Z^r is stacked if $\sup F(Z^r) \geq \sup F(Z^l)$, otherwise, Z^l is stacked.

Considerable savings in computer time have been observed due to the introduction of the additional rejection conditions in $6°$. In the case of multiprocessors, an efficient way to share the current value of $m(t)$ is necessary, and the rejection of regions in which function values are too large would be carried out in step $1°$.

In case the function f can attain smaller values than m on the boundary ∂X of X at points which are not critical, an alteration has to be made in the above procedure. The value of $m(t)$ is updated only when regions X^* containing critical points x^* are computed by interval iteration. If $\sup F(x^*) < m(t-1)$, then we set $m(t) = \sup F(x^*)$, otherwise $m(t) = m(t-1)$. The rejection criterion $\inf F(Z) > m(t)$ remains unaltered. The algorithm will generally be slower than the one given above in this case, but usually still faster than an exhaustive search for all critical points.

In the same way, a function $M(t)$ giving the lower bound for the global maximum M of f is constructed by setting $M(0) = -$MAXREAL, and updating by $M(t)$ to be the maximum of $M(t-1)$ and $\inf F(x)\}$, assuming that the global maximum is critical. Intervals are rejected if $\sup F(X) < M(t)$. Otherwise, $M(t)$ is updated only at critical points, as above. The modification of the choice algorithm for bisected intervals is done by reversing inequality signs and interchanging infs with sups in the above.

9. Use of the Algorithm for Validation. In addition to its use for global searching, the algorithm given in §6 can be used to validate solutions to optimization problems given by other algorithms. For example, suppose that $\hat{x}$ is an approximate critical point of f found by Newton's method or some other numerical technique. Then, the initial region X

can be taken to be, say

(9.1)
$$X = \hat{x} + \frac{\delta}{2} \cdot e \cdot [-1, 1],$$

where $\delta > 0$ and $e = (1, 1, \ldots, 1) \in \mathbf{R}^n$ denotes the vector with all components equal to one. If (5.5) holds for this value of X, then all components of $\hat{x}$ are validated to be accurate to a tolerance of δ. Furthermore, the interval iteration (6.1) will give approximations of possibly increased accuracy for the critical point, as well as bounds the for function, gradient vector, and Hessian matrix values. It should be noted that the interval calculations furnish upper and lower bounds for maximum and minimum values of the function, while some other methods give only one-sided bounds.

10. Numerical Examples. The functions selected for numerical computation were the n-dimensional Rosenbrock function [1],

(10.1)
$$f_n(x) = \sum_{i=1}^{n-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \right],$$

and the "three-humped camel" function [3],

(10.2)
$$g(x) = 2x_1^2 - 1.05x_1^4 + \frac{1}{6}x_1^6 - x_1 x_2 + x_2^2.$$

The program used was written in Pascal-SC for a microcomputer with a Z80 processor and the CP/M operating system. This was done to take advantage of support for interval arithmetic, an already written library of operators and functions for type IHESSIAN, and the utility procedure LGLI for solving linear systems with interval coefficient matrices and right sides. On the other hand, the small amount of storage available in this machine (64 kilobytes) limited the values of n for the Rosenbrock function (10.1) to $n = 2, 3$. The actual machine used was also rather slow, with a 1MHz system clock, giving typical times for floating-point interval addition and subtraction of 13.5 milliseconds, multiplication, 57.5 milliseconds, and division, 77.5 milliseconds. Nevertheless, the results given below were obtained in a reasonable amount of time.

The most time-consuming part of the computation is the performance of the Krawczyk transformation $K(X)$ (actually, $\tilde{K}(X)$), using the Pascal-SC utility program LGLI to solve the system (5.1) with interval coefficient matrix and right side. A count is made of the number of times this transformation is performed, the number of critical points found, the number of regions rejected, and the number of regions (if any) in which exceptions are encountered. The sum of the number of regions rejected (which cannot contain critical points), the number of regions in which critical points are found, and the number of exceptional regions gives the total number of subregions examined. The Krawczyk transformation may be applied to a given region several times before it is accepted as containing a critical point, or rejected.

The Rosenbrock function (10.1) has the global minimum $f_n(x^*) = 0$ at the critical point $x^* = e = (1, 1, \ldots, 1)$. It is easy to find x^* by Newton's method, but methods which

try to reduce $f_n(x)$ at each step find this function rather difficult, particularly in higher dimensions. Four cases were considered:

1. $n = 2$, $X_0 = ([0.9, 1.2], [0.8, 1.1])$;
2. $n = 2$, $X_0 = ([-3.7, 1.4], [-1.6, 3.5])$;
3. $n = 3$, $X_0 = ([0.9, 1.2], [0.8, 1.1], [0.9, 1.2])$;
4. $n = 3$, $X_0 = ([-3.7, 1.4], [-1.6, 3.5], [-3.7, 1.4])$.

The value $\epsilon = 0$ was taken in each case, and no exceptions occured in any of the examples given here. Searching for all critical points gave the following results:

	Case 1	Case 2	Case 3	Case 4
Transformations	242	957	553	1976
Rejected	167	685	432	1567
Critical Points	1	1	1	1
Transformations to Locate	128	187	328	1672

The algorithm was very busy in the neighborhood of the critical points $x^* = (1, 1)$ and $x^* = (1, 1, 1)$. The region in which (5.5) holds turned out to be rather small, and nearby regions not containing x^* had to be made very small before they could be rejected with certainty. The increase in area of X_0 by a factor of 289 between cases 1 and 2 increased the number of Krawczyk transformations required by a factor of less than four, while the increase in volume of X_0 between cases 3 and 4 by a factor of 4913 resulted in an even smaller increase in the number of transformations, less than 3.6. Going from two to three dimensions increased the number of transformations required to search the entire initial region by a factor of about two in each case.

The modification of the program to search for a global critical minimum gave the following results:

	Case 1	Case 2	Case 3	Case 4
Transformations	92	140	346	219
Rejected	75	110	169	172
Critical Points	1	1	1	1
Transformations to Locate	83	103	330	201

These results show a considerable improvement over the search for all critical points. Once the global minimum value has been found, remaining regions are generally rejected quickly. The algorithm was very effective for the largest problem considered, Case 4 above.

The required increase in minimum function values in the search for a global critical maximum forced the algorithm toward the boundary of X_0, where regions were quickly rejected. The corresponding results were:

	Case 1	Case 2	Case 3	Case 4
Transformations	9	16	27	37
Rejected	9	16	28	38
Critical Points	0	0	0	0

In the last two cases, the final regions were rejected without having to perform a **Krawczyk transformation**.

In two dimensions, the interval iteration to the critical point x^* **converged to**

(10.3) $\qquad X^* = ([0.999999999999, 1.00000000001], [0.999999999999, 1.00000000001])$

with

(10.4) $\qquad\qquad\qquad F_2(X^*) = [0.00, 9.62 \times 10^{-20}],$

(10.5) $\qquad\qquad\qquad F'_2(X^*) = \begin{pmatrix} [-4.802 \times 10^{-9}, 1.242 \times 10^{-8}] \\ [-6.200 \times 10^{-9}, 2.400 \times 10^{-9}] \end{pmatrix},$

and

(10.6)

$$F''_2(X^*) = \begin{pmatrix} [801.999999987, 802.000000030] & [-400.000000004, -399.999999998] \\ [-400.000000004, -399.999999998] & 200 \end{pmatrix}.$$

The midpoint of X^* was calculated to be $\bar{x} = (1,1)$, with $F_2(\bar{x}) = 0$, $F'_2(\bar{x}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and

the Hessian matrix $F''_2(\bar{x}) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix}$, which are validated to be the **exact values** of x^*, $f_2(x^*)$, $\nabla f_2(x^*)$, and $H f_2(x^*)$ by the above.

The results in three dimensions were completely similar, with the **midpoint $\bar{x} = (1,1,1)$** of X^* giving the exact values $f_3(\bar{x}) = 0$, $\nabla f_3(\bar{x}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and

(10.7) $\qquad\qquad H f_3(\bar{x}) = \begin{pmatrix} 802 & -400 & 0 \\ -400 & 1002 & -400 \\ 0 & -400 & 200 \end{pmatrix}.$

The three-humped camel function $g(x)$ given by (10.2) has **five critical points:**

(10.8) $\qquad\qquad\qquad x^* = (0,0),$

which is its global minimum point,

(10.9) $\qquad\qquad\qquad \pm y^* = \pm \left(\sqrt{2.1 - \sqrt{0.865}}, \frac{1}{2}\sqrt{2.1 - \sqrt{0.865}} \right),$

896

which are saddle points, and

$$(10.10) \qquad \pm z^* = \pm\left(\sqrt{2.1 + \sqrt{0.865}}, \frac{1}{2}\sqrt{2.1 + \sqrt{0.865}}\right),$$

which are relative minimum points. One has

$$(10.11) \qquad 0 = g(x^*) < g(-z^*) = g(z^*) < g(-y^*) = g(y^*).$$

The search for all critical points was conducted in the inital region

$$(10.12) \qquad X_0 = ([-2.0, 1.8], [-0.9, 1.0]),$$

with $\epsilon = 0$. The total number of Krawczyk transformations required was 82, 48 intervals were rejected, all 5 critical points were found, and there were no exceptions. This function is less of a computational challenge than the Rosenbrock function $f_2(x)$; however, the critical points $\pm y^*$ and $\pm z^*$ tend to shield x^* from straightforward iterative procedures, such as Newton's method. Letting $T(\cdot)$ denote the number of Krawczyk transformations required to locate a given critical point, the results were:

$$(10.13) \qquad \begin{aligned} T(x^*) &= 6, \\ T(y^*) &= 21, \\ T(z^*) &= 45, \\ T(-y^*) &= 69, \\ T(-z^*) &= 79. \end{aligned}$$

The search for the global minimum of $g(x)$ in X_0 required 60 Krawczyk transformations, 46 intervals were rejected, 2 critical points were located in order of decreasing function value, and there were no exceptions. The critical points found were first $-z^*$ and then the global minimum point x^*, with

$$(10.14) \qquad \begin{aligned} T(-z^*) &= 10, \\ T(x^*) &= 57. \end{aligned}$$

Obviously, the search took an entirely different path than the exhaustive search (10.13) for all critical points of $g(x)$ in X_0.

The search for the global critical maximum of $g(x)$ was somewhat slower, due to the fact that $g(x)$ attains it maximum at a noncritical point on the boundary ∂X_0 of X_0. Consequently, the function $M(t)$ which gives a lower bound for the critical maximum was updated only at critical points. This computation required 76 Krawczyk transformations, 50 intervals were rejected, and three critical points (x^*, y^*, and $-y^*$) were found in order of nondecreasing function values. The number of transformations required were:

$$(10.15) \qquad \begin{aligned} T(x^*) &= 6, \\ T(y^*) &= 21, \\ T(-y^*) &= 64. \end{aligned}$$

The critical points $\pm z^\star$ were also located, but rejected, because the values of $g(x)$ at these relative minimum points is smaller than at the saddle points $\pm y^\star$. Investigation of the nature of critical points is done by finding bounds for the eigenvalues of all symmetric matrices A such that $A \in G''(X^\star)$, using the Pascal-SC procedure EIGEN. Denoting the eigenvalues of a 2×2 symmetric matrix A by $\lambda_1(A)$ and $\lambda_2(A)$, then intervals $\Lambda_1(X^\star)$ and $\Lambda_2(X^\star)$ are computed by EIGEN such that

$$(10.16) \qquad \{\lambda_i(A) \mid A \in G''(X^\star)\} \subseteq \Lambda_i(X^\star), \qquad i = 1, 2.$$

The character of the critical point $x^\star \in X^\star$ can be decided on the basis of these bounds. The results of the interval iteration to critical points were:

$$X^\star = ([-2.0 \times 10^{-99}, 2.0 \times 10^{-99}], [-2.0 \times 10^{-99}, 2.0 \times 10^{-99}]),$$
$$G(X^\star) = [-2.05 \times 10{-}99, {}^{-}.00 \times 10^{-99}],$$
$$(10.17) \qquad G'(X^\star) = \begin{pmatrix} [-1.52 \times 10^{-98}, 1.52 \times 10^{-98}] \\ [-6.00 \times 10^{-99}, 6.00 \times 10^{-99}] \end{pmatrix},$$
$$G''(X^\star) = \begin{pmatrix} [3.99999999999, 4.00000000000] & -1 \\ -1 & 2 \end{pmatrix}.$$

The eigenvalues of $Hg(x^\star)$ are contained in the intervals

$$(10.18) \qquad \begin{aligned} \Lambda_1(X^\star) &= [4.142135623, 4.142135625], \\ \Lambda_2(X^\star) &= [1.58578643759, 1.58578643766], \end{aligned}$$

which proves conclusively that the critical point $x^\star \in X^\star$ is a minimum point, because both eigenvalues of $Hg(x^\star) \in G''(X^\star)$ must be positive [1]. The midpoint of $X^\star$ is $\bar{x} = x^\star = (0, 0)$, with $G(\bar{x}) = g(x^\star) = 0$, $G'(\bar{x}) = \nabla g(x^\star) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $G''(\bar{x}) = Hg(x^\star) = \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix}$.

Next,

$$Y^\star = ([1.07054229181, 1.07054229185], [0.535271145904, 0.535271145921]),$$
$$G(Y^\star) = [0.877361557501, 0.877361558041],$$
$$(10.19) \qquad G'(Y^\star) = \begin{pmatrix} [-5.81 \times 10{-}10, 5.26 \times 10^{-10}] \\ [-5.00 \times 10^{-11}, 4.00 \times 10^{-11}] \end{pmatrix},$$
$$G''(Y^\star) = \begin{pmatrix} [-3.87308929305, -3.87308929080] & -1 \\ -1 & 2 \end{pmatrix}.$$

The eigenvalues of $Hg(y^\star)$ are contained in the intervals

$$(10.20) \qquad \begin{aligned} \Lambda_1(Y^\star) &= [-4.03868818, -4.03868818], \\ \Lambda_2(Y^\star) &= [2.165598876, 2.165598884], \end{aligned}$$

898

so that the critical point $y^* \in Y^*$ is indubitably a saddle point.

The next values obtained were

$$Z^* = \left([1.74755234581, 1.74755234586]\right),$$

$$G(Z^*) = [0.298638440884, 0.298638443572],$$

(10.21) $$G'(Z^*) = \begin{pmatrix} [-2.18 \times 10^{-10}, 3.82 \times 10^{-10}] \\ [-1.00 \times 10^{-11}, 0.00] \end{pmatrix},$$

$$G''(Z^*) = \begin{pmatrix} [1.21530892921, 1.21530892930] & -1 \\ -1 & -2 \end{pmatrix}.$$

The eigenvalues of $Hg(z^*)$ are contained in the intervals

(10.22) $$\begin{aligned} \Lambda_1 &= [5.33440787, 5.33440793], \\ \Lambda_2 &= 2.681868136, 2.681868143], \end{aligned}$$

and thus the critical point z^* is a relative minimum point.

The computed intervals

(10.23)
$$-Y^* = \left([-1.070544229185, -1.07054229181], [-0.535271145923, -0.535271145906]\right)$$

and

(10.24)
$$-Z^* = \left([-1.74755234585, -1.74755234580], [-0.873776172925, -0.873776172902]\right)$$

contain the critical points $-y^*$ and $-z^*$, respectively. The function, gradient, and Hessian values on these intervals do not differ significantly from the corresponding ones for Y^* and Z^*. In particular, the eigenvalues of $Hg(-y^*)$ lie in the intervals (10.20), while the eigenvalues of $Hg(-z^*)$ belong to the intervals (10.22). Thus, $-y^*$ is guaranteed to be a saddle point, and $-z^*$ a relative minimum point of g.

References

1. P. E. Gill, W. Murray and M. H. Wright. Practical Optimization. Academic Press, New York, 1981.

2. E. R. Hansen. Global optimization using interval analysis—the multi-dimensional case. Numer. Math. **34** (1980), 247–270.

3. E. R. Hansen. Global optimization with data perturbations. Comput. & Ops. Res. **11**, no. 2 (1984), 97–104.

4. E. R. Hansen and S. Sengupta. Global constrained optimization using interval analysis. *Interval Mathematics 1980*, ed. by K. Nickel, pp. 25–47. Academic Press, New York, 1980.

5. R. Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing **4** (1969), 187–201.

6. G. P. McCormick. Nonlinear Programming. Wiley, New York, 1983.

7. R. E. Moore. A test for existence of solutions to nonlinear systems. SIAM J. Numer. Anal. **14** (1977), 611–615.

8. R. E. Moore. Methods and Applications of Interval Analysis. SIAM Studies in Applied Mathematics 2, Soc. Ind. Appl. Math., Philadelphia, 1979.

9. R. E. Moore and S. T. Jones. Safe starting regions for iterative methods. SIAM J. Numer. Anal. **14** (1977), 1051–1065.

10. L. B. Rall. A comparison of the existence theorems of Kantorovich and Moore. SIAM J. Numer. Anal. **17**, no. 1 (1980), 148–161.

11. L. B. Rall. Automatic Differentiation: Techniques and Applications. Lecture Notes in Computer Science No. 120. Springer-Verlag. Berlin-Heidelberg-New York, 1981.

12. L. B. Rall. A theory of interval iteration. Proc. Amer. Math. Soc. **86**, no. 4 (1982), 625–631.

13. L. B. Rall. Differentiation and generation of Taylor coefficients in Pascal-SC. *A New Approach to Scientific Computation*, ed. by U. W. Kulisch and W. L. Miranker, pp. 291–309. Academic Press, New York, 1983.

14. L. B. Rall. Mean value and Taylor forms in interval analysis. SIAM J. Math. Anal **14**, no. 2 (1983), 223–238.

15. S. M. Rump. Solving algebraic problems with high accuracy. *A New Approach to Scientific Computation*, ed. by U. W. Kulisch and W. L. Miranker, pp. 51–120. Academic Press, New York, 1983.

COMPUTING K-TERMINAL RELIABILITY
IN TIME POLYNOMIAL IN THE NUMBER OF (S,K)-QUASICUTS

Michael O. Ball [†]
College of Business and Management
University of Maryland
College Park, MD 20742
and
J. Scott Provan [*]
Curriculum in Operations Research and Systems Analysis
University of North Carolina
Chapel Hill, NC 27514

ABSTRACT: In this paper, we present an algorithm for computing k-terminal reliability in directed or undirected networks. The running time of the algorithm is bounded by a polynomial in the number of (s,K)-quasicuts, where an (s,K)-quasicut is defined to be a minimal (s,K^*)-cut for some K^* contained in K. The algorithm can be implemented by applying our cut based 2-terminal algorithm to a network to which a super-sink node has been added. Thus, a 2-terminal code based on our cut based algorithm could be used to solve the k-terminal problem.

1. INTRODUCTION: We present a new algorithm for computing the k-terminal network reliability measure. The algorithm is a generalization of an algorithm we previously described in (Provan and Ball 1984) for the 2-terminal reliability measure. The 2-terminal algorithm was bounded by a polynomial in the size of the network and the number of minimal (s,t)-cuts. An analogous result for the k-terminal problem would be an algorithm bounded by a polynomial in the size of the network and the number of minimal (s,K)-cuts. We showed in (Provan and Ball 1984) that the existence of such an algorithm is unlikely since it would imply that P=NP. The algorithm given in this paper has a time bound which is polynomial in the size of the network and the number of (s,K)-quasicuts, which are arc sets closely related to minimal (s,K)-cuts. A desirable feature of this algorithm and its development is that the algorithm can be realized by applying the 2-terminal algorithm to a graph in which a super sink node has

---

been added. Thus, a 2-terminal code based on our cut procedure could be used to solve the k-terminal problem.

2.   PRELIMINARY DEFINITIONS: For network reliability problems, we are given an underlying network G=(N,A), either directed or undirected, with node set N=(v₁,v₂,...,vₘ) and arc set A=(e₁,e₂,...,eₙ). An arc is denoted by (v_i, v_j), which is taken to be an ordered or unordered pair depending on whether G is directed or undirected. We assume that each arc $e \in A$ fails independently with known failure probability p_e. Our reliability measures are defined by the existence of __operating paths__, paths of operating arcs, between certain pairs of nodes in G. We define, for nodes s and t of G, the event

$$EP(s,t) = [\text{there exists an operating path from s to t}].$$

The path is taken to be directed when G is directed, and there is always an operating path from s to itself. The problem of computing $Pr[EP(s,t)]$ is called the __(s,t)-connectedness problem__ or the __2-terminal reliability problem__. We extend this definition to a general terminal set as follows. Given a $K \subseteq N$ and an $s \in K$, define the event

$$EP(s,K) = [\text{there exists an operating path from s to every node of K}]$$
$$= \cup_{t \in K} EP(s,t) \quad .$$

The problem of computing $Pr[EP(s,K)]$ is called the __(s,K)-connectedness problem__ or the __k-terminal reliability problem__ where k=|K|. When K=N, it is also called the __all-terminal reliability problem__.

Each of the measures given above has the property that the underlying system is coherent with respect to that measure; that is, if the system operates when a set S of components (arcs) operates, then it operates when any superset of S operates. Coherent systems can be described completely by listing either (1) the collection of minimal sets of components whose operation allows system operation, or (2) the collection of minimal sets of components whose failure causes system failure. We call these sets, respectively, the __pathsets__ and __cutsets__ of the system. In the case of the (s,t)-connectedness problem, the pathsets are simply the (s,t)-paths, and the cutsets are the (s,t)-cuts, i.e. __minimal__ sets of arcs whose removal disconnects s and t. We denote the set of all (s,t)-cuts by Ç(s,t). In the case of the (s,K)-connectedness problem, the pathsets are (s,K)-trees (also called K-trees), i.e. minimal sets of arcs that include paths from s to all members of K. The cutsets are (s,K)-cuts, i.e. minimal sets of arcs that disconnect s from some node in K. We denote by Ç(s,K) the set of all (s,K)-cuts.

Many approaches to reliability analysis problems involve either the explicit or implicit enumeration of all cutsets or

pathsets. Surprisingly, even if one is willing to invest the effort required to enumerate all pathsets or cutsets, the problem of computing system reliability may still be difficult, specifically NP-hard (see Ball 1986 for a discussion of these and other complexity issues). The algorithm we give in (Provan and Ball 1984) has a running time bounded by a polynomial in the network size and the number of (s,t)-cuts. Thus, for the 2-terminal problem, if one is willing to expend the effort to enumerate cutsets then one can efficiently compute system reliability. On the other hand, we show that such an algorithm can exist for the k-terminal problem only if P=NP. These results and related results presented in that paper all assume that the network itself is given as input. In (Ball and Provan 1986), we address a related problem where the input is given as an explicit list of all pathsets or cutsets.

In this paper, we extend our 2-terminal algorithm to the k-terminal problem. The time bound obtained is not a polynomial in the number cutsets, (s,K)-cuts in this case, but rather a polynomial in the number of (s,K)-quasicuts, which are arc sets closely related to (s,K)-cuts. We now give definitions required to present the algorithm and its time bound.

3. THE ALGORITHM: The algorithm itself is a particular application of the 2-terminal algorithm. The 2-terminal algorithm can be described in terms of a recursive update formula. After giving some necessary definitions we state the update formula and then show how it can be applied to the k-terminal problem and derive its time bound.

Given a network $G=(N,A)$ and a subset of nodes S, define the subnetwork G(S) induced by S to be the node set S together with all arcs between nodes of S. For node sets U and V, define $A(U,V)=\{(u,v)\in A:u\in U, v\in V\}$. For any (s,t)-cut C, we identify the two sets:

$SN(C)$ = $\{u\in N:$there exists a path from s to u containing no arcs of C$\}$,

$TN(C)$ = $\{v\in N:$there exists a path from v to t containing no arcs of C$\}$

and note that (i) SN(C) and TN(C) are disjoint (although they do not necessarily comprise all nodes of G), and (ii) $C=A(SN(C),TN(C))$. The set of exit nodes associated with C is defined to be $SE(C)=\{u\in SN(C):$ there exists an arc (u,v) with $v\in TN(C)\}$.

We now define some random events necessary for the statement of the update formula. For any $C\subseteq A$, let E(C) be the event that all arcs in C fail. For any $C\in\mathcal{G}(s,t)$ define

$EC(C)$ = $EP(s,SE(C))\cap E(C)$
= {there is an operating path from s to all nodes in SE(C), but to no node of TN(C)}.

The following two results, which are proved in (Provan and Ball 1984) form the basis of the 2-terminal algorithm.

Theorem 1:

$$Pr[EP(s,t)] = 1 - \sum_{C\in\mathcal{C}(s,t)} Pr[EC(C)] . \quad (1)$$

Theorem 2: For any $C\in\mathcal{C}(s,t)$

$$Pr[EC(C)] =$$

$$\prod_{e\in C} p_e \left(1 - \sum_{\substack{C'\in\mathcal{C}(s,t) \\ \text{with } SN(C')\subset SN(C)}} Pr[EC(C')]/\prod_{e\in C'\cap C} p_e \right). \quad (2)$$

where $\prod_{e\in C'\cap C} p_e$ is defined to be 1 if $C'\cap C=\phi$.

Expression (1) is a formula for computing $Pr[EP(s,t)]$ in which the number of terms equals the number of (s,t)-cuts. Expression (2) is an update formula for recursively computing each of the terms in expression (1). Note that in order to compute $Pr[EC(C)]$ using (2) a large number of "previous" (s,t)-cuts C', must be considered. As a result expressions (1) and (2) lead to an algorithm whose complexity is proportional to the number of (s,t)-cuts squared. The specific complexity given in (Provan and Ball 1984) is $O((n+m)\mu^2)$ where $\mu=|\mathcal{C}(s,t)|$.

Let us now consider applying the 2-terminal algorithm to the k-terminal problem. Given the input network $G=(N,A)$ together with the source node s, terminal set K and probability vector $\{p_e\}$, we modify the network using a "super sink" node t as follows. Construct network $G'=(N',A')$ from G by setting $N'=N\cup\{t\}$ and $A'=A\cup C_0$, where $C_0=\{(x,t):x\in K\}$. For each $e\in C_0$ we set p_e to 0. Figure 1 illustrates this transformation.

Now consider the effect of applying the 2-terminal algorithm to G'. Since the algorithm computes $Pr[EC'(C)]$ for all (s,t)-cuts C, it will, in particular, compute $Pr[EC'(C_0)]$, where $'$ indicates an event or set associated with G'. But we have

$$Pr[EC'(C_0)] = Pr[EP(s,SE'(C_0))]*Pr[E'(C_0)]$$
$$= Pr[EP(s,K)]*1.$$

The last term is, of course, the k-terminal reliability measure computed over G. Although Figure 1 illustrates the transformation for directed networks, it also applies to undirected networks.

The number of terms in expression (1) will be the number of (s,t)-cuts in G'. In order to interpret this quantity relative to other k-terminal algorithms we need to characterize this quantity in terms of G and the k-terminal problem. Given G, s and K, we define an <u>(s,K)-quasicut</u> to be a set of arcs that is an (s,K^*)-cut for some $K^*\subseteq K$. We denote by $\mathcal{C}_q(s,K)$ the set of all (s,K)-quasicuts. Note that $\mathcal{C}_q(s,K)\supseteq\mathcal{C}(s,K)$. Of course, all (s,K)-quasicuts disconnect s from some member of K, however, not
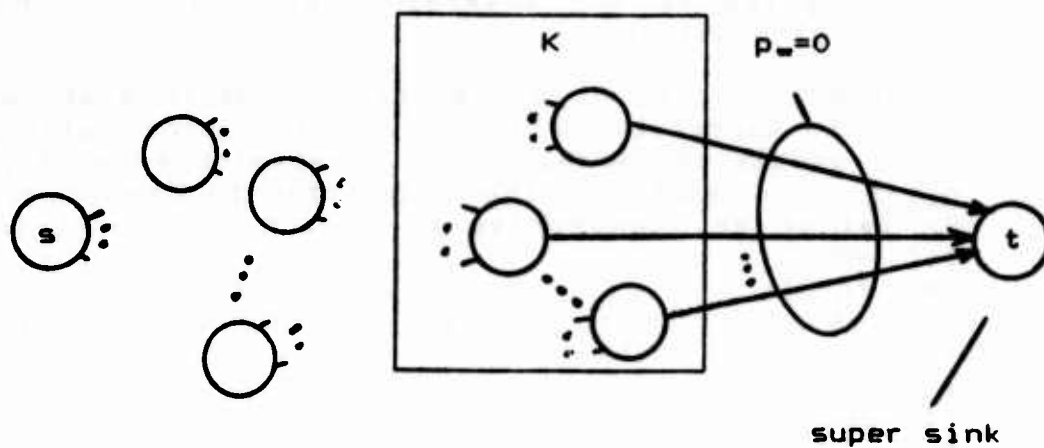
904

FIGURE 1:   Transformation for solving k-terminal problem using
cut based 2-terminal algorithm.

all quasicuts are minimal with respect to this property. For any
(s,K)-quasicut C, we define

    $J(C) = K \cap SN(C)$
    $K(C) = K - SN(C)$ $(\neq \phi)$.

We now show that the work performed by the 2-terminal algorithm
when applied to the k-terminal problem is proportional to the
number of quasicuts.

Theorem 3: Given G, G', K and s as defined above, the number of
terms in expression (1) when the 2-terminal algorithm is applied
to G' equals the number of (s,K)-quasicuts in G' plus one.
<u>proof</u>: We again use the ' notation to denote quantities in G'.
We define an isomorphism $\Phi: \mathcal{C}'(s,t)-\{C_0\} \rightarrow \mathcal{C}_q(s,K)$ by setting
$\Phi(C)=C \cap A$.
(Φ into):   Let $C \neq C_0$ be an (s,t)-cut in G'.   Then
$C=A'(SN'(C),TN'(C))$ and $TN'(C) \supset \{t\}$.   Then it must be that
$TN'(C) \cap K \neq \phi$, so that $C \cap A$ is an $(s,TN'(C) \cap K)$-cut, and hence an
(s,K)-quasicut.
(Φ one to one):   Let $C_1, C_2 \in \mathcal{C}'(s,t)-\{C_0\}$, and suppose
$C_1 \cap A = C_2 \cap A = C$. Then there are paths comprised of arcs of A-C which
go from s to every node of J(C), and no such paths going to any
node of K(C). Thus, the set $D=\{(x,t):x \in J(C)\}$ must be contained
in both C_1 and C_2, and no arc of C_0-D can be contained in either
C_1 or C_2. It follows that $C_1-A=C_2-A=D$, so that $C_1=C_2$.
(Φ onto):   Let $C \in \mathcal{C}_q(s,K)$ and set $C'=C \cup \{(x,t):x \in J(C)\}$.   Then
clearly, C' is an (s,t)-cut in G', and since $K(C) \neq \phi$, $C' \neq C_0$.
                                                              ∎

It immediately follows that,

Corollary 4: Given appropriate G, s, K and $\{p_e\}$, the k-terminal
network reliability problem can be solved in $O((n+m)\mu_q^2)$ time
where $\mu_q=|\mathcal{C}_q(s,K)|$, by applying the cut based 2-terminal
algorithm to G' as described above.
        We should note that the construction illustrated in Figure 1
does not provide a general way of transforming a k-terminal
problem into a 2-terminal problem. It is useful in this case
only because of the particular intermediate calculations made by
our cut based algorithm.
        This approach to the k-terminal problem provides a common
generalization of our cut based algorithm for the 2-terminal
problem and Buzacott's algorithm for the all-terminal problem
(see Buzacott 1980, 1983). That is, when the approach described
in this paper is applied to the 2-terminal problem its steps are
essentially equivalent to the steps of our cut based 2-terminal
algorithm and when it is applied to the all-terminal problem <u>on a
complete network</u> its steps are essentially equivalent to the
steps of Buzacott's algorithm. In this case, the running times
of both algorithms are polynomial in the number of <u>cutsets</u> since
for complete networks (s,K)-quasicuts are (s,K)-cuts. We should
note that for non-complete networks our algorithm will require

substantially less computational effort than Buzacott's.

There are two interesting areas for further research suggested by these results. First of all, it would be quite useful to develop a cut based algorithm that was linear in the number of cuts rather than quadratic. Such an algorithm would be particularly significant when one considers the exponential growth rate of the number of (s,t)-cuts in a network. A second line of research involves the union of products problem defined in (Ball and Provan 1986). The input to this problem is either a list of pathsets or cutsets. We were able to adapt an all-terminal network reliability algorithm whose running time is proportional to the number of pathsets (spanning trees in this case) to obtain an efficient general approximate algorithm and an efficient special case exact algorithm for this problem. It would seem that the cut based algorithm could also be adapted to the union of products problem context.

REFERENCES

Ball, Michael O., 1986, "Computational Complexity of Network Reliability Analysis: An Overview", IEEE Transactions on Reliability, R-35, 230-239.

Ball, Michael O. and J. Scott Provan, 1986, "Disjoint Products and Efficient Computation of Reliability", Working Paper #86-019, College of Business and Management, University of Maryland at College Park.

Buzacott, John A., 1980, "A Recursive Algorithm for Finding Reliability Measures Related to the Connection of Nodes in a Graph", Networks, 10, 311-327.

Buzacott, John A., 1983, "A Recursive Algorithm for Directed Graph Reliability", Networks, 13, 241-246.

Provan, J. Scott and Michael O. Ball, 1984, "Computing Network Reliability in Time Polynomial in the Number of Cuts", Operations Research, 32, 516-526.

WEAK GREEDY HEURISTICS FOR PERFECT MATCHING*

M. D. Grigoriadis, B. Kalantari, and C. Y. Lai

Department of Computer Science

Rutgers University, New Brunswick, NJ 08903

Abstract. *We consider relaxations of the greedy heuristic for computing minimum-weight perfect matching of complete graphs with edge weights satisfying the triangle inequality. We analyze their worst-case error bounds and time complexities, and report on computational results for a set of randomly-generated points in the Euclidean plane.*

Keywords: Approximate algorithms, complexity, heuristics, weighted perfect matching.

1. Introduction.

A *perfect matching* of an edge-weighted complete graph K_n, n even, is a subset M of $n/2$ of its edges such that no two edges in M are incident upon the same vertex. Its weight $w(M)$ is the total weight of its edges. The *minimum weight perfect matching* problem is to find a matching M^* of minimum weight. The problem may be solved *exactly* in $O(n^3)$ time by Edmonds' algorithm [3, 4] as suggested by Gabow and Lawler [5, 10].

When the edge weights satisfy the triangle inequality, the well-known (ordinary) *greedy* heuristic may be used to obtain an approximate perfect matching. This heuristic repeatedly selects an edge of least weight into the matching, and removes from the graph its two vertices and all edges incident upon them. Reingold and Tarjan [11] have shown that the weight of the so-obtained approximate solution is at most $(\frac{4}{3}) n^{\log (3/2)} - 1$ times that of the optimal matching.

The edge selection criterion at each stage $j = 1, ..., n/2$ may also be viewed as identifying a set S_j of edges from *each* vertex to its least-weight neighbor and then selecting an edge of minimum weight in S_j. In this context, the question naturally arises whether examining fewer vertices and weakening the minimization requirement may lead to greedy heuristics of

---

reasonable performance. In this paper, we discuss the behavior of an exponential number of such greedy heuristics proposed in [7], having time complexities ranging from $O(n^2)$ to $O(n^2 \log n)$, and with worst-case error bounds ranging from $2^{n/2}-1$ to $n/2$. The latter is a bound for a subclass of these heuristics to which the ordinary greedy belongs.

In the next section, we define the class of weak greedy heuristics which select *one edge at each stage*, and we derive worst-case error bounds. In section 3, we examine a number of special cases of interest. In section 4, we discuss the time complexity of these heuristics. In the last section, we present computational results for randomly generated points in the Euclidean plane. Surprisingly, these experiments indicate that there is very little variation in the average performance of these heuristics, regardless of the amount of work spent in selecting the matching edge at each stage. We conclude with a brief discussion of a much stronger class of matching heuristics which select *several* edges at each stage [8].

2. The class of weak greedy heuristics

As a measure of performance, we shall use the ratio $f(n)$ of the weight of the approximate perfect matching M, produced by any heuristic, to that of the optimal matching M^*, i.e. $f(n) = w(M)/w(M^*)$. We define $f(n) = 1$ if $w(M) = w(M^*) = 0$. We shall refer to any *heuristic* algorithm that sequentially selects into the matching M, edges e_j, $j = 1, ..., n/2$, one at a time, as *weak greedy*. If, at any of its stages, a weak greedy selects an edge $e_j = (u, v)$ which is *not* a least-weight edge among those incident on u or v, then $f(n)$ may become infinite. For instance, consider the case where the n vertices consist of $n/2$ distinct points in the plane, and their duplicates. Thus, the weight of each selected edge e_j must be a finite multiple of M_j^*, where M_j^* is an optimal matching of the vertices unmatched by M immediately prior to the selection of e_j. We denote by $GREEDY(\alpha_1, ..., \alpha_{n/2})$ those weak greedy heuristics for which $w(e_j) \leq \alpha_j w(M_j^*)$, where $0 \leq \alpha_j < \infty$, $j = 1, ..., n/2$. Furthermore, we let M_j be the set of matching edges $\{e_j, ..., e_{n/2}\}$ that would be selected by such a heuristic at stages j through $n/2$, and let the corresponding ratios be $f(n-2j+2) = w(M_j) / w(M_j^*)$, where M_j^* is as defined earlier.

Theorem 1: For any given $GREEDY(\alpha_1, ..., \alpha_{n/2})$,

$$f(n-2j+2) \leq \alpha_j + (1+\alpha_j) f(n-2j), \quad j = 1, ..., n/2-1, \quad \text{and} \quad f(2) \leq \alpha_{n/2}.$$

Proof: For $j = 1, ..., n/2-1$ we have $f(n-2j+2) = w(M_j) / w(M_j^*) = (w(e_j) + w(M_{j+1})) / w(M_j^*) \leq \alpha_j + w(M_{j+1}) / w(M_j^*)$. It suffices to show that $w(M_{j+1}) / w(M_j^*) \leq (1 + \alpha_j) f(n-2j)$, or equivalently that

$$w(M_{j+1}^*) \leq (1 + \alpha_j) w(M_j^*). \tag{1}$$

Let $e_j = (u, v)$ be the edge selected at the $j-th$ stage of the heuristic. If an optimal perfect matching also contains e_j, then (1) holds trivially, since in this case $w(e_j) + w(M_{j+1}^*) = w(M_j^*)$. Otherwise, the optimal matching must contain edges $e' = (u, u')$ and $e'' = (v, v')$. Let e be the edge (u', v'). Then, $M' = M_j^* \cup \{e\} \setminus \{e', e''\}$ forms a perfect matching of the vertices matched by M_{j+1}^*. Clearly, $w(M') \leq w(M_j^*) + w(e) - w(e') - w(e'')$. By triangle inequality we have $w(e) \leq w(e') + w(e'') + w(e_j)$, and hence (1) holds for all $j = 2, ..., n/2$. Clearly, $f(2) \leq \alpha_{n/2}$. ∎

In order to derive a bound on the $f(n)$ of $GREEDY(\alpha_1, ..., \alpha_{n/2})$ heuristics, we require the following lemma:

Lemma 1: Suppose that $g(n-2j+2) = \alpha_j + (1+\alpha_j) g(n-2j)$ for $j = 2, ..., n/2-1$ and $g(2) = \alpha_{n/2}$. Then, $g(n) = \prod_{j=1}^{n/2} (1 + \alpha_j) - 1$.

Proof: Let $g^*(n-2j+2) = g(n-2j+2) + 1$ for $j = 1, ..., n/2$. Then,

$$g^*(n-2j+2) = (1+\alpha_j) g^*(n-2j) \text{ for } j = 2, ..., n/2-1, \text{ with } g^*(2) = 1 + \alpha_{n/2}. \tag{2}$$

Solving (2) recursively, we obtain $g^*(n) = \prod_{j=1}^{n/2} (1 + \alpha_j)$. ∎

Theorem 2: For any $GREEDY(\alpha_1, ..., \alpha_{n/2})$, the ratio $f(n) \leq \prod_{j=1}^{n/2} (1 + \alpha_j) - 1$.

Proof: Considering the definitions of $f(n)$ in Theorem 1 and $g(n)$ in Lemma 1, it suffices to show that $g(n) \geq f(n)$. This is certainly the case for $n = 2$. Suppose $g(n-2) \geq f(n-2)$. Then, $g(n) = \alpha_1 + (1 + \alpha_1) g(n-2) \geq \alpha_1 + (1 + \alpha_1) f(n-2) \geq f(n)$. ∎

3. Special cases

Let k_j be an integer satisfying $1 \leq k_j \leq n-2j+2$ for each $j = 1, ..., n/2$. At each *stage* j, such a heuristic arbitrarily selects k_j distinct vertices, determines the multiset S_j of edges which connect these k_j vertices to their least-weight neighbors, and selects an edge, say, $e_j \in S_j$, whose weight $w(e_j)$ does not exceed the *average* of all edges in S_j. Then, it places e_j into the matching, and removes from the graph its two vertices, along with all edges incident upon them.

Lemma 2: For each $j=1,...,n/2$, the parameter $\alpha_j \le 1$ if $k_j=1$, and $\alpha_j \le 2/k_j$ otherwise.

Proof: Clearly, for $k_j=1$ we have $w(e_j) \le w(M_j^*)$. Now, suppose $k_j > 1$, and let (u,v) be an edge in M_j^*. If u' and v' are least-weight neighbors of u and v, respectively, then $w(u,u') + w(v,v') \le 2w(u,v)$. At each stage j, we must then have $\sum_{e \in S} w(e) \le 2w(M_j^*)$, and due to the particular selection of edge e_j, we have $k_j w(e_j) \le 2w(M_j^*)$. ∎

For notational convenience, we denote the above heuristics by $GREEDY(k_1,...,k_{n/2})$. Clearly, there are exponentially many such heuristics. We will consider some special cases. At one extreme is $GREEDY(1,...,1)$, which is the "weakest". At the other extreme is $GREEDY(n,n-2,...,2)$, which forms the "strongest" subclass of $GREEDY(k_1,...,k_{n/2})$. $GREEDY(n,n-2,...,2)$, with the additional stipulation that, at each stage j, an edge $e_j \in S_j$ of *minimum weight* is selected, is precisely the ordinary greedy described in section 1. In between, there are several interesting heuristics. For instance, one may examine no more than a *fixed number* of vertices at each stage: For a given constant $3 \le k \le n$ let $k_j=k$ for all j such that $k \le n-2j+2$, and $k_j=n-2j+2$ otherwise. This results in $GREEDY(k,...,k,2\lceil k/2 \rceil-2,...,2)$. Alternately, one may examine the least-weight neighbors of a *fixed portion* of the vertices at each stage, i.e. $k_j = \lceil (n-2j+2)/p \rceil$ for $j=1,...,n/2$, where p is a positive integer.

Theorem 3: The ratios $f(n)$ for $GREEDY(1,...,1)$, $GREEDY(k,...,k,2\lceil k/2 \rceil-2,...,2)$, $GREEDY(n,n-2,...,2)$, and $GREEDY(\lceil n/p \rceil, \lceil (n-2)/p \rceil,...,\lceil 2/p \rceil)$, are bounded above by $2^{n/2}-1$, $\lceil k/2 \rceil(1+2/k)^{n/2-\lceil k/2 \rceil+1}-1$, $n/2$, and $2^p(p!)^2\binom{\frac{n}{2}+p}{p}/(2p)!-1$, respectively.

Proof: For each case we compute the bounds on the α_j's from Lemma 1, and apply Theorem 2: For the first one of these heuristics, we have $\alpha_j \le 1$ for all j. For the second, we have $\alpha_j \le 2/k$ for $j=1,...,n/2-\lceil k/2 \rceil+1$, and $\alpha_j \le 2/(n-2j+2)$ for $j=n/2-\lceil k/2 \rceil+1,...,n/2$. For the third, the bound is directly obtained from the previous case by selecting $k=n$. For the last one, $\alpha_j \le 2p/(n-2j+2)$ for $j=1,...,(n-2p+2)/2-1$, and $\alpha_j \le 1$ for $j=(n-2p+2)/2,...,n/2$. ∎

Although the error bounds for the first three of these heuristics are tight [7], more restrictive edge-selection criteria may result in improved error bounds, e.g. the ordinary-greedy implementation of $GREEDY(n,n-2,...,2)$ results in an $f(n) \le (\frac{4}{3})n^{\log(3/2)}-1$ (see [11]).

912

It is worthwhile to note that distinct members of $GREEDY(k_1, ..., k_{n/2})$ may have identical worst-case error bounds that may be tight [7]. Such is the case for all $GREEDY(k_1, ..., k_{n/2})$ with $k_j = 1$ or 2.

4. Time complexity

The time complexity of all $GREEDY(k_1, ..., k_{n/2})$ heuristics is bounded above by that of the ordinary greedy which is $O(n^2 \log n)$ for general weights and $O(n^{1.5} \log n)$ for Euclidean problems in the plane (see Bentley and Saxe [1]). However, an improvement is possible for some special cases. For instance, $GREEDY(1, ..., 1)$ can be implemented in $O(n^2)$ time for general weights satisfying the triangle inequality. Furthermore, if the vertices are selected *randomly* at each stage, the average time complexity for the Euclidean case is $O(n \log n)$: Initially construct the Delaunay triangulation of the given vertices in $O(n \log n)$ time (see e.g., Guibas and Stolfi [9]), and update this data structure after each stage. The latter requires $O(d \log d)$ time per edge deletion, where d is the sum of the degrees of its two vertices. The average degree of a vertex in the triangulation is 6. Also, the average degree of a vertex that is a nearest neighbor of another vertex is at most 36, since no vertex can be the nearest neighbor of more than 6 vertices. Hence, the average of the sum of the degrees of a vertex and of its nearest neighbor is at most 42, and the average time to implement the deletion of an edge in the triangulation is $O(1)$. Similarly, $GREEDY(k, ..., k, 2\lceil k/2 \rceil - 2, ..., 2)$ can be implemented to run in $O(kn^2)$ time for general weights. The previous argument may be generalized to show that, on the average, the sum of the degrees of the vertices, upon which the edges in S_j (defined in section 3) are incident, is $O(k)$. Thus, the average time complexity of this heuristic for Euclidean problems in the plane is $O(\max \{n \log n, n k \log k\})$.

5. Discussion

In spite of the large spread in the theoretical error bounds of these greedy heuristics, our preliminary computational results with vertices uniformly distributed in the plane indicate that their average performance is surprisingly similar. For each $n = 50, 100, 200, 300$, we generated 10 problems and solved each one by the two "extreme" greedy heuristics, i.e. $GREEDY(1, ..., 1)$ and the ordinary greedy. The ratio of the average weight of the solutions

obtained by the former to that obtained by the latter was found to be within $[1.08, 1.15]$. We also solved these problems optimally using an implementation of Edmonds' algorithm given in Burkard and Derigs [2]. The ratio of solutions obtained by $GREEDY(1, ..., 1)$, and by the ordinary greedy, to the optimal solution was found to be within $[1.29, 1.45]$ and $[1.17, 1.29]$, respectively. Based upon these preliminary results, we conclude that the performance exhibited by $GREEDY(1, ..., 1)$ could be satisfactory for large practical instances of the class of problems we have tested.

The worst-case error of the heuristics discussed in this paper can be considerably reduced by selecting more than one edge into the matching at each stage. Grigoriadis and Kalantari [8], have proposed such a class of heuristics, called *"ONE-THIRD(k)"*, where $0 \leq k \leq \lfloor \log_3 n \rfloor$ is the number of heuristic stages. If $k = 0$, the heuristic is bypassed and the problem is solved by an exact algorithm. Otherwise, it selects a matching S_1 of K_n, with $|S_1| = \lfloor \frac{1}{3} n \rfloor$ such that $w(S_1)$ does not exceed a certain portion of the optimal weight. Then, it discards all vertices matched by S_1, as well as, their incident edges. This results in a reduced complete graph. The heuristic repeats this process $k-1$ times, and at its $(k+1)$-st stage, it obtains an *optimal* perfect matching of the remaining complete graph.

At each stage j, the set of matching edges S_j is obtained by first extracting a subgraph whose edges are determined by computing the least-weight neighbors of its vertices. Then, the edges of each connected component of this graph are duplicated to give an Eulerian multigraph, and an Euler tour is constructed which is subsequently reduced to a Hamiltonian tour. From the union H_j of these Hamiltonians, a subset S_j of size $|S_j| = \lfloor \frac{1}{3} | H_j | \rfloor$ is selected with $w(S_j) \leq \frac{1}{3} w(H_j)$.

This class of heuristics produces an approximate solution with weight of at most $2 (\frac{7}{3})^k - 1$ times the optimal weight, in $O(\max \{n^2, (n/3^k)^3\})$ time. For Euclidean problems, this time can be reduced to $O(\max \{n \log n, (n/3^k)^3\})$. These are substantial improvements over the class of weak greedy heuristics which select a single edge at each stage. In particular, when $k = \lfloor \frac{1}{3} \log_3 (n^2/\log_3 n) \rfloor$, we have $f(n) \leq 2 (n^2/\log_3 n)^\alpha - 1$, where $\alpha = \frac{1}{3} \log_3 (\frac{7}{3}) = 0.2571$. The time complexity is $O(n^2)$ for general weights and $O(n \log n)$ for problems in the Euclidean plane. This case is "optimal" with respect to time, since in the worst case, no

heuristic with lower time complexity can produce a finite-valued $f(n)$ [6]. Whether these theoretical improvements over existing matching heuristics may also offer superior performance in practice is the subject of further computational experimentation.

6. References

1. J. L. Bentley and J. B. Saxe, "Decomposable searching problems 1: Static to dynamic transformations", *Journal of Algorithms* 1(1980) 301-358.

2. R. E. Burkard and U. Derigs, *Assignment and matching problems: Solution Methods with FORTRAN Programs*, Lecture Notes in Economics and Mathematical Systems 184, Springer-Verlag, Berlin, Heidelberg, New York (1980).

3. J. Edmonds, "Matching and a polyhedron with 0 - 1 vertices", *Journal of Research National Bureau of Standards* 69B(1965) 125-130.

4. J. Edmonds, "Paths, trees and flowers", *Canadian Journal of Mathematics* 17(1965) 449-467.

5. H. N. Gabow, "Implementation of algorithms on nonbipartite graphs", Ph. D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA (1973).

6. M. D. Grigoriadis and B. Kalantari, "A lower bound to the complexity of Euclidean and rectilinear matching algorithms", *Information Processing Letters* 22 (1986) 73-76.

7. M. D. Grigoriadis and B. Kalantari, "On the existence of weak greedy matching heuristics", LCSR TR-75, Laboratory for Computer Science Research, Department of Computer Science, Rutgers University, New Brunswick, NJ 08903 (1985). To appear in *Operations Research Letters*.

8. M. D. Grigoriadis and B. Kalantari, "A new class of heuristic algorithms for weighted perfect matching", LCSR TR-76, Laboratory for Computer Science Research, Department of Computer Science, Rutgers University, New Brunswick, NJ 08903 (1985).

9. L. J. Guibas and J. Stolfi, "Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams", *Proc. Fifteenth Annual ACM Symposium on Theory of Computing* (1983) 221-234.

10. E. L. Lawler, *Combinatorial optimization: Networks and matroids*, Holt, Rinehart and Winston, New York (1976).

11. E. M. Reingold and R. E. Tarjan, "On a greedy heuristic for complete matching". *SIAM Journal on Computing* 10(1981) 676-681.

SENSITIVITY ANALYSIS

FOR

STATIONARY PROBABILITIES OF MARKOV CHAINS

Peter W. Glynn[1]
University of Wisconsin

ABSTRACT

This paper considers the problem of evaluating the sensitivity of a steady-state cost $\alpha(\theta)$ to underlying uncertainty in a parameter vector θ governing the probabilistic dynamics of the system under consideration. We show that the gradient $\nabla\alpha(\theta)$ plays a fundamental role in the parametric statistical theory for Markov processes. We then survey numerical methods available for evaluating $\nabla\alpha(\theta)$ and introduce a new Monte Carlo estimator for $\nabla\alpha(\theta)$, which is applicable to Markov processes of substantial generality.

1. INTRODUCTION

Let $X = \{X_n : n \geq 0\}$ be an irreducible positive recurrent Markov chain governed by a transition kernel $P(\theta)$, where θ is a parameter vector taking values in $\mathbb{R}^d$. If $\pi(\theta)$ is the stationary measure of $P(\theta)$ and $f(\theta,x)$ is the cost of running the chain while in state x, then $\alpha(\theta) \equiv \int f(\theta,x)\, \pi(\theta,dx)$ is the long-run average cost of running X under parameter choice θ. In many applications settings, it is of interest to compute the sensitivity of α to (infinitesimal) changes in the parameter θ. Specifically, it is frequently useful to be able to evaluate $\nabla\alpha(\theta)$, the gradient of $\alpha(\cdot)$ evaluated at $\theta \in \mathbb{R}^d$. Since it is generally impossible to analytically evaluate $\nabla\alpha(\theta)$ (except for simple models), this paper will concentrate on numerical methods for determining $\nabla\alpha(\theta)$.

This paper is organized as follows. In Section 2, we introduce an important statistical application for these methods. We show that the numerical methods discussed here offer the opportunity to do statistical point, variance, and interval estimation for highly complex functionals of analytically intractable Markov processes. Section 3 is devoted to the formal derivation of an expression for $\nabla\alpha(\theta)$ and describes, for finite state Markov chains, a set of linear equations which characterizes $\nabla\alpha(\theta)$. For complicated stochastic processes, the corresponding linear systems are too complex to solve via standard numerical methods, and Monte Carlo techniques therefore become relevant. Thus, Section 4 provides a (new) Monte Carlo estimator for $\nabla\alpha(\theta)$, which is applicable to Markov chains of substantial generality. Finally, Section 5 offers a brief summary of the paper.

2. STATISTICAL RELEVANCE OF THE GRADIENT

Suppose that the transition kernel P governing the Markov chain X is determined by a finite family of distributions $(F_1, \ldots, F_n)$ $(F_1(\gamma_1), \ldots, F_m(\gamma_m))$, where each $F_i(\gamma_i)$ is a probability distribution associated with a known parametric family in which $\gamma_i \in \mathbb{R}^{d_i}$. If $\theta = (\gamma_1, \ldots, \gamma_m)$, then P can be viewed as a function of θ, namely $P = P(\theta)$.

In statistical contexts, the vector $\theta \in \mathbb{R}^d$ $(d = d_1 + \cdots + d_m)$ is, in general, unknown. Most of the literature on statistical inference for Markov processes has concentrated on estimation of the "true" parameter θ^* (i.e., estimation of θ when the observed chain X is governed by $P(\theta^*)$) and on related issues such as production of variance estimates and confidence intervals. However, in many applications settings, it is of more practical importance to estimate not θ^* but some associated steady-state cost $\alpha(\theta^*)$.

(2.1) EXAMPLE. Let $X = \{X_n : n \geq \}$ be the Markov chain consisting of waiting times of consecutive customers in the M/M/1/∞ queue. (See HEYMAN and SOBEL (1982) for a description.) Arrivals follow an $\exp(\gamma_1)$ distribution, whereas service times are distributed $\exp(\gamma_2)$. Suppose that the long-run customer waiting time $\alpha(\theta)$ is of importance, when $\theta = (\gamma_1, \gamma_2)$. The objective is to produce estimates for $\alpha(\theta^*)$, as well as variance and interval estimates, from observed inter-arrival times $Y_{11}, \ldots, Y_{n_1}$ as well as observed service times $Y_{21}, \ldots, Y_{2n_2}$. Note that in certain settings, the inter-arrival times and service times may have been collected

from two independent sources, so that no waiting times for the system are available. For example, the queue might correspond to a telephone switching system being designed, in which historical inter-arrival data exists and service time data for the proposed switching device is available.

(2.2) EXAMPLE. Virtually any general discrete-event stochastic system can be formulated as a generalized semi-Markov process (GSMP). A GSMP can, in turn, be viewed as a Markov chain $X = \{X_n : n \geq 0\}$, where $X_n = (S_n, C_n)$ records the "physical state" S_n (e.g., configuration of customers in a queue) and clock readings C_n (e.g., remaining service times for each of the customers in the system) at the nth transition of the GSMP. (For further details, see GLYNN (1983).) GSMP's are characterized probabilistically by certain distributions $F_1, \ldots, F_\ell$ governing the way clocks are reset (e.g., service times in a queue) and by routing probabilities $p_1, \ldots, p_k$ (e.g., the proportion of customers who visit station j after receiving service at station i).

In many applications environments, the distributions $F_1, F_2, \ldots, F_\ell$ and routing probabilities $p_1, \ldots, p_k$ are unknown and must be estimated via statistical methods. If one models the distributions $F_1, \ldots, F_\ell$ as belonging to parametric families (i.e., $F_i = F_i(\gamma_i)$), then the transition function P governing X can be viewed as $P = P(\theta)$, where $\theta = (\gamma_1, \ldots, \gamma_\ell, p_1, \ldots, p_k)$. The performance of a stochastic system is often assessed by considering a long-run average cost α for the system which, in this context, can be regarded as a function $\alpha = \alpha(\theta)$ of the unknown parameter θ associated with P. Consequently, an important

statistical objective involves point and interval estimation of $\alpha(\theta^*)$, where θ^* is the "true" parameter governing the system.

We will now outline a method for obtaining point and interval estimates for $\alpha(\theta^*)$, which is applicable to very general stochastic systems. Let $\hat{\theta} = (\hat{\theta}_1(n_1), \ldots, \hat{\theta}_d(n_d))$ be an estimator for $\theta^* = (\theta_1^*, \ldots, \theta_d^*)$ (n_1 is the sample size associated with estimation of θ_1^*.) Such estimators $\hat{\theta}$ are frequently available for complex systems. In particular, one can often appeal to maximum likelihood estimation (MLE) methods for estimating θ^*. Under very general conditions, $\hat{\theta}$ will be asymptotically normal, in the sense that

$$(2.3) \qquad \hat{\theta} \overset{\mathcal{D}}{\approx} N(\theta^*, C(n_1, \ldots, n_d))$$

where $N(\theta^*, C(n_1, \ldots, n_d))$ is a multivariate normal r.v. with mean θ^* and covariance matrix $C(n_1, \ldots, n_d)$. ($\overset{\mathcal{D}}{\approx}$ denotes "has approximately the distribution of".) In certain design settings (see Example 2.1), the data for each of the different components θ_i^* is gathered from independent sources. In this case, $C(n_1, \ldots, n_d)$ takes the diagonal form

$$(2.4) \qquad C(n_1, \ldots, n_d) = \begin{pmatrix} \sigma_1^2/n_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d^2/n_d \end{pmatrix}.$$

If α is continuously differentiable in a neighborhood of θ^*, then a Taylor expansion of α around θ^* shows that (2.3) yields

$$(2.5) \qquad \alpha(\hat{\theta}) \overset{\mathcal{D}}{\approx} N(\alpha(\theta^*), \nabla\alpha(\theta^*)^t \, C(n_1, \ldots, n_d) \, \nabla\alpha(\theta^*))$$

where $\nabla\alpha(\theta^*)$ is the (column) gradient of α evaluated at θ^*. (This is the so-called "delta method" of statistics.)

Relation (2.5) shows that if $n_1, \ldots, n_d$ are large, then $\alpha(\hat{\theta})$ is a good point estimator for $\alpha(\theta^*)$. Let $\hat{C}(n_1, \ldots, n_d)$ be an estimator for $C(n_1, \ldots, n_d)$ (such variance estimators are commonly available for MLE point estimates $\hat{\theta}$). Then, (2.5) proves that

$$(2.6) \qquad \hat{v} \equiv \nabla\alpha(\hat{\theta})^t \, \hat{C}(n_1, \ldots, n_d) \, \nabla\alpha(\hat{\theta})$$

is an estimator for the variance of $\alpha(\hat{\theta})$ and,

$$(2.7) \qquad [\alpha(\hat{\theta}) - z(\delta)\hat{v}^{1/2}, \alpha(\hat{\theta}) + z(\delta)\hat{v}^{1/2}\}$$

is an approximate $100(1-\delta)\%$ confidence interval for $\alpha(\theta^*)$, where $z(\delta)$ is the solution of $P\{N(0,1) \leq z(\delta)\} = 1 - \delta/2$. Thus, provided that $\alpha(\hat{\theta})$ and $\nabla\alpha(\hat{\theta})$ can be evaluated (either analytically or numerically), (2.6) and (2.7) provide a solution to the variance and interval estimation problems discussed above.

In the case that the covariance matrix $C(n_1, \ldots, n_d)$ takes the form (2.4), $\hat{v}$ can be expressed as

$$(2.8) \qquad \hat{v} = \sum_{i=1}^{d} \left(\frac{\partial}{\partial\theta_i} \alpha(\hat{\theta})\right) \hat{\sigma}_i^2/n_i \; .$$

Relation (2.8) shows that the contribution of uncertainty in θ_i^* to the variance of $\alpha(\hat{\theta})$ is given by $(\partial\alpha(\hat{\theta})/\partial\theta_i)^2 \, \hat{\sigma}_i^2/n_i$. This can be used to determine which component to additionally sample if the current estimator of $\alpha(\theta^*)$ is too "noisy."

(2.1) EXAMPLE (continued). Because of the simplicity of the $M/M/1/\infty$ queue, α can be analytically determined in closed form, namely $\alpha(\gamma_1, \gamma_2) = \gamma_2(\gamma_2-\gamma_1)^{-1}$ for $\gamma_1 < \gamma_2$ (∞ for $\gamma_1 \geq \gamma_2$). If $\hat{\gamma}_1 < \hat{\gamma}_2$, (2.8) reduces to

$$\hat{v} = \frac{\hat{\gamma}_2^2}{(\hat{\gamma}_2-\hat{\gamma}_1)^4}\, \hat{\sigma}_1^2/n_1 + \frac{\hat{\gamma}_1^2}{(\hat{\gamma}_2-\hat{\gamma}_1)^4}\, \hat{\sigma}_2^2/n_2\ ,$$

where $\hat{\sigma}_1^2(\hat{\sigma}_2^2)$ is a variance estimate for $\gamma_1(\gamma_2)$ formed from $Y_{11}, \ldots, Y_{1n_1}(Y_{21}, \ldots, Y_{2n_2})$.

For more complicated systems, such as that described in Example 2.2, $\alpha(\cdot)$ cannot be determined analytically, and so one must turn to numerical algorithms. These algorithms will be described in the remaining sections of this paper.

3. A FORMULA FOR THE GRADIENT OF THE STEADY-STATE

Let $P(\theta)$ be the transition function for X under parameter θ, so that $P(\theta, x, A)$ is the corresponding conditional probability that $X_{n+1} \in A$, given that $X_n = x$. For an initial distribution $\mu(\theta)$, let P_θ be the probability measure on the path-space of X associated with $P(\theta)$, namely

$$P_\theta\{X_0 \in A_0, \ldots, X_n \in A_n\} = \int_{A_0} \mu(\theta, dx_0) \int_{A_1} P(\theta, x_0, dx_1) \cdots \int_{A_n} P(\theta, x_{n-1}, dx_n)\ .$$

If X is Harris recurrent under $P(\theta)$ (see REVUZ (1984)), then there exists a unique probability measure $\pi(\theta)$ such that

$$(3.1) \qquad \frac{1}{n} \sum_{k=0}^{n-1} f(\theta, X_k) \rightarrow \int_S f(\theta, x) \, \pi(\theta, dx) \, P_\theta \text{ a.s.}$$

as $n \rightarrow \infty$ (for a large class of $f(\theta)$'s). The measure $\pi(\theta)$ is stationary for $P(\theta)$, in the sense that

$$(3.2) \qquad \pi(\theta, \cdot) = \int_S P(\theta, x, \cdot) \, \pi(\theta, dx) \, .$$

(S is the state space of X.) In fact, $\pi(\theta)$ is the unique probability measure satisfying (3.2). Our goal is to numerically compute $\alpha(\theta)$ and $\nabla \alpha(\theta)$, where $\alpha(\theta)$ is the steady-state limit

$$(3.3) \qquad \alpha(\theta) = \int_S f(\theta, x) \, \pi(\theta, dx) \, .$$

Since (3.2) only determines $\pi(\theta)$ up to a multiplicative constant, it is necessary to add an additional constraint stating that the total mass $\pi(\theta, S)$ equals 1. The quantity $\alpha(\theta)$ is then the unique solution of the integral equation system

$$
\begin{aligned}
\pi(\theta, \cdot) &= \int_S P(\theta, x, \cdot) \, \pi(\theta, dx) \\
\pi(\theta, S) &= 1 \\
\alpha(\theta) &= \int_S f(\theta, x) \, \pi(\theta, dx) \, .
\end{aligned}
$$

(3.4)

The system (3.4) is well known and has been extensively studied. If S is finite, then $P(\theta)$ is a finite matrix and (3.4) becomes

$$\pi(\theta)^t = \pi(\theta)^t P(\theta)$$

(3.5)
$$\pi(\theta)^t e = 1$$

$$\alpha(\theta) = \pi(\theta)^t f(\theta)$$

(all vectors are column vectors; e is the vector consisting of 1's).

As we shall see, a similar system describes the gradient $\nabla\alpha(\theta)$ of α. Let us formally suppose that the transition function $P(\theta)$ can be expanded as

$$P(\theta + he_i) = P(\theta) + hQ_i(\theta) + o(h)$$

where e_i is the ith unit vector in $\mathbf{R}^d$. Assume that $\pi(\theta+he_i)$ is formally differentiable at $h = 0$, so that there exists a signed measure $\eta_i(\theta)$ such that

(3.6)
$$\pi(\theta + he_i) = \pi(\theta) + h\eta_i(\theta) + o(h) .$$

The stationarity equation (3.2) implies that $\eta_i(\theta)$ must satisfy

(3.7)
$$\eta_i(\theta,dx) - \int_S \eta_i(\theta,dx) \; P(\theta,x,\cdot) = \int_S Q_i(\theta,x,\cdot) \; \pi(\theta,dx)$$

(formally differentiate both sides of (3.2)). (The equation (3.7) is Poisson's equation for the kernel $P(\theta)$.) These formal calculations can be made rigorous, even in general state space; such arguments will appear elsewhere.

In finite state space, the arguments are more straightforward and have previously appeared in SCHWEITZER (1968), GOLUB and MEYER (1986), and MEYER

and STEWART (1986). We give a very elementary proof in the Appendix to this paper; our argument uses only elementary Markov chain theory. Note that in finite state space, (3.7) becomes $\eta_i(\theta)^t(I-P(\theta)) = \pi(\theta)^t Q_i(\theta)$. This does not uniquely identify $\eta_i(\theta)$, since $\eta_i(\theta) + \delta\pi(\theta)$ also satisfies the equation, for all δ. Note that since $\pi(\theta)^t e = 1$ for all θ, it follows that $\eta_i(\theta)^t e = 0$ (see (3.6)). Let $\Pi(\theta)$ be the matrix in which all rows are identical to $\pi(\theta)$. It is easily verified that since $\eta_i(\theta)^t e = 0$, $\eta_i(\theta)^t \Pi(\theta) = 0$. Consequently, $\eta_i(\theta)$ also satisfies

$$(3.8) \qquad \eta_i(\theta)^t(I - P(\theta) + \Pi(\theta)) = \pi(\theta)^t Q_i(\theta) \ .$$

It is well known (see KEMENY and SNELL (1960), p. 100) that $(I-P(\theta)+\Pi(\theta))$ has an inverse, called the fundamental matrix, which we shall denote $F(\theta)$. Hence, in finite state space, the ith component of $\nabla\alpha(\theta)$ can be computed as the solution of the system

$$(3.9) \qquad \begin{aligned} \eta_i(\theta)^t &= \pi(\theta)^t Q_i(\theta) F(\theta) \\ \frac{\partial}{\partial\theta_i} \alpha(\theta) &= \pi(\theta)^t f_i'(\theta) + \eta_i(\theta)^t f(\theta) \end{aligned}$$

where $f_i'(\theta)$ is the vector in which the jth component is $\partial f(\theta,j)/\partial\theta_i$.

Consequently, when S is finite, the systems of linear equations (3.5) and (3.9) may be solved numerically to obtain $\alpha(\theta)$ and $\nabla\alpha(\theta)$. If S is not finite (or if the number of elements in S is large), numerical methods not dependent on explicit solution of linear equations must be considered. In the next section, we show how Monte Carlo methods can be used to advantage here.

4. MONTE CARLO EVALUATION OF STEADY-STATE GRADIENTS

A critical assumption underlying the analysis of this section is that it is possible to generate sample trajectories of X under the measure P_θ. For the examples that we have in mind (see particularly Example 2.2), this assumption is clearly in force.

Assuming now that X has distribution P_θ, relation (3.1) states that

$$(4.1) \qquad \frac{1}{n} \sum_{k=0}^{n-1} f(\theta, X_k) \rightarrow \alpha(\theta) \qquad P_\theta \text{ a.s.}$$

as $n \rightarrow \infty$. In other words, rather than solving the integral equation system (3.4), one may numerically approximate $\alpha(\theta)$ by the sample average appearing on the left-hand side of (4.1). The simplicity of this numerical procedure, as well as its broad applicability, is the source of the power of the Monte Carlo method. Our objective here is to obtain a similar Monte Carlo algorithm for evaluation of the gradient $\nabla\alpha(\theta)$.

Observe that (at least formally) we have

$$(4.2) \qquad \frac{\partial}{\partial\theta_i} \alpha(\theta) = \int_S \frac{\partial}{\partial\theta_i} f(\theta,x)\ \pi(\theta,dx) + \int_S f(\theta,x)\ \eta_i(\theta,dx)\ .$$

A Monte Carlo estimator for the first term appearing on the right-hand side of (4.2) is given by the sample mean

$$(4.3) \qquad \frac{1}{n} \sum_{k=0}^{n-1} \frac{\partial}{\partial\theta_i} f(\theta, X_k)\ .$$

It remains to obtain an estimator for the second term.

As in the finite state space context, one expects that the signed measure $\eta_1(\theta)$ will satisfy $\eta_1(\theta,S) = 0$. As a consequence, it follows from (3.7) that $\eta_1(\theta)$ should satisfy

$$(4.4) \qquad \eta_1(\theta,\cdot) - \int_S \eta_1(\theta,dx)\, P(\theta,x,\cdot) + \int_S \eta_1(\theta,dx)\, \pi(\theta,\cdot)$$
$$= \int_S Q_1(\theta,x,\cdot)\, \pi(\theta,dx) .$$

Letting $\Pi(\theta)$ be the operator $\Pi(\theta,x,\cdot) = \pi(\theta,\cdot)$, one can write (4.4) symbolically as

$$(4.5) \qquad \eta_1(\theta)(I - P(\theta) + \Pi(\theta)) = \pi(\theta)\, Q_1(\theta) .$$

(This is the general state space analogue of (3.8).) The formal inverse of $(I-P(\theta)+\Pi(\theta))$ is given by

$$\sum_{k=0}^{\infty} (P(\theta) - \Pi(\theta))^k .$$

Because of the stationarity of $\pi(\theta)$ and the independence of $\Pi(\theta,x,\cdot)$ from x, it follows that $(P(\theta)-\Pi(\theta))^k = P(\theta)^k - \Pi(\theta)$, for $k \geq 1$. Hence, a formal analysis of (4.5) shows that

$$\eta_1(\theta) = \pi(\theta)\, Q_1(\theta) + \sum_{k=1}^{\infty} \pi(\theta)\, Q_1(\theta)(P^k(\theta) - \Pi(\theta)) .$$

For the same reason that $\eta_1(\theta,S) = 0$, $Q_1(\theta,x,S) = 0$ and hence $Q_1(\theta)\Pi(\theta) = 0$. Consequently,

928

$$(4.6) \qquad \eta_1(\theta)f(\theta) = \sum_{k=0}^{\infty} \pi(\theta) \, Q_1(\theta) \, P(\theta)^k \, f(\theta) \ .$$

Suppose that the measures $P(\cdot,x,dy)$ are absolutely continuous with respect to $P(\theta,x,dy)$ in a neighborhood of θ. Then, one expects that $Q_1(\theta,x,dy)$ has a density with respect to $P(\theta,x,dy)$, call it $q_1(\theta,x,y)$. A typical term on the right-hand side of (4.6) then takes the form

$$\int_S \pi(\theta,dx) \int_S q_1(\theta,x,y) \, P(\theta,x,dy) \int_S P^k(\theta,y,dz) \, f(\theta,z)$$

which can be represented probabilistically as an expectation:

$$\widetilde{E}_\theta[q_1(\theta,X_0,X_1)f(\theta,X_{k+1})]$$

where $\widetilde{E}_\theta(\cdot)$ is the expectation corresponding to $\widetilde{P}_\theta$, and $\widetilde{P}_\theta$ is the probability on path-space associated with initial distribution $\pi(\theta)$ and transition function $P(\theta)$. Thus, the second term in (4.2) has the formal representation

$$(4.7) \qquad \sum_{k=0}^{\infty} \widetilde{E}_\theta[q_1(\theta,X_o,X_1)f(\theta,X_{k+1})] \ .$$

The formula (4.7) is the key to the Monte Carlo analysis.

Each term in (4.7) can be consistently estimated (under suitable hypotheses) via

$$\frac{1}{n} \sum_{j=0}^{n-1} q_1(\theta,X_j,X_{j+1})f(\theta,X_{j+k+1})$$

when X evolves according to transition function $P(\theta)$ (regardless of X's initial distribution). In order to estimate the infinite sum, a standard device is to consider an estimator of the form

$$(4.8) \qquad \sum_{k=0}^{\ell(n)} \frac{1}{n-\ell(n)} \sum_{j=0}^{n-\ell(n)-1} q_i(\theta, X_j, X_{j+1}) \, f(\theta, X_{j+k+1})$$

where the truncation point $\ell(n)$ is keyed to the sample size n in such a way that $\ell(n) \to \infty$ with $\ell(n)/n \to 0$. The particular choice of $\ell(n)$ effects a compromise between bias and variance effects in estimating the infinite sum (4.7).

Since $q_i(q,x,y)$ is generally easily computable (for S countable, $q_i(\theta,j,k) = (\partial P(\theta,j,k)/\partial\theta_j) \cdot P(\theta,j,k)^{-1}$), (4.7) provides a Monte Carlo solution to estimating the appropriate gradient.

It turns out that (4.6) is closely related to a formula which one obtains when one uses likelihood ratio change-of-measure ideas to evaluate gradients. These connections will be explored more fully in a future paper.

5. SUMMARY

We have shown that the gradient $\nabla\alpha(\theta)$ of steady-state quantity α plays a critical role in the variance and interval estimation theory for steady-state estimators $\alpha(\hat\theta)$ of complex stochastic systems. In some sense, the large-sample variance and interval estimation theory is fully solved given that one can evaluate $\alpha(\hat\theta)$ and $\nabla\alpha(\hat\theta)$. Numerical methods for dealing with $\alpha(\hat\theta)$ when the system is Markov are, of course, well

930

known. However, numerical algorithms for evaluating $\nabla\alpha(\hat\theta)$ are a recent development. We have therefore provided a self-contained exposition of the relevant theory, and discuss both Monte Carlo (see (4.3) and (4.8)) and non-Monte Carlo (see (3.9)) approaches to solving the problem.

APPENDIX

Let $P(\cdot)$ be a family of $n \times n$ stochastic matrices which are:

(i) irreducible in a neighborhood of θ

(ii) differentiable at θ.

Under (i), $P(\cdot)$ has a unique stationary distribution $\pi(\cdot)$ in a neighborhood of θ. Our goal is to rigorously verify the first equation in (3.9).

Given the existence of the inverse matrix $F(\theta) = (I-P(\theta)+\Pi(\theta))^{-1}$, (3.9) follows immediately once the differentiability of $\pi(\theta)$ is established. Note that for h sufficiently small,

$$\pi(\theta+he_i) - \pi(\theta) = \pi(\theta+he_i)^t[P(\theta) + hQ_i(\theta) + o(h)] - \pi(\theta)^t P(\theta)$$

so

$$[\pi(\theta+he_i) - \pi(\theta)]^t(I-P(\theta)) = h\pi(\theta+he_i)^t Q_i(\theta) + o(h)$$

(note that $o(h)\pi(\theta+he_i) = o(h)$ since all terms in $\pi(\theta+he_i)$ are uniformly (in h) bounded by 1). Since $\Pi(\theta)$ has identical rows and $\pi(\theta+he_i)$ is stochastic for $h \geq 0$, it follows that $[\pi(\theta+he_i)-\pi(\theta)]\Pi(\theta) = 0$. Hence,

931

(A1) $$[\pi(\theta+he_1) - \pi(\theta)]^t = h\pi(\theta+he_1)^t \; Q_1(\theta) \; F(\theta) + o(h) \; .$$

Again, since $\pi(\theta+he_1)$ is uniformly bounded in h, it is evident from (A1) that $\pi(\theta+he_1)$ is continuous at h = 0. Thus, (A1) implies that

$$[\pi(\theta+he_1) - \pi(\theta)]^t = h\pi(\theta)^t \; Q_1(\theta) \; F(\theta) + o(h)$$

$$\text{i.e., } \eta_1(\theta)^t = \pi(\theta)^t \; Q_1(\theta) \; F(\theta) \; ,$$

which is the required result.

REFERENCES

1. GLYNN, P.W. (1983). On the role of generalized semi-Markov processes
    in simulation output analysis. Proc. of the 1983 Winter Simulation
    Conference, 38-42.

2. GOLUB, G.H. and MEYER, C.D. (1986). Using the QR factorization and
    group inversion to compute, differentiate, and estimate the sensitiv-
    ity of stationary probabilities for Markov chains. SIAM J. Alg.
    Disc. Meth. 7, 273-281.

3. HEYMAN, D.P. and SOBEL, M.J. (1982). Stochastic Models in Operations
    Research, Volume 1. McGraw-Hill, New York.

4. KEMENY, J.G. and SNELL, J.L. (1960). Finite Markov Chains. Van
    Nostrand, Princeton, New Jersey.

5. MEYER, C.D. and STEWART, G.W. (1986). Derivatives and perturbations
    of eigenvectors. Unpublished manuscript.

6. REVUZ, D. (1984). Markov Chains. North Holland, New York.

7. SCHWEITZER, P.J. (1960). Perturbation theory and finite Markov
    chains. J. Appl. Prob. 5, 401-413.

Stochastic Differential Forms

Eugene Wong[1]

1. Introduction

A substantial body of results on stochastic integration with respect to multiparameter martingales now exists. Yet, as it stands, the theory is not entirely satisfactory in a number of ways. In particular, the *calculus* for stochastic integration, already complicated in two dimension, becomes prohibitively so in higher dimensions. In retrospect, the source of the difficulty seems to be that integration over n-dimensional volumes in n-space is only a very small part of a complete theory of integration in n-space. What seems to be needed is a theory of differential forms involving martingales and integration of such forms on sets of appropriate dimensionality. To embark on a course to develop such a theory is the objective of the work reported here.

Our approach to stochastic differential forms follows the general approach of Whitney [2] and forms are defined as function on chains or functions parametrized by chains satisfying certain continuity conditions. While the flat cochains defined by Whitney ([2], ch. IX) have the representation

$$X(\sigma) = \int_{\sigma} x(t) \, dt_{i_1} \,\hat{}\,\cdots\,\hat{}\, dt_{i_r} \tag{1.1}$$

we cannot expect such a representation to hold for any class of stochastic forms that includes the Wiener process. However, as we intend to show in this paper, an exterior calculus for martingale forms can be constructed without such a representation.

The focus of this paper will be on the conceptual framework needed for the development of a theory of differential forms. Details on some aspects of this paper can be found in a forthcoming paper by M. Zakai and the author [5].

---

[1]Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720.

2. Co-Chains and Form

Let a_j denote a finite interval open to the left and closed to the right on the t_j axis. For $i < j$, $a_i \hat{\ } a_j$ will denote a possibly oriented 2-dimensional rectangle with sides a_i and a_j and $a_j \hat{\ } a_i = -a_i \hat{\ } a_j$ will denote the same rectangle with a negative orientation. In general, let $a_{i_1}, a_{i_2}, \cdots \cdots, a_{i_r}$ denote intervals as above. Then $a_{i_1} \hat{\ } a_{i_2} \hat{\ } \cdots \hat{\ } a_{i_r}$ will denote an r dimensional rectangle with sides $a_{i_1}, a_{i_2}, \cdots \cdots, a_{i_r}$. The orientation is positive if an even permutation of $(i_1, i_2, \cdots \cdots, i_r)$ puts it into increasing order, and the orientation is negative otherwise. We call such rectangles *oriented r—rectangles* and refer to [1] as the direction of $a_{i_1} \hat{\ } a_{i_2} \hat{\ } \cdots \hat{\ } a_{i_r}$.

We note that the boundary $\partial \sigma$ of an oriented $(r+1)$ rectangle σ is a collection of oriented r-rectangles that overlap at most on boundaries. Subdivision of an r-rectangle produces a collection of r-rectangles. It is useful to denote such a collection by a sum $\sigma_1 + \sigma_2 + \cdots \cdots + \sigma_m$. Furthermore if σ is an oriented r-rectangle it is useful to denote by $-\sigma$ the same rectangle with the opposite orientation. It is therefore useful to introduce linear combinations

$$A = \sum_{k=1}^{m} \alpha_k \sigma_k \tag{2.1}$$

where α_k are real numbers taking values in $\{-1, 1\}$ and σ_k are oriented r-rectangles. We shall call any sum of the form (2.1) an *r—chain*.

Let $X(\sigma)$ be a real-valued random function defined on $(\Omega, \underline{\underline{F}}, P)$ and parametrized by oriented r-rectangles such that

(a) $X(\sigma)$ is defined for every oriented r-rectangle σ

(b) $X(\sigma) = -X(-\sigma)$ and for disjoint rectangles $X(\sum_{k=1}^{m} \sigma_k) = \sum_{k=1}^{m} X(\sigma_k)$

We can extend X to all rectangular r-chains by linearity and X so extended is termed a *random r—cochain*.

Intuitively we would like to write a random r-cochain $X(\sigma)$ as an integral over σ

$$X(\sigma) = \int_{\sigma} X$$

where the integrand X is a "stochastic differential r-form." For the concept of stochastic differential forms to be useful, we need to integrate X not only on rectangular chains, but also on suitable r-surfaces. For this purpose, we need an appropriate topology on chains and a corresponding continuity condition on c0-chains with respect to such topology.

Let $|\sigma|$ denote the r-dimensional volume of the oriented rectangle σ with $|\sigma| = 1$ for r=0. For A defined by (2.1) with disjoint σ_k, $k = 1,\cdots,n$, the *mass* of a chain A is defined as

$$|A| = \sum_1 |\alpha_k| \cdot |\sigma_k| .$$

Turning to another norm, let $\{A_m, m = 1,2,\cdots\}$ be a sequence of r chains, we say that the sequence is a *Cauchy sequence* if either

$$|A_m - A_k| \underset{m,k \to \infty}{\to} 0$$

or, if for every m,k there is an $r+1$ chain $B_{m,k}$ such that $\partial B_{m,k} = A_m - A_k$ and

$$|B_{m,k}| \underset{m,k \to \infty}{\to} 0$$

Note that for the convergence of an n-chain in R^n, only the first type of convergence makes sense, while for the convergence of a 1-chain in R^2 to a curve the second type of convergence is necessary. Therefore, it is useful to define the *flat norm* $|A|^\cdot$ for an r-chain in R^n by ([2], p. 154)

$$|A|^\cdot = inf\{|A - \partial B| + |B|\} \tag{2.2}$$

where the infimum is over all $r+1$ chains B. It is shown in [2] that $|A + B|^\cdot \leq |A|^\cdot + |B|^\cdot$ and $|A|^\cdot = 0$ if and only if $A = \emptyset$. Hence, $|\cdot|^\cdot$ is a norm. Furthermore, $|\cdot|^\cdot$ satisfies: (see [2])

$$|\partial A|^\cdot \leq |A|^\cdot \leq |A| \tag{2.3}$$

Note that for $r = n$, $|A|^\cdot = |A|$. For $r = 0$ and A a point in R^n, $|A|^\cdot = 1$. For the case where A is the difference of two points, s and t, $|A|^\cdot = min(2, |(s,t)|)$

We can now define a stochastic differential r-form as the formal integrand of a stochastic r-cochain that is continuous in probability with respect to the flat norm defined in the previous section, i.e.,

$$X(A_m) \xrightarrow{p} 0 \quad \text{whenever} \quad |A_m|^{\cdot} \to 0 \quad \text{as} \quad m \to \infty \; . \tag{2.4}$$

Similarly a stochastic differential form is said to be an L_q form or a q-integrable form if

$$E |X(A)|^q < \infty \tag{2.5}'$$

and

$$E |X(A_m)|^q \to 0 \tag{2.5}''$$

whenever $|A_m|^{\cdot} \to 0$. In (2.4) and (2.5) we extend the definition of X to limits of chains under the flat norm by adjoining $X(A_\infty)$.

As an example let η be "Gaussian white noise" on R_+^2 defined as follows:

(a) $\eta(\sigma)$ is a Gaussian random function parametrized by oriented 2-rectangles σ on R_+^2

(b) $E \eta(\sigma) = 0$

(c) $E \eta(\sigma) \eta(\sigma') = \mu(\bar{\sigma} \cap \bar{\sigma}')$ if σ and σ' are similarly oriented

$$= -\mu(\bar{\sigma} \cap \bar{\sigma}') \quad \text{otherwise}$$

where $\bar{\sigma}$ denotes σ without orientation and μ denotes the Lebesgue measure.

A Wiener process W_t, $t \in R_+^2$, is defined by

$$W_t = \eta(A_t)$$

where A_t is the rectangle $\{s : 0 \leq s \leq t\}$. The white noise η is a random rectangular 2-cochain. Since $E \eta^2(\sigma) = |\sigma|$, (2.5) is satisfied. The Wiener process is a 0-cochain satisfying (2.5).

Next, we define the exterior derivative dX of a random r-cochain X (via the Stokes theorem) as follows. Set

$$dX(A) = X(\partial A) \tag{3.8}$$

for all oriented $(r+1)$ chains A. The exterior derivative of a stochastic differential form as defined by (2.4) and (2.5) is also a stochastic differential form of the same type, this follows directly from

936

the definition of dX and the fact that $|\partial A| \leq |A|$

As we have defined them, random differential r-forms are random currents of Ito [1], but not every Ito current is an r-form in our sense. While linear operations are definable on all random currents, nonlinear operations (e.g., exterior products) are not. The r-forms that we have defined have the right degree of localization to allow exterior products to be defined.

If X is a regular (nongeneralized) differential form, then X can be represented as

$$X_t = \sum_{[i]} \alpha_{[i]}(t) dt_{[i]} \tag{2.7}$$

where the differentials $dt_{[i]} = dt_{i_1} \cdot dt_{i_2} \cdot \cdots \cdot dt_{i_r}$ provide a local coordinate system. For a random current such a representation is in general not possible, but a useful representation similar to this one still exists. For a random cochain X, define $X_{[i]}$ as the cochain such that for every rectangle σ

$$\begin{aligned} X_{[i]}(\sigma) &= X(\sigma) \quad \text{if } \sigma \text{ has the direction } [i] \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{2.8}$$

Then for any rectangular chain A

$$X(A) = \sum_{[i]} X_{[i]}(A) \tag{2.9}$$

and if X is a random differential form so is $X_{[i]}$. Hence we can write

$$X = \sum_{[i]} X_{[i]} \tag{2.10}$$

and this is the equivalent of (2.7) for random differential forms.

The rectangular coordinate system provides an alternate but equivalent definition for the exterior derivative as follows. Define $d_k X_{[i]}$ for rectangles σ as

$$\begin{aligned} d_k X_{[i]}(\sigma) &= dX_{[i]}(\sigma) \quad \text{if } k \text{ is not in } [i] \text{ and } \sigma \text{ has direction } [k,[i]] \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{2.11}$$

Then we can write

$$d_k X = \sum_{[i]} d_k X_{[i]} \tag{2.12}$$

and

$$dX = \sum_k d_k X \tag{2.13}$$

For an example, consider a Wiener process W_t , $t \in R_+^2$. Take σ_1 and σ_2 to be the horizontal and vertical 1-rectangles

$$\sigma_1 = ((t_1,t_2),(t_1+a,t_2))], \qquad \sigma_2 = ((t_1,t_2),(t_1,t_2+b)]$$

oriented from the left (from below) to the right (to above). We have

$$d_1 W(\sigma_1) = W_{t_1+a,t_2} - W_{t_1,t_2}$$

$$d_2 W(\sigma_2) = W_{t_1,t_2+b} - W_{t_1,t_2}$$

$$d_1 W(\sigma_2) = d_2 W(\sigma_1) = 0$$

Now, take a positively oriented 2-rectangle σ with $\underline{t} = (t_1,t_2)$ and $\overline{t} = (t_1+a,t_2+b)$. Its boundary $\partial\sigma$ is given by:

$$\partial\sigma = \{\, \sigma_1 \,, \quad -\sigma_2 \,, \quad ((t_1+a,t_2),(t_1+a,t_2+b)] \,, \quad -((t_1,t_2+b),(t_1+a,t_2+b)]$$

Hence

$$d(d_1 W)(\sigma) = (W_{t_1+a,t_2} - W_{t_1,t_2}) - (W_{t_1+a,t_2+b} - W_{t_1,t_2+b})$$
$$= -\eta(\sigma) \tag{2.14}$$

and

$$d(d_2 W)(\sigma) = \eta(\sigma) \tag{2.15}$$

We can interpret (2.14) and (2.15) as follows

$$d(d_1 W) = d_1 d_1 W + d_2 d_1 W$$

$$d(d_2 W) = d_1 d_1 W + d_1 d_2 W$$

with $d_1 d_1 W = d_2 d_2 W = 0$, $d_2 d_1 W = -d_1 d_2 W$ and $d_1 d_2 W = d_{12} W = \eta$. Observe that

$$ddW = d(d_1 W + d_2 W) = d_2 d_1 W + d_1 d_2 W = 0$$

as it should be.

938

Finally, we note that the Hodge star operator $*$ is a linear operator defined on all Ito random currents. Hence $*X$ is well defined as an Ito random current for any r-cochain X considered as an Ito random current. However, $*X$ is not necessarily a cochain (equivalently a differential form) and for many interesting cases it is not. For example, let η be an n-cochain representing Gaussian white noise, for $*\eta$ to be a 0-cochain it must be a continuous random function. However

$$\eta(\sigma) = \int_\sigma *\eta \; dt_1 {}^\hat{} dt_2 {}^\hat{} \cdots {}^\hat{} dt_n$$

so that $*\eta$ cannot be a continuous random function and hence is not a 0-cochain.

3. Markovian Currents

It is well known [3] that an isotropic Gaussian random field with a covariance function given by

$$EX_t X_s = \int_{R^n} e^{i(\nu,t)} \frac{1}{\alpha^2 + |\nu|^2} d\nu$$

is Markov. Indeed, it is one of the few known examples of Markovian fields. Yet, because $EX^2 = \infty$, X is a random current (generalized field) rather than an ordinary field. Insofar the Markov property requires the consideration of "surface data," how it can be applied to a random current requires a careful interpretation.

Let D_p denote the space of all ordinary p-forms with coefficients that are C^∞ functions with compact support. A random r-current X in R^n is a continuous linear random function on D_{n-r}. For a random r-current X and $f \in D_p$ with $p + r \leq n$, $X \cdot f$ is well defined by

$$(X \cdot f)(g) = X(f \cdot g) \qquad \text{for all } g \in D_{n-p-r} \; .$$

Now we can define localizable currents and Markovian currents. We say an r-current X is *localizable* if $X \cdot f$ is a stochastic $(n-1)$ form for all $f \in D_{n-r-1}$.

Suppose Γ is an $(n-1)$-surface separating R^n into a bounded part B^- form and compounded part B^+. For any current X we can define:

$$past(X) = \left\{ X(f), \quad support(f) \subset B^- \right\}$$

$$future(X) = \left\{ X(f), \quad support(f) \subset B^+ \right\}$$

For a localizable r-current X, we can further define

$$present(X) = \left\{ (X \cdot f)(\sigma), \quad f \in \mathbf{D}_{n-r-1}, \ \sigma \subset \Gamma \right\}$$

We say a localizable 0-current X is Markov if for any Γ, past (X) and future (X) are conditionally independent given present (X). For $1 \le r \le n-1$, we say an r-current X is Markov if both X and $^*X$ are localizable and,

$$\left(past(X), past(^*X) \right) \quad and \quad \left(future(X), future(^*X) \right)$$

are conditionally independent given $\left(present(X), present(^*X) \right)$. An n-current X is said to be Markov if $^*X$ is Markov. Eq. (3.1) yields an 0-current that is Markov.

4. Exterior Product

If X and Y are stochastic differential forms, then any definition of the exterior product $X \cdot Y$ would involve the "product" of generalized processes, not a well defined quantity. For example, if η is a white noise n-form and X is an 0-form, the $(X \cdot \eta)(\sigma)$ is a stochastic integral

$$\int_\sigma x_t \ \eta(dt)$$

Thus, to define exterior products requires a generalization to stochastic integration. One way of defining the exterior product is to define martingale forms, and to require that, for martingale p and r forms X and Y, the exterior product $X \cdot Y$ be a martingale $(p + r)$ form. Defined this way, the exterior product is a generalization of both the martingale stochastic integral for one-parameter processes and the stochastic integrals of types 1 and 2 of Wong and Zakai [4]. The situation can be summarized as follows:

940

	p	t	$X \cdot Y$
n = 1	0	0	product
	0	1	martingale integral
n = 2	0	0	product
	0	1	martingale integral on paths
	0	2	type 1 integral
	1	1	type 2 integral

Details of how $X \cdot Y$ is defined can be found in [5].

References

[1]   K. Ito, Isotropic random currents, Proc. 3rd Berkeley Symp. on Math. Stat. and Prob., pp. 125-132, 1956.

[2]   H. Whitney, Geometric Integration Theory, Princeton Univ. Press, 1957.

[3]   E.Wong, Homogeneous Gauss-Markov random fields, Ann.Math.Stat., Vol.40, pp. 1625-1634, 1969.

[4]   E. Wong and M.Zakai, Martingales and stochastic integrals for processes with a multidimensional parameter, 7. für Wahrsch. v. Geb., Vol.29, pp. 109-122, 1974.

[5]   E. Wong and M. Zakai, Multiparameter martingale differential forms, to appear.

FILTERING AND CONTROL FOR WIDE BANDWIDTH
NOISE AND 'NEARLY' LINEAR SYSTEMS

H. J. Kushner
Lefschetz Center for Dynamical Systems
Division of Applied Mathematics
Brown University
Providence, Rhode Island 02912

and

W. Runggaldier
Institute of Analysis
University of Padua
Padua, Italy

ABSTRACT. Typically, modern stochastic control theory uses ideal white noise driven systems (Itô equations), and if the observed data is corrupted by noise, that noise is usually assumed to be 'white Gaussian'. If the models are linear, a Kalman-Bucy filter is then used to estimate the state, and a control based on this estimate is computed. Actually, the noise processes are rarely 'white', and the system is only approximated in some sense by a diffusion. But, owing to lack of 'computable' alternatives, one still uses the above procedure. Then the 'filter' estimates might be quite far from being optimal. We examine the sense in which such estimates are useful, in order to justify the the use of the commonly used procedure. For the filtering problem where the signal is a 'near' Gauss-Markov process and the observation noise is wide band, it is shown that the usual filter is 'nearly optimal' with respect to a very natural class of alternative data processors. The asymptotic (in time and bandwidth) problem is treated, as is the conditional Gaussian case.

The paper is an outline of some of the work reported in [9].

I. INTRODUCTION. Typical models in modern filtering theory are of the following type, where $W(\cdot)$ are standard Wiener processes, $u(\cdot)$ is a control, and b_z, σ_z, etc., are appropriate functions. We let $z(\cdot)$ denote a reference signal and the noise corrupted observation.

$$(1.1) \qquad dz = b_z(z)dt + \sigma_z(z)dW_z$$

$$(1.2) \qquad dy = h(x,z)dt + dW_y$$

The actual physical system, which we denote by $z^\epsilon(\cdot)$, $y^\epsilon(\cdot)$ is not of the form (1.1) - (1.2). The reference signal $z^\epsilon(\cdot)$ might be only

approximately representable by (1.1), and the noise in the control and observation system would rarely be 'white'. But, via some approximation or identification procedure, one chooses a model of the form (1.1) - (1.2), then computes a good filter for that model, and then applies this filter to the actual physical system. One must question the value of the filter output when applied to the 'physical' problem. The filter output might not be even nearly optimal for use in making estimates of $z^\epsilon(\cdot)$. Such questions are basic to the relevence of much theoretical work. We will deal with these questions here, when the approximating system (1.1), (1.2) is linear - for which a fairly complete theory can be obtained.

Owing to the usual lack of 'near optimality' (when applied to the physical system) of the filter which is obtained by using (1.1) - (1.2), one should ask the question: with respect to which alternative filters (called 'data processors' below) for the physical system is the chosen one nearly optimal? It turns out that this alternative class of filters is quite large and quite reasonable. The basic mathematical techniques used here are those of the theory of weak convergence of probability measures [1], [3], [4], a technique which is quite useful for problems in the approximation of random processes [1], [5] - [8], [12], [13].

When the ideal model is linear - one would usually use the Kalman-Bucy filter appropriate for the ideal model, but whose input is the physical observation. Obviously, the filter does not usually yield the conditional distribution of the $z^\epsilon(t)$ given the data $y^\epsilon(s)$, s ≤ t. In Section 2, we discuss some counter examples to illustrate the sort of difficulties which arise in such approximations, and in Section 3 the approximation theorem is given, together with the class of alternative data processors. Section 4 concerns the average filter error per unit time - or the errors for large time. The symbol ⇒ denotes weak convergence. A fuller development appears in [9], together with the conditional Gaussian case and a treatment of certain non-linear observations. For the weak convergence, we work with the space $D^k[0,\infty)$, the space of R^k-valued functions which are right continuous and have left-hand limits, and endowed with the Skorohod topology. (See [1], [3], [4].) Reference [2] deals with similar approximations for the non-linear filtering problem, and reference [10] concerns the approximation problem for the non-linear control problem.

II. LINEAR FILTERING: PRELIMINARIES. Consider the following filtering problem: For each $\epsilon > 0$, $z^\epsilon(\cdot)$ is a *signal* process, $\xi_y^\epsilon(\cdot)$ is a 'wide-bandwidth' observation noise, and the two are mutually independent. The actual observation process is:

(2.1) $$\dot{y}^\epsilon(t) = H_s z^\epsilon(t) + \xi_y^\epsilon(t), \ y^\epsilon(0) = 0 .$$

944

All 'noise' processes are assumed to be right continuous and have left-hand limits. Define $y^\epsilon(t) = \int_0^t \dot{y}^\epsilon(s)ds$ and $W_y^\epsilon(t) = \int_0^t \zeta_y^\epsilon(s)ds$. Let $z(\cdot)$ satisfy (for matrices A_s , etc.)

$$(2.2) \qquad\qquad dz = A_s z\, dt + B_s dW_s,$$

Since $\zeta_y^\epsilon(\cdot)$ is to be 'nearly' white noise, and $z^\epsilon(\cdot)$ 'nearly' a Gauss-Markov diffusion, let

$$(2.3) \qquad (z^\epsilon(\cdot), W_y^\epsilon(\cdot)) \rightarrow (z(\cdot), W_y(\cdot)) \quad \text{as} \quad \epsilon \rightarrow 0 ,$$

where $W_y(\cdot)$ is a non-degenerate Wiener process. The $W_s(\cdot)$ and $W_y(\cdot)$ must be independent. Also $y^\epsilon(\cdot) \rightarrow y(\cdot)$, where

$$(2.4) \qquad\qquad dy = H_s z\, dt + dW_y , \quad y(0) = 0 .$$

The actual physical system is, of course, 'fixed' and corresponds to some small $\epsilon > 0$. The use of weak convergence here is just a way of embedding the *actual data* in a sequence - so that an approximation method can be used. The approximation of the values of expectations of functions of $z^\epsilon(\cdot)$, conditioned on the data $y^\epsilon(\cdot)$ is not easy in general. Furthermore, we cannot restrict ourselves to Gaussian noise, since it itself is only an approximation to the physical processes.

For (2.2), (2.4), the filter equations are

$$(2.5) \qquad d\hat{z} = A_s \hat{z}\, dt + Q(t)\, [dy - H_s \hat{z}\, dt]$$

$$Q(t) = \Sigma(t) H_s' R_0^{-1}$$

$$(2.6) \qquad \dot{\Sigma} = A_s \Sigma + \Sigma A_s' + B_s B_s' - \Sigma H' R_0^{-1} H \Sigma ,$$

where R_0 = covariance matrix of observation 'noise' $W_y(1)$, which we set to I, unless mentioned otherwise. In practice, with signal $z^\epsilon(\cdot)$ and noise $\zeta_y^\epsilon(\cdot)$, one normally uses (2.6) and (2.5_{WB}):

$$(2.5_{WB}) \qquad\qquad \dot{\hat{z}}^\epsilon = A_s \hat{z}^\epsilon + Q(t)\, [\dot{y}^\epsilon - H_s \hat{z}^\epsilon] .$$

This system is not necessarily even a nearly optimal filter for the physical observation. But, as will be seen, it makes a great deal of sense and is quite appropriate in a specific but important way.

Some illustrations will illustrate the problems that we must contend with, particularly concerning the *possible lack of continuity in the optimal estimators* as the noise bandwidth goes to ∞. Let (X_n, Y_n) be bounded real-valued random variables which converge in distribution to (X, Y). Generally $E(X_n | Y_n) \nrightarrow E(X|Y)$. For example, let $X_n = X$, $Y_n = X/n$. Next, let $Z_n = Z_n(Y)$, where Y is a random varible and $(Z_n, Y) \Rightarrow (Z, Y)$. Then Z is *not* necessarily a function of Y, and might even be independent of Y, as illustrated by the following:

Let Y be uniformly distributed on $[0,1]$. Define $Z_n = nY$ for $0 \leqslant Y < 1/n$ and, in general, define $Z_n = (nY - k)$ on $k/n \leqslant Y < (k+1)/n$, $k = 0,1,...,n-1$. Then $(Z_n, Y) \Rightarrow (Z, Y)$ where Z is independent of Y, and both Z and Y are uniformly distributed on $[0,1]$. Clearly $E(Z_n | Y) \nrightarrow E(Z | Y)$ in any sense.

Even though $W_y^\epsilon(\cdot) \Rightarrow W_y(\cdot)$, a non-degenerate Wiener process, $y^\epsilon(\cdot)$ might contain a *great deal* more information about $z^\epsilon(\cdot)$ than $y(\cdot)$ does about $z(\cdot)$. See [9] for an example where as $\epsilon \to 0$, we can calculate $z^\epsilon(t)$ nearly exactly from the data $y^\epsilon(\cdot)$. In general we have

Let $(X_n, Y_n) \Rightarrow (X, Y)$ (X_n-real valued, Y_n with values in R^ϵ). Then

(2.7)
$$\overline{\lim_n} \, E[X_n - E(X_n | Y_n)]^2 \leqslant E[X - E(X|Y)]^2 .$$

In the above examples, the inequality is strict. The examples do caution us to take considerable care in dealing with information processing with wide bandwidth noise disturbances.

III. THE 'APPROXIMATELY OPTIMAL' LINEAR FILTERING PROBLEM. For the ideal filtering problem (2.2), (2.4), the optimal decisions are functions of $\hat{z}(\cdot)$, $\Sigma(\cdot)$, since these completely determine the conditional distribution. There are no functions of the data which give better estimates. This is *not so* with estimates based on $\Sigma(\cdot)$, $\hat{z}^\epsilon(\cdot)$ for the system $z^\epsilon(\cdot)$, $y^\epsilon(\cdot)$. We now define a class of functions of the observed data $y^\epsilon(\cdot)$ with respect to which functions of $\hat{z}^\epsilon(\cdot)$, $\Sigma(\cdot)$ *are* 'nearly optimal' for small $\epsilon > 0$. We need to specify both a criterion of comparison; i.e., a cost function. Although we use one particular cost function, the general idea and possible extensions should be clear .

Let $\mathfrak{D}$ denote the class of measurable functions on $C[0,\infty]$, the space of real valued continuous functions on $[0,\infty)$ (with the topology of uniform convergence on bounded intervals), which are continuous w.p.1 relative to Wiener measure (hence, with repect to the measure of $y(\cdot)$). Let $\mathfrak{D}_t$ denote the subclass which depends only on the function values up to time t. For arbitrary $F(\cdot) \in \mathfrak{D}$ or in $\mathfrak{D}_t$, we will use $F(y^\epsilon(\cdot))$ as an *alternative estimator* of a functional of $z^\epsilon(\cdot)$. The class is quite large.

First, note that $\mathfrak{D}$ contains all continuous functions and that the $\hat{z}(\cdot)$ of (2.5) can be written as a continuous function of the integral of the driving force $y(\cdot)$. Thus, continuous functions of $\hat{z}^\epsilon(\cdot)$ are admissible estimators. Many important functionals are only continuous w.p.1 (relative to Wiener measure). Let $\tau(x(\cdot))$ denote the first time that a closed set A with a piecewise differential boundary is reached by $x(\cdot)$. Then the function with values $T \cap \tau(x(\cdot))$ is in $\mathfrak{D}_T$ for any $T < \infty$. Thus, our alternative estimators can involve stopping times. This is essential in sequential decision problems, since there the cost function involves first entrance times of a function of $y(\cdot)$ into a decision set.

$\mathfrak{D}$ and $\mathfrak{D}_t$ do not contain 'wild' functions such as those involving differentiation. We consider $\mathfrak{D}$ and $\mathfrak{D}_t$ as a class of *data processors*. It seems to contain a large enough class for practical applications when the corrupting noise is 'white'.

We now state the 'model' 'robustness' or 'approximation' result. For a function $q(z)$, we write (P_t^ϵ, q) for the integral of $q(z)$ with respect to the *Gaussian distribution* with mean $\hat{z}^\epsilon(t)$ and covariance $\Sigma(t)$ - the *ersatz conditional measure* of $z^\epsilon(\cdot)$.

The theorem states that (for a small ϵ) the ersatz conditional distribution is 'nearly optimal' with respect to a specific (but broad) class of alternative estimators. The alternative class includes those that make sense to use when the corrupting noise is white. If the noise is wide band, then it migh: not make sense to exploit its detailed structure and use other 'better' estimators. Doing so might, in practical cases, cause processing errors and other (unmodelled) noise effects.

Theorem 3.1. *Assume the conditions on* $z^\epsilon(\cdot)$, $W_y^\epsilon(\cdot)$ *of Section 2. Then* $(\hat{z}^\epsilon(\cdot),$ $z^\epsilon(\cdot), W_y^\epsilon(\cdot)) \Rightarrow (\hat{z}(\cdot), z(\cdot), W_y(\cdot))$. *Let* $F(\cdot) \in \mathfrak{D}_t$ *be bounded, and* $q(\cdot)$ *bounded continuous and real valued. Then (the limits all exist)*

(3.1)
$$\lim_\epsilon E[q(z^\epsilon(t)) - F(y^\epsilon(\cdot))]^2$$

$$\geqslant \lim_\epsilon E[q(z^\epsilon(t)) - (P_t^\epsilon, q)]^2.$$

947

Remark. The assertion concerning the weak convergence is necessary, since we need to know that the limit of the cited ϵ-triple represents a true filtering problem. The result would not make sense if only 2 out of the 3 components converged.

Proof. By the weak convergence and the w.p.1 continuity of $F(\cdot)$,

$$\left[q(z^\epsilon(t)),\ F(y^\epsilon(\cdot)),\ (P_t^\epsilon,q)\right] \Rightarrow \left[q(z(t)),\ F(y(\cdot)),\ (P_t,q)\right],$$

where $(P_t,q) = \int q(z)dN(\hat{z}(t),\ \Sigma(t);dz)$, and $N(\hat{z},\Sigma;\cdot)$ is the normal distribution with mean $\hat{z}$ and covariance Σ. Thus, the left and right sides of (3.1) converge to, respectively,

$$(3.2) \qquad E[q(z(t)) - F(y(\cdot))]^2, \quad E\left[q(z(t)) - E[q(z(t))|y(s),\ s \leqslant t]\right]^2 \qquad .$$

Since the conditional expectation is the optimal estimator, the second expression is no greater than the first. This yields the theorem.

$$\text{Q.E.D.}$$

IV. FILTERING THE LARGE TIME PROBLEM *(Large* t, *small* ϵ).

The filtering system often operates over a very long time interval. For the model (2.2), (2.4), or with (2.6), (2.5_{WB}), one would then use the stationary filter. But with the system $y^\epsilon(\cdot)$, $z^\epsilon(\cdot)$, *two limits are involved* since both $t \to \infty$ and $\epsilon \to 0$, and it is important that the results *not depend on how* $t \to \infty$ and $\epsilon \to 0$, and that the use of the stationary limit filter is justified. We make some additional assumptions.

C4.1. A_s *is stable.* (A_s,H_s) *is observable and* (A_s,B_s) *controllable.*

C4.2. $\xi_y^\epsilon(t)$ *takes the form* $\xi_y^\epsilon(t) = \xi_y(t/_\epsilon 2)/_\epsilon$, *where* $\xi_y(\cdot)$ *is a second order stationary process with integrable covariance function* $R(\cdot)$. *Also, if* $t_\epsilon \to \infty$ *as* $\epsilon \to 0$, *then* $W_y^\epsilon(t_\epsilon + \cdot) - W_y^\epsilon(t_\epsilon) \Rightarrow W_y(\cdot$

Remark. The model (C4.2) is a common way of modelling wide bandwidth noise, and is used to simplify a calculation below, and to avoid the details involved with other models. It can be extended in many ways. We also make the rather unrestrictive assumption that the initial time is not important and that the $z^\epsilon(\cdot)$ processes do not explode:

C4.3. *If* $\{z^\epsilon(t_\epsilon)\}$ *converges weakly to a random variable* $z(0)$ *as* $\epsilon \to 0$, *then* $z^\epsilon(t_\epsilon + \cdot) \Rightarrow z(\cdot)$ *with initial condition* $z(0)$. *Also*

$$\sup_{\epsilon,t}\ E|z^\epsilon(t)|^2 < \infty.$$

Consistency. In order that $\hat{z}(\cdot)$, $\Sigma(\cdot)$, be a filter for $z(\cdot)$, $y(\cdot)$, it is necessary that the *initial conditions* be *consistent.* Let $N(\hat{z},\Sigma;A)$ denote the probability that the normal random variable (with mean $\hat{z}$, and covariance Σ) takes values in the set A. By *consistency*, we mean that $P\{z(0) \in A | \hat{z}(0), \Sigma(0)\} = N(\hat{z}(0), \Sigma(0);A)$. One cannot choose the initial (random) conditions arbitrarily. It should be obvious that if $\Sigma(0) = \bar{\Sigma}$ and $(z(0), \hat{z}(0))$ are the *stationary* random variables for (stable) (2.2) and (2.5), then the initial conditions are consistent.

The question of consistency arises because when we study the asymptotics as $t \to \infty$ and $\epsilon \to 0$, we will start the filter at some large t_ϵ and do not know a-priori what the *limits* of $(\hat{z}^\epsilon(t), z^\epsilon(t))$ are. The initial condition of the limit equations must be consistent for the problem to make sense. Fortunately, they will be consistent.

Theorem 4.1. *Assume the conditions of Section 2 and (C4.1) - (C4.3). Let $q(\cdot)$ be bounded and continuous and let $F(\cdot) \in \mathfrak{D}_t$. Define $y^\epsilon(s) = 0$, for $s \leqslant 0$ and define $y^\epsilon(-\infty,t,\cdot)$ to be the 'reversed' function - with values $(0 \leqslant \tau < \infty)$ $y^\epsilon(-\infty,t;\tau) = y^\epsilon(t-\tau)$. Then, if $t_\epsilon \to \infty$ as $\epsilon \to 0$,*

$$(4.1) \qquad \{z^\epsilon(t_\epsilon + \cdot), \hat{z}^\epsilon(t_\epsilon + \cdot), W_y^\epsilon(t_\epsilon + \cdot) - W_y^\epsilon(t_\epsilon)\} \to$$

$$(z(\cdot), \hat{z}(\cdot), W_y(\cdot))$$

satisfying (2.3), (2.5), and $z(\cdot)$, $\hat{z}(\cdot)$ are stationary. Also (3.1) holds in the form

$$(4.2) \qquad \lim_{\epsilon,t} E [q(z^\epsilon(t)) - F(y^\epsilon(-\infty,t;\cdot))]^2$$

$$\geqslant \lim_{\epsilon,t} E[q(z^\epsilon(t)) - (P_t^\epsilon,q)]^2 .$$

The limit of (P_t^ϵ,q) is the expectation with respect to the stationary $(\hat{z}(\cdot), \bar{\Sigma})$ system.

Proof. Suppose that $\{\hat{z}^\epsilon(t), \epsilon > 0, t < \infty\}$ is tight. Then, by the hypothesis, $\{\hat{z}^\epsilon(t), z^\epsilon(t), \epsilon > 0, t < \infty\}$ is tight and each subsequence of $\{z^\epsilon(t_\epsilon+\cdot), \hat{z}^\epsilon(t_\epsilon+\cdot), W_y^\epsilon(t_\epsilon+\cdot) - W_y^\epsilon(t_\epsilon), t_\epsilon < \infty, \epsilon > 0\}$ has a weakly convergent subsequence with limit satisfying (2.2), (2.5). Choose a weakly convergent subsequence (with $t_\epsilon \to \infty$) also indexed by ϵ and with limit denoted by $z(\cdot), \hat{z}(\cdot), W_y(\cdot)$. Suppose, for the moment, that $z(\cdot), \hat{z}(\cdot)$ is stationary. (Clearly, $\Sigma(t) \to \bar{\Sigma}$ as $t \to \infty$.) If all limits are stationary, then the subsequence is irrelevant since the stationary solution is unique. Also, since the initial conditions of $\hat{z}(\cdot)$ and $z(\cdot)$ are consistent (owing to the stationarity), $(\hat{z}(\cdot), \bar{\Sigma})$ is the optimal filter for $y(\cdot)$, $z(\cdot)$. Inequality (4.2) is a consequence of this and the weak convergence.

949

We next prove tightness of $(\hat{z}^\epsilon(t), \epsilon > 0, t < \infty)$, and then the stationarity will be proved. We have

(4.3)
$$\dot{\hat{z}}^\epsilon = [A_s - Q(t)H_s]\hat{z}^\epsilon + Q(t)\zeta(t/\epsilon^2)/\epsilon + Q(t)H_s z^\epsilon(t) .$$

Let $\Phi(t,\tau)$ denote the fundamental matrix for $[A_s - Q(t)H_s]$. There are $K < \infty$, $\lambda > 0$ such that $|\Phi(t,\tau)| \leq K \exp - \lambda(t-\tau)$. We have

$$\hat{z}^\epsilon(t) = \Phi(t,0)z^\epsilon(t) + \int_0^t \Phi(t,\tau) Q(\tau)\zeta(\tau/_\epsilon 2)d\tau/\epsilon$$

$$+ \int_0^t \Phi(t,\tau)Q(\tau)Hz^\epsilon(\tau)d\tau .$$

A straightforward calculation using (C4.2 - C4.3) and the change of variable $\tau/\epsilon^2 \to \tau$ in the first integral yields

$$E |\hat{z}^\epsilon(t)|^2 \leq \text{constant } (1 + E|z^\epsilon(0)|^2) ,$$

giving the desired tightness.

To prove the stationarity of the limit of any weakly convergent subsequence, we need only show stationarity of the limit values $(z(0), \hat{z}(0))$ of the $(z^\epsilon(t_\epsilon), \hat{z}^\epsilon(t_\epsilon))$. For this, we use a 'shifting' argument.
Fix $T > 0$ and take a weakly convergent subsequence of (indexed also by ϵ,

and with $t_\epsilon \overset{\epsilon}{\to} \infty$)

$$(z^\epsilon(t_\epsilon+\cdot), \quad \hat{z}^\epsilon(t_\epsilon+\cdot), \quad W_y^\epsilon(t_\epsilon+\cdot) - W_y^\epsilon(t_\epsilon), \quad z^\epsilon(t_\epsilon-T+\cdot), \quad z^\epsilon(t_\epsilon-T+\cdot),$$

$$W_y^\epsilon(t_\epsilon-T+\cdot) - W_y^\epsilon(t_\epsilon-T))$$

with limit denoted by $(z(\cdot), \hat{z}(\cdot), W_y(\cdot), z_T(\cdot), \hat{z}_T(\cdot), W_{y,T}(\cdot))$. We have $\hat{z}_T(T) = \hat{z}(0)$ and $z_T(T) = z(0)$. We do not yet know what $\hat{z}_T(0)$ or $z_T(0)$ are - but, *uniformly in* T, they belong to a tight set, owing to the tightness of $(\hat{z}^\epsilon(t), \epsilon > 0, t < \infty)$. Write (where $W_{s,T}(\cdot)$ 'drives' the equation for dz_T)

$$z(0) = z_T(T) = (\exp A_s T)z_T(0) + \int_0^T \exp A_s(T-\tau) \cdot B_s dW_{s,T}(\tau)$$

950

$$\hat{z}(0) = \hat{z}_T(T) = (\exp [A_s - Q(\bullet)H_s]T)\hat{z}_T(0)$$

$$+ \int_0^T \exp [A_s - Q(\bullet)H_s](T-\tau) \cdot (dW_{y,T}(\tau) + H_s z_T(\tau)d\tau)$$

Since T is arbitrary and the set of all possible $\{z_T(0)\}$ is tight, the stability of A_s and $(A_s - Q(\bullet)H_s)$ implies that $z(0)$ is the stationary random variable, hence $z(\cdot)$ is stationary. Similarly, the pair $(z(\cdot), \hat{z}(\cdot))$ is stationary.

<div align="right">Q.E.D.</div>

REFERENCES

[1] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, M.I.T. Press, Cambridge, U.S.A., 1984.

[2] H.J. Kushner and Hai Huang, "Approximate and Limit Results for Nonlinear Filters with Wide Bandwidth Observation Noise", Stochastics, Feb., 1986.

[3] P. Billingsley, *Convergence of Probability Measures*, 1968, Wiley, New York.

[4] T.G. Kurtz, *Approximation of Population Processes*, 1981, Vol. 36 in CBMS-NSF Regional Conf. Series in Appl. Math, Soc. for Ind. and Appl. Math, Phila.

[5] A. Benveniste, "Design of Monostep and Multistep Adaptive Algorithms for the Tracking of Time Varying Systems," Proc., 23 Conf. on Dec. and Control, 1984, IEEE Publications, New York.

[6] M. El-Ansary and H. Khalil, "On the Interplay of Singular Perturbations and Wide-band Stochastic Fluctuations", SIAM J. on Control, *24*, 1986, 83-98.

[7] H.J. Kushner and Hai Huang, "Averaging Methods for the Asymptotic Analysis of Learning and Adaptive Systems with Small Adjustment Rate", SIAM J. on Control and Optim., *19*, (1981), 635-650.

[8] H. Kushner, "Jump Diffusion Approximations for Ordinary Differential Equations with Wideband Random Right Hand Sides", SIAM J. on Control and Optimization, *17*, 1979, 729-744.

[9] H. J. Kushner and W. Runggaldier, "Filtering and Control for Wide Bandwidth Noise and 'Nearly' Linear Systems", LCDS Rept. #86-8, 1986, Brown Univ.; to appear in IEEE Trans. on Aut. Control.

[10] H. Kushner and W. Runggaldier, "Nearly Optimal State Feedback Controls for Stochastic Systems with Wideband Noise Disturbances", to appear SIAM J. on Control and Optimization. Also, LCDS Rept. #85-23, 1985, Brown Univ.

[11] A.V. Skorohod, "Limit Theorems for Stochastic Processes", Theory of Probability and Its Applications, *1*, 1956, 262-290.

[12] H.J. Kushner, "Diffusion Approximations to Output Processes of Nonlinear Systems with Wide-band Inputs, and Applications", IEEE Trans. on Inf. Theory, *IT-26*, 1980, 715-725.

[13] G.B. Blankenship and G.C. Papanicolaou, "Stability and Control of Stochastic Systems with Wide Band Noise Disturbances", SIAM J. Appl. Math *34*, 1978, 437-476.

Adaptive Kalman Filtering for Instrumentation Radar

Charles K. Chui

Department of Mathematics
Texas A & M University
College Station, Texas 77843

Robert E. Green

Instrumentation Directorate
White Sands Missile Range
New Mexico 88002

ABSTRACT

The optimization criterion of the adaptive Kalman filter for instrumentation tracking radars is slightly modified to change nonlinear matrix equations to linear matrix operations, and the resulting equations are implemented with parallel processing for efficient real-time applications.

1. The tracking radar

We only consider monopulse tracking radars. The pencil-beam antenna of a monopulse tracking radar consists of a reflector and a cluster of four feed horns from which four single-pulses are transmitted simultaneously. In tracking a target, two of the echo signals are used for determining the magnitude and direction of its azmuthal angular position, the other two for determining the magnitude and direction of its elevation angular position, and the four together are used to determine the range of the target (see [2] and [7] for references). A transmitted signal is described graphically by a lobe as shown in Fig.1. If a target happens to be located along the beam axis, the voltage response measured from the echo signal at the tracking-radar receiver will be of highest value since the radiated power is concentrated in the direction of the beam axis. If the target is not detected along the beam axis , the resulting voltage response will be somewhat smaller.
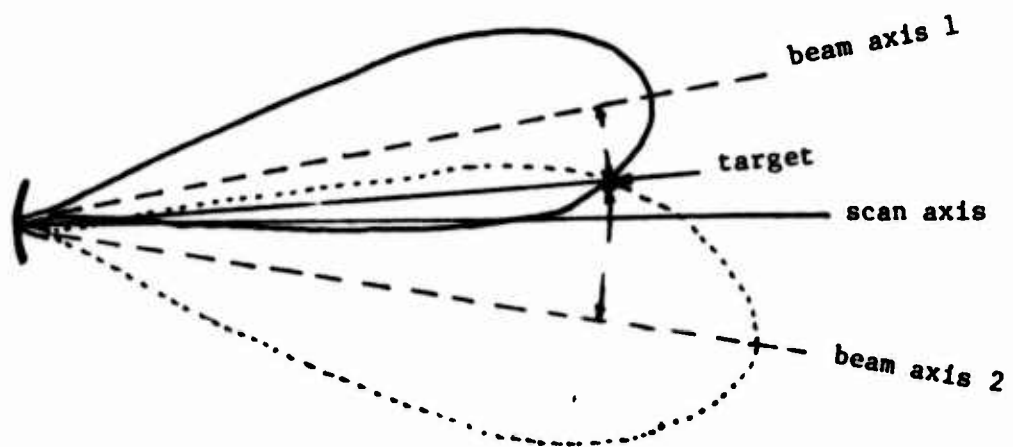
Fig.1.

A polar representation of the pencil-beam from two horns which are used to determine the azimuthal or elevation angular position of the target is shown in Fig.1, and is translated to the rectangular coordinates with voltage as angular measurement in Fig.2.
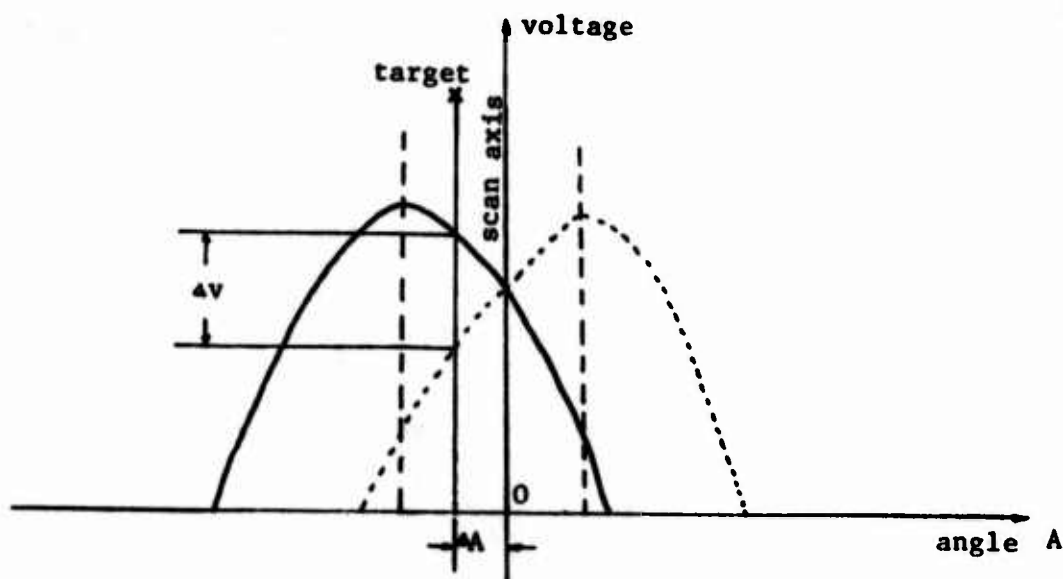
Fig.2.

The information on the difference Δv in amplitude between the voltage responses at the two positions of two beam-lobes, usually called the error signal, yields a measurement of the angular displacement (or angular error) of the target from the scan

axis. Hence, both azimuthal and elevation angular positions are observed. In addition, the tracking radar is designed so that the sum of the echo signals provides the range measurement of the displacement of the target. In the monopulse tracking radar we are discussing, the echo signals are combined so that both the sum and the difference signals are obtained simultaneously. Hence, the range Σ, the azimuthal angular error ΔA and the elevational angular error ΔE are all obtained simultaneously. Because the azimuth angle at the k th instance is $A_k = a_k + \Delta A_k$, where a_k is the horizontal angle of the scan axis of the antenna measured from some reference axis (cf. Fig.3) and ΔA_k is the value of ΔA at the k th instance, A_k can be determined immediately. Similarly, $E_k = e_k + \Delta E_k$, where e_k is the vertical angle of the scan axis of the antenna measured from the same reference axis (cf. Fig.3), is also obtained.

Fig.3

956
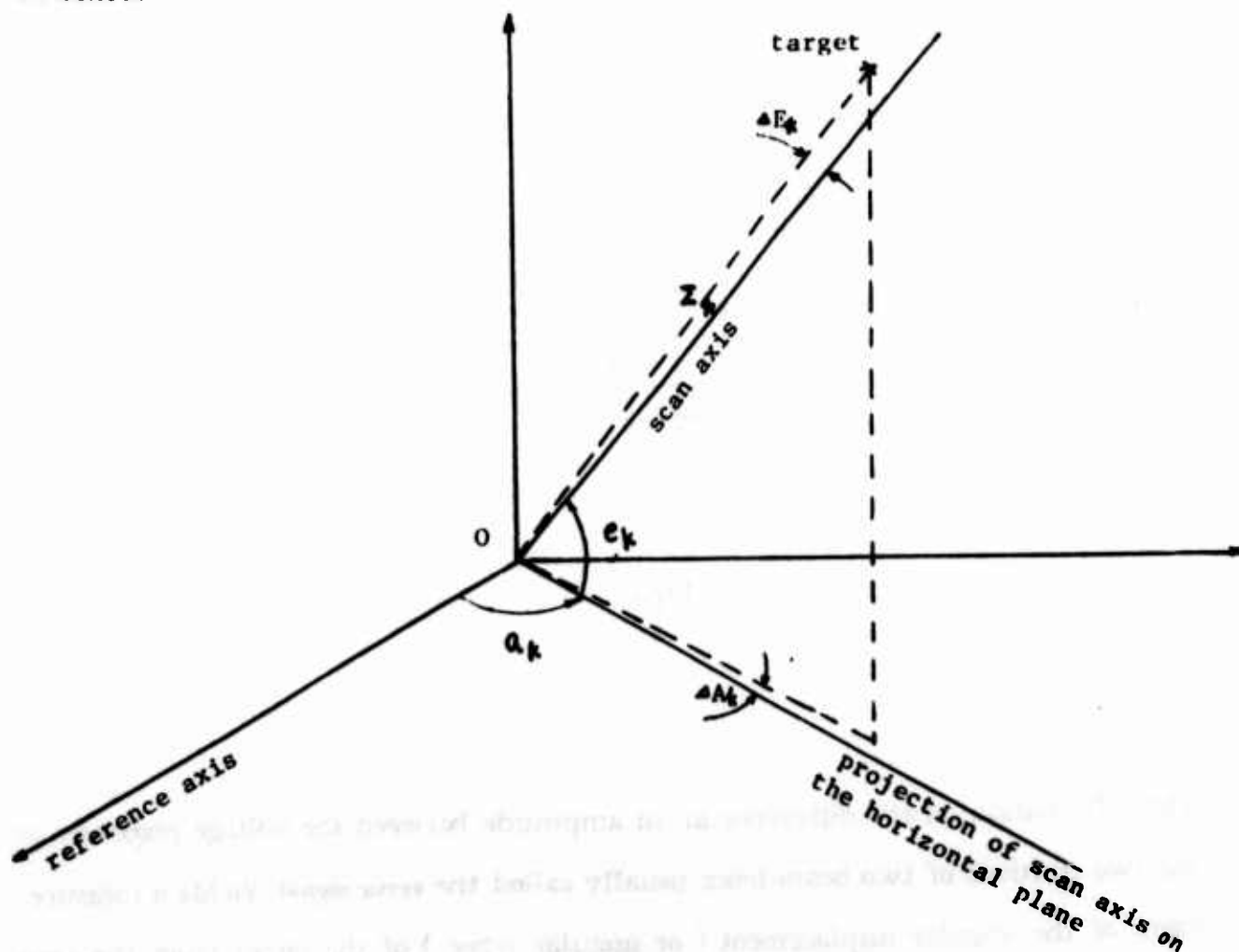
The principle of generating the error signal and determining the azimuth angle A, the elevation angle E, and the range Σ by the monopulse tracking techniques we discuss here already solves the tracking problem in an enviroment without any interference or noise. In practice, however, there are different sources of noise such as the solar or galactic noise, the ground noise from the environment of the radar and the electronic or mechanical equipment in the radar itself. Hence, the data $\bar{\Sigma}$, $\Delta\bar{A}$ and $\Delta\bar{E}$ we obtain from the tracking radar system are associated with random errors which must be filtered out in order to be able to determine the real values of Σ, ΔA and ΔE. This real-time problem is usually tackled by applying Kalman filtering.

2. Kalman Filtering

A general linear mathematical model for the control-observation system is given by

$$\begin{cases} y_{k+1} = A_k y_k + B_k u_k + \Gamma_k \xi_k , & y_0 = E(y_0) \\ w_k = C_k y_k + D_k u_k + \eta_k , \end{cases} \tag{1}$$

where, for each $k = 0, 1, \ldots, A_k, B_k, C_k, D_k$, and Γ_k are known constant matrices, $E(y_0)$ is given, $\{u_k\}$ is a sequence of predesigned control functions, and $\{\xi_k\}$ and $\{\eta_k\}$ are white noise processes; that is, $E(\xi_k) = 0$, $E(\xi_k \xi_j^T) = Q_k \delta_{kj}$, $E(\eta_k) = 0$, $E(\eta_k \eta_j^T) = R_k \delta_{kj}$, $E(\xi_k \eta_j^T) = 0$, for $k, j = 0, 1, 2, \ldots$. Here, as usual,

$$\delta_{kj} = \begin{cases} 1 & k = j \\ 0 & k \neq j . \end{cases}$$

As is well known, this system can be uncoupled into two systems

$$\begin{cases} x_{k+1} = A_k x_k + \Gamma_k \xi_k , & x_0 = y_0 \\ v_k = C_k x_k + \eta_k , \end{cases}$$

and

$$\begin{cases} z_{k+1} = A_k z_k + B_k u_k , & z_0 = 0 \\ \tilde{v}_k = C_k z_k + D_k u_k , \end{cases}$$

where $y_k = x_k + z_k$ and $w_k = v_k + \tilde{v}_k$. Now, the state vector z_k can be computed using the formula

$$z_k = \sum_{i=1}^{k} (A_{k-1} \ldots A_i) B_{i-1} u_{i-1}$$

and the observation vector v_k becomes

$$v_k = w_k - C_k z_k - D_k u_k .$$

This information is used to estimate x_k. That is, we use it in studying the stochastic state-space decomposition :

$$\begin{cases} x_{k+1} = A_k x_k + \Gamma_k \xi_k \ , \quad x_o = y_o \\ v_k = C_k x_k + \eta_k \ . \end{cases} \tag{2}$$

Of course, if $\hat{x}_k$ is the optimal estimation of x_k, then $\hat{y}_k = \hat{x}_k + z_k$ is the optimal estimate of y_k in the original system (1).

The stochastic linear system for radar tracking can be described as follows. Let Σ, ΔA, ΔE be the range, the azimuthal angular error, and the elevational angular error, respectively, of the target, with the radar being located at the origin (cf. Fig.3), and consider Σ, ΔA, and ΔE as functions of time with first and second derivatives denoted by $\dot{\Sigma}$, $\Delta \dot{A}$, $\Delta \dot{E}$, $\ddot{\Sigma}$, $\Delta \ddot{A}$, $\Delta \ddot{E}$, respectively. Let $h > 0$ be the sampling time unit and set $\Sigma_k = \Sigma(kh)$, $\dot{\Sigma}_k = \dot{\Sigma}(kh)$, $\ddot{\Sigma}_k = \ddot{\Sigma}(kh)$, etc. Then, using the second degree Taylor polynomial approximation, the radar tracking model takes on the following linear stochastic state-space description:

$$\begin{cases} x_{k+1} = A x_k + \Gamma_k \xi_k \\ v_k = C x_k + \eta_k \end{cases} \tag{3}$$

where

$$x_k = [\ \Sigma_k \ \dot{\Sigma}_k \ \ddot{\Sigma}_k \ \Delta A_k \ \Delta \dot{A}_k \ \Delta \ddot{A}_k \ \Delta E_k \ \Delta \dot{E}_k \ \Delta \ddot{E}_k \ ]^T \ ,$$

$$A = \begin{vmatrix} 1 & h & m & & & & & & \\ 0 & 1 & h & & & & & & \\ 0 & 0 & 1 & & & & & & \\ & & & 1 & h & m & & & \\ & & & 0 & 1 & h & & & \\ & & & 0 & 0 & 1 & & & \\ & & & & & & 1 & h & m \\ & & & & & & 0 & 1 & h \\ & & & & & & 0 & 0 & 1 \end{vmatrix}, \ m = h^2/2,$$

$$C = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{vmatrix},$$

$$v_k = C x_k + \eta_k = \begin{vmatrix} \Sigma_k \\ \Delta A_k \\ \Delta E_k \end{vmatrix} + \eta_k .$$

and $\{\xi_k\}$ and $\{\eta_k\}$ are white noise processes. This is a time-invariant model of (2).

Observe that if we let

$$x_k = \begin{vmatrix} x_k^1 \\ x_k^2 \\ x_k^3 \end{vmatrix}, \quad x_k^1 = \begin{vmatrix} \Sigma_k \\ \dot{\Sigma}_k \\ \ddot{\Sigma}_k \end{vmatrix}, \quad x_k^2 = \begin{vmatrix} \Delta A_k \\ \Delta \dot{A}_k \\ \Delta \ddot{A}_k \end{vmatrix}, \quad x_k^3 = \begin{vmatrix} \Delta E_k \\ \Delta \dot{E}_k \\ \Delta \ddot{E}_k \end{vmatrix},$$

$$\xi_k = \begin{vmatrix} \xi_k^1 \\ \xi_k^2 \\ \xi_k^3 \end{vmatrix}, \quad \eta_k = \begin{vmatrix} \eta_k^1 \\ \eta_k^2 \\ \eta_k^3 \end{vmatrix}, \quad v_k = \begin{vmatrix} v_k^1 \\ v_k^2 \\ v_k^3 \end{vmatrix},$$

$$\tilde{A} = \begin{vmatrix} 1 & h & m \\ 0 & 1 & h \\ 0 & 0 & 1 \end{vmatrix}, \quad \tilde{C} = [1\ 0\ 0],$$

and assume that

$$\Gamma_k = \begin{vmatrix} \Gamma_k^1 & & \\ & \Gamma_k^2 & \\ & & \Gamma_k^3 \end{vmatrix}, \quad Q_k = \begin{vmatrix} Q_k^1 & & \\ & Q_k^2 & \\ & & Q_k^3 \end{vmatrix}, \quad R_k = \begin{vmatrix} R_k^1 & & \\ & R_k^2 & \\ & & R_k^3 \end{vmatrix},$$

where Γ_k^i are 3×3 submatrices, Q_k^i are 3×3 nonnegative definite symmetric submatrices, and R_k^i are 3×3 positive definite symmetric submatrices, for $i = 1, 2, 3$, then, system (3) can be split into three subsystems:

$$\begin{cases} x_{k+1}^i = \tilde{A} x_k^i + \Gamma_k^i \xi_k^i \\ v_k^i = \tilde{C} x_k^i + \eta_k^i, \end{cases} \quad i = 1, 2, 3.$$

Hence, for our radar tracking problem, it is sufficient to study the following system:

$$\begin{cases} x_{k+1} = A x_k + \Gamma_k \xi_k \\ v_k = C x_k + \eta_k \end{cases} \tag{4}$$

where, for each k, x_k and ξ_k are 3-vectors, v_k and η_k are scalars,

$$A = \begin{vmatrix} 1 & h & m \\ 0 & 1 & h \\ 0 & 0 & 1 \end{vmatrix}, \; m = h^2/2,$$

$$C = [1 \; 0 \; 0],$$

and R_k is a scalar.

Now, suppose that the appropriate statistical properties of the noise sequences $\{\xi_k\}$, $\{\eta_k\}$ and the initial state x_0 are known. More precisely, let

$$E(\xi_k) = 0, \; E(\eta_k) = 0,$$

$$E(\xi_k \xi_j^T) = \Gamma_k Q_k \Gamma_k^T \delta_{kj}, \; E(\eta_k \eta_j^T) = R_k \delta_{kj}, \; E(\xi_k \eta_j^T) = 0,$$

$$E(x_0 \xi_k^T) = 0, \; E(x_0 \eta_k^T) = 0,$$

where Q_k and R_k as well as $E(x_0)$ and $Var(x_0)$ are given. Then the Kalman filter can be described by the following recursive formulae (cf., for example, Anderson and Moore [1] or Chui and Chen [3]):

$$\begin{cases} \hat{x}_{k/k} = \hat{x}_{k/(k-1)} + G_k(v_k - C\hat{x}_{k/(k-1)}) \\ \hat{x}_{0/0} = E(x_0), \end{cases} \tag{5.1}$$

$$\hat{x}_{k/(k-1)} = A\hat{x}_{(k-1)/(k-1)}, \tag{5.2}$$

$$G_k = P_{k,k-1}C^T(CP_{k,k-1}C^T + R_k)^{-1}, \tag{5.3}$$

$$P_{k,k-1} = AP_{k-1,k-1}A^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T, \tag{5.4}$$

$$\begin{cases} P_{k,k} = P_{k,k-1} - P_{k,k-1}C^T(CP_{k,k-1}C^T + R_k)^{-1}CP_{k,k-1}, \\ P_{0,0} = Var(x_0). \end{cases} \tag{5.5}$$

In tracking the azimuthal and elevational angles, that is, when

$$\hat{x}_{k/k} = \begin{vmatrix} \Delta A_k \\ \Delta \dot{A}_k \\ \Delta \ddot{A}_k \end{vmatrix} \quad \text{and} \quad \hat{x}_{k/k} = \begin{vmatrix} \Delta E_k \\ \Delta \dot{E}_k \\ \Delta \ddot{E}_k \end{vmatrix}, \qquad 961$$

respectively, are obtained, it is advisable to use the filtered outputs

$$A_k = a_k + \Delta A_k$$

and

$$E_k = e_k + \Delta E_k$$

to give meaningful tracking images. Of course, a_k and e_k are simply the horizontal and vertical angles of the scan axis as shown in Fig.3.

3. Adaptive Kalman Filtering

The statistical properties of the noise sequences $\{\xi_k\}$, $\{\eta_k\}$ and the initial state x_0 in the Kalman filter discussed above are assumed to be given before the process is performed. In practice, however, these statistical properties are usually unknown and even unpredictable. Hence, adaptive filtering is essential. This means that we must estimate the statistical properties at each stage so that Kalman filtering can be performed using these estimates. In this note, we assume that the initial statistical properties $E(x_0)$, $Var(x_0)$, Q_0, and R_0 of the state and noise sequences are known so that the filtering process can get started. The adaptive filter associated with system (3) can be described by

$$\begin{cases} \hat{x}_k = A \hat{x}_{k-1} + \hat{G}_k (v_k - CA \hat{x}_{k-1}) \\ \hat{x}_0 = E(x_0), \end{cases} \tag{6}$$

where $\hat{G}_k$ is a real-time estimate of the gain matrix G_k at the kth instance, in the sense that $\hat{x}_k = \hat{x}_k(\hat{G}_k)$ satisfies

$$tr \|\hat{x}_k(\hat{G}_k) - x_k\|^2 = \min_G tr \|\hat{x}_k(G) - x_k\|^2 , \tag{7}$$

where, for any 3-dimensional random vector z, $\|z\|^2 = <z,z> = Var(z)$. However, the computation of $\hat{x}_k$ defined in (7) is extremely complicated and is not suitable for real-time problems (cf. Chui and Chen [3]). Instead, we will adopt the following slightly weaker optimality criterion:

$$tr \|\hat{x}_k(\hat{G}_k) - x_{k-3}\|^2 = \min_G tr \|\hat{x}_k(G) - x_{k-3}\|^2 . \tag{8}$$

It should be remarked that estimation of x_{k-1} by $\hat{x}_k$ has been studied in a different situation. For instance, estimation of x_{k-1} by $\hat{x}_k$ was done in Jazwinski [4] in non-adaptive Kalman filtering with colored input.

To see that (8) can be used instead of (7) without too much loss in the order of

963

estimation, we have the following

LEMMA 1 The error variance $\|\hat{x}_k - x_k\|^2$ is equivalent to $\|\hat{x}_k - x_{k-3}\|^2$ in the sense that

$$\frac{1}{2}\|\hat{x}_k - x_k\|^2 - C_k \leq \|\hat{x}_k - x_{k-3}\|^2 \leq 2\|\hat{x}_k - x_k\|^2 + 2C_k ,$$

where the constant $C_k = \|x_k - x_{k-3}\|^2$ depends only on k.

Next, using the notation

$$N_{CA} = \begin{vmatrix} C \\ CA \\ CA^2 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & h & m \\ 1 & 2h & 4m \end{vmatrix}, \quad m = h^2/2, \quad \text{and} \quad \mathbf{v}_{k-1,k-3} = \begin{vmatrix} v_{k-3} \\ v_{k-2} \\ v_{k-1} \end{vmatrix},$$

we have the following

LEMMA 2

$$\|\hat{G}_k (v_k - CA\hat{x}_{k-1}) - (N_{CA}^{-1}\mathbf{v}_{k-1,k-3} - A\hat{x}_{k-1})\|^2 = \|\hat{x}_k - x_{k-3}\|^2 + D_k ,$$

where D_k is a constant symmetric matrix depending only on k .

In view of Lemmas 1 and 2, instead of using the criterion (7), we will consider the minimization problem:

$$\min_{G_k} tr \ \| G_k (v_k - CA\hat{x}_{k-1}) - (N_{CA}^{-1}\mathbf{v}_{k-1,k-3} - A\hat{x}_{k-1}) \|^2 . \tag{9}$$

which is equivalent to (8). Under this criterion, we have the following result.

THEOREM 1 Let $\hat{G}_k$ be a solution of the minimization problem (9). Then $\hat{G}_k$ is uniquely determined by

$$\hat{G}_k = [Var(v_k - CA\hat{x}_{k-1})]^{-1} N_{CA}^{-1} E[(\mathbf{v}_{k-1,k-3} - N_{CA} A\hat{x}_{k-1})(v_k - CA\hat{x}_{k-1})] . \tag{10}$$

To give a recursive algorithm for computing $\hat{G}_k$, the following lemma is necessary.

964

LEMMA 3 The estimate $\hat{G}_k$ of G_k in (10) can be rewritten as

$$\hat{G}_k = \frac{1}{CP_{k,k-1}C^T + \hat{R}_{k-1}}[J_{k-3,k} + N_{CA}^{-1}M_{k-3,k}]C^T ,$$

where

$$J_{k-3,k} = E((\mathbf{x}_{k-3} - A\hat{\mathbf{x}}_{k-1})(\mathbf{x}_k - A\hat{\mathbf{x}}_{k-1})^T) ,$$

$$M_{k-3,k} = E\left(\begin{vmatrix} \eta_{k-3} \\ C\Gamma_{k-3}\xi_{k-3} + \eta_{k-2} \\ CA\Gamma_{k-3}\xi_{k-3} + C\Gamma_{k-2}\xi_{k-2} + \eta_{k-1} \end{vmatrix}(\mathbf{x}_k - A\hat{\mathbf{x}}_{k-1})^T\right) ,$$

and

$$P_{k,k-1} = \|\mathbf{x}_k - A\hat{\mathbf{x}}_{k-1}\|^2 .$$

Using this lemma, we can derive the following recursive computational scheme :

$$\begin{cases} \hat{G}_k = \dfrac{1}{h_{k-1}}[J_{k-3,k} + N_{CA}^{-1}M_{k-3,k}]C^T , \\[3mm] \hat{G}_0 = \dfrac{Var(\mathbf{x}_0)C^T}{C\,Var(\mathbf{x}_0)C^T + \hat{R}_0} , \end{cases}$$

where

$$\begin{cases} P_{k,k-1} = F_{k-1}P_{k-1,k-2}F_{k-1}^T + \Gamma_{k-1}\hat{Q}_{k-1}\Gamma_{k-1}^T + A\hat{G}_{k-1}\hat{R}_{k-1}\hat{G}_{k-1}^T A^T \\[3mm] P_{1,0} = A\,Var(\mathbf{x}_0)A^T + \Gamma_0\hat{Q}_0\Gamma_0^T , \end{cases}$$

$$J_{k-3,k} = AJ_{k-4,k-1}F_{k-1}^T + A\hat{G}_{k-1}[h_{k-1}\hat{G}_{k-1}^T - CP_{k-1,k-2}]A^T + \Gamma_{k-4}\hat{Q}_{k-4}\Gamma_{k-4}^T F_{k-3}^T F_{k-2}^T F_{k-1}^T ,$$

$$k \geqslant 4 ,$$

$$J_{k-3,k} = AJ_{k-4,k-1}F_{k-1}^T + A\hat{G}_{k-1}h_{k-1}\hat{G}_{k-1}^T A^T - \hat{G}_{k-1}CP_{k-1,k-2}A^T , \qquad k = 2, 3,$$

$$J_{-2,1} = 0_{3 \times 3} \, .$$

$$M_{-2,1} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_0 \hat{G}_0^T A^T \end{vmatrix} \, ,$$

$$M_{-1,2} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_1 \hat{G}_1^T A^T \end{vmatrix} + \begin{vmatrix} \tilde{M}_{-2,1} \\ \\ C\Gamma_0 \hat{Q}_0 \Gamma_0^T \end{vmatrix} F_{-1}^T \, ,$$

$$M_{k-3,k} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_{k-1} \hat{G}_{k-1}^T A^T \end{vmatrix} +$$

$$\begin{vmatrix} \tilde{M}_{k-4,k-1} \\ [CA\Gamma_{k-3}\hat{Q}_{k-3}\Gamma_{k-3} + C\Gamma_{k-2}\hat{Q}_{k-2}\Gamma_{k-2}^T F_{k-1}^T] F_{k-2}^T \end{vmatrix} F_{k-1}^T \, , \quad k \geqslant 3 \, ,$$

with

$$\tilde{M} = \begin{vmatrix} m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{vmatrix} \, ,$$

$$F_j = A(I - \hat{G}_j C) \, , \quad j = 0, 1, \ldots, k \, ,$$

and

$$h_j = CP_{j,j-1}C^T + \hat{R}_{j-1} \, , \quad j = 1, 2, \ldots, k \, .$$

Here $\hat{Q}_k$ and $\hat{R}_k$ are determined by the relation :

$$\Gamma_k \hat{Q}_k \Gamma_k^T = P_{k,k-1} - F_k P_{k,k-1} F_k^T - A\hat{G}_k \hat{R}_k \hat{G}_k^T A^T \, ,$$

and

$$\hat{R}_k = \begin{vmatrix} \dfrac{1}{\hat{G}_k^T \hat{G}_k} \hat{G}_k^T - C \end{vmatrix} P_{k,k-1} C^T \, .$$

966

We remark that we are supposed to know the initial conditions $E(\mathbf{x}_0)$, $Var(\mathbf{x}_0)$, $\hat{Q}_0$ and $\hat{R}_0$ in order to apply this algorithm. However, even if they are unkown, any rough prior estimates of them would do the job. As the process is being performed, we still have a near-optimal adaptive filtering.

4. Systolic Implementation

The near-optimal adaptive Kalman filter for our radar-tracking system is given by

$$\begin{vmatrix} \hat{\mathbf{x}}_t = A\,\hat{\mathbf{x}}_{t-1} + \hat{G}_t\,(v_t - CA\,\hat{\mathbf{x}}_{t-1}) \\ \\ \hat{\mathbf{x}}_0 = E(\mathbf{x}_0) , \end{vmatrix}$$

where the adaptive Kalman gain $\hat{G}_t$ is obtained using the following procedure :

Step 1 : Start with $E(\mathbf{x}_0)$, $Var(\mathbf{x}_0)$, $\hat{Q}_0 = Q_0$, $\hat{R}_0 = R_0$, and

$$N_{CA}^{-1} = \begin{vmatrix} 1 & 0 & 0 \\ -3/2h & 2/h & -1/2h \\ 1/h^2 & -2/h^2 & 1/h^2 \end{vmatrix} .$$

Step 2 : Set $\hat{G}_0 = \dfrac{Var(\mathbf{x}_0)\,C^T}{C\,Var(\mathbf{x}_0)\,C^T + \hat{R}_0}$.

Step 3 : Compute

$$P_{1,0} = A\,Var(\mathbf{x}_0)\,A^T + \Gamma_0\hat{Q}_0\Gamma_0 ,$$

$$J_{-2,1} = 0_{3\times3} ,$$

and

$$M_{-2,1} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_0\hat{G}_0^T A^T \end{vmatrix} .$$

Step 4 : Compute $h_1 = CP_{1,0}C^T + \hat{R}_0$.

Step 5 : Set $\hat{G}_1 = \dfrac{1}{h_1}\,[J_{-2,1} + N_{CA}^{-1}M_{-2,1}]\,C^T$.

968

Step 6 : Compute $F_1 = A(I - \hat{G}_1 C)$.

Step 7 : Compute

$$\hat{R}_1 = \left| \frac{\hat{G}_1^T}{\hat{G}_1^T \hat{G}_1} - C \right| P_{1,0} C^T \ ,$$

and

$$\Gamma_1 \hat{Q}_1 \Gamma_1^T = P_{1,0} - F_1 P_{1,0} F_1^T - A \hat{G}_1 \hat{R}_1 \hat{G}_1^T A^T \ .$$

Step 8 : Compute

$$P_{2,1} = F_1 P_{1,0} F_1^T + \Gamma_1 \hat{Q}_1 \Gamma_1^T + A \hat{G}_1 \hat{R}_1 \hat{G}_1^T A^T \ ,$$

$$J_{-1,2} = A J_{-2,1} F_1^T + A \hat{G}_1 [h_1 \hat{G}_1^T - C P_{1,0}] A^T \ ,$$

and

$$M_{-1,2} = \left| \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_1 \hat{G}_1^T A^T \end{matrix} \right| + \left| \begin{matrix} \tilde{M}_{-2,1} \\ C \Gamma_0 \hat{Q}_0 \Gamma_0^T \end{matrix} \right| F_1^T \ .$$

Step 9 : Compute $h_2 = C P_{2,1} C^T + \hat{R}_1$.

Step 10 : Set $\hat{G}_2 = \frac{1}{h_2} [J_{-1,2} + N_{CA}^{-1} M_{-1,2}] C^T$.

Step 11 : Compute $F_2 = A(I - \hat{G}_2 C)$.

Step 12 : Compute

$$\hat{R}_2 = \left| \frac{\hat{G}_2^T}{\hat{G}_2^T \hat{G}_2} - C \right| P_{2,1} C^T \ ,$$

and

969

$$\Gamma_2 \hat{Q}_2 \Gamma_2^T = P_{2,1} - F_2 P_{2,1} F_2^T - A\hat{G}_2 \hat{R}_2 \hat{G}_2^T A^T .$$

Step 13 : For $k \geqslant 3$, repeat the following.

   (1) Compute

$$P_{k,k-1} = F_{k-1} P_{k-1,k-2} F_{k-1}^T + \Gamma_{k-1} \hat{Q}_{k-1} \Gamma_{k-1}^T + A\hat{G}_{k-1} \hat{R}_{k-1} \hat{G}_{k-1}^T A^T ,$$

$$J_{k-3,k} = AJ_{k-4,k-1} F_{k-1}^T + A\hat{G}_{k-1}[h_{k-1} \hat{G}_{k-1}^T - CP_{k-1,k-2}]A^T + \Gamma_{k-4} \hat{Q}_{k-4} \Gamma_{k-4}^T F_{k-3}^T F_{k-2}^T F_{k-1}^T ,$$

with $\hat{Q}_{-1} = 0$ and

$$M_{k-3,k} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_{k-1} \hat{G}_{k-1}^T A^T \end{vmatrix} +$$

$$\begin{vmatrix} \tilde{M}_{k-4,k-1} \\ [CA^2 \Gamma_{k-3} \hat{Q}_{k-3} \Gamma_{k-3}^T + CA \Gamma_{k-2} \hat{Q}_{k-2} \Gamma_{k-2}^T F_{k-1}^T]F_{k-2}^T \end{vmatrix} F_{k-1}^T .$$

   (2) Compute

$$h_k = CP_{k,k-1} C^T + \hat{R}_{k-1} .$$

   (3) Compute

$$\hat{G}_k = \frac{1}{h_k} [J_{k-3,k} + N_{CA}^{-1} M_{k-3,k}] C^T .$$

   (4) Compute

$$F_k = A(I - \hat{G}_k C) .$$

   (5) Compute

$$\hat{R}_k = \begin{vmatrix} \dfrac{\hat{G}_k^T}{\hat{G}_k^T \hat{G}_k} - C \end{vmatrix} P_{k,k-1} C^T ,$$

970

and

$$\Gamma_k \hat{Q}_k \Gamma_k^I = P_{k,k-1} - F_k P_{k,k-1} F_k^I - A\hat{G}_k \hat{R}_k \hat{G}_k^T A^T .$$

We again remark that when

$$\hat{x}_{k,k} = \begin{vmatrix} \Delta A_k \\ \Delta \dot{A}_k \\ \Delta \ddot{A}_k \end{vmatrix} \quad \text{and} \quad \hat{x}_{k,k} = \begin{vmatrix} \Delta E_k \\ \Delta \dot{E}_k \\ \Delta \ddot{E}_k \end{vmatrix}$$

are considered, we always use the filtered information

$$A_k = a_k + \Delta A_k$$

and

$$E_k = e_k + \Delta E_k \quad ,$$

respectively.

A flow chart is given below.

$$E(\mathbf{x}_0), \quad Var(\mathbf{x}_0), \quad \hat{Q}_0 = Q_0, \quad \hat{R}_0 = R_0, \quad N_{CA}^{-1}$$

$$\hat{G}_0$$

$$k := 1$$

$$P_{k,k-1}, \quad J_{k-3M}, \quad M_{k-3M}$$

$$h_k$$

$$\hat{G}_k$$

$$F_k$$

$$\hat{R}_k$$

$$\hat{Q}_k$$

$$k := k + 1$$

$$\hat{\mathbf{x}}_k = A\hat{\mathbf{x}}_{k-1} + \hat{G}_k(v_k - CA\hat{\mathbf{x}}_{k-1})$$

$$\mathbf{x}_0 = E(\mathbf{x}_0)$$

972
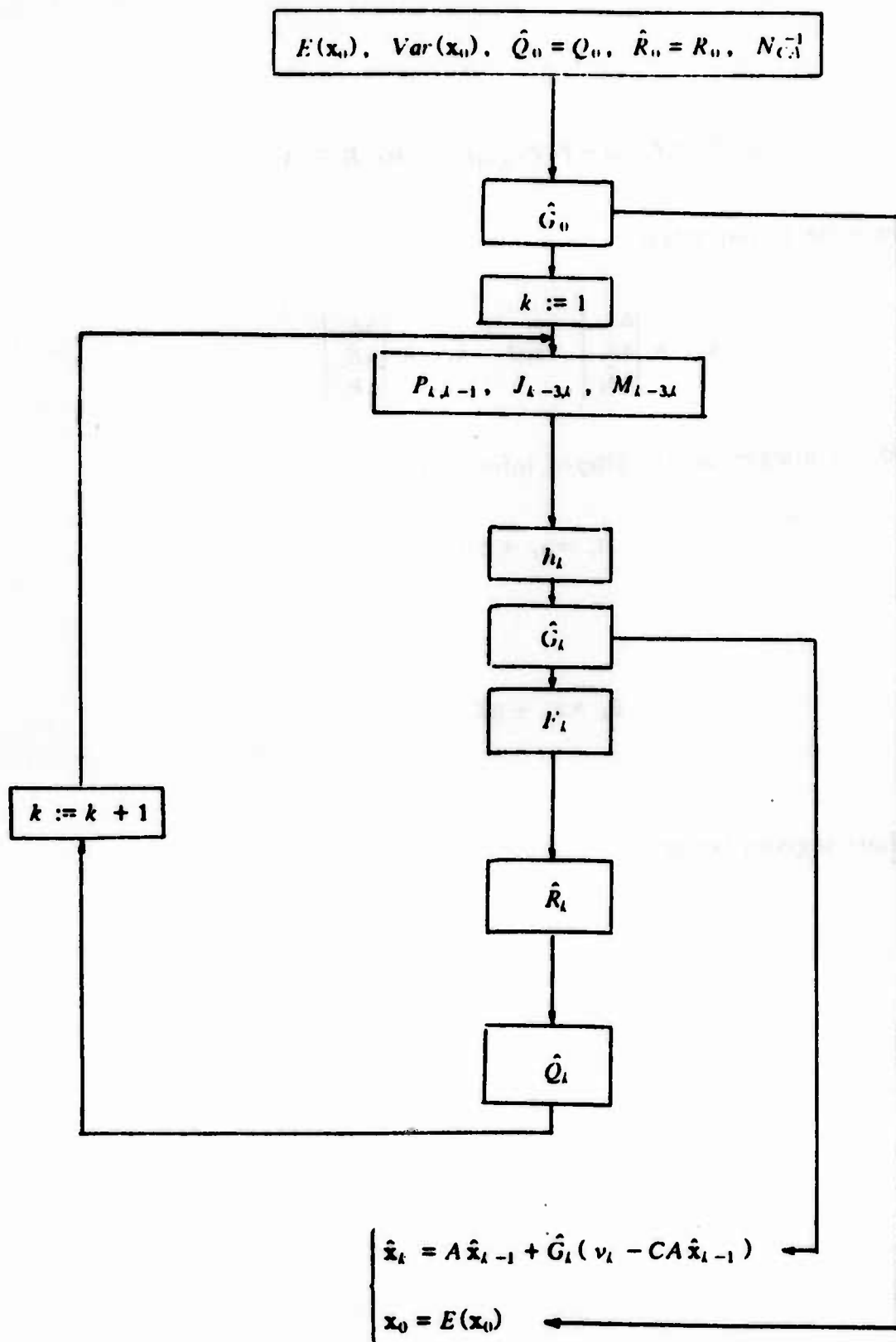
We conclude this note by discussing parallel processing implementation to our adaptive Kalman filtering. Three sets of processors are simultaneously used. The number of operations of matrix-matrix multiplications etc. in each set of processors is listed in the following table where all matrices and vectors are 3-dimensional. Systolic arrays can be used here to perform fast parallel operation (cf. [5,6,8,9]).

Number of Operations

Type of Operations	Set 1	Set 2	Set 3
matrix-matrix multiplication	4	3	4
matrix-vector multiplication	4	2	2
vector-vector multiplication	1	4	0
matrix addition	1	3	3
vector addition	0	1	0
scalar multiplication	1	1	1
scalar addition	3	1	0

$$k = 0$$

$$E(x_0), Var(x_0), \hat{Q}_0, \hat{R}_0, N_{CA}^{-1} = \begin{vmatrix} 1 & 0 & 0 \\ -3/2h & 2/h & -1/2h \\ 1/h^2 & -1/2h & 1/h^2 \end{vmatrix}$$

$$\hat{G}_0 = \frac{Var(x_0) C^T}{C \, Var(x_0) C^T + \hat{R}_0}$$

$k = 1$

$$M_{-2,1} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_{1,1} A \hat{G}_{1,1} y \end{vmatrix}$$

$$J_{-2,1} = O_{3x3}$$

$$T_1 = J_{-2,1} + N^t_{1,1} M_{-2,1}$$

$$P_{1,0} = A \, Var(x_0) A^T + \Gamma_0 \hat{Q}_0 \Gamma^t_0$$

$$s_1 = C P_{1,0} C^T$$

$$h_1 = s_1 + \hat{R}_0$$

$$\hat{G}_1 = \frac{T_1 C^T}{h_1}$$

$$a_1 = A \hat{G}_1$$

$$F_1 = A - a_1 C$$

$$K_1 = F_1 P_{1,0} F^t_1$$

$$g_1 = \hat{G}^t_1 \hat{G}_1$$

$$\hat{R}_1 = g_1 t_1 C^T - s_1$$

$$H_1 = a_1 \hat{R}_1 a^t_1$$

$$\Gamma_1 \hat{Q}_1 \Gamma^t_1 = P_{1,0} - K_1 - H_1$$

$$t_1 = \hat{G}^t_1 P_{1,0}$$

$k \geq 3$

$$M_{i-1} := \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\hat{R}_{i-1} & -a_{i-1}^T \end{vmatrix}_{i-1}$$

$$\hat{M}_{i-1} := \; + \; CA[A\Gamma_{i-1}\hat{Q}_{i-3}\Gamma_{i-1}^T + \Gamma_{i-2}\hat{Q}_{i-2}\Gamma_{i-2}^T \bar{F}_{i-1}^T] \Big|_{i-1}$$

$$J_{i-1} := AJ_{i-1}\bar{F}_{i-1}^T$$

$$J_{i-1} := J_{i-1} + a_{i-1}(h_{i-1}\hat{G}_{i-1} - CP_{i-1})A^T$$

$$T_i = J_{i-1} + N_i^{-1}\hat{M}_{i-1}$$

$$P_{i,i-1} := \bar{F}_{i-1}P_{i-1}\bar{F}_{i-1}^T$$

$$P_{i,i-1} := P_{i,i-1} + \Gamma_{i-1}\hat{Q}_{i-1}\Gamma_{i-1}^T$$

$$P_{i,i-1} := P_{i,i-1} + H_{i-1}$$

$$s_i = CP_{i,i-1}C^T$$

$$h_i = s_i + \hat{R}_{i-1}$$

$$\hat{G}_i = \frac{T_i C^T}{h_i}$$

$$a_i = A\hat{G}_i$$

$$F_i = A - a_i C$$

$$K_i = F_i P_{i,i-1}\bar{F}_i^T$$

$$t_i = \hat{G}_i^T P_{i,i-1}$$

$$g_i = \hat{G}_i^T \hat{G}_i$$

$$\hat{R}_i = g_i t_i C^T - s_i$$

$$H_i = a_i \hat{R}_i a_i^T$$

$$\Gamma_i \hat{Q}_i \Gamma_i^T = P_{i,i-1} - K_i - H_i$$

977

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, Optimal Filtering. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1979.

[2] D. K. Barton, Radar System Analysis. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964.

[3] C. K. Chui and G. Chen, Kalman Filtering and Real-Time Tracking. To appear.

[4] A. H. Jazwinski, Stochastic Processing and Filtering Theory. Academic Press, Inc., New York, 1970.

[5] H. T. Kung, Why systolic architecture ? Computer, Vol.15, No.1 (1982), 37-46.

[6] S. Y. Kung, VLSI array processors. IEEE ASSP Magazine, Vol.2, No.3 (1985), 4-22.

[7] M. I. Skolnik, Introduction to Radar Systems. McGRAW-Hall, New York, 1962.

[8] J. M. Speiser and H. J. Whitehouse, A review of signal processing with systolic arrays. Real-Time Signal Processing VI, Proceeding of the SPIE International Symposium, Vol.431 (1984), 2-6.

[9] R. H. Travassos, Real-time implementation of systolic Kalman filters. Real-Time Signal Processing VI, Proceeding of the SPIE International Symposium, Vol.431 (1984), 97-104.

OPTIMAL IMPULSE - CORRECTION OF A RANDOM LINEAR OSCILLATOR*

P.L. Chow and J.L. Menaldi

Department of Mathematics, Wayne State University

Detroit, Michigan 48202

ABSTRACT. Consider the impulse-correction problem to minimize the randomly excited vibrations in a simple mechanical system. The system is modeled by a damped linear oscillator under a white-noise perturbation. By the dynamic programming approach, a set of variational (quasi-variational) inequalities are derived for the optimal cost function. Some analytical properties of the optimal cost function and the optimal control law are described. A numerical approximation procedure is proposed for computing the optimal cost functions. It is an iteration procedure which is shown to be convergent and stable. Some numerical results are given.

I. INTRODUCTION. The control of undesirable vibrations in a mechnical or electrical system is a problem of practical interest. For instance, in the design of light-weight robotic arm, the vibration of the flexible arm in the presence of external noise must be reduced to an acceptable level for satisfactory performance. In its simpliest form, a lumped parameter model is given by the optimal correction of a damped linear oscillator excited by a white-noise. In an earlier paper [1], we have studied this kind of problem, where the control process is either continuous or with jump discontinuities. Numerical solution of such problem was briefly discussed in [2] and was described in detail in a subsequent paper [3].

From the practical viewpoint, a continuous or piecewise continuous control process is more difficult to implement. This has led us to investigate the possibility of applying impulsive controls which, under suitable assumptions, require only a finite number of switching actions. Thus it becomes much easier to implement.

The optimal impulse-correction problem was treated by Gorbunov [4], among others, by the minimax principle. In contrast we shall analyze the problem by the dynamic programming principle and the associated quasi-variational inequalities. For the general mathematical techniques involved, one may consult the references [5] and [6].

This paper briefly summarizes some preliminary results of our investigation into this subject. Both analytical solution and the related numerical approximation will be discussed.

II. __OPTIMAL__ __IMPULSE-CORRECTION__ __PROBLEM__. We consider an impulse control of the undesirable mechanical vibrations in a randomly excited linear oscillator with damping:

(1)
$$\begin{cases} \ddot{x} + p\dot{x} + q^2 x = r\,\dot{w}_t + \dot{\nu}_t, & 0 < t \leqslant T, \\ x(0) = x_0, \quad \dot{x}(0) = y_0, \end{cases}$$

where p, q are the damping and spring constants; x_0, y_0 the initial position and velocity; r the intensity of the white noise $\dot{w}_t$, and $\dot{\nu}_t$ is an impulsive control. It is the formal derivative of the jump process:

(2)
$$\nu_t = \sum_{i=1}^{\infty} \xi_i \, H(t - \theta_i),$$

where $H(t) = 1$ for $t \geqslant 0$, 0 otherwise, being the Heaviside function, and ξ_i is the correctional impulse applied at the time θ_i. So ν_t represents the total correctional momentum up to the time t.

In order to avoid unnecessary control actions, one may either impose

980

penalty for excessive control actions or require a minimum change in state before applying an impulse correction. These considerations lead to the following alternative conditions:

(3) $\begin{cases} \text{(i)} \quad \text{the cost function } k(\xi) \text{ for an impulse magnitude } \xi \text{ satisfies} \\ \qquad k(\xi) \geqslant k_0 + k_1 |\xi|, \text{ for some constants } k_0, k_1 > 0, \text{ for all } \xi \in \mathbb{R}, \\ \text{(ii)} \quad \text{there exist constants } \delta_0, \delta_1 > 0 \text{ such that} \end{cases}$

$$|x(\theta_{i+1}) - x(\theta_i)| \geqslant \delta_0 + \delta_1 |x(\theta_i)|, \text{ for } i = 1, 2, \ldots.$$

It seems plausible to anticipate that the condition (i) implies the condition (ii) for sufficiently small δ_0, δ_1. For, if it costs to switch on the control, one will wait until a noticeable change in state occurs before taking another correctional impulse. Since the condition (i) causes less technical difficulty, we assume the condition (i) holds in this paper. The problem under the condition (ii) will be discussed elsewhere.

Setting $y = \dot{x}$, the state equation (1) can be rewritten in the integral form:

(4) $\begin{cases} x_t = x_0 + \int_0^t y_s \, ds, \\ \\ y_t = y_0 - \int_0^t (py_s + q^2 x_s) \, ds + \nu_t + r w_t, \end{cases}$

where $\langle w_t, t \geqslant 0 \rangle$ is the standard Wiener process in one dimension, and the control process ν_t is given by (2) subject to the condition (3), which will depend on $\langle w_s, s \leqslant t \rangle$.

For each control policy $v = \langle \xi_i, \theta_i \rangle$, let J denote the average cost function defined by

(5) $J(x, y, t, v) = E_{xy} \langle \int_t^T f(x_s, y_s) \, ds + g(x_T, y_T)$

$$+ \sum_{i=1}^{\infty} k(\xi_i) H(\theta_i - t) \rangle, \quad 0 \leqslant t \leqslant T, \ x \in \mathbb{R},$$

where f, g denote the running and terminal costs, respectively, and $k(\xi)$ is the cost for impulse control. They are assumed to be positive definite and bounded. The symbol E_{xy} stands for the conditional expectation with $x_t = x, \dot{x}_t = y$. Our goal is to find an optimal policy $\hat{v}$, in the form of a feedback law, which minimizes the average cost J, that is

$$(6) \qquad u(x,y,t) = J(x,y,t,\hat{v})$$

$$= \inf \langle J(x,y,t,v): v \rangle.$$

To analyze this problem, we will appeal to the principle of dynamic programming to derive a set of (quasi)variational inequalities.

III. **OPTIMAL COST FUNCTION AND VARIATIONAL INEQUALITIES**. By the dynamic programming approach, [6] it is possible to derive a set of variational inequalities governing the optimal cost function u. First, when there is no impulse at time t, u must satisfy the differential inequality:

$$(7) \qquad \begin{cases} -\dfrac{\partial u}{\partial t} + A\,u \leqslant f, \quad 0 \leqslant t < T, \ -\infty < x, y < \infty, \\[2mm] u|_T = g, \quad -\infty < x, y, < \infty, \end{cases}$$

where

$$(8) \qquad A\,u = -\frac{r^2}{2}\frac{\partial^2 u}{\partial y^2} + (q^2 x + p\,y)\frac{\partial u}{\partial y} - y\frac{\partial u}{\partial x},$$

and $u|_T = u(.,.,T)$.

If we decide to produce an impulse at t and then proceed with whatever is optimal, we get

$$(9) \qquad u \leqslant Mu \quad \text{for } 0 \leqslant t < T, \ -\infty < x, y < \infty,$$

where

$$(10) \qquad Mu(x,y,t) = \inf\langle k(\xi) + u(x, y+\xi, t) : |\xi| < \infty \rangle.$$

Since at each instant one must decide on one of these two options, one of the two inequalities (7) and (9) must be an equation. Summing up, the

982

optimal cost function u is the solution of the following optimality system:

(11)
$$
\begin{cases}
-\dfrac{\partial u}{\partial t} + Au \leqslant f, \\[6pt]
u \leqslant M\,u, \quad 0 \leqslant t < T, \;\; -\infty < x,y < \infty, \\[6pt]
\left(-\dfrac{\partial u}{\partial t} + A\,u - f\right)(u - Mu) = 0, \\[6pt]
u|_T = g, \quad -\infty < x,y < \infty.
\end{cases}
$$

We note that both sides of the inequality (9) involves the unknown function u. This is known as a quasi-variational inequality, in contrast with the variational inequality (13) below.

To solve the system (11), we may proceed by a successive approximation procedure. Define a sequence of approximate cost functions $\{u_n\}$ as follows:

(12)
$$
\begin{cases}
-\dfrac{\partial u^0}{\partial t} + Au^0 = f, \quad 0 \leqslant t < T, \\[6pt]
u^0|_T = g, \quad -\infty < x,y < \infty,
\end{cases}
$$

and, for $n \geqslant 1$

(13)
$$
\begin{cases}
-\dfrac{\partial u^n}{\partial t} + A\,u^n \leqslant f, \\[6pt]
u^n \leqslant M\,u^{n-1}, \quad 0 \leqslant t < T, \;\; -\infty < x,y < \infty, \\[6pt]
\left(-\dfrac{\partial u^n}{\partial t} + A\,u^n - f\right)(u^n - Mu^{n-1}) = 0, \\[6pt]
u^n|_T = g, \quad -\infty < x,y < \infty,
\end{cases}
$$

Under some suitable assumptions on the functions f, g and k it can be shown that the sequence of approximations u^n converges pointwise to the optimal cost u. In fact we have the following estimate

$$
0 \leqslant u^n(x,y,t) - u(x,y,t) \leqslant \delta^n, \quad n = 1,2,\ldots.
$$

where δ is a constant with $0 < \delta < 1$. This suggests an iterative numerical method of solution which will be described in the next section.

To indicate the optimal control law, we define the continuation set

$$C = \{(x,y,t) : u(x,y,t) < (Mu)(x,y,t)\},$$

and denote the associated free boundary of C by Γ. Also we let $\xi = \hat{\xi}(x,y,t)$ satisfy

$$(Mu)(x,y,t) = k(\hat{\xi}) + u(x,y+\hat{\xi},t).$$

Then, starting from the continuation set C, the system evolves freely in C until it first reaches the point (x,y) on the boundary Γ at $t = \theta_1$. Then the first impulse of the size $\hat{\xi}_1(x,y,\theta_1)$ is applied to push the system into the region C. Then the process is repeated as many times as necessary over the finite horizon T. On the other hand, if the initial state is not in C, an impulse correction should be made to bring the state into the region C and then coninue the process as before. It is possible to show the rule indicated above will yield the optimal feedback control. Therefore the constuction of the optimal policy depends on the solution $u(x,y,t)$ of the optimality system (11).

III. **NUMERICAL APPROXIMATION**. For numerical solution, we replace the unbounded x-y-t space by a rectangular box:

$$B = \{(x,y,t) \text{ in } \mathbb{R}^3 : |x| \leqslant a, |y| \leqslant b, 0 \leqslant t \leqslant T\}.$$

By a finite-difference scheme, we approximate the variational inequalities (11) by a discrete system and introduce some appropriate conditions on the boundary of B. To this end we set

$$\Delta x = \frac{a}{I}, \ \Delta y = \frac{b}{J}, \ \Delta t = \frac{T}{N},$$

for some positive integers I, J and N. Denote by $Q(I,J,N)$ the set of mesh points in B, i.e.

$$Q(I,J,N) = \{(x_i,y_j,t_n) : x_i = i\Delta x, \ y_j = j\Delta y, \ t_n = n\Delta t,$$

$$i = 0,\pm 1,\ldots,\pm I; \ j=0,\pm 1,\ldots,\pm J, \ n=0,1,\ldots,N\}.$$

984

The pivotal value of the approximate solution $\tilde{u}$ to u at the mesh point (x_i, y_j, t_n) is given by

$$u_{i,j}^n = \tilde{u}(x_i, y_j, t_n).$$

Next we discretize the differential operator in (7) as follows:

$$u_{i,j}^n \leqslant c_1(i,j)\, u_{i,j}^{n+1} + c_2(i,j) u_{i,j+1}^n + c_3(i,j) u_{i,j-1}^n$$

$$+ c_4(i,j) u_{i+1,j}^n + c_5(i,j) u_{i-1,j}^n.$$

$$\equiv (M,u)_{i,j}^n.$$

where

$$c_1 = [\Delta t\, c_0(i,j)]^{-1},$$

$$c_2 = \langle (\tfrac{r}{\Delta y})^2 + [q^2 i(\tfrac{\Delta x}{\Delta y}) + p\ j]^- \rangle / c_0(i,j),$$

$$c_3 = \langle (\tfrac{r}{\Delta y})^2 + [q^2 i(\tfrac{\Delta x}{\Delta y}) + pj]^+ \rangle / c_0(i,j),$$

$$c_4 = (j)^+\, \Delta y / \Delta x\, c_0(i,j),$$

$$c_5 = (j)^-\, \Delta y / \Delta x\, c_0(i,j),$$

$$c_0 = (\tfrac{1}{\Delta t}) + \frac{2r}{(\Delta y)^2} + |q^2 i \tfrac{\Delta x}{\Delta y} + pj| + |j|\, \tfrac{\Delta y}{\Delta x}.$$

and

$[x]^+([x]^-) = 1$ if $x>(<)0$; 0, otherwise.

Note that

$$c_1 + c_2 + c_3 + c_4 = 1, \text{ and } c_i > 0,\ i=1,\ldots,4.$$

This property is crucial and gives rise to a discrete maximum principle.

Next we discretize the quasi-variational inequality (9) as follows:

$$u_{i,j}^n \leqslant \min\ (k(\ell) + u_{i,j+\ell}^n : \ell \text{ and } |j+\ell| \leqslant J)$$

$$\equiv (M_2 u)_{i,j}^n.$$

Here we set

$$u = - \langle u^n_{i,j} \rangle,$$

regarded as a vector in the space E defined by

$$E = \langle u \in \mathbb{R}^{(2I+1) \times (2J+1) \times N+1} : \|u\| < \infty \rangle,$$

where

$$\|u\| = \max_{i,j,n} |u^n_{i,j}|.$$

The discrete version of the optimality system (13) can now be written as:

(i) $u \leqslant M_\ell u$, $\ell = 1, 2$,

(ii) $\max_\ell \langle u^n_{i,j} - (M_\ell u)^n_{i,j} \rangle = 0$, for each i, j, n,

(iii) $u^N_{i,j} = g_{i,j}$,

(iv) $u^n_{i,j} = v_{i,j}$, for $|i| = I$ or $|j| = J$, $n = 0, 1, \ldots, N-1$,

in which $g_{i,j} = g(x_i, y_j)$ and $v_{i,j}$ is an apparent boundary value imposed on the artifical boundary. The boundary value v should be chosen to approximate the asymptotic value of u as $|x|, |y| \to \infty$.

For numerical solution, we will present an iteration procedure. Let Q be an operator on E defined by

$$(Qu)^n_{i,j} = \begin{cases} \min_\ell \langle (M_\ell u)^n_{i,j}, \ n < N, \ i \neq \pm I, \ j \neq \pm J \rangle, \\ v_{i,j}, \ n < N, i = \pm I \text{ or } j = \pm J, \\ g_{i,j}, \ n = N. \end{cases}$$

We start with an initial guess $u^{(0)} = \langle u^{n,0}_{i,j} \rangle$ which satisfies the conditions (i), (iii) and (iv) of the system (14). Then we compute

$$u^{(1)} = Qu^{(0)}.$$

For $\ell > 1$, we compute successively

$$u^{(\ell)} = Qu^{(\ell-1)}, \ \ell = 2, 3, \ldots.$$

Given a pre-assigned error $\varepsilon > 0$, the iteration terminates at the m-th step when

986

$$\|u^{(m)} - u^{(m-1)}\| < \varepsilon.$$

By invoking the build-in maximum principle, we are able to show that the sequence of iteates $\langle u^{(\ell)}\rangle$ converges monotonically to the solution of the discrete system (14). Furthermore the numerical procedure is stable.

As an example, numerical computation was carried out for the so-called "cheap control" problem, where $k \equiv 0$. For some special values of p,q,r, numerical results are shown in Figs. 1-3. Here we also assume $f(x,y) = x^2 + y^2$ and $q \equiv 0$. Fig. 1 and Fig. 2 show the y - and x - section curves for the minimal cost function u at $n = 1$, respectively. The continuation set C, marked by the letter c's, is displayed in Fig. 3. Numerical calculation for the general problem will be undertaken in the near future.

REFERENCES

[1] P.L. Chow and J.L. Menaldi, Optimal Correction of a Damped Linear Oscillator under Random Perturbations, Trans. 2nd Army Conf. on Appl. Math. and Comp., (1985), pp. 149-158.

[2] ——————————————, On the Numerical Solution of of an Optimal Correction Problem, Trans. 3rd. Army Conf. or Appl. Math. and Comp., (1986), pp. 531-558.

[3] M.C. Bancora-Imbert, P.L. Chow and J.L. Menaldi, (Submitted for publication).

[4] V.K. Gorbunov, Minimax Impulsive Correction of Perturbations of a Linear Damped Oscillator, Appl. Math. and Mech. (PMM), 40 (1976), pp. 252-259.

[5] W.H. Fleming and R.W. Rishel, Deterministic and Stochastic Optimal Control, Springer-Verlag, New York, 1975.

[6] A. Bensoussan and J.L. Lions, Impulse Control and Quasi-variational Inequalities, (Translation in English), Gauthier-Villars, Paris, 1984.
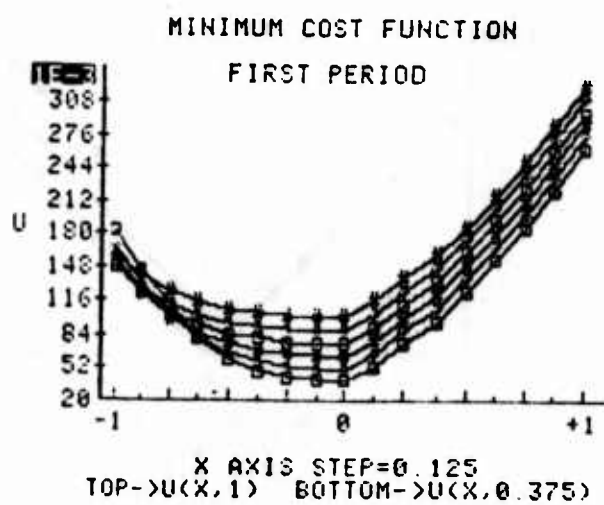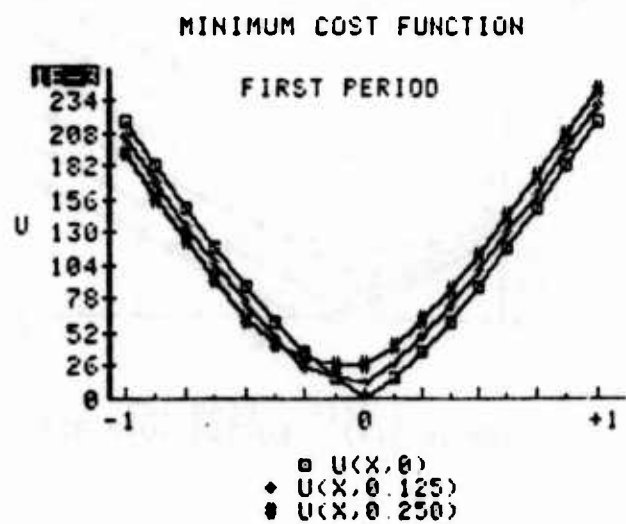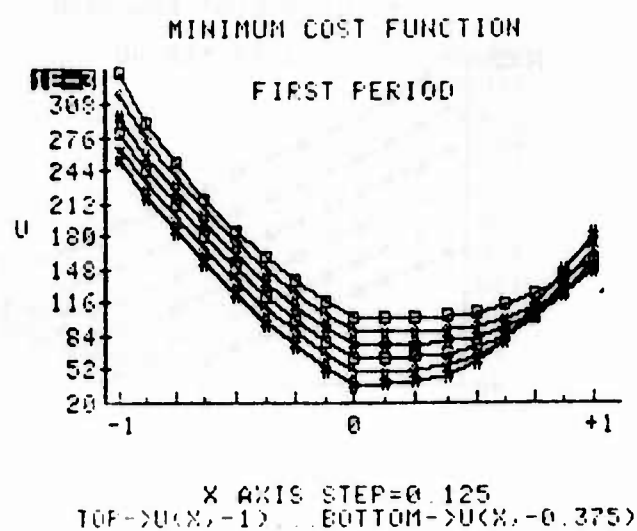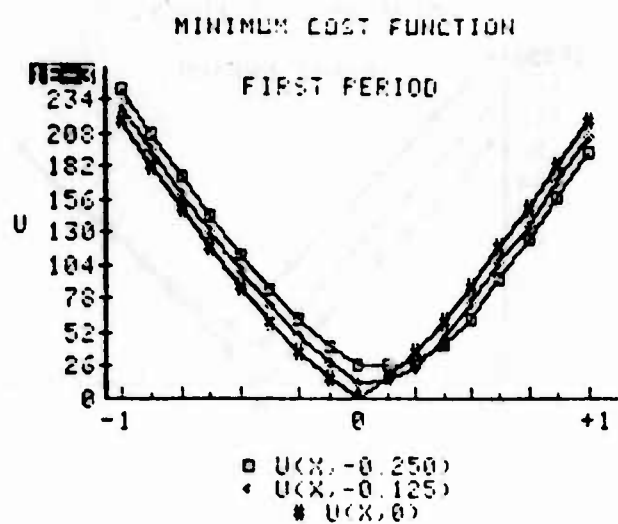
FIG. 1. MINIMUM COST FUNCTION: Y-SECTION CURVES
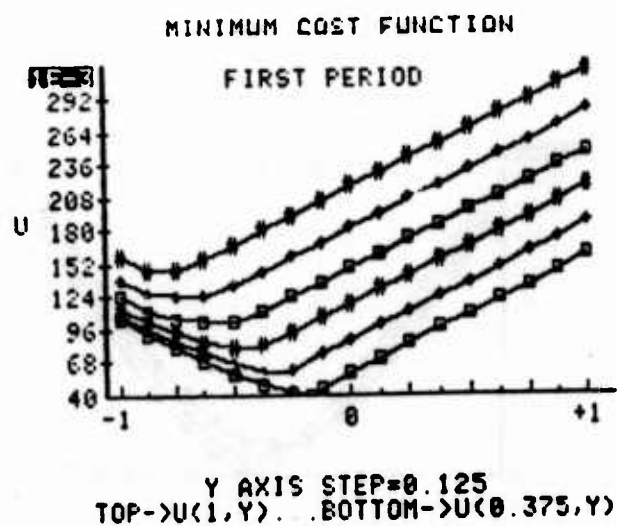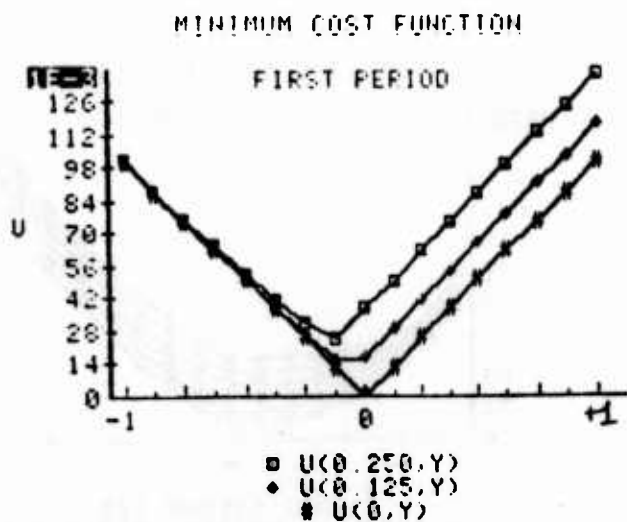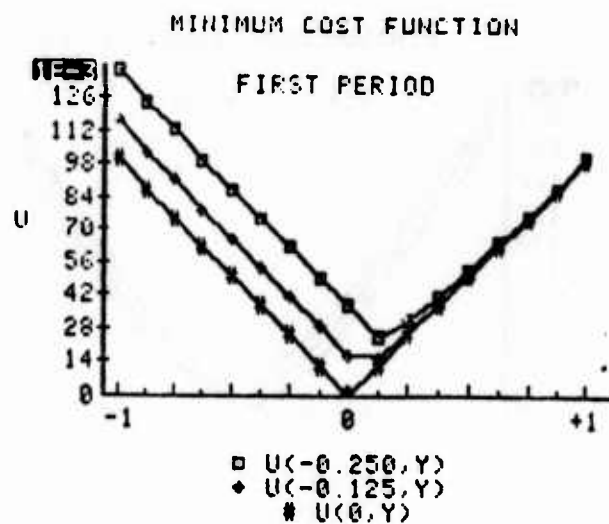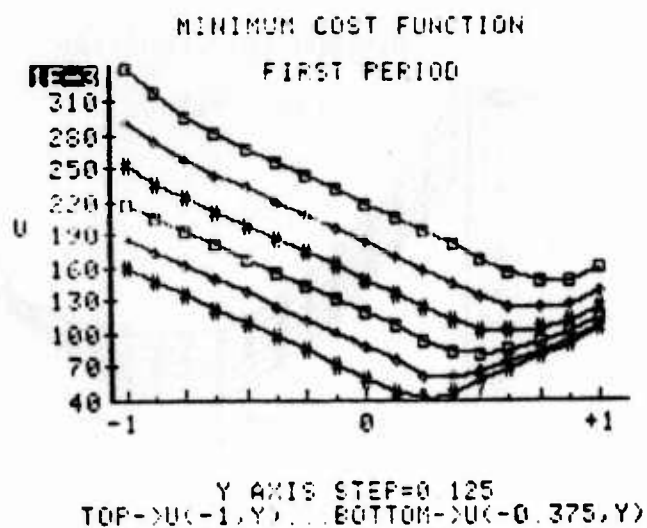
FIG. 2. MINIMUM COST FUNCTION: X-SECTION CURVES
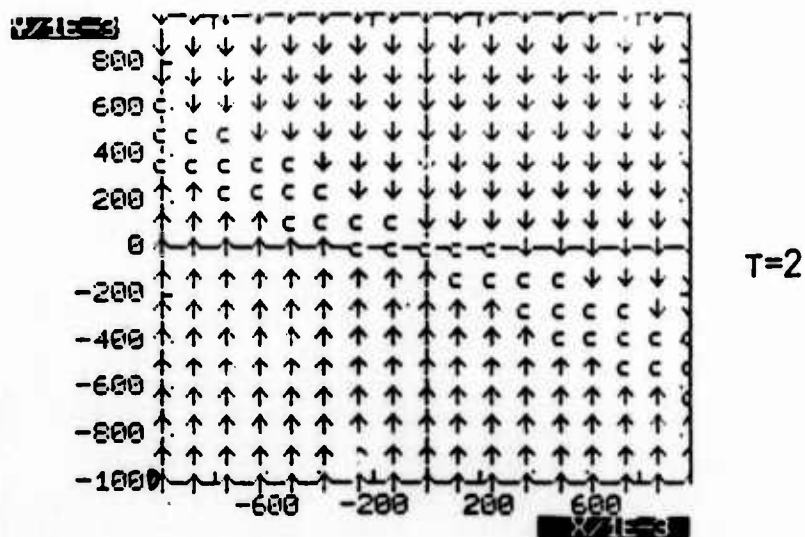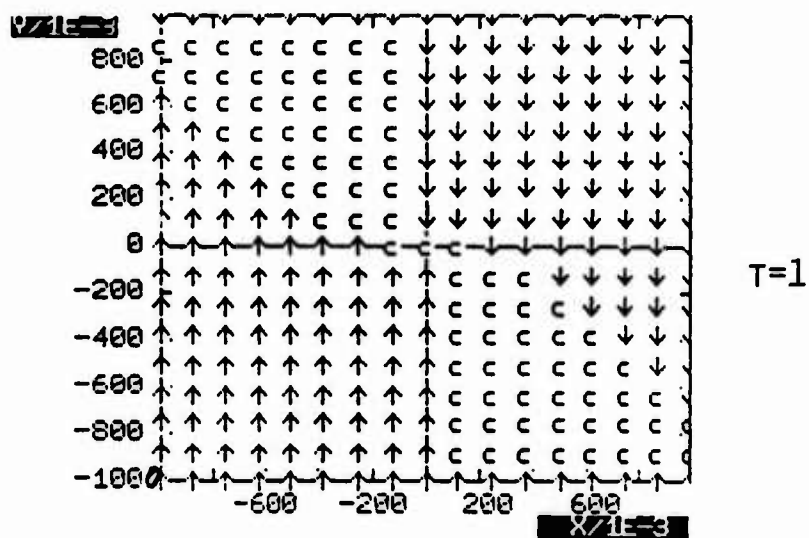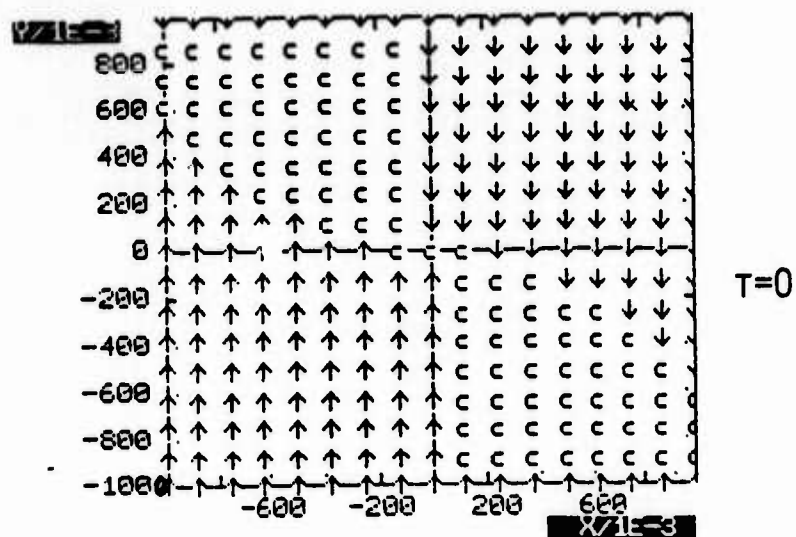
FIG. 3. COMPUTED FREE-BOUNDARIES   991

PULSE - ARRIVAL TIME FOR WAVES IN TURBULENT MEDIA

P.L. Chow and J.L. Menaldi

Department of Mathematics, Wayne State University

Detroit, Michigan 48202

ABSTRACT. The pulse-arrival time for waves in turbulent media is treated by the probabilistic method. The wave-travel time is introduced as a probabilistic hitting time by an ideal particle moving along a characteristic curve. By a simple-wave approximation, the mean travel-time and its variance are investigated and found to satisfy appropriate boundary-value problem for elliptic equations in the diffusion limit. As an example, the mean travel-time for a plane wave and its variance are calculated explicitly.

I. <u>INTRODUCTION</u>. In wave propagation through a turbulent medium, the statistical problem for pulse-arrival time has been studied by many authors (for references, see [1]), due to its important applications, e.g. the propagation of laser beam through the atmosphere. In the engineering literature, a popular method of treating such problems is the so-called "temporal moment" method. Specifically we let $u(x,t,\omega)$ be the random wave function of the pulse, and define

$$p(x,t) = \frac{E|u(x,t)|^2}{\int_0^\infty E|u(x,s)|^2 ds},$$

where the initial pulse shape is $u(x,0) = f(x)$. Here one considers the pulse arrival time τ at x from the orgin as a random variable with $p(x,t)$ as its probability density function. Thereby the m-th moment is calculated by

$$T_m = \int_0^\infty t^m \, p(x,t) \, dt.$$

In practice the above quantity is not experimentally measurable and difficult to compute. These obvious shortcomings have suggested the need for a new investigation.

In this paper we shall introduce a probabilistic method for treating the problem. This will give the pulse-arrival time a proper meaning conceptually. At the same time a method of approximation will be proposed for computing the associated moments more efficiently. Some computional results will be provided for illustration.

II. PLUSE-ARRIVAL TIME. Consider the propagation of a pulse-wave through a turbulent medium in the free space. Then the wave function $u(x,t,\omega)$ satisfies the random wave equation:

$$(1) \qquad \frac{\partial^2 u}{\partial t^2} = c^2(x,t,\omega) \, \nabla^2 u, \quad t>0, \quad |x|<\infty,$$

which is subject to the initial conditions

$$(2) \qquad u(x,0,\omega) = f(x),$$

$$\frac{\partial}{\partial t} u(x,0,\omega) = 0.$$

Here $x = (x_1, x_2, x_3)$ is the position vector; $c(x,t,\omega)$ the random local wave speed; ∇^2 the Laplacian operator, and $f(x)$ is the initial pulse form which is concentrated near the origin. For the wave equation (1), the motion of wave-front is governed by the random Hamilton-Jacobi equation [3]:

$$(3) \qquad \left(\frac{\partial \phi}{\partial t}\right)^2 = c^2 \, |\nabla \phi|^2.$$

By the method of characteristics, the above yields the characteristic equation

$$(4) \qquad \frac{dx_t}{dt} = c(x_t, t, \omega) \hat{p}_t, \quad x_0 = x,$$

$$\frac{dp_t}{dt} = -\nabla c(x_t, t, \omega) \, |p_t|, \quad p_0 = p,$$

where $\hat{p}_t$ denotes the unit vector along the vector p_t. Physically the characteristic curves

$$\Gamma_x: \; y = x_t, \quad t \geq 0$$

is called rays emanating from the point x.

Let B be a spatial region of influence by rays through x. We first define the "travel-time" from x to B by

(5)
$$\tau_x(B) = \inf_{\Gamma_x} \{t > 0 \; : \; x_t \text{ in } B\}.$$

By definition, $\tau_x(B)$ is a probabilistic hitting time for the region B by the random ray x_t. In terms of $\tau_x(B)$, it is reasonable to define the pulse-arrival time $\tau(B)$ to the region B as a random variable with the conditional probability distribution

$$P\{\tau(B) < t \,|\, x_0 = x\} = P\{\tau_x(B) < t\}.$$

The initial distribution $p_0(x)$ may incorporate the pulse-shape by, for instance, setting

(6)
$$p_0(x) = \frac{|f(x)|}{\int |f(y)| \, dy}.$$

Therefore, from now on, we will only be concerned with the wave-travel time $\tau_x(B)$, instead of the pulse-arrival time $\tau(B)$.

If x_t is a Markov process, the mathematical theory for the hitting time or, in general, a stopping time is a well-developed subject [4]. For example, if

$$x_t = x + w_t,$$

where w_t is the standard Brownian motion is space, then the mean hitting time

$$T(x) = E \, \tau_x(B)$$

995

is known to satisfy the Dirichlet problem

$$\begin{cases} \frac{1}{2} \nabla^2 T(x) = -1, & x \text{ in } B', \\ T(x) = 0, & x \text{ on } \partial B \end{cases}$$

where B' is the exterior region and ∂B the boundary of B. Similarly, for a general diffusion process, the computation of the mean travel-time or a higher moment reduces to the solution of the appropriate boundary value problem. Therefore, in order to employ the method of differential equations, we are going to seek a diffusion approximation to the ray process x_t. This procedure has been used to study the progressing waves in random media [5].

III. DIFFUSION APPROXIMATION. Suppose the flutuation of the wave speed due to turbulence is weak so that

(7) $$c^\varepsilon(x,t,\omega) = c_0 + \varepsilon \, \xi(x,t,\omega),$$

where c^ε designates the depedence of c on the small fluctuation amplitude $\varepsilon > 0$, c_0 is the constant average wave speed, and ξ is a random field satisfying

(8) $$\begin{cases} E\xi = 0, \\ E\xi(x,t)\xi(y,s) = R(x-y,t-s). \end{cases}$$

In view of (7), the ray equation (4) can be rewritten as

(9) $$\begin{cases} \dfrac{dx_t^\varepsilon}{dt} = c^\varepsilon(x_t^\varepsilon,t,\omega)\hat{p}_t, & x_0^\varepsilon = x, \\[2mm] \dfrac{dp_t^\varepsilon}{dt} = \varepsilon \, \nabla\xi(x_t,t,\omega)|p_t^\varepsilon|, & p_0^\varepsilon = p. \end{cases}$$

Since the right-hand side of (10) is small, as a first approxmation, it is neglected. Then the equation (9) becomes

(11) $$\dfrac{dx_t^\varepsilon}{dt} = [c_0 + \varepsilon\xi(x_t^\varepsilon,t,\omega)] \, \hat{p}, \quad x_0^\varepsilon = x.$$

Let us rename x_t^0 the unperturbed solution

(12)
$$x_t^0 = x + c_0 t \hat{p},$$

and set

(13)
$$y_t^\epsilon = x_t^\epsilon - x_t^0.$$

Then in view of (12) and (13), the equation (11) yields

(14)
$$\begin{cases} \dfrac{dy_t^\epsilon}{dt} = \epsilon \, \xi \, (y_t^\epsilon + x_t^0, \ t, \ \omega) \hat{p}, \\[2mm] y_0^\epsilon = 0. \end{cases}$$

To obtain a diffusion approximation for y_t^ϵ, we assume that the random field $\xi(x,t,\omega)$ satisfies a strong-mixing condition in space and time so that $\xi(x,t,\cdot)$ and $\xi(y,s,\cdot)$ become asymptotically independent when either $|t-s|$ or $|x-y|$ becomes large. Then one can invoke a limit theorem due to Khasminskii [6] to get the diffusion approxmation. That is, for small ϵ and large t with $r = \epsilon^2 t$ fixed,

(15)
$$y_y^\epsilon = y^\epsilon(r/\epsilon^2) \sim y(r),$$

where $y(r)$ is a diffusion process with known diffusion coefficients b_{ij} and the drift a_j. They are defined by

(16)
$$\begin{cases} b_{ij} = b \, \hat{p}_i \hat{p}_j, \\[2mm] a_j = a \, \hat{p}_j, \end{cases}$$

where

(17)
$$\begin{cases} b = \lim_{T \to \infty} \dfrac{1}{T} \int_0^T \int_0^T R[c_0(t-s)\hat{p}, t-s] \, dt \, ds, \\[3mm] a = \lim_{T \to \infty} \dfrac{1}{T} \int_0^T \int_0^T \sum_{i=1}^3 R_i [c_0(t-s)\hat{p}, t-s] \, \hat{p}_i \, dt \, ds, \end{cases}$$

with $R_i(x,t) = \dfrac{\partial R(x,t)}{\partial x_i}$.

Alternatively the process $y(r)$ satisfies the Ito equation:

$$(18) \qquad \begin{cases} dy_i(t) = a_i dr + \sum_{j=1}^{3} b_{ij} \, dw_j(r) \\ y_i(0) = 0, \quad i=1,2,3. \end{cases}$$

Here $w_i(r)$, $i=1,2,3$, are independent standard Brownian motions. Returning to the original process x_t^ε, from (13), we get the diffusion approximation:

$$(19) \qquad x_t^\varepsilon \sim x_t \equiv x_t^0 + y \, (\varepsilon^2 t).$$

Therefore, under this approxmation, the travel-time $\tau_x^\varepsilon(B) \sim \tau_x(B)$ with

$$(20) \qquad \tau_x^\varepsilon (B) = \inf \langle t>0: x_t^\varepsilon \in B \rangle$$

$$(21) \qquad \tau_x(B) = \inf \langle t>0 : x_t \in B \rangle.$$

As mentioned before, the computation of statistical properties of τ_x for the Markov process x_t is much easier.

IV,. <u>COMPUTATION OF STATISTICS</u>. We will compute the first two moments of the travel-time $\tau_x(B)$ as a diffusion approximation. To this end let L be the generator of the diffusion process x_t gien by

$$(22) \qquad L \, u = \frac{1}{2} \varepsilon^2 \sum_{j=1}^{3} b_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{3} (c_0 \hat{p}_i + \varepsilon^2 a_i) \frac{\partial u}{\partial x_i},$$

and let $T(x,b)$ and $T_2(x,B)$ denote the first and second moments of $\tau_x(B)$, respectively. Then it has been shown [7] that they satisfy the following boundary-value problems:

$$(23) \qquad \begin{cases} L \, T(x,B) = -1, \text{ for } x \text{ in } B', \\ T(x,B) = 0, \text{ for } x \text{ on } \partial B, \end{cases}$$

$$(24) \qquad \begin{cases} L \, T_2(x,b) = -2 \, T(x,B), \quad x \text{ in } B', \\ T_2(x,B) = 0, \text{ for } x \text{ on } \partial B, \end{cases}$$

998

where B' means the exterior and ∂B the boundary of B. Thereby the computation of the first two moments of τ_x is reduced to solving the respective boundary-value problems (23) and (24). In fact all higher moments can be obtained in this way, (see [7]).

For explicit results, it is interesting to consider a special case. Suppose initially the wave propagates in the x_1-direction, under the simple wave approximation, so that

$$(25) \qquad \qquad \hat{p} = (1,0,0).$$

Then the operator L in (22) reduces to

$$(26) \qquad L_1 u = \frac{1}{2} b\epsilon^2 \frac{\partial^2 u}{\partial x_1^2} + (c_0 + a\epsilon^2) \frac{\partial u}{\partial x_1}.$$

Consequently, if we are interested in the travel time for a nearly plane wave from $x_1 = \ell_1$ to $B = \langle x_1 \geqslant \ell_2 \rangle$, the equations (23) and (24) are reducible to problems in one dimension, which can be solved easily. The results are given by:

$$(27) \qquad \qquad T(\Delta \ell) = \frac{\Delta \ell}{\hat{c}}.$$

$$(28) \qquad \qquad T_2(\Delta \ell) = (\frac{\Delta \ell}{\hat{c}})^2 + \frac{\sigma_\epsilon^2 \Delta \ell}{2(\hat{c})^3},$$

where $\Delta \ell = (\ell_2 - \ell_1)$ and

$$(29) \qquad \qquad \hat{c} = (c_0 + a\epsilon^2),$$

$$(30) \qquad \qquad \sigma_\epsilon^2 = b\epsilon^2,$$

which may be interpreted as the "effective" mean and variance parameters, respectively.

We note that $\hat{c}(\epsilon) < c_0$ for $\epsilon > 0$. Physically the drift a, in view of (17), should be negative as the correlation R is a decreasing function of x so that the mean travel-time increases due to random scattering. From

999

(27) and (28) we obtain the standard deviation of the travel-time:

(31)
$$\sigma_\tau(\Delta \ell) = \frac{(b\Delta \ell)^{1/2}}{\sqrt{2}(\hat{c})^{3/2}} \epsilon ,$$

which is proportional to the correlation length $b^{1/2}\epsilon$ and the square-root of distance $(\Delta \ell)^{1/2}$, and inversely, to $(\hat{c})^{3/2}$.

As a numerical example, we set $c_0 = 1$ and

(32)
$$R(r,t) = (\frac{e^{-\alpha t}}{1+r^2}), \quad \alpha > 0.$$

Some numerical results are displayed in four graphs, Figs. 1-4 in the following page. In Fig. 1, the mean travel-time T is plotted against the distance $x = \Delta \ell$ for $\alpha = 0.7$ and $\epsilon = 0$, 0.2 and 0.5, respectively, while α is changed to 1.0 in Fig. 2. They showed clearly the decrease of T as ϵ increases. In Figs 3 and 4, the three lines correspond to the mean travel-time T and its deviations $(T \pm \sigma_\tau)$ v.s. $x = \Delta \ell$ with $\epsilon = 0.5$ fixed, for $\alpha = 0.7$ and 1.0, respectively. These results show an increase of σ_τ with the distance but a decrease of σ_τ as the decay exponent α increases. Even though these results are obtained for a special case, the qualitative feature may hold in general. For details, one is referred to the paper [7].

Fig.3 X=x/a = 0.7

Fig.1 X=x/a = 0.7

Fig.4 X=x/a = 1.0

Fig.2 X=x/a = 1.0
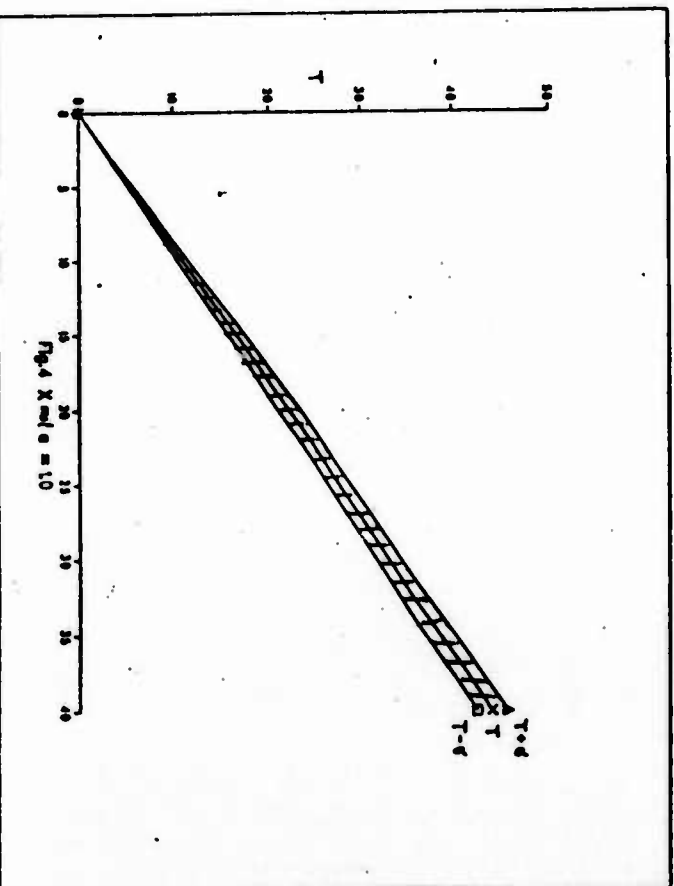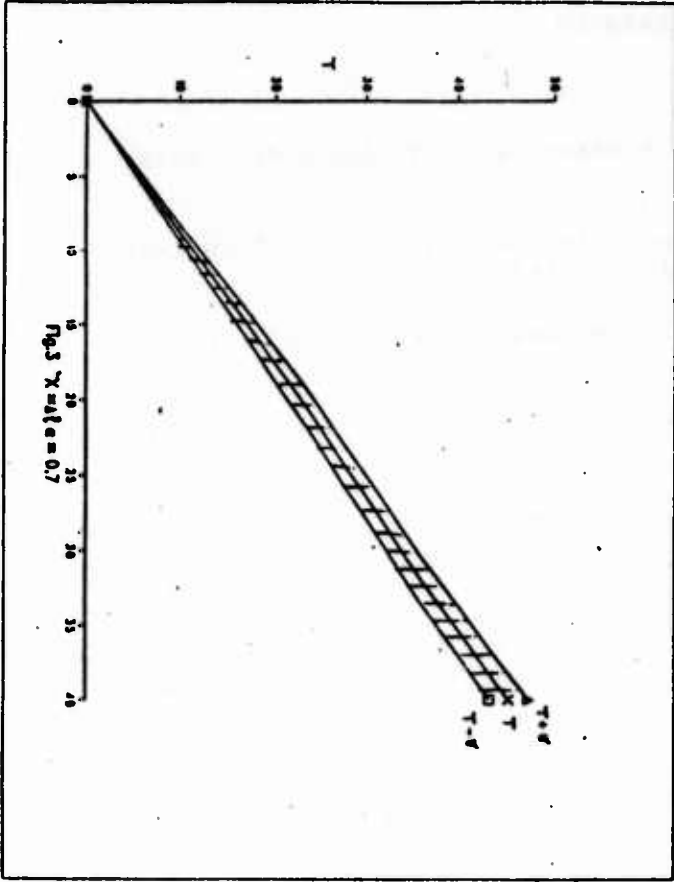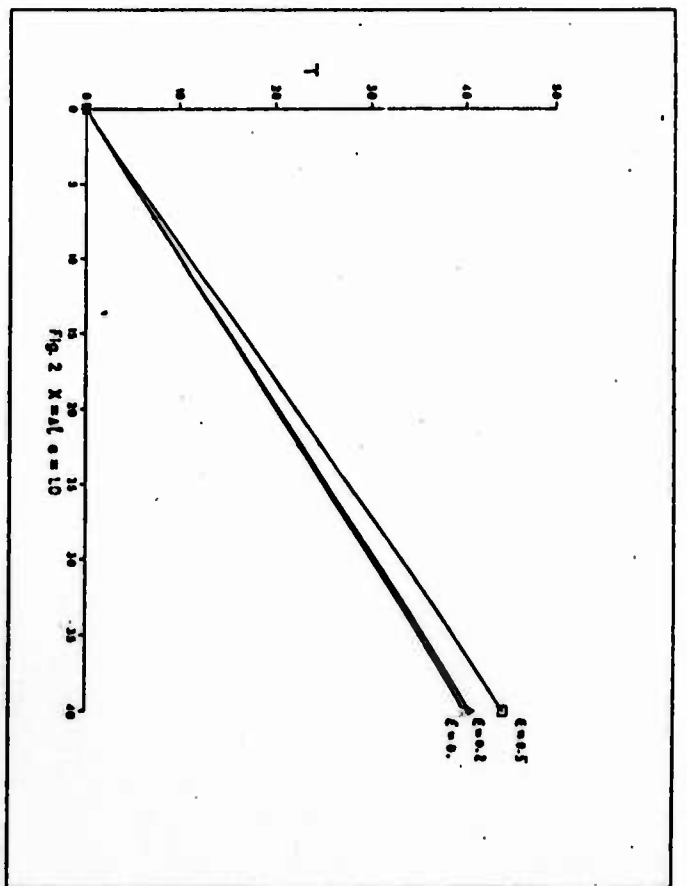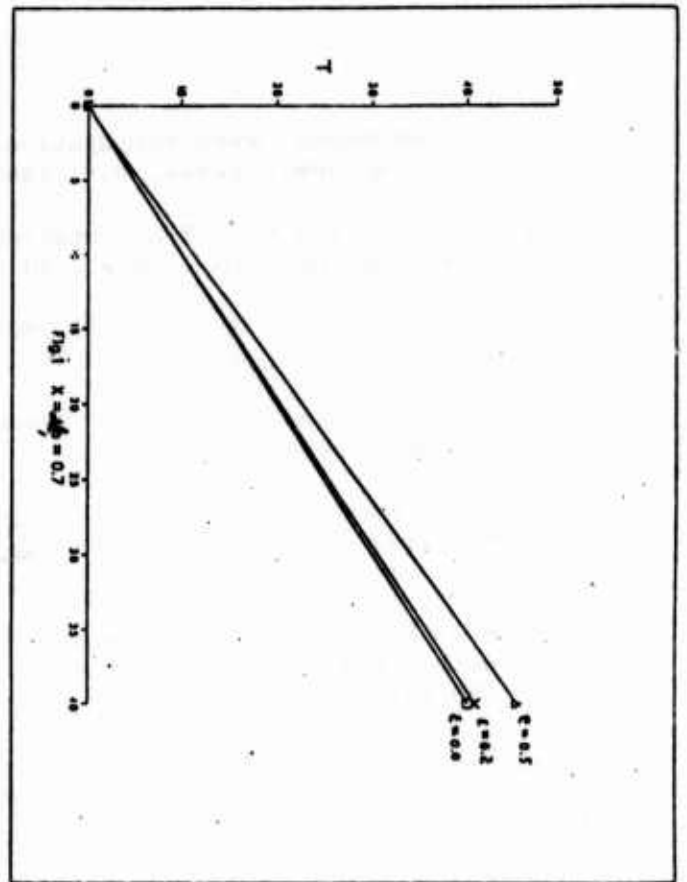
REFERENCES

1. A. Ishimaru, Wave Propagations and Scattering in Random Media, Vols. I,II, Academic press, N.Y. 1982.

2. C.H. Liu and K.C. Yeh, Statistics of Pulse Arrival Time in Turbulent Media, J. Opt. Soc. Amer., $\underline{70}$ (1980), pp. 168-172.

3. R. Courant and D. Hilbert, Methods of Mathematical Physics, vol. II, Wiley-Interscience, N.Y. 1962.

4. S.C. Port and C.J. Stone, Brownian Motion and Classical Potential Theory, Academic press, N.Y. 1978.

5. P.L. Chow, Simple Progressing Waves in Random Meedia, in proc. Symp. on Multiple Scatterings in Random Media and Random Rough Surfaces, Technomic pub. co. Lancaster, PA., 1986.

6. R.Z. Khasminskii, A Limit Theorem for Solutions of Differential Equations with a Random Right-Hand Side, Theory Prob. and Appl. $\underline{11}$ (1966), pp. 390-406.

7. P.L. Chow, Travel-Time Problems for Waves in Random Media, Proc. IMA Workshop on Random Media, Springer-Verlag, 1986 (to appear).

THE TRANSITION FROM PHASE LOCKING TO DRIFT
IN A SYSTEM OF TWO WEAKLY COUPLED VAN DER POL OSCILLATORS*

Tapesh Chakraborty and Richard H. Rand
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca NY 14853

ABSTRACT

We investigate the slow flow resulting from the
application of the two variable expansion perturbation method
to a system of two linearly coupled van der Pol oscillators.
The slow flow consists of three nonlinear coupled ode's on the
amplitudes and phase difference of the oscillators. We obtain
regions in parameter space which correspond to phase locking,
phase entrainment and phase drift of the coupled oscillators.
In the slow flow, these states correspond respectively to a
stable equilibrium, a stable limit cycle and a stable
libration orbit.

Phase entrainment, in which the phase difference between
the oscillators varies periodically, is seen as an
intermediate state between phase locking and phase drift. In
the slow flow, the transitions between these states are shown
to be associated with Hopf and saddle-connection bifurcations.

INTRODUCTION

In this work we shall be concerned with the behavior of
two coupled oscillators. We begin by introducing some
terminology. We suppose that the outputs of the two
oscillators are of the form

$$(1.1) \qquad x_1(t) = R_1(t) \cos(t - \theta_1(t))$$

$$(1.2) \qquad x_2(t) = R_2(t) \cos(t - \theta_2(t))$$

in which $R_1(t)$ represents amplitude modulation and $\theta_1(t)$
represents frequency or phase modulation. We shall define the
terms phase locking, phase drift and phase entrainment for two
functions of the form (1). Although these definitions can be
generalized to apply to a wider class of functions than (1),
we shall restrict our attention to such functions since the
approximate solutions which we are interested in this work
will have this form.

We define the phase difference $\varphi(t)$ as

$$(2) \qquad \varphi(t) = \theta_1(t) - \theta_2(t)$$

Then the functions (1) will be said to be <u>1:1 phase locked</u> if $\varphi(t)$ is constant. If, on the other hand, the oscillators are running at unequal average frequencies, then $\varphi(t)$ will grow unbounded, defining the condition of <u>1:1 phase drift</u>. An intermediate situation exists when $\varphi(t)$ varies periodically, a condition which we shall call <u>1:1 phase entrainment</u> [1].

TWO WEAKLY COUPLED VAN DER POL OSCILLATORS

We shall be interested in the following system of two coupled van der Pol oscillators

$$(3.1) \qquad \frac{d^2 x_1}{dt^2} + x_1 - \epsilon (1-x_1^2) \frac{dx_1}{dt} = \epsilon \, \alpha \, (x_2 - x_1)$$

$$(3.2) \qquad \frac{d^2 x_2}{dt^2} + (1 + \epsilon \, \Delta) \, x_2 - \epsilon (1-x_2^2) \frac{dx_2}{dt} = \epsilon \, \alpha \, (x_1 - x_2)$$

Here $\epsilon \ll 1$ and Δ and α are parameters. Δ is related to the (small) difference in linearized frequencies, and α represents the strength of the coupling. In a previous work [2], the two variable expansion perturbation method was utilized to obtain an approximate solution to eqs.(3), valid to order ϵ:

$$(4.1) \qquad x_1 = R_1(\eta) \cos(t - \theta_1(\eta)) + O(\epsilon)$$

$$(4.2) \qquad x_2 = R_2(\eta) \cos(t - \theta_2(\eta)) + O(\epsilon)$$

where the slow time variable η is given by

$$(5) \qquad \eta = \epsilon \, t$$

and where the amplitudes R_1 and R_2, and the phase angle $\varphi = \theta_1 - \theta_2$ are given by the slow flow on the space $M : R^+ \times R^+ \times S^1$

$$(6.1) \qquad 2 \frac{dR_1}{d\eta} = -R_1 \left[\frac{R_1^2}{4} - 1 \right] + \alpha \, R_2 \sin \varphi$$

$$(6.2) \qquad 2 \frac{dR_2}{d\eta} = -R_2 \left[\frac{R_2^2}{4} - 1 \right] - \alpha \, R_1 \sin \varphi$$

$$(6.3) \qquad 2 \frac{d\varphi}{d\eta} = \Delta + \alpha \left[\frac{R_2}{R_1} - \frac{R_1}{R_2} \right] \cos \varphi$$

1004

Eqs. (6) are invariant under the transformation

(7) $R_1 \to R_2$, $R_2 \to R_1$, $\varphi \to \varphi - \pi$

and hence possess the corresponding symmetry. Thus if there is an equilibrium point at $(R_1^o, R_2^o, \varphi^o)$, then there also one at $(R_2^o, R_1^o, \varphi^o - \pi)$. In order to simplify the following discussion, we shall only talk about half of the system, i.e., if we say that the system (6) contains an equilibrium point (or a periodic orbit), then it actually contains two equilibria (or periodic orbits), the other one being located at the symmetrical position in M under the transformation (7).

RESULTS

Our goal is to classify the various qualitatively distinct behaviors of the system (6) as we change the values of the parameters α and Δ. We summarize here the results obtained in [3].

The system (6) possesses two other symmetries based on invariance under the transformations

(8.1) $\alpha \to -\alpha$, $\varphi \to \varphi - \pi$, and

(8.2) $\Delta \to -\Delta$, $\varphi \to \pi - \varphi$

Thus it turns out that the qualitative behavior is invariant under both $\alpha \to -\alpha$ and $\Delta \to -\Delta$, and we present our results in the $\alpha^2 - \Delta^2$ parameter plane, see Fig.1.

The system (6) contains 3 equilibria in region R in Fig.1, and one equilibrium elsewhere. Region R can be shown [3] to be bounded by curves having the equation:

(9) $\Delta^6 + (6\alpha^2 + 2) \Delta^4 + (12\alpha^4 - 10\alpha^2 + 1) \Delta^2 + 8\alpha^6 - \alpha^4 = 0$

In obtaining (9) as well as many of the other results in this work, we used the computer algebra system MACSYMA [4].

The presence of limit cycles in (6) may be investigated by looking for Hopf bifurcations, i.e. by linearizing in the neighborhood of each of the equilibria, and requiring that there exist a pair of pure imaginary eigenvalues. This leads to the curve H in Fig.1, which can be shown [3] to have the equation:
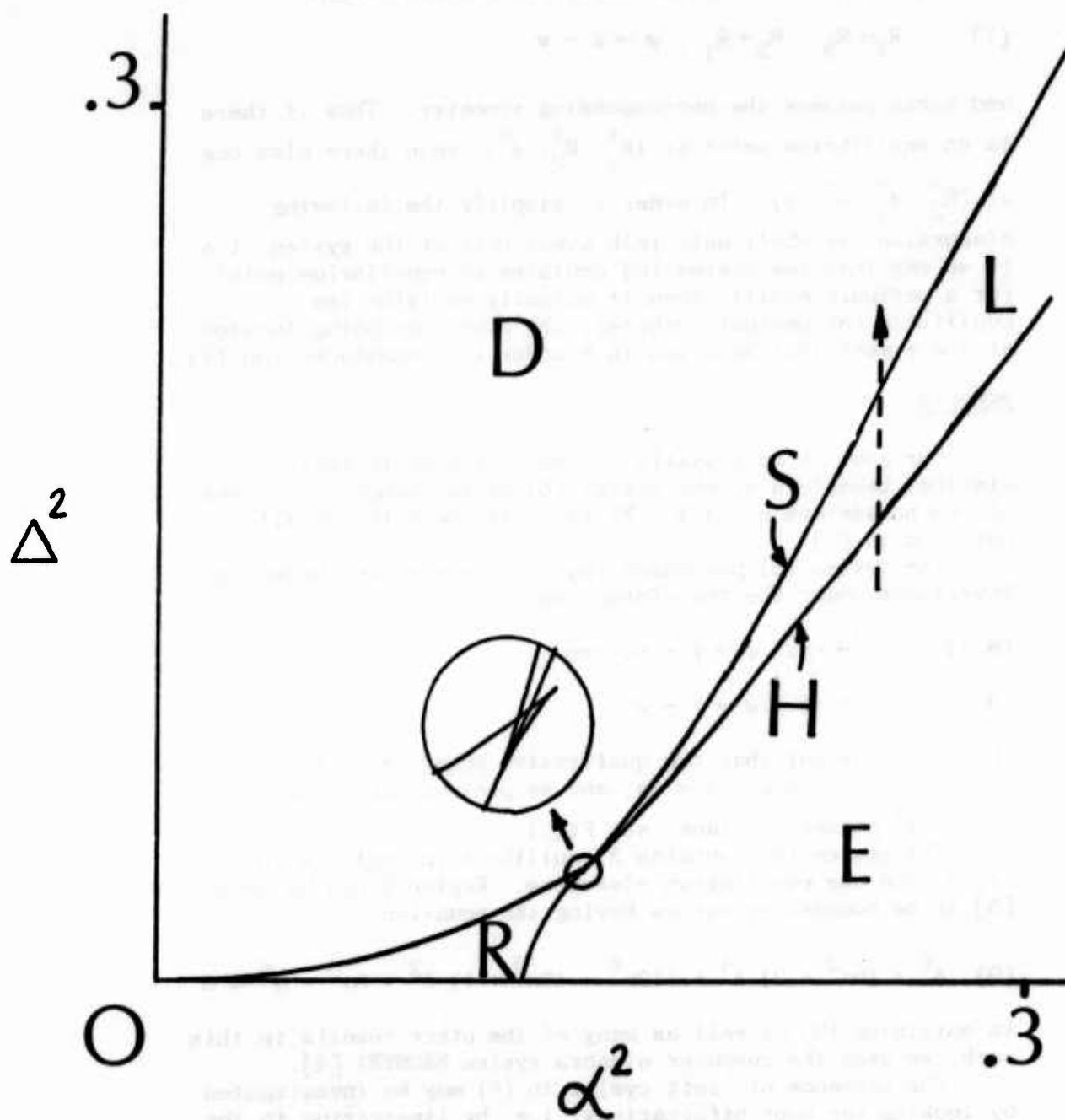
Fig.1. Bifurcation diagram for the slow flow (6):
        E = Region containing a single stable equilibrium point
        L = Region containing a stable limit cycle
        D = Region containing a stable libration orbit
        R = Region containing 3 equilibria
        H = Curve of Hopf bifurcations
        S = Curve of saddle-connection bifurcations

(10) $49 \Delta^8 + (266 \alpha^2 + 238) \Delta^6 + (88 \alpha^4 + 758 \alpha^2 + 345) \Delta^4$

$+ (-1056 \alpha^6 + 1099 \alpha^4 + 892 \alpha^2 + 172) \Delta^2$

$+ (-1152 \alpha^8 - 2740 \alpha^6 - 876 \alpha^4 + 16) = 0$

We showed that stable limit cycles occur in the region lying above curve H in Fig.1 by using center manifold theory and normal forms [3].

The region L of limit cycles is bounded on the other side by curve S, which we shall show later in this paper to involve saddle-connection bifurcations. As one crosses curve S, the limit cycle becomes a libration orbit, i.e. a closed trajectory in M, along which φ increases without bound. The nature of this bifurcation will be discussed in greater detail in the rest of this paper.

The region D of libration orbits is bounded by the curve (9) defining region R. As one crosses from region D into region R, a pair of equilibria are born on the libration orbit, which becomes a nonperiodic saddle-connection.

Fig.1 describes the behavior of the slow flow (6). In terms of the original eqs.(3), region E corresponds to 1:1 phase locking, region L to 1:1 phase entrainment, and region D to 1:1 phase drift. Thus as one moves in the parameter space along the dashed line (corresponding to holding the coupling strength α fixed while increasing the frequency difference Δ), the system (3) passes from phase locking to phase entrainment and then to drift.

THE ENTRAINMENT-DRIFT BIFURCATION

In order to better understand the nature of the bifurcation which results as one crosses from region L to region D in Fig.1, we display the results of some numerical integrations of the system (6). Figs.2-7 show these results in a portion of the phase space M in which

(11) $0 \leq R_1 \leq 4, \quad 0 \leq R_2 \leq 4, \quad -\pi \leq \varphi \leq \pi$

In each of Figs.2-7, $\alpha^2 = 0.25$, while Δ^2 varies from 0.15 in Fig.2 (in region E) to 0.21 in Fig.7 (in region D). Thus this sequence of views corresponds to moving along the dashed line in Fig.1. Fig.2 shows the asymptotic approach to steady state equilibrium, while Figs.3-7 show only the steady state periodic motions. Between Figs.2 and 3 a Hopf bifurcation has occurred corresponding to passage across curve H in Fig.1. Figs.3-6 show the gradual increase in size of the limit cycle, and the curve S in Fig.1 is crossed between Figs.6 and 7.
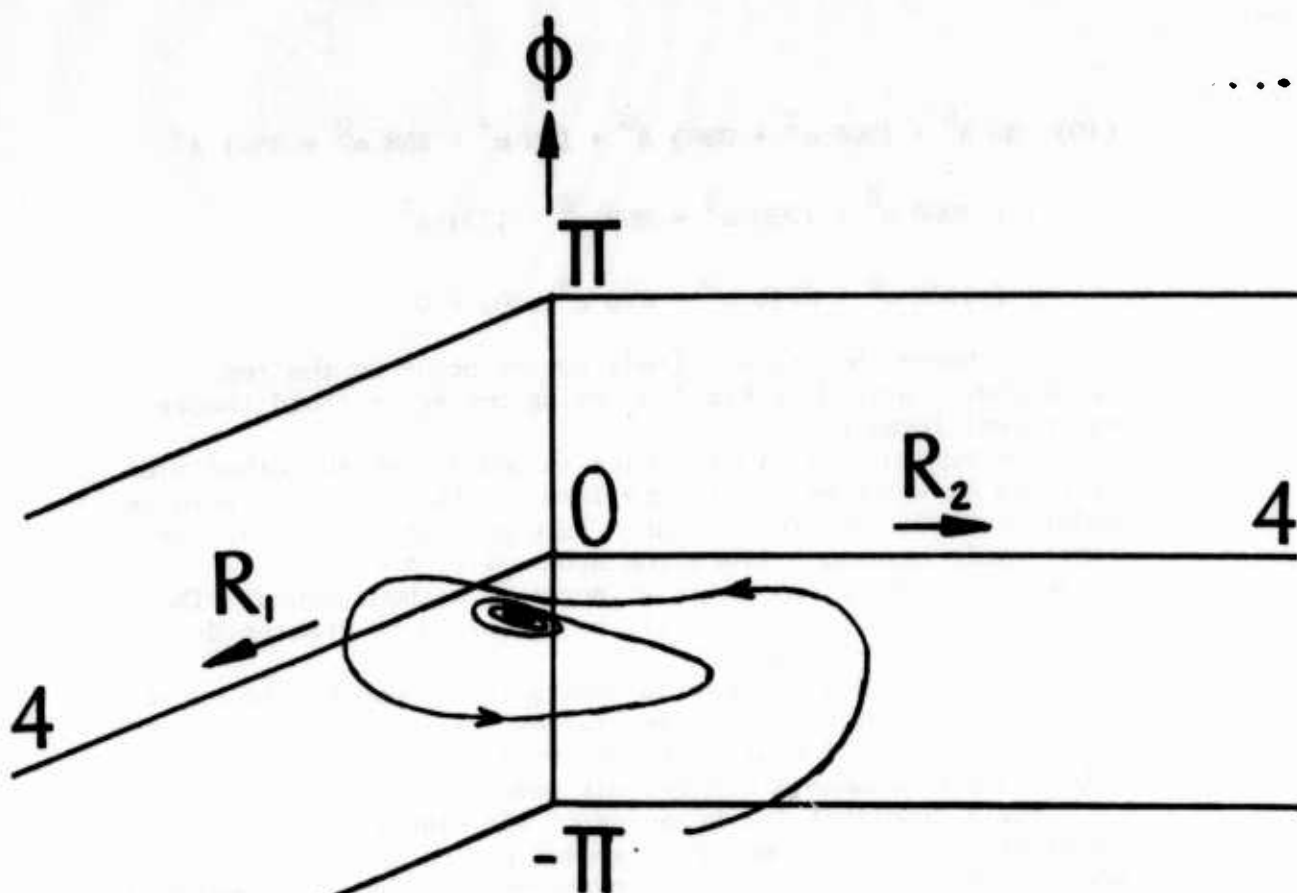
Fig.2. Numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.15$. Note the asymptotic approach to equilibrium. The labeling of the axes in this Figure applies to Figs.3-7 and 9-11.
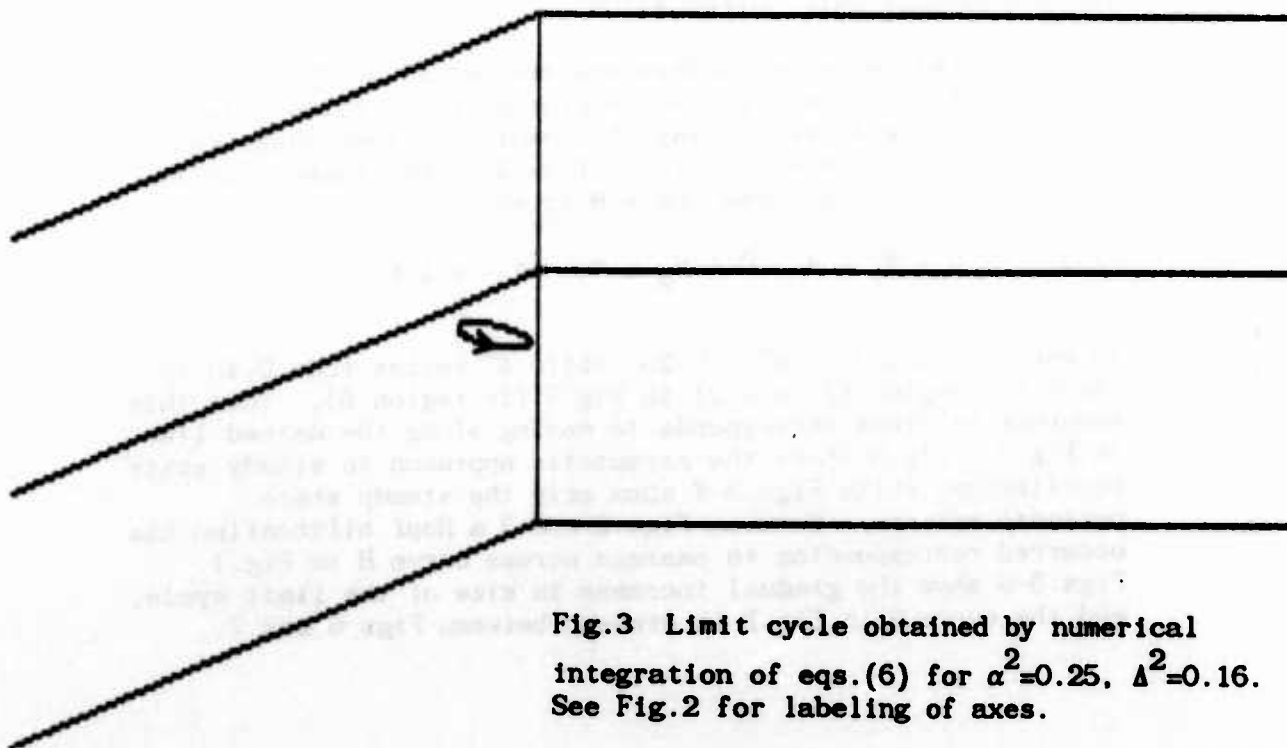
Fig.3 Limit cycle obtained by numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.16$. See Fig.2 for labeling of axes.
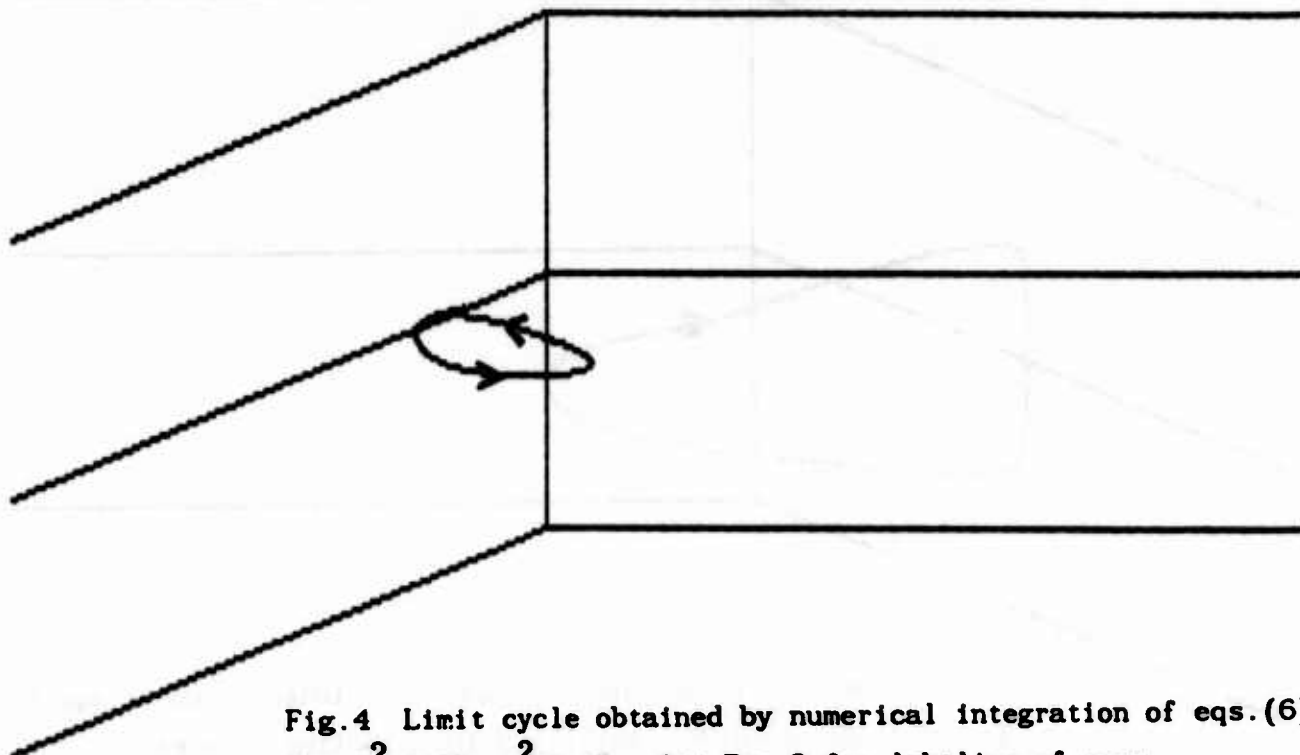
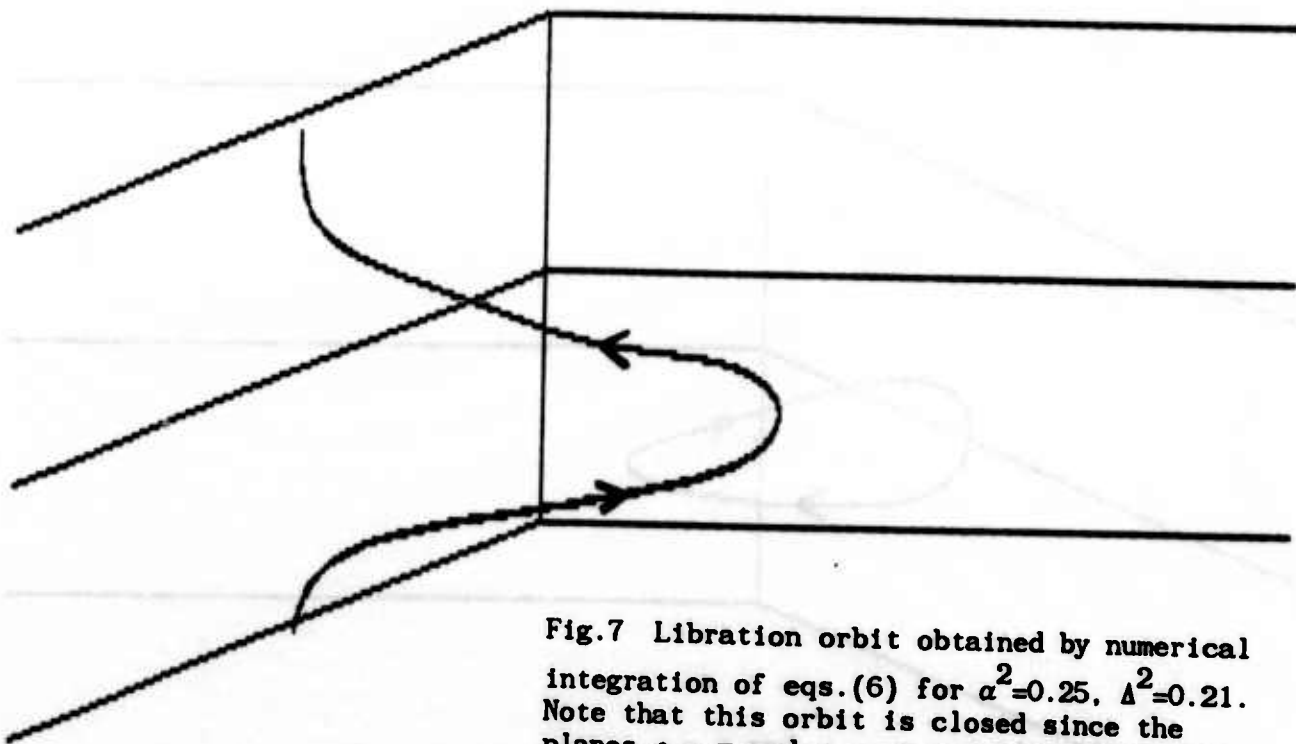Fig.4 Limit cycle obtained by numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.17$. See Fig.2 for labeling of axes.
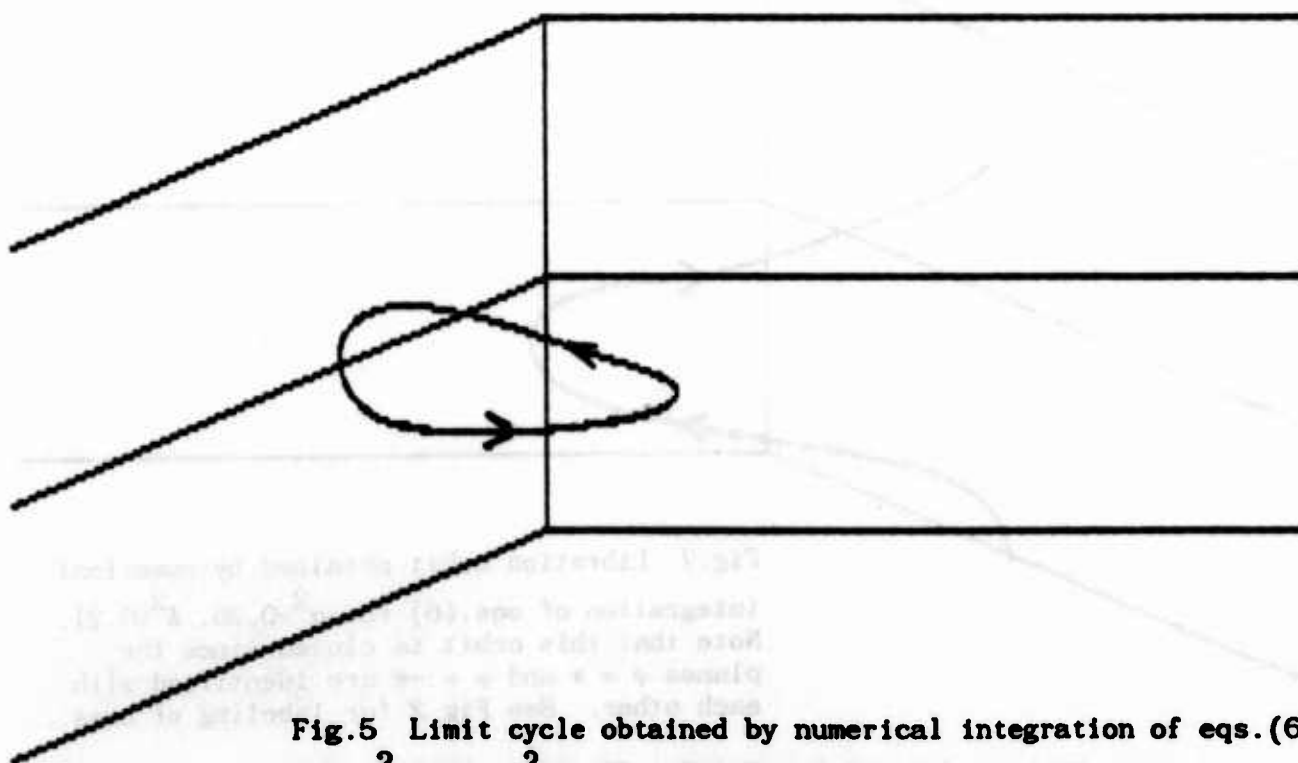
Fig.5 Limit cycle obtained by numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.185$. See Fig.2 for labeling of axes.
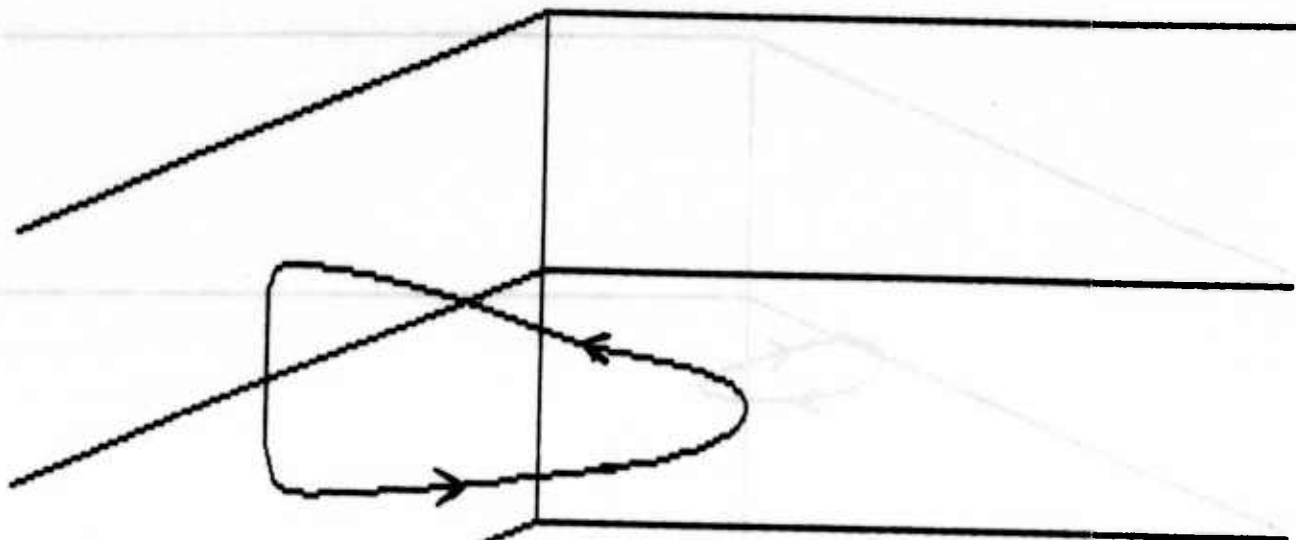
Fig.6 Limit cycle obtained by numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.20$. See Fig.2 for labeling of axes.

Fig.7 Libration orbit obtained by numerical integration of eqs.(6) for $\alpha^2=0.25$, $\Delta^2=0.21$. Note that this orbit is closed since the planes $\varphi = \pi$ and $\varphi = -\pi$ are identified with each other. See Fig.2 for labeling of axes.

We note from the numerical integrations that the transition between the limit cycle in Fig.6 and the libration orbit in Fig.7 involves a drastic change in the neighborhood of the surface $R_2 = 0$, but little change in the steady state

motions elsewhere. In Fig.6, the portion of the limit cycle near $R_2 = 0$ is a rapid motion at nearly constant R_1, i.e. a

jump down in φ. In Fig.7, on the contrary, the libration orbit jumps up in φ near $R_2 = 0$.

This observation leads us to examine the behavior of the system (6) when $R_2 = 0$. Since $R_2 = 0$ is a singular surface for

the system (6), we change independent variables from η to τ, where

$$(12) \qquad d\tau = \frac{d\eta}{R_2}$$

This transformation "blows up" the singularity at $R_2 = 0$ ([5],

§7.2), and while it reparametrizes the motion along the trajectories in M, it does not change the shape of the trajectories. Under (12), the system (6) becomes:

$$(13.1) \quad 2\frac{dR_1}{d\tau} = -R_1\, R_2 \left[\frac{R_1^2}{4} - 1\right] + \alpha\, R_2^2\, \sin\varphi$$

$$(13.2) \quad 2\frac{dR_2}{d\tau} = -R_2^2 \left[\frac{R_2^2}{4} - 1\right] - \alpha\, R_1\, R_2\, \sin\varphi$$

$$(13.3) \quad 2\frac{d\varphi}{d\tau} = \Delta\, R_2 + \alpha \left[\frac{R_2^2}{R_1} - R_1\right]\cos\varphi$$

We note that the surface $R_2 = 0$ is an invariant manifold of (13), which becomes when $R_2 = 0$:

$$(14.1) \qquad \frac{dR_1}{d\tau} = 0$$

$$(14.2) \qquad \frac{dR_2}{d\tau} = 0$$

$$(14.3) \quad 2\frac{d\varphi}{d\tau} = -\alpha\, R_1\, \cos\varphi$$

Eqs.(14) show that R_1 remains constant in time when $R_2 = 0$, and that only φ changes. Eq.(14.3) has the general solution

(15)
$$\tan\left[\frac{\varphi}{2} + \frac{\pi}{4}\right] = C\, e^{-\alpha\, R_1 \tau/2}$$

where C is an arbitrary constant. Fig. 8 shows the flow (14.3) on the φ-circle (since R_1 and R_2 are constant). Both Fig.8 and eq.(15) show that as $\tau \to +\infty$, $\varphi \to -\pi/2$, while as $\tau \to -\infty$, $\varphi \to +\pi/2$. This represents the previously referred to jump in φ, see Fig.9.

Thus the system (13) has two lines of nonisolated equilibria at $R_2 = 0$, $\varphi = +\pi/2$ and at $R_2 = 0$, $\varphi = -\pi/2$, shown as dashed lines in Fig.9. In order to better understand the bifurcation, we consider the nature of these equilibria.

We linearize eqs.(13) about the equilibrium

(16)
$$R_1 = R_1^o = \text{any}, \quad R_2 = 0, \quad \varphi = \pm\pi/2$$

and obtain the variational equations:

(17.1)
$$2\frac{d}{d\tau}\delta R_1 = -R_1^o \left[\frac{R_1^{o^2}}{4} - 1\right]\delta R_2$$

(17.2)
$$2\frac{d}{d\tau}\delta R_2 = \stackrel{-}{+}\, \alpha\, R_1^o\, \delta R_2$$

(17.3)
$$2\frac{d}{d\tau}\delta\varphi = \Delta\, \delta R_2 \pm \alpha\, R_1^o\, \delta\varphi$$

which have the general solution:

(18)
$$\begin{bmatrix}\delta R_1 \\ \delta R_2 \\ \delta\varphi\end{bmatrix} = k_1 \begin{bmatrix}1 \\ 0 \\ 0\end{bmatrix} + k_2 \begin{bmatrix}0 \\ 0 \\ 1\end{bmatrix} e^{\pm R_1^o \alpha\, \tau/2}$$

$$+ k_3 \begin{bmatrix}R_1^{o^2} - 4 \\ \pm 4\,\alpha \\ -2\,\Delta/R_1\end{bmatrix} e^{\stackrel{-}{+} R_1^o \alpha\, \tau/2}$$
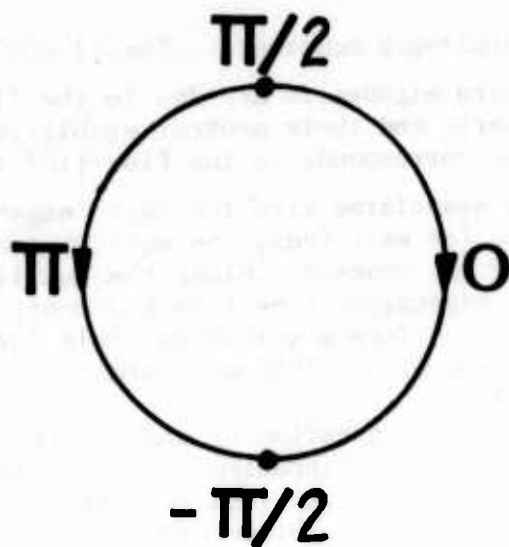
Fig.8. Flow on the φ-circle given by eq.(14.3). Arrows show the
direction of the flow. Dots represent the equilibria $\varphi = \pm \frac{\pi}{2}$.
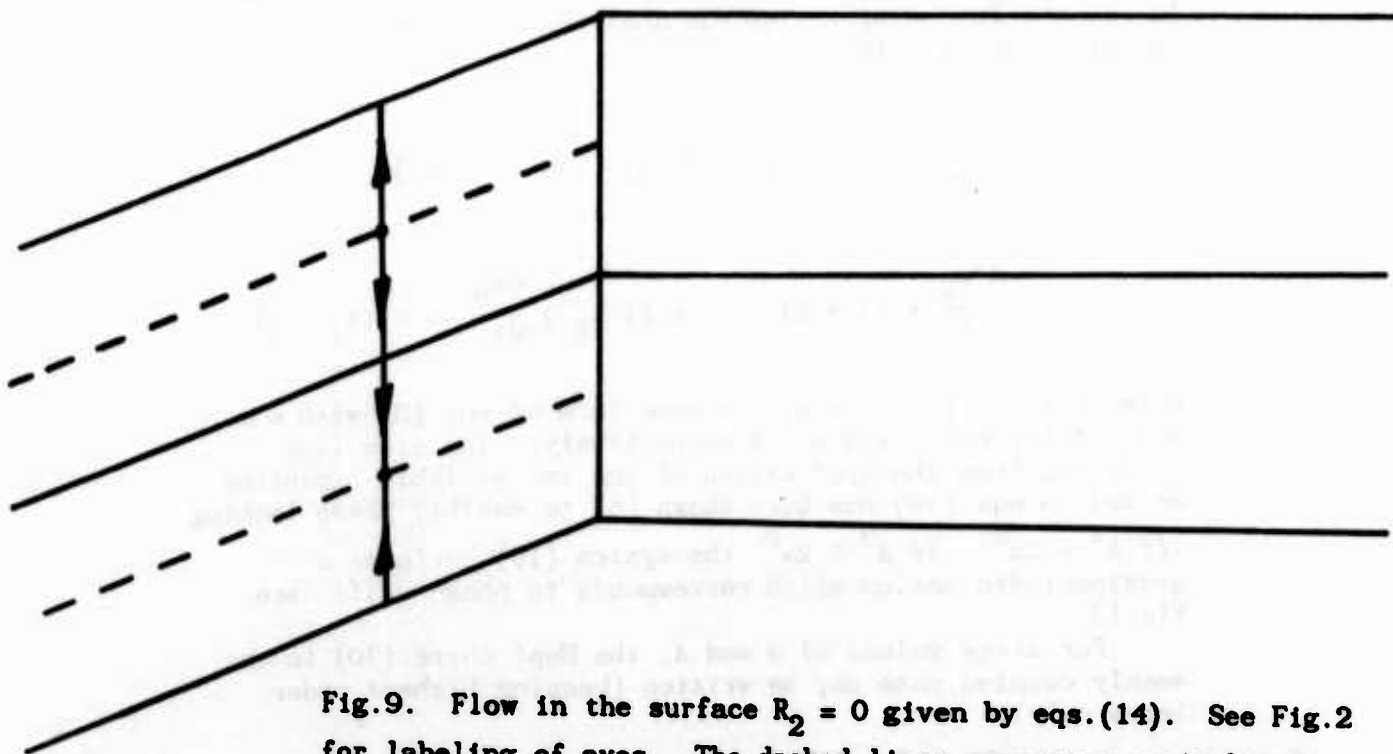
Fig.9. Flow in the surface $R_2 = 0$ given by eqs.(14). See Fig.2
for labeling of axes. The dashed lines represent nonisolated
equilibria. The lines with arrows represent the flow along
R_1 = constant, cf. Fig.8.

where the k_i are arbitrary constants. The (1 0 0) eigenvector and accompanying zero eigenvalue are due to the line of nonisolated equilibria and their neutral stability. The (0 0 1) eigenvector corresponds to the flow (15) in the $R_2 = 0$ plane. The motion associated with the third eigenvector permits approach to (or exit from) the equilibrium (16) from (or to) the rest of the space M. Since the equilibria are nonisolated, these eigendirections form a 2-manifold in M, see Fig.10. These surfaces form a stable manifold for the line of equilibria $\varphi = \pi/2$ and an unstable manifold for the line of equilibria $\varphi = -\pi/2$.

We return now to the question of the bifurcation from the limit cycle of Fig.6 to the libration orbit of Fig.7. In Fig.11 we superimpose Figs.6 and 7 on the eigenmanifolds of Fig.10. Fig.11 shows that the bifurcation involves a saddle-connection orbit.

The bifurcation may be further understood by projecting the limit cycle and libration motions on the $\varphi-\eta$ cylinder, see Fig.12.

CONCLUSIONS

It is interesting to compare the results of this study of weakly coupled van der Pol oscillators with those of an earlier study [6] of strongly coupled van der Pol oscillators. In [6] the following system was analyzed by using perturbations to $O(\epsilon)$:

$$(19.1) \qquad \frac{d^2 x_1}{dt^2} + x_1 - \epsilon (1-x_1^2) \frac{dx_1}{dt} = \alpha (x_2 - x_1)$$

$$(19.2) \qquad \frac{d^2 x_2}{dt^2} + (1 + \Delta) x_2 - \epsilon (1-x_2^2) \frac{dx_2}{dt} = \alpha (x_1 - x_2)$$

Note that eqs.(19) are of the same form as eqs.(3) with $\epsilon \alpha$ and $\epsilon \Delta$ replaced by α and Δ respectively. The slow flow resulting from the application of the two variable expansion method to eqs.(19) has been shown [6] to exhibit phase locking iff $\Delta^2 < 2\alpha^2$. If $\Delta^2 > 2\alpha^2$, the system (19) performs a quasiperiodic motion which corresponds to phase drift, see Fig.13.

For large values of α and Δ, the Hopf curve (10) in the weakly coupled case may be written (keeping highest order terms only):

$$(20) \quad 49 \, \Delta^8 + 266 \, \alpha^2 \, \Delta^6 + 88 \, \alpha^4 \, \Delta^4 - 1056 \, \alpha^6 \, \Delta^2 - 1152 \, \alpha^8 = 0$$
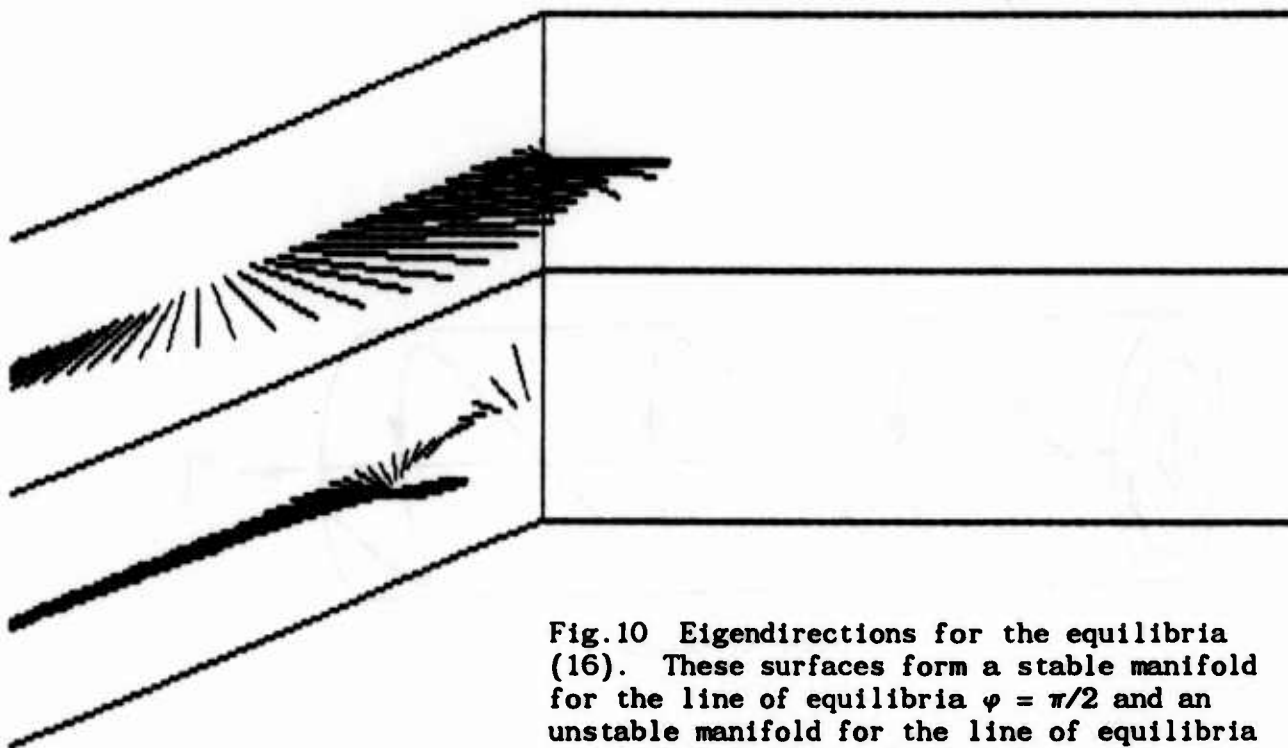
Fig.10 Eigendirections for the equilibria
(16). These surfaces form a stable manifold
for the line of equilibria $\varphi = \pi/2$ and an
unstable manifold for the line of equilibria
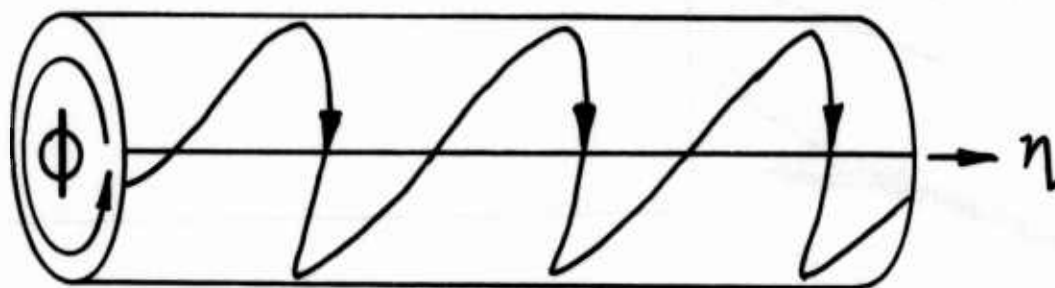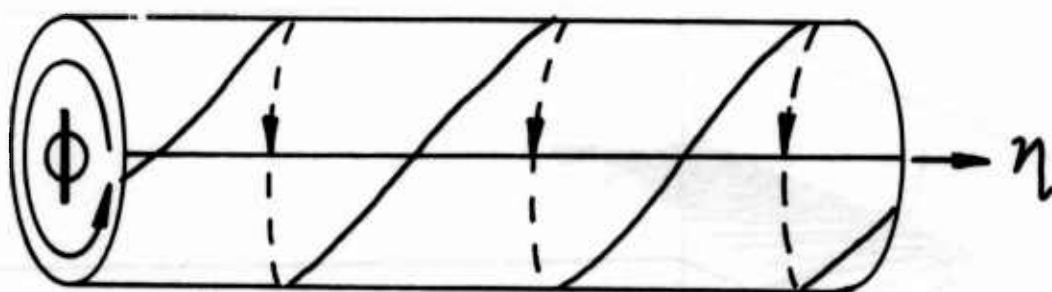$\varphi = -\pi/2$. See Fig.2 for labeling of axes.
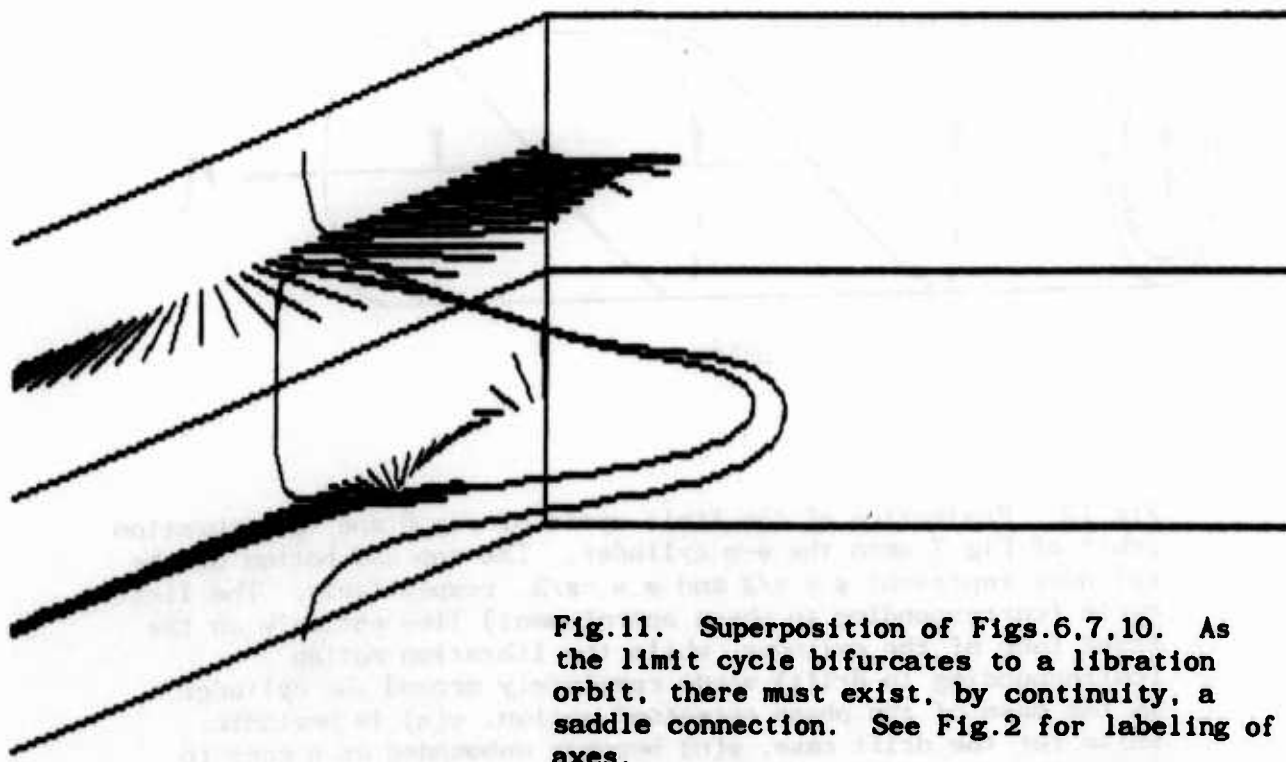
Fig.11. Superposition of Figs.6,7,10. As
the limit cycle bifurcates to a libration
orbit, there must exist, by continuity, a
saddle connection. See Fig.2 for labeling of
axes.

PHASE ENTRAINED

DRIFT

Fig.12. Projection of the limit cycle of Fig.6 and the libration orbit of Fig.7 onto the φ-η cylinder. The top and bottom of the cylinder represent $\varphi = \pi/2$ and $\varphi = -\pi/2$, respectively. The limit cycle (corresponding to phase entrainment) lies entirely on the front face of the cylinder, while the libration motion (corresponding to drift) winds completely around the cylinder. In the case of the phase entrained motion, $\varphi(\eta)$ is periodic, while for the drift case, $\varphi(\eta)$ becomes unbounded as η goes to infinity.
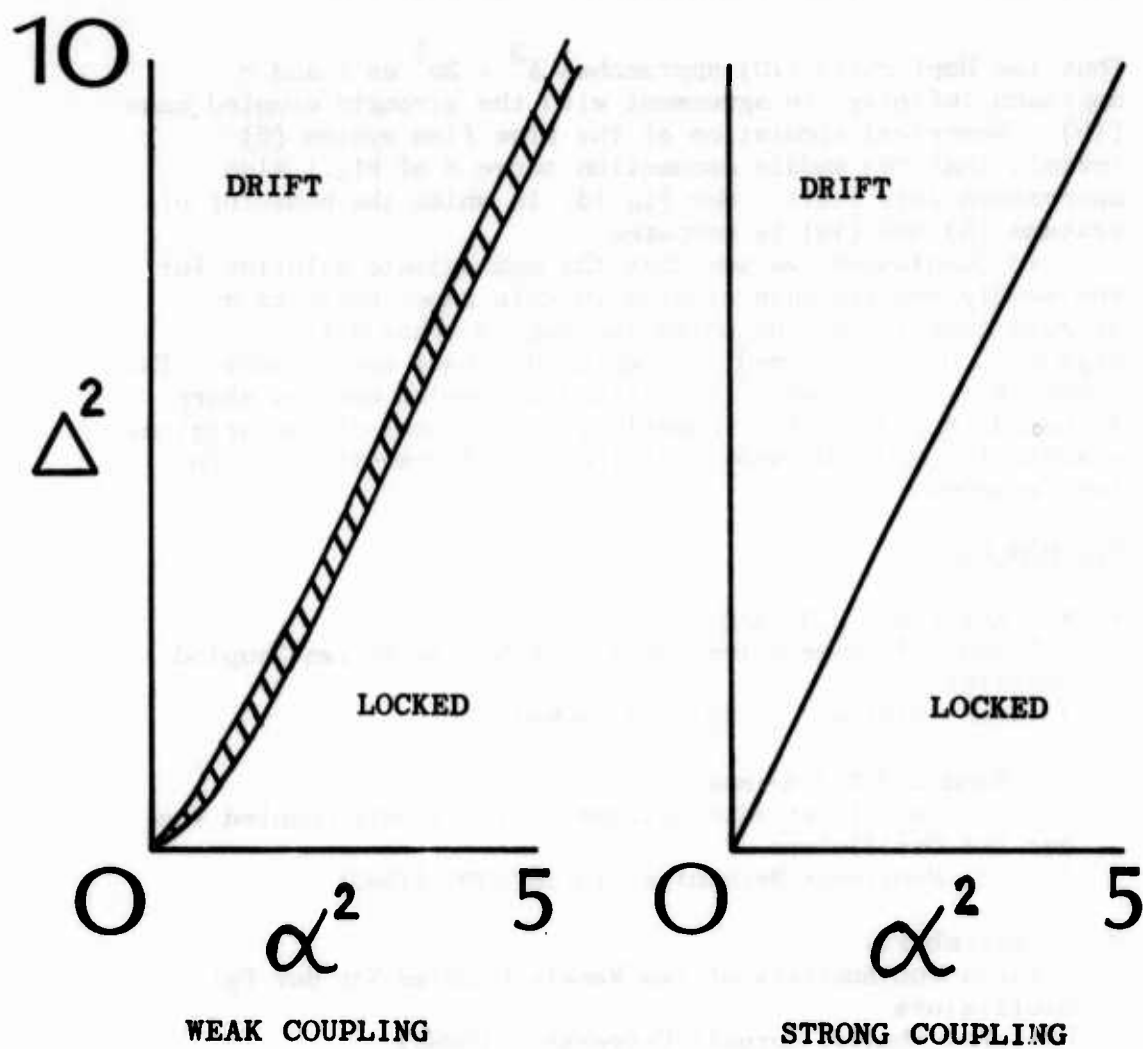
Fig.13.   Comparison of the dynamics of the weakly coupled system
(3) studied in this work with the strongly coupled system (19)
studied in [6].   The shaded region represents phase entrainment.

which may be factored to give

$$(21) \quad (\Delta^2 - 2\,\alpha^2)\,(\Delta^2 + 4\,\alpha^2)\,(7\,\Delta^2 + 12\,\alpha^2)^2 = 0$$

Thus the Hopf curve (10) approaches $\Delta^2 = 2\alpha^2$ as Δ and α approach infinity, in agreement with the strongly coupled case (19). Numerical simulation of the slow flow system (6) reveals that the saddle connection curve S of Fig.1 also approaches this limit. See Fig.13, in which the behavior of systems (3) and (19) is compared.

In conclusion, we see that the approximate solution for the weakly coupled case studied in this paper exhibits a gradual transition from phase locking to phase drift, separated by an intermediate region of phase entrainment. The comparable transition in the strongly coupled case is sharp. As noted in [3] and [6], numerical simulations of the original systems (3) and (19) show that the actual transition is in fact gradual.

REFERENCES

1. W.L.Keith and R.H.Rand
   1:1 and 2:1 Phase Entrainment in a System of Two Coupled Oscillators
   J. Math. Biology, 20:133-152 (1984)

2. R.H.Rand and P.J.Holmes
   Bifurcation of Periodic Motions in Two Weakly Coupled Van der Pol Oscillators
   Int. J. Nonlinear Mechanics, 15:387-399 (1980)

3. T.Chakraborty
   Bifurcation Analysis of Two Weakly Coupled Van der Pol Oscillators
   Doctoral thesis, Cornell University (1986)

4. R.H.Rand
   Computer Algebra in Applied Mathematics: An Introduction to MACSYMA
   Research Notes in Mathematics No.94, 181 pp.
   Pitman Publishing Inc. (1984)

5. J.Guckenheimer and P.Holmes
   Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields
   Applied Math. Sciences Vol.42, 453 pp.
   Springer-Verlag (1983)

6. D.W.Storti and R.H.Rand
   Dynamics of Two Strongly Coupled Van der Pol Oscillators
   Int. J. Nonlinear Dynamics, 17:143-152 (1982)

DESIGN AND IMPLEMENTATION OF A MULTIVARIABLE
ADAPTIVE CONTROL SYSTEM FOR AIRCRAFT/WEAPON APPLICATIONS

Pak T. Yip and David Ngo
Research and Development Branch
Fire Control Division
Fire Support Armaments Center
ARDEC , USAAMCCOM
Dover , NJ 07801

ABSTRACT. The design of a digital controller for weapon pointing and stabilization requires an accurrate plant and disturbance model in order to achieve optimal performance. However, these models are often unknown. Moreover, the disturbance and plant dynamics (such as helicopter turret motion) in weapon pointing systems are often time varying. Adaptive control, which permits on-line parameter identification update, can improve weapon pointing accuracy under these conditions. An adaptive weapon control module has been developed to perform minimum variance self-tuning control for advanced armament system applications. A weighted control scheme is used to obtain a stable control law for non-minimum phase plants which frequently arise in digital control implementations. The adaptive control module consists of parallel Intel 286/287 and 86/87 processors for identification and control updates. Preliminary test results using a laboratory test fixture will be discussed which demonstrate the performance capabilities of the system. Final evaluation of the performance of the adaptive weapon control module will be carried out in FY86 using a 30 mm automatic cannon mounted on a Cobra aircraft.

INTRODUCTION. Multivariable adaptive control [1] can significantly improve the performance of control systems where the plant or the disturbance model is not completely known. In particular, to achieve high performance in gun pointing control system for aircraft/weapon application where the disturbance model and aircraft dynamics are time varying, adaptive control may render greater pointing accuracy over broader performance envelopes.

To design a robust, high bandwidth digital weapon controller requires precise knowledge and analytical models of both high and low frequency system dynamics, including the platform disturbance environment. Typically, the high frequency system dynamics is not known a priori, or in many cases, the relevant model parameters will exhibit time varying behavior, resulting in poor performance, or even instability, in systems with fixed gain controllers. Some of the more commonplace situations include (1) variations in structural characteristics due to changes in weapon firing angle/engagement geometry, barrel heating and nonlinear effects, (2) variations in the disturbance frequency spectrum due to changes in air/ground speed, turbulence, ground roughness, firing rates, maneuvers, etc. and (3) variations in plant dynamics due to sensor/actuator failure or malfunction, and component degradation. It is under these conditions

that on-line adaptive control is used to maintain peak operating performance.

We will review the algorithms for adaptive control in the sense of minimum variance in the next section. Then we will describe the adaptive control with weighted control scheme. Following will be a general description of the adaptive control module configuration. Then the implementation of the adaptive control law will be presented and discussed.

ALGORITHMS OF ADAPTIVE CONTROL.   The basic structure for an adaptive control system is shown in Figure 1. The plant represents the physical weapon system which is being controlled including actuators, sensors, platform, weapon and disturbance dynamics. The vector Y consists of all platform sensor outputs which are processed by a variable parameter controller. This controller consists of a parameter identification update algorithm which identifies all model parameters from the measurement vector Y and the control vector u. The approach discussed in this paper uses UDU factored recursive least squares identification algorithm [2,3]. The model parameters are then passed to a control design algorithm which predicts the next plant output and computes the control input u to the plant for pointing and stabilization.

The plant dynamics of a typical weapon pointing system can be represented in discrete form by a system of equations expressed in observer canonical form

$$X(j+1)=AX(j) + Bu(j) \tag{1}$$

$$Y(j)=CX(j) \tag{2}$$

where

$$A = \begin{bmatrix} -a_1 & 1 & 0 & . & . & . & . & . & 0 \\ . & 0 & 1 & & & & & & . \\ . & & & & & & & & . \\ . & & & & & & & & . \\ . & & & & & & & & . \\ . & & & & & & & & 0 \\ . & & & & & & & & 1 \\ -a_n & 0 & . & . & . & . & . & . & 0 \end{bmatrix} \tag{3}$$

$$B = [b_1 \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ b_n]^T$$

$$C = [1 \ 0 \ . \ . \ . \ . \ . \ . \ . \ . \ . \ 0]$$

1020

and X(j) is the current state vector of the system, Y(j) the measurement and n the number of system states. To identify the system model parameters a_s and b_s one may rewrite the measurement equation as

$$Y(j) = \phi^T(j)\theta + e(j) \qquad (4)$$

$$\theta = [\, a_1 , \ldots , a_n , b_1 , \ldots , b_n \,]^T \qquad (5)$$

$$\phi(j) = [-Y(j-1), \ldots , -Y(j-n), u(j-1), \ldots , u(j-n)]^T \qquad (6)$$

$$e(j) = Y(j) - \hat{Y}(j) \qquad (7)$$

where $\hat{Y}(j)$ is the estimated measurement.

The parameters can then be estimated with the well established weighted recursive least square method given by the equations

$$\hat{\theta}(j) = \hat{\theta}(j-1) + L(j-1)E(j) \qquad (8)$$

$$E(j) = Y(j) - \phi^T(j-1)\hat{\theta}(j-1) \qquad (9)$$

$$L(j-1) = \frac{P(j-2)\phi(j-1)}{\alpha(j-1) + \phi^T(j-1)P(j-2)\phi(j-1)} \qquad (10)$$

$$P(j-1) = \frac{1}{\alpha(j-1)} [P(j-2) - L(j-1)\phi^T(j-1)P(j-2)] \qquad (11)$$

$$\alpha(j) = \alpha_0 \alpha(j-1) + (1-\alpha_0) \qquad (12)$$

where $\hat{\theta}(0)$, $P(-1) > 0$

$\alpha(0) = 0.95$

$\alpha_0 = 0.99$

In the factored version of the weighted recursive least squares algorithm, the covariance matrix P is factored as

$$P = UDU^T \qquad (13)$$

where U is a unit upper triangular matrix and D is a diagonal matrix. The matrices U, D, and L(j) are updated in such a way that the

positive definite property of the diagonal matrix D is maintained. This is essential to maintain the computational stability in the presence of truncation and round off errors associated with microprocessor implementation.

Given the estimate $\hat{\theta}(j)$ of the plant parameter vector $\theta(j)$, one may use a pole placement procedure such as the one described in reference [4] or one may compute directly the minimum variance control which can be shown to be equivalent to solving the system of linear equations

$$Y^*(j+d) = \phi^T(j)\hat{\theta}(j) \qquad (14)$$

where Y^* is the desired reference trajectory and d is the delay between the input and plant output.

ADAPTIVE CONTROL WITH WEIGHTED CONTROL SCHEME. Minimum variance control will generally exist poor performance when the physical system has unstable zeros. In this case, it is often helpful to use a weighted control scheme to get good adaptive control. The ARMAX model representing the system given in equation (4) is

$$A(q^{-1})Y(j+1) = B(q^{-1})u(j) + C(q^{-1})V(j+1) \qquad (15)$$

$$A(q^{-1}) = 1 + a_1 q^{-1} + a_2 q^{-2} + \ldots\ldots\ldots a_n q^{-n} \qquad (16)$$

$$B(q^{-1}) = b_1 + b_2 q^{-1} + b_3 q^{-2} + \ldots\ldots b_n q^{-n-1}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + c_2 q^{-2} + \ldots\ldots\ldots c_n q^{-n}$$

where $V(j+1)$ is a white noise process,

q^{-1} is the time delay operator.

Minimizing the weighted control index

$$J = \text{expectation of } [|Y(j+1) - Y_{ref}(j+1)|^2 + \rho^2 |u(j)|^2] \qquad (17)$$

at time j gives a control law that requires

$$[B(q^{-1}) + \frac{\rho^2}{b_0} A(q^{-1})] \text{ to be stable instead of } B(q^{-1}).$$

Where b_0 is the weight on the $u(j+1)$.

Hence adaptive control with weighted control scheme provides stable control for systems that have unstable zeros.

An alternate form of implementing the weighted control scheme is to replace $Y(j+1)$ by

$$\overline{Y}(j+1) = Y(j+1) + \gamma u(j) \tag{18}$$

$$Y_{ref}(j+1) = \overline{Y}(j+1) - V(j+1) \tag{19}$$

in the usual minimum variance adaptive control law where $\gamma = \rho^2/b_0$.

From equations (15), (18), and (19)

$$A(q^{-1})\overline{Y}(j+1) = [B(q^{-1}) + \gamma A(q^{-1})]u(j) + C(q^{-1})V(j+1) \tag{20}$$

$$[B(q^{-1})+\gamma A(q^{-1})]u(j) = [A(q^{-1})-C(q^{-1})]V(j+1)+A(q^{-1})Y_{ref}(j+1) \tag{21}$$

In order to achieve good performance, choose such that the poles of $[B(q^{-1})+ \gamma A(q^{-1})]$ must lie inside the unit circle.

However, the weighted control scheme generally creates steady state errors in the system outputs which are unacceptable for tracking/pointing applications. One way to deal with this steady state error problem is to insert a frequency shaping at the input and the output of the system such that the overall control becomes one of the model following adaptive control.

The frequency shaping $N(q^{-1})/D(q^{-1})$ changes the modified output

$$\overline{Y}(j+1) = \frac{N(q^{-1})}{D(q^{-1})} Y(j+1) + \gamma R(q^{-1})u(j) \tag{22}$$

where $N(.)$, $D(.)$, and $R(.)$ are monic.

Now, it is required that $N(.)$ and $D(.)$ have stable roots and the roots of $[N(q^{-1})B(q^{-1})+\gamma R(q^{-1})A(q^{-1})D(q^{-1})]$ are stable. The $\gamma R(q^{-1})$ term is used to stabilize the unstable zeros of $B(q^{-1})$.

ADAPTIVE CONTROL MODULE CONFIGURATION. The parallel architecture of the adaptive control processor module is shown in Figure 2. The Intel 8086/8087 contains all analog to digital (A/D) and digital to analog (D/A) circuitry and is programmable for easy scaling and manipulation of data. It operates as a fast system for controller/prediction operation. The Intel 80286/80287 processor board with 1 megabyte of RAM serves as the master processor and also supports floating point implementation of the factored weighted recursive least squares identification update and control law update algorithms. This processor will compute control law updates and identification updates every 8500 μ sec for a 12 state, 2 input - 2 output system.

ADAPTIVE CONTROL LAW IMPLEMENTATION. The controller software is developed with a VAX 11/780 host computer. The developed program is downloaded to the adaptive control processor module which will output control commands to a test fixture, in this case an inertia wheel. Angular measurement of the system and reference command are fed to the control module as inputs.

The test fixture consists of two DC torque motors which drive the inertia wheel and a resolver which measures the shaft angle of the wheel. Through an electronic circuitry, the resolver signal is amplified and demodulated before entering the A/D converter in the control module. The analog control command from the module actuates the system through a torque drive amplifier in the electronic circuitry.

Accounting for the effect of a zero order hold in the A/D converter, the open-loop response of the test fixture in z-domain is

$$Y(z) = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} U(z) \tag{23}$$

The weighted control adaptive control design is implemented with the inertia wheel test fixture. A square wave sequence is fed to the system as a reference command signal. The sampling rate and the control rate are set to 100 Hz. The identification update and control law update are processed at 10 Hz. The initial values of the parameters are estimated from the open-loop data of the system. Figure 3 shows the input square wave command Y_{ref}, the control signal u, and the output Y of the system. Here the output approaches the reference command with good stability property. In this case, the weighting factor γ in equation (18) is tuned to -2.85. When the factor γ is set to -5.0, Figure 4 shows the steady state error of the output Y from the reference signal Y_{ref}.

DISCUSSION. We have reviewed the minimum variance adaptive control algorithm as well as the weighted recursive least square identification method. Also, we have described the weighted control scheme to get around the non-minimum phase problem and talked about the model following/frequency shaping which will help us deal with the tracking problem. The basic configuration of the control module using parallel Intel 286/287 and 86/87 processors was presented. Some preliminary test results of using the weighted control adaptive control algorithm interfaced with the laboratory test fixture were shown to have good stability property. Based on computer simulations and the results of this implementation study, we strongly believe that digital adaptive control for aircraft/weapon pointing applications is a practical and viable option. A final evaluation of the performance of the adaptive weapon control module using a 30 mm automatic cannon mounted on a Cobra aircraft has been scheduled for FY86.
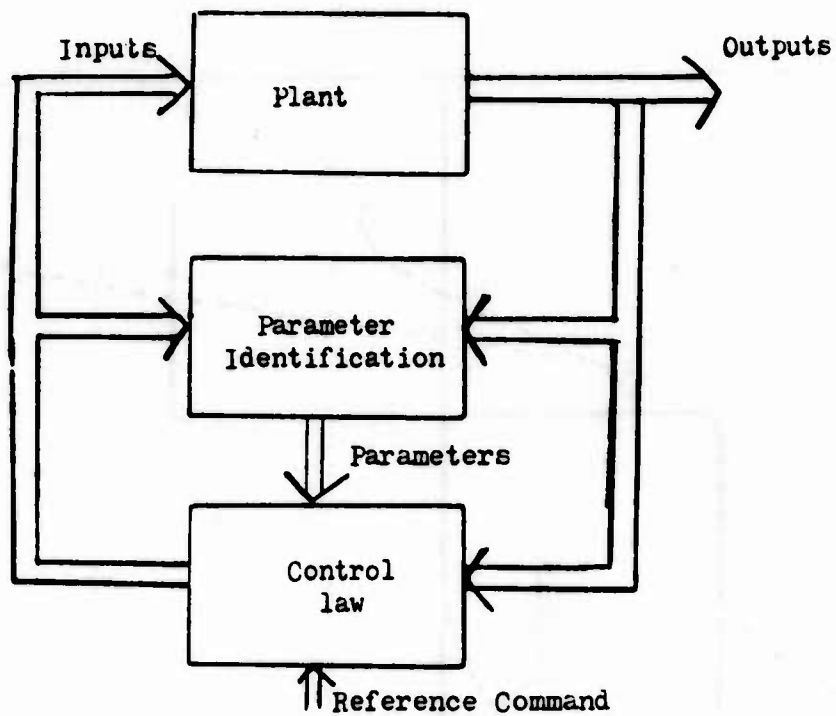
1024

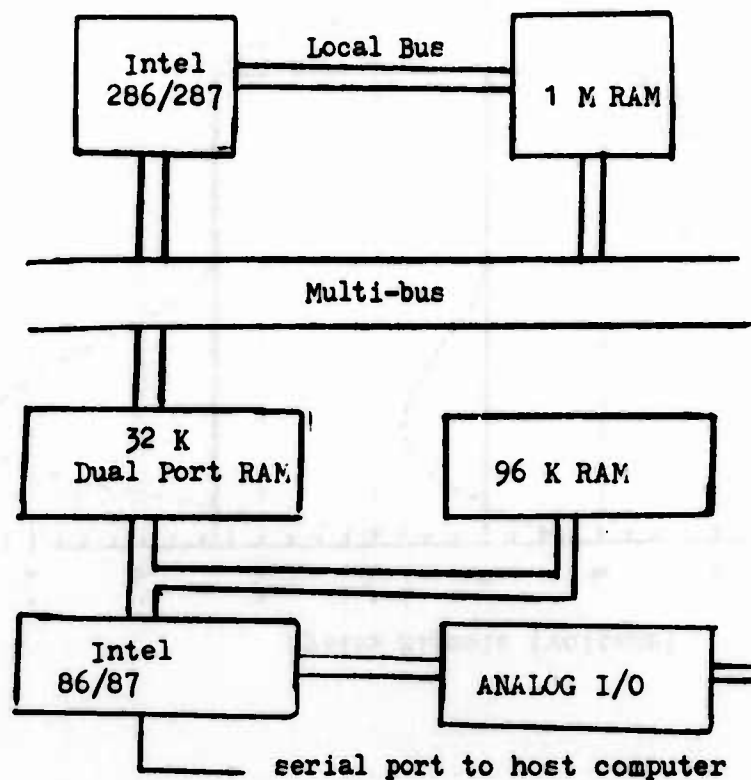Figure 1. Basic Structure of Adaptive Control

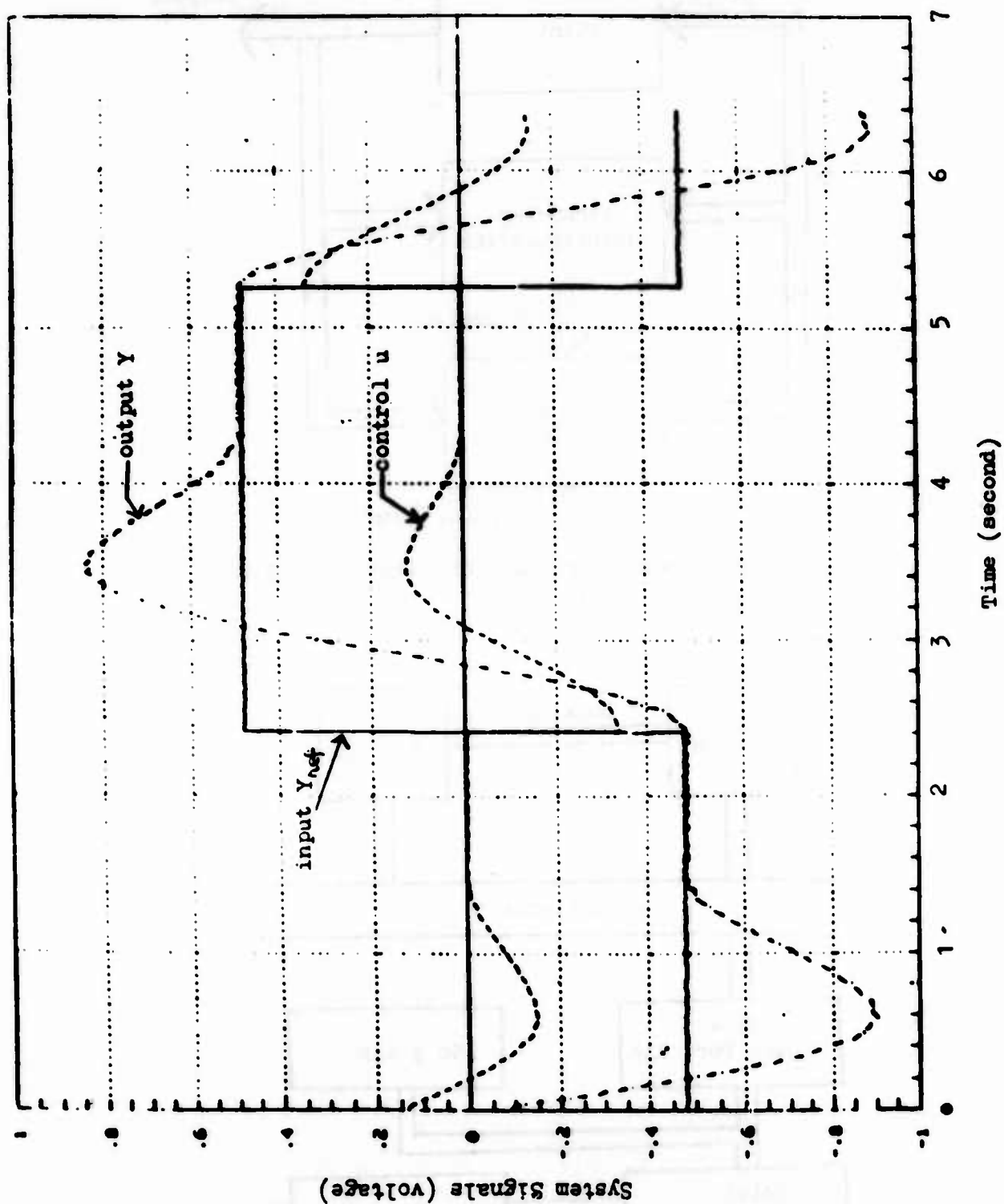Figure 2. Architecture of Adaptive Controller Module

1025

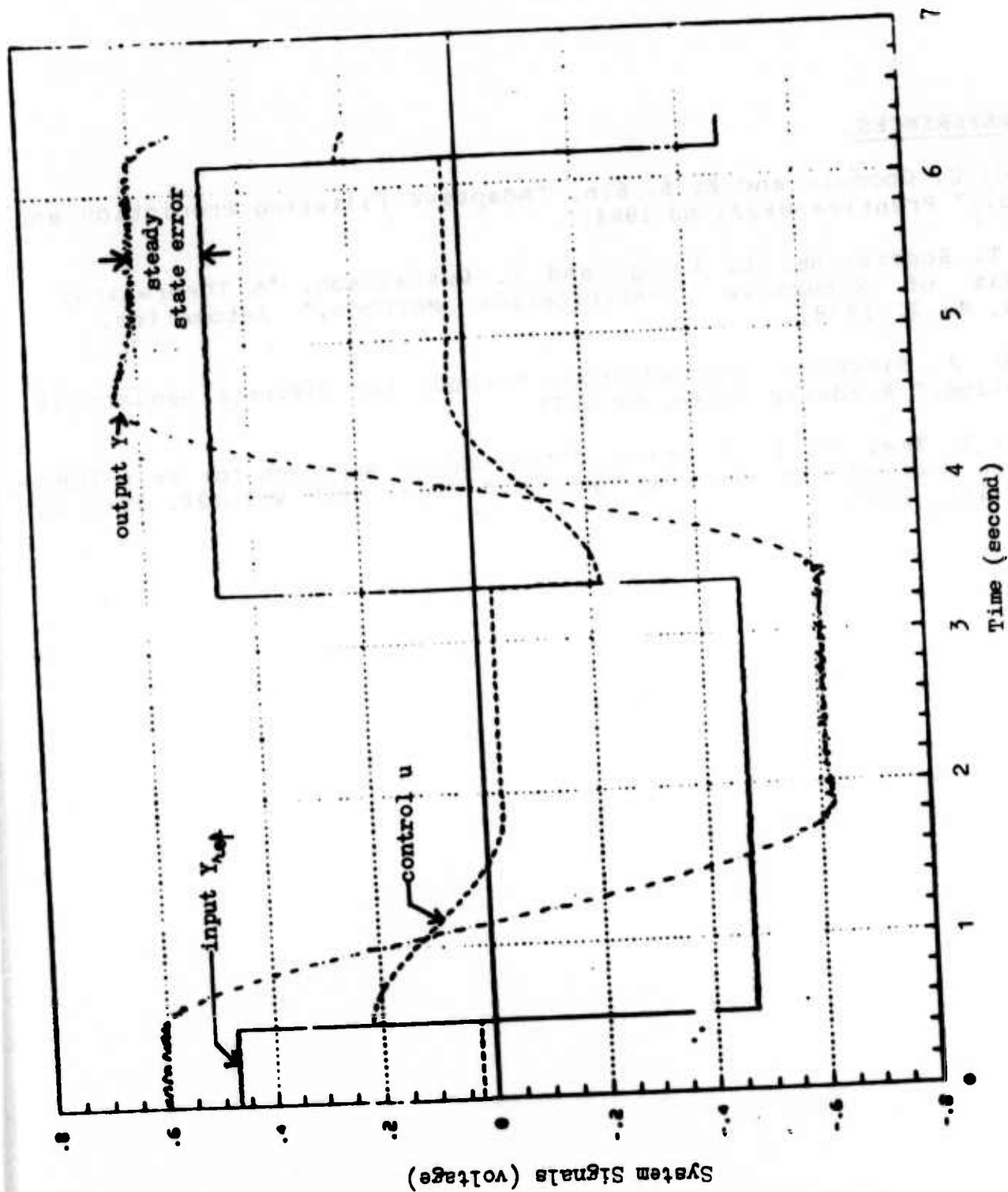Figure 3. The Time Response Of The System With Weighting Factor $\gamma = -2.85$

Figure 4. The Steady State Error In Output With Weighting Factor $\gamma = -5.0$

REFERENCES.

[1] G. C. Goodwin and K. S. Sin, "Adaptive Filtering Prediction and Control," Prentice-Hall, NJ 1984.

[2] T. Soderstrom, L. Ljung, and I. Gustavsson, "A Theoretical Analysis of Recursive Identification Methods," Automatica, vol.14, No.3, 1978.

[3] G. J. Bierman, "Factorization Methods for Discrete Sequential Estimation," Academic Press, NY 1977.

[4] Y. T. Tsay and L. S. Shieh, "State-Space Approach for Self-Tuning Feedback Control with Pole Assignment," Proc. IEE, Vol.128, Part D, pp. 93-101, 1981.

DOMAIN CONTRACTIONS IN FINITE-DIFFERENCE COMPUTATIONS OF POISSON'S EQUATION BY MEANS OF INFINITE NETWORK THEORY

A.H. Zemanian

Department of Electrical Engineering

State University of New York at Stony Brook

Stony Brook, N.Y. 11794-2350

and

T.S. Zemanian

Department of Chemical Engineering

Cornell University

Ithaca, N.Y. 14853

ABSTRACT. A finite-difference analysis of the exterior problem for Poisson's equation has an infinite-electrical-network analog. Concurrently with the recent numerical attacks upon that problem, a theory for infinite electrical networks has been developing and provides an alternative approach. Examples taken from the flow of petroleum into an oil well and from well-logging with the resistivity method illustrate the unusual kinds of connections at infinity that an infinite electrical network can have. A technique is described for solving the exterior problem for a three-dimensional anomaly contained within a sphere S. Infinite electrical network theory yields a driving - point conductance matrix Y that replicates the effect of a spherical grid arising from a finite-difference analysis for the region exterior to S. This allows the domain of analysis to be contracted to S and its interior. The resulting analysis remains exact in the sense that the solution for the spherical grid within and on S coupled with Y is precisely the same as that for the infinite spherical grid.

I. INTRODUCTION. The purpose of this paper is to briefly survey a relatively new area of research, namely, the search for the solutions for certain kinds of infinite electrical networks, and to point out how this can lead to better computational methods for the exterior problem for Possion's equation. That problem arises when the domain of interest is infinite in extent. During the past decade or so, the exterior problem for the numerical analysis of partial differential equations has been the subject of a substantial amount of research (Bayliss et al 1982, Canuto et al 1985, Engquist and Majda 1977, Fix and Marin 1978, Goldstein 1979 and 1982, Gustafsson and Kreiss 1979, Hagstrom and

Keller 1984 and in press). The common procedure in all these works is the introduction of an artificial boundary that renders the domain finite and then a search for an appropriate boundary condition on that boundary.

For Poisson's equation (as well as for other partial differential equations), there is another way to solve this problem. It is well known that a five-point finite-difference approximation of Laplace's operator can be represented by a rectangular grid of positive resistors. As a result, the exterior problem for Poisson's equation has an infinite-electrical-network analog. So, instead of inserting an artificial boundary, one might search for the solution to the infinite network. Concurrently with the appearance of the aforementioned literature on the exterior problem, a theory for infinite electrical networks has been developing. Almost all of the earlier work was directed toward existence and uniqueness theorems but more recently solutions have been found for certain infinite gridlike resistive networks (Zemanian 1978, 1981, 1982, 1985c; Zemanian and An 1985; Zemanian and Subramaniam 1983; Zemanian and Zemanian in press). These are solution for rectangular, cylindrical, and spherical resistive grids of both the grounded and ungrounded varieties. It is required however that the grid's parameters vary with only one or at most two coordinates. This approach to the discretized exterior problem has also been applied to polarized electromagnetic waves (Zemanian 1985a,1985b), but now an RLC grid ensues. In every one of these applications an operational calculus can be exploited to obtain a rapid means of computing the solution.

In the next section we examine one aspect of infinite-electrical-network theory in its current state of development. That aspect concerns the variety of conditions at infinity that may arise in different applications. Infinite network theory leads naturally to a consideration of these conditions, whereas the analogous problem does not seem to have been taken up in the numerical approaches to the exterior problem listed above.

In the last section, we consider the operator $\nabla \cdot \sigma \nabla$ in spherical coordinates for a medium whose parameter σ may vary only with the radial and latitudinal coordinates r and θ but not with the logitudinal coordinate ϕ. We show how an exact solution for the corresponding spherical grid coupled with an appropriate operational calculus yields a fast computational method for solving Pisson's equation in all of three-dimensional space for the case where the medium is uniform except in a bounded three-dimensional region.

II. CONDITIONS AT INFINITY. Strange things can happen in an infinite electrical network. We present three examples to show how some of these peculiarities arise naturally out of practical applications.

II.1. Example. For finite domains the resistive-network analogs of Poisson's equation have finite nodes, that is, no more than a finite number of branches are incident to any node. For infinite domains however this need not be so. Figure 1(a) shows two perfect conductors that extend out to infinity. The electrostatic potential in the region between the conductors is determined by Laplace's equation, a five-point discretization of which yields a resistive grid whose nodes away from the conductors each have four incident branches. However, the nodes of any conductor are all shorted together, and thus there are two nodes having an infinity of incident branches. This is shown in Figure 1(b).

Unfortunately, Kirchhoff's current law need not hold at an infinite node (Zemanian 1974 page 275). Situations can arise where, if one adds all the incident currents flowing toward an infinite node, as well as all the incident currents flowing away from that node, one finds a total of one ampere flowing toward it and zero amperes flowing away. A perspective on this anomaly can be obtained by examining a sequence of networks that approach the infinite network with the branches that carry currents away from the node being added one by one. What can happen is that those currents individually tend to zero while their sum remains one. In the limit each such branch carries zero current, and so the infinite series of those zero values must equal zero, not one. Nonstandard analysis may provide a means of avoiding this paradox by allowing the infinite series of infinitesmal currents to add up to one. However, the approach used in our rigorous theory of infinite networks is to allow the nonsatisfaction of Kirchhoff's current law at certain infinite nodes (Zemanian 1974), the "nonrestraining" nodes (Zemanian, 1986).

II. 2. Example. The second example concern the flow of petroleum into an oil well through a completion zone of fractured rock surrounded by porous virgin rock. See Figure 2(a). The pressure P is assumed to be governed by Laplace's equation $\nabla \cdot \nabla P = 0$ with the orifice of the oil-well pipe as a sink for the flow. The pipe's surface is a boundary with a null Neumann condition. Given the flow into the orifice, the pressure variation is desired. A discretization in cylindrical coordinates yields a cylindrical grid, a cross-section of which through the pipe's centerline is shown in Figure 2(b). The encircled arrows therein represent flow sinks (i.e. current generators) whose values are known.

For finite networks the extraction of currents at some nodes must be balanced by the injection of currents at other nodes. For infinite networks this principle is upheld by assuming (often tacitly) that the current is returned at infinity to the network. -- But where at infinity? Does it matter? Sometimes it does and sometimes it doesn't (Zemanian 1986). If the series of (positive) resistance values along a one-ended path P_1 to infinity converges -- as can happen if the rock's permeability increases fast enough in some direction toward infinity -- and if that series for another such path P_2 diverges and moreover P_2 is not "shorted at infinity " to other such paths (to be more precise we should refer to "pathlike extremities" (Zemanian 1986 Section II)), then it truly does matter where at infinity the returns for the sinks are located. This must be specified if the pressure is to be uniquely determined.

In general, one has to specify which paths to infinity are "shorted together at infinity", which paths to infinity are left "open-circuited at infinity", (Zemanian 1975), and which voltage (i.e., pressure) or current (i.e., flow) sources are connected at infinity to which such paths (Zemanian 1986). All this requires a thorough reworking of standard electrical network theory and the introduction of some new concepts such as nodes and branches that are connected to certain extremities of the network at infinity. Standard infinite-graph theory does not allow such entities, and so that too must be reworked.

II. 3. Example. Figure 3(a) shows a well-logging tool for a normal-log resistivity measurement. A current probe injects an electrical current h into the rock surrounding a borehole and that current is extracted at a return probe on the earth's surface. A potential probe adjacent to the down-hole current probe measures the resulting electrical potential in the borehole. The probes are mounted on an insulated sonde but the rest of

the borehole is filled with drilling mud, which has a high electrical conductivity. The borehole is assumed to extend infinitely above as well as infinitely below the sonde. A discretization of Laplace's equation in cylindrical coordinates yields a cylindrical grid, whose cross-section on a plane through the borehole is shown in Figure 3(b) only for the section to the right of the borehole. Even though the earth's surface is taken to be infinitely far from the down-hole probes, we may have a situation that may be modeled by a conductivity that increases rapidly as a path is traced upward indefinitely. Indeed, such a model may be suitable if there is a highly (more precisely, a perfectly) conducting overburden (e.g., a salt marsh). For such a model, the return probe of the current source must be specified as being connected to the grid's "extremities in the upward direction" -- and not elsewhere at infinity.

As a further idealization, we might take it that the drilling mud is also perfectly conducting. In this case, the borehole above the sonde is at a constant potential, the same potential as the overburden. This means that there is an infinite node n_u, which is moreover shorted to the "extremities in the upward direction". This is illustrated in Figure 3(c). There is another infinite node n_l representing the borehole below the sonde, but this one is not shorted to any extremities at infinity.

Still another complication that arises with infinite electrical networks concerns the application of Kirchhoff's voltage law around infinite loops. Such loops must now be considered because of the possible connections at infinity. Our theory designates those loops (that is the finite loops and also the "perceptible extended" loops (Zemanian 1986 pages 33-34)) on which Kirchhoff's voltage law holds, and those loops (the "imperceptible extended" loops) on which it need not hold.

III. DOMAIN CONTRACTIONS AROUND A THREE-DIMENSIONAL ANOMALY.
We now describe how the theory of infinite electrical networks can be used to generate a fast computational method for solving a particular exterior problem.

III. 1. The Model. We consider again the flow of petroleum into an oil well, but this time we let the completion zone be a single perforated one. Such a completion zone is made by firing a projectile through the side of the oil-well pipe to fracture a balloon-shaped region of rock adjacent to the borehole. This is illustrated in Figure 4. Here too, the borehole is taken to be infinitely long in both the upward and downward directions. The flow of petroleum into the well is given, and the pressure variation in the surrounding rock is desired.

That pressure P is determined by Poisson's equation

$$\nabla \cdot (\sigma \nabla P) = -H \tag{1}$$

where H represents the flow sources measured positively as flow injected into the medium and the parameter σ is the ratio of medium permeability divided by the petroleum's viscosity. In taking (1) as the governing equation, we are actually assuming that the fluid is incompressible, Newtonian, and saturates the porous rock and that gravitational and inertial effects are negligible. As for σ, it is taken to be a constant outside the borehole and completion zone but varies three-dimensionally and at considerably higher values within the completion zone. The interior of the borehole's pipe is not part of the problem's domain.

III. 2. The Spherical Grid.

We choose a spherical coordinate system (r, θ, ϕ) whose polar axis ($\theta = 0$ and $\pi/2$) coincides with the centerline of the borehole and whose origin is directly adjacent to the center of the perforation. This too is indicated in Figure 4. Upon choosing a discretization of the coordinates and computing the corresponding finite differences for the partial derivatives of (1), we obtain an infinite spherical grid of positive resistors that are fed by current sources at several points of the pipe's perforated orifice. (See Zemanian and Zemanian (in press) for the specifics of these manipulations.) The returns for those current sources are at infinity; all the extremities of the grid are taken to be shorted together at infinity. The grid's nodes lie upon the concentric sphere S_1, S_2, S_3, ... shown in Figure 4 and are distributed upon each sphere as shown in Figure 5. However, a few nodes lie within the borehole. Some of these, namely, those on S_1 are simply removed along with their incident branches. The remaining nodes - those on S_2 - are maintained, but the resistances of their incident branches are increased. For our choice of increments this yields 48 nodes on S_1 and 72 nodes on all the other S_j. Outside S_5 the medium is uniform except for the borehole. We simply ignore the borehole outside S_5 because there it occupies only a small fraction of the medium. For our numerical example we have on each sphere 6 equilatitudinal circles and 12 nodes on each such circle; thus, L = 12 in Figure 5.

III. 3. A Matrix - valued Infinite Ladder Network.

Our objectives are to determine a solution (that is, the nodal pressures measured with respect to the ground node) for this infinite spherical grid and to find a fast computational method for computing numerical values for that solution. Our primary interest is in the nodal pressures within and near the completion zone, but this does not mean that we can simply truncate the medium at some arbitrarily chosen remove from S_5. Instead, we shall seek a driving-point conductance matrix Y_5 that exactly represents the effect of the infinite spherical grid beyond S_5 as measured at the nodes of S_5.

To this end, we decompose the spherical grid into an infinite ladder network of n-ports where $n = 72$. This is shown in Figure 6 where for our model $J = 5$. For example, $R_{j+1/2}$ is the $n \times n$ resistance matrix of the n-port consisting of all radial branches connecting the nodes on S_j to the nodes on S_{j+1}. Thus, $R_{j+1/2}$ is a diagonal matrix. If $j \geq 5$, then -- because of the numbering system of Figure 5 -- $R_{j+1/2}$ has six 12×12 main-diagonal blocks of the form cI, where c is a positive constant equal to the resistance of a radial branch and I is the 12×12 identity matrix. This is an especially simple form of a circulant matrix (Davis 1979).

Furthermore, G_j the conductance matrix of the n-port obtained by pairing each node on S_j with a hypothetical ground node and letting the branches on S_j comprise the interior of the n-port. In terms of the theory described in Section II, the hypothetical ground node is in fact a node that is shorted to all the grid's extremities at infinity. Because the n-port contains no branches to that ground, G_j is a singular matrix. However, again because of the numbering system of Figure 5, G_j can be written in the form of 6×6 blocks, each block being a 12×12 submatrix. For $j > 5$, the six main-diagonal blocks are circulants with exactly three nonzero entries in each row, and the 10 adjacent blocks are also of the form cI, where now c is a negative constant. All the other blocks are null.

The facts that for $j > 5$ both $R_{j-1/2}$ and G_j have this block form and that those blocks are circulant matrices will be invoked to make use of an operational calculus that will speed our computations considerably.

III. 4. An exact Solution. The basic idea in our search for an exact solution for the chosen grid is the replacement of the infinite spherical grid outside S_5 by a set of resistors such that there is a resistor connected between every pair of nodes on S_5 as well as a resistor connected between every node on S_5 and the ground node -- 2628 resistors altogether. These resistors will be called the "terminating resistors" on S_5 and will be so chosen that the driving-point conductance matrix Y_5 for the 72-port consisting of the infinite spherical grid as seen from the 72 nodes on S_5 with the ground node as the common node for all 72 ports remains unchanged under the stated replacement. When this is so, the vector $I_{5+1/2}$ of currents entering the spherical grid outside S_5 at the 72 nodes of S_5 is exactly determined by $I_{5+1/2} = Y_5 V_5$, where V_5 is the vector of node voltages on S_5.

In view of the ladder network of Figure 6, where now $J = 5$, we can write Y_5 as an infinite continued fraction of matrices:

$$Y_5 = \cfrac{1}{R_{5+1/2}} + \cfrac{1}{G_6} + \cfrac{1}{R_{6+1/2}} + \cfrac{1}{G_7} \quad \cdots \qquad (2)$$

This expression does not exist as the common limit of its even and odd truncations. In fact, the even truncations do not exist because the G_j are all singular. However, the odd truncations

$$Y_5^{(N)} = \cfrac{1}{R_{5+1/2}} + \quad \cdots \quad + \cfrac{1}{G_N} + \cfrac{1}{R_{N+1/2}} \qquad (3)$$

do exist for every N and correspond to the driving-point conductance matrix of the finite spherical grid whose nodes on the sphere S_{N+1} are all shorted to the ground node. In effect, assuming that V_5 is given, we are seeking $I_{5+1/2}$ as the limit of the $I_{5+1/2}^{(N)}$ for the finite spherical grids shorted to ground at their spheres S_{N+1}. Actually, the theory cited in Section II dictates a unique $I_{5+1/2}$ for the infinite grid. Moreover, it has been proven (Zemanian and Zemanian, in press, Section 9) that the $I_{5+1/2}^{(N)} \to I_{5+1/2}$ as $N \to \infty$, hence our interest in determining (2) by computing (3) for sufficiently large N.

One may ask, "Since the grid is being truncated at this point of the analysis, why not truncate it at the beginning and seek appropriate boundary conditions on the sphere S_N for some N?" One answer is that it is considerably easier to control the error of truncation by truncating the exact solution (2) rather than imposing a truncation before an analysis is made. Another is that it is with this application of infinite network theory that one is naturally led to a derivation of the exact solution; such an expression does not seem to have arisen in the numerical approaches to the exterior problem listed in the Introduction.

Actually, we are not done until an analysis of the grid within and on S_5 is also made. However, with the terminating resistors at the nodes of S_5 in hand, one need merely make a standard nodal analysis to obtain the nodal pressures on and within S_5 exactly. This solution can then be extended to as many nodes outside S_5 as desired by using pressure transfer ratios (Zemanian and Zemanian, in press, Section 7).

Alternatively, an elementary form of a limb analysis (Zemanian 1978), which is in fact a marching technique, can be used to obtain the nodal presures on the first several spheres beyond S_5 most rapidly indeed, but this can not be carried too far because that marching technique is computationally unstable.

III. 5. An operational Calculus. The computation of (3) for a fixed N can be facilitated by employing the discrete Fourier transform to render the circulant blocks in the $R_{j+1/2}$ and G_j into functions on a set of discrete points in the unit circle -- in this case, twelve discrete points. Thus, manipulations of the circulant matrices implicit in (3) are replaced by corresponding manipulations of ordinary functions. This operational calculus replaces the 72×72 matrices by 6×6 matrices dependent upon the twelve points in the unit circle. The corresponding continued fraction is then computed twelve times and the inverse discrete Fourier transform is applied to obtain the terminating resistors. This provides a fast method of computing Y_5 because of the fast-Fourier-transform algorithm.

III. 6. Computational Advantages. The method described herein is especially advantageous when one wishes to examine many shapes and permeabilty variations for different completion zones, all contained within a fixed sphere, say, S_5. This entails the recomputation of nodal pressures many times. Since Y_5 and thereby the terminating resistors on S_5 do not depend upon the parameters within and on S_5, we can compute those terminating resistors once and for all of the different models, given a fixed spherical discretization for the medium outside S_5. In this way, the domain for our computations has effectively been contracted down to S_5 and its interior, and thus a much smaller system of equations need be solved for each model. It should be noted however that, because the terminating resistors connect all pairs of nodes on S_5, the matrix for those equations is sparse except for a full 72×72 block corresponding to the nodes of S_5. This contrasts with the usual nodal equations for a standard finite-difference analysis within S_N, its matrix being sparse throughout. Nonetheless, whenever N is much larger than J ($= 5$), our method will yield a considerable saving in computation time.

One way of choosing a suitable N is to successively increase its choice until the changes in the resulting terminating resistors become acceptably small. In testing our method for a particular choice of model (see Zemanian and Zemanian, in press, Section 8), we compared results for three different choices of N: $N = 30$, 55, and 105. For example, the high and low values of pressure at the nodes of S_3 were .03106 and .01364 for $N = 30$, .03115 and .01373 for $N = 55$, and .03117 and .01375 for $N = 105$. Using pressure transfer ratios for the computation of nodal pressures outside S_5, we found that the CPU times on a UNIVAC 1100 computer for the determination of the pressures at all the nodes within S_N were 46 seconds for $N = 30$, 1 minute and 19 seconds for $N = 55$, and 2 minutes and 27 seconds for $N = 105$. We note again that in all these cases, the equations needing solution were for the nodes within and on S_5 and thereby had a matrix of order 336×336. In contrast to this, a standard finite difference analysis for the nodes within and on S_N has a matrix of order 2064×2064 for $N = 30$, 3864×3864 for $N = 55$, and 7464×7464 for $N = 105$.

REFERENCES

A. Bayliss, M. Gunzburger, and E. Turkel (1982), "Boundary conditions for the numerical solution of elliptic equations in exterior regions," *SIAM J. Appl.Math.*, vol.42, pp. 430-451.

C. Canuto, S.I. Hariharan, and L. Lustman (1985), "Spectral methods for exterior elliptic problems," *Numerische Mathematik*, vol. 46, pp. 505-520.

P.M. Davis (1979), *Circulant Matrices*, John Wiley, New York.

B. Engquist and A.Majda (1977), "Absorbing boundary conditions for the numerical simulation waves," *Methods of Computation*, vol. 31, pp. 629-651.

G.J. Fix and S.P. Marin (1978), "Variational methods for underwater acoustic problems," *J. Computational Physics*, vol. 28, pp. 253-270.

C.I. Goldstein (1979), "Numerical methods for Helmholtz-type equations in unbounded regions," *Mathematical Methods and Applications of Scattering Theory*, J.A. DeSanto, A.W. Saenz, and W.W. Zachary (Editors), Springer-Verlag Lecture Notes in Physics, No. 130.

C.I. Goldstein (1982), "A finite element method for solving Helmholtz-type equations in waveguides and other unbounded domains," *Mathematics of Computation*, vol. 39, pp. 309-324.

B. Gustafsson and H.O. Kreiss (1979), "Boundary conditions for time dependent problems with an artifical boundary," *J. Computational Physics*, vol. 30, pp. 333-351.

T.M. Hagstrom and H.B. Keller, (1984), *Numerical Solution of Semi-linear Elliptic Problems in Unbounded Domains*, MRC Technical Summary Report No. 2754, University of Wisconsin, Madison, Wisconsin.

T.M. Hagstrom and H.B. Keller (in press), "Exact boundary conditions at an artificial boundary for partial differential equations in cylinders," *SIAM J. Math. Anal.*, vol. 17, 1986.

A.H. Zemanian (1974), "Countably infinite networks that need not be locally finite," *IEEE Trans. Circuits and System*, vol. CAS-21, pp. 274-277.

A.H. Zemanian (1975) "Connections at infinity of a countable resistive network," *Circuit Theory and Applications*, vol. 3, pp. 333-337.

A.H. Zemanian (1978), "The limb analysis of countably infinite electrical networks," *Journal of Combinatorial Theory, Series B*, Vol. 24, pp. 76-93.

A.H. Zemanian (1981), "The characteristic-resistance method for grounded semi-infinite grids," *SIAM. J. Math. Anal.,* vol 12, pp. 115-138.

A.H. Zemanian (1982), "Nonuniform semi-infinite grounded grids." *SIAM J. Math. Anal.,* vol. 13, pp. 770-788.

A.H. Zemanian (1985a), "Operator-valued transmission lines as models of a horizontally layered earth under transient two-dimensional electromagnetic excitation," *IEEE Trans. Antennas and Propagation,* vol. AP-33, pp. 346-350.

A.H. Zemanian (1985b), "Operator-valued transmission lines in the analysis of two-dimensional anomalies embedded in a horizontally layered earth under transient polarized electromagnetic excitation," *SIAM J. Appl. Math.,* vol. 45, pp. 591-620.

A.H. Zemanian (1985c), *Models of Borehole Resistivity Measurements Using Infinite Electrical Grids,* CEAS Technical Report 472, State University of New York At Stony Brook, Stony Brook, N.Y.

A.H. Zemanian (1986), *Infinite Electrical Networks with Finite Sources at Infinity,* CEAS Technical Report 476, State University of New York at Stony Brook, Stony Brook, N.Y.

A.H. Zemanian and H.K. An (1985), *Finite Difference Analysis of Borehole Flows Involving Domain Contractions around Three-Dimensional Anomalies,* CEAS Technical Report 473, State University of New York at Stony Brook, Stony Brook, N.Y.

A.H. Zemaniar and P. Subramaniam (1983), "A theory for ungrounded electrical grids and its application to the geophysical exploration of layered strata," *Studia Mathematica,* vol. LXXVII, pp. 163-181.

A.H. Zemanian and T.S. Zemanian (in press), "Domain contractions around three-dimensional anomalies in spherical finite-difference computations of Poisson's equations," *SIAM. J. App. Math.*

(a)

(b)

Figure 1

pipe

completion zone

virgin rock

(a)

To infinity

(b)

Figure 2

P

h

Return
probe

Borehole

Current
probe

Sonde

Potential
probe

(a)

Figure 3a

1040
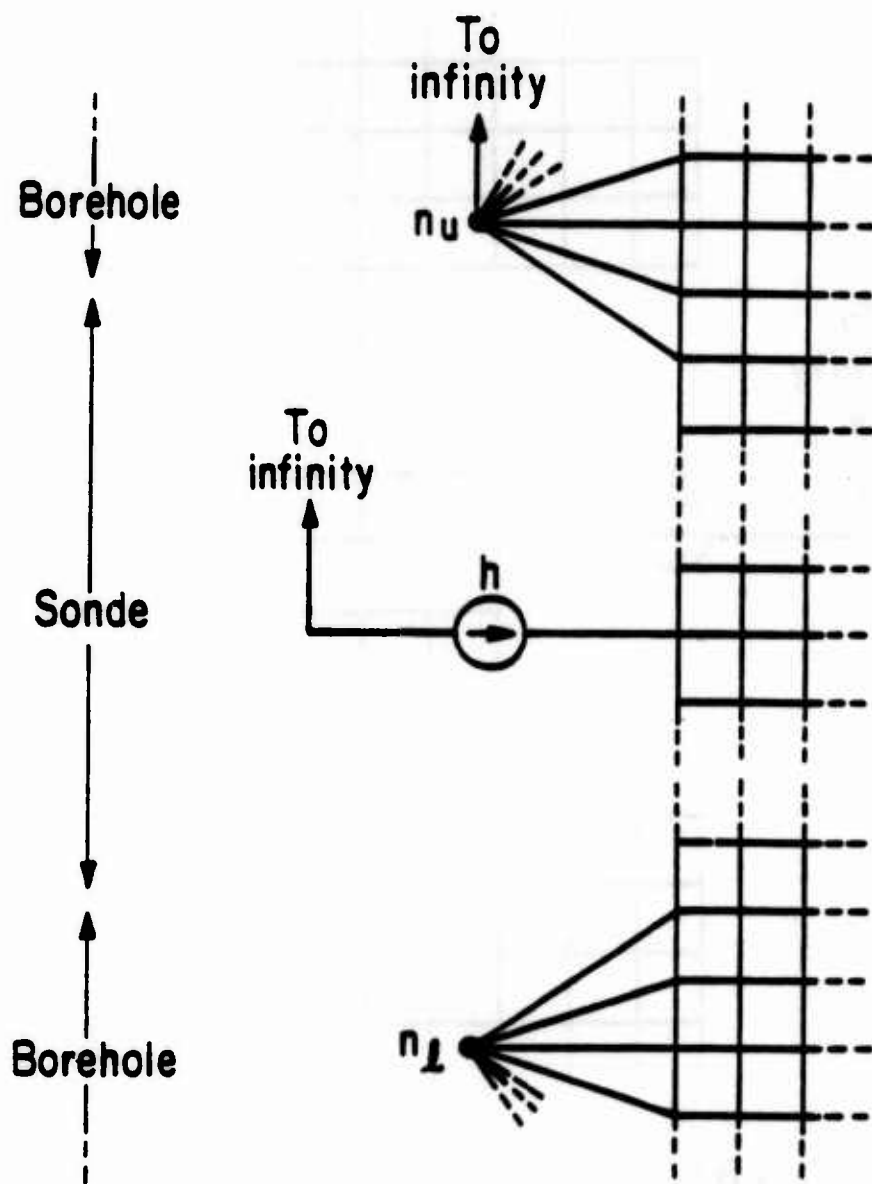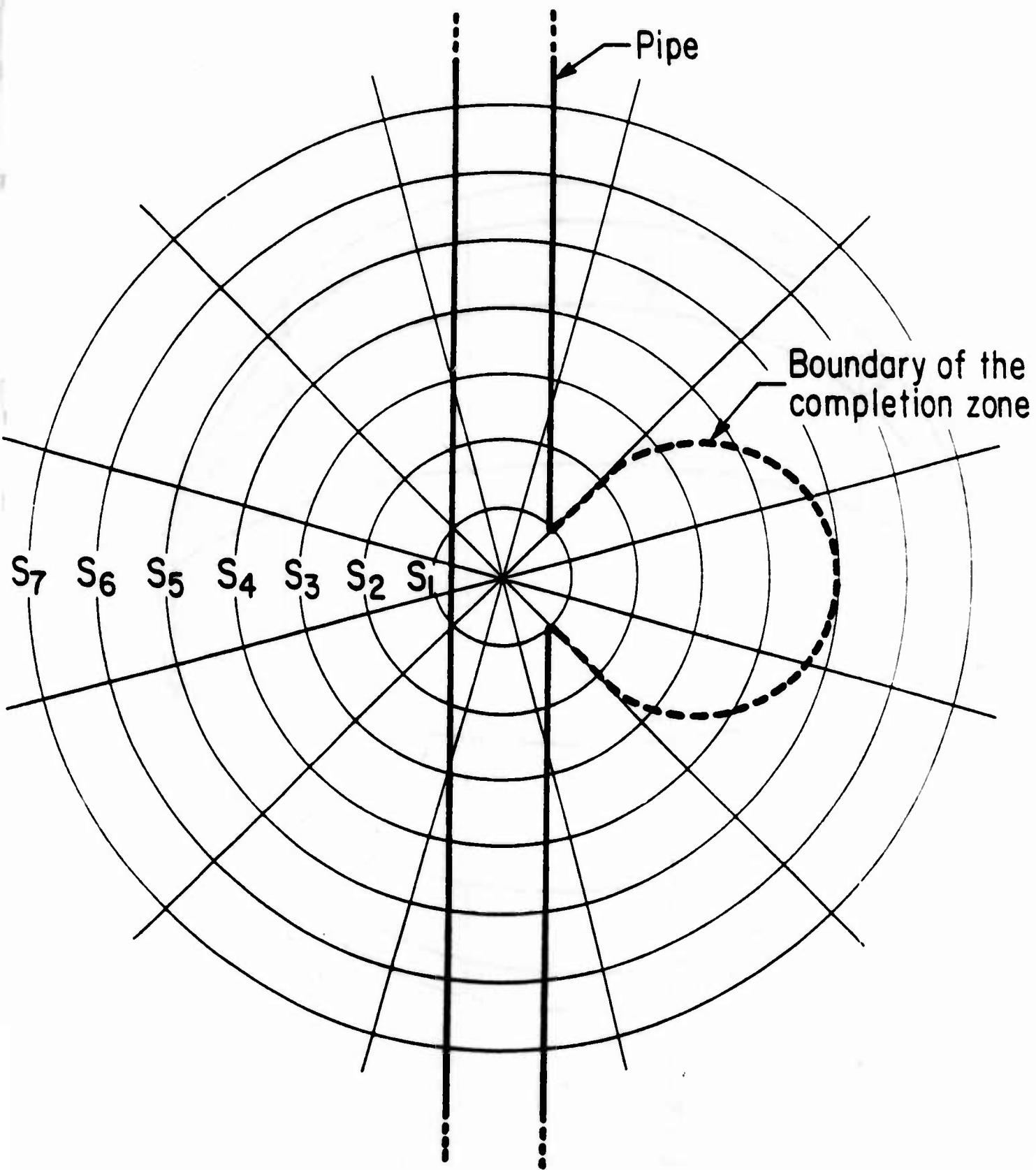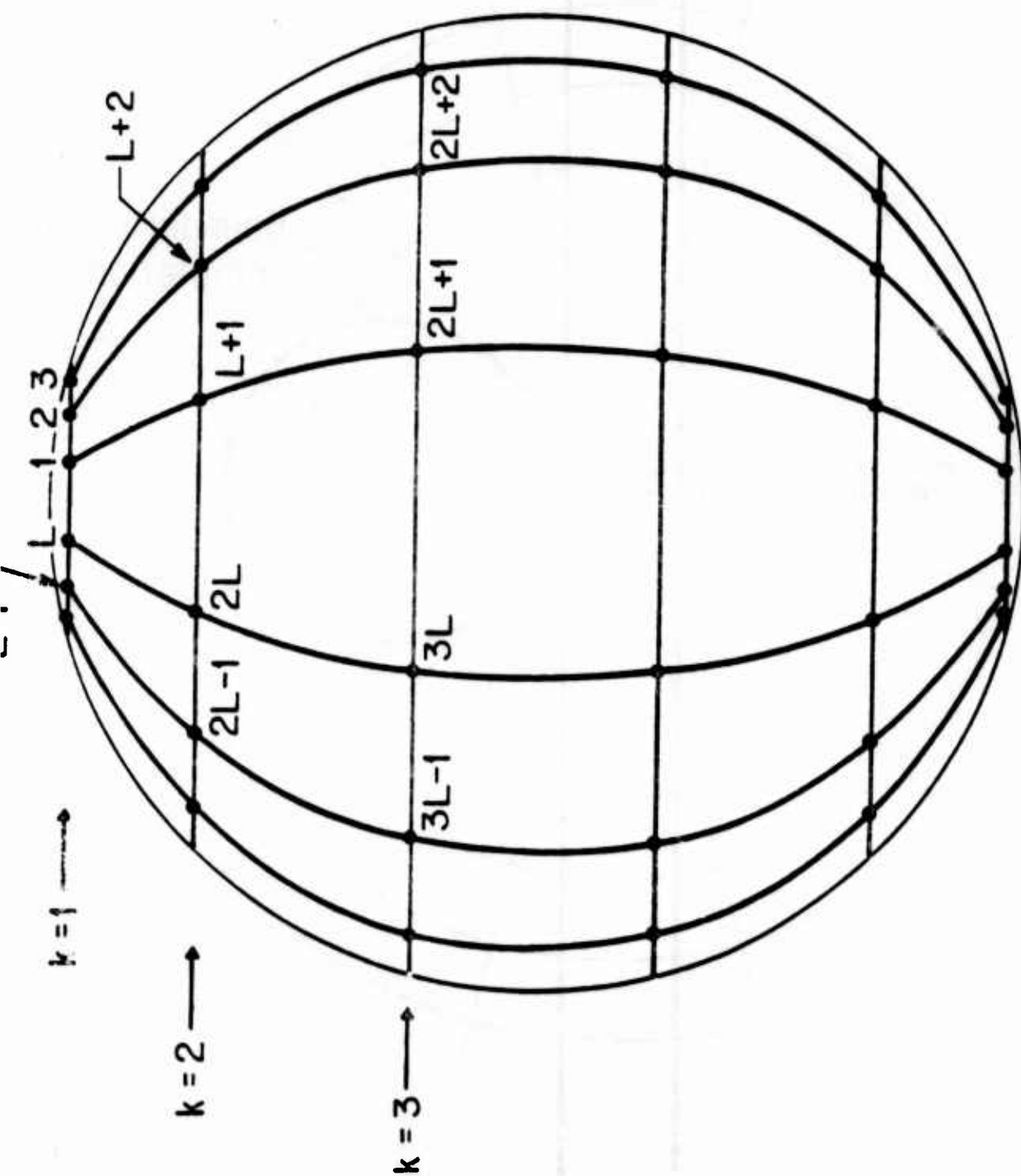
(b)

Figure 3b

Figure 3c

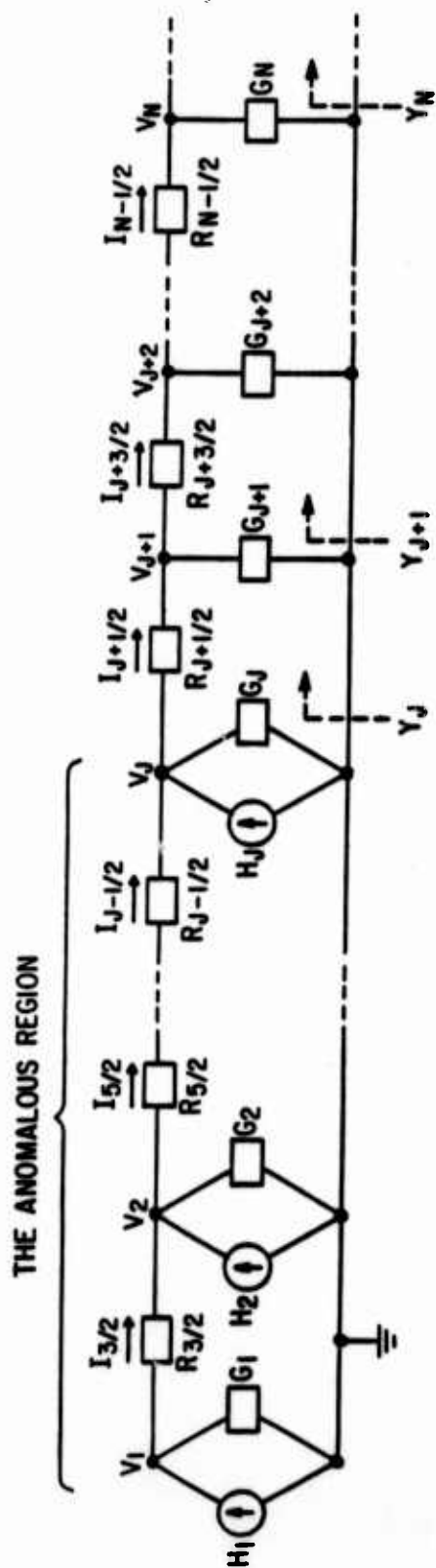Figure 4

1043
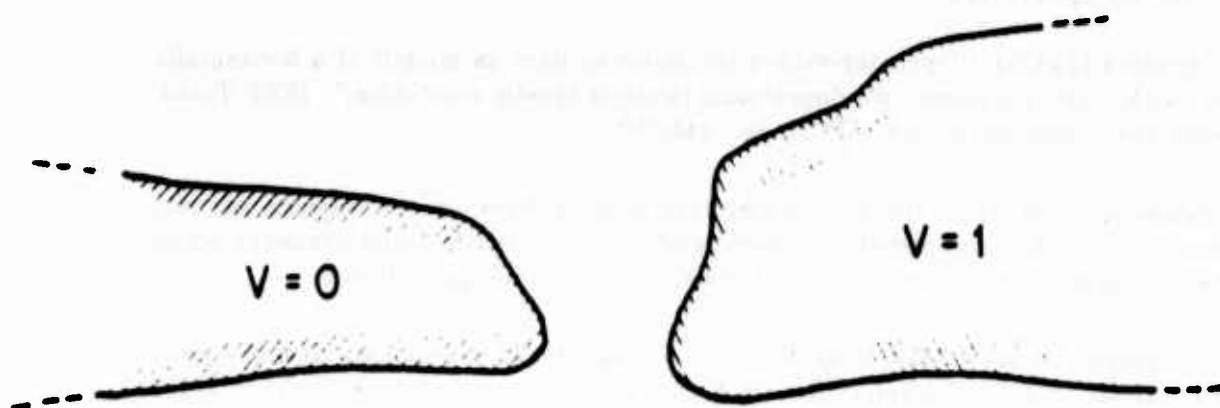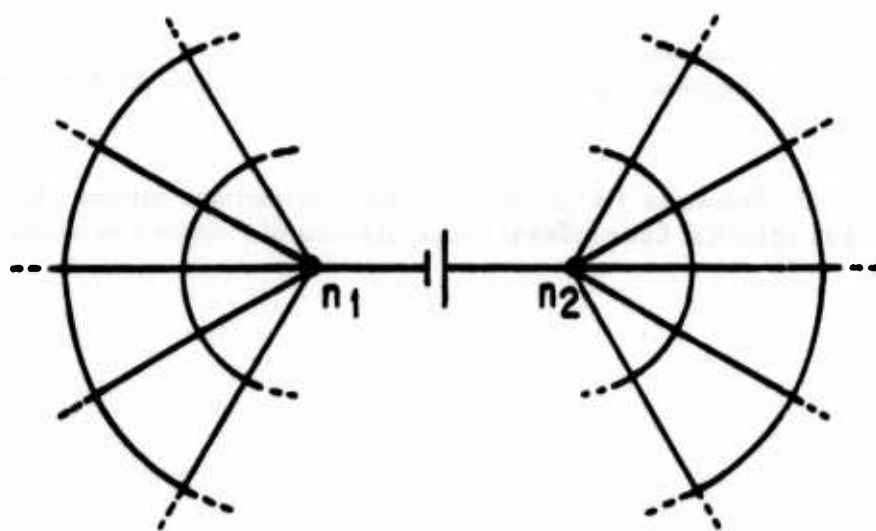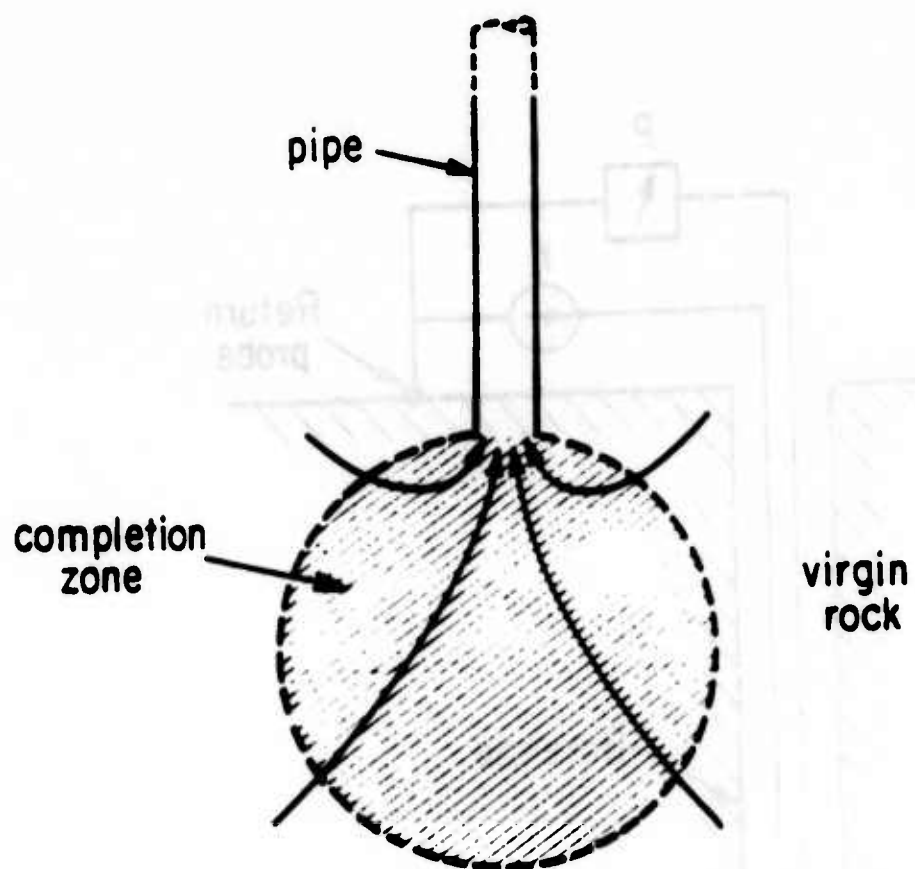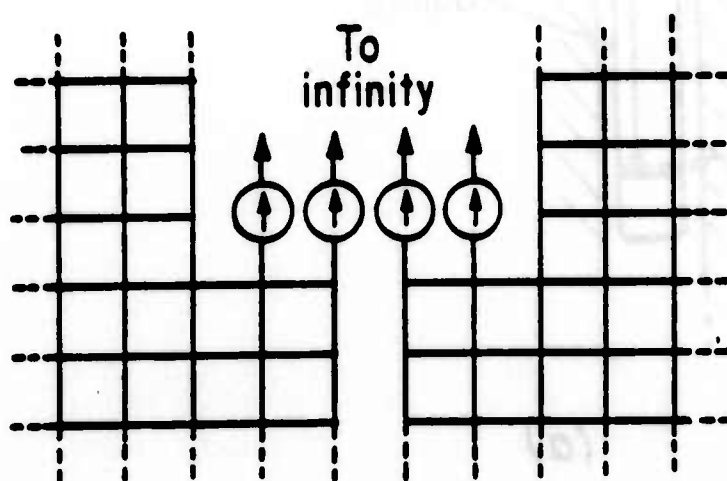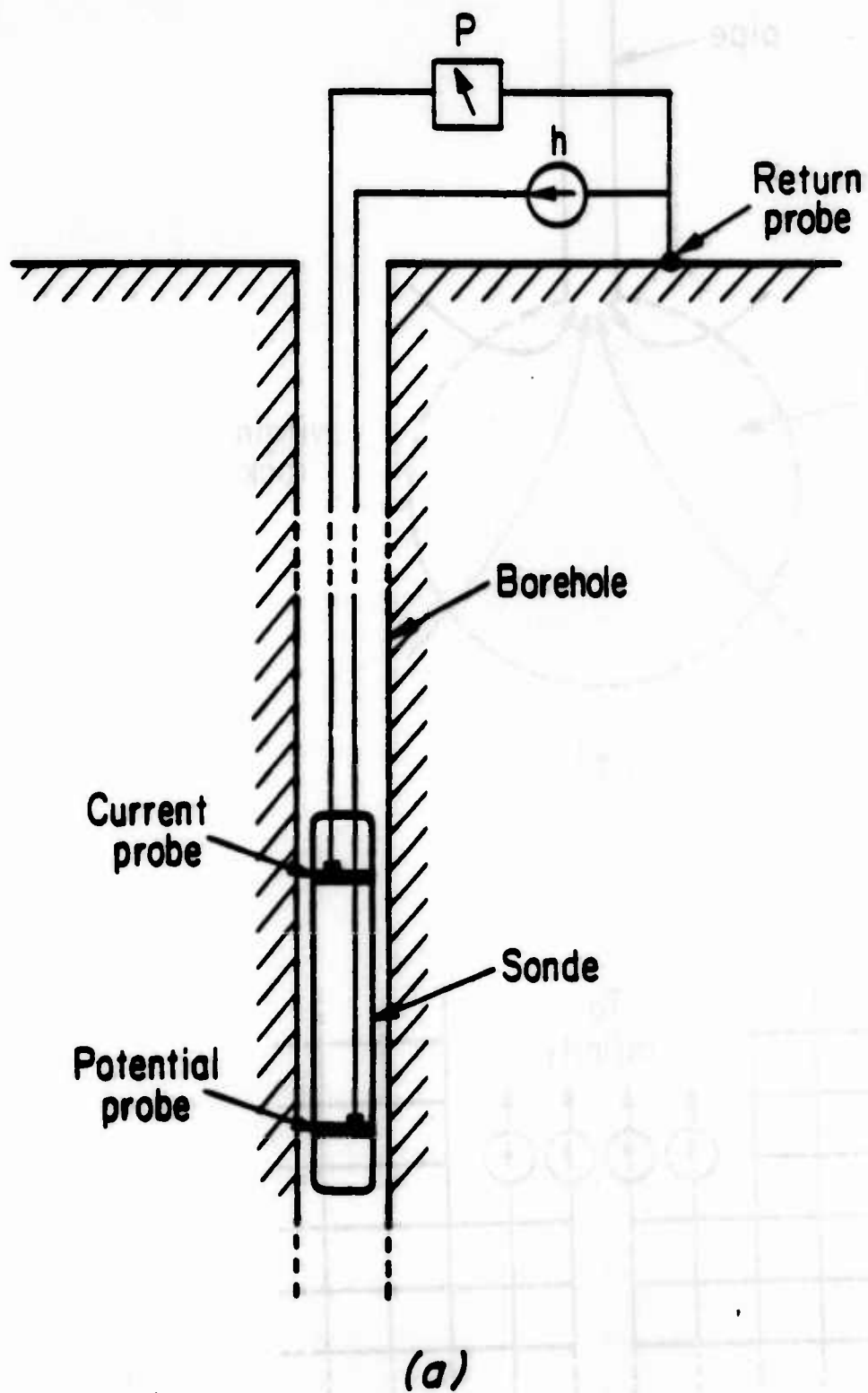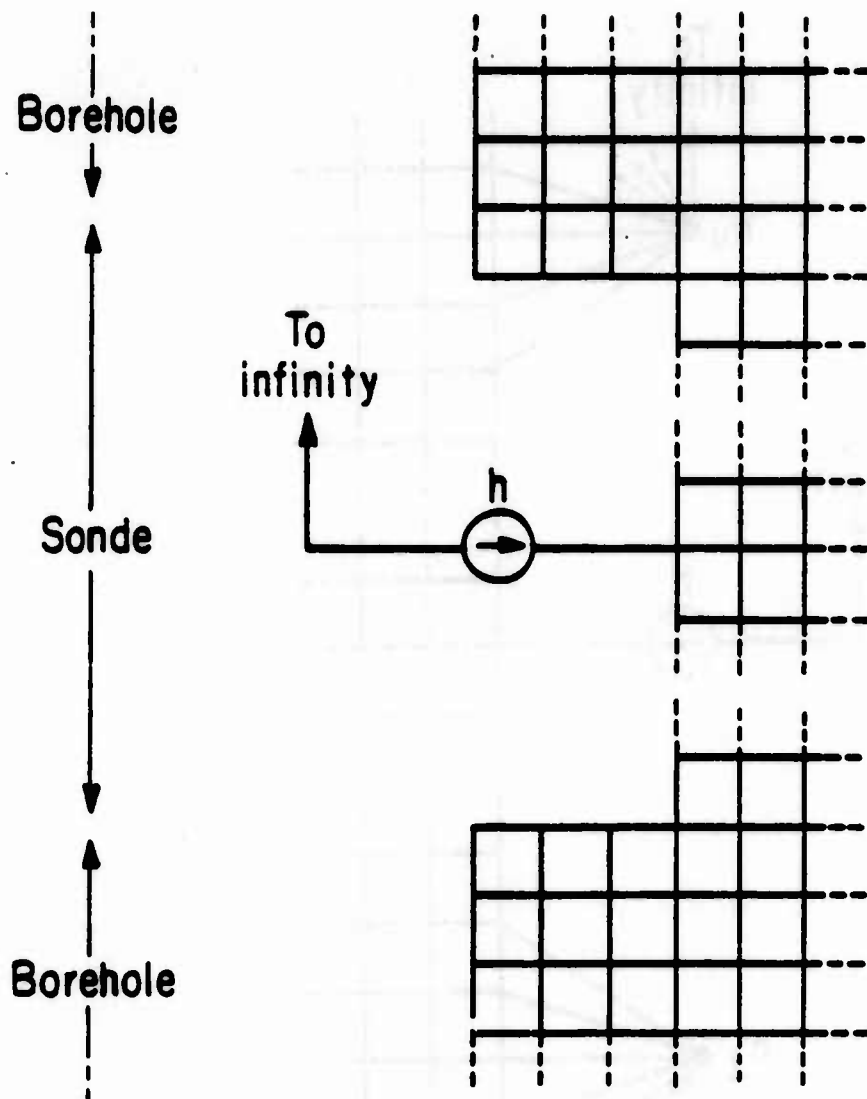
Figure 5

Figure 6

UPWIND DIFFERENCING AND MHD EQUATIONS

M. Brio[1]
Department of Mathematics, UCLA

C.C. Wu[1]
Department of Physics, UCLA

A. Harten[2]
Department of Mathematics, UCLA and
School of Mathematical Sciences, Tel-Aviv University

S. Osher[2]
Department of Mathematics, UCLA

ABSTRACT. We describe recently developed by A. Harten and S. Osher uniformly high order accurate essentially non-oscillatory schemes. The design involves an essentially non-oscillatory piecewise polynomial reconstruction of the solution from its cell averages, time evolution through an approximate solution of the resulting initial value problem, and averaging of this approximate solution over each cell. To _olve this reconstruction problem we use a new interpolation technique that when applied to piecewise smooth data gives high-order accuracy whenever the function is smooth but avoids a Gibbs phenomenon at discontinuities.

As a result of a construction and numerical experiment with the second order upwind schemes for the MHD equations M. Brio and C.C. Wu found that MHD equations are not convex. Specifically, as the transverse magnetic field goes through zero, the slow (fast) field becomes degenerate if the sound speed is greater (less) than the Alfvén speed. This property is called nonconvexity. As a consequence, the solution to the Riemann problem may contain a compound wave, namely a shock followed by a rarefaction wave of the same family and transverse component of the magnetic field changes its sign across the shock wave. As an example, we show a numerical solution which contains a compound wave. For the same Riemann problem, the traditional solution includes a 180° Alfvén wave instead.

1. UNIFORMLY HIGH ORDER ACCURATE ESSENTIALLY NON-OSCILLATORY SCHEMES. Consider hyperbolic initial value problem (IVP).

$$u_t + f(u)_x = 0$$
$$u(x,0) = u_0(x), \tag{1}$$

Here u and f are m vectors. The Jacobian matrix, is assumed to have only real eigenvalues and a complete set of linearly independent eigenvectors. Let $x_j = jh, t_n = n\tau$.

Integrating the partial differential equation (1) over the computational cell $(x_{j-1/2}, x_{j+1/2}) \times (t_n, t_{n+1})$, we get

$$\bar{u}_j^{n+1} = \bar{u}_j^n - \lambda[\hat{f}_{j+1/2}(u) - \hat{f}_{j-1/2}(u)], \tag{2}$$

where

$$\hat{f}_{j+1/2}(u) = \frac{1}{\tau} \int_{t^n}^{t^{n+1}} f(u(x_{j+1/2}, t))dt \tag{3}$$

and

$$\bar{u}_j^n = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_n)dx.$$

We denote by v_j^n the numerical approximation to the cell averages $\bar{u}_j^n$ of the exact solution to (2) and set v_j^0 to be the cell averages of the initial data. Given $v^n = \{v_j^n\}$, we compute v^{n+1} as follows:

First we construct $u(x, t_n)$ out of its approximate cell-averages $\{v_j^n\}$ to the appropriate accuracy and denote the result by $L(x; v^n)$. Next we solve the IVP:

$$v_t + f(v)_x = 0, v(x,0) = L(x; v^n) \tag{4}$$

and denote its solution by $v(x, t)$. Finally we obtain v_j^{n+1} by taking cell averages of $v(x, \tau)$:

$$v_j^{n+1} = \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} v(x, \tau)dx. \tag{5}$$

We define its total variation in x to be:

$$TV(v^n) = TV(v_h(\cdot, t_n)) = \sum_j |v_{j+1}^n - v_j^n|.$$

1048

where $|\ |$ denotes any norm on R^m.

Our goal is that, if the initial data $u_0(x)$ are *piecewise* smooth, then for h sufficiently small

$$TV(v_h(\cdot, t + \Delta t)) \leq TV(v_n(\cdots, t)) + O(h^{N+1})$$

where N is the order of accuracy of (2).

The averaging procedure does not increase the total variation, therefore, the design of ENO high order accurate schemes boils down to a problem on the level of approximation of functions: that of constructing an essentially non-oscillatory high-order accurate interpolant of a piecewise smooth function from its cell averages.

First we consider the scalar case of (1), $m = 1$. In [5] we have constructed an essentially non-oscillatory piecewise polynomial of order N, $Q^N(x; w)$, that interpolates a piecewise-smooth function $w(x)$ at the cell interface points:

$$Q^N(x_{j+1/2}; w) = w(x_{j+1/2}) \tag{6}$$

and satisfies, wherever $w(x)$ is smooth

$$\left(\frac{d}{dx}\right)^r q^N(x \pm 0; u) = \left(\frac{d}{dx}\right)^r w(x) + O(h^{N+1-r}), r = 1, ..., N. \tag{7}$$

We shall use this polynomial together with two different approaches to design ENO schemes. These methods are:

RP: Reconstruction via the primitive function.

RD: Reconstruction via deconvolution.

We describe here the first one, for the second approach the reader is referred to [5]. Let $W(x)$ be the primitive function of $u(x)$

$$W(x) = \int_a^x u(s)ds. \tag{8}$$

Since we wish to reconstruct $u(x)$ out of its approximate cell averages v_j, (dropping the t or n dependence), we have an approximation to $W(x_{j+1/2})$

$$W(x_{j+1/2}) = \sum_{k=0}^{j} v_k^h. \tag{9}$$

In each cell $I_j : \{x/x_{j-1/2} \leq x < x_{j+1/2}\}$, $Q^N(x; w)$ is a polynomial of degree N which interpolates $w(x_{j+1/2})$; i.e. for all j

$$Q^N(x_{j+1/2}; w) = w(x_{j+1/2}).$$ (10)

Thus $Q^N(x, w)$ is a continuous piecewise polynomial, and both of $d/dx \, Q^N(x\pm 0; w)$ are globally well defined.

Our approximation to (1) can be obtained by solving (4) with

$$v(x, w) = d/dx \, Q^n(x; w^n) = L(x; v^n),$$

obtaining $v(x, t)$, $0 \leq t \leq \tau$, and then computing cell averages (5). This can be rewritten, using the divergence theorem, as:

$$v_j^{n+1} = v_j^n - \lambda(\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n),$$ (11)

since

$$\frac{Q^N(x_{j+1/2}; w^n) - Q^N(x_{j-1/2}; w^n)}{h} = v_j^n$$

because of (6) and (9).

Here $\hat{f}_{j+1/2}^n$ is computed by averaging the flux function $f(u)$ applied to $v(x_{j+1/2}, t)$ as in (3).

For general $f(u)$ the explicit solution to (4) can be difficult to obtain and various approximations might be applied [5].

In the next section we will describe some results of the application of the upwind schemes for the MHD equations.

2. <u>NONCONVEXITY OF THE MHD EQUATIONS</u>. The equations of ideal magnetohydrodynamics (MHD) characterize the flow of conducting fluid in the presence of magnetic field. They represent coupling of the fluid dynamical equations with Maxwell's equations of electrodynamics. By neglecting displacement current, electrostatic forces, effects of viscosity, resistivity, and heat conduction, one obtains the following ideal MHD equations [1]:

$$\rho_t + \vec{\nabla} \cdot (\rho\vec{u}) = 0$$

$$(\rho\vec{u})_t + \vec{\nabla} \cdot (\rho\vec{u}\vec{u} + \vec{I}P^* - \vec{B}\vec{B}) = 0,$$

$$\vec{B}_t + \vec{\nabla} \cdot (\vec{u}\vec{B} - \vec{B}\vec{u}) = 0,$$

$$E_t + \vec{\nabla} \cdot ((E + P^*)\vec{u} - \vec{B}(\vec{B} \cdot \vec{u})) = 0,$$

1050

with the additional requirement that $\vec{\nabla} \cdot \vec{B} = 0$, which is satisfied if it is satisfied initially. In the above equations the following notations are used: ρ for density, $\vec{u}$ for velocity, $\vec{B}$ for magnetic field, P for static pressure, P^* for full pressure, $P^* = P + \frac{1}{2} |\vec{B}|^2$, E for energy, $E = \frac{1}{2}\rho |\vec{u}|^2 + P/(\gamma - 1) + \frac{1}{2}|\vec{B}|^2$, and γ for the ratio $\frac{1}{2}|B^*|^2, \frac{1}{2}|u^2|^2$ of specific heats. We consider one-dimensional MHD equations which are obtained from the above system by assuming that all variables depend on x and t only. The resulting equations are:

$$\rho_t + (\rho u)_x = 0,$$
$$(\rho u)_t + (\rho u^2 + P^*)_x = 0,$$
$$(\rho v)_t + (\rho u v - B_x B_y)_x = 0,$$
$$(\rho w)_t + (\rho u w - B_x B_z)_x = 0,$$
$$(B_y)_t + (B_y u - B_x v)_x = 0,$$
$$(B_z)_x + (B_z u - B_x w)_x = 0,$$
$$E_t + ((E + P^*)u - B_x(B_x u + B_y v + B_z w))_x = 0.$$

$B_x \equiv$ const, $u, v,$ and w are three components of the velocity field.

The eigenvalues of the Jacobian matrix can be written in nondecreasing order as

$$u - c_f, \quad u - c_a, \quad u - c_s, \quad u, \quad u + c_s, \quad u + c_a, \quad u + c_f,$$

where c_f, c_a, c_s are called the fast, Alfvén, and the slow characteristic speeds, respectively. They can be expressed as:

$$c_a{}^2 = b_x{}^2,$$
$$c_{f,s}{}^2 = \frac{1}{2}((a^*)^2 \pm \sqrt{(a^*)^4 - 4a^2 b_x^2}), \tag{12}$$

with the following notations:

$$b_x = B_x/(\rho)^{\frac{1}{2}},$$
$$b_y = B_y/(\rho)^{\frac{1}{2}},$$
$$b_z = B_z/(\rho)^{\frac{1}{2}},$$
$$b^2 = b_x^2 + b_y{}^2 + b_z{}^2,$$
$$(a^*)^2 = (\gamma - 1)(H - \frac{1}{2}(u^2 + v^2 + w^2)) + (2 - \gamma)b^2,$$

and a is the sound speed which is related to a^* by

$$a^2 = (a^*)^2 - b^2.$$

In equation (12) the plus sign is for c_f and the minus sign for c_s. There are two points where these eigenvalues may coincide:

(1) At $B_x = 0$, $c_s = c_a = 0$, thus u is an eigenvalue of multiplicity 5.

(2) At $B_y{}^2 + B_z{}^2 = 0$, $c_f{}^2 = \max(a^2, b_x{}^2)$, and $c_s{}^2 = \min(a^2, b^2)$. Therefore, for the case $a^2 \neq b_x{}^2$, either $c_f{}^2 = b_x{}^2$ or $c_s{}^2 = b_x{}^2$, thus multiplicity of $u \pm c_a$ is 2; and for the case $a^2 = b_x{}^2$, $c_f{}^2 = c_s{}^2 = b_x{}^2$ and the multiplicity of $u \pm c_a$ is 3.

The usually used eigenvectors [1] do not form a complete set, i.e., near the points where either $B_x = 0$ or $B_y{}^2 + B_z{}^2 = 0$, this set is not well-defined and the matrix with these eigenvectors as its columns becomes singular. However, by proper renormalization a complete set can be obtained [2]. Specifically, near the points that $B_t = 0$ $(B_t{}^2 = B_y{}^2 + B_z{}^2)$, the normalization factor is proportional to B_t for the fast wave if $a^2 < b_x{}^2$ and for the slow wave if $a^2 > b_x{}^2$. Due to the required normalization, these waves are not genuinely nonlinear as usually believed. For example, using the set of right eigenvectors given by Jeffrey and Taniuti it follows that both fast and slow waves are genuinely nonlinear (Theorem E.1 in Ref. 1), i.e., $(\nabla\lambda) \cdot R \neq 0$ for these waves. Now, by using our complete set of eigenvectors, one gets $(\nabla\lambda) \cdot R \propto B_t$, when B_t is small for the fast (slow) wave if $a^2 < b_x{}^2(a^2 > b_x{}^2)$. Thus, when B_t is zero, either the slow or fast wave becomes degenerate. Threfore, they are nonconvex!

As a consequence of the nonconvexity, there exist solutions to some coplanar MHD Riemann problems, whose initial transverse magnetic fields on the left and right states have opposite signs, such that the transverse magnetic field change its sign through the slow (fast) compound wave, which consists of a slow (fast) shock wave and attached to it slow (fast) rarefaction wave. The slow (fast) wave can exist if condition $a^2 > b_x{}^2$ $(a^2 < b_x{}^2)$ holds. These solutions satisfy physical entropy condition and Liu's admissibility criteria [3] and are suggested by the mathematical theory of the scalar nonconvex conservation law as well as by Liu's work on nonconvex Euler's equations of hydrodynamics. Figure 1 illustrates a solution to the coplanar MHD Riemann problem containing a slow compound wave (SM) obtained by an upwind numerical scheme [2].

The initial data for the problem is as follows: $\rho_\ell = 1.$, $u_\ell = .0$, $v_\ell = .0$, $p_\ell = 1.$, $p_r = .125$, $u_r = 0.$, $v_r = .0$, $p_r = .1$, $B_x = .75$. The initial discontinuity is in the middle of the computational interval. The solution is shown after 800 steps with $\Delta t = 0.2$. It consists of a fast rarefaction wave (FR) and slow compound wave (SM) moving to the left; contact (C), slow shock (SS) and a weak fast rarefaction wave (FR) moving to the right. The numerical solution is in good agreement with the appropriate Rankine-Hugoniot jump relations and Riemann invariants across and discontinuities and rarefaction waves.

To illustrate the behavior of the eigenvalue along the shock curve, we use the numerical data for the left state with respect to the slow shock, ($\rho = 0.6763$, $u =$

1052

0.6366, $v = 0.2333$, $B_y = 0.5849$, $p = 0.4574$), and resolve the jump relations using $(B_y)_r$ as a parameter [4]. The dependence of the shock speed(s) and the slow characteristic speed $(u - c_s)$ with respect to $(B_y)_r$ is shown in Figure 2 for the entropy nondecreasing shocks. Point (b) denotes the intersection of these two curves. The values of the variables at the intersection point are as follows: $\rho = 0.7935$, $u = 0.4983$, $v = -1.290$, $B_y = -.307$, $p = 0.667$, and $s = 0.2995$.

The portion of the solution containing compound waves which consists of a slow shock (SS) and attached to it rarefaction waves (SR) is shown in Figure 3 for the u variable. The continuous line in this figure shows the position of the slow shock and attached to it rarefaction wave using the above calculations for the shock position and appropriate Riemann invariants to find position of a rarefaction wave. If the right state is obtained using Riemann invariants and intersection point as a left ∴ then the values are as follows: $\rho = 0.6965$, $u = 0.5987$, $v = -1.583$, $B_y = -0.5341$, $p = 0.5157$. The dotted line shows the values obtained by the second order upwind scheme. Those values are $\rho = 0.6962$, $u = 0.5997$, $v = -1.578$, $p = 0.5133$, which has a maximum deviation of 0.4% in density and pressure from the state obtained by using the left numerical state and appropriate Rankine-Hugoniot relations and Riemann invariants. The agreement is within the numerical accuracy used for those calculations.

Figure 4 illustrates the relation between the shock speed (s) and the characteristic speeds (λ_s) for different points on a shock curve in Fig. 2, using $x - t$ diagram. The left state corresponds to the origin in Fig. 2, and the right state corresponds to the points denoted by a, b, c, and d on a shock curve in this figure. The first case (a) illustrates the shock with convergent characteristics and is similar to the shocks usually encountered for Euler equation which correspond to genuinely nonlinear fields. The second case (b) has the right characteristic speed equal to the shock speed. This allows for a rarefaction wave to be attached to such a shock as in compound wave considered in the above numerical example. Case (c) shows a shock having divergent characteristics on the right hand side. In this case, two waves of the same family can travel in the same direction without one being overtaken by another. The last diagram (d) shows a particular case of the previous one, the characteristic speed is constant across the shock. It corresponds to a 180° Alfvén wave, namely, density, pressure, x-component of the velocity are constant across the shock and the transverse magnetic field reverses its sign.

Traditional solution to this problem is to include an 180° Alfvén wave as suggested by 3D solution of the MHD Riemann problem. This is because, as we have mentioned before, all waves, except Alfvén waves, for the case $B_x \neq 0$ are coplanar, and therefore Alvén waves have to be introduced if the MHD Riemann problem is not coplanar.

Numerically, Lax-Friedrichs and upwind schemes give the solution containing

compound wave; while Lax-Wendroff scheme seems to give a one-parameter family of solutions depending on the magnitude of the artificial or physical viscosity which approaches solution containing an 180° Alfvén wave as resistivity becomes smaller.

Therefore, one of the approaches to pick "physical" solution is to study those problems for the full resistive 3D MHD equations as resistivity and B_s tend to zero in arbitrary order. A valuable tool in this investigation can be an arbitrary high order essentially non-oscillatory scheme which will allow to study the effects of resistivity.

BIBLIOGRAPHY.

1. A. Jeffrey and T. Taniuti, Non-linear wave propagation, Academic Press, Inc., 1964.

2. M. Brio and C. C. Wu, "An upwind differencing scheme for the equtions of magnetohydrodynamics," UCLA PPG report, submitted to J. Comp. Phys. 1985.

3. T. P. Liu, Amer. Math. Soc. Memoirs, 240, 1981.

4. J. Bazer and W. B. Ericson, Astrophys. J., 129 (1959), 758.

5. A. Harten, S. Osher, "Uniformly high order accurate essentially non-oscillatory schemes," MRC Technical Summary Report #2823, May 1985, submitted to SINUM.

6. P. L. Roe, J. Comput. Phys., 43 (1981), 357.
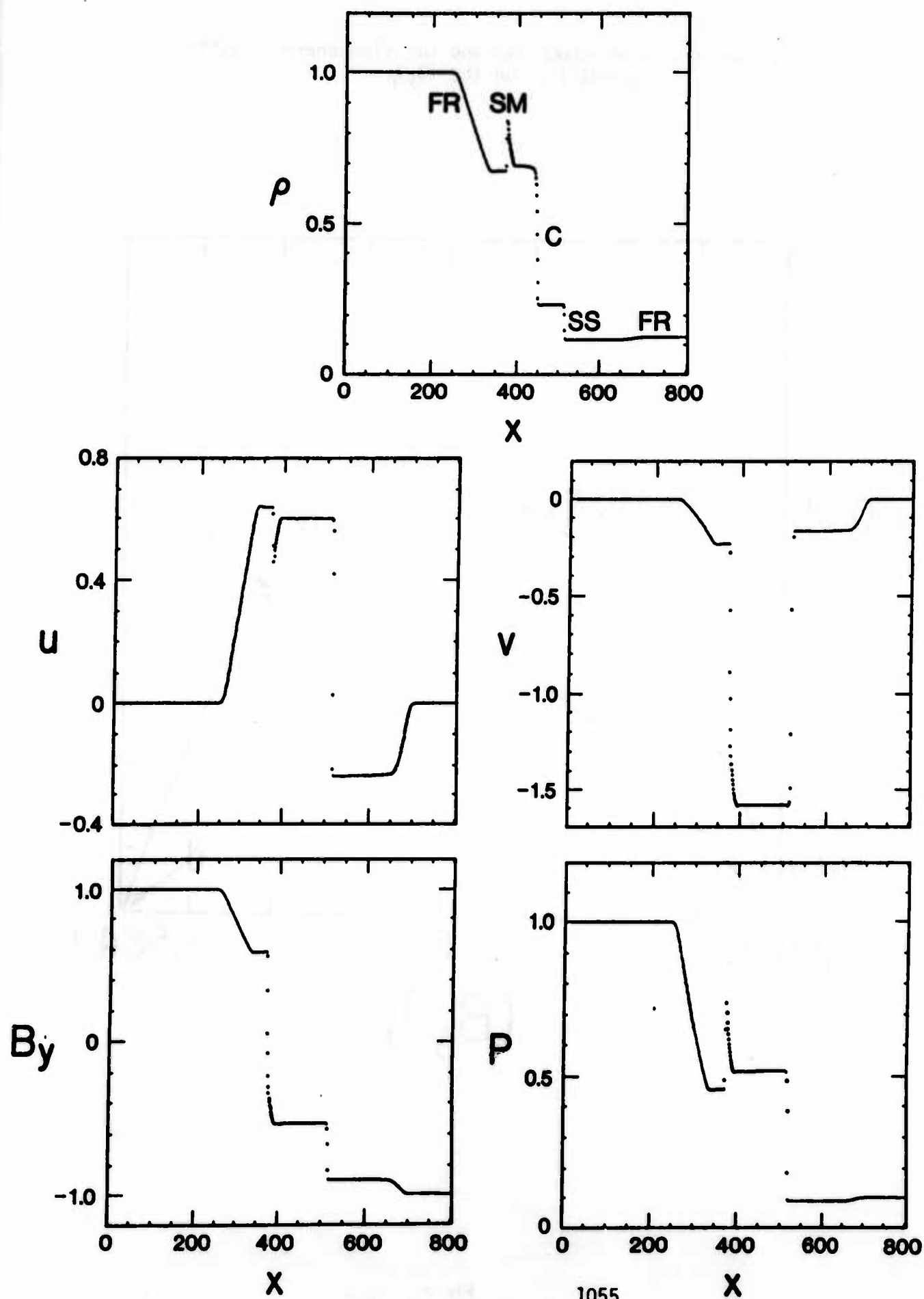
Second order upwind scheme

Fig. 1

Dependence of the schock (s) and the slow characteristic speeds (λ_s) on the $(B_y)_r$.

$\lambda_s = u - c_s$

s

a

b

c

d

.5849   0   -.5849
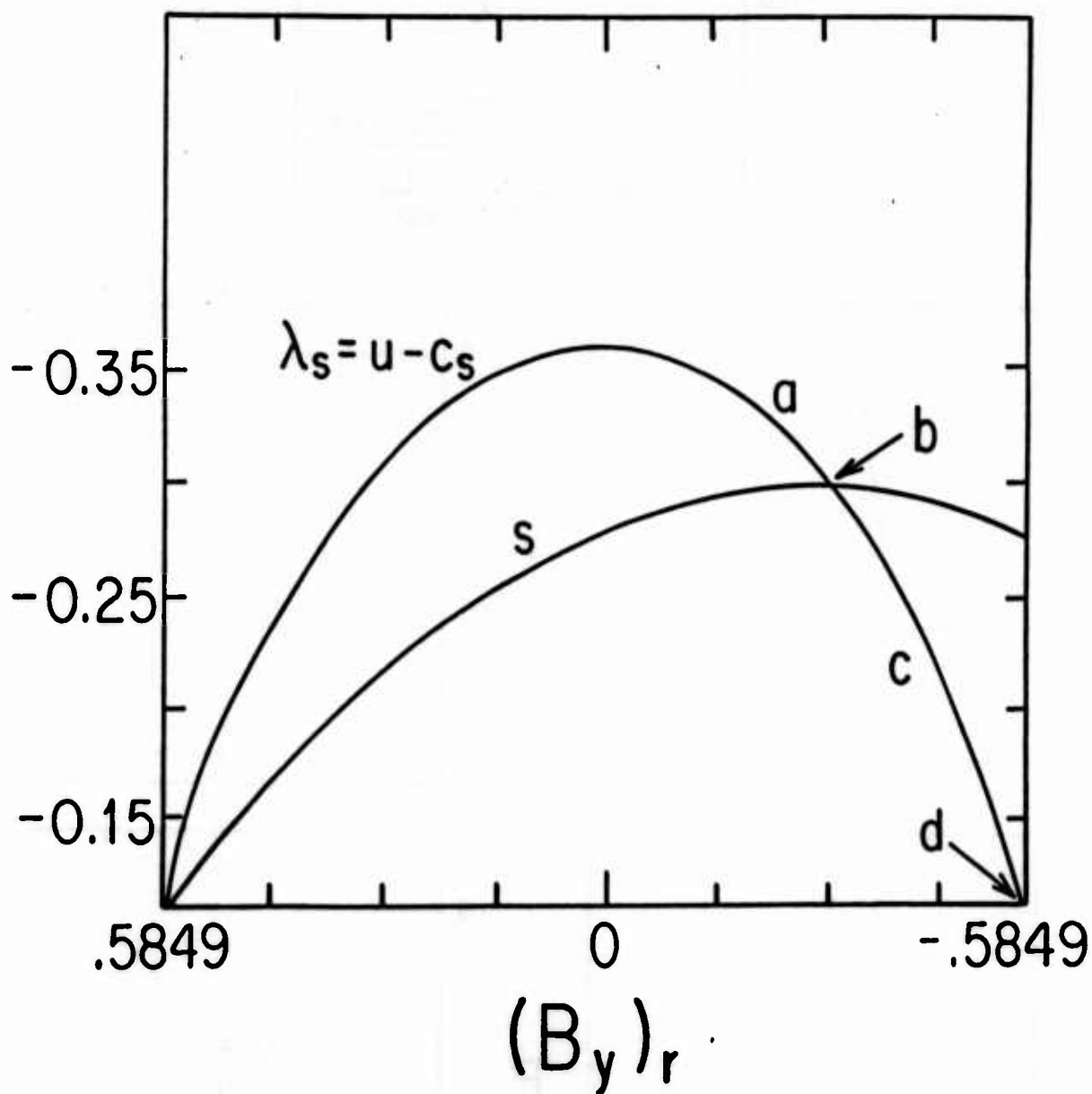
$(B_y)_r$

Fig. 2   1056

Slow compound wave (x-velocity component).
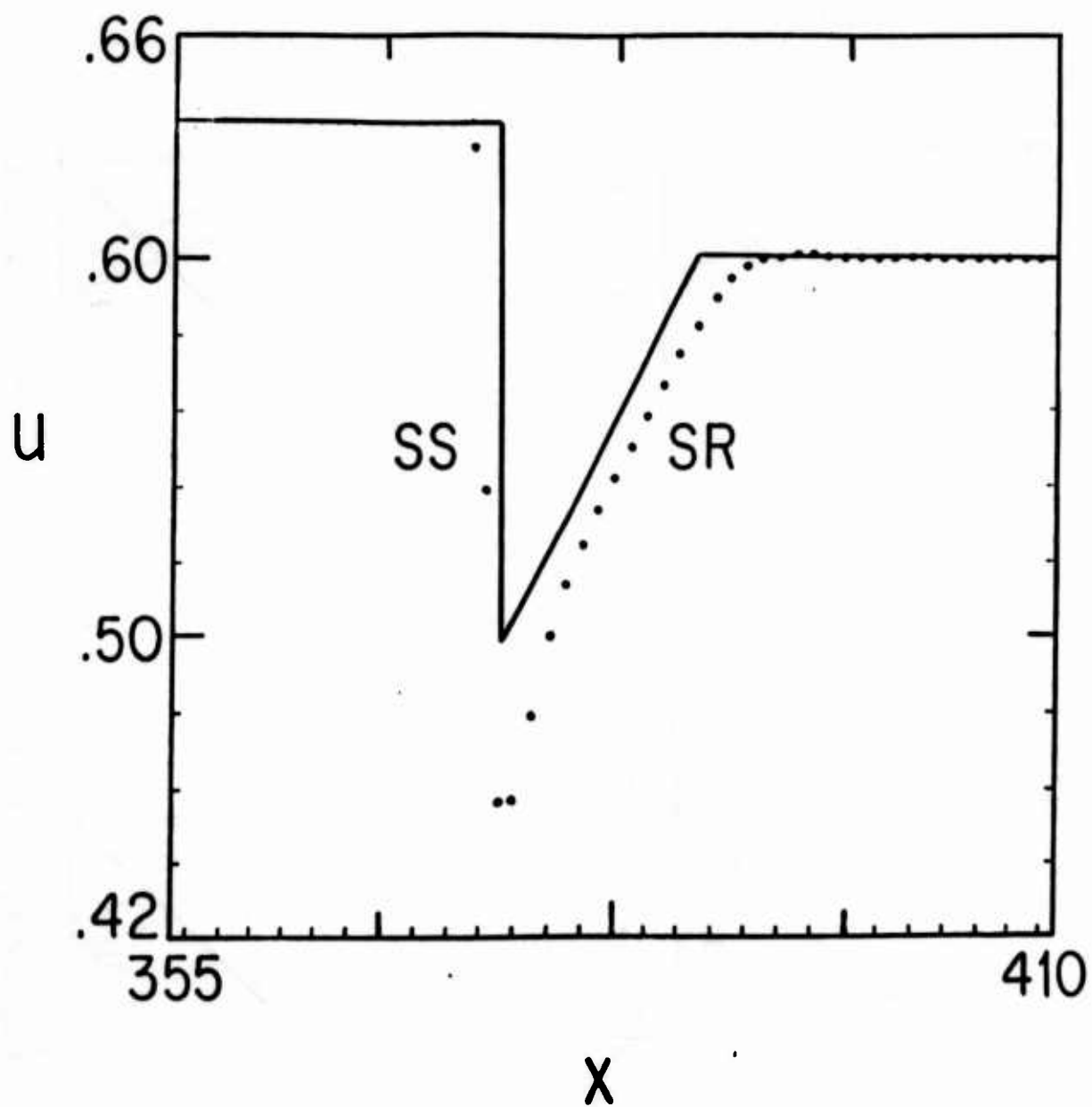
Fig. 3   1057

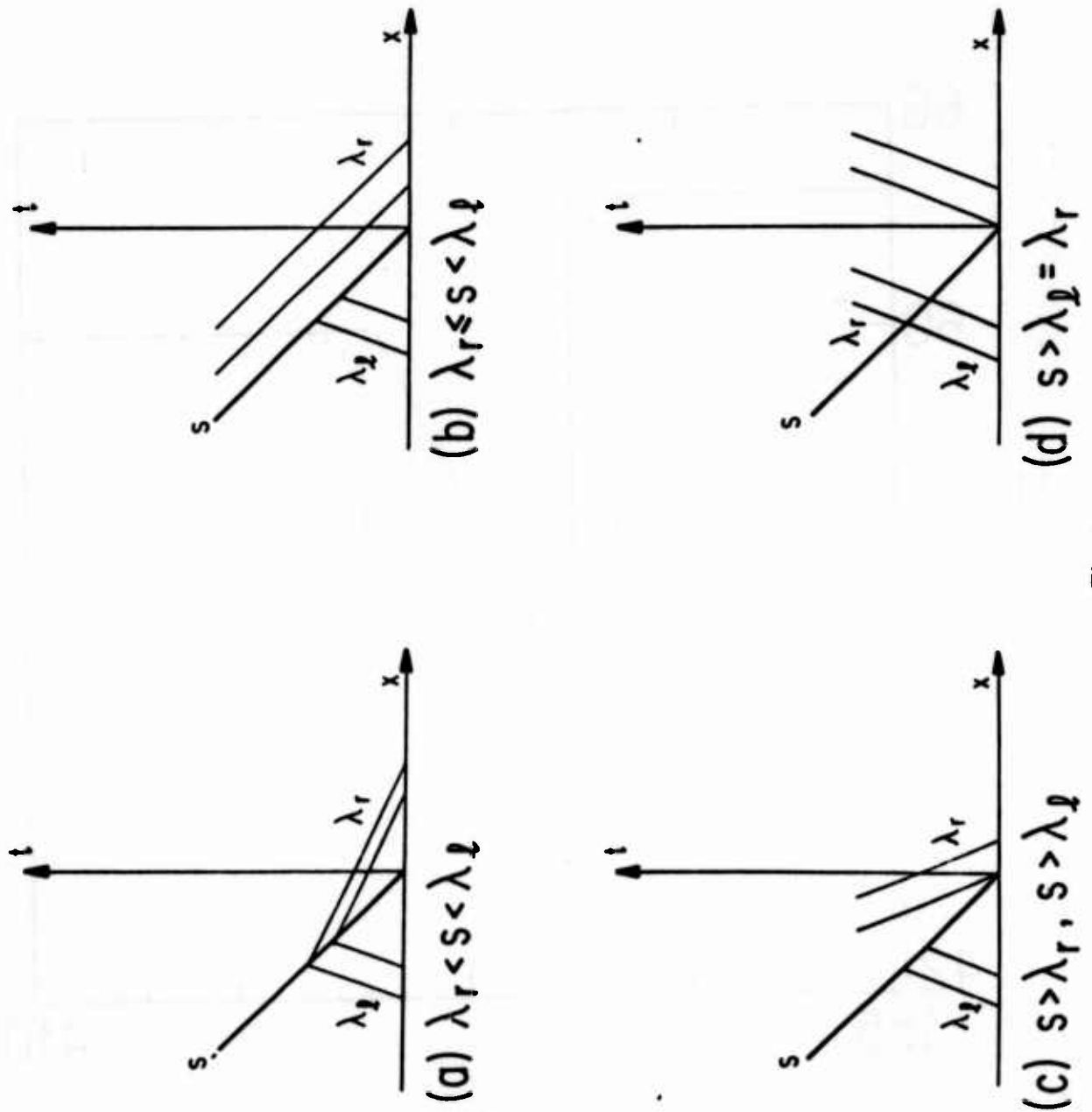Relation between the shock (s) and characteristic speeds (λ_s) for the different points on a shock curve.

(a) $\lambda_r < s < \lambda_\ell$

(b) $\lambda_r \leqslant s < \lambda_\ell$

(c) $s > \lambda_r, s > \lambda_\ell$

(d) $s > \lambda_\ell = \lambda_r$

Fig. 4

FINITE DIFFERENCE METHODS FOR
POLAR COORDINATE SYSTEMS

John C. Strikwerda and Yvonne M. Nagel

Mathematics Research Center
University of Wisconsin - Madison
610 Walnut St.
Madison, WI 53705

Abstract

We discuss finite difference methods for partial differential equations on polar and spherical coordinate systems. The distinctive feature of these coordinate systems is the coordinate system singularity at the origin. We show how to accurately and conveniently determine the solution at the origin for both scalar and vector fields. We also discuss the Fourier method to approximate derivatives with respect to the angular variable in polar coordinates. Computational examples are presented illustrating the accuracy and efficiency of the method for hyperbolic and elliptic equations, and also for the computation of vector fields at the origin.

1. Introduction

In this paper we consider the use of polar and spherical coordinates with finite difference methods, determining how to achieve accurate results with convenience. Although the use of finite difference methods with polar coordinates is not at all new there are several features of their use that are not well known among numerical analysts, computational scientists, and engineers. In particular, the accurate treatment of vector fields with polar coordinates is not widely known.

It is the aim of this paper to bring together the pertinent information and present it in an organized way. As such, this paper presents few new ideas, but it is hoped that it will be a useful addition to the literature on numerical methods.

Much of what is presented here also applies to axially symmetric problems in polar or spherical coordinates; the common feature of these problems is the singular nature of the coordinate system at the origin.

---

2. The Center Formulas

Consider the plane with a polar coordinate system. Each point is determined by its polar coordinates (r, ϕ) which, for points other than the origin, is unique up to integer multiples of 2π in ϕ. However the origin has the coordinates $(0, \phi)$ for all angles ϕ, and it is this lack of uniqueness in the coordinates of the origin that introduces difficulties for numerical methods. These difficulties are displayed in the Jacobian of the coordinate map which takes ordered pairs in $[0, \infty) \times \mathbb{R}$ to points in the plane. The Jacobian vanishes on $\{0\} \times \mathbb{R}$. However it is important to realize that this singularity is present in the coordinate map and polar representations of functions and need not be present in the functions themselves. In this paper we shall only consider functions which are smooth in the domain being considered.

The singular behavior of the polar coordinate system at the origin usually precludes the direct use of finite difference approximations to differential equations at that point. We consider therefore the use of interpolation formulas to accurately determine the solution at the origin. We begin by considering a function defined in the plane, without considering a coordinate system.

Consider a smooth function u defined in a neighborhood of a point P in the plane. We wish to express $u(P)$ in terms of averages, $\bar{u}(P, \rho)$, on circles of radius ρ centered at P. We begin by expanding u in a Taylor series in cartesian coordinates with the origin at P,

$$u(x, y) = \sum_{k,l=0}^{N} \frac{x^k y^l}{k!\, l!} \frac{\partial^{k+l} u}{\partial x^k \partial y^l}(0,0) + R_N. \tag{2.1}$$

where R_N is the remainder term. Then,

$$\bar{u}(P, \rho) = \frac{1}{2\pi} \int_0^{2\pi} u(\rho\cos\phi, \rho\sin\phi)\, d\phi \tag{2.2}$$

and using (2.1)

$$\bar{u}(P, \rho) = \sum_{l=0}^{N} c_l \rho^{2l} \nabla^{2l} u(P) + \bar{R}_N \tag{2.3}$$

where $c_l = 1/4^l(l!)^2$. Formula (2.3), which is independent of a coordinate system, is the basis for determining a function value at the origin given values of the function at points nearby and given the differential equation satisfied by u.

Consider now a uniform finite difference grid with grid points (r_i, ϕ_j) for integers i and j with $i \geq 0$ and $0 \leq j \leq J - 1$ where $r_i = i\Delta r$, $\phi_j = j\Delta\phi$ for $\Delta r > 0$ and $\Delta\phi = 2\pi/J$. For a function u defined in a neighborhood of the origin P, we have

$$\bar{u}(l\Delta r) = \bar{u}(P, l\Delta r) = \frac{1}{J} \sum_{j=0}^{J-1} u_{l,j} + \mathcal{O}(\Delta\phi^m) \tag{2.4}$$

where $u_{l,j} = u(r_l, \phi_j)$ and m is a positive integer whose value will be considered in section 5. Using (2.3) we then have the relations

$$u(P) = \bar{u}(\Delta r) + \mathcal{O}(\Delta r^2) + \mathcal{O}(\Delta\phi^m) \tag{2.5}$$

and

$$u(P) = \frac{1}{3}(4\bar{u}(\Delta r) - \bar{u}(2\Delta r)) + \mathcal{O}(\Delta r^4) + \mathcal{O}(\Delta\phi^m) \tag{2.6}$$

which can be used with finite difference methods to determine values at the origin. Higher order formulas can be obtained by similar means.

3. The Laplacian with Polar Grids

When the differential equation being solved involves the laplacian operator then formula (2.3) can be used to a special advantage. We consider as examples the Poisson equation

$$\nabla^2 u = f \tag{3.1}$$

and the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \nabla^2 u \tag{3.2}$$

on a disk of unit radius.

For the Poisson equation (3.1) consider the semi-discrete finite difference approximation

$$\frac{1}{r_i \Delta r}\left(r_{i+\frac{1}{2}}\frac{u_{i+1}(\phi) - u_i(\phi)}{\Delta r} - r_{i-\frac{1}{2}}\frac{u_i(\phi) - u_{i-1}(\phi)}{\Delta r}\right) + \frac{1}{r_i^2}\frac{\partial^2 u_i(\phi)}{\partial\phi^2} = f_i(\phi) \text{ for } i > 0, \tag{3.3}$$

where we discretize only the radial direction. Employing (2.3) we have at the origin

$$u_0 = \bar{u}(\Delta r) - \frac{\Delta r^2}{4}\nabla^2 u(0) + \mathcal{O}(\Delta r^4)$$

or

$$u_0 = \bar{u}(\Delta r) - \frac{\Delta r^2}{4}f(0) - \mathcal{O}(\Delta r^4). \tag{3.4}$$

This formula maintains the second-order accuracy of the scheme and is easy to use. However, even when the equation being solved involves the laplacian, (2.6) may be more accurate or convenient to use than formulas such as (3.4). Formula (3.4) has been used by

Swarztrauber and Sweet [6] for solving the Poisson equation in a disk, and by Swarztrauber [7] for the Poisson equation on a sphere.

For the wave equation (3.2) we also consider a semi-discrete approximation in which only time and the radial direction are discretized with the angular variation continuous. Let $u_i^n(\phi)$ be the approximation to $u(n\Delta t, r_i, \phi)$. At the origin we have

$$
\begin{aligned}
u_0^n &= \bar{u}_0^n(\Delta r) - \frac{1}{4}\Delta r^2 \nabla^2 u_0^n + \mathcal{O}(\Delta r^4) \\
&= \bar{u}_0^n(\Delta r) - \frac{1}{4}\Delta r^2 \frac{\partial^2 u}{\partial t^2}\Big|_0^n + \mathcal{O}(\Delta r^4) \\
&= \bar{u}_0^n(\Delta r) - \frac{1}{4}\left(\frac{\Delta r}{\Delta t}\right)^2 (u_0^{n+1} - 2u_0^n + u_0^{n-1}) \\
&\quad + \mathcal{O}(\Delta r^4) + \mathcal{O}(\Delta r^2 \Delta t^2).
\end{aligned}
$$

using a central difference approximation in time. This gives the formula at the origin as

$$
u_0^{n+1} = 2u_0^n - u_0^{n-1} + 4\left(\frac{\Delta t}{\Delta r}\right)^2 (\bar{u}_0^n(\Delta r) - u_0^n) \tag{3.5}
$$

which maintains the second-order accuracy of the scheme. Example 1 in section 7 shows that this formula gives accurate results. Similar methods can also be used with parabolic equations.

4. Vector Fields with Polar Grids

In addition to the coordinate singularity at the origin, the polar coordinate representation of vector fields introduces an additional difficulty. Let $\vec{F}$ be a vector field defined on a domain on which there is a polar coordinate system. The polar coordinate representation assigns to each vector $\vec{F}(P)$ the component in the radial direction and the component in the direction of increasing angle. This representation is unique at all points other than the origin.

At the origin the vector $\vec{F}(0)$ has a different representation for each choice of the radial direction. This is best illustrated using the mapping between the polar and cartesian representations. Let (U, V) be the usual cartesian representation of the vector field $\vec{F}$ which is uniquely determined, then the polar representation (u, v) is given by

$$
\begin{aligned}
u &= U\cos\phi + V\sin\phi \\
v &= -U\sin\phi + V\cos\phi.
\end{aligned} \tag{4.1}
$$

Since at the origin the pair (U, V) is single valued, (4.1) shows the multivalued nature of the polar representation.

Using a polar grid the vector field F will be represented, and approximated, by vectors (u_{ij}, v_{ij}) at each grid point (r_i, ϕ_j). At the origin, there is a representation (u_{0j}, v_{0j}) for each coordinate direction $(0, \phi_j)$. For consistency these representations must be related by the formulas (4.1). That is, there are values (U_0, V_0) such that

$$u_{0j} = U_0 \cos\phi_j + V_0 \sin\phi_j$$

$$v_{0j} = -U_0 \sin\phi_j + V_0 \cos\phi_j. \tag{4.2}$$

The values of U_0 and V_0 can be obtained by formulas such as (2.6). For example, on a uniform grid define

$$\bar{U}(i\Delta r) = \frac{1}{J} \sum_{j=0}^{J-1} u_{ij} \cos\phi_j - v_{ij} \sin\phi_j \tag{4.3}$$

$$\bar{V}(i\Delta r) = \frac{1}{J} \sum_{j=0}^{J-1} u_{ij} \sin\phi_j + v_{ij} \cos\phi_j.$$

Then U_0 and V_0 can be approximated by

$$U_0 = \frac{1}{3}(4\bar{U}(\Delta r) - \bar{U}(2\Delta r)) \tag{4.4}$$

$$V_0 = \frac{1}{3}(4\bar{V}(\Delta r) - \bar{V}(2\Delta r)).$$

These values can then be used in (4.2) to give the values of (u_{0j}, v_{0j}). Example 3 in section 7 demonstrates the accuracy of this method as applied to the Stokes equations. This method has been used in Strikwerda [4] and Nagel and Strikwerda [3] with excellent results.

For finite difference grids which are not uniform in the angular variable formulas (4.3) should be replaced by

$$\bar{U}(i\Delta r) = \frac{1}{\sigma}\left(\sum_{j=0}^{J-1} u_{ij}(\sin\phi_{j+1} - \sin\phi_{j-1}) + v_{ij}(\cos\phi_{j+1} - \cos\phi_{j-1})\right) \tag{4.5}$$

$$\bar{V}(i\Delta r) = \frac{1}{\sigma}\left(\sum_{j=0}^{J-1} u_{ij}(\cos\phi_{j+1} - \cos\phi_{j-1}) + v_{ij}(\sin\phi_{j+1} - \sin\phi_{j-1})\right) \tag{4.6}$$

where

$$\sigma = 2 \sum_{j=0}^{J-1} \sin(\phi_{j+1} - \phi_j).$$

The formulas (4.5) and (4.6) are exact for the case when the vector field has constant cartesian components in a neighborhood of the origin.

5. The Fourier Method

We now consider the Fourier method for the approximation of derivatives with respect to ϕ. Consider a periodic discrete function f_j defined on grid points $\phi_j = j\Delta\phi$ with $\Delta\phi = 2\pi/J$. The object of both the finite difference and Fourier methods is to obtain approximations to $\partial f/\partial\phi$ at the grid points. The Fourier method begins with the finite Fourier series representation of f_j, i.e. for the case when J is an even integer

$$f_j = b_0 + \sum_{k=1}^{J/2-1} (a_k \sin k\phi_j + b_k \cos k\phi_j) + b_{\frac{J}{2}} \cos(\tfrac{J}{2}\phi_j). \tag{5.1}$$

Note that $\cos(\tfrac{J}{2}\phi_j) = (-1)^j$. Replacing ϕ_j in (5.1) by a continuous variable ϕ we can approximate $\partial f/\partial\phi$ at ϕ_j as

$$\frac{\partial f}{\partial\phi}\Big|_j \simeq \sum_{k=1}^{J/2-1} a_k k \cos k\phi_j - b_k k \sin k\phi_j \tag{5.2}$$

and similarly

$$\frac{\partial^2 f}{\partial\phi^2}\Big|_j \simeq -\sum_{k=1}^{J/2-1} (a_k k^2 \sin k\phi_j - b_k k^2 \cos k\phi_j) - (-1)^j (\tfrac{J}{2})^2 b_{\frac{J}{2}}.$$

The coefficients a_k and b_k are easily obtained by

$$b_k = \frac{2}{J} \sum_{j=0}^{J-1} f_j \cos k\phi_j \tag{5.3}$$

$$a_k = \frac{2}{J} \sum_{j=0}^{J-1} f_j \sin k\phi_j$$

for $0 < k < J/2$, and

$$b_0 = \frac{1}{J} \sum_{j=0}^{J-1} f_j \tag{5.4}$$

$$b_J = \frac{1}{J} \sum_{j=0}^{J-1} (-1)^j f_j.$$

The Fourier method has the advantage that it gives far higher accuracy for a given number of grid points than do finite difference methods (Gottlieb and Orzag [2]). Alternatively to attain a given accuracy the Fourier method requires significantly fewer grid points than do finite difference methods. For example 2 of section 7, finite difference methods would require at least three times as many grid points in the angular direction to obtain comparable accuracy.

The efficiency gained by the Fourier method over the finite difference method for the angular variation is due to the natural periodicity in the variable ϕ. Spectral methods can be used with the radial variation but not necessarily with the same gain in efficiency, Gottlieb and Orzag [2].

Line successive-over-relaxation (LSOR) can easily be used to solve elliptic boundary value problems in polar coordinates in which the Fourier method is used to approximate the derivatives with respect to ϕ. The basic formula for LSOR as applied to the semi-discrete approximation (3.3) is

$$\left(-\frac{2}{\Delta r^2} + \frac{1}{r_i^2}\frac{\partial^2}{\partial\phi^2}\right)(u_i^{\nu+1}(\phi) - u_i^{\nu}(\phi)) \tag{5.5}$$

$$= -\omega\left(\frac{1}{r_i\Delta r}\left(r_{i+\frac{1}{2}}\frac{u_{i+1}^{\nu}(\phi) - u_i^{\nu}(\phi)}{\Delta r} - r_{i-\frac{1}{2}}\frac{u_i^{\nu}(\phi) - u_{i-1}^{\nu+1}(\phi)}{\Delta r}\right) + \frac{1}{r_i^2}\frac{\partial^2 u_i^{\nu}}{\partial\phi^2} - f_i(\phi)\right).$$

In (5.5) the order of progression through the grid for the LSOR is in the direction of decreasing radius. When the Fourier method is used to approximate the derivatives with respect to ϕ, the Fourier coefficients of the update, $u_i^{\nu+1} - u_i^{\nu}$ can be easily obtained from the coefficients of the right-hand side of (5.5). That is, the right-hand side of (5.5) is evaluated for each value of j, then the Fourier coefficients are calculated. Dividing the coefficients of the kth node by $-(2/\Delta r^2 + k^2/r_i^2)$ gives the coefficients of the update, from which the update is determined at each value of j. This method is used in examples 2 and 3 of section 7.

6. Quadrature Formulas

The approximation of integrals by sums arises in several contexts in the use of finite difference methods on polar grids. As we have seen the approximations at the origin (2.5) and (2.6) use integrals in ϕ at various values of r. Also, the accurate determination of integral quantities over the domain requires quadrature formulas in r and ϕ.

We begin by considering integration in the angular variable only. We consider a 2π-periodic function $f(\phi)$. We first consider the error resulting from approximating the integral

$$\int_0^{2\pi} f(\phi)d\phi \tag{6.1}$$

by the sum

$$\sum_{j=0}^{J-1} f(\phi_j)\Delta\phi \tag{6.2}$$

where $\Delta\phi = 2\pi/J$ and $\phi_j = j\Delta\phi$. By the theory of Fourier series we have

$$f(\phi) = \sum_{n=-\infty}^{\infty} a_n e^{in\phi} \tag{6.3}$$

where

$$a_n = \frac{1}{2\pi} \int_0^{2\pi} f(\phi) e^{-in\phi} d\phi. \tag{6.4}$$

Thus

$$\sum_{j=0}^{J-1} f(\phi_j) \Delta\phi = \sum_{j=0}^{J-1} \sum_{n=-\infty}^{\infty} a_n e^{in\phi_j} \Delta\phi = \sum_{k=-\infty}^{\infty} a_{kJ},$$

since $\sum_{j=0}^{J-1} e^{in\phi_j}$ vanishes unless n is a multiple of J. The integral (6.1) is precisely a_0 thus the error in the approximation (6.2) is

$$2\pi \sum_{k=-\infty, k\neq 0}^{\infty} a_{kJ}.$$

By the definition of the a_n in (6.4) we have

$$a_n = \frac{1}{2\pi} (in)^{-m} \int_0^{2\pi} f^{(m)}(\phi) e^{-in\phi} d\phi$$

if f is m times differentiable. Thus

$$|a_n| \leq C_m |n|^{-m}.$$

for some constant C_m depending on f. Thus the error in the approximation (6.2) is bounded by

$$2C_m \sum_{k=1}^{\infty} |kJ|^{-m} = O(J^{-m}) = O(\Delta\phi^m). \tag{6.5}$$

We now consider quadrature for the unit circle using uniform spacing in r and ϕ. We consider

$$\int_0^1 \int_0^{2\pi} f(r,\phi) \, r \, dr \, d\phi = 2\pi \int_0^1 \bar{f}(r) \, r \, d\phi \tag{6.6}$$

where $\bar{f}(r)$ may be approximated to $O(\Delta\phi^m)$ as in (6.2). Using the trapezoid formula we have

$$2\pi \int_0^1 f(r) r \, dr = 2\pi \sum_{i=0}^{I-1} \frac{1}{2} (\bar{f}_i r_i + \bar{f}_{i+1} r_{i+1}) \Delta r + O(\Delta r^2)$$

$$= 2\pi \sum_{i=1}^{I-1} \bar{f}_i r_i \Delta r + \pi \bar{f}_I r_I \Delta r + O(\Delta r^2).$$

Hence

$$\int_0^1 \int_0^{2\pi} f(r,\phi) r\, dr\, d\phi = \sum_{i=1}^{I-1} \sum_{j=0}^{J-1} f_{ij} r_i \Delta r \Delta \phi + \frac{1}{2} \sum_{j=0}^{J-1} f_{Ij} r_I \Delta r \Delta \phi + O(\Delta r^2, \Delta \phi^m). \tag{6.7}$$

7. Computational Results

In this section we present results of computations using the formulas discussed in the previous sections applied to three test problems. The first test problem is to solve the second-order wave equation. The two formulas (2.6) and (3.5) for determining the solution at the origin are compared. The second test problem is to solve an elliptic equation using the LSOR method given in section 4 to solve the discrete equations. The third test problem uses the Stokes equations to illustrate the use of the formulas for vector fields at the origir.

The first test problem was to solve the second-order wave equation

$$u_{tt} = \nabla^2 u \tag{7.1}$$

in the unit disk for $0 \le t \le 1$. The exact solution we used was

$$u(t, x, y) = \cos(t - .6x - .8y). \tag{7.2}$$

The equation for the time advancement is

$$u_{ij}^{n+1} = 2u_{ij}^n - u_{ij}^{n-1} + (\Delta t)^2 \nabla_h^2 u_{ij}^n, \tag{7.3}$$

where the discrete laplacian, ∇_h^2, is given by the left-hand side of (3.3) and the derivatives with respect to ϕ are approximated by the Fourier method. The formula for the first time-step is based on a Taylor series in time and is

$$u_{ij}^1 = u_{ij}^0 + \Delta t (u_t)_{ij}^0 + \frac{1}{2}(\Delta t)^2 \nabla_h u_{ij}^0. \tag{7.4}$$

where u_{ij}^0 and $(u_t)_{ij}^0$ were obtained from the exact solution.

Both the interpolation formula (2.6) and the formula (3.5) were used to determine the solution at the origin. The interpolation formula was applied using $\bar{u}(\Delta r)$ and $\bar{u}(2\Delta r)$ at the given time level to compute u at the origin for that same time level. The results of four test cases are displayed in Table 1, where I and J are the number of radial and angular grid points, respectively, and K is the number of time steps. Both the L^2 norm of

the error and the error at the origin are shown for each case. The two formulas are seen to be comparable in accuracy, but the interpolation formula is slightly more accurate. This was also observed in all the other cases in which these two formulas were compared. Since formulas (2.6) and (3.5) yielded comparable results, we used the simpler formula (2.6) for all subsequent runs.

Table I. Two Center Methods

GRID	FORMULA 1		FORMULA 2	
I $\quad J$ $\quad K$	$\|u_{err}\|$	c_{err}	$\|u_{err}\|$	c_{err}
21 16 160	.5586(-4)	.1010(-3)	.5685(-4)	.1069(-3)
41 20 220	.1662(-4)	.3177(-4)	.1623(-4)	.3105(-4)

A list of cases using formula (2.6) is given in Table II. The data show that errors are relatively insensitive to J, the number of angular grid points, for the chosen values of J. The number of time steps must be chosen so that the scheme is stable. No attempt was made to determine the stability condition for this example. Because the accuracy of the scheme depends on the three parameters I, J, and K it is difficult to discern the order of accuracy of the scheme.

Table II. Wave Equation Results

I $\quad J$ $\quad K$	$\|u_{err}\|$	c_{err}
11 12 40	.2050(-3)	.3511(-3)
21 12 80	.5314(-4)	.9447(-4)
21 16 100	.5294(-4)	.9525(-4)
21 20 120	.5397(-4)	.9727(-4)
41 12 160	.1595(-4)	.2998(-4)
41 16 180	.1623(-4)	.3117(-4)
41 20 320	.1662(-4)	.3177(-4)

Tests were made using a non-uniform radial grid for this test case. We found that in all cases it degraded the accuracy. Further study is needed to determine if non-uniform grids can be used to give good accuracy and less restrictive stability limitations on the time step.

The second test problem was to solve the elliptic equation

$$\nabla^2 u - c(x,y)u = 0. \tag{7.5}$$

on the unit disk with u specified on the boundary. The exact solution was given by

$$u(x,y) = \exp((x - 0.1)(y - 0.5)) \tag{7.6}$$

with

$$c(x,y) = (x - 0.1)^2 + (y - 0.5)^2. \tag{7.7}$$

The equation (7.1) was approximated using the left-hand side of (3.3) for the laplacian with the Fourier method being used to approximate the derivatives with respect to ϕ. The solutions were obtained with the LSOR method discussed at the end of section 5. If finite difference methods are used in the angular variable, then a direct solver such as that of Swarztrauber and Sweet [6] can be used.

The results of several test runs are displayed in Table III. The number of grid points is given along with the iteration parameter ω and the tolerance on the updates. The iterative procedure was stopped when

$$\|u^{n+1} - u^n\|/\omega \leq tol. \tag{7.8}$$

The number of iterations required for convergence is seen to be dependent on the number of radial grid points and not on the number of angular grid points. The accuracy at the origin is seen to be relatively independent of the value of J, as is expected from the analysis of section 5. This was also noted for test problem 1. The norm of the error is, however, dependent on J for small values of J. If J is sufficiently large then the second-order accuracy of the scheme is seen. A center formula similar to (3.4) gave results comparable to those obtained by (2.6). The results shown were obtained by using center formula (2.6).

The third test problem was to solve the Stokes equations

$$\nabla^2 u - \frac{u}{r^2} - 2\frac{1}{r^2}\frac{\partial v}{\partial \phi} - \frac{\partial p}{\partial r} = 0$$

$$\nabla^2 v - \frac{v}{r^2} + 2\frac{1}{r^2}\frac{\partial u}{\partial \phi} - \frac{\partial p}{r\partial \phi} = 0 \tag{7.9}$$

$$\frac{1}{r}\frac{\partial ru}{\partial r} + \frac{1}{r}\frac{\partial v}{\partial \phi} = 0$$

Table III. Poisson equation results

I	J	iter	ω	tol	$\|u_{err}\|$	c_{err}
11	12	42	1.5	1(-5)	.13(-2)	-.11(-2)
11	16	42	1.5	1(-5)	.13(-2)	-.11(-2)
21	8	53	1.8	1(-6)	.39(-2)	-.63(-3)
21	12	52	1.8	1(-6)	.38(-3)	-.28(-3)
21	16	52	1.8	1(-6)	.35(-3)	-.29(-3)
41	12	131	1.9	1(-7)	.19(-3)	-.74(-4)
41	16	131	1.9	1(-7)	.90(-4)	-.75(-4)
41	20	131	1.9	1(-7)	.90(-4)	-.75(-4)
61	12	281	1.9	1(-7)	.17(-3)	-.26(-4)
61	16	283	1.9	1(-7)	.32(-4)	-.27(-4)
61	20	281	1.9	1(-7)	.30(-4)	-.26(-4)
61	24	286	1.9	1(-7)	.32(-4)	-.27(-4)

on the unit disk with the velocity components u and v given on the boundary.

The exact solution was given by

$$u(r,\phi) = r\sin\phi(r\cos\phi - a)(r - a\cos\phi)/R^4 + \frac{1}{2}(r - a\cos\phi)/R^2$$

$$v(r,\phi) = a\,r\sin^2\phi(r\cos\phi - a)/R^4 + \frac{1}{2}a\sin\phi/R^2$$

$$p(r,\phi) = 2r\sin\phi(r\cos\phi - a)/R^4$$

with

$$R^2 = r^2 + a^2 - 2ar\cos\phi \tag{7.10}$$

where a had the value 1.5. Notice that for this solution the polar representation of the solution at the origin is multiply valued with

$$(u(0,\phi), v(0,\phi)) = (\frac{\cos\phi}{2a}, \frac{\sin\phi}{2a}). \tag{7.11}$$

which corresponds to a vector of magnitude $(2a)^{-1}$ in the direction of the negative x-axis

The system (7.9) was approximated using the discrete laplacian as given in (3.3) with the Fourier approximation of derivatives with respect to ϕ. The system was solved with the iterative method as given in Strikwerda [1984b] with the LSOR method used to update the velocities. Explicitly the formulas are:

$$(-\frac{2}{\Delta r^2} + \frac{1}{r_i^2}\frac{\partial^2}{\partial\phi^2})(\Delta u_{i,j}^\nu) \tag{7.12}$$

$$= -\omega(\frac{1}{r_i\Delta r}\left(r_{i+1/2}\frac{(u_{i+1,j}^\nu - u_{i,j}^\nu)}{\Delta r^2} - r_{i-1/2}\frac{(u_{i,j}^\nu - u_{i-1,j}^{\nu+1})}{\Delta r^2}\right)$$

$$+ \frac{1}{r_i^2}\frac{\partial^2 u_{i,j}^\nu}{\partial\phi^2} - \frac{u_{i,j}^\nu}{r_i^2} - \frac{2}{r_i^2}\frac{\partial v_{i,j}^\nu}{\partial\phi}$$

$$- \left(\frac{p_{i+1,j}^\nu - p_{i-1,j}^\nu}{2\Delta r} - \frac{p_{i+2,j} - 3p_{i+1,j} + 3p_{i,j} - p_{i-1,j}}{6\Delta r}\right))$$

and

$$(-\frac{2}{\Delta r^2} + \frac{1}{r_i^2}\frac{\partial^2}{\partial\phi^2})(\Delta v_{i,j}^\nu) \tag{7.13}$$

$$= -\omega(\frac{1}{r_i\Delta r}\left(r_{i+1/2}\frac{(v_{i+1,j}^\nu - v_{i,j}^\nu)}{\Delta r^2} - r_{i-1/2}\frac{(v_{i,j}^\nu - v_{i-1,j}^{\nu+1})}{\Delta r^2}\right)$$

$$+ \frac{1}{r_i^2}\frac{\partial^2 v_{i,j}^\nu}{\partial\phi^2} - \frac{v_{i,j}^\nu}{r_i^2} + \frac{2}{r_i^2}\frac{\partial u_{i,j}^\nu}{\partial\phi} - \frac{1}{r}\frac{\hat{\partial}p_{i,j}}{\partial\phi})$$

with

$$u_{i,j}^{\nu+1} = u_{i,j}^\nu + \Delta u_{i,j}^\nu$$

and

$$v_{i,j}^{\nu+1} = v_{i,j}^\nu + \Delta v_{i,j}^\nu$$

The pressure was updated by

$$p_{i,j}^{\nu+1} = p_{i,j}^\nu - \gamma\left(\frac{1}{r_i}\frac{(r_{i+1}u_{i+1,j}^{\nu+1} - r_{i-1}u_{i-1,j}^{\nu+1})}{2\Delta r}\right. \tag{7.14}$$

$$\left. - \frac{u_{i+1,j}^{\nu+1} - 3u_{i,j}^{\nu+1} + 3u_{i-1,j}^{\nu+1} - u_{i-2,j}^{\nu+1}}{6\Delta r} + \frac{1}{r_i}\frac{\hat{\partial}v_{i,j}^{\nu+1}}{\partial\phi}\right),$$

where γ is an iteration parameter. as described in Strikwerda [5] The third-order differences with respect to γ in (7.12) and (7.14) are necessary to preserve the regularity of the scheme

and hence the smoothness of the solution, e.g. Bube and Strikwerda [1] and Strikwerda [4]. The derivatives with respect to ϕ which are marked with a carat in (7.13) and (7.14) are computed as in the Fourier method with the addition of the term

$$\pm b_{J/2}\left(\frac{J}{4}\right)(-1)^j. \tag{7.15}$$

where the plus sign is used in (7.14) and the minus sign in (7.13). These terms are included to ensure the regularity of the scheme and hence the smoothness of the solution. Without terms such as these the solution would contain Fourier modes with wavelength $2\Delta\phi$ of sufficient amplitude to affect the accuracy of the solution.

The results of test problem 3 are displayed in Tables IV and V. In Table IV the l^2 norms of the error are displayed for the velocity components and the pressure. That is,

$$\|u_{\text{err}}\| = \left(\frac{1}{\pi}\sum_{i=1}^{I-1}\sum_{j=0}^{J-1}|u(r_i,\phi_j) - u_{i,j}|^2 r_i\Delta r\Delta\phi\right)^{1/2}$$

where the initial factor of π^{-1} is included to normalize by the area of the disc. The error for v is computed similarly. The expression $u(r_i,\phi_j)$ is the exact solution evaluated at (r_i,ϕ_j) and $u_{i,j}$ is the computed solution at that grid point.

Table IV. Norm Errors for Problem 3

I	J	$\|u_{\text{err}}\|$	$\|v_{\text{err}}\|$	$\|p_{\text{err}}\|$	δ_h
11	16	1.1(-3)	9.1(-4)	.34	-1.5(-3)
11	24	3.9(-5)	3.3(-5)	3.2(-2)	-5.9(-5)
21	32	2.1(-6)	2.1(-6)	4.0(-3)	-2.3(-6)
41	40	1.1(-7)	1.2(-7)	3.6(-4)	-1.5(-7)

Because the pressure is defined only to within an additive constant we use

$$\|p_{\text{err}}\| = \left(\frac{1}{\pi}\sum_{i=1}^{I}\sum_{j=0}^{J-1}|p(r_i,\phi_j) - p_{i,j} - \overline{\Delta p}|^2 \varepsilon_i r_i\Delta r\Delta\phi\right)^{1/2} \tag{7.16}$$

where

$$\varepsilon_i = \begin{cases} 1/2, & \text{if } i = 0 \text{ or } I; \\ 1, & \text{otherwise}, \end{cases}$$

as required by the trapezoid rule (6.7), and $\overline{\Delta p}$ is the average value of $p(r_i, \phi_j) - p_{i,j}$ computed over the disc, i.e.

$$\overline{\Delta p} = \frac{1}{\pi} \sum_{i=1}^{I} \sum_{j=0}^{J-1} (p(r_i, \phi_j) - p_{i,j}) \varepsilon_i r_i \Delta r \Delta \phi. \tag{7.17}$$

Also displayed in Table IV is the value of δ_h which is the average of the discrete approximation to the divergence of the velocity field. The finite difference and Fourier scheme does not enforce the condition

$$\vec{\nabla} \cdot \vec{u} = 0$$

on the discrete solution, rather the iterative method converges to a solution with

$$\vec{\nabla}_h \cdot \vec{u}_{i,j} = \delta_h \tag{7.18}$$

where δ_h is the average of the left-hand side of (7.18). Thus the value of δ_h is an

a posteriori indicator of the accuracy of the discrete solution. As seen in Table IV the numerical method gives very good solutions to the Stokes equation.

Table V. Errors at the center for Problem 3

I	J	u	v	p
11	16	2.8(-3)	8.1(-4)	-1.8(-7)
11	24	8.6(-5)	1.8(-3)	1.7(-7)
21	32	7.3(-5)	4.0(-4)	9.9(-8)
41	40	7.2(-6)	1.1(-4)	1.2(-7)

Table V displays the errors at the center for test problem 3. The errors displayed for u and v are the errors for ϕ equal to 0 at the origin. The error in the pressure is

$$p(0) - p_{i,j} - \overline{\Delta p}$$

where $\overline{\Delta p}$ is the average of the difference between $p(r_i, \phi_j)$ and $p_{i,j}$ taken over the whole disc given by (7.17). Note that the computation of $\overline{\Delta p}$ does not use the values at the origin.

Table VI. Iteration Parameters for Problem 3

I	J	γ	ω	tol	iterations
11	16	.4	1.6	10(-4)	73
11	24	.4	1.6	10(-4)	65
21	32	.2	1.7	10(-4)	188
41	40	.1	1.8	10(-5)	596

Table VI gives the iteration parameters and the resulting number of iterations for each of the cases reported in Tables IV and V. The iterative method was considered to have converged when the successive changes in u and v were less than than *tol* times ω and when the changes in p deviated from its average value by less than *tol* times γ. That is, when the value of

$$\left(\sum_{i=0}^{I} \sum_{j=0}^{J-1} (\Delta p_{i,j}^{\nu} - \overline{\Delta p^{\nu}})^2 \varepsilon_i r_i \Delta r \Delta \phi \right)^{1/2} \tag{7.19}$$

was less than *tol* times γ the solution was considered converged.

The values of the expressions (7.16) and (7.19) were each computed in one pass through the data by the following modification of West's algorithm [West 8]. To compute the value of

$$Q = \sum_{i,j=0}^{I,J-1} (X_{ij} - \overline{X})^2 \alpha_{ij}$$

where

$$\overline{X} = \sum_{i,j=0}^{I,J-1} X_{ij} \alpha_{ij} / \sum_{i,j=0}^{I,J-1} \alpha_{ij}$$

initialize with

$$A = 0$$
$$Q = 0$$
$$\overline{X} = 0$$

Then for $i = 0$ to I, and $j = 0$ to $J - 1$,

$$Q \longleftarrow Q + A\alpha_{ij}(X_{ij} - \overline{X})^2/(A + \alpha_{ij})$$
$$\overline{X} \longleftarrow (A\overline{X} + \alpha_{ij}X_{ij})/(A + \alpha_{ij})$$
$$A \longleftarrow A + \alpha_{ij}.$$

(7.20)

Thus, to compute the expression (7.19) we have

$$X_{ij} = \Delta p_{i,j}^{\nu}$$

and

$$\alpha_{ij} = \begin{cases} r_i\Delta r\Delta\phi, & i \neq I \\ \frac{1}{2}r_i\Delta r\Delta\phi, & i = I, 0. \end{cases}$$

This algorithm makes the computation of expressions (7.16) and (7.19) only slightly more difficult than the computation of the usual norm.

Conclusions

The results of the test problems show that the methods presented in this paper can be used to compute accurate solutions to equations on domains with polar grids. The basic formulas can be used with most numerical procedures. The use of the Fourier method, while not essential to the center formulas, is very convenient and efficient to use with polar grids.

References

1. K.P. Bube and J.C. Strikwerda, Interior regularity estimates for elliptic systems of difference equations. SIAM J. Num. Anal., 20. (1983), pp. 639-656.

2. D. Gottlieb and S.A. Orszag. Numerical Analysis of Spectral Methods: Theory and Applications, SIAM, Philadelphia, PA.(1977)

3. Y. Nagel and J.C. Strikwerda, A numerical study of the flow in a spinning and coning cylinder, to appear.

4. J.C. Strikwerda, Finite difference methods for the Stokes and Navier-Stokes equations. SIAM J. Sci. Stat. Comput. 5, (1984), pp. 56-68.

5. J.C. Strikwerda, An iterative method for solving finite difference approximations to the Stokes equations. SIAM J. Numer. Anal. 21, (1984), pp. 447-458.

6. P.N. Swarztrauber and R.A. Sweet, The direct solution of the discrete Poisson equation on a disk, SIAM J. Numer. Anal. 10, (1973), pp. 900-907.

7. P.N. Swarztrauber, The direct solution of the discrete Poisson equation on the surface of a sphere, J. Comp. Phys. 15, (1974), pp. 46-54.

8. D.H.D. West, Updating mean and variance estimates: An improved method, Comm. ACM. 22, (1979), pp. 532-535.

ADAPTIVE FINITE ELEMENT METHODS FOR PARABOLIC SYSTEMS IN ONE AND TWO SPACE DIMENSIONS[1]

Slimane Adjerid

Department of Computer Science
and
Center for Applied Mathematics and Advanced Computation
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. Munition. and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

Abstract. We discuss adaptive finite element methods for solving initial-boundary value problems for vector systems of parabolic partial differential equations in one and two space dimensions.

One-dimensional systems are discretized using piecewise linear finite element approximations in space and a backward difference code for stiff ordinary differential systems in time. A spatial error estimate is calculated using piecewise quadratic approximations that employ nodal superconvergence to increase computational efficiency. This error estimate is used to move and refine the finite element mesh in order to equidistribute a measure of the total spatial error and to satisfy a prescribed error tolerance. Ordinary differential equations for the spatial error estimate and the mesh motion are integrated in time using the same backward difference software that is used to determine the finite element solution.

Two-dimensional systems are discretized using piecewise bilinear finite element approximations in space and backward difference software in time. A spatial error estimate is calculated using piecewise cubic approximations that take advantage of nodal superconvergence. This error estimate is used to locally refine a stationary finite element mesh in order to satisfy a prescribed spatial error tolerance.

Some examples are presented in order to illustrate the effectiveness of our error estimation technique and the performance of our adaptive algorithm.

---

1. Introduction. Adjerid and Flaherty [1-3] developed adaptive finite element methods for solving m-dimensional vector systems of partial differential equations having the form

$$\mathbf{M}(\mathbf{x},t)\mathbf{u}_t + \mathbf{f}(\mathbf{x},t,\mathbf{u},\nabla\mathbf{u}) = \sum_{k=1}^{d} [\mathbf{D}^k(\mathbf{x},t,\mathbf{u})\mathbf{u}_{z_k}]_{z_k}, \quad \mathbf{x} \in \Omega, \quad t > 0, \qquad (1a)$$

subject to the initial and boundary conditions

$$\mathbf{u}(\mathbf{x},0) = \mathbf{u}^0(\mathbf{x}), \quad \mathbf{x} \in \Omega \bigcup \partial\Omega. \qquad (1b)$$

$$\text{either} \quad u_i(\mathbf{x},t) = c_i(\mathbf{x},t) \quad \text{or} \quad \sum_{k=1}^{d}\sum_{j=1}^{m} D_{ij}^k u_{j_{z_k}}(\mathbf{x},t)\nu_k = c_i(\mathbf{x},t),$$

$$\text{for } z \in \partial\Omega, \quad t > 0, \quad i = 1, 2, ..., m. \qquad (1c)$$

They considered problems in one $(d = 1)$ and two $(d = 2)$ spatial dimensions with $\mathbf{x} = [x_1, ..., x_d]^T$ denoting a position vector in $\mathbf{R}^d$, t denoting time, and Ω being either a segment of the real line or a rectangle. The subscripts t and x_k denote temporal and spatial partial derivatives, respectively, and $\nu = [\nu_1, ..., \nu_d]^T$ denotes the unit outer normal vector to the boundary $\partial\Omega$ of Ω. Problems were assumed to be parabolic and have an isolated solution; thus, $\mathbf{M}$ and $\mathbf{D}^k$, $k = 1, ..., d$, are positive definite $m \times m$ matrices.

Adjerid and Flaherty discretized (1) in space using Galerkin's method with a piecewise linear polynomial basis in one dimension and piecewise bilinear polynomials in two dimensions. An a posteriori estimate of the spatial discretization error was calculated using Galerkin's method with piecewise quadratic functions in one dimension and piecewise cubic functions in two dimensions. In each case, a nodal superconvergence property of the finite element method was used to neglect errors at nodes and, thus, improve computational efficiency. The error estimate was used to control global [1] and local [2, 3] refinement procedures that added and/or deleted finite elements to the mesh in order to satisfy a prescribed global measure of the spatial discretization error. For one-dimensional problems, the error estimate was further used to move the finite element mesh so as to equidistribute the global error measure. Ordinary differential equations for the finite element solution, error estimate, and, in one dimension, mesh motion were integrated in time using the backward difference code DASSL [18] for stiff differential and algebraic systems.

Initially, a global refinement procedure was used in combination with mesh motion to satisfy prescribed error tolerances in the H^1 norm [1]. This procedure was replaced by a more efficient local mesh refinement strategy and some problem dependent parameters were removed from the mesh moving scheme [2]. In particular, numerical experiments indicated that the performance of the error estimation procedure could deteriorate when the system of equations governing mesh motion was too stiff. Adjerid and Flaherty [2] remedied this defect by limiting the stiffness of the mesh moving equations and using refinement, instead of mesh motion, to equidistribute the error estimate in these situations. They subsequently extended their finite element, error estimation and adaptive local refinement procedures to two-dimensional parabolic problems [3] and proved that the error estimate of [1, 2] converged to the true discretization error in H^1 as the mesh is refined for linear one-dimensional parabolic systems [4].

In Section II of this paper, we review the one-dimensional adaptive procedures of Adjerid and Flaherty [1, 2], describe some improvements to their mesh refinement scheme, and present some examples that illustrate the relationship and interaction between mesh motion and refinement. The essential details of Adjerid and Flaherty's [3] two-dimensional procedure and the dynamic data structures used in its implementation are summarized in Section III. The results of a nonlinear two-dimensional example are also presented in Section III. Finally, in Section IV, we discuss our results and suggest some future directions.

II. ONE-DIMENSIONAL ADAPTIVE PROCEDURES.

The one-dimensional version of problem (1) consists of solving

$$\mathbf{M}(x,t)\mathbf{u}_t + \mathbf{f}(x,t,\mathbf{u},\mathbf{u}_x) = [\mathbf{D}(x,t,\mathbf{u})\mathbf{u}_x]_x, \quad x \in (a,b), \quad t > 0. \tag{2a}$$

$$\mathbf{u}(x,0) = \mathbf{u}^0(x), \quad x \in [a,b] \tag{2b}$$

$$\text{either} \quad u_i(x,t) = c_i(t) \quad \text{or} \quad \sum_{j=1}^{m} D_{i,j} u_{j,x}(x,t) = c_i(t).$$
$$\text{for } x = a, b, \quad t > 0, \quad i = 1, 2, \ldots, m. \tag{2c}$$

The unit subscripts on x and superscripts on $\mathbf{D}$ have been omitted for simplicity.

The procedure for discretizing (2) and estimating the spatial discretization error of its solution are identical to our earlier work [1, 2] and are briefly summarized in Section II.1. The essential details of our current adaptive procedure are presented in Section II.2 and some examples illustrating its capabilities and the interplay between mesh motion and refinement are presented in Section II.3.

II.1. Discrete System. We construct a weak form of (2) by assuming $\mathbf{u} \in H_E^1$, selecting a test function $\mathbf{v} \in H_0^1$, multiplying (2a) by $\mathbf{v}$, integrating it on $a \leqslant x \leqslant b$, and integrating the diffusive term by parts to obtain

$$(\mathbf{v}, \mathbf{M}\mathbf{u}_t) + (\mathbf{v}, \mathbf{f}) + A(\mathbf{v}, \mathbf{u}) = \mathbf{v}^T \mathbf{D}(x,t,\mathbf{u})\mathbf{u}_x \,|_a^b, \quad \text{for all } \mathbf{v} \in H_0^1, \quad t > 0, \tag{3a}$$

where

$$(\mathbf{v}, \mathbf{u}) = \int_a^b \mathbf{v}(x,t)^T \mathbf{u}(x,t)\, dx, \quad A(\mathbf{v}, \mathbf{u}) = \int_a^b \mathbf{v}_x^T \mathbf{D}(x,t,\mathbf{u})\mathbf{u}_x \, dx. \tag{3b,c}$$

Recall that the Sobolev space H^1 consists of functions that are square integrable and have square integrable first spatial derivatives. Functions belonging to H_E^1 are further restricted to satisfy any essential (Dirichlet) boundary conditions in (2c), while functions in H_0^1 must satisfy homogeneous versions of any essential boundary conditions.

Initially $\mathbf{u}$ must satisfy

$$(\mathbf{v}, \mathbf{u}) = (\mathbf{v}, \mathbf{u}^0), \quad \text{for all } \mathbf{v} \in H_0^1, \quad t = 0. \tag{3d}$$

1079

and any natural (Neumann) boundary conditions in (2c) should be used to replace $\mathbf{D}u_x$ in the last term of (3a).

Finite element solutions of (3) are constructed by selecting finite dimensional approximations $\mathbf{U} \in S_E^N \subset H_E^1$ and $\mathbf{V} \in S_0^N \subset H_0^1$ of $\mathbf{u}$ and $\mathbf{v}$, respectively, and finding $\mathbf{U}$ such that

$$(\mathbf{V}.\mathbf{M}\mathbf{U}_t) + (\mathbf{V}.\mathbf{f}) + A(\mathbf{V}.\mathbf{U}) = \mathbf{V}^T \mathbf{D}(x,t,\mathbf{U})\mathbf{U}_x |_a^b, \quad \text{for all } \mathbf{V} \in S_0^N, \ t > 0.$$

(4a)

$$(\mathbf{V},\mathbf{U}) = (\mathbf{V},\mathbf{u}^0), \quad \text{for all } \mathbf{V} \in S_0^N, \ t > 0.$$

(4b)

Specifically, we introduce a partition

$$\pi(t,N) := \{ a = x_0(t) < x_1(t) < \cdots < x_N(t) = b \}$$

(5)

of $[a,b]$ into N moving subintervals $(x_{i-1}(t),x_i(t))$, $i = 1, 2, \ldots, N$, $t \geqslant 0$, and select S_E^N and S_0^N to consist of piecewise linear polynomials with respect to this partition. The system of ordinary differential equations that result from this spatial discretization can be integrated in time using one of the many excellent software packages for solving stiff differential systems. We found that the backward difference code DASSL (cf. Petzold [18]) for differential and algebraic systems best fit our purposes.

The spatial discretization error of the finite element solution

$$e(x,t) = \mathbf{u}(x,t) - \mathbf{U}(x,t)$$

(6)

satisfies (3) with $\mathbf{u}$ replaced by $\mathbf{U} + e$, i.e.,

$$(\mathbf{v}.\mathbf{M}(\mathbf{U}_t + e_t)) + (\mathbf{v}.\mathbf{f}(\cdot,t,\mathbf{U}+e,\mathbf{U}_x+e_x)) + A(\mathbf{v}.\mathbf{U}+e) =$$

$$\mathbf{v}^T \mathbf{D}(x,t,\mathbf{U}+e)(\mathbf{U}_x + e_x)|_a^b, \quad \text{for all } \mathbf{v} \in H_0^1, \quad t > 0.$$

(7a)

$$(\mathbf{v},e) = (\mathbf{v},\mathbf{u}^0 - \mathbf{U}), \quad \text{for all } \mathbf{v} \in H_0^1, \ t = 0.$$

(7b)

We approximate e by a function $\mathbf{E} \in \hat{S}_0^N$, where $\hat{S}_0^N$ is a finite dimensional subspace of H_0^1 consisting of piecewise quadratic functions that vanish on $\pi(t,N)$. We further approximate $\mathbf{v}$ by $\mathbf{V} \in \hat{S}_0^N$ and determine $\mathbf{E}$ as the solution of

$$(\mathbf{V},\mathbf{M}(\mathbf{U}_t + \mathbf{E}_t)) + (\mathbf{V}.\mathbf{f}(\cdot,t,\mathbf{U}+\mathbf{E},\mathbf{U}_x+\mathbf{E}_x)) + A(\mathbf{V}.\mathbf{U}+\mathbf{E}) = 0,$$

$$\text{for all } \mathbf{V} \in \hat{S}_0^N, t > 0,$$

(8a)

$$(\mathbf{V},\mathbf{E}) = (\mathbf{V},\mathbf{u}^0 - \mathbf{U}), \quad \text{for all } \mathbf{V} \in \hat{S}_0^N, \ t = 0.$$

(8b)

In constructing the error estimate $\mathbf{E}(x,t)$, we assumed the superconvergence of the piecewise linear finite element solution $\mathbf{U}(x,t)$, i.e., we assumed that $\mathbf{U}(x,t)$ converges at a faster rate on $\pi(t,N)$ than elsewhere on $a < x < b$. This

superconvergence property was established by Thomee [19] and the convergence of **E** to **e** has been proven for linear problems by Adjerid and Flaherty [4].

The error estimate $\mathbf{E}(x,t)$ is used to control the refinement/coarsening strategy and the motion of $\pi(t,N)$. We determine mesh motion by solving the ordinary differential system

$$\dot{x}_i(t) - \dot{x}_{i-1}(t) = -\lambda(W_i - \overline{W}), \quad i = 1, 2, \dots, N, \tag{9a}$$

where λ is a non-negative parameter, W_i is an error indicator on the subinterval (x_{i-1}, x_i). and $\overline{W}$ is the average of W_i, $i = 1, 2, \dots, N$. We shall take W_i to be the square of the local error estimate in H^1, i.e.,

$$W_i(t) = \|\mathbf{E}\|_{1,i}^2 := \int_{x_{i-1}}^{x_i} [\mathbf{E}^T\mathbf{E} + \mathbf{E}_x^T\mathbf{E}_x] \, dx; \tag{9b}$$

however. other local measures can be used [2].

When $\lambda > 0$ and $W_i > \overline{W}$, the right-hand side of (9a) is negative and the nodes x_i and x_{i-1} move closer to each other. Similarly. the nodes $x_i(t)$ and $x_{i-1}(t)$ move apart when $\lambda > 0$ and $W_i < \overline{W}$. Coyle et al. [16] studied the stability of (9a) with respect to small perturbations from an equidistributing mesh (i.e.. one where $W_i(t) = \overline{W}(t)$. $i = 1, 2, \dots, N$, $t \geqslant 0$) and showed that such perturbations could only grow by a bounded amount when $\lambda > 0$. $W_i > 0$. $i = 1, 2, \dots, N$. and the velocity of the equidistributing mesh remained finite for $t \geqslant 0$. They further showed that the mesh obtained by solving (9a) stayed closer to the equidistributing system when λ was large. This, however. introduces stiffness into the system which makes its solution expensive and, as noted. causes some difficulties with our error estimate. Adjerid and Flaherty [2] studied (9) and developed an adaptive algorithm for selecting λ as a function of t that balanced stiffness and equidistribution. The procedure for selecting λ will not be discussed further. but it has been used in the examples of Section II.3.

In order to maintain sparsity. we eliminate $\overline{W}$ by combining (9a) on two neighboring intervals and solve the scalar tridiagonal system

$$\dot{x}_{i-1} - 2\dot{x}_i + \dot{x}_{i+1} = -\lambda(W_{i+1} - W_i). \quad i = 1, 2, \dots, N-1. \quad t > 0. \tag{10}$$

The ordinary differential equations resulting from (8) and (10) are solved using the same backward difference software that is used to integrate the finite element system (4).

II.2. Adaptive Algorithms. In addition to controlling mesh motion. the error estimate **E** described in Section II.1 is used as an error indicator in conjunction with procedures that locally refine or coarsen the mesh. A top-level description of an adaptive local refinement/coarsening algorithm is presented in Figure 1 in a pseudo-PASCAL language. The procedure *adapfem* integrates the system (4, 8, 10) from time *tinitial* to *tfinal* and attempts to keep the spatial error estimate $\|\mathbf{E}\|_1 < TOL$, where TOL is a prescribed tolerance. The time steps that are selected by the temporal integration routine (e.g.. DASSL) are denoted as $\Delta t[m]$. $m = 1, 2, \dots$, and the corresponding times are $tout[m]$, $m = 0, 1, \cdots$. with $tout[0]$ initially set to *tinitial*. The integration is halted every *nstep* steps or when

$tout[m] = tfinal$ and the arrays Δt and tout are recomputed with $tout[0]$ reset to the last computed time, i.e.. $tout[nstep]$ or $tfinal$.

procedure adapfem ($tinitial$, $tfinal$, $nstep$, TOL):

  begin
    Calculate the initial conditions and an initial mesh;

        { Integrate the system from $tinitial$ to $tfinal$. }
    $m := 0$;
    $tout[0] := tinitial$;
    while $tout[m] < tfinal$ **do**
      begin
        $m := m + 1$;
        $redone := $ **false**;
        Integrate (4, 8, 10) for one time step $\Delta t[m]$;
        $tout[m] := tout[m-1] + \Delta t[m]$;

        { Check the error estimate. }

      if $(m = nstep)$ **or** $(tout[m] = tfinal)$ **then**
        begin
          Compute a new value of λ, if necessary:

          { Refine the mesh. }

          while $\|E(\cdot, tout[m])\|_1 > TOL$ **do**
            begin
              Add elements to the mesh:
              Redo the integration on the refined mesh from
                $t = tout[0]$ to $tout[m]$;
              $redone := $ **true**
            end:

          { Coarsen or regenerate the mesh. }

          if $(\|E(\cdot, tout[m])\|_1 < TOL / 3)$ **or** (redone)
            then Delete elements from the mesh. if possible
            else Generate a new mesh, if necessary:

          $tout[0] := tout[m]$:
          $m := 0$
        end { **if** $m = nstep$ ... }
      end { **while** $tout[m] < $ tfinal }
  end { adapfem }:

Figure 1. Top-level description of an adaptive finite element procedure with mesh motion and/or local mesh refinement/coarsening.

The spatial error estimate $\|\mathbf{E}\|_1$ is checked whenever the temporal integration is halted. If $\|\mathbf{E}\|_1 > TOL$, the last m integration steps are rejected and the mesh is refined by adding

$$N[i] := \max \{ round_{\beta}[\|\mathbf{E}\|_{1,i} / \bar{E}] - 1.0 \} \qquad (11a)$$

elements uniformly to (x_{i-1}, x_i), $i = 1, 2, \dots, N$. Here,

$$round_{\beta}(x) := \begin{cases} trunc(x) + 1, & \text{if } x - trunc(x) \geqslant \beta \\ trunc(x), & \text{otherwise}. \end{cases} \qquad (11b)$$

where $0 < \beta < 1$, $trunc(x)$ evaluates the integer part of x, and

$$\bar{E} := 0.9 \, TOL / N. \qquad (11c)$$

The choice of $\beta = 0.2$ in (11) seemed to produce refined meshes that reliably reduced $\|\mathbf{E}\|_1$ to approximately TOL the next time that the error estimate was checked. Further justification for this value of β is given in Adjerid and Flaherty [4].

The integration is redone from $tout[0]$ to $tout[m]$, where m is either $nstep$ or such that $tout[m] = tfinal$, on the refined mesh which has

$$N_r = N + \sum_{i=1}^{N} N[i] \qquad (12)$$

elements. This process is repeated until $\|\mathbf{E}(\cdot, tout[m])\|_1 \leqslant TOL$.

Elements can be deleted from a mesh whenever $\|\mathbf{E}(\cdot, tout[m])\|_1 < TOL / 3$ or whenever refinement was necessary to integrate from $tout[0]$ to $tout[m]$. The need to refine often indicates that the spatial error pattern has changed and that fine grids may no longer be needed in some portions of the domain. A mesh is coarsened by uniting successive pairs of elements, (x_{i-1}, x_i) and (x_i, x_{i+1}), when $\|\mathbf{E}(\cdot, tout[m])\|_{1,j} < TOL / 3N$, $j = i, i+1$. This union of elements is only performed when a significant percentage of elements may be removed from a mesh. This strategy avoids the overhead associated with restarting the temporal integration routine.

If $TOL / 3 \leqslant \|\mathbf{E}(\cdot, tout[m])\|_1 \leqslant TOL$, we continue the temporal integration with the existing mesh provided that its speed is not too great and it is not close to equidistributing the local error indicators. A mesh where the error indicators are not equilibrated indicates that mesh motion and/or refinement are being performed in a suboptimal manner and that a new mesh may be more efficient. We use the following indicator to measure the effectiveness of a mesh $\pi(t, N)$ with respect to equidistributing the local error indicators:

$$\mu(\pi(t, N)) = \frac{2}{N \bar{W}} \left| \sum_{i=1}^{N} \left| \sum_{j=1}^{i} W_j - i \bar{W} \right| \right|. \qquad (13)$$

If π is a mesh that equidistributes W_j, $j = 1, 2, \dots, N$, then $W_j = \bar{W}$, $j = 1, 2, \dots, N$ and $\mu(\pi) = 0$. Larger values of μ indicate increasing departures

from equidistribution. For example, suppose all of the error is concentrated in the first element, i.e., $W_1 = NW$ and $W_j = 0$, $j = 2, 3, ..., N$. Then $\mu(\pi) \approx N$, which we interpret as meaning that at least one element (the last one) will have to cross N elements in order to equidistribute the error indicators. If all of the error were concentrated in the $N/2$ th element, then $\mu(\pi) \approx N/2$, indicating that one element has to cross $N/2$ elements of π. Whenever the mesh speed is too fast and $\mu(\pi(tout[m],N)) > 0.1N$, we generate a new mesh that approximately equidistributes the error indicators by iteratively removing elements with small error indicators and refining those having large error indicators.

Additional details of our procedures, such as the generation of new initial conditions whenever the number of elements in a mesh changes, are as described in Adjerid and Flaherty [2].

In Section II.3, we present some calculations performed on stationary meshes. These were done by using a code based on adapfem with the mesh moving parameter $\lambda = 0$. Additionally, we only generated new meshes when $\mu(\pi(tout[m],N)) > 0.4N$ in order to avoid excessive restarting of the temporal integration routine.

II.3. Computational Examples. We conclude this section by presenting some examples that illustrate our adaptive strategies and also attempt to appraise the relative advantages of mesh moving and local refinement. There are several potential reasons why an adaptive procedure that combines mesh moving with refinement would be very efficient. Mesh moving techniques are inexpensive relative to refinement (cf. Arney and Flaherty [6]) and the use of mesh motion should reduce the need for refinement. Mesh motion can also reduce the necessity of restarting the temporal integrator, which is an important consideration in a method of lines approach such as ours. Some refinement is essential, however, since mesh motion alone cannot generally satisfy prescribed error tolerances. Furthermore, rapid mesh motion, e.g., towards an evolving region of high error, can severely restrict time steps and diminish the efficiency of an adaptive procedure (cf. Adjerid and Flaherty [2]). Finally, many numerical techniques converge at higher rates on uniform meshes than they do on nonuniform moving meshes.

There is, thus, a need to quantify the optimal use of mesh moving with local refinement; however, this is a very difficult problem and there have been very few attempts in this direction. Arney and Flaherty [6] presented some computational results comparing mesh moving and local refinement procedures for two-dimensional hyperbolic systems. Bieterman, Flaherty, and Moore [15] attempted to compare adaptive local refinement and method of lines procedures for one-dimensional parabolic problems and noted the difficulties in finding appropriate performance measures. Herein, we apply a code based on our adaptive procedures to two computational examples and compare results on moving and stationary meshes. We use the total number of space-time cells to integrate the partial differential system from $tinitial$ to $tfinal$ as a measure of performance. A similar measure of computational complexity was used by Arney and Flaherty [6]. It has several apparent deficiencies, such as not providing an indication the effort devoted to the various segments of the adaptive algorithm.

Example 1. Consider the linear heat conduction problem

$$u_t + u_x + g(x,t) = u_{xx}, \quad -1 < x < 1, \quad t > 0. \tag{14a}$$

$$u(x,0) = u^0(x), \quad -1 \leqslant x \leqslant 1. \tag{14b}$$

1084

$$u(-1,t) = c_1(t), \quad u(1,t) = c_2(t), \quad t \geqslant 0. \tag{14c,d}$$

We select g, u^0, c_1, c_2, so that the exact solution of (14) is

$$u(x,t) = 1 - \frac{1}{2}\{tanh\,[10(x-t+0.8)] + tanh\,[20(x+2t-1.6)]\}. \tag{15}$$

Equation (15) represents two wave fronts initially centered at $x = -0.8$ and $x = 1.6$ and moving with speeds 1 and -2, respectively. The center of the fastest front enters the domain (-1,1) at $x = 1$ and $t = 0.3$.

We solved (14) for $0 \leqslant t \leqslant 1.2$ using tolerances of 2^{-k}, $k = 2, 3, 4, 5$, in H^1 with adaptive procedures on moving and stationary meshes. The total number of space-time cells used on $0 \leqslant t \leqslant 1.2$, the exact error $\|e\|_1$ at $t = 1.2$, and the effectivity index

$$\theta := \|E\|_1 / \|e\|_1 \tag{16}$$

at $t = 1.2$ are presented in Table 1. The moving and stationary mesh trajectories that were used to solve (14) with a tolerance of $1/8$ are shown in Figure 2.

Solutions on moving meshes used less than half of the space-time cells of those on stationary meshes. A larger number of cells are needed with a stationary mesh because the temporal integration must be restarted more often and more time steps must be redone due to a failure to satisfy the error tolerance. In each case, the actual error was less than the prescribed tolerance and fine meshes were concentrated in high-error regions. The effectivity index is a common method of appraising the performance of an error estimation technique (cf., e.g., Babuska et al. [9]). Ideally, θ should not deviate appreciably from unity and should approach unity as N increases. The results of Table 1 suggest that this is the case. The performance of our error estimate seems to be slightly better on a stationary mesh than on a uniform mesh.

Tol.	Stationary Mesh			Moving Mesh		
	No. Cells $\times 10^{-4}$	$\|e\|_1$	θ	No. Cells $\times 10^{-4}$	$\|e\|_1$	θ
1/4	4.11	0.1803	0.983	1.60	0.1725	0.979
1/8	8.67	0.0848	0.994	2.79	0.0903	0.993
1/16	26.94	0.0413	0.996	9.50	0.0566	0.988
1/32	47.72	0.0230	0.998	20.27	0.0282	0.996

Table 1. Number of space-time cells on $0 \leqslant t \leqslant 1.2$, spatial discretization error at $t = 1.2$, and effectivity index at $t = 1.2$ as functions of error tolerance using stationary and moving mesh methods to solve Example 1.

Example 2. Consider the reaction-diffusion system
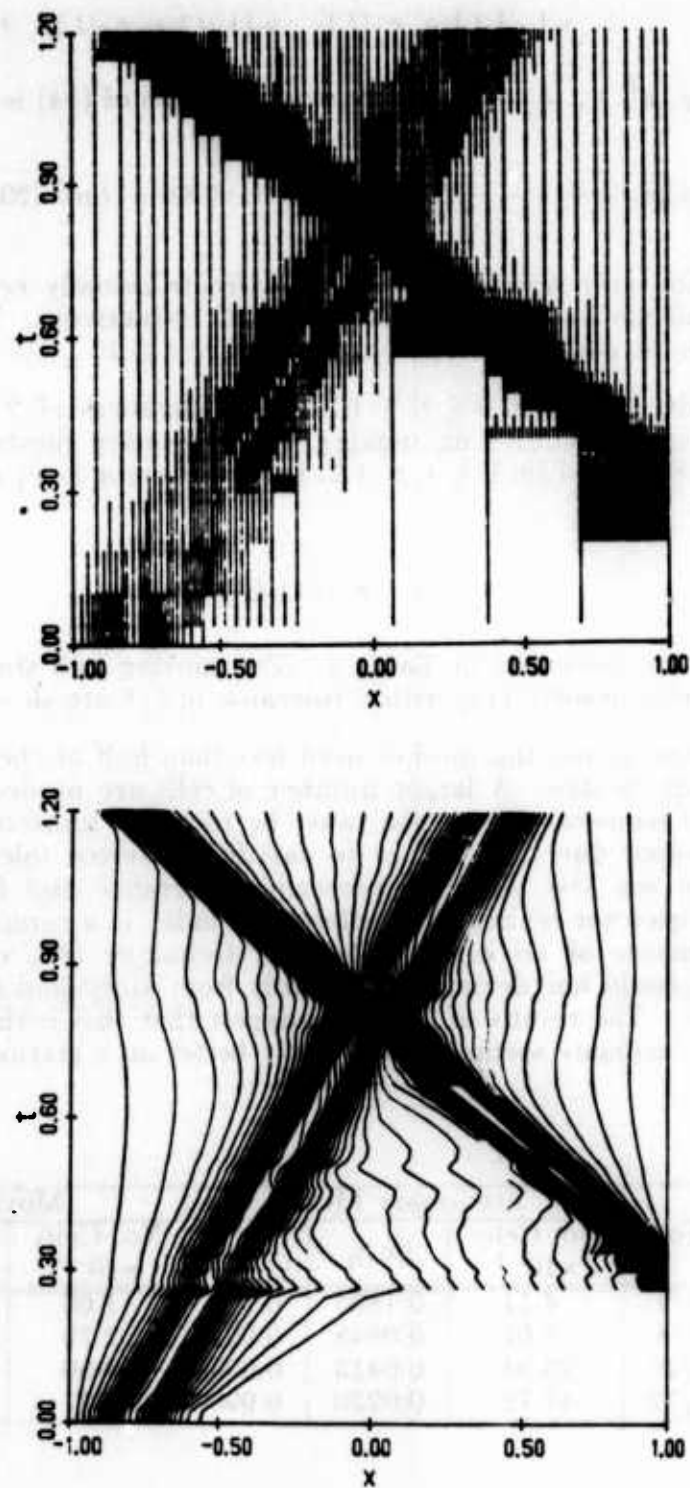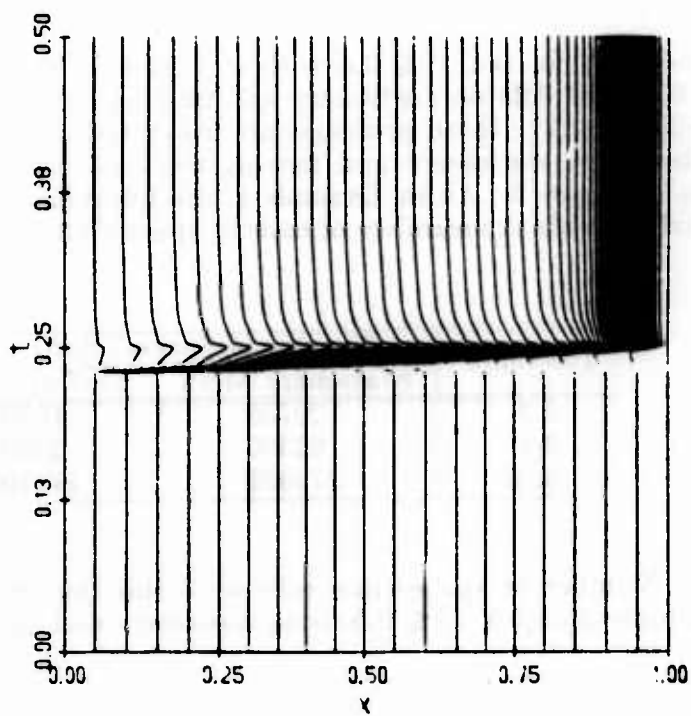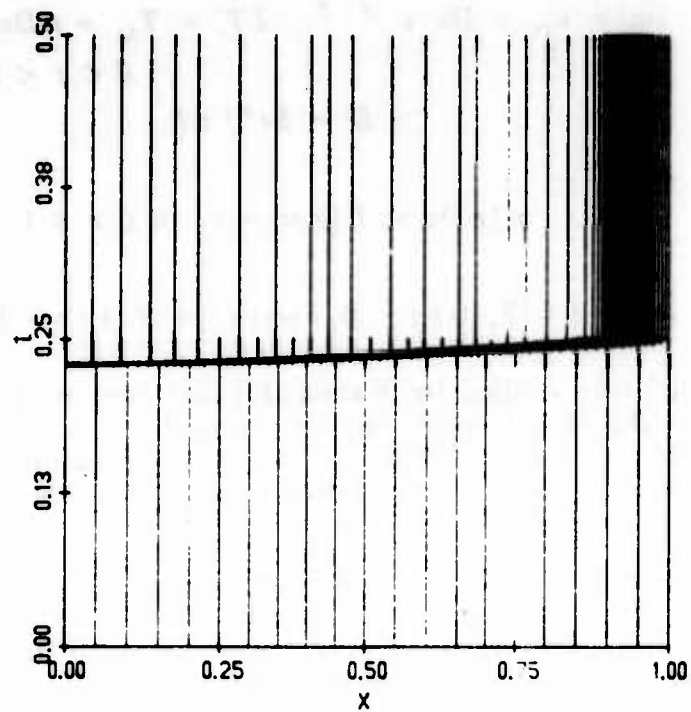
Figure 2. Mesh trajectories used to solve Example 1 with a tolerance of 1/8 on stationary (upper) and moving (lower) meshes.

1086

$$u_t = u_{xx} - Du \, e^{-\delta/T}, \quad LT_t = T_{xx} + \alpha Du e^{-\delta/T},$$

$$0 < x < 1, \quad t > 0, \qquad (17a,b)$$

$$D = Re^{\delta}/\alpha\delta. \qquad (17c)$$

$$u(x,0) = T(x,0) = 1, \quad 0 \leqslant x \leqslant 1, \qquad (17d,e)$$

$$u_x(0,t) = T_x(0,t) = 0, \quad u(1,t) = T(1,t) = 1, \quad t > 0. \qquad (17f,g,h,i)$$

This model was studied by Kapila [17] and used to describe a single one-step reaction $(A \to B)$ of a mixture in the region $0 < x < 1$. The quantity u is the mass fraction of the reactant, T is the reactant temperature, L is the Lewis number, α is the heat release, δ is the activation energy, D is the Damkohler number, and $R > 0.88$ is the reaction rate.

When L is near unity, the temperature slowly increases with a "hot spot" forming at $x = 0$. At some time $t > 0$, ignition occurs and the temperature at $x = 0$ jumps rapidly from near unity to near $1 + \alpha$. A steep flame front then forms and propagates towards $x = 1$ with speed proportional to $e^{\alpha\delta/2(1-\alpha)}$. In practical problems, α is about unity and δ is large; thus, the flame front moves exponentially fast after ignition. The solution tends to a steady state once the flame has reached $x = 1$.

We solved (17) for $0 \leqslant t \leqslant 0.5$ with $\alpha = 1$, $\delta = 20$, and $R = 5$ using tolerances of 0.2, 0.1, and 0.05 on stationary and moving meshes. The number of space-time cells needed to solve these problems are presented in Table 2 and the mesh trajectories for both the stationary and moving mesh calculations with a tolerance of 0.1 are shown in Figure 3. As in Example 1, the number stationary space-time cells is approximately double the number of moving space-time cells.

Tolerance	Number of Space-Time Cells	
	Stationary Mesh	Moving Mesh
0.2	29700	16300
0.1	62300	30300
0.05	170800	118100

Table 2. Number of space-time cells as a function of error tolerance to solve Example 2 on $0 \leqslant t \leqslant 0.5$ using stationary and moving meshes.

III. TWO-DIMENSIONAL ADAPTIVE PROCEDURES.

Our finite element, error estimation, and local refinement procedures for two-dimensional partial differential systems closely parallel our one-dimensional methods and are briefly summarized in Section III.1 and III.2. The representation of data and its management are much more complicated in two dimensions and we use a dynamic tree data structure to store and retrieve information about the mesh, solution, and error estimate. Similar structures have been used by other investigators (cf., e.g., Babuska et

Figure 3. Mesh trajectories used to solve Example 2 with a tolerance of 0.1 on stationary (upper) and moving (lower) meshes.

al. [8-10] and Bank et al. [12-14]) to design adaptive procedures for elliptic systems and they have been shown to be an effective means of reducing storage and access overhead. The essential details of our tree structure are described in Section III.2 and a two-dimensional combustion problem, similar to Example 2, is presented in Section III.^

III.1. Discrete System. A weak form of (1) is constructed in the manner described in Section II.1 for one-dimensional problems. Thus, we seek to determine $u \in H_E^1$ such that

$$(v,u_t) + (v,f(\cdot,t,u,\nabla u)) + A(v,u) = \int_{\partial\Omega} [v^T D^1 u_{z_1} \nu_1 + v^T D^2 u_{z_2} \nu_2] d\sigma.$$

$$\text{for } all \ v \in H_0^1, \ t > 0. \quad (18a)$$

$$(v,u) = (v,u^0). \quad \text{for } all \ v \in H_0^1. \quad t = 0, \quad (18b)$$

where

$$(v,u) = \int_\Omega v(x,y,t)^T u(x,y,t) dx_1 dx_2. \quad (18c)$$

$$A(v,u) = \int_\Omega [v_{z_1}^T D^1(x,t,u)u_{z_1} + v_{z_2}^T D^2(x,t,u)u_{z_2}] dx_1 dx_2. \quad (18d)$$

We have set the mass matrix $\mathbf{M}$ in (1) to the identity matrix for simplicity.

The functions $\mathbf{u}$ and $\mathbf{v}$ are approximated by $\mathbf{U} \in S_E^N \subset H_E^1$ and $\mathbf{V} \in S_0^N \subset H_0^1$, respectively, where S_E^N and S_0^N are spaces of bilinear polynomials with respect to a piecewise rectangular partition of the rectangular domain Ω. The finite element solution $\mathbf{U}$ is obtained by solving

$$(V,U_t) + (V,f(\cdot,t,U,\nabla U)) + A(V,U) = \int_{\partial\Omega} [V^T D^1 U_{z_1} \nu_1 + V^T D^2 U_{z_2} \nu_2] d\sigma.$$

$$\text{for } all \ V \in S_0^N. \ t > 0. \quad (19a)$$

$$(V,U) = (V,u^0). \quad \text{for } all \ V \in S_0^N. \ t = 0. \quad (19b)$$

As in the one-dimensional case, the spatial error $e(x,t) := u(x,t) - U(x,t)$ is approximated by $\mathbf{E} \in \hat{S}_E^N \subset H_E^1$. In two dimensions, we select the finite dimensional space $\hat{S}_E^N$ to consist of piecewise cubic functions with respect to a piecewise rectangular partition of Ω. The cubic functions are biquadratic polynomials that are missing their quartic terms (i.e., serendipity functions in the terminology of Zienkiewicz [20]) and further vanish at the vertices of each element. Thus, once again we take advantage of nodal superconvergence to simplify our approximation of the discretization error. However, there is very little theoretical justification of the superconvergence property for two-dimensional problems and we are relying on computational evidence [3] and our one-dimensional theory [4].

The approximate error $\mathbf{E}$ is determined by replacing $\mathbf{u}$ and $\mathbf{v}$ in (18) by $\mathbf{U} + \mathbf{E}$ and $\mathbf{V} \in \hat{S}_0^N \subset H_0^1$, respectively, where $\hat{S}_0^N$ is composed of the same cubic functions as $\hat{S}_E^N$, and solving

$$(\mathbf{V} \cdot \mathbf{U}_t + \mathbf{E}_t) + (\mathbf{V} \cdot \mathbf{f}(\cdot, t \cdot \mathbf{U} + \mathbf{E} \cdot \nabla(\mathbf{U} + \mathbf{E})) + A(\mathbf{V} \cdot \mathbf{U} + \mathbf{E}) =$$

$$\int_{\partial\Omega} [\mathbf{V}^T \mathbf{D}^1(\mathbf{U}+\mathbf{E})_{,\mathbf{z}} \nu_1 + \mathbf{V}^T \mathbf{D}^2(\mathbf{U}+\mathbf{E})_{,\mathbf{z}} \nu_2] d\sigma, \quad \text{for all } \mathbf{V} \in \hat{S}_0^N, \quad t > 0. \quad (20\text{a})$$

$$(\mathbf{V} \cdot \mathbf{E}) = (\mathbf{V} \cdot \mathbf{u}^0 - \mathbf{U}^0), \quad \text{for all } \mathbf{V} \in \hat{S}_0^N, \quad t > 0. \qquad (20\text{b})$$

The resulting ordinary differential equations (19) and (20) for the solution and error estimate are integrated in time using a code for stiff systems (e.g., DASSL).

III.2. Local Refinement Algorithms. A top-level description of our two-dimensional adaptive procedure closely resembles the one-dimensional algorithm shown in Figure 1. except that we have no mesh moving procedures, as yet. Initially, the domain Ω is partitioned into a "base" mesh of $N \times M$ rectangular elements, which is the coarsest mesh that can be used to solve the problem. Refinement is performed by bisecting the edges of a coarser element, thus, creating four elements where there was previously one. A base mesh having four elements and a refined mesh obtained by bisecting one of them is shown in Figure 4.

The refinement process may be repeated. i.e., elements may be bisected again to create four new elements. Additionally, quartets of elements that were created by refinement may be subsequently deleted if they are no longer needed to maintain accuracy. Bilinear approximations in S_E^N and S_0^N and cubic approximations in $\hat{S}_E^N$ and $\hat{S}_0^N$ are constrained to be linear and quadratic, respectively, on edges between elements of different levels in order to maintain continuity of $\mathbf{U}$ and $\mathbf{E}$ on Ω.

The mesh is organized as a tree structure with the domain Ω being the root of the tree and the $N \times M$ elements of the base mesh being offsprings of the root. All nonleaf nodes of the tree. other than the root node. have four offsprings which correspond to the four elements created by refining its parent element. The domain Ω is referred to as level zero of the tree. the elements of the base mesh are level one. and the levels increase as elements are recursively refined. The tree structure for the mesh shown in the lower portion of Figure 4 is displayed in Figure 5.

Each node of the tree contains the following information:

i.   the element number. say k. of the finite element.

ii.   the level l of the tree.

iii. pointers to the four vertex nodes of element k.

iv.   pointers to the four midside nodes of element k. which are needed to represent $\mathbf{E}$.

v.   pointers to the four elements neighboring element k. with a null pointer used when an edge of element k is on the boundary.

vi.   a pointer to the parent of element k. and

vii. pointers to the four sons of element k. with null pointers used when element k is a leaf node of the tree.

As in the one-dimensional algorithm of Figure 1. elements are added to a mesh when $\|\mathbf{E}\|_1 > TOL$ and deleted from a mesh when either $\|\mathbf{E}\|_1 < TOL / 3$ or when
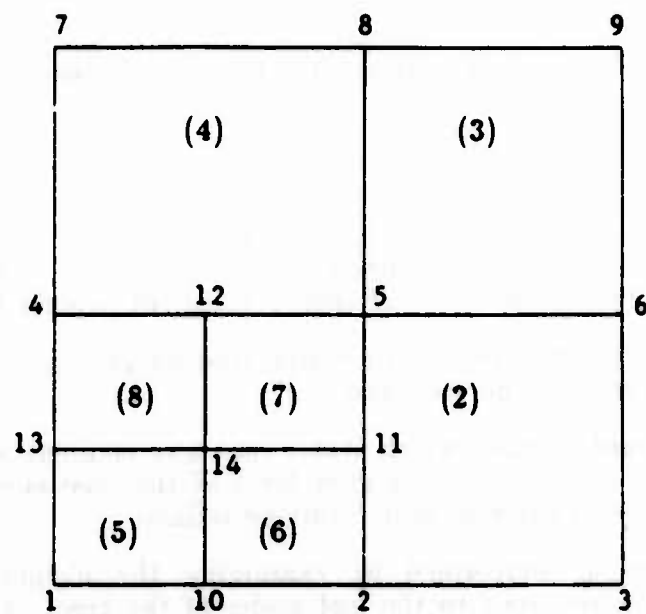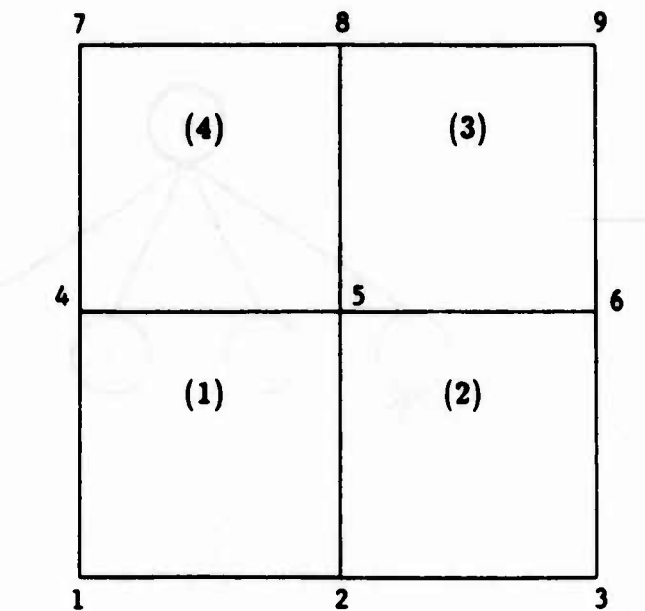
Figure 4. A coarse mesh with four elements numbered 1 to 4 (top) and the resulting mesh after refining element 1 (bottom).
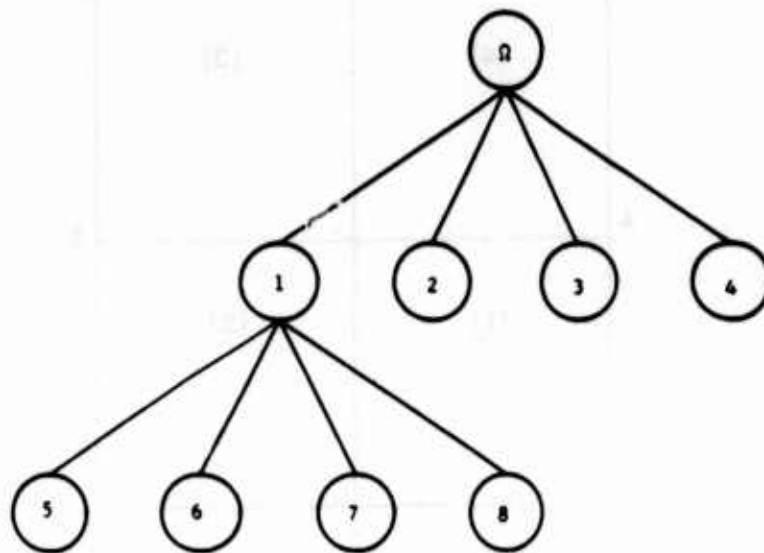
Figure 5. Tree representation of the mesh shown in the lower portion of Figure 4.

refinement was necessary to integrate to the current time. Our refinement and deletion procedures impose the following two rules, which Bank et al. [12-14] found to aid the efficiency and accuracy of their refinement process for elliptic systems:

i. the 1−*irregular* rule. which states that neighboring elements can differ by at most one level of the tree, and

ii. the 3−*neighbor* rule. which states that any element where the number of edges containing elements at a higher level of the tree and the number of boundary edges totals to three or more must be refined.

Refinement is performed by examining the elements of a mesh by levels, proceeding from the root to the leaf nodes of the tree. An element k is refined by dividing it into four subelements whenever $\|\mathbf{E}\|_{1,k} > TOL / \sqrt{N_e}$, where N_e is the number of elements in the mesh. Elements are deleted from meshes. other than the base $N \times M$ mesh. by pruning the tree. A quartet of elements having the same parent is deleted if:

i. every element in the quartet has no offsprings.

ii. every neighbor of the elements in the quartet are at the same or lower level of the tree. and

iii. the average error estimate of the four elements is less than $TOL / 3\sqrt{N_e}$.

Additional details pertaining to other aspects of our adaptive procedures are presented in Adjerid and Flaherty [3].

III.3. Computational Example. A code based on our two-dimensional local mesh refinement procedure has been written and applied to several problems [3]. Herein, we present the results of a two-dimensional version of the model combustion problem considered in Example 2.

Example 3. Consider the partial differential system on the rectangular domain $\Omega := \{ (x_1, x_2) \mid 0 < x_1, x_2 < 1 \}$

$$T_t = T_{x_1 x_1} + T_{x_2 x_2} + D(1 + \alpha - T)e^{-\delta/T}, \quad (x, y) \in \Omega, \quad t > 0, \quad (21a)$$

$$T(\mathbf{x}, 0) = 1, \quad \mathbf{x} \in \Omega \bigcup \partial \Omega, \quad (21b)$$

$$T_{x_1}(0, x_2, t) = 0. \quad T(1, x_2, t) = 1, \quad 0 \leqslant x_2 \leqslant 1, \quad t > 0, \quad (21c,d)$$

$$T_{x_2}(x_1, 0, t) = 0. \quad T(x_1, 1, t) = 1, \quad 0 \leqslant x_1 \leqslant 1, \quad t > 0. \quad (21e,f)$$

All of the parameters are as described in Example 2. The Lewis number L has been set to unity and, in this case, the mass fraction $u = 1 + (1 - T)/\alpha$.

We solved (21) with $\alpha = 1$, $\delta = 20$, and $R = 5$ using a spatial error tolerance of 0.2. Mesh refinement had to be restricted to a maximum of two levels because of virtual memory restrictions on our computing system. The meshes that were created at $t = 0.2867$, 0.2979, 0.3055, and 1 are shown in Figure 6. Surface and contour plots of the calculated temperatures at $t = 0.28674$ and 0.3115 are presented in Figures 7 and 8, respectively.

The temperature slowly increases until ignition occurs at approximately $t = 0.28$. The temperature at the origin then jumps from near unity to near two. A circularly shaped reaction front forms and moves radially with a speed of approximately 30 towards the boundaries. A steady state is reached at about $t = 0.32$. Refinement is confined to the vicinity of the reaction front. The results of Figures 7 and 8 show some small oscillations in the temperature ahead of the reaction front. At present, we are unsure if these oscillations are caused by interpolation inaccuracies in our plotting routines, inadequate resolution of the finite element solution due to our restricting the number of levels of refinement, or an instability of the reaction front. We plan to explore these matters further using a combination of numerical and asymptotic techniques.

Our results on this difficult nonlinear problem are very encouraging; however, we anticipate that greater efficiency could be achieved by combining local mesh refinement with mesh moving as in the one-dimensional procedures described in Section II.

IV. DISCUSSION OF RESULTS AND CONCLUSIONS. We have developed adaptive local mesh refinement finite element procedures for solving vector systems of parabolic partial differential equations in one and two dimensions. The nodal superconvergence property of the finite element method on parabolic systems has been used to calculate an estimate the spatial discretization error and mesh
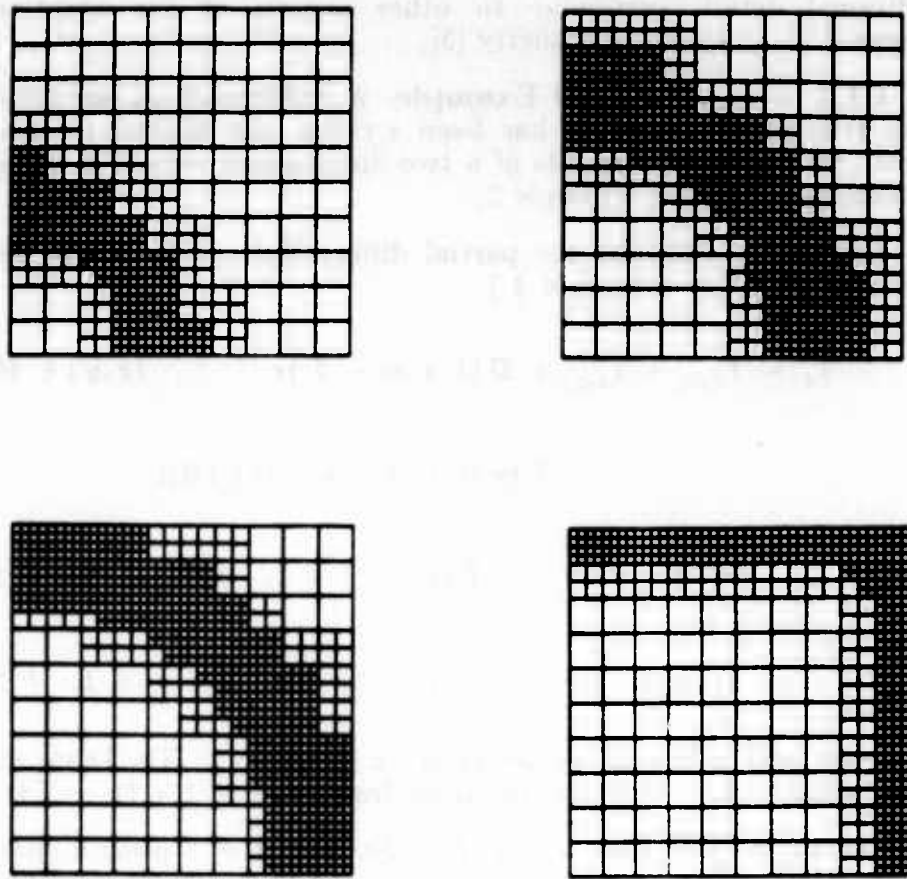
Figure 6. Meshes that were used for Example 3 at $t = 0.2867$ (upper left), 0.2979 (upper right), 0.3055 (lower left), and 1 (lower right).

motion has been combined with local mesh refinement for one-dimensional problems.

Examples 1 and 2 were designed to illustrate the performance of our one-dimensional procedures and to characterize the importance of mesh motion as an adaptive technique relative to mesh refinement. These experiments indicate that our combination of mesh moving and refinement can obtain solutions with about one-half the total number of space-time cells of calculations performed using only refinement. We emphasize the preliminary nature of these results. Many more experimental and theoretical investigations will be necessary before firm conclusions can be reached regarding the optimal combination of mesh motion and refinement. Appropriate performance measures and optimality conditions are yet to be specified. There is also a strong temptation to compare coded implementations of procedures and, at this stage, we are interested in more theoretical bounds on an algorithm's performance.

Comparisons of the exact and estimated errors, presented in Example 1 and in [1-4], give us some confidence in the accuracy of our error estimate. Additionally, the results of Example 3 provide an indication of the robustness of our methods.
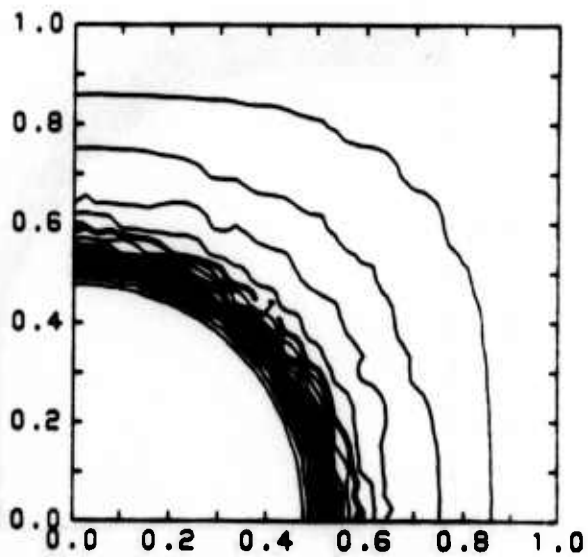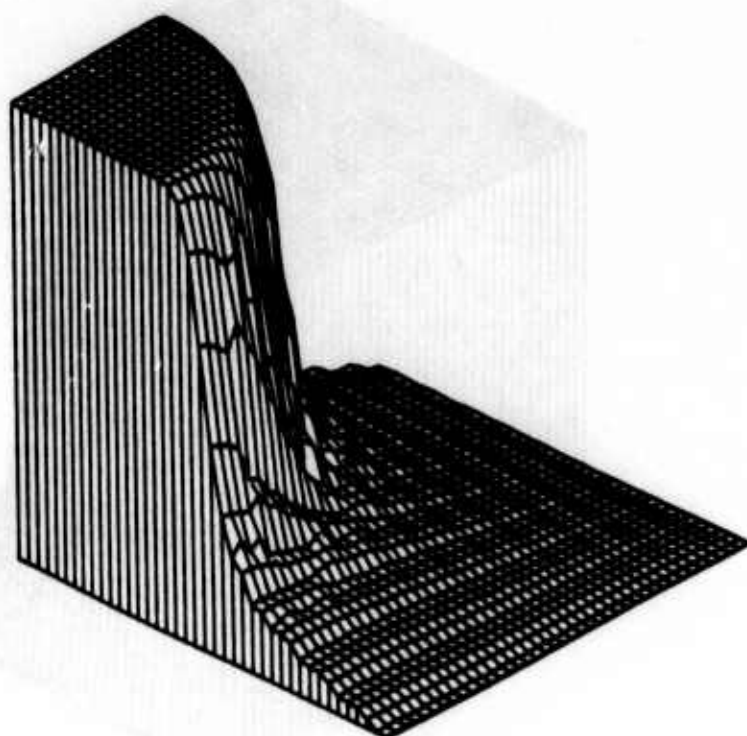
Figure 7. Surface (top) and contour (bottom) plots of calculated temperature for Example 3 at $t = 0.28674$.
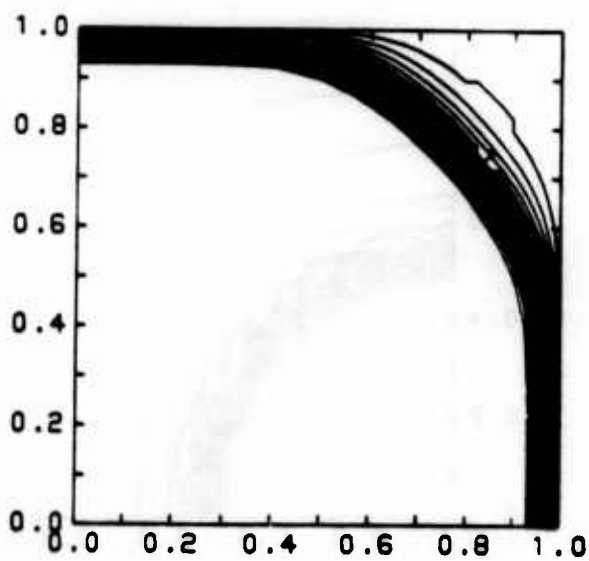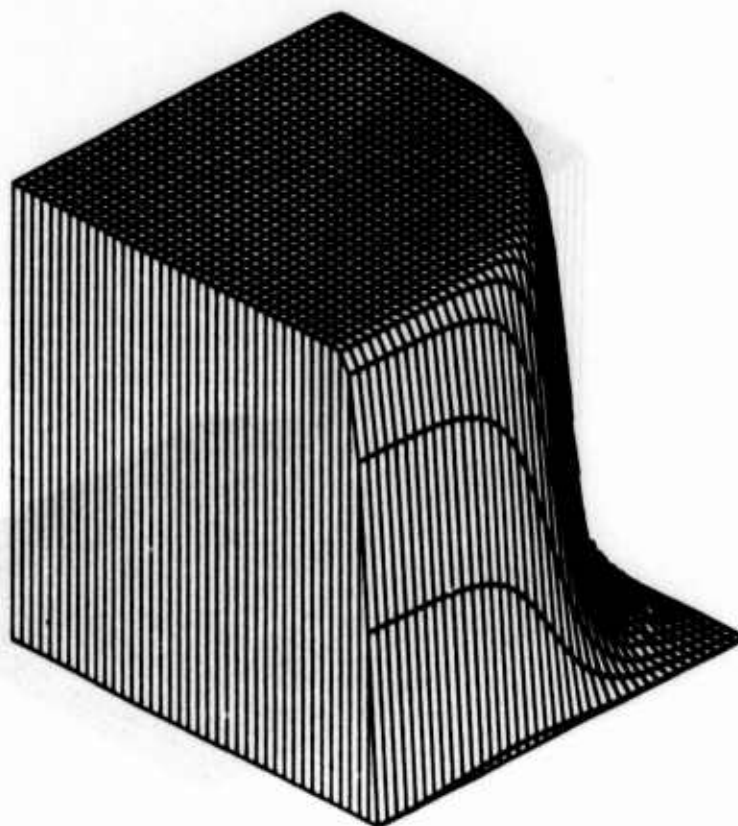
Figure 8. Surface (top) and contour (bottom) plots of calculated temperature for Example 3 at $t = 0.3115$.

This is a difficult nonlinear two-dimensional problem; yet, we solved it without intervention and a priori knowledge of the solution. No special initial mesh was used, fine meshes were automatically added to the vicinity of the reaction front, and the fine meshes followed the dynamics of the problem. Indeed, our one- and two-dimensional techniques, seem to be well-suited for the automatic solution of reaction-diffusion systems.

Despite our preliminary success, there is a great deal more that should be done to justify and improve the performance of our procedures. As noted, rigorous analyses of the convergence of our error estimate to the true error have only been done for one-dimensional linear parabolic problems on stationary meshes [4]. Dimensional, nonlinear, and refinement effects should be included in a complete analysis. This is a difficult task, as very few analyses of two-dimensional time dependent problems with refinement have appeared in the literature.

Several computational procedures in our approach might also be improved. For example, a sparse Gaussian elimination procedure was used to solve the linear algebraic systems associated with the temporal integration of (4), (8), and (10) in one dimension and (19) and (20) in two dimensions. The solution of linear systems is a significant part of the total computational effort, and it is possible that iterative schemes, such as multigrid methods, could substantially improve performance and reduce storage. Multigrid iteration was used successfully in the adaptive PLTMG package for elliptic systems by Bank et al. [13].

We are also studying the addition of mesh moving capabilities to our two-dimensional algorithm, the use of higher-order finite element approximations, and implementations of our procedures on vector and parallel computers. A simple, stable and explicit mesh moving technique, that may be useful for our purposes, was developed by Arney and Flaherty [5] for two-dimensional hyperbolic systems. This procedure dramatically reduced errors and enhanced the resolution of their solutions (cf. Arney and Flaherty [6]). We are developing procedures for two-dimensional parabolic problems that use piecewise biquadratic finite element approximations as solution spaces and piecewise cubic approximations as error estimates. Babuska [7] has shown that the error associated with even-degree polynomial finite element approximations for elliptic problems is principally due to the error in the interior of the element. Thus, the error on element boundaries may be neglected. Babuska and Yu [11] have implemented procedures for elliptic systems based on this theory and we are studying their utility for parabolic problems. Finally, our tree structure is well-suited for parallel computation and we are exploring its use on a variety of parallel computing systems.

REFERENCES

1. S. Adjerid and J.E. Flaherty. A moving finite element method with error estimation and refinement for one-dimensional time dependent partial differential equations. *SIAM J. Numer. Anal.*, 23 (1986), pp. 778-796.

2. S. Adjerid and J.E. Flaherty. A moving mesh finite element method with local refinement for parabolic partial differential equations. *Comp. Meths. Appl. Mech. Engr.*, 56 (1986), pp. 3-26.

3. S. Adjerid and J.E. Flaherty. A local refinement finite element method for two-dimensional parabolic systems. Tech. Rep. No. 86-7. Department of

Computer Science, Rensselaer Polytechnic Institute. Troy, 1986.

4. S. Adjerid and J.E. Flaherty, Local refinement finite element methods on stationary and moving meshes for one-dimensional parabolic systems, in preparation.

5. D.C. Arney and J.E. Flaherty, A two-dimensional mesh moving technique for time dependent partial differential equations. Tech. Rep. No. 85-9, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, 1985. Also *J. Comput. Phys.*, to appear.

6. D.C. Arney and J.E. Flaherty, An adaptive method with mesh moving and refinement for time-dependent partial differential equations, *Trans. Fourth Army Conf. Appl. Maths. and Comput.*, U. S. Army Research Office, Research Triangle Park, NC, to appear.

7. I. Babuska, personal communication, 1986.

8. I. Babuska and A. Miller. A posteriori error estimates and adaptive techniques for the finite element method. Tech. Note BN-968, Institute for Physical Science and Technology. University of Maryland, College Park, MD, 1981.

9. I. Babuska. A. Miller, and M. Vogelius. Adaptive methods and error estimation for elliptic problems of structural mechanics, in *Adaptive Computational Methods for Partial Differential Equations*. I. Babuska. J. Chandra, J.E. Flaherty. eds., SIAM, Philadelphia, 1983, pp. 57-73.

10. I. Babuska and W. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15 (1978), pp. 736-734.

11. I. Babuska and D. Yu. Asymptotically exact a-posteriori error estimator for biquadratic elements. Tech. Note BN-1050, Institute for Physical Science and Technology. University of Maryland, College Park, MD, 1986.

12. R.E. Bank. PLTMG users' guide, June 1981 version. Technical Report, Department of Mathematics, University of California at San Diego, La Jolla, CA, 1982.

13. R.E. Bank and A. Sherman, An adaptive multi-level method for elliptic boundary value problems. *Computing*. 26 (1981), pp. 91-105.

14. R.E. Bank. A.H. Sherman, and A. Weiser. Refinement algorithms and data structures for regular local mesh refinement, *Mathematics and Computers in Simulation*. to appear.

15. 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.

16. J.M. Coyle. J.E. Flaherty. and R. Ludwig. On the stability of mesh equidistribution strategies for time-dependent partial differential equations, *J. Comput. Phys.*, 62 (1986), pp. 26-39.

17. A.K. Kapila, *Asymptotic Treatment of Chemically Reacting Systems*, Pitman Advanced Publishing Program, Boston, 1983.

18. L.R. Petzold, A description of DASSL: a differential/algebraic system solver, Sandia Report No. Sand. 82-8637, Sandia National Laboratory, Livermore, CA, 1982.

19. V. Thomee, Negative norm estimates and superconvergence in Galerkin methods for parabolic problems, *Math. Comp.*, 34 (1980), pp. 93-113.

20. O.C. Zienkiewicz, *The Finite Element Method: Third Edition*, McGraw Hill. London, 1977.

A POSTERIORI ERROR ESTIMATION IN A FINITE ELEMENT METHOD FOR PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS

J. M. Coyle and J. E. Flaherty*
U.S. Army Armament, Munitions, and Chemical Command
Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

ABSTRACT. Superconvergence properties and quadratic polynomials are used to derive a computationally inexpensive approximation to the spatial component of the error in a piecewise linear finite element method for one-dimensional parabolic partial differential equations. This technique is coupled with time integration schemes of successively higher orders to obtain an approximation of the temporal and total discretization errors. Computational results indicate that these approximations converge to the exact discretization errors as the mesh is refined. The approximate errors are used to control an adaptive mesh refinement strategy.

I. INTRODUCTION. Adjerid and Flaherty [1,2] developed an a posteriori estimate of the spatial discretization error in a finite element method of lines for solving vector systems of parabolic partial differential equations. They discretized the system in space using Galerkin's method with piecewise linear finite element approximations. The error estimate was calculated using Galerkin's method with piecewise quadratic functions. A nodal superconvergence property of the finite element method was used to neglect errors at nodes and, thus, improve computational efficiency. Ordinary differential equations (ODEs) for the finite element solution and error estimate were then integrated in time using the backward difference code DASSL [3].

Adjerid and Flaherty [1,2] assumed that the temporal discretization error associated with DASSL was negligible compared to the spatial error. Thus, their estimate of the spatial discretization error could be regarded as an estimate of the total error. They used their error estimate to control mesh moving and local mesh refinement procedures that attempted to equidistribute the error estimate and satisfy a prescribed global error tolerance. Similar mesh refinement strategies have been used by Bieterman and Babuska [4,5].

Our goal is to develop techniques that simultaneously estimate the temporal and spatial discretization errors. To this end, we consider M-dimensional partial differential systems of the form

$$u_t(x,t) + f(x,t,u,u_x) = (D(x,t)u_x(x,t))_x \quad , \quad a < x < b, \ t > 0 \ , \quad (1a)$$

subject to the initial conditions

$$u(x,0) = u^0(x) \quad , \quad a \leq x \leq b \ , \quad (1b)$$

and boundary conditions

$$A_L(t)u(a,t) + B_L(t)u_x(a,t) = g_L(t) \ ,$$

$$A_R(t)u(b,t) + B_R(t)u_x(b,t) = g_R(t) \ , \ t > 0 \ . \quad (1c)$$

The variables x and t represent spatial and temporal coordinates and denote partial differentiation when they are used as subscripts; u, f, u^0, g_L, and g_R are M-vectors; and D, A_L, B_L, A_R, and B_R are $M \times M$ matrices.

We, like Adjerid and Flaherty [1,2], discretize Eq. (1) in space using Galerkin's method with piecewise linear finite elements. Temporal discretization, however, is performed by the backward Euler method as opposed to using an ODE code. A second solution is calculated using trapezoidal rule integration in time and the difference between the two solutions is used to furnish an estimate of the temporal discretization error. A third solution is obtained using Adjerid and Flaherty's [1,2] quadratic finite elements and the trapezoidal rule in time. This solution is higher order in space and time than the original piecewise linear finite element-backward Euler solution. Hence, it can be used to provide an estimate of the total discretization error of the piecewise linear finite element-backward Euler solution. Furthermore, the difference between the piecewise linear and quadratic solutions calculated by the trapezoidal rule can be used to furnish an estimate of the spatial discretization error.

At first sight, the above procedure seems to be very expensive; however, nodal superconvergence significantly reduces computational complexity. Defect correction methods can also be used to reduce costs associated with the temporal integration.

The estimates of the temporal, spatial, and total discretization errors of the piecewise linear finite element-backward Euler solution are used to control a global refinement procedure that attempts to keep an estimate of the total discretization error per time step in H^1 below a prescribed limit. Depending on the proportions of the temporal and spatial error estimates to the total error estimate, we refine the time step, finite element mesh, or both.

The piecewise linear and quadratic finite element procedures and the temporal integration schemes are described in Section II. Our error estimation procedures are presented in Section III. Adjerid and Flaherty [6] proved that their spatial error estimate converges to the exact error as the mesh is refined when temporal integration is exact for linear parabolic problems. Similar results have not yet been established when temporal errors are present; however, computational results of Section III indicate that convergence of our temporal, spatial, and total error estimates are likely. Our global refinement strategy is presented in Section IV and it is applied to an unstable heat conduction problem. Finally, in Section V we discuss our results and suggest some future investigations.

II. DISCRETE SYSTEM. We simplify the presentation slightly by assuming that only Dirichlet data is prescribed; thus, $B_L(t) = B_R(t) = 0$, $t > 0$, in Eq. (1c). A weak form of Eq. (1) is then constructed by multiplying Eq. (1a) by a test function $v(x,t) \epsilon H_0^1$, integrating the result with respect to x from a to b, and integrating the diffusive term by parts to obtain

$$(v,u_t) + (v,f) + A(v,u) = 0 \quad , \quad t > 0 \text{ , for all } v \epsilon H_0^1 \ . \tag{2a}$$

The inner product (v,u) and strain energy $A(v,u)$ are defined as

$$(v,u) = \int_a^b v^T u\, dx \quad , \quad A(v,u) = \int_a^b v_x^T D u_x\, dx \ . \qquad (2b,c)$$

Functions v belonging to H_0^1 are required to have finite values of (v,v) and (v_x,v_x) and vanish at $x = a$ and b. Any weak solution $u \in H_E^1$ of Eq. (2a) must also satisfy the Dirichlet (essential) boundary conditions

$$u(a,t) = A_L^{-1}(t)g_L(t) \quad , \quad u(b,t) = A_R^{-1}(t)g_R(t), \ t > 0 \ , \qquad (2d,e)$$

and initial conditions obtained by multiplying Eq. (1b) by v and integrating with respect to x, i.e.,

$$(v,u) = (v,u^0) \quad , \quad t = 0 \quad , \quad \text{for all } v \in H_0^1 \ . \qquad (2f)$$

A discrete version of the weak system, Eq. (2), is constructed by using finite element-Galerkin procedures in space (cf. Section II.1) and finite difference techniques in time (cf. Section II.2).

II.1. Spatial Discretization. In order to discretize Eq. (2a) in space, we introduce a partition

$$\Pi_N := \{a = x_0 < x_1 < \ldots < x_N = b\} \qquad (3)$$

of (a,b) into N subintervals (x_{i-1},x_i), $i = 1,2,\ldots,N$, and approximate u and v by piecewise polynomial functions U and V, respectively, with respect to this partition. Thus, the spatially-discrete form of Eq. (2) consists of finding $U \in S_E^N \subset H_E^1$ such that

$$(V,U_t) + (V,f) + A(V,U) = 0 \quad , \quad t > 0 \quad , \quad \text{for all } V \in S_0^N \subset H_0^1 \ , \qquad (4a)$$

$$(V,U) = (V,u^0) \quad , \quad t = 0 \quad , \quad \text{for all } V \in S_0^N \subset H_0^1 \ . \qquad (4b)$$

The spaces S_E^N and S_0^N will be chosen to consist of either piecewise linear or piecewise quadratic polynomial functions. The spaces of piecewise linear polynomials are denoted as $S_E^{N,1}$ and $S_0^{N,1}$ and are easily constructed in terms of the familiar "hat" functions

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} \ , & x_{i-1} \leqslant x < x_i \\[2mm] \dfrac{x_{i+1} - x}{x_{i+1} - x_i} \ , & x_i \leqslant x < x_{i+1} \ , \quad i = 0,1,\ldots,N \ . \\[2mm] 0 \ , & \text{otherwise} \end{cases} \qquad (5)$$

The piecewise linear finite element solution $U^1 \in S_E^{N,1}$ is written in the form

$$U^1(x,t) = \sum_{i=0}^N c_i(t)\phi_i(x) \qquad (6)$$

and determined by solving the ordinary differential system

$$(V^1, U^1_t) + (V^1, f) + A(V^1, U^1) = 0 \quad , \quad t > 0 \quad , \quad \text{for all } V^1 \epsilon S_0^{N,1} \quad , \quad (7a)$$

$$(V^1, U^1) = (V^1, u^0) \quad , \quad t = 0 \quad , \quad \text{for all } V^1 \epsilon S_0^{N,1} \quad , \quad (7b)$$

where the piecewise linear test functions $V^1 \epsilon S_0^{N,1}$ have a form similar to Eq. (6).

Piecewise quadratic approximations $U^2 \epsilon S_E^{N,2}$ are constructed by adding a "hierarchical" correction $E^2(x,t)$ to U^1, i.e.,

$$U^2(x,t) = U^1(x,t) + E^2(x,t) \qquad (8a)$$

where

$$E^2(x,t) = \sum_{i=1}^{N} d_{i-\frac{1}{2}}(t) \psi_{i-\frac{1}{2}}(x) \quad . \qquad (8b)$$

The basis $\psi_{i-\frac{1}{2}}(x)$, $i = 1, 2, \ldots, N$, for the quadratic correction has the form

$$\psi_{i-\frac{1}{2}}(x) = \begin{cases} -2\left(\dfrac{x - x_{i-1}}{x_i - x_{i-1}}\right)\left(\dfrac{x_i - x}{x_i - x_{i-1}}\right) \quad , \quad x_{i-1} \leqslant x \leqslant x_i \\[4mm] 0 \qquad\qquad\qquad\qquad , \quad \text{otherwise} \end{cases} \quad , \quad i = 1, 2, \ldots, N \quad . \quad (9)$$

Piecewise quadratic solutions are determined by solving

$$(V^2, U^2_t) + (V^2, f) + A(V^2, U^2) = 0 \quad , \quad t > 0 \quad , \quad \text{for all } V^2 \epsilon S_0^{N,2} \quad , \quad (10a)$$

$$(V^2, U^2) = (V^2, u^0) \quad , \quad t = 0 \quad , \quad \text{for all } V^2 \epsilon S_0^{N,2} \quad , \qquad (10b)$$

where, once again, V^2 has a form similar to Eq. (8).

II.2. Temporal Discretization. The finite element systems, Eqs. (4), (7), or (10), are discretized in time for the time step $[t_{n-1}, t_n]$ using a weighted two-step method, which for Eq. (4) has the form

$$\left(V, \frac{U^n - U^{n-1}}{\Delta t_n}\right) + \theta[(V, f^n) + A(V, U^n)] +$$

$$(1-\theta)[(V, f^{n-1}) + A(V, U^{n-1})] = 0 \quad , \quad \text{for all } V \epsilon S_0^N \quad . \qquad (11)$$

The scalar parameter θ is selected on $[0,1]$, $U^n(x) := U(x, t_n)$, etc., and $\Delta t_n := t_n - t_{n-1}$. For simplicity, the test function V has been assumed to be independent of time, although this will not be strictly correct when refinement is incorporated into the finite element method (cf. Section IV).

The two particular choices of θ that are appropriate for our investigation are $\theta = 1$, which yields the backward Euler method, and $\theta = \frac{1}{2}$, which yields the trapezoidal rule. It is well known [7] that the local discretization error of the backward Euler method is $O(\Delta t_n^2)$ and that of the trapezoidal rule is $O(\Delta t_n^3)$. We will use this difference in the orders of accuracy of the two methods to estimate the local temporal discretization error of the finite element solution.

III. ERROR ESTIMATION. Local and global estimates of the discretization error have been successfully used to control refinement algorithms that attempt to solve partial differential systems to prescribed levels of accuracy [1,2,4-6,8-12]. Our goal is to estimate the discretization error per time step in solutions of Eq. (2) obtained by using piecewise linear finite element approximations in space and the backward Euler method in time. It seems most appropriate to gage errors

$$e := u - U \tag{12}$$

in the H^1 norm

$$\|e\|_1 := [\int_a^b (e_x^T e_x + e^T e)dx]^{\frac{1}{2}} \tag{13}$$

however, other measures may also be used. An error estimate that is global in space and local in time may at first seem unusual, but it is commonly used when spatial finite element approximations are combined with temporal finite difference methods (cf., e.g., Thomee [13]).

Let the piecewise linear finite element solution obtained by using backward Euler temporal integration be denoted as $U_{BE}^{1,n}(x)$ at time t_n. Likewise, let $U_T^{1,n}(x)$ and $U_T^{2,n}(x)$ denote solutions obtained at t_n by trapezoidal rule integration with piecewise linear and quadratic approximations, respectively.

It is known [14] that $\|u(\cdot,t_n) - U_{BE}^{1,n}(\cdot)\|_1 = O(\Delta t_n^2) + O(\Delta t_n/N)$. Since $\|u(\cdot,t_n) - U_T^{2,n}\|_1 = O(\Delta t_n^3) + O(\Delta t_n/N^2)$, we should be able to use the difference between $U_T^{2,n}$ and $U_{BE}^{1,n}$ to estimate the error in $U_{BE}^{1,n}$; thus,

$$\|u - U_{BE}^{1,n}\|_1 \leq \|U_T^{2,n} - U_{BE}^{1,n}\|_1 + \|u - U_T^{2,n}\|_1$$

$$\leq \|U_T^{2,n} - U_{BE}^{1,n}\|_1 + O(\Delta t_n^3) + O(\Delta t_n/N^2) . \tag{14}$$

The main problem in using $\|U_T^{2,n} - U_{BE}^{1,n}\|_1$ as an a posteriori estimate of $\|u - U_{BE}^{1,n}\|_1$ is the computational effort required to obtain $U_T^{2,n}$. This cost can be reduced considerably by using the superconvergence property of the finite element method for one-dimensional parabolic systems. In the present context, superconvergence implies that finite element solutions converge at a faster rate on Π_N than elsewhere on (a,b). Hence, the error at the nodes may be neglected relative to the error in the interior of the elements when N is sufficiently large.

Nodal superconvergence has been used by several investigators as a means of constructing a posteriori error estimates in finite element approximations.

In particular, Adjerid and Flaherty [1,2] used it in conjunction with their adaptive finite element method of lines. Their situation was somewhat more restrictive than ours as they also required the temporal error to be negligible relative to the spatial error.

The use of the nodal superconvergence property enables us to approximate $U_T^{2,n}$ as

$$U_T^{2,n} \approx U_T^{1,n} + E_T^{2,n} \qquad (15)$$

where $U_T^{1,n}$ is obtained by solving Eq. (7) using trapezoidal rule integration and $E_T^{2,n}$ is obtained by solving Eq. (10a) by trapezoidal rule integration with U^2 replaced by Eq. (15). Furthermore, it is only necessary to test Eq. (10a) against functions $V^2 \epsilon \hat{S}_0^{N,2}$, where $\hat{S}_0^{N,2}$ is a space of quadratic polynomials that vanish on Π_N.

To summarize, our procedure for obtaining the finite element solution $U_{BE}^{1,n}$ and its error estimate $U_T^{1,n} + E_T^{2,n} - U_{BE}^{1,n}$ for the time step $[t_{n-1},t_n]$ consists of:

(i) discretizing Eq. (7a) by the backward Euler method and determining $U_{BE}^{1,n}$ as the solution of

$$(V^1, \frac{U_{BE}^{1,n} - U_{BE}^{1,n-1}}{\Delta t_n}) + (V^1, f(\cdot,t_n,U_{BE}^{1,n})$$

$$+ A(V^1,U_{BE}^{1,n}) = 0 , \text{ for all } V^1 \epsilon S_0^{N,1} , \qquad (16a)$$

(ii) discretizing Eq. (7a) by the trapezoidal rule and determining $U_T^{1,n}$ as the solution of

$$(V^1, \frac{U_T^{1,n} - U_{BE}^{1,n-1}}{\Delta t_n}) + \frac{1}{2}[(V^1,f(\cdot,t_n,U_T^{1,n}) + A(V^1,U_T^{1,n})$$

$$+ (V^1,f(\cdot,t_{n-1},U_{BE}^{1,n-1})) + A(V^1,U_{BE}^{1,n-1})] = 0 , \text{ for all } V^1 \epsilon S_0^{N,1}, \quad (16b)$$

(iii) discretizing Eq. (10a) by the trapezoidal rule and determining $E_T^{2,n}$ as the solution of

$$(V^2, \frac{U_T^{1,n} + E_T^{2,n} - U_{BE}^{1,n-1} - E_T^{2,n-1}}{\Delta t_n})$$

$$+ \frac{1}{2}[(V^2,f(\cdot,t_n,U_T^{1,n} + E_T^{2,n})) + A(V^2,U_T^{1,n}+E_T^{2,n})$$

$$+ (V^2,f(\cdot,t_{n-1},U_{BE}^{1,n-1}+E_T^{2,n-1})) + A(V^2,U_{BE}^{1,n-1}+E_T^{2,n-1})] = 0 ,$$

$$\text{for all } V^2 \epsilon \hat{S}_0^{N,2} . \qquad (16c)$$

Temporal error estimation is local; thus, we use $U_{BE}^{1,n-1}$ as an initial condition for the trapezoidal rule integrations in Eqs. (16b) and (16c). Nodal superconvergence and the hierarchical formulation has uncoupled the piecewise linear and quadratic components of $U_T^{2,n}$. The spatial error estimate $E_T^{2,n}$ on the subinterval (x_{i-1}, x_i) is furthermore uncoupled from the error on other subintervals and this significantly reduces the computational complexity associated with solving Eq. (16c). The solution of Eq. (16b), noted in step (ii), is necessary in order to increase the temporal accuracy of the solution because superconvergence only increases the order of accuracy in space. Some computational savings can generally be obtained, especially for nonlinear problems, by calculating $U_T^{1,n}$ as a defect correction to the backward Euler solution $U_{BE}^{1,n}$.

As described above,

$$\bar{e}^n := \|U_T^{1,n} + E_T^{2,n} - U_{BE}^{1,n}\|_1 \tag{17}$$

furnishes an estimate to the error $\|u - U_{BE}^{1,n}\|_1$ of the backward Euler solution. Equation (17) suggests the inequality

$$\bar{e}^n \leq \|U_T^{1,n} - U_{BE}^{1,n}\|_1 + \|E_T^{2,n}\|_1 . \tag{18}$$

The term $\|U_T^{1,n} - U_{BE}^{1,n}\|_1$ is the difference between two piecewise linear solutions computed with temporal integration schemes of different orders and can be regarded as a measure of the temporal discretization error. In a similar manner, $\|E_T^{2,n}\|_1$ can be regarded as a measure of the spatial discretization. Indeed, when the finite element system, Eq. (7), is integrated exactly in time, Adjerid and Flaherty [6] proved that $\|E^2\|_1$ converges to the exact spatial discretization error $\|u - U^1\|_1$ as $N \to \infty$ for linear parabolic problems.

We conclude this section by presenting an example that indicates that $\bar{e}^n$, $\|U_T^{1,n} - U_{BE}^{1,n}\|_1$, and $\|E_T^{2,n}\|_1$ provide good estimates of the total, temporal, and spatial discretization errors, respectively.

Example 1: Consider the linear heat conduction problem

$$u_t = u_{xx}/\pi^2 \quad , \quad 0 < x < 1 \quad , \quad t > 0 \ , \tag{19a}$$

$$u(x,0) = \sin \pi x \quad , \quad 0 \leq x \leq 1 \ , \tag{19b}$$

$$u(0,t) = u(1,t) = 0 \quad , \quad t > 0 \ . \tag{19c,d}$$

The exact solution of this simple problem is

$$u(x,t) = e^{-t} \sin \pi x \tag{20}$$

We solved Eq. (19) on a uniform mesh with N finite elements for one time step Δt using the methods described above and several choices of N and Δt. The effectivity index

$$\theta := \bar{e}^1 / \|u(\cdot, \Delta t) - U_{BE}^{1,1}\|_1 \tag{21}$$

(cf., e.g., Babuska et al. [15]), is used as a means of gaging the accuracy of the error estimate $\bar{e}^1$. Ideally, we would like θ not to differ appreciably from unity and to approach unity as $N \to \infty$ and $\Delta t \to 0$.

We present a summary of results for the reciprocal of the effectivity index for a sequence of calculations performed with $N = 2^{p+1}$ and $\Delta t = 5^{-p}/2$, $p = 0,1,\ldots,5$, in Figure 1. These results strongly suggest that $\theta \to 1$ as $p \to \infty$.

We use the temporal effectivity index

$$\theta_t := \|U_T^{1,1} - U_{BE}^{1,1}\|_1 / \|u(\cdot,\Delta t) - U_{BE}^{1,1}\|_1 \qquad (22)$$

as a method of appraising the accuracy of the temporal error estimate $\|U_T^{1,1} - U_{BE}^{1,1}\|_1$. For fixed Δt, $\theta_t \to K_t(\Delta t)$ as $N \to \infty$ and the limiting value $K_t(\Delta t) \to 1$ as $\Delta t \to 0$.

We solve Eq. (19) for a single time step using a sequence of meshes with $N = 2^p$, $p = 3,4,\ldots,10$, finite elements and time steps of $\Delta t = 0.7$, 0.49, 0.343, and 0.2401. We present our findings for the temporal effectivity index θ_t as a function of p for the four time steps in Figure 2. As expected, θ_t tends to a limiting value $K_t(\Delta t)$ for large N, which approaches unity as $\Delta t \to 0$.

Finally, we define the spatial effectivity index as

$$\theta_s := \|E_T^{2,1}\|_1 / \|u(\cdot,\Delta t) - U_{BE}^{1,1}\|_1 \qquad (23)$$

and use it as a measure of the spatial error estimate $\|E_T^{2,1}\|_1$. For fixed N, $\theta_s \to K_s(N)$ as $\Delta t \to 0$ and the limiting value $K_s(N) \to 1$ as $N \to \infty$.

Again, we solve Eq. (19) and present results for the reciprocal of the spatial effectivity index as a function of $\Delta t = 5^{-p}/2$, $p = 1,2,\ldots,7$, for meshes with $N = 2$, 4, and 8 finite elements. These results suggest that for a fixed N, $\theta_s \to K_s(N)$ as $p \to \infty$, and that $K_s(N)$ is reasonably close to unity. Futhermore, it appears that $K_s(N) \to 1$ as N increases.

IV. MESH REFINEMENT. The error estimates developed in Section III are used to control a simple global mesh refinement procedure that keeps $\bar{e}^n$ below a specified tolerance TOL. Suppose that a solution $U_{BE}^{1,n-1}$ and error estimates $E_T^{2,n-1}$ and $\bar{e}^{n-1}$ have been calculated at time t_{n-1} using a mesh with N elements and time step Δt_{n-1}. Further suppose that $\bar{e}^{n-1} < $ TOL and calculate solutions and error estimates at time $t_n = t_{n-1} + \Delta t_n$ using a mesh with N elements and time step $\Delta t_n = \Delta t_{n-1}$. Our refinement strategy consists of checking $\bar{e}^n$ and proceeding as follows:

(i) if $\bar{e}^n \leqslant$ TOL, continue to the next time step;

(ii) if $\bar{e}^n >$ TOL, and $0.3 < \|E_T^{2,n}\|/\bar{e}^n < 0.7$, double N, reduce Δt_n by thirty percent, and redo the integration;

(iii) if $\bar{e}^n >$ TOL and $0.7 \leqslant \|E_T^{2,n}\|/\bar{e}^n$, double N and redo the integration; and

(iv) if $\bar{e}^n$ > TOL and $\|E_T^{2,n}\|/\bar{e}^n$ ⩽ 0.3, reduce Δt_n by thirty percent and redo the integration.

Steps (ii) through (iv) are repeated until step (i) is satisfied.

The main advantage of this refinement procedure is that the separate estimates of the spatial and temporal errors allow different strategies to be used depending upon the dominant component of the error. Thus, if the spatial component of the error, as measured by $\|E_T^{2,n}\|/\bar{e}^n$ is large, then only spatial refinement is used to reduce the total error. The opposite situation arises when the spatial component of the error is small.

It is important to note that the error estimates used in the refinement procedure are, at best, only asymptotically correct. Thus, they will not produce reliable estimates on coarse meshes or when errors are large. With this in mind, it may be best to replace $\bar{e}^n$ by $\|U_T^{1,n} - U_{BE}^{1,n}\|_1 + \|\dot{E}_T^{2,n}\|_1$ in the refinement procedure.

The specific choice of the limiting values 0.3 and 0.7 that are used to determine the dominant component of the error in our refinement procedure are basically arbitrary. Under normal circumstances, the spatial error measure $\|E_T^{2,n}\|/\bar{e}^n \epsilon (0,1)$; thus, it is reasonable to divide $(0,1)$ approximately into thirds, i.e., $(0,0.3)$, $(0.3,0.7)$, and $(0.7,1)$ corresponding, respectively, to only temporal refinement, temporal and spatial refinement, and only spatial refinement. This strategy may not be appropriate in all situations and further analysis and experimentation is needed to determine optimal refinement criteria.

A local refinement strategy, such as those considered in [2,4-6,8-12], is usually more efficient than the global strategy presented herein. Our plans are to combine refinement with a mesh moving method that equidistributes a global error measure on a mesh with a fixed number of finite elements [16,17]. It may be possible to use a simple global refinement strategy in conjunction with such a mesh moving method since the local error measure will be approximately the same on every subinterval.

Doubling the number of finite elements whenever spatial refinement is performed simplifies interpolation issues, but may add more nodes than necessary. Reducing the time step by thirty percent keeps temporal accuracy comparable to spatial accuracy, since the temporal convergence rate is $O(\Delta t_n^2)$, while the spatial convergence rate is $O(1/N)$ (cf. Section III). Thus, doubling N would correspond to reducing Δt_n by $1/\sqrt{2}$, which is approximately thirty percent.

We apply the above refinement procedure to the following singular parabolic problem.

Example 2: Consider the partial differential system

$$u_t + u/[2(1-t)] = -u_{xx}/4 \quad , \quad 0 < x < \infty \quad , \quad 0 < t < 1 \quad , \tag{24a}$$

$$u(x,0) = e^{-x^2} \quad , \quad 0 \leqslant x < \infty \quad , \tag{24b}$$

$$u_x(0,t) = \lim_{s \to \infty} u_x(s,t) = 0 \quad , \quad 0 < t < 1 \quad . \tag{24c,d}$$

1107

The exact solution of this problem is

$$u(x,t) = e^{-x^2/(1-t)} .$$

(25)

This problem was motivated by our interest in solving the nonlinear Schrödinger equation in cylindrical coordinates [18]. It is known [19] that the solution of the Schrödinger equation can "self-focus," i.e., its solution can become infinite at one point, while decaying elsewhere on the domain. Problems of this type occur in laser optics.

Such problems are difficult to solve by traditional numerical methods and illustrate the need for adaptive strategies. The model given by Eq. (24) was developed as a simple approximation of the behavior of the Schrödinger equation. Its solution "focuses" in the sense that $u(x,t) \rightarrow 0$, $x > 0$, as $t \rightarrow 1$, while $u(0,t) = 1$.

This problem was solved for values of TOL of 0.2, 0.1, and 0.05 and a summary of the results are presented in Tables 1, 2, and 3, respectively. These tables present the relevant refinement data for each time step of the solution process. The time and numerical parameters Δt and N at the beginning of a solution step are found in the columns labelled "Initial Time," "Initial Δt," and "Initial N," respectively. The refined values of Δt and N, necessary to complete the solution step within the given tolerance, are found in the columns labelled "Refined Δt," and "Refined N," respectively. The resulting time at the end of a solution step is found in the column labelled "Final Time." The last column, labelled "Total Error Estimate," lists the value of e^n at the successful completion of a time step. The rows of each table outline the solution process as it advances through time.

These results indicate that it is sometimes possible to reduce the total error by refining only in space or only in time, and that the error estimates e^n, $\|U_T^{1,n} - U_{BE}^{1,n}\|_1$, and $\|E_T^{2,n}\|_1$ can be used to detect when these situations arise.

TABLE 1. NUMERICAL PARAMETERS AT THE BEGINNING AND THE END OF EACH TIME STEP FOR SOLVING EXAMPLE 2 WITH TOL = 0.2

Initial Time	Initial Δt	Initial N	Refined Δt	Refined N	Final Time	Total Error Estimate
0.0000	0.1250	16	0.1250	16	0.1250	0.0885
0.1250	0.1250	16	0.1250	16	0.2500	0.1496
0.2500	0.1250	16	0.0875	32	0.3375	0.1062
0.3375	0.0875	32	0.0875	32	0.4250	0.1931
0.4250	0.0875	32	0.0613	64	0.4863	0.1381
0.4863	0.0613	64	0.0429	128	0.5291	0.0782
0.5291	0.0429	128	0.0429	128	0.5720	0.1144
0.5720	0.0429	128	0.0429	128	0.6149	0.1404
0.6149	0.0429	128	0.0429	128	0.6578	0.1911
0.6578	0.0429	128	0.0300	256	0.6878	0.1052

TABLE 2. NUMERICAL PARAMETERS AT THE BEGINNING AND THE END OF EACH TIME STEP
FOR SOLVING EXAMPLE 2 WITH TOL = 0.1

Initial Time	Initial Δt	Initial N	Refined Δt	Refined N	Final Time	Total Error Estimate
0.0000	0.1250	16	0.1250	16	0.1250	0.0885
0.1250	0.1250	16	0.0875	32	0.2125	0.0496
0.2125	0.1250	32	0.0875	32	0.3000	0.0615
0.3000	0.0875	32	0.0613	64	0.3613	0.0390
0.3613	0.0613	64	0.0613	64	0.4225	0.0549
0.4225	0.0613	64	0.0429	64	0.4654	0.0604
0.4654	0.0429	64	0.0300	128	0.4954	0.0340
0.4954	0.0300	128	0.0300	128	0.5254	0.0528
0.5254	0.0300	128	0.0300	128	0.5554	0.0847
0.5554	0.0300	128	0.0210	128	0.5764	0.0782

TABLE 3. NUMERICAL PARAMETERS AT THE BEGINNING AND THE END OF EACH TIME STEP
FOR SOLVING EXAMPLE 2 WITH TOL = 0.5

Initial Time	Initial Δt	Initial N	Refined Δt	Refined N	Final Time	Total Error Estimate
0.0000	0.1250	16	0.1250	32	0.1250	0.0451
0.1250	0.1250	32	0.0875	64	0.2125	0.0261
0.2125	0.0875	64	0.0875	64	0.3000	0.0377
0.3000	0.0875	64	0.0613	64	0.3613	0.0375
0.3613	0.0613	64	0.0429	128	0.4041	0.0211
0.4041	0.0429	128	0.0429	128	0.4470	0.0300
0.4470	0.0429	128	0.0300	256	0.4470	0.0176
0.4470	0.0300	256	0.0300	256	0.5070	0.0246
0.5070	0.0300	256	0.0210	256	0.5280	0.0202
0.5280	0.0210	256	0.0210	256	0.5490	0.0272

V. DISCUSSION. We developed methods for calculating a posteriori esti-
mates of the total, spatial, and temporal discretization errors when a vector
system of parabolic partial differential equations is solved using piecewise
linear finite elements in space and the backward Euler method in time. The
error estimates are obtained by using higher order methods, with nodal supercon-
vergence being used to improve computational efficiency.

The three estimates were used to control a global refinement procedure that
attempts to keep a global measure of the error per time step below a prescribed
tolerance. Refinement can be performed in space, time, or both space and time
depending on the dominant component of the error estimate.

Comparison of the exact and estimated errors, presented in Example 1, give
us some confidence in the accuracy of our error estimates. Additionally, the
results of Example 2 provide an indication of the utility of these estimates in

an adaptive procedure. In certain situations, only spatial or temporal refinement was needed to keep the total error within the prescribed tolerance and our error estimates could be used to determine when these situations arise.

This is the first attempt that we know of which simultaneously addresses spatial and temporal errors with different refinement strategies. Some researchers [8-11] have used binary refinement in space and time, but did not attempt to determine the dominant component of the total discretization error. As noted, method of lines techniques [1,2,4,5] typically assume that temporal integration is exact and refine based on estimates of spatial errors. There is a great potential for techniques that utilize different spatial and temporal refinement strategies, particularly with problems having singularities. Our work, however, is still very preliminary and there is still a great deal to be done. Rigorous convergence results for our error estimates are yet to be established. The refinement algorithm of Section IV is very simple and will likely benefit from further experimental and theoretical analyses. We also anticipate that the inclusion of a mesh moving procedure based on equidistributing a global error measure [16] will dramatically improve the performance of our adaptive solution technique. In the future, we would like to extend our techniques to multi-dimensional problems and to consider higher order spatial and temporal discretization methods.

REFERENCES

1. S. Adjerid and J. E. Flaherty, "A Moving Finite Element Method With Error Estimation and Refinement for One-Dimensional Time Dependent Partial Differential Equations," SIAM J. Numer. Anal., Vol. 23, 1986, pp. 778-796.

2. S. Adjerid and J. E. Flaherty, "A Moving Mesh Finite Element Method With Local Refinement For Parabolic Partial Differential Equations," Comp. Meths. Appl. Mech. Engr., Vol. 56, 1986, pp. 3-26.

3. L. R. Petzold, "A Description of DASSL: A Differential/Algebraic System Solver," Sandia Report No. Sand. 82-8637, Sandia National Laboratory, Livermore, CA, 1982.

4. M. Bieterman and I. Babuska, "The Finite Element Method For Parabolic Equations, I: A Posteriori Error Estimation," Numer. Math., Vol. 40, 1982, pp. 339-371.

5. M. Bieterman and I. Babuska, "The Finite Element Method For Parabolic Equations, II: A Posteriori Error Estimation and Adaptive Approach," Numer. Math., Vol. 40, 1982, pp. 373-406.

6. S. Adjerid and J. E. Flaherty, "Local Refinement Finite Element Methods on Stationary and Moving Meshes For One-Dimensional Parabolic Systems," in preparation.

7. C. W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations, Prentice-Hall, Englewood Cliffs, NJ, 1971.

8. M. Berger and J. Oliger, "Adaptive Mesh Refinement For Hyperbolic Partial Differential Equations," J. Comput. Phys., Vol. 53, 1984, pp. 484-512.

9. M. Berger, "Data Structures For Adaptive Mesh Refinement," in Adaptive Computational Methods For Partial Differential Equations, (I. Babuska, J. Chandra, J. E. Flaherty, eds.), SIAM, Philadelphia, 1983.

10. M. Bieterman, J. E. Flaherty, and P. K. Moore, "Adaptive Local Refinement Methods For Nonlinear Parabolic Partial Differential Equations," 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, England, 1986, Chapter 19.

11. J. E. Flaherty and P. K. Moore, "A Local Refinement Finite Element Method For Time Dependent Partial Differential Equations," Transactions of the Second Army Conference on Applied Mathematics and Computing, ARO Report 85-1, US Army Research Office, Research Triangle Park, NC, 1985, pp. 585-596.

12. D. Gannon, "Self Adaptive Methods For Parabolic Partial Differential Equations," Ph.D. Thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, 1980.

13. V. Thomee, "Negative Norm Estimates and Superconvergence in Galerkin Methods for Parabolic Problems," Math. Comp., Vol. 34, 1980, pp. 93-113.

14. G. Strang and G. Fix, An Analysis of the Finite Element Method, Prentice-Hall, Englewood Cliffs, NJ, 1973.

15. I. Babuska, A. Miller, and M. Vogelius, "Adaptive Methods and Error Estimation For Elliptic Problems of Structural Mechanics," in Adaptive Computational Methods For Partial Differential Equations, (I. Babuska, J. Chandra, J. E. Flaherty, eds.), SIAM, Philadelphia, pp. 57-73.

16. J. M. Coyle, J. E. Flaherty, and R. Ludwig, "On the Stability of Mesh Equidistribution Strategies For Time-Dependent Partial Differential Equations," J. Comput. Phys., Vol. 62, 1986, pp. 26-39.

17. S. Davis and J. E. Flaherty, "An Adaptive Finite Element Method For Initial-Boundary Value Problems For Partial Differential Equations," SIAM J. Sci. Stat. Comput., Vol. 3, 1982, pp. 6-27.

18. J. E. Flaherty, J. M. Coyle, R. Ludwig, and S. F. Davis, "Adaptive Finite Element Methods For Parabolic Partial Differential Equations," in Adaptive Computational Methods For Partial Differential Equations, (I. Babuska, J. Chandra, and J. E. Flaherty, eds.), SIAM, Philadelphia, 1983, pp. 144-164.

19. K. Konno and H. Suzuki, "Self-Focusing of Laser Beams in Nonlinear Media," Physica Scripta, Vol. 20, 1979, pp. 382-386.
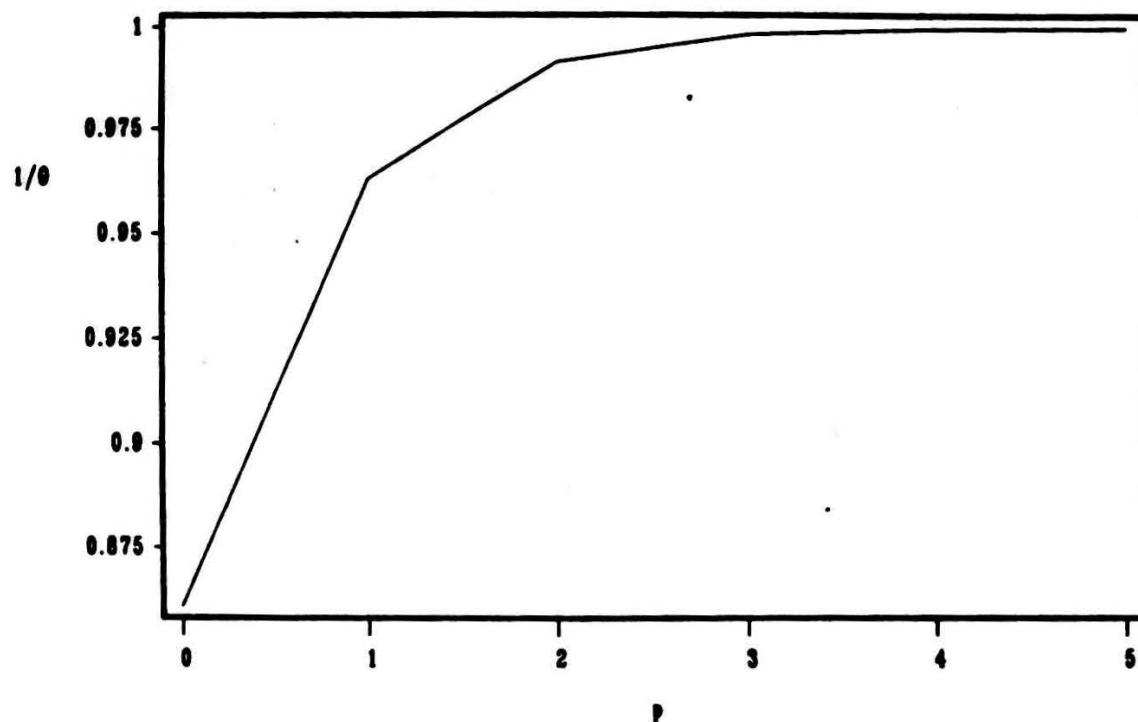
Figure 1. Reciprocal of the effectivity index θ versus p, where $N = 2^{p+1}$ and $\Delta t = 5^{-p}/2$ when solving Example 1. Note that θ approaches 1 as p increases.
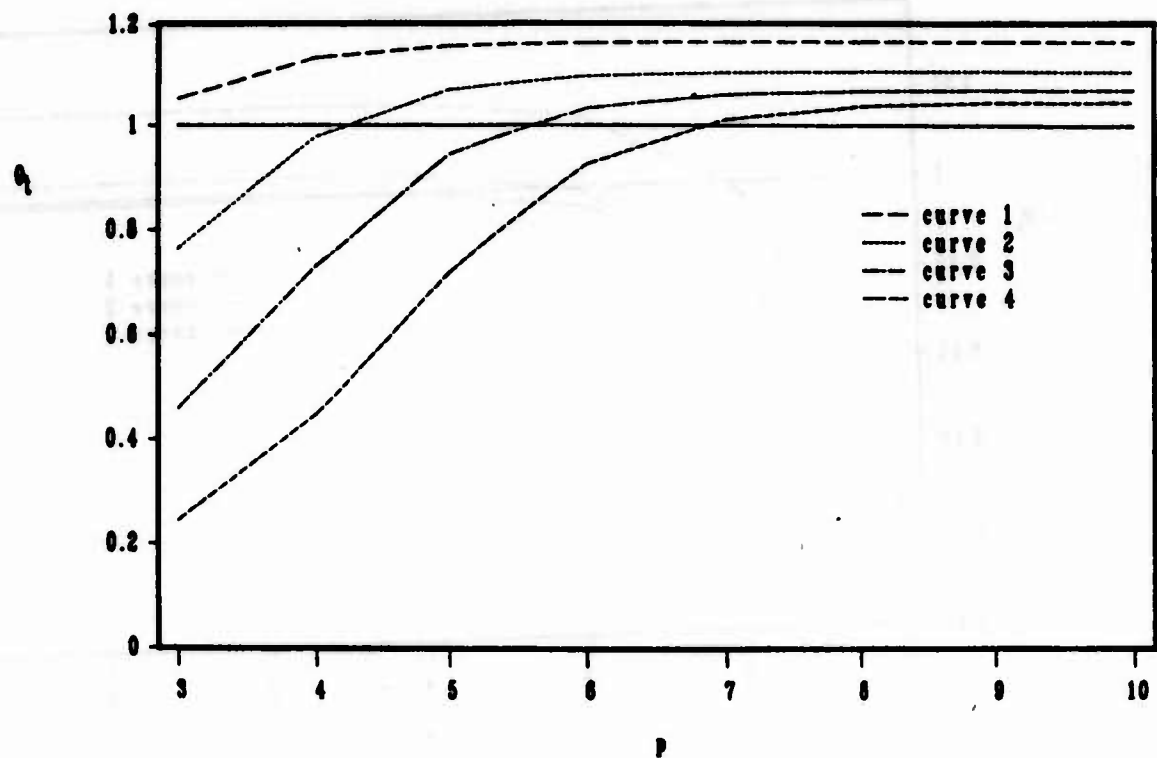
Figure 2. Temporal effectivity index θ_t for $\Delta t = 0.7$ (curve 1), $\Delta t = 0.49$ (curve 2), $\Delta t = 0.343$ (curve 3), and $\Delta t = 0.2401$ (curve 4), versus p, where $N = 2^p$ when solving Example 1. Note that θ_t approaches 1 as Δt decreases.
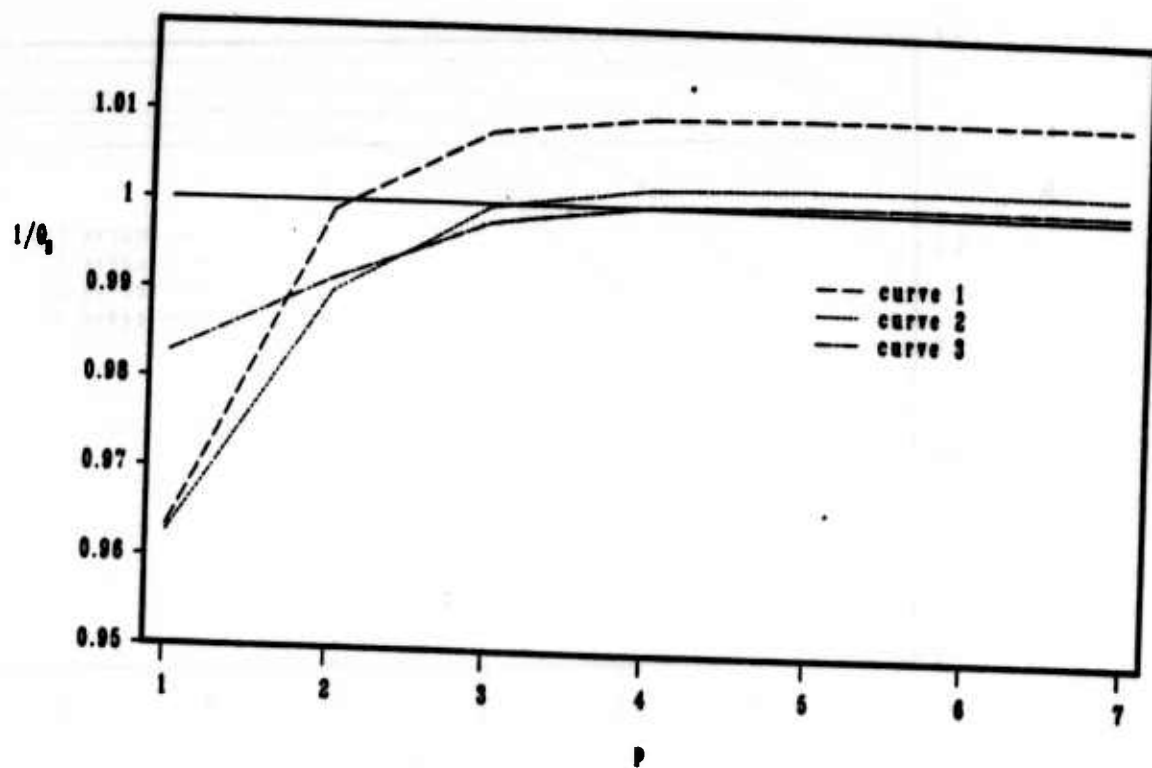
Figure 3. Reciprocal of the spatial effectivity index θ_s for $N = 2$ (curve 1), $N = 4$ (curve 2), and $N = 8$ (curve 3) versus p, where $\Delta t = 5^{-p}/2$ when solving Example 1. Note that θ_s approaches 1 as N increases.

AN ADAPTIVE METHOD WITH MESH MOVING AND LOCAL MESH REFINEMENT FOR TIME-DEPENDENT PARTIAL DIFFERENTIAL EQUATIONS[1]

David C. Arney

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

and

Joseph E. Flaherty

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

and

U.S. Army Armament, Munition, and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

ABSTRACT. We discuss mesh moving, static mesh regeneration, and local mesh refinement algorithms that can be used with a finite difference or finite element scheme to solve initial-boundary value problems for vector systems of time-dependent partial differential equations in two space dimensions and time. A coarse base mesh of quadrilateral cells is moved by an algebraic mesh movement function so as to follow and isolate spatially distinct phenomena. The local mesh refinement method recursively divides the time step and spatial cells of the moving base mesh in regions where error indicators are high until a prescribed tolerance is satisfied. The static mesh regeneration procedure is used to create a new base mesh when the existing ones becomes too distorted.

In order to test our adaptive algorithms, we implemented them in a system code with an initial mesh generator, a MacCormack finite difference scheme for hyperbolic systems, and an error indicator based upon estimates of the local discretization error obtained by Richardson extrapolation. Results are presented for several computational examples.

I. INTRODUCTION. Many initial-boundary value problems for time-dependent partial differential equations involve fine-scale structures that develop, propagate, decay, and/or disappear as the solution evolves. Some examples are shock waves in compressible flows, boundary and shear layers in viscous flows, and reaction zones in combustion processes. The numerical solution of these problems is usually difficult because the nature, location, and duration of the structures are often not known in advance. Thus, conventional numerical approaches that calculate solutions on a prescribed (typically uniform) mesh often fail to adequately resolve the fine-scale phenomena, have excessive

---

computational costs, or produce incorrect results. Adaptive procedures that evolve with the solution offer a robust, reliable, and efficient alternative. Such techniques have been the subject of a great deal of recent attention (cf. Babuska et al. [7, 9]) and are generally capable of introducing finer meshes in regions where greater resolution is needed [1, 2, 3, 6, 8, 10, 15, 16], moving meshes in order to follow isolated dynamic phenomena [1, 2, 5, 21, 23, 24, 25, 30], or changing the order of methods in specific regions of the problem domain [18, 22]. The utility of such adaptive techniques is greatly enhanced when they are capable of providing an estimate of the accuracy of the computed solution. Local error estimates are often used as refinement indicators and to produce solutions that satisfy either local or global accuracy specifications [1, 2, 3, 6, 8, 10, 15, 16]. Successful error estimates have been obtained using h-refinement [6, 15, 16], where the difference between solutions on different meshes is used to estimate the error, and p-refinement [1, 2, 3, 8, 16, 22] where the difference between methods of different orders are used to estimate the error.

We discuss an adaptive procedure that combines mesh movement and local refinement for m-dimensional vector systems of partial differential equations having the form

$$\mathbf{u}_t + \mathbf{f}(x,y,t,\mathbf{u},\mathbf{u}_x,\mathbf{u}_y) = [\mathbf{D}^1(x,y,t,\mathbf{u})\mathbf{u}_x]_x + [\mathbf{D}^2(x,y,t,\mathbf{u})\mathbf{u}_y]_y,$$
$$\text{for } t > 0, \quad (x,y) \in \Omega, \tag{1a}$$

with initial conditions

$$\mathbf{u}(x,y,0) = \mathbf{u}^0(x,y), \quad \text{for } (x,y) \in \Omega \cup \partial\Omega, \tag{1b}$$

and appropriate well-posed boundary conditions on the boundary $\partial\Omega$ of a rectangular region Ω.

We suppose that a numerical method is available for calculating approximate solutions and error indicators of (1) at each node of a moving mesh of quadrilateral cells. Any appropriate numerical method is applicable and the error indicator can either be an estimate of the local discretization error or another function (e.g., an estimate of the solution gradient or curvature) that is large where additional resolution is needed and small where less resolution is desired. Our adaptive algorithm consists of three main parts: (i) movement of a coarse base mesh, (ii) local refinement of the base mesh in regions where resolution is inadequate, and (iii) creation and regeneration of the base mesh when it becomes overly distorted. Our experience (cf. Section III) and that of others [25] indicates that mesh motion can substantially reduce errors for a very modest computational cost. Mesh motion alone, however, cannot produce a solution that will satisfy a prescribed error tolerance in all situations. For this reason, we have combined mesh motion with local mesh refinement and recursively solve local problems in regions where error tolerances are not satisfied. The local solution scheme successively reduces the domain size and, thus, further reduces the cost of the computation. Some problems, e.g., those with severe material deformations, can result in tangling and distortion of the moving base mesh. Therefore, we have created a procedure that automatically generates a new base mesh whenever the old one is unsuitable.

The adaptive procedures described in this paper combine our earlier work on mesh moving techniques [5] and local refinement procedures [6]. The inclusion of a static mesh regeneration scheme adds greater reliability and efficiency to these methods. The three components of our adaptive algorithm are described in Section II; however, frequent references are made to our previous investigations [5, 6]. A computer code based on the adaptive algorithm of Section II has been combined with a MacCormack finite difference scheme and an error indicator based on Richardson extrapolation. It has been used to

1116

solve a sequence of hyperbolic problems (i.e., problems having the form (1) with $D^1 = D^2 := 0$) and our findings on three examples, where we have attempted to appraise the relative costs and benefits of the mesh moving and local refinement portions of our adaptive algorithm, are reported in Section III. We have also compared solutions obtained by adaptive techniques to those obtained using stationary uniform meshes. In all three examples, solutions obtained by adaptive techniques cost less than solutions obtained on stationary uniform meshes having approximately the same accuracy. The mesh moving technique added approximately ten percent to the computational time of the adaptive algorithm and greatly improved the results. Most of the computational time was devoted to calculating the solution and error indicators, and not to the overhead induced by the refinement procedure. Although we are greatly encouraged by our results, our adaptive procedures are far from complete. Some possible improvements and future considerations are discussed in Section IV.

II. ALGORITHM DESCRIPTION. A top-level description of our adaptive procedure is presented in Figure 1 in a pseudo-PASCAL language. This procedure is called *adaptive_PDE_solver* and it integrates a system of partial differential equations from time *tinit* to *tfinal* and attempts to keep the local error indicators below a tolerance of *tol*. The base level time step Δt is initially specified, but may be changed, as needed, during the integration.

```
procedure adaptive_PDE_solver(tinit, Δt, tfinal, tol: real; M, N: integer);

    begin
        Generate an initial base mesh;
        t := tinit;

        while t < tfinal do
            begin
            Move the base mesh for the time step t to t + Δt;
            local_refine(0, t, Δt, tol);
            t := t + Δt;
            Select an appropriate Δt;
            if base mesh is too distorted then regenerate a base mesh
            end
    end { adaptive_PDE_solver };
```

Figure 1. Pseudo-PASCAL description of an adaptive procedure to solve the partial differential system (1) from $t = tinit$ to $tfinal$ to within a tolerance of *tol*.

The rectangular domain Ω is initially discretized into a coarse moving spatial grid of $M \times N$ quadrilateral cells. An initial base mesh is generated from this mesh by increasing the values of M and N, as necessary, and moving the mesh so that it is concentrated in regions where error indicators are large (cf. Section II.3). The base mesh is moved for each base time step Δt (cf. Arney and Flaherty [5] and Section II.1) and the partial

differential system (1) is solved on this mesh for a base time step. This is followed by a recursive local mesh refinement in regions where error indicators are larger than *tol*. The local refinement procedure *local_refine* was described in Arney and Flaherty [6] and its major features are summarized in Section II.2. The integration for each base-mesh time step is concluded by the selection of a new value of Δt for the subsequent time step and the generation of new base mesh (cf. Section II.3), if necessary.

The mesh moving, local refinement, and mesh regeneration algorithms are uncoupled from each other as well as from the procedures used to solve the partial differential system and calculate local error indicators. This reduces computational costs and provides a great deal of flexibility. Thus, individual modules can easily be replaced, omitted, or combined with other software.

II.1. Mesh Moving Algorithm. Mesh moving strategies should produce a smooth mesh where the sizes of neighboring computational cells vary slowly and cell angles differ only by modest amounts from right angles. It is, of course, essential for the nodes of the mesh to remain within Ω and for cells not to overlap. Meshes that violate these conditions can produce large discretization errors that overwhelm the positive effects of mesh moving. Our mesh moving procedure is based on an intuitive approach rather than more analytic error equidistribution (cf., e.g., Coyle et al. [19] or Dwyer [23]) and variational approaches (cf. Brackbill and Saltzman [17]). The essential idea is to move the mesh so as to follow isolated nonuniformities, such as wave fronts, shock layers, and reaction zones. This generally reduces dispersive errors and allows the use of larger time steps while maintaining accuracy and stability.

At each base time, we scan the $M \times N$ base mesh of quadrilateral cells and locate "significant error nodes" as those having error indicators greater than twice the mean nodal error indicator and also greater than ten percent of *tol*. This empirical strategy avoids having the mesh respond to fluctuations when error indicators are too small, but is sensitive enough to avoid missing dynamic phenomena associated with large error indicators. If there are no significant error nodes, computation proceeds on a stationary mesh. The nearest neighbor clustering algorithm of Berger and Oliger [15] is then used to gather the significant error nodes into clusters. In this iterative algorithm, a cluster is first defined to consist of one arbitrary significant error node. Other significant error nodes are added to the cluster if they are within a specified minimum intercluster distance from the nearest node in the cluster. We take the minimum intercluster distance to be the length of a cell diagonal. New clusters are established for nodes that do not belong to any existing cluster. Clusters are united when a node is determined to belong to more than one of them. Upon completion of the algorithm, (i) nodes in different clusters will be separated by at least the minimum intercluster distance, and (ii) no node in a cluster with more than one node will be further than the minimum intercluster distance from its nearest neighbor in the cluster.

Following Berger and Oliger [15], we generate near minimum area rectangles that contain each cluster. The principal axes of each rectangle are the major and minor axes of an enclosed ellipse having the same first and second moments as the nodes in the cluster. Thus, if (x_i, y_i) are the coordinates of a node and (x_m, v_m) are the mean coordinates of all nodes in the cluster, then the axes of the rectangle are in the directions of the eigenvectors of the symmetric (2×2) matrix

$$\begin{bmatrix} \sum (x_i^2 - x_m^2) & \sum (x_i y_i - x_m y_m) \\ \sum (x_i y_i - x_m y_m) & \sum (y_i^2 - y_m^2) \end{bmatrix}, \tag{2}$$

1118

where the summations range over all nodes in the cluster.

For many problems, there may be too small a percentage of significant error nodes within a cluster. In order to reduce this inefficiency and provide some alignment with, e.g., curved wave fronts, the rectangles are checked for efficiency by determining the percentage of significant error nodes in each rectangle. If a fifty-percent efficiency is not achieved, the rectangle is iteratively bisected in the direction of its major axis until all clusters have at least a fifty-percent efficiency.

We determine node movement from the velocity of propagation, the orientation, and the size of error clusters. We assume that nodes in the same cluster have related solution characteristics, so that we can determine individual node movement from the propagation of the center of the error cluster. Each cluster moves according to the differential equation

$$\ddot{\mathbf{r}}_m + \lambda \dot{\mathbf{r}}_m = 0, \tag{3}$$

where $\mathbf{r}_m(t) = [x_m(t), y_m(t)]^T$ is the position of the center of an error cluster and $(\dot{\ }) := d(\ )/dt$. The choice of the parameter λ can be critical in certain situations. If λ is selected too large, the system (3) will be stiff and computationally expensive. On the other hand, if λ is too small, the mesh can oscillate from time step-to-time step. Coyle et al. [19] and Adjerid and Flaherty [2] suggested some adaptive procedures for choosing λ; however, we found no appreciable differences in results or computation times when λ varied significantly. The examples of Section III were calculated with $\lambda = 1$.

We solve (3) for each base time step and each cluster using an explicit numerical method. The center of an error cluster is moved a distance $\Delta \mathbf{r}_m = \mathbf{r}_m(t+\Delta t) - \mathbf{r}_m(t)$ at the base time t. Let Δr_{m_1} and Δr_{m_2} denote the projections of $\Delta \mathbf{r}_m$ onto the major and minor axes of the rectangular cluster. We use the one-dimensional piecewise linear function

$$
d_{i,inside} = \begin{cases}
\Delta r_{m_i}(3/2 + x_i/w_i), & \text{if } -3w_i/2 \leq x_i \leq -w_i/2 \\
\Delta r_{m_i}, & \text{if } -w_i/2 < x_i < w_i/2 \\
\Delta r_{m_i}(3/2 - x_i/w_i), & \text{if } w_i/2 \leq x_i \leq 3w_i/2 \\
0, & otherwise
\end{cases}
, \quad i = 1, 2, \tag{4}
$$

to move the nodes of the mesh along the two principal axial directions of the error clusters. The cluster referred to in (4) has dimensions $w_1 \times w_2$ and (x_1, x_2) are local Cartesian coordinates of a node in the principal directions of the cluster relative to its center. For $i = 1$, nodes in the range of the cluster $(-3w_1/2 \leq x_1 \leq 3w_1/2, -w_2/2 \leq x_2 \leq w_2/2)$ are moved a distance $d_{1,inside}$. This situation is shown in Figure 2.

In order to maintain smooth mesh motion throughout the domain, nodes outside the range of a cluster move in a distance

$$d_{i,outside} = d_{i,inside}[1 - (2z/D)], \quad i = 1, 2, \tag{5}$$

where z is the shortest distance to the range of the cluster (cf. Figure 2) and D is the diagonal of Ω. For each cluster, the mesh is moved in the direction of the major axis ($i = 1$) using (4) and (5). This is followed by a similar procedure in the direction of the

1119

domain Ω
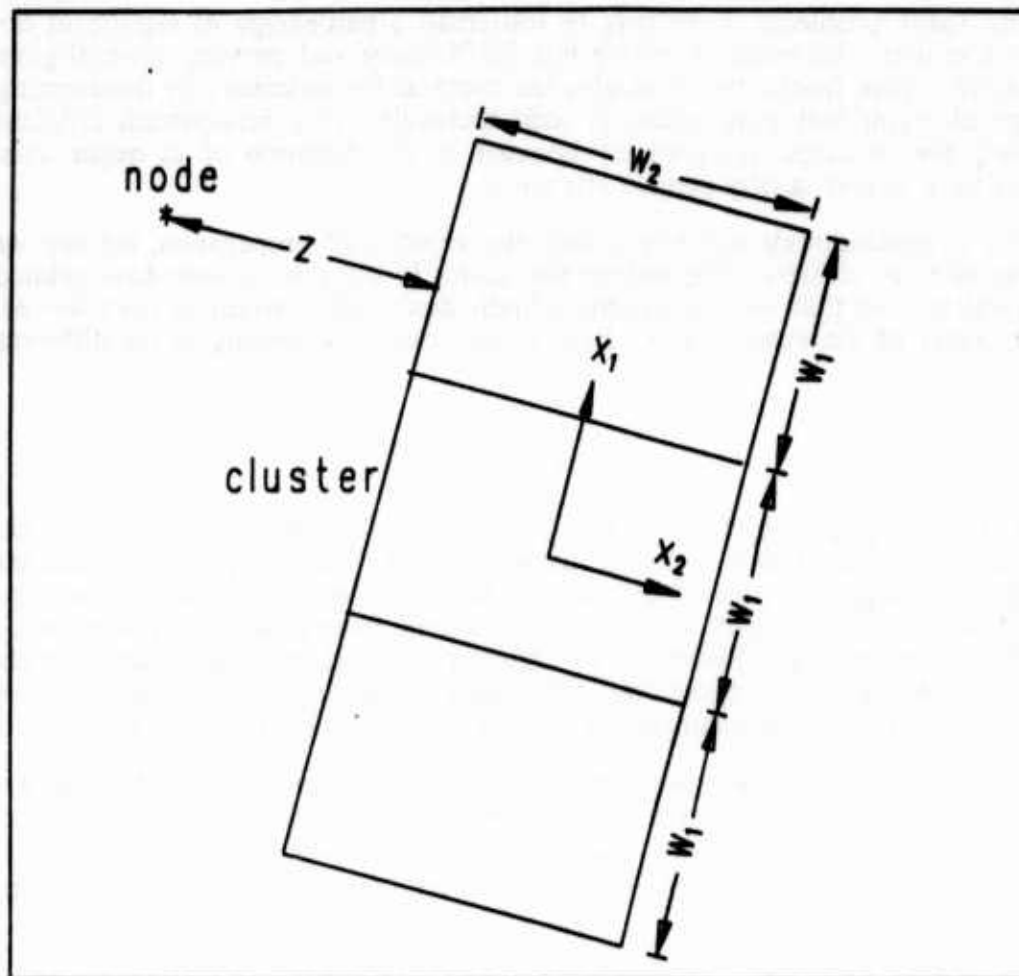
Figure 2. A rectangular $w_1 \times w_2$ error cluster. Nodes within the range of the cluster, $3w_1 \times w_2$, are moved a distance $d_{1,inside}$ in the x_1 principal direction according to eq. (4). Nodes outside the range of the cluster are moved a distance $d_{1,outside}$ in the x_1 direction according to eq. (5). The distance z is the shortest distance to the range of the cluster.

minor axis ($i = 2$). The distances $d_{i,inside}$ and $d_{i,outside}$ are reduced near $\partial\Omega$ in order to prevent nodes from leaving Ω. In particular, we recalculate $d_{i,j}$ as $d_{i,j}[min(1, b/c)]$, $i = 1, 2$, $j = inside, outside$, where b is the distance of the node to the boundary and c is twice the length of a cell diagonal on a uniform mesh having the same number of cells as the moving mesh. Nodes on domain boundaries, except corner nodes, which are not moved, are restrained to move along the boundary. Finally, the mesh moving algorithm is not restricted to the functions given by (4) and (5), and several other choices are possible.

II.2. Local Refinement Algorithm. As shown in Figure 1, the local refinement procedure is invoked after the base mesh has been moved for a base time step. Our

1120

refinement strategy consists of first calculating a preliminary solution on the base mesh for a base time step. An error indicator is used to locate regions where greater resolution is needed. Finer grids are adaptively created in these high-error regions by locally bisecting the time step and the sides of the quadrilateral cells of the base grid and the solution and error indicators are computed on the finer grids. The refinement scheme is recursive; thus, fine subgrids may be refined by adaptively creating even finer subgrids. This relationship leads naturally to a tree data structure. Information regarding the geometry, solution, and error indicators of the base grid is stored as the root node or level 0 of the tree. Subgrids of the base grid are offsprings of the root node and are stored as level 1 of the tree. The structure continues, with a grid at level l having a parent coarser grid at level $l - 1$ and any finer offspring grids at level $l + 1$. Grids at level l of the tree are given an arbitrary ordering and we denote them as $G_{l,j}$, $j = 1, 2, ..., N_l$, where N_l is the number of grids at level l. Our refinement procedures permit grids at the same level of a two-dimensional problem to intersect and overlap; however offspring grids must be properly nested within the boundaries of their parent grid. A one-dimensional grid with its appropriate tree structure for a base time step is shown in Figure 3.

A top-level pseudo-PASCAL description of a recursive local refinement algorithm that solves systems of the form (1) on the tree of grids described above is presented in Figure 4. The procedure local_refine integrates partial differential equations on the grids $G_{l,j}$, $j = 1, 2, ..., N_l$, at level l of the tree from time $tinit$ to $tinit + \Delta t$ and attempts to satisfy a prescribed local error tolerance tol. For each grid at level l, a solution and error indicators are calculated at time $tinit + \Delta t$. Additional finer grids are introduced in regions where the error indicators exceed the prescribed tolerance tol and the differential system is solved again on the finer grids using two time steps of duration $\Delta t/2$ and a tolerance of $tol/2$. Observe that the solution, error indicators, and refined subgrids are calculated for all grids at level l before calculating any solutions at level $l + 1$. Implicit in local_refine are the assumptions that a solution can be computed on any grid and that refinement terminates. If either of these assumptions are violated, the procedure terminates in failure.

Our technique for introducing finer subgrids consists of four steps: (i) an initial scan of each level l grid to locate "untolerable-error" nodes as those where the error indicator exceeds the prescribed tolerance tol, (ii) clustering any untolerable nodes into rectangular regions, (iii) buffering the clustered regions in order to reduce problems associated with prescribing initial and boundary conditions at coarse/fine grid interfaces, and (iv) cellularly refining the level l meshes and time step within the buffered clusters. Of course, if there are no untolerable-error nodes, the solution is acceptable and further refinement is unnecessary.

The same clustering algorithm of Berger and Oliger [15] that was used to move the base mesh is also used to group untolerable-error nodes for refinement. Each rectangular error cluster is enlarged by increasing its major and minor axes by twice the size of the average cell edge within the cluster. The region between the enlarged and original error clusters provides a buffer so that artificial internal boundary conditions (that are discussed below) will be prescribed at low-error nodes as far as possible and fine-grid errors will not propagate through the buffer in a time step.

Refined subgrids are created by bisecting the time step and edges of each cell of the parent mesh that intersects the buffered rectangular error clusters. Coarse mesh motion is maintained on the refined grids so that after two time steps of size $\Delta t/2$, cells of the refined grids will be properly nested within those of their parent grid. Additional details of the refinement algorithm and data structures are presented in Arney and Flaherty [6].

Artificial initial and boundary data must be determined from solutions on other grids in order to calculate the solution and error indicators on refined subgrids. Furthermore,
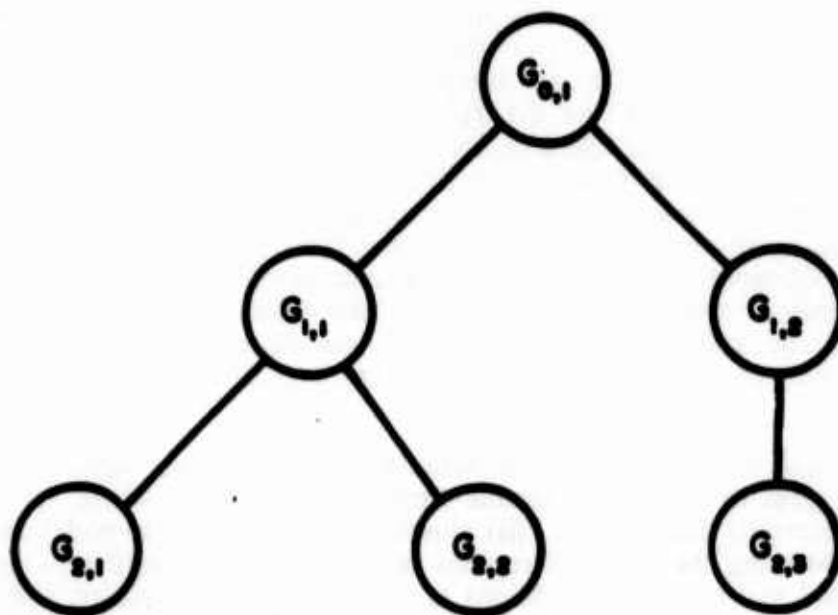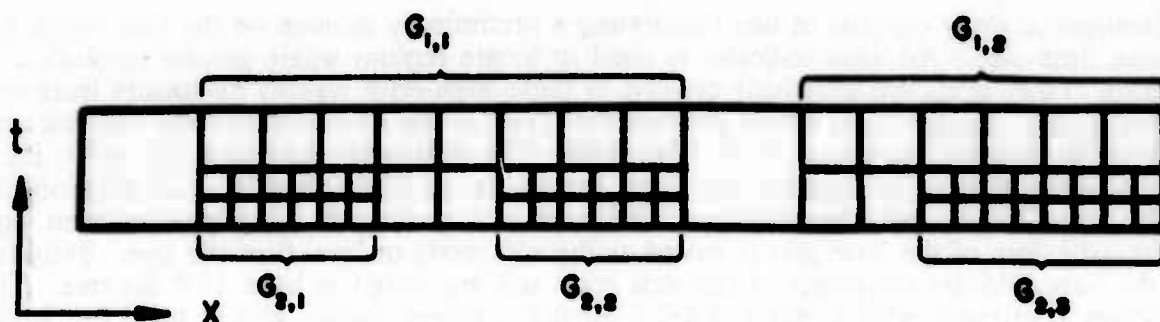
1121

Figure 3. Coarse and refined grids (top) and their tree representation (bottom) for a one-dimensional example.

solutions on finer grids are used to replace those on coarser grids at common nodal locations.

Initial data for a subgrid is calculated directly from the initial function $u^0(x,y)$ at $t = 0$. For $t > 0$, initial data is obtained by interpolation using the solution at the same time on the finest available mesh. In order to provide data for this interpolation, we save all solution values on previous subgrids until they are no longer needed due to advancement in time of an acceptable solution. Bilinear functions using the solution values at the four vertices of the finest existing cell are used to obtain the solution at the nodes of cells

```
procedure local_refine( l: integer; tinit, Δt, tol: real );

begin
    for j := 1 to N[l] do
        begin
            Integrate the partial differential system from tinit to tinit + Δt
                on grid G[l,j];
            Calculate error indicators at tinit + Δt at all nodes of grid
                G[l,j];
            if any error indicators > tol then introduce level l + 1 subgrids
                of G[l,j]
        end { for };

    if any error indicators > tol then
        begin
            local_refine(l + 1, tinit, Δt/2, tol/2);
            local_refine(l + 1, tinit + Δt/2, Δt/2, tol/2)
        end
end { local_refine };
```

Figure 4. Pseudo-PASCAL description of a recursive local refinement procedure to find a solution of the partial differential system (1) on all grids at level l of the tree.

of the refined mesh. Further analysis is needed regarding the effects on accuracy and stability and the proper order of this interpolation. Bieterman, Flaherty, and Moore [16] give an example where the fine-scale structure of a solution was lost by interpolation from too coarse a mesh

In a similar manner, boundary data for refined meshes are calculated directly from the prescribed boundary conditions on portions of subgrids that intersect $\partial\Omega$. Dirichlet boundary data is prescribed on the edges of subgrids that are in the interior of Ω by interpolating the solution from coarser meshes. Bilinear functions using the solution values at the four vertices of the adjacent face of the finest existing space-time cell are used to obtain solution values for the nodes of refined cells.

Acceptable fine-mesh solutions are used to replace solutions at the nodes of coarser grids that lie within the untolerable-error portions of clusters. Solutions at low-error nodes in the buffer zones of clusters are not replaced in order to avoid possible contamination of accurate solutions. When fine grids overlap each other in an untolerable-error region, the average value of the solutions at common fine-grid nodes is used to replace the appropriate coarse grid solution. Boundary effects do not propagate through a sufficiently large buffer and, thus, have no effect on the solution within the untolerable-error region of a cluster when an explicit numerical scheme is used for the integration. Greater care is needed when implicit integration methods are used, since artificial boundary conditions can affect the accuracy, convergence, and stability of the solution at all nodes in the cluster regardless of the size of the buffer.

Stability and conservation of, e.g., fluxes at interfaces between coarse and fine meshes must be investigated further, particularly in two dimensions. For one-dimensional problems, Berger and Oliger [15] showed that linear interpolation of solutions from a coarse to a fine mesh produced no instabilities in the the Lax-Wendroff scheme. Berger [14] also discussed conservation at mesh interfaces and proposed explicit enforcement of conserved quantities at coarse/fine mesh boundaries. Rai [29] presented some finite difference schemes that maintained conservation at grid interfaces for two-dimensional compressible flow problems.

II.3. Initial Mesh Construction and Regeneration. The efficiency of our adaptive mesh moving and refinement strategies are dependent on our ability to generate a suitable initial mesh and to regenerate a new base mesh should it become severely distorted at later times. The proper base mesh can reduce the need for refinement and, thus, increase efficiency.

The two essential elements of a mesh generation or regeneration procedure are the determination of the number of nodes and their optimal location. A base mesh having too few nodes will result in excessive refinement while one having too many nodes will reduce efficiency. Many mesh generation procedures have been developed (cf., e.g., Thompson [31] or Brackbill and Saltzman [17]); however, the best one to use in conjunction with an adaptive procedure is still far from being established. Our current approach to mesh generation is to use the error indicators computed by a trial solution to determine an initial mesh that approximately equidistributes the error indicators.

To begin, we create a uniform $M \times N$ rectangular mesh using prescribed values of M and N that reflect the coarsest mesh that should be used to calculate a solution. We solve the system (1) for a base time step Δt on the uniform stationary base mesh and compute the solution and error indicators. Local mesh refinement is performed as described in Section II.2 until the prescribed tolerance is attained. We use this solution to determine the number of nodes K in a new base mesh as

$$K = MN + \sum_{l=1}^{n} (3/4)^l K_l, \tag{6a}$$

where K_l is the number of nodes introduced at level l and n is the total number of levels in the tree. Having computed K, we calculate the dimensions of a new $\overline{M} \times \overline{N}$ mesh as

$$\overline{M} = \sqrt{KM/N}, \quad \overline{N} = \sqrt{KN/M}. \tag{6b}$$

The bars have been omitted on M and N in the algorithms displayed in Figures 1 and 4 and in all further discussions.

Node placement for the new base mesh is accomplished by locating all nodes of the original base mesh having error indicators that are greater than twice the mean error indicator. These nodes are then grouped into rectangular clusters using the clustering algorithm of Section II.1. A uniform base mesh is generated when there are no nodes having error indicators that are greater than twice the mean error indicator.

Nodes are moved towards the center of the nearest error cluster unless they are within a two-cell diagonal range of two or more error clusters. In the former case, a node is moved four-tenths of its distance to the center of the nearest cluster unless this distance is greater than 12.5 times the average cell diagonal, in which case it is moved five times the average cell diagonal. Nodes that are within a two-cell diagonal range of two or more clusters are moved by four-tenths of a weighted average of the distances to centers of the

involved clusters. Nodes on ∂Ω remain on ∂Ω. Nodes near the boundary move a reduced distance in order to prevent the formation of large elements. When an error cluster intersects opposite boundaries of Ω, nodes are not moved in the direction of the major axis of the cluster. This construction generates a base mesh that depends on the solution of the partial differential system as well as its initial condition.

The base mesh can become severely distorted for some problems (cf. Arney and Flaherty [5]) and we would like a capability for generating a new base mesh whenever this happens. Since the new mesh is created at a specific time, rather than by mesh motion, we refer to this process as static mesh regeneration. Our static mesh regeneration procedure consists of three steps: (i) determining that there is a need for a new base mesh, (ii) creating the new base mesh, and (iii) interpolating the solution from the old to the new base mesh.

A mesh is regenerated when any interior angle of a cell is less than 50 or greater than 130 degrees, the aspect ratio of any cell is greater than 15, or the mesh ratio of adjacent cells exceeds 5 or is less than 1/5. In the present context, the aspect ratio is defined as the average length divided by the average width of a cell and the mesh ratios are defined as the ratio of the lengths and widths of adjacent cell sides.

A new base mesh, having the same number of nodes as the old one, is generated using the procedure described above for creating an initial base mesh. The error clusters for the existing mesh are used to generate the new base mesh, so that new clusters do not have to be computed. This process appears to reduce angle deviations from ninety degrees, control aspect ratios, and mollify adjacent mesh ratios.

Once a new base mesh has been constructed, the solution on the old one is interpolated to the new one by using bilinear interpolation with respect to the cells of the old base mesh. The order and nature of the interpolation needs further investigation and we are studying methods that conserve, e.g., fluxes (cf. Berger [14] or Rai [29]).

III. COMPUTATIONAL EXAMPLES.

In order to demonstrate the capabilities of the adaptive procedure described in Section II, we applied it to three hyperbolic systems. We used a two-step MacCormack finite difference method (cf. Arney and Flaherty [5], Hindman [26], or MacCormack [27]) to integrate the partial differential equations and Richardson's extrapolation (cf. Arney [4] or Berger and Oliger [15]) to indicate local errors. Base mesh geometry was prescribed as indicated in each example. If the base mesh time step failed to satisfy the Courant, Friedrichs, Lewy theorem, it was automatically reduced to the maximum allowed by the Courant condition (cf. Arney [4] and Arney and Flaherty [6]). This procedure should also satisfy the Courant condition on all subgrids when the characteristic speeds vary slowly.

Numerical results obtained on uniform stationary grids are compared with those obtained by adaptive strategies that use (i) mesh moving only, (ii) local refinement only, and (iii) the combination of mesh moving and refinement discussed in Section II. The examples are designed to determine the relative cost, accuracy, and efficiency of our adaptive algorithm and each of its components. Accuracy is appraised by computing the difference e between the exact and numerical solutions of a problem in either the maximum or L_1 norms, i.e., by computing either

$$\|e(\cdot,\cdot,t)\|_\infty := \max_{1 \le i \le K} \max_{1 \le j \le m} |e_j(x_i, y_i, t)|, \tag{7a}$$

or

$$\|e(\cdot,\cdot,t)\|_1 = \iint_\Omega P \sum_{j=1}^m |e_j| \, dxdy, \tag{7b}$$

1125

respectively. Here, K is the number of nodes in the mesh at time t and P is a piecewise constant interpolation operator with respect to the cells of the base mesh that, on each cell, has the average value of the errors at the vertices of the cell. We use either the total CPU time or the maximum number of nodes used in a base time step as measures of the computational complexity of a procedure. All calculations were performed in double precision arithmetic on an IBM 3081/D computer at the Rensselaer Polytechnic Institute.

Solutions are displayed by drawing either level lines or wire-frame perspective renditions. Meshes are displayed by showing the complete two-dimensional spatial discretization at specified times with finer subgrids overlaying coarser ones. This portrayal does not show the reduced time steps that are used for the subgrid calculations. The broken-line rectangles in the figures indicates the error cluster(s) that are used to move the base mesh.

Example 1. Consider the linear initial-boundary value problem that was proposed as a test problem by McRae et al. [28]:

$$u_t - yu_x + xu_y = 0, \quad t > 0, \quad (x,y) \in \Omega, \tag{8a}$$

$$u(x,y,0) = \begin{cases} 0, & \text{if } (x-1/2)^2 + 1.5y^2 \geq 1/16 \\ 1 - 16((x-1/2)^2 + 1.5y^2), & \text{otherwise}, \end{cases}$$
$$(x,y) \in \Omega \cup \partial\Omega, \tag{8b}$$

and

$$u(x,y,t) = 0, \quad t > 0, \quad (x,y) \in \partial\Omega, \tag{8c}$$

where $\Omega := \{ (x,y) \mid -1.2 < x, y < 1.2 \}$.

The exact solution of (8) is an elliptical cone that rotates about the origin in the counterclockwise direction with period 2π. It can be written in the form

$$u(x,y,t) = \begin{cases} 0, & \text{if } C < 0 \\ C, & \text{if } C \geq 0, \end{cases} \tag{9a}$$

where

$$C = 1 - 16[(x\cos t + y\sin t - 1/2)^2 + 1.5(y\cos t - x\sin t)^2]. \tag{9b}$$

Five adaptive and uniform mesh solutions of (8) were calculated for $0 < t \leq 3.2$ and our findings are summarized in Table 1. Solutions 3 and 4, with refinement, were calculated using an error tolerance of 0.0002 and a maximum of two levels of refinement. The tolerance and maximum level of refinement were selected so that the high-error region under the cone would maintain approximately the same mesh spacing as the uniform mesh used to obtain Solution 5. The grids that were used to obtain Solution 4 are shown in Figure 5 at $t = 0.56, 1.68, 2.24,$ and 3.2. A new base mesh was introduced at $t = 2.82$. The meshes that were used to obtain Solutions 2, 3, and 4 at $t = 3.2$ are shown in Figure 6. Finally, surface and contour plots of Solutions 1, 2, and 3 and of Solutions 4 and 5 at $t = 3.2$ are shown in Figures 7 and 8. respectively.

Solution 1 bears no resemblance to the exact solution and demonstrates the devastating effects of large dissipative and dispersive errors. Solution 2, with mesh moving only, provides a dramatic improvement in the results for approximately one-half the cost of using both mesh motion and refinement. Solution 5 took more than three-times longer to

Ref. No.	Strategy	Base Mesh	Base Time Step	$\|e\|_1$	$\|e\|_\infty$	CPU Time (sec.)
1	Stationary uniform mesh	14×14	0.056	0.2560	0.78	46
2	Moving mesh	32×32	0.026	0.0301	0.20	458
3	Stationary mesh with refinement	14×14	0.056	0.0832	0.48	852
4	Moving mesh with refinement	14×14	0.056	0.0249	0.18	904
5	Stationary uniform mesh	56×56	0.014	0.0759	0.48	2647

Table 1. Errors at $t = 3.2$ and computational costs for five solutions of Example 1.

calculate than Solution 4 for approximately the same accuracy; thus, demonstrating the efficiency of the refinement process. The subgrids for the refined Solutions 3 and 4 are concentrated in the region of the cone and are aligned with its principal axes as it rotates. Dissipative and dispersive errors cause a "wake" of spurious oscillatory information to follow the moving cone (cf. Figures 7 and 8). Some mesh refinement is performed in the wake region and this greatly reduces the magnitude of the oscillations.

Example 2. Consider the uncoupled linear initial-boundary value problem

$$u_{1_t} + u_{1_x} = 0, \quad u_{2_t} - u_{2_x} = 0, \quad t > 0, \quad (x,y) \in \Omega, \tag{10a}$$

$$u_1(x,y,0) = \begin{cases} 1 - 16((x-1/2)^2 + 1.5y^2), & \text{if } (x-1/2)^2 + 1.5y^2 \le 1/16 \\ 0, & \text{otherwise}, \end{cases}$$
$$(x,y) \in \Omega \cup \partial\Omega, \tag{10b}$$

$$u_2(x,y,0) = \begin{cases} 1 - 16((x+1/2)^2 + 1.5y^2), & \text{if } (x+1/2)^2 + 1.5y^2 \le 1/16 \\ 0, & \text{otherwise}, \end{cases}$$
$$(x,y) \in \Omega \cup \partial\Omega, \tag{10c}$$

$$u_1(x,y,t) = u_2(x,y,t) = 0, \quad t > 0, \quad (x,y) \in \partial\Omega, \tag{10d}$$

and $\Omega := \{ (x,y) \mid -1 \le x \le 1, -0.6 \le y \le 0.6 \}$.

The solution of this problem consists of two moving cones that collide and pass through each other. We selected it in order to determine how the various adaptive strategies could cope with interacting phenomena.

1127
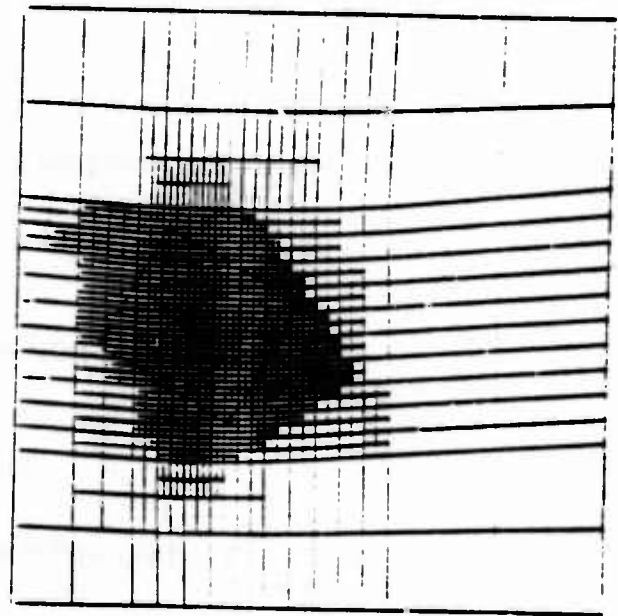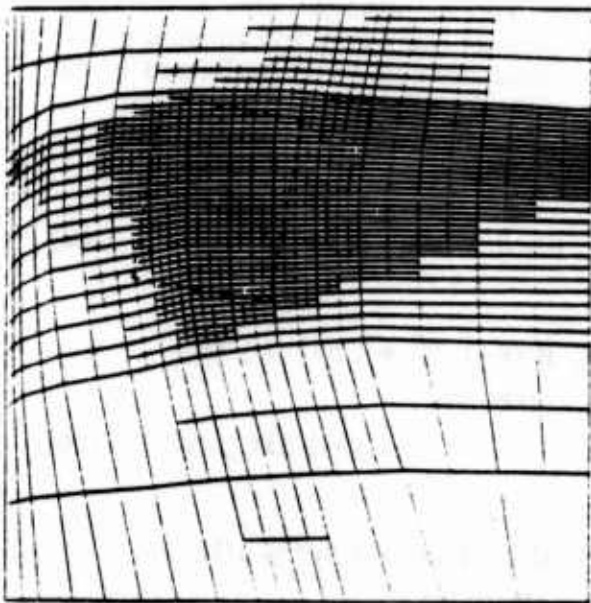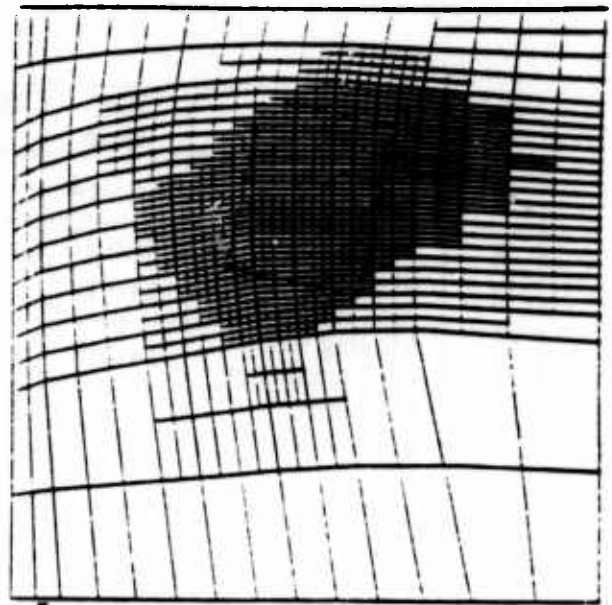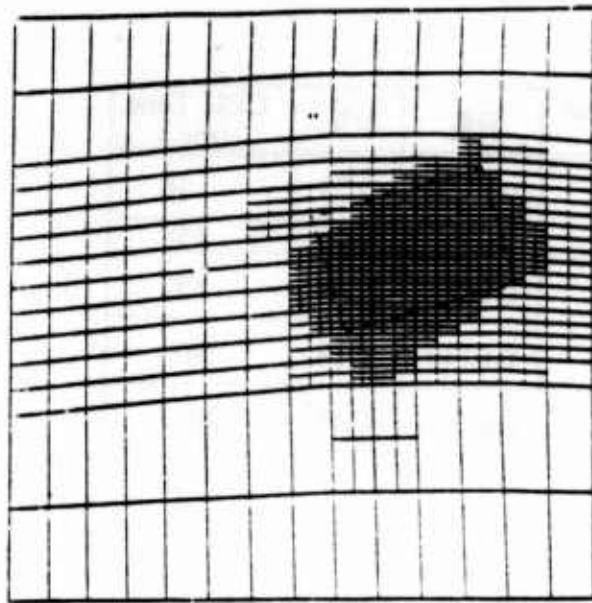
Figure 5. Grids created for Solution 4 of Example 1 at $t = 0.056$ (upper left), 1.68 (upper right), 2.24 (lower left), and 3.2 (lower right).
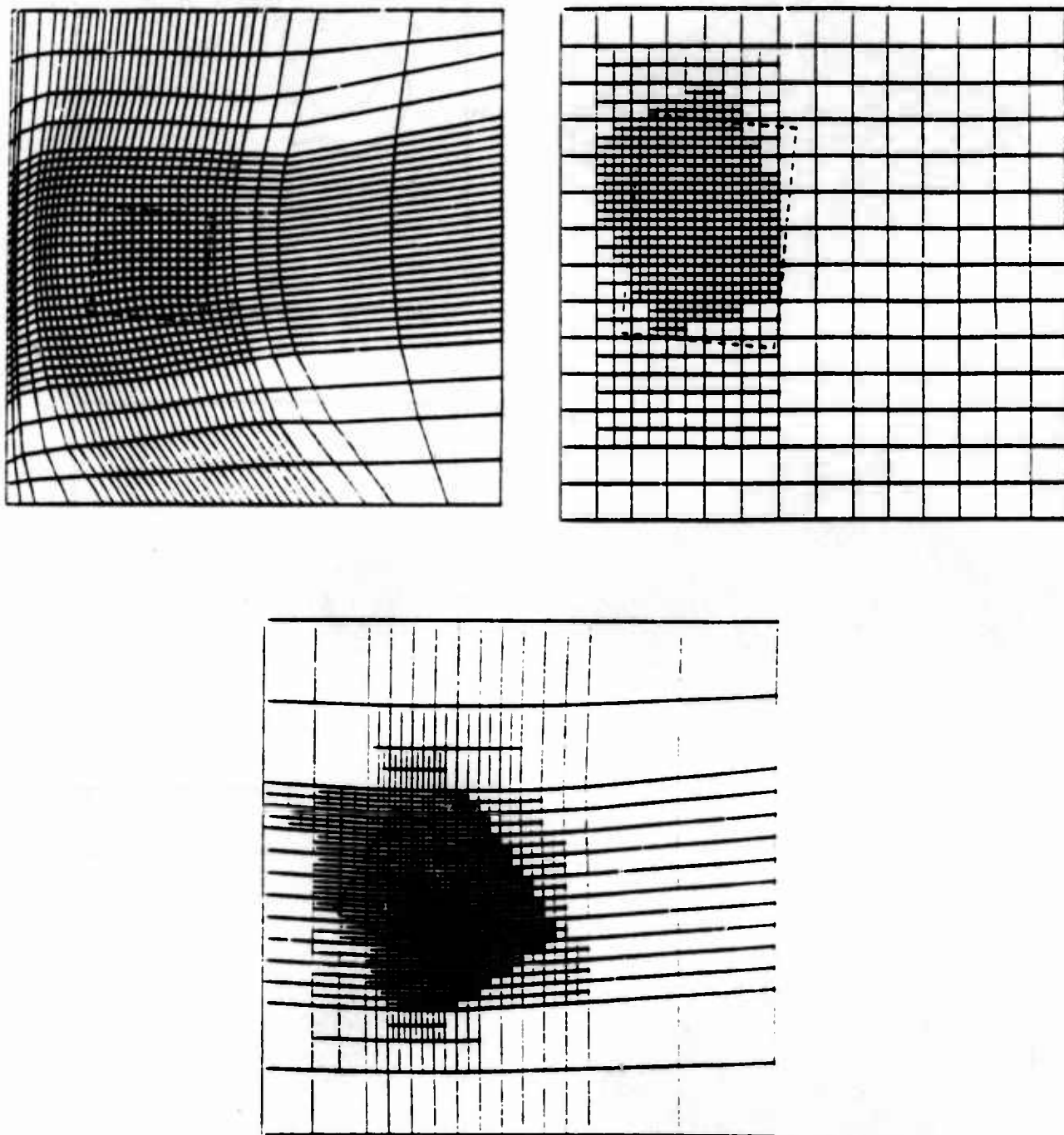
Figure 6. Grids created for Solutions 2, 3, and 4 of Example 1 at $t = 3.2$.
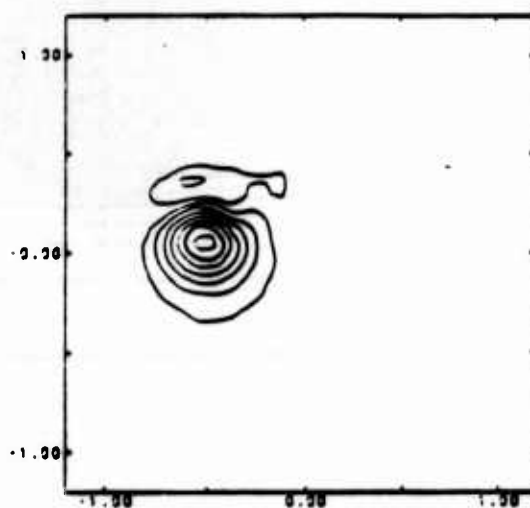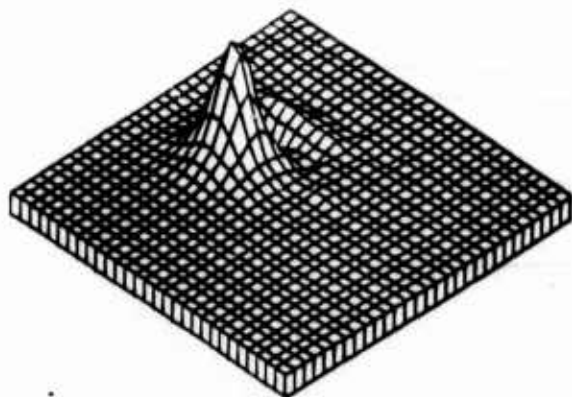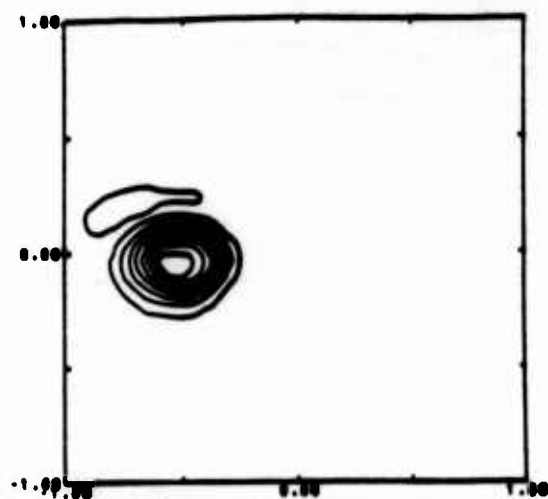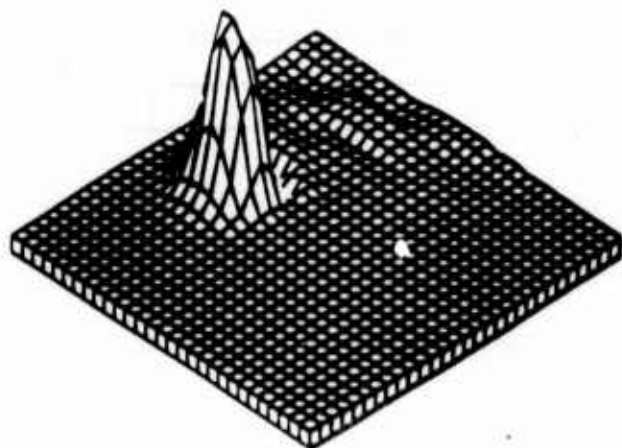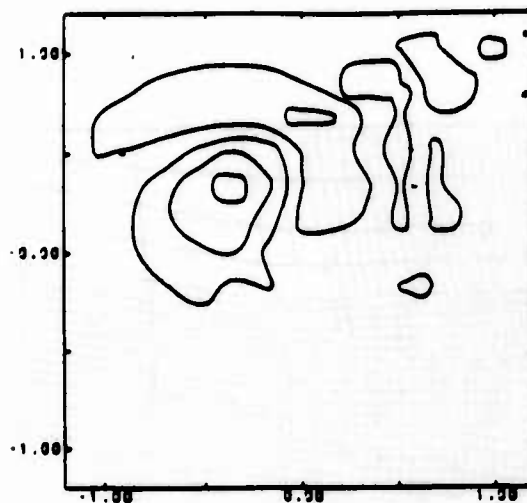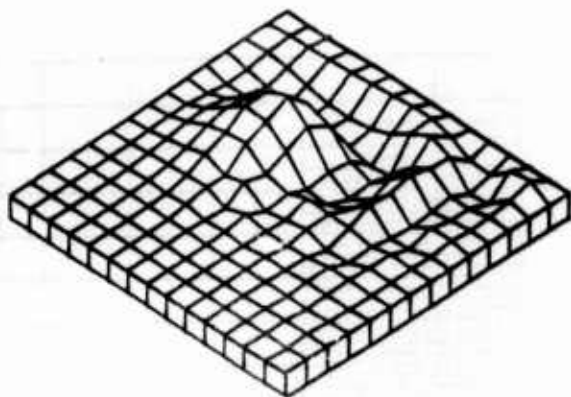
Figure 7. Surface and contour plots for Solutions 1, 2, and 3 (top to bottom) at $t = 3.2$ of Example 1.
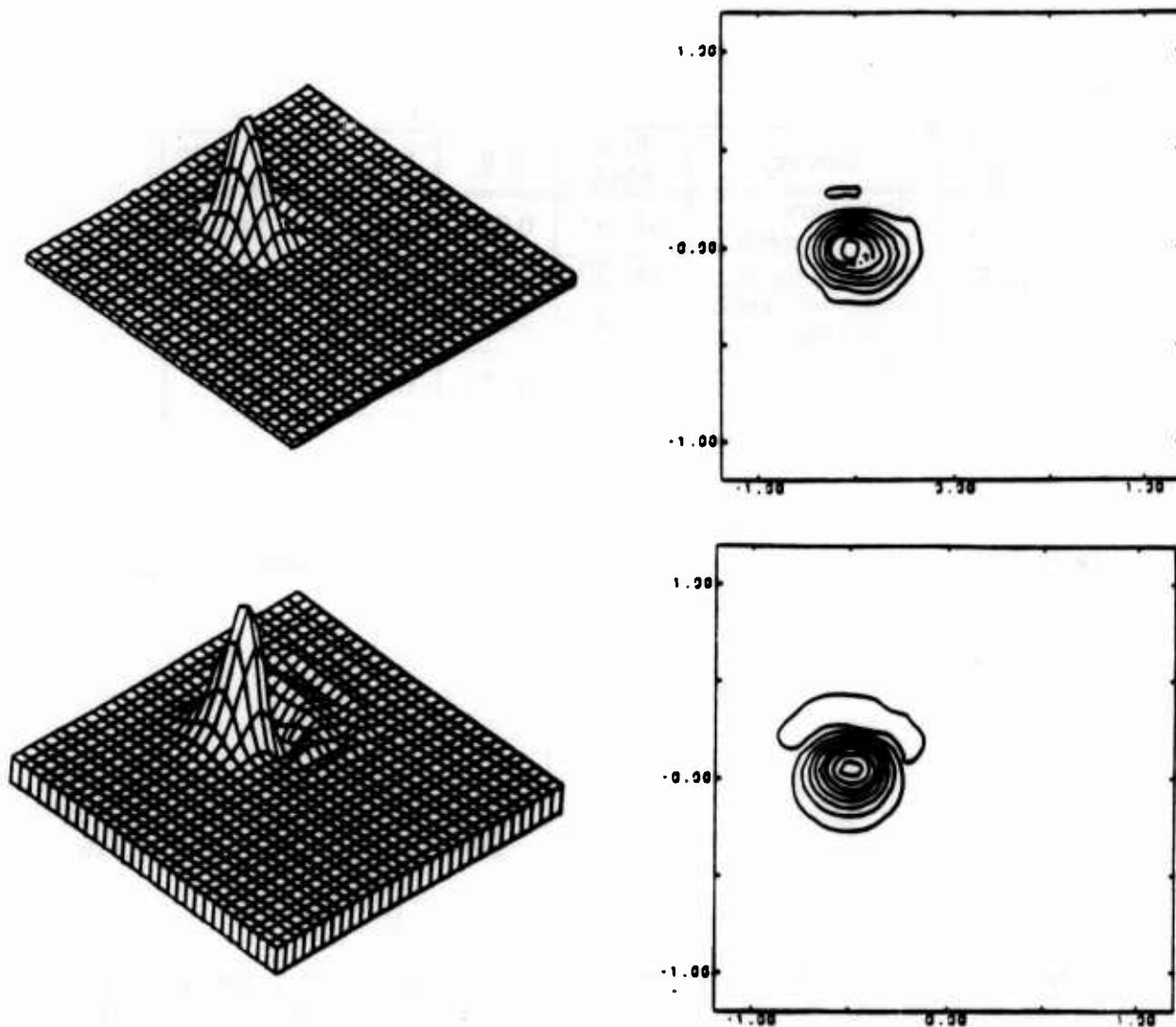
Figure 8. Surface and contour plots for Solutions 4 (top) and 5 (bottom) at $t = 3.2$ of Example 1.

One uniform mesh and three adaptive solutions of (10) were calculated for $0 < t \leq 1.2$ and our findings are summarized in Table 2. The solutions involving refinement were computed with a tolerance of 0.0038. All solutions were designed to have approximately the same accuracy. The grids that were used to obtain Solution 4 are shown in Figure 9 at $t = 0, 0.23, 0.46, 0.92$, and 1.2.

The results of Table 2 demonstrate the efficiency of the mesh moving strategy on this example. Solution 2 with mesh moving was slightly more accurate than Solution 1 obtained on a uniform mesh, and it required less than one-half of the computation time. Solution 3 with refinement on a stationary mesh shows only a modest improvement over Solution 1; however, the combination of mesh moving and refinement computed in Solution 4 again shows a significant gain in efficiency. We suspect that the high accuracy achieved by mesh moving on this example is due to the reduction in dispersive errors that results when the mesh follows the cones with approximately the correct velocity.

Example 3. Consider the Euler equations for a perfect inviscid compressible fluid

$$u_t + f_x(u) + g_y(u) = 0, \tag{11a}$$

1131

Ref. No.	Strategy	Base Mesh	$\|e\|_1$	$\|e\|_\infty$	CPU Time (sec.)
1	Stationary uniform mesh	64×34	0.066	0.26	710
2	Moving mesh	44×20	0.056	0.18	340
3	Stationary mesh with refinement	44×20	0.055	0.23	719
4	Moving mesh with refinement	44×20	0.039	0.16	609

Table 2. Errors at $t = 1.2$ and computational costs for four solutions of Example 2.

where

$$
\mathbf{u} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ e \end{bmatrix}, \quad
\mathbf{f(u)} = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ u(e+p) \end{bmatrix}, \quad
\mathbf{g(u)} = \begin{bmatrix} \rho u \\ \rho u v \\ \rho v^2 + p \\ v(e+p) \end{bmatrix}. \tag{11b,c,d}
$$

Here, u and v are the velocity components of the fluid in the x and y directions, ρ is the fluid density, e is the total energy of the fluid per unit volume, and p is the fluid pressure. For an ideal gas

$$
p = (\gamma - 1)[e - \rho(u^2 + v^2)/2], \tag{11e}
$$

where γ is the ratio of the specific heat at constant pressure to that at constant volume.

We solve a problem where a Mach 10 shock in air ($\gamma = 1.4$) moves down a channel containing a wedge with a half-angle of thirty degrees. This problem was used by Woodward and Collela [32] to compare several finite difference schemes on uniform grids. Like them, we orient a rectangular computational domain, $-0.3 \leq x \leq 3.4$, $0 \leq y \leq 1$, so that the top edge of the wedge is on the bottom of the domain in the interval $y = 0$, $1/6 \leq x \leq 3.4$. Thus, in the computational domain it appears like a Mach 10 shock is impinging on a flat plate at an angle of sixty degrees. The initial conditions that are appropriate for this situation are

$$
\rho = 8.0, \quad p = 116.5, \quad e = 563.5, \quad u = 4.125\sqrt{3}, \quad v = -4.125,
$$
$$
\text{if } y < \sqrt{3}(x - 1/6), \tag{12a}
$$

and
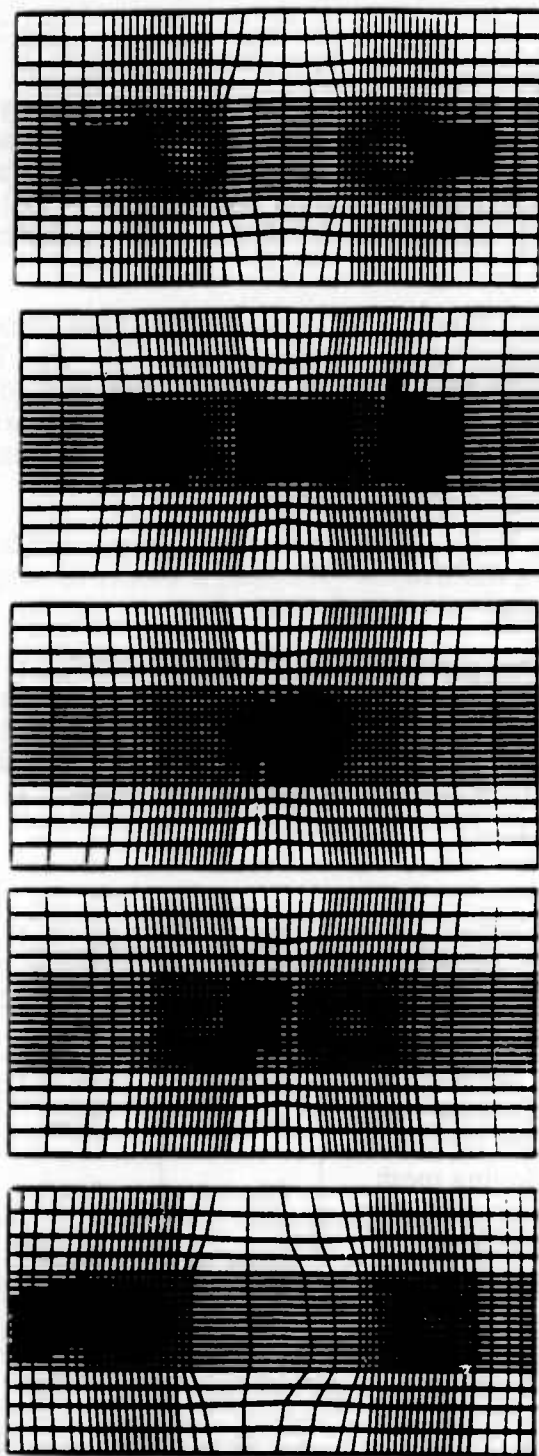
Figure 9. Grids created for Solution 4 of Example 2 at $t = 0, 0.23, 0.46, 0.92,$ and 1.2 (top to bottom).

$$\rho = 1.4, \quad p = 1.0, \quad e = 2.5, \quad u = 0, \quad v = 0,$$
$$\text{if } y \geq \sqrt{3}(x - 1/6). \tag{12b}$$

Along the left boundary ($x = -0.3$) and the bottom boundary to the left of the wedge ($y = 0$, $-0.3 \leq x \leq 1/6$), we prescribe Dirichlet boundary conditions according to (12); along the top boundary ($y = 1$), values are prescribed that describe the exact motion of an undisturbed Mach 10 shock; along the right boundary ($x = 3.4$), all normal derivatives are set to zero; and along the wedge ($y = 0$, $1/6 \leq x \leq 3.4$) reflecting boundary conditions are used.

The solution of this problem is a complete self-similar structure called a double-Mach reflection that was described in Ben-Dor and Glass [12, 13]. Two reflected Mach shocks form with their associated Mach stems and contact discontinuities. The geometry of these structures are very fine and are primarily confined to a small region that moves along the wedge with the incident shock. One of the two contact discontinuities is so weak that it is usually not noticed in computations.

The MacCormack finite difference scheme needs artificial viscosity to "capture" shocks without excessive oscillations. We used a model developed by Davis [20] which is total variation diminishing in one space dimension.

Five solutions of this problem were calculated for $0 < t \leq 1.9$ as indicated in Table 3. Refinement was restricted to a maximum of two levels and a tolerance of 0.6 in the maximum norm was prescribed. A pointwise error indicator based on the assumption of smooth solutions, like the present one, is not appropriate for problems having discontinuities. Without restricting the maximum level of refinement, we could refine indefinitely in the vicinity of a discontinuity.

Ref. No.	Strategy	Base Mesh	Max. No. Nodes	CPU Time (sec.)
1	Stationary uniform mesh	63×29	1827	2130
2	Moving mesh	63×29	1827	2220
3	Stationary mesh with refinement	29×11	2782	3254
4	Moving mesh with refinement	29×11	3540	3725
5	Stationary uniform mesh	120×40	4800	6861

Table 3. Maximum number of nodes in any base time step and computational costs for five solutions of Example 3.

Solutions 2 through 5 were intended to be of comparable accuracy and we shall attempt to appraise the computational cost of each adaptive strategy. The maximum number of nodes that was introduced in any base time step and the total CPU time are

presented as measures of computational complexity in Table 3. Contours of the density at $t = 0.19$ are shown for all five solutions in Figure 10 and the grids that were generated for Solution 4 at $t = 0.038, 0.076, 0.114, 0.152,$ and 0.19 are shown in Figure 11.

As in the previous two examples, the mesh-moving strategy of Solution 2 does a great deal to improve the results of the static Solution 1 for approximately a five-percent increase in computational cost. Comparing the top two contours of Figure 10, we see that the resolution of the incident and reflected shocks is much finer with solution 2 than with Solution 1. Additional detail of the structures in the Mach stem region and of the contact discontinuities are present in Solution 2, but not in the nonadaptive Solution 1. Finally, Solutions 1 and 5 display more oscillatory behavior behind the incident shock near the upper boundary. This is undoubtedly due to our maintaining a discontinuity where the shock intersects the upper boundary.

The use of refinement on a stationary mesh again does not give the dramatic improvement obtained by mesh moving (cf. the second and third contours of Figure 10). Initally the fine meshes were following the incident and reflected shock structures and better results were obtained; however, by $t = 0.19$ refinement is being performed over much of the domain and two levels of refinement are not sufficient for adequate resolution (cf. Arney and Flaherty [6]). The combination of mesh motion and refinement depicted by Solution 4 in Figure 10 provides a marked improvement in resolution. The sequence of meshes shown in Figure 11 shows that the coarse mesh is able to follow the differing dynamic structures and that refinement is only performed in the vicinity of discontinuities. Initially, only one rectangular cluster is needed to follow the incident shock (cf. Arney and Flaherty [5]). As time progresses, two clusters are created in order to follow the incident and reflected shocks (cf. the upper three meshes of Figure 11). A third cluster is created as time increases further in order to follow the evolving activity in the region of the Mach stem (cf. the lower two meshes of Figure 11).

Severe distortion of the mesh in the reflected shock region caused a static mesh regeneration to occur for Solution 4 at $t = 0.162$. The base meshes before and after the static regeneration are shown in Figure 12. Thus, Solution 4 demonstrates all of the capabilities of our adaptive procedure. The results presented in Table 3 and Figure 10 also show that Solution 4 provided greater resolution than the uniform mesh Solution 5 for approximately one-half of the cost. Solution 4 also shows many of the same characteristics as the solution computed by Woodward and Collela [32] using MacCormack's method on a 240×120 uniform grid. We were unable to compute a solution on such a fine mesh due to virtual memory limitations on our computer; however, we estimate that it would have used 14,400 nodes and 40,000 CPU seconds.

The results presented for this problem demonstrate the power and efficiency of our adaptive techniques; however, we would have preferred to allow more than two levels of refinement and a finer base mesh. These calculations would have produced better resolution of the discontinuities and other fine-scale structures that further demonstrate the computational advantages of adaptive methods relative to uniform mesh techniques. As noted, restrictions of our computing environment prevented us from doing this in a reasonable manner. We hope to perform these calculations in the future using a larger computing system.

IV. DISCUSSION OF RESULTS AND CONCLUSIONS. We have described an adaptive procedure for solving systems of time-dependent partial differential equations in two-space dimensions that combines existing mesh moving [5] and local refinement [6] techniques. The algorithm also contains procedures for initial mesh generation and static mesh regeneration. It can be used with a wide variety of finite difference or finite element schemes and error indicators.
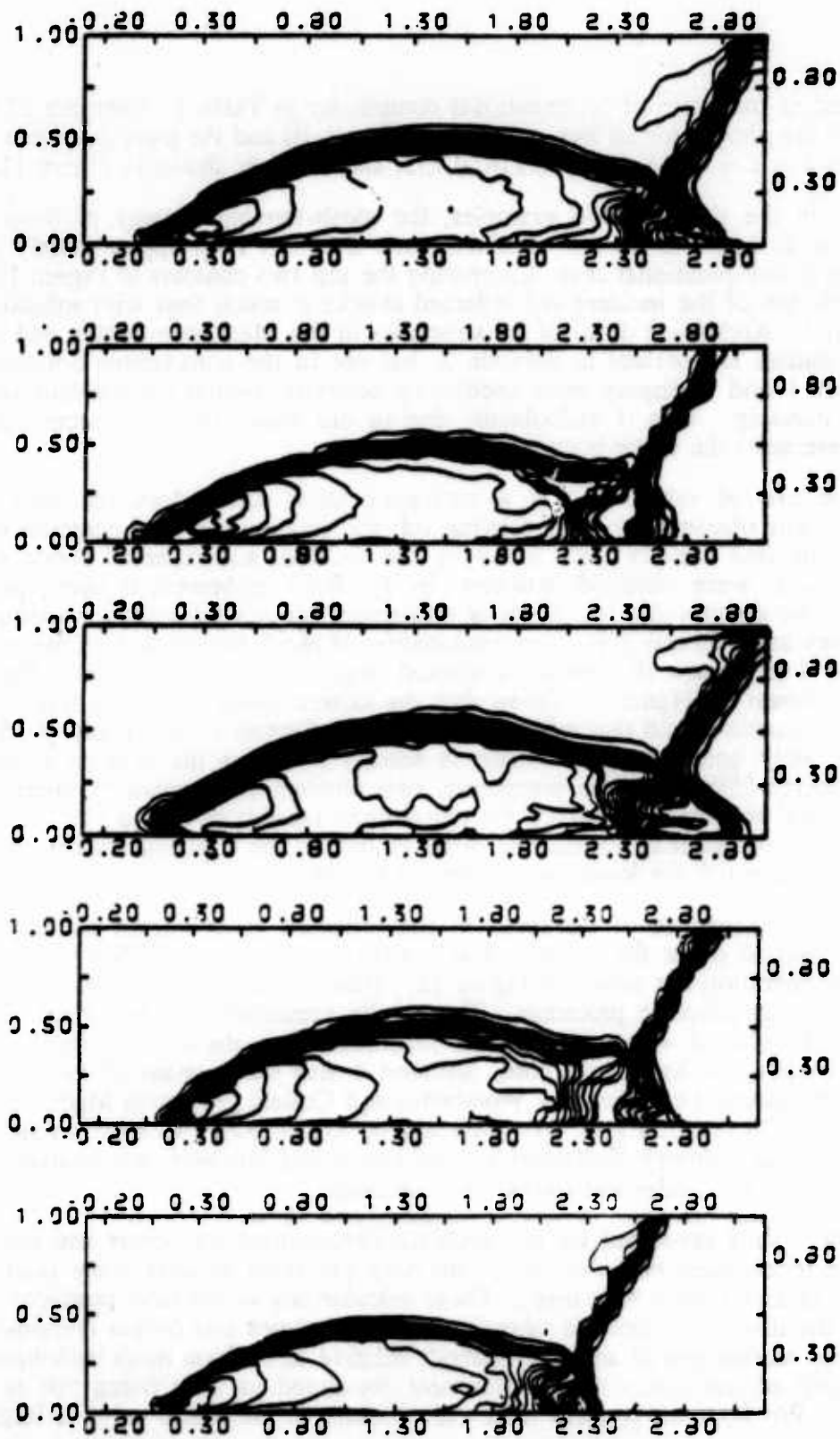
Figure 10. Contours of the density at $t = 0.19$ for Solutions 1 to 5 (top to bottom) of Example 3.
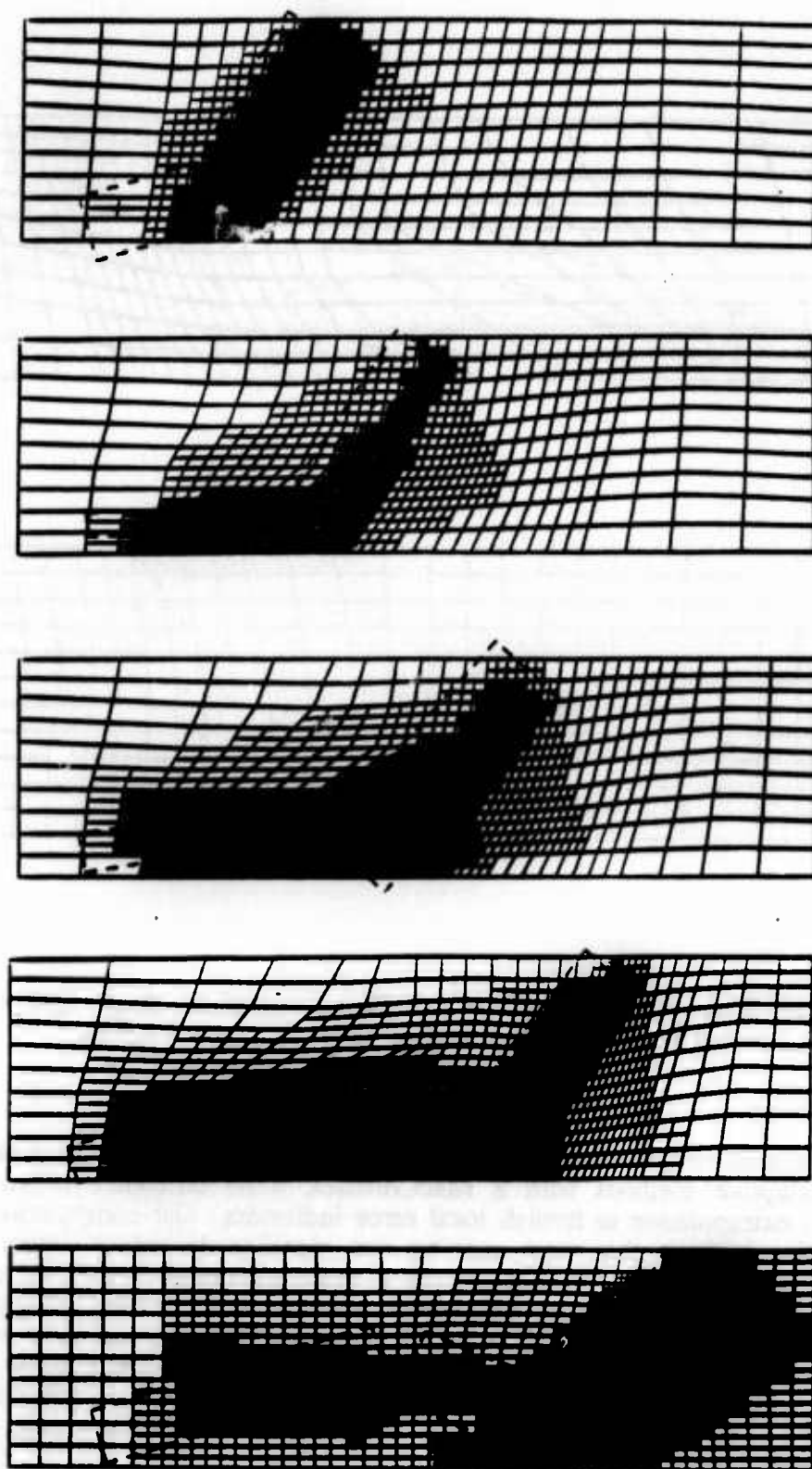
Figure 11. Grids created for Solution 4 of Example 3 at t = .038, 0.076, 0.114, 0.152, and 0.19 (top to bottom).
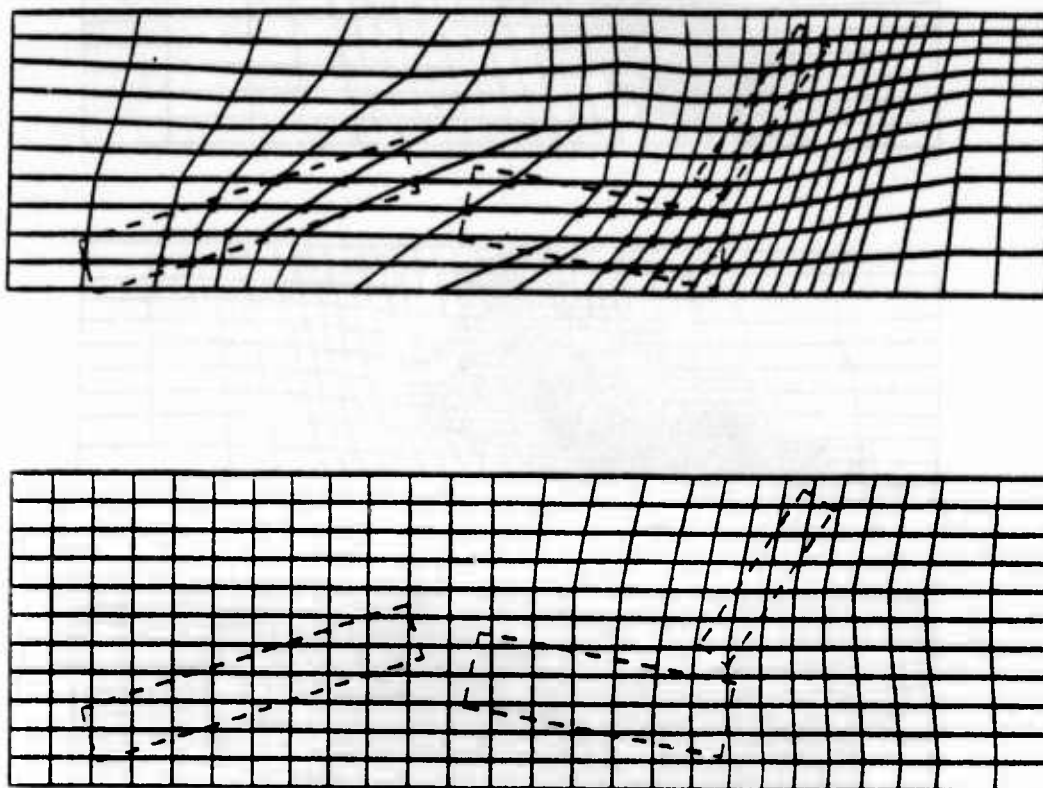
Figure 12. Base grids before (top) and after (bottom) the static mesh regeneration that was performed for Solution 4 of Example 3 at $t = 0.162$.

We obtained computational results for hyperbolic systems of conservation laws by using our adaptive methods with a MacCormack finite difference scheme and using Richardson's extrapolation to furnish local error indicators. Our computational results on three examples indicate that mesh moving can significantly reduce errors for approximately a ten-percent increase in cost relative to computations performed on stationary uniform meshes. The use of local refinement without mesh moving provided increased efficiency relative to uniform-mesh calculations, although not as dramatic as those found using mesh moving. The combination of mesh moving and local refinement provided reliable results while costing significantly less than stationary-mesh calculations. Thus, the overhead associated with the dynamic data structures is less than the time to calculate a comparable solution on a uniform mesh.

The results of Section III and others (cf. Arney and Flaherty [5, 6]) indicate that our mesh moving procedures perform better alone than with refinement. This is because the projection of fine-mesh solutions onto coarser meshes reduces the errors at base mesh nodes and mesh motion based on controlling small or zero local discretization errors either fails or results in no movement. Erratic mesh motion can also occur with some techniques when movement indicators are small. This topic is discussed in Coyle et al. [19] and a

possible remedy for one-dimensional problems is suggested in Adjerid and Flaherty [2]. Further experimentation and analysis are being performed in order to determine the best way to combine mesh moving and refinement.

There are several other ways to improve the efficiency, reliability, and robustness of our adaptive methods. The present Richardson's extrapolation based error indicator is expensive and we are seeking ways of replacing it by techniques using p-refinement. Such methods have been shown [1, 2, 3, 8, 10, 16, 22] to have an excellent cost performance ratio when used in conjunction with finite element methods. An appropriate error indicator or estimator can be used to control a differential refinement algorithm, where different refinement factors (i.e., other than binary) are used in different high-error clusters. If the error indicator is capable of providing separate estimates of the spatial and temporal errors, as the present one does, then different refinement factors can also be used in space and time. We also hope to demonstrate the flexibility of our refinement procedure by using it with a finite difference or finite element scheme for parabolic problems.

The greater reliability and efficiency of adaptive techniques will be most beneficial in three dimensions. These techniques must be able to take advantage of the latest advances in vector and parallel computing hardware. The tree is a highly parallel structure and we have been developing solution procedures that exploit this in a variety of parallel computing environments.

REFERENCES

1. S. Adjerid and J. E. Flaherty, A moving finite element method with error estimation and refinement for one-dimensional time dependent partial differential equations, *SIAM J. Numer. Anal.*, 23 (1986), to appear.

2. S. Adjerid and J. E. Flaherty, A moving mesh finite element method with local refinement for parabolic partial differential equations, *Comp. Meths. Appl. Mech. Engr.*, 56 (1986), pp. 3-26.

3. S. Adjerid and J. E. Flaherty, A local refinement finite element method for two-dimensional parabolic systems, Tech. Rep. No. 86-7, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, 1986.

4. D.C. Arney, *An Adaptive Mesh Algorithm for Solving Systems of Time-Dependent Partial Differential Equations*, Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, 1985.

5. D.C. Arney and J.E. Flaherty, A two-dimensional mesh moving technique for time dependent partial differential equations, Tech. Rep. No. 85-9, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, 1985. Also *J. Comput. Phys.*, to appear.

6. D.C. Arney and J.E. Flaherty, An adaptive local mesh refinement method for time-dependent partial differential equations, Tech. Rep. No. 86-10, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, 1986.

7. I. Babuska, J. Chandra, and J.E. Flaherty, Eds., *Adaptive Computational Methods for Partial Differential Equations*, SIAM, Philadelphia, 1983.

8. I. Babuska, A. Miller, and M. Vogelius, Adaptive methods and error estimation for elliptic problems of structural mechanics, in *Adaptive Computational Methods for*

Partial Differential Equations, I. Babuska, J. Chandra, J. E. Flaherty, Eds., SIAM, Philadelphia, 1983, pp. 57-73.

9. 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.

10. R.E. Bank and A. Weiser, Some a posteriori error estimates for elliptic partial differential equations, *Maths. Comp.*, 44 (1985), pp. 283-301.

11. J.B. Bell and G.R. Shubin, An adaptive grid finite difference method for conservation laws, *J. Comput. Phys.*, 52 (1983), pp. 569-591.

12. G. Ben-Dor and I.I. Glass, Non-stationary oblique shock-wave reflections: Actual iso-pycnics and numerical experiments, *AIAA J.*, 16 (1978), pp. 1146-1153.

13. G. Ben-Dor and I.I. Glass, Domains and boundaries of non-stationary oblique shock-wave reflections. 1. Diatomic gas, *J. Fluid Mech.*, 92 (1979), pp. 459-496.

14. M. Berger, On conservation at grid interfaces, ICASE Rep. No. 84-43, ICASE, NASA Langley Research Center, Hampton, 1984.

15. M. Berger and J. Oliger, Adaptive mesh refinement for hyperbolic partial differential equations, *J. Comput. Phys.*, 53 (1984), pp. 484-512.

16. 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.

17. J.U. Brackbill and J.S. Saltzman, Adaptive zoning for singular problems in two dimensions, *J. Comput. Phys.*, 46 (1982), pp. 342-368.

18. S.R. Chakravarthy and S. Osher, Computing with high-resolution upwind schemes for hyperbolic equations, in *Large-Scale Computations in Fluid Mechanics*, B.E. Engquist, S. Osher, and R.C.J. Somerville, Eds., Lectures in Applied Mathematics, 22-1, AMS, Providence, 1985, pp. 57-86.

19. J.M. Coyle, J.E. Flaherty, and R. Ludwig, On the stability of mesh equidistribution strategies for time-dependent partial differential equations, *J. Comput. Phys.*, 62 (1986), pp. 26-39.

20 S. Davis, TVD finite difference schemes and artificial viscosity, ICASE Rep. No. 84-20, NASA CR No. 172373, ICASE, NASA Langley Research Center, Hampton, 1984.

21. S.F. Davis and J.E. Flaherty, An adaptive finite element method for initial-boundary value problems for partial differential equations, *SIAM J. Sci. Stat. Comput.*, 3 (1982), pp. 6-27.

22. M. R. Dorr, The approximation theory for the p-version of the finite element method, I, *SIAM J. Numer. Anal.*, 21 (1984), pp. 1180-1207.

23. H. A. Dwyer, Grid adaption for problems with separation, cell Reynolds number, shock-boundary layer interaction, and accuracy, AIAA paper No. 83-0449, AIAA 21st Aerospace Sciences Meeting, Reno, 1983.

24. R.J. Gelinas, S.K. Doss, and K. Miller, The moving finite element method: applications to general partial differential equations with multiple large gradients, *J. Comput. Phys.*, 40 (1981), pp. 202-249.

25. A. Harten and J.M. Hyman, Self-adjusting grid methods for one-dimensional hyperbolic conservation laws, *J Comput. Phys.*, 50 (1983), pp. 235-269.

26. R. Hindman, Generalized coordinate forms of governing fluid equations and associated geometrically induced errors, *AIAA J.*, 20 (1982), pp. 1359-1367.

27. R.W. MacCormack, The effect of viscosity in hypervelocity impact cratering, AIAA Paper 69-354, 1969.

28. G. McRae, W. Goodin, and J. Seinfeld, Numerical solution of the atmospheric diffusion equation for chemically reacting flows, *J. Comput. Phys.*, 45 (1982), pp. 1-42.

29. M.M. Rai, Patched-grid calculations with the Euler and Navier-Stokes equations, SIAM National Meeting, Boston, July 1986.

30. M.M. Rai and D. Anderson, Grid evolution in time asymptotic problems, *J. Comput. Phys.*, 43 (1981), pp. 327-344.

31. J.F. Thompson, Ed., *Numerical Grid Generation*, North-Holland, New York, 1982.

32. P. Woodward and P. Collela, The numerical simulation of two-dimensional fluid flow with strong shocks, *J. Comput. Phys.*, 54 (1984), pp. 115-173.

VORTEX FISSION AND FUSION

Karl Gustafson
Dept. of Mathematics, University of Colorado
Boulder, Colorado USA

Abstract. Vortex fission-fusion sequences are found in full Navier-Stokes flow at higher Reynold's numbers and higher aspect ratios. Some of these represent transient bifurcations whereas others persist, indicating Hopf bifurcations in the final states. Such sequences appear to be initiated by a sublayer viscous-inviscid bursting effect caused by a wall-eddy development near a separation point. Parity rules and the relative proximities of provocation points and obstructions play fundamental roles in the further evolution of the vortex shedding and coalescence dynamics. As intensities drop off sharply in the secondary structures, numerical resolution considerations become paramount. For the latter a new multigrid localization procedure currently under development has proven to be remarkably robust.

1. INTRODUCTION

Recent studies (see Benjamin and Mullin [1], Cliff and Mullin [2], Bolstad and Keller [3], and the references therein) have been concerned with questions of flow multiplicity higher than previously expected in the Taylor Problem of flow between rotating cylinders. For the most part these studies consider steady cellular flows at Reynolds numbers reasonably near those at which the Taylor vortices appear. Quoting

> "[1, p. 219] A prime contention of the previous discussions has been that although the realistic hydrodynamic problem modelling the Taylor experiment is yet unsolved in closed form, it must have a high multiplicity of isolated solutions when R lies well above the quasi-critical range wherein Taylor cells are first easily demonstrable by standard flow-visualization techniques.

> "[2, p. 256] A striking feature of anomalous modes, particularly those with a larger number of cells, is the distortion of the cell boundary adjacent to the anomalous cell."

> "[3, p. 16] A new phenomenon is ... the splitting of the extra vortices into two smaller vortices."

Re [1], while admitting that I have only recently become aware of these recent new higher multiplicities found for the Taylor problem, nontheless I would first like to advance here the hypothesis that in some cases the end effects in the

Taylor Problem imply even higher multiplicities that just haven't been found yet, in some situations infinite multiplicities. All of this depends on the exact experimental or numerical model employed, but when a corner with no slip conditions prevailing is encountered, or a corner with slip conditions on one side only, e.g., an intersurface separation interface, and when the angle is not too large, one should expect an infinite set of smaller vortices descending into it. An example of a sequence of ten of these that we have found in a corner will be given below.

A second thought I advance here is that the existence of higher multiplicities in real flow depends more on certain "parity rules" established by the fluid during its actual dynamic evolution, than on the bifurcation parameter homotopy arguments followed in [1,2,3]. The latter "homotopy model" is a valuable technique in connection with the numerical continuation methods used in [1,2,3] to enable the tracking of the "full" bifurcation diagram as, say, the Reynolds number Re or the aspect ratio A is varied. But in the end it would appear to be limited to the analyses of the steady flow equations and can therefore generate mathematically valid but physically spurious solutions. I will illustrate below the development of such a "parity rule" structure governing a full Navier-Stokes flow. Moreover, as will be seen, the parity rules explain the cell boundary distortion referred to in [2].

Finally, I will illustrate the mechanisms of the splitting of vortices into smaller vortices. This can occur [6] as a function of the varying of the key parameters (e.g., Re, A) of the problem in a steady flow as in [3] but more interestingly is found to occur dynamically [6,7] in unsteady flow, with both splitting and coalescence sequences found.

2. END EFFECTS AND CORNER VORTEX SEQUENCES

As pointed out in [2, p. 257], the anomalous modes are not surprising, should be expected, and are due to the end effects on the Taylor annulus.

In [4,5] we concentrated on finding similar "anomalous modes" for corner flow in a driven unit cavity, and thus far have succeeded in finding twenty of them. There are (mathematically) an infinite number there, although (computationally) they will, depending on the precision carried, drop into the noise level because their intensities fall off $O(10^{-4})$, and (physically) experimentally only three or four at most have been seen. For full details about this interesting problem see [8,9]. Here are the first 10 corner modes reported in [4], measured both by stream function intensity ψ_i and in terms of the zeros z_i between them on the 45° diagonal angle bisector extending out from the lower left

1144

corner of the cavity. There are such vortex sequences in other cavity corners but we will omit discussion of those here. The sign changes on the ψ_i intensities are in accordance with the parity rules I will discuss next.

Local Maximum Stream Function Intensity	Stream Function Zero Measured Along Diagonal
1.0006×10^{-1}	6.97×10^{-2}
-2.232×10^{-1}	4.205×10^{-3}
6.165×10^{-11}	2.534×10^{-4}
-1.703×10^{-15}	1.536×10^{-5}
4.71×10^{-20}	9.247×10^{-7}
-1.30×10^{-24}	5.602×10^{-8}
3.59×10^{-29}	3.370×10^{-9}
-9.93×10^{-34}	2.040×10^{-10}
2.75×10^{-38}	1.236×10^{-11}
-7.59×10^{-43}	7.421×10^{-13}

3. PARITY RULES AND PROXIMITY LIMITATIONS

As pointed out in [3, p. 4], the demonstration of additional "hidden" vortices remove all difficulties with "wrong" odd numbers of vortices found in previous experiments.

Such "hidden" very weak vortices are known in the aerodynamic literature as "ornamentation" vortices. I preferred the term "intermediating" vortices in [8] to indicate that they are not ornamental in any sense of the word but are in fact topological necessities to the flow. Aerodynamics is characterized by open regions and often the smaller vortices do indeed flow away, but in a closed flow such as the Taylor geometry of [1,2,3] or in the cavity geometry of [6,7] they do not disappear once they have managed to enter the flow. Whether or not they can enter appears to depend not only on their parity but also on the proximity of their potential development region to ends, corners, walls, and even to intersurface separation lines. Here are some details of their evolution as reported in [6,7]. Note that the deformed cell contours occur very naturally in terms of the parity signs plotting along their boundaries.
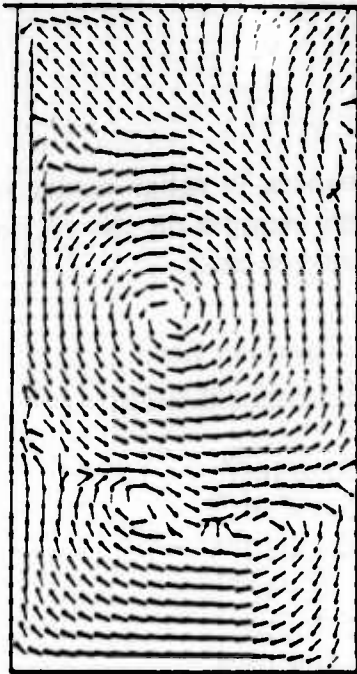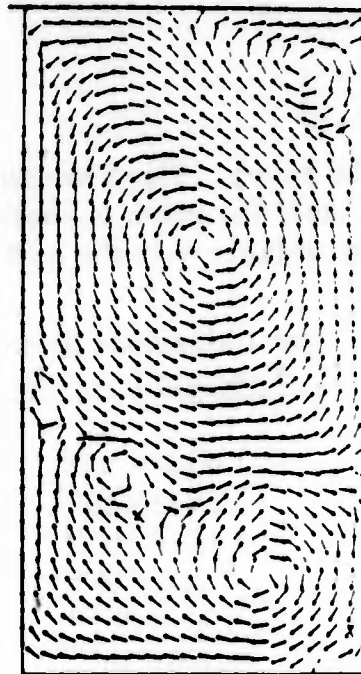
(a) t = 35 seconds (b) t = 50 seconds
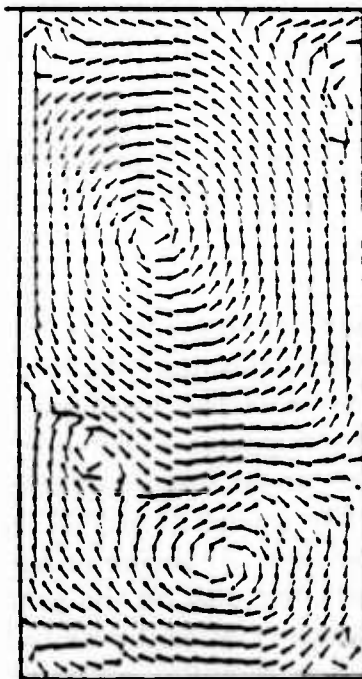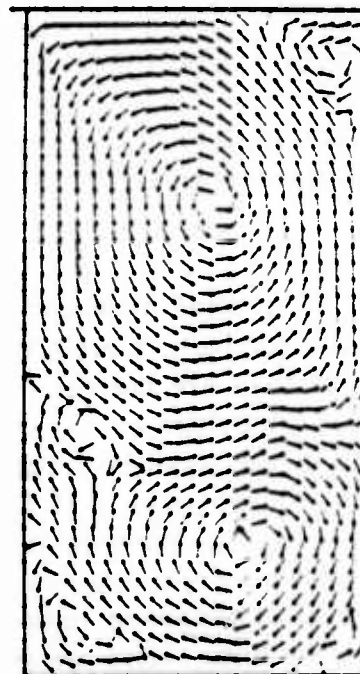
(c) t = 60 seconds (d) t = 75 seconds

Figure 1. Separation Region "Lubrication" Dynamics

(a) Fission into 3, almost 4, tertiary eddies for "self-lubrication".
(b) Right "corner lubrication" begins, to continue the energy cascade.
(c) The last two eddies report in, causing "temporary mass confusion".
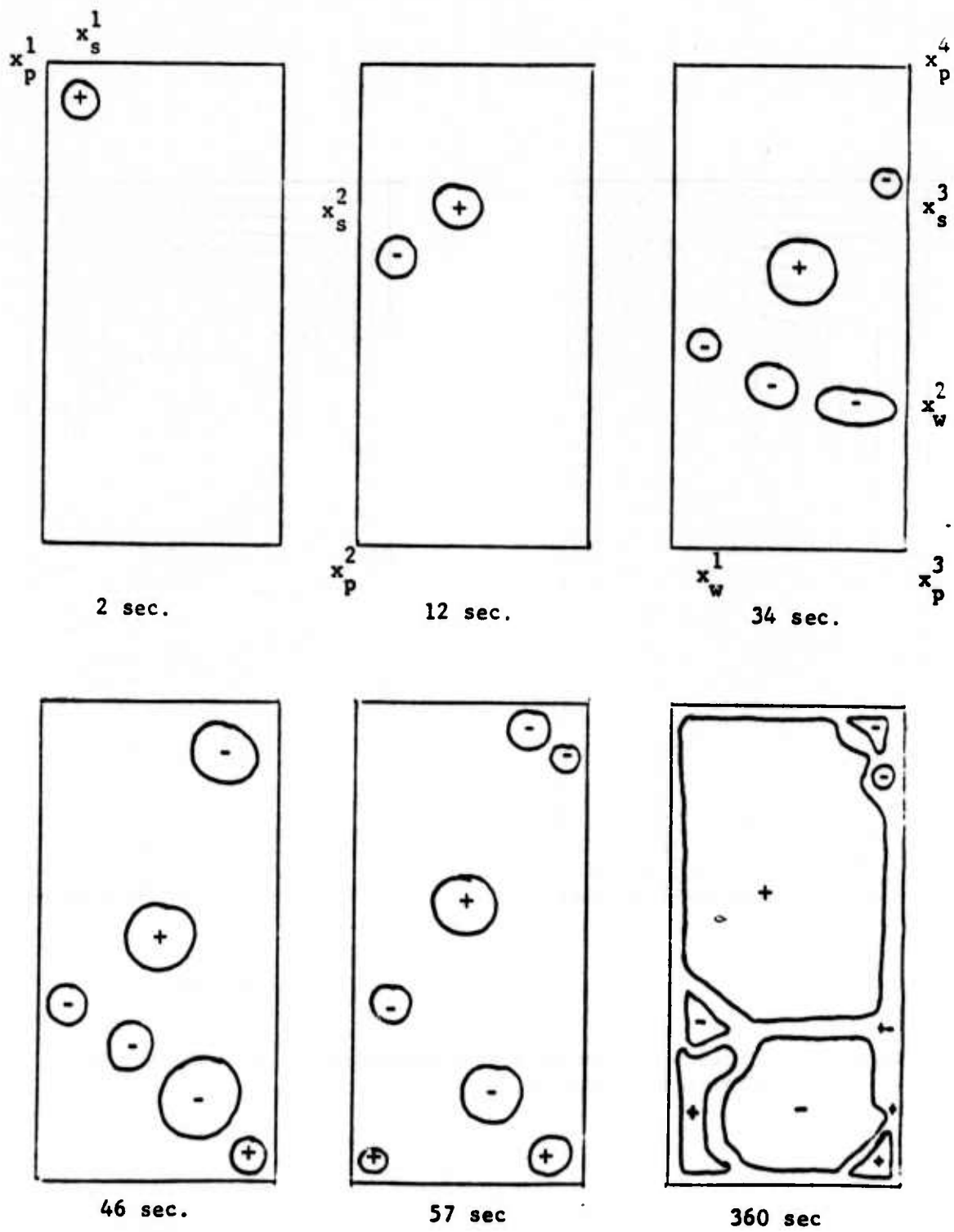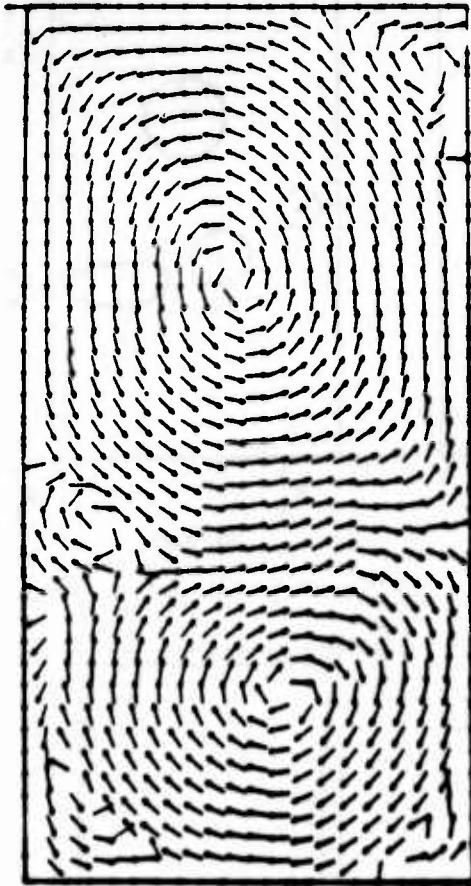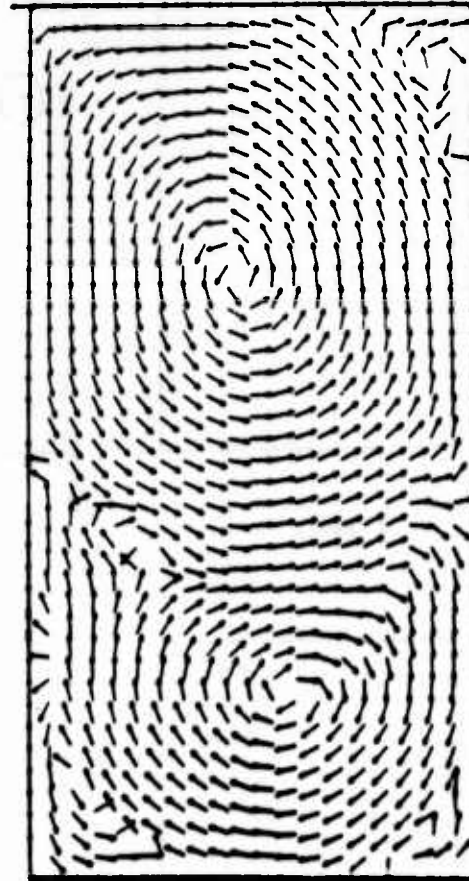(d) "Final Resting Place", as the basic final flow topology is determined.

1146

Figure 2. Wall, corner, provocation, separating, and intermediating effects in cavity flow dynamics.

359 sec.   360 sec.

Figure 3.   Final wavyness of vortex dynamics at   Re = 10000   in a cavity of depth   2.

4. INTERMITTENT BIFURCATION AND HOPF BIFURCATIONS

As can be seen from the above, Figures 1, 2, 3, for full Navier-Stokes flow at Re = 10,000 and in a depth 2 cavity, there are many intermittent bifurcations (e.g., wall bursting, the splitting-coalescence sequences) that appear during a dynamic development of a flow. These must be taken into account in any attempt at any physically correct understanding of any final steady flow bifurcation diagram. Some of these sequences are transient, thus "ornamental" in one sense, yet their temporary existence is absolutely essential in the "mediating" of the development of a final flow pattern.

The full flow history (see [6,7]) of the Re = 10,000 cavity dynamics indicates that the repeating pattern at the left midwall represents a final periodic solution. Our results thus imply the existence of a Hopf bifurcation, for flow in a depth 2 cavity, at some critical Reynolds number Re strictly between 2000 and 10,000. One should distinguish this Hopf bifurcation of the discretized equations from a claim for the cases of the continuous equations and the actual physics. Moreover, the aspect ratio A = depth/width enters as a second bifurcation parameter of considerable importance. That is, holding Re = 10,000 as studied in [7], from [6,7] there is indicated a Hopf bifurcation at some critical A strictly between 1 and 2. A recent study [10] found no periodic solution for flow at Re = 10,000 in a unit (A = 1) cavity.

In addition to the pronounced oscillation on the left, we noted also very small tertiary nonstationarities at three points along the right wall: (i) just below the upper right corner, where there earlier was a definite tertiary eddy; (ii) just below the right midwall, where the fluid separates into the upper and lower regions; (iii) at the top of the lower right corner eddy, see Figure 3. Whether or not these represent minor numerical or fluid instabilities or very small tertiary features of a periodic final solution is a further interesting question. One would imagine a first Hopf bifurcation at some Re_1 in which the flow would settle into an oscillation tracking the principal separation point movement on the left wall, and then higher critical Re_2, Re_3, Re_4, at which one or more of the just mentioned three tertiary features may maintain itself into the final state.

5. SOME NUMERICAL CONSIDERATIONS

Due to the multidirectional nature of the unsteady flow dynamics in a cavity, any incorporation of upwinding could result in a computational quagmire and would likely distort the unsteady flow transient details. On the other hand, in employing a forward Euler-MAC scheme as we did in [6,7], care is needed in

choosing the discrete steps δt, δx, and δy to assure both stable time integration and adequate spatial resolution of the dynamics under consideration. We required that the ratio $4(\delta t)\nu/(\delta x)^2$ not exceed a critical value $K(\nu)$, where we have taken $\delta x = \delta y$. Because the equations are nonlinear, $K(\nu)$ must in general be determined experimentally by simulation runs on a coarse grid. For example the value $\nu = 0.01$ yielded $K(\nu) \approx 1.0$, while $\nu = 0.0025$ yielded $K(\nu) \approx 0.36$. Our experiments with $K(\nu)$ indicated that $K(\nu) \rightarrow 0$ as $\nu \rightarrow 0$, in a nonlinear way. Instability, when it occurred, manifested itself first in the downstream left lid corner, resulting quickly in a subsequent dissolution of the primary vortex accompanied by a rapid buildup of large pressure gradients. A well established primary vortex appeared to guarantee stability thereafter.

Proper resolution of boundary layer effects at high Reynolds numbers requires in general very fine discretizations. However, the flow velocities near the midwall and corner vortex structures are considerably smaller than the driving velocity. For $Re = \dfrac{1}{\nu} = 10{,}000$, at the final time $t = 360$ seconds the effective local Reynold's number in the left midwall region was found to be about

$$(Re)_{effective} \sim \frac{U_0^{local} L}{\nu} \lesssim \frac{(0.1)(1)}{(10^{-4})} \sim 1000 \, .$$

Examination of the velocity matrices during the whole time history revealed corresponding or smaller values of $(Re)_{effective}$ in the neighborhoods of all secondary vortex structures. In particular, the transient wall vortex separation dynamics, i.e., bursting and separation point movement on the left wall, would appear to be accurately resolved.

To obtain the very high resolution steady Stokes flow corner subvortices (20 of them thus far) found in [4,5], care to avoid underflow and to maintain good subdomain residuals required several innovations on the FAC multigrid schemes. Roughly, the additional considerations arise in accurately passing information back and forth from local to global grids when treating a problem which is biharmonic rather than harmonic. Challenging and interesting investigations of several proposed N-Processor (e.g., Hypercube configuration) algorithms for simultaneous subdomain computation present themselves in this context, although we have not yet launched such a study.

6. OTHER GEOMETRIES AND APPLICATIONS

In another paper presented at this conference, Ghoniem [11] reports similar studies by a different method on another model, that of flow over a backward facing step. From our discussions at this conference, it would appear that our

results and those of Ghoniem et al. [11,12,13] support and indeed reinforce the conclusions found separately in each of the two geometries.

Vortex splitting dynamics with properties very consistent with those that we have found have been recently observed physically in unsteady flow over airfoils [14]. Our methods could be applied to those and other geometries (e.g., corners of arbitrary angle, and effects, general obstructions) by means of new high resolution grid generation mapping schemes utilizing the same elliptic solver techniques described herein. Other future work includes a better understanding of the mechanisms under which higher multiplicities and cell boundary distortions develop in a flow, and in particular how such dynamical features depend fundamentally on the new parity rules that we have reported here.

REFERENCES

1. T. Brooke Benjamin and T. Mullin, Notes on the multiplicity of flows in the Taylor experiment, *J. Fluid Mech.* 121 (1982), 219-230.

2. K. Cliffe and T. Mullin, A numerical and experimental study of anomalous modes in the Taylor experiment, *J. Fluid Mech.* 153 (1985), 243-258.

3. J. Bolstad and H. Keller, Computation of anomalous modes in the Taylor experiment, *J. Computational Physics,* to appear.

4. K. Gustafson and R. Leben, Multigrid computation of subvortices, *Applied Math. and Computation* (1986), to appear.

5. K. Gustafson and R. Leben, in preparation.

6. K. Gustafson and K. Halasi, Vortex dynamics of cavity flows, *J. Computational Physics* 64 (1986), June.

7. K. Gustafson and K. Halasi, Cavity flow dynamics at higher Reynolds number and higher aspect ratio, *J. Computational Physics,* to appear.

8. K. Gustafson, Vortex separation and fine structure dynamics, *Applied Numerical Mathematics* (1986), to appear.

9. K. Gustafson, *Partial Differential Equations,* 2nd Edition, Wiley, New York, to appear.

10. P. Gresho, S. Chan, R. Lee, C. Upson, *International J. Num. Meth. Fluids* 4 (1984), 619.

11. A. Ghoniem, Computing unsteady reacting flows using vortex methods, these *Proceedings.*

12. A. Ghoniem and Y. Gagnon, Vortex simulation of laminar recirculating flows, *J. Computational Physics*, to appear.

13. A. Ghoniem and J. Sethian, Dynamics of turbulent structure in a recirculating flow: a computational study, AIAA-85-0146, *AIAA 23rd Aerospace Sciences Meeting*, January 1985.

14. P. Freymuth, W. Bank, and M. Palmer, Vortices around airfoils, *American Scientist* 72 (1984), 242-248.

INCIPIENT SINGULARITIES IN THE NAVIER-STOKES EQUATIONS

Alain Pumir & Eric D. Sigga

Laboratory of Atomic and Solid States Physics

Cornell University

Ithaca, NY 14853

In this paper, we examine one of the challenging phenomena in
3-dimensional incompressible fluid dynamics, namely the streching of the
theoretical difficulties. Mathematically, it has proven to be impossible
so far to show that the solutions of the 3-dimensional Navier-Stokes
equations do not blow-up when the viscosity is very small, even in the
absence of any external forces. In this respect, the 2-dimensional
situation is simpler, due precisely to the absence of any vorticity
streching. We report here several results, which suggest that solutions
to the fluid equations can get close to a finite time singularity. Our
constructions proceed as follows. In a first part, we study several
models for the evolution of vortex filaments, in an inviscid fluid. The
results suggest that vorticity blows-up in a finite time. We then try
to reconstruct from the filaments solutions a solution of the Euler equations,
by using asymptotic techniques. Finally, we consider the role of viscosity.
It is argued that viscosity is barely able to prevent the collapse.

A reasonable model for the evolution of a slender vortex filament
describes a vortex tube by a curve, with an internal degree of freedom,
the core, size. The evolution equation reads

(1) $$\frac{dr(\theta,t)}{dt} = \frac{\Gamma}{4\pi} \int \frac{\frac{dr(\theta',t)}{d\theta'} \times (r(\theta,t) - r(\theta',t)) \, d\theta'}{\{(r(\theta) - r(\theta'))^2 + \sigma(\theta)^2 + \sigma(\theta')^2\}^{3/2}}$$

(2) $$|\sigma^2 ds/d\theta| = cst$$

Where σ measures the core-size of the slender tube, and acts as a cutoff. Equation (1) can be rigorously deduced in the limit of an infinitely thin tube (namely, $\sigma \ll r_c$, where r_c is the radius of curvature of the filament). In what follows, it will be referred to as the Biot-Savart equation. θ denotes a Lagrangian parameter, and s is the arclength along the curve. The equation for the core-size, (2), insures the local conservation of volume of the vortex tube. We emphasize that it is an approximation, applying when the time scale of the evolution of the vortex filament is large compared with the time scale of the internal dynamics of the core. With these assumptions, the vorticity scale is $1/\sigma^2$.

Starting from a simple, non planar curve, a pairing occurs and leads to a smooth curve composed of two anti-parallel filaments[1]. This new curve stretches very rapidly and generates smaller and smaller scales. Typically, two anti-parallel arcs of filament grow like two anti-parallel expanding circles, until an instability develops, and folds the curve into smaller pieces. These smaller pieces start stretching, and the process keeps repeating itself. The radius of curvature is always significantly larger than the core size, and the distance between the two filaments remains of the same order as the core-size. The minimum of the square of the core size decreases according a linear law: $\sigma^2 \simeq (t^*-t)$, suggesting a

1154

divergence of the vorticity like $1/(t^*-t)$ (with possibly logarithmic corrections)[2]. This result also follows from naive dimensional analysis of the vorticity equation. Various local approximations to the Biot-Savart formula have been derived and suggest that the singularity found for Biot-Savart is very robust, at least within a filament approximation.

The fact that, in all numerical simulations, the ratio $\varepsilon=\sigma/r_c$ is small allows us to look for solutions of the Euler equations by an asymptotic expansion, in powers of ε. We observe that the two paired filaments look locally like a 2-dimensional dipole. Such dipoles are known to be very stable. The zeroth order solution in powers of ε is constructed with an exact dipole solution of the 2-D Euler equations. Various corrections come in at higher order. Our asymptotic formalism allows us to recover the equations resulting from local approximation to the Biot-Savart equations. However, we are unable to rule out a systematic deformation of the basic 2-D dipole, that would make the filament approximations inappropriate. Even so the reason why the core could be destroyed is rather subtle.

If the collapse does proceed at least partway, according to the Biot-Savart approximation for an inviscid fluid, it is not obvious that the viscous effects can prevent it[2]. The viscosity, which is commonly believed to wipe out any perturbation smaller than a dissipation scale can play a subtler role in this problem. A naive expression for the core size in the presence of viscosity is:

$$(3) \qquad d\sigma^2/dt = v-\sigma^2 d\ell n(ds/d\theta)/dt$$

The first term in the right-hand side of (3) represents fattening of the core due to viscous diffusion. The other term is due to the stretching; a crude estimate of it leads to: $\sigma^2 d\ell n(ds/d\theta)/dt \simeq \Gamma x f(\sigma,r_c)$, where f is a

dimensionless function of the characteristic lengths, σ and r_c, say. As ν and Γ have the same dimension and the ratios between the different lengths do not vary much, (3) can be approximated by $d\sigma^2/dt = \nu-\Gamma \times cst$. On the basis of this naive argument, the viscosity may not be able to prevent the collapse. A more rigorous analysis of the strained 2-dimensional Navier-Stokes equations confirms that the usual Laplacian dissipation is only marginally able to control the stretching.

A complementary numerical simulation of the full fluid equations allows one to identify different ways the core can be destroyed. When ε is not small enough shocks may appear on the tube, or the core can be deformed into thin sheets, that roll-up. Such mechanisms are the most likely impediments to the collapse[3].

In any case we have worked out a situation where some strong stretching effects do occur. Whether or not the stretching leads to a singularity has not been proven by our study. The relevance of our particular flow to the experimental situation is rather unclear. This study, however, sheds some light on an important feature of 3-dimensional, incompressible hydrodynamics.

References:

(1) E.D. Siggia, Phys. Fluids 28, 794(1985)

(2) E.D. Siggia and A. Pumir, Phys. Rev. Letters 55, 1749(1985)

(3) in preparation

Finite element approximation of a

reaction-diffusion equation

Part II: Approximation of the spontaneous bifurcation

and error estimates uniform in time.

Sat Nam S. Khalsa

Department of Mathematics
Iowa State University
Ames, Iowa   50011

Abstract.   The initial-boundary value problem for a reaction-diffusion equation

(*) $u_t = u_{xx} -G(L,u)$, $0 < x < 1$, $u(0,t) = u(1,t) = 0$, $t > 0$,

$$G(L,u) = 4L^2 u(u-b)(u-1), \quad 0 < b < 1/2,$$

was recently analyzed by Conley and Smoller. We study the large time behavior of the semidiscrete finite element approximations, with interpolation of the coefficients in the nonlinear terms. In Part I we have established that the properties of the semidiscrete approximations are completely analogous to those of the solutions of (*), and the asymptotic, as $t \to \infty$, optimal order convergence has been proved. In this paper we approximate the "spontaneous bifurcation (with L as a parameter) for the steady-state problem. For the semi-discrete approximations of (*) we establish error estimates that hold uniformly on the infinite time interval $[t_0, \infty)$, $0 > t_0$, for nonsmooth or incompatible initial data.

1. Introduction.

We are interested in numerical analysis of the nonlinear reaction-diffusion equation

$$u_t = u_{xx} - G(L,u), \quad 0 < x < 1, \ t > 0, \qquad (1.1.a)$$

$$u(x,0) = u^0(x), \quad 0 < x < 1, \qquad (1.1.b)$$

$$u(0,t) = u(1,t) = 0, \quad t > 0, \qquad (1.1.c)$$

which is a paradigm example of nonlinear phenomena such as multiple steady state solutions and bifurcation in mathematical biology.

The corresponding steady-state problem is

$$u_{xx} - G(L,u) = 0, \quad 0 < x < 1, \qquad (1.2.a)$$

$$u(0) = u(1) = 0. \qquad (1.2.b)$$

Here

$$G(L,u) = 4L^2 u(u-b)(u-1), \quad 0 < b < 1/2. \qquad (1.3)$$

The continuous problem was analyzed by Conley and Smoller and [2] and Smoller [10]. Their result is

Theorem 1.1 [10, Theorem 24.13]. Let G be defined by (1.3), and let $L > L_0$. Then there are exactly three steady-state solutions $u_i \in C^\infty$, $i = 0, 1, 2$ of (1.1.a), (1.1.c):
$0 \equiv u_0(x) < u_1(x) < u_2(x) < 1$, $|x| < 1$; u_0 and u_2 are attractors for (1.1.a), (1.1.c); and the linearized operators Q_0 and Q_2, where

$$Q_k = d^2/dx^2 + g_k, \quad g_k(x) = G'(u_k(x)), \quad k = 0, 1, 2,$$

1158

together with the boundary conditions (1.2.b), have only negative

eigenvalues. Q_1 has precisely one positive eigenvalue, and u_1 has a 1-

dimensional unstable manif-ld which consists of orbits connecting u_1 to each

of the other rest points. Initial data $u(x,0)$ which satisfies

$u_1(x) < u(x,0) < u_2(x)$ (resp. $0 < u(x,0) < u_1(x)$) on $|x| < 1$ is in the

stable manifold of u_2 (resp. 0).

The problem (1.1) is discretized in space [4] by the finite element

method with the interpolation of coefficients for the evaluation of nonlinear

terms, the so called "product approximation".

For $I = [0,1]$ let $H^s = H^s(I)$, $H_0^s = H_0^1 \cap H^s$, for s real, and

$L^\infty = L^\infty(I)$ be the usual Sobolev spaces with the norms $|\cdot|_s$ and

$|\cdot|_{0,\infty}$, respectively; if $s = 0$, we write $|\cdot|_0 = |\cdot|$.

Let $\Delta = \{0 = \bar{x}_0 < \bar{x}_1 < \ldots \bar{x}_{N+1} = 1\}$ be a partition of

$I = [0,1]$. Set $I_j = [\bar{x}_{j-1}, \bar{x}_j]$, $h_j = \bar{x}_j - \bar{x}_{j-1}$, $j = 1, \ldots, N + 1$ and

$h = \max\limits_{1 \leq j \leq N+1} h_j$. It is assumed that as the meshes vary they are quasi-uniform,

i.e. $\max\limits_{1 \leq j \leq N+1} h\, h_j^{-1} < \sigma$, for some $\sigma > 1$. Define the finite element space S_0^h

by

$S_0^h = M_0^k(\Delta) \equiv \{v \in C^0(I) : v(0) = v(1) = 0, v \in P_k(I_j), j = 1, \ldots, N+1\}$,

where for any interval $E \subseteq I$, $P_k(E)$ denotes the space of polynomials of

degree $\leq k$ restricted to E. Every interval $I_j = 1, \ldots, N + 1$ is divided

into k subintervals $[\bar{x}_{j-1+(i-1)/k}, \bar{x}_{j-1+i/k}]$, $i = 1, \ldots, k$, where the

nodes within each I_j are chosen to be the Gauss-Lobatto points. We relabel

$\bar{x}_{j+i/k}$ as x_{kj+i}, $j = 0, 1, \ldots, N$ and set $M = (N+1)k-1$.

Let $Q_h : g \in H_0^1 \rightarrow Q_h g \in S_0^h$ be the usual interpolation operator defined

by

$$(Q_h g)(x_i) = g(x_i), \quad i = 0, \ldots, M + 1.$$

Then the continuous time Galerkin approximation with the product approximation $u^h(x,t)$ to the solution of (1.1) is defined as a differentiable map $u^h : (0,\infty) \to S_0^h$ such that

$$(u_t^h, \chi) = -(u_x^h, \chi_x) - (Q_h G(L, u^h), \chi), \quad t > 0, \quad \chi \in S_0^h,$$

$$(1.4)$$

$$u^h(\cdot, 0) \quad \text{given in} \quad S_0^h.$$

The corresponding approximate steady-state problem is to find $u^h \in S_0^h$ from

$$(u_x^h, \chi_x) + (Q_h G(L, u^h), \chi) = 0, \quad \chi \in S_0^h. \quad (1.5)$$

If $\psi_i(x)$, $i = 1, \ldots, M$, is the usual interpolotory basis for S_0^h, then to compute the solutions of (1.4) and (1.5) one has to solve the problems

$$\sum_{i=1}^{M} \dot{u}_i (\psi_i, \psi_j) = -\sum_{i=1}^{M} u_i (\psi_i', \psi_j') - \sum_{i=1}^{M} G(L, u_i)(\psi_i, \psi_j), \quad j = 1, \ldots, M,$$

$$(1.6)$$

$$u_i(0), \quad i = 1, \ldots, M \quad \text{given.}$$

For $i = 0, 1, 2$ the linearized discrete eigenvalue problems are to find $v^h \in S_0^h$ from

$$-B_h^i(v^h, \chi) \equiv (v_x^h, \chi_x) + (a_h^i v^h, \chi) =$$

$$(1.7)$$

$$\lambda(h)(v^h, \chi), \quad a_h^i(x) = Q_h G_u(L, u_i^h(x)), \quad \chi \in S_0^h,$$

In [4] we have established the following result.

Theorem 1.2. Let G be defined by (1.3) and $L > L_0$. Then there exists

$h_0 > 0$ such that for $h < h_0$

(i) The approximate steady-state problem (1.5) has exactly three solutions $u_i^h \in S_0^h$, $i = 0, 1, 2$, satisfying $0 \equiv u_0^h < u_1^h < u_2^h < 1 + Ch^k$ and

$$\|u_i - u_i^h\|_1 < Ch^r \|u_i\|_{r+2}, \quad i = 1, 2, \quad r = 1, \ldots, k. \quad (1.8)$$

(ii) u_0^h and u_2^h are the attractors, and the linearized problem (1.7) has only negative eigenvalues for $i = 0, 2$; u_1^h has 1-dimensional unstable manifold which consists of orbits connecting u_1^h to each of the other rest points, and the linearized problem (1.7) has precisely one positive eigenvalue and the rest negative for $i = 1$.

(iii) Let $u(x,t)$ and $u^h(x,t)$ be the solutions of (1.1) and (1.4), respectively. Initial data which satisfies $u_1^h(x) < u^h(x,0)$ is in the stable manifold of u_2^h, and there holds:

$$\lim_{t \to \infty} \|u(\cdot,t) - u^h(\cdot,t)\|_1 < ch^r \|u_2\|_{r+2}, \quad r = 1, \ldots, k. \quad (1.9)$$

Initial data which satisfies $u^h(x,0) < u_1^h(x)$ is in the stable manifold of $u_0^h \equiv 0$, and there holds:

$$\lim_{t \to \infty} \|u^h(\cdot,t)\|_1 = 0. \quad (1.10)$$

In [10] the "spontaneous" bifurcation (i.e., the bifurcation whereby the solution suddenly "appears" when a parameter, in our case the length L of the interval, crosses a certain critical value) was analyzed for the steady-state problem (1.2). For the practical solution of our problem we can first find the stable steady-state solution u_2^h as a limit as $t \to \infty$ of an appropriate solution of the time dependent problem. We then obtain the unstable steady state solution u_1^h by following the bifurcation diagram.

The outline of the rest of the paper is as follows. In Section 2 we establish error estimates in the numerical approximation of the spontaneous bifurcation using general methods of approximation of nonlinear problems in Brezzi, Rappaz, and Raviart [1].

In Section 3 we establish large time error estimates in the semi-discrete solution extending the results in Khalsa [5] to take into account the numerical integration.

Recently in Manoranjan and Mitchell [9] numerical studies of the problem (1.1.a)-(1.1.c) were carried out, using the finite element discretization with the product appproximation, and estimates for the critical length L_0 were obtained. In Larsson [8] error estimates on an infinite time interval in the case of convex f were derived for a semilinear parabolic problem using piecewise linear finite elements. The results in [8] have been extended [5] to the case when the solution being approximated is asymptotically stable, instead of the convexity assumption on f, initial data is nonsmooth or incompatible, piecewise polynomial finite elements being used. See [5] for other references and discussion.

2. Approximation of the spontaneous bifurcation.

Following [10] we introduce a parameter

$$\beta = u(1/2),$$

where u is a solution of (1.2)

Lemma 2.1 [10, pp. 185-190] (i) There exists a unique branch $\{L(\beta), u(\beta); 0 < \beta < 1\}$ of solutions of (1.2) with

$$L''(\beta) < 0, \quad 0 < \beta < 1. \tag{2.1}$$

1162

(ii) There exists a unique $\beta_0 \in (0,1)$ with

$$L'(\beta_0) = 0. \tag{2.2}$$

(iii) Set $L_0 = L(\beta_0)$. Then (1.2) has one solution $u_0 \equiv 0$ for $L < L_0$, two solutions u_0 and u_1 for $L = L_0$ and three solutions u_0, u_1 and u_2 for $L > L_0$.

We next rewrite the problems (1.2) and (1.5) in the operator form. If $T : H_0^1 \to H_0^1$ and $T_h : H_0^1 \to S_0^h$ are the continuous linear operators defined, for $f \in H_0^1$, by

$$Tf = u \in H_0^1 \text{ if } (u_x, v_x) = (f, v), \forall v \in H_0^1,$$

and

$$T_h f = u^h \in S_0^h \text{ if } (u_x^h, \chi_x) = (Q_h f, \chi), \forall \chi \in S_0^h,$$

then (1.2) and (1.5) are, respectively, equivalent to the problems to find the pairs (L, u) in $\mathbb{R} \times H_0^1$ and (L, u^h) in $\mathbb{R} \times S_0^h$ such that

$$F(L, u) \equiv u + TG(L, u) = 0, \tag{2.3}$$

$$F_h(L, u^h) \equiv u^h + T_h G(L, u^h) = 0. \tag{2.4}$$

Lemma 2.2

(i) $F^0 \equiv F(L_0, u_1) = 0$

(ii) $D_u F^0 \equiv D_u F(L_0, u_1) = I + TD_u G(L_0, u_1) \in \mathcal{L}(H_0^1, H_0^1)$, is singular and -1 is an eigenvalue of the compact operator $TD_u G(L_0, u_0)$ with algebraic multiplicity 1.

(iii) there exists a unique branch $\{(L(\alpha),u(\alpha)); |\alpha| < \alpha_0$, for some

$\alpha_0 > 0\}$ of solutions of (2.3) in a neighborhood of the point

(L_0,u_1), where $\alpha \to L(\alpha)$ and $\alpha \to u(\alpha)$ are C^2 functions given by

$$\left. \begin{array}{c} L(\alpha) = L_0 + \xi(\alpha), \\ \\ u(\alpha) = u_0 + \eta(\alpha)\phi_0 + v(\xi(\alpha),\alpha). \end{array} \right\} \qquad (2.5)$$

Here $\xi(0) = \eta(0) = v(0,0) = 0$,

$H_0^1 = \text{Ker}(D_u F^0) \oplus \text{Range } (D_u F^0) \equiv V_1 \oplus V_2$,

$\phi_0 \in V_1$, $\|\phi_0\|_1 = 1$, $v \in V_2$.

(iv) (L_0,u_1) is a turning point, in particular, $\xi(\alpha)$ has a local minimum

at $\alpha = 0$ and

$$\xi'(0) = 0, \ \xi''(0) > 0. \qquad (2.6)$$

Proof. (i) follows from (2.3) and Lemma 2.1. Next set

$$\alpha = \beta - \beta_0.$$

Then the solution branch $\{L(\alpha), u(\alpha)\}_1 -\beta_0 < \alpha < 1-\beta_0\}$ of (1.2) provided by

Lemma 2.1 is clearly a smooth solution branch of (2.3) passing through

(L_0,u_1) at $\alpha = 0$. Differentiating (2.3) along this branch we find that

$$0 = \frac{d}{d\alpha} F^0 = D_u F(L_0,u_1) \frac{du}{d\alpha} (0) + D_L F(L_0,u_1) \frac{dL}{d\alpha} (0). \qquad (2.7)$$

Since by (2.2) $\frac{dL}{d\alpha} (0) = 0$, $D_u F^0$ is singular, and -1 is an eigenvalue of the

compact operator $TD_u G(L_0,u_1)$. Let ϕ be a corresponding eigenvector.

Since all the eigenvectors of $TD_u G(L_0,u_1)$ are smooth we also have

1164

$$\mathcal{L}\phi \equiv \phi'' + D_u G(L_0, u_1) \phi = 0,$$

$$\phi(0) = \phi(1) = 0.$$

Since all eigenvalues of $\mathcal{L}$ are simple, the corresponding eigenvalues of $TD_u G(L_0, u_1)$ are also simple, and we arrive at (ii).

The part (iii) follows from (i), (ii) and the classical theory of compact operators. Finally, (iv) follows from (iii), (2.1) and (2.2).

Using the Degree Theory [7] one can obtain from the above Lemmas existence, uniqueness and convergence of a branch $\{(L_h(\alpha), u_h(\alpha)); |\alpha| < \alpha_0\}$ of solutions of (2.4),. To establish the optimal rate of convergence we make an additional assumption (which can be verified numerically).

<u>Theorem 2.1</u> (i) Assume that (L_0, u_1) is a simple limit point of F, i.e. $D_L F^0 \notin V_2$ or, equivalently, $(D_L F^0, \phi_0) \neq 0$, where $\phi_0 \in V_1$ and V_2 are as in Lemma 2.2. Then there exists a unique branch $\{(L_h(\alpha), u^h(\alpha)); |\alpha| < \alpha_0\}$ of solutions of (2.4) (or (1.5)) in the neighborhood of the branch $\{(L(\alpha), u(\alpha); |\alpha| < \alpha_0\}$. This branch is of class C^∞, and for all $\alpha \in [-\alpha_0, \alpha_0]$ and all integer $m > 0$ we have the error estimate

$$|L_h^{(m)}(\alpha) - L^{(m)}(\alpha)| + |(u^h)^{(m)}(\alpha) - u^{(m)}(\alpha)|_1$$

$$\tag{2.8}$$

$$\leq C h^r \sum_{\ell=0}^{m} |u^{(\ell)}(\alpha)|_{r+2}, \quad r = 1, \ldots, k.$$

(ii) There exists a unique nondegenerate turning point (L_h^0, u_1^h) of F_h in a sufficiently small neighborhood of (L_0, u_1), and we have the error estimate

$$|L_h^0 - L_0| + |u_1^h - u_1|_1 \leq C h^r \sum_{\ell=0}^{1} |u^{(\ell)}(\alpha)|_{r+2}, \quad r = 1, \ldots, k, \tag{2.9}$$

<u>Proof</u>. In [4] we have proved that

$$\|(T-T_h)g\|_1 < c\ h^r \|g\|_r, \quad r = 1, \ldots, k, \ \forall\ g \in H^r. \quad (2.10)$$

By (i)-(iii) of Lemma 2.2 and (2.10) the assumptions of Theorem 3 in [1] are satisfied, and we have the existence of a unique branch $\{L_h(\alpha), u^h(\alpha)\}; |\alpha| < \alpha_0\}$ and the error estimate

$$|L_h^{(m)}(\alpha) - L^{(m)}(\alpha)| + \|(u^h)^{(m)}(\alpha) - u^{(m)}(\alpha)\|_1$$

$$< c \sum_{\ell=0}^{m} \|(T-T_h)\frac{d^\ell}{d\alpha^\ell} G(L(\alpha), u(\alpha))\|_1. \quad (2.11)$$

Hence (2.8) follows from (2.10), (2.11) and (1.2). Using also (2.6), by [1, Section 4] there exists a unique nondegenerate turning point (L_h^0, u_1^h) of F_h in a sufficiently small neighborhood of (L_0, u_1), and by [1, Theorem 4] we have the error estimate

$$|L_h^0 - L_0| + \|u_1^h - u_1\|_1 < c \sum_{\ell=0}^{1} \| (T-T_h)\frac{d^\ell}{d\alpha^\ell} G(L(\alpha), u(\alpha)) \big|_{\alpha=0}\|_1. (2.12)$$

Using again (2.10) and (1.2) gives (2.9).

<u>Remark</u>. Results in [6] imply optimal order error estimates in (2.8) and (2.9) when the H^1 norm of the error is replaced by the L_2 or L_∞ norm.

3. <u>Large time error estimates for the parabolic problem</u>.

In this section we derive large time error estimates for (1.4) where the initial data is in the domain of attraction of a stable steady-state solution $u*$, where $u*$ is u_0 or u_2. Our approach is a modification of the one in [5] where similar results have been proved without considering numerical integration.

We define the linearized operator L and the corresponding bilinear form

1166

by setting

$$Av = -v'' - g'(u^*(x))v,$$

$$A(v,w) = (v',w')-(g'(u^*)v,w),$$

where $g(u) \equiv -G(L,u)$. We assume that for some $\lambda_1 > 0$ there holds

$$\lambda_1 \|v\|^2 < A(v,v), \qquad \forall \ v \in H_0^1. \tag{3.1}$$

Lemma 3.1 [5] (i) There exists a neighborhood $N \subset H_0^2$ of u^* and α with $0 < \alpha < \lambda_1$ such that if u is a solution of (1.1) with $u(t_0) \in N$, $t_0 > 0$, then $u \in H_0^2$ exists for all $t > 0$ and we have

$$\|u(t) - u^*\|_2 < c \ e^{-\alpha(t-t_0)} \|u(t_0) - u^*\|_2, \ t > t_0. \tag{3.2}$$

(ii) There exists $c = c(m) > 0$ such that

$$\|u(t)\|_{C^m} < c, \quad m = 1, 0, \ldots, \quad t > t_0.$$

(iii) There exist $\lambda = \lambda(m) > 0$ and $c = c(m) > 0$ such that

$$\|u_t\|_m < c \ e^{-\lambda(t-t_0)}, \ t > t_0, \quad m = 0, 1, 2, \ldots \tag{3.4}$$

Let $u = u(t)$ and $u_h = u_h(t)$ be the solutions of (1.1) and (1.4), respectively. Set

$$g(u_h) - g(u) = w(t)(u_h-u),$$

$$\bar{w}(t) = w(t) - g'(u^*) = \int_0^1 \int_0^1 g''(u^*+\sigma(u+s(u_h-u) - u^*))d\sigma$$

$$\cdot(u-s(u_h-u)-u^*)ds,$$

where

1167

$$w(t) = \int_0^1 g'(u + s(u_h - u))ds.$$

We also set

$$N_\delta(u) = \{v \in L_\infty : |u-v|_{L_\infty} < \delta\} .$$

Lemma 3.2 For any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that for $u \in N_{\delta/2}(u^*)$ and $u_h \in N_{\delta/2}(u)$ we have $u_h \in N_\delta(u^*)$ and $|\bar{w}(t)|_{L_\infty} < \varepsilon$.

Proof. Obviously, the lemma holds with $\delta = \dfrac{\varepsilon}{2|g''|_{L_\infty}}$. ∎

Set

$$e(t) = u_h(t) - u(t).$$

Lemma 3.3. Let t_0 be as in Lemma 3.1. Suppose that for given ε with $0 < \varepsilon < \lambda_1$ we can find $t_1 = t_1(\varepsilon) > t_0$ such that for some $t^* > t$, the semidiscrete problem (1.4) has a solution $u_h(t)$, $0 < t < t^*$, and for $t_1 < t < t^*$ Lemma 3.2 holds.

Then there exist $\gamma = \gamma(\varepsilon) > 0$ and $h_0 > 0$ such that for $h < h_0$

$$|e(t)| < c_1(e^{-\gamma(t-t_1)}|e(t_1)|+h^{k+1}), \quad t_1 < t < t^*, \qquad (3.6)$$

$$|e(t)|_{L_\infty} < c_2 h^{-1/2} |e(t)|, \quad t_1 < t < t^*, \qquad (3.7)$$

where the constant c_1 and c_2 do not depend on h and t^*.

Proof. Following a standard procedure we set

$$e = u_h - u = (u_h - P_1 u) + (P_1 u - u) = \theta + \rho,$$

where $P_1 : H_0^1 \to S_0^h$ is the elliptic projection defined by $((P_1 v - v)_x, \chi_x) = 0, \forall \chi \in S_0^h$. It is well known that

$$\|P_1 v - v\|_1 + h^{-1} \|P_1 v - v\| \leq c \, h^s \, \|v\|_s, \quad \forall v \in H_0^s, \quad 1 < s \leq k. \quad (3.8)$$

From (1.1) and (1.4) we have the equation for θ:

$$(\theta_t, \chi) = -(\theta_x, \chi_x) + (Q_h g(u_h) - g(u), \chi) - (\rho_t, \chi), \quad (3.9)$$

with $g \equiv -G(L, u)$. Next write

$$Q_h G(u_h) - g(u) = Q_h \big(g(u_h) - g(u)\big) - (I - Q_h) g(u).$$

Using (3.5),

$$Q_h \big(g(u_h) - g(u)\big) = Q_h w(\theta + \rho) = Q_h w \rho + Q_h \big(\bar{w} + g'(u^*)\big)\theta$$

$$= Q_h w \rho + Q_h \bar{w}\theta - (I - Q_h) g'(u^*)\theta + g(u^*)\theta.$$

And thus (3.9) is written as

$$(\theta_t, \chi) + (\theta_x, \chi_x) - (g'(u^*)\theta, \chi)$$

$$\quad (3.10)$$

$$= \big(Q_h \bar{w}\theta + Q_h w \rho - (I - Q_h)(g(u) + g'(u^*)) - \rho_t, \chi\big).$$

Taking $\chi = \theta$,

$$\frac{1}{2} \frac{d}{dt} |\theta|^2 + |\theta_x|^2 - (g'(u^*)\theta, \theta) - (Q_h \bar{w}\theta, \theta)$$

$$= \big(Q_h w \rho - (I - Q_h)(g(u) + g'(u^*)) - \rho_t, \theta\big)$$

Using (3.1), this reduces to

$$|\theta| \frac{d}{dt} |\theta| + \lambda_1 |\theta|^2 - |Q_h \bar{w}|_{L_\infty} |\theta|^2$$

$$\leq \big(|Q_h w|_{L_\infty} |\rho| + |\rho_t| + |(I - Q_h)(g(u)$$

1169

$$+ g'(u^*))|) |\theta|.$$

Since $|Q_h|_{L_\infty} = 1$, $|w|_{L_\infty} < C$ and by Lemma 3.2 $|w|_{L_\infty} < \epsilon$, we have, setting $\gamma = \lambda_1 - \epsilon$,

$$\frac{d}{dt} (e^{\gamma t} |\theta|) < c e^{\gamma t}(|\rho| + |\rho_t|$$

$$+ |(I-Q_h)g(u) + g'(u^*)| \equiv c e^{\gamma t} \phi(t),$$

or after integration

$$|\theta(t)| < c e^{-\gamma(t-t_1)} |\theta(t_1)| + c \int_{t_1}^t e^{-\gamma(t-s)} \phi(s) ds. \quad (3.11)$$

Hence we conclude that altogether

$$|e(t)| < |\theta(t)| + |\rho(t)| < e^{-\gamma(t-t_1)} |e(t_1)| + c \sup_{t_1 < s < t} \phi(s)$$

Taking into account (3.8) and the fact that (3.8) also holds for P_1 replaced by Q_h,

$$|e(t)| < e^{-\gamma(t-t_1)} |e(t_1)| + c h^{k+1} \max_{t_1 < s < t} (|u(s)|_{k+1} + |u_t(s)|_{k+1}). \quad (3.12)$$

Finally, estimating u and u_t by Lemma 3.1, we arrive at (3.6). By using the approximation properties of S and an inverse inequality (see e.g. [5]), we also obtain (3.7).

From Lemma 3.3 we obtain in the same way as in [5]

Theorem 3.1. There exist $t_1 > 0$, $\gamma > 0$ and $h_0 > 0$ such that for $h < h_0$

$$|u_h(t)-u(t)| < c(\ell e^{-\gamma(t-t_1)} |u_h(t_1)-u(t_1)| + h^{k+1}), \quad t < t_1. \quad (3.13)$$

The order of convergence for $u_h(t_1)-u(t_1)$ depends on smoothness of

$u(t)$ on $[0, t_1]$, which in turn depends on the compatibility and smoothness of the initial data v. In particular, for $v \in H_0^\alpha$, $\alpha > 5/2$ (in the absence of nonlocal compatibility conditions on v) we have [3,8] for any $\varepsilon_1 > 0$

$$\|u_t(s)\|_\beta \leq c\, s^{-1+(5/2-\varepsilon_1-\beta)} . \qquad (3.14)$$

By a slight modification, to incorporate the numerical integration, of a standard procedure we can obtain the estimate

$$\|u_h(t)-u(t)\| \leq c\, \|v_h - v\| + c\, h^\beta \{\|v\|_\beta + \int_0^t \|u_t\|_\beta \, ds\}, \quad 0 < t < t_1, \quad 1 < \beta < k+1. \qquad (3.15)$$

In order for the integral in (3.15) to converge we must have $\beta < 5/2 - \varepsilon$, $\varepsilon > 0$. We thus obtain from Theorem 3.1

<u>Corollary</u>. Let $v \in H_0^\alpha$, $\alpha > 5/2$. Then for some $t_1 > 0$, $\gamma > 0$, $h_0 > 0$, any $\varepsilon > 0$, $h < h_0$ we have

$$\|u_h(t)-u(t)\| \leq c\, (e^{-\gamma(t-t_1)} h^{5/2-\varepsilon} + h^{k+1}), \quad t > t_1. \qquad (3.16)$$

<u>Remark</u>. By a modification, to incorporate numerical integration, of the argument in [5] one can also establish error estimates similar to (3.13), (3.16) for the gradient of the error and maximum norm error estimates.

References

[1] Brezzi, F., Rappaz, J., and Raviart, P. Finite dimensional approximation of nonlinear problems. Part II. Limit points. Numer. Math., 37 (1981), 1-28.

[2] Conley, C. and Smoller, J. Remarks on th stability of steady state solutions of reation-diffusion equations. In: Bifurcation phenomena in mathematical physics and related phenomena, C. Bardos and D. Bessis, eds., Reidel: Dodrecht, 1980, 47-56.

[3] Johnson, C., Larsson, S., Thomée, V., and Wahlbin, L. B. Error
     estimates for spacially discrete approximations of semilinear
     parabolic equations with nonsmooth initial data. Preprint (1985).

[4] Khalsa, S. N. S. Finite element approximation of a reation-diffusion
     equation. Part I: Application of topological techniques to the
     analysis of asymptotic behavior of the semidiscrete solutions .
     Quarterly Appl. Math. (to appear).

[5] Khalsa, S. N. S. Large time error estimates for spatially discrete
     approximations of asymptotically stable solutions of semilinear
     parabolic equations with nonsmooth initial data. (in preparation).

[6] Khalsa, S. N. S. A remark on the convergence of the perturbed Galerkin
     solutions of nonlinear problems. (in preparation).

[7] Krasnoselskii, M. A., Vainikko, G. M., Zabbreiko, P. P., Rututchii, J.
     B. and Stecenko, V. Ja. Approximate solution of operator equations,
     Walters-Noordhoff, Groningen, 1972.

[8] Larsson, S. On reaction-diffusion equations and their approximation by
     finite elements methods. Thesis, Chalmers University of Technology
     Goteborg, Sweden (1985).

[9] Manoranjan, V. S., Mitchell, A. R. and Sleeman, B. D. Bifurcation
     studies in reaction-diffusion. J. Comp. and Appl. Math. 11 (1984) 27-
     337.

[10] Smoller, J. Shock waves and reation-diffusion equations, Springer-
     Verlag: New York, 1983.

HIGH RESOLUTION, MINIMAL STORAGE ALGORITHMS FOR CONVECTION DOMINATED, CONVECTION-DIFFUSION EQUATIONS

V. Ervin
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332

and

W. Layton[*]
Department of Mathematics
and Statistics
University of Pittsburgh
Pittsburgh, PA 15260

ABSTRACT. Several new methods for the numerical solution of convection dominated, convection-diffusion equations are presented. These methods are high accuracy methods and, in some cases, monotone schemes. Further, they can be implemented in a way as to require only an asymptotically negligible increase in storage over usual first order methods. Thus they are promising candidates for vectorization. Numerical experiments are presented and some error estimates, proven by the authors for these schemes, are reviewed.

1. INTRODUCTION. In this paper we consider the problem of solving convection-diffusion equations, with dominant convection terms, by a high accuracy numerical method. There are two problems which provide convenient models for these effects, which we now consider.

1.1. The Model Problems. Let $\Omega \subset \mathbb{R}^2$ be a domain, then we seek $u(x,y)$, for $(x,y) \in \Omega$, satisfying

$$-\varepsilon \Delta u + v_1(x,y)u_x + v_2(x,y)u_y + g(x,y)u = q(x,y), \quad (x,y) \in \Omega,$$

(1.1)

$$u = 0 \text{ on } \partial\Omega.$$

In the above, u typically denotes a concentration or temperature that is convected through Ω by the velocity field $\underline{v}(x,y) = (v_1,v_2)$, $\varepsilon > 0$ is a measure of the diffusion effects relative to $\underline{v}$, $q(x,y)$ is a source and $g(x,y)u$ represents a loss term. In many typical applications $\varepsilon \ll |v|$ and we will particularly be interested in cases when (the cell Peclet number) $h|\underline{v}|/2\varepsilon \ll O(1)$, where h is a typical meshwidth used by a numerical scheme for (1.1). Boundary conditions other than Dirichlet are also possible. We will focus on Dirichlet conditions for our exposition here.

We also consider the 1-D version of (1.1), given by

$$-\varepsilon u'' + f(x)u' + g(x)u = q(x), \quad 0 < x < 1, u = u(x),$$

$$u(0) = \alpha, \quad u(1) = \beta.$$

(1.2)

However, we consider (1.1) to be the basic model problem. Methods that can only be applied to the 1-D case are not interesting for the problems we have in mind.

The algorithm we consider is a modification of the defect-correction method originally proposed by Hemker [29], [30]. It is extremely promising as a high accuracy, minimum storage algorithm for convection diffusion problems. Before we introduce the method in Section 2 we now consider a few typical applied problems to which it could be applied. We will then outline the criteria which we used to judge the potential success or failure of a numerical method applied to these problems.

1.2. Two Representative Applied Problems. It is important to note that convection-diffusion equations frequently arise in more complicated contexts: in systems of equations, nonlinear problems, complicated geometries, with reaction terms present, etc. To illustrate this consider the following two examples:

Example 1: The Navier-Stokes Equations. The stream-function-vorticity formulation of steady incompressible plane flow gives the following system of partial differential equations in a region $\Omega \subset \mathbb{R}^2$

$$-\nu \Delta \omega + v_1 \omega_x + v_2 \omega_y = 0$$

$$\omega = \Delta \psi, \quad v_1 = \psi_y, \quad v_2 = -\psi_x,$$

(1.3)

which consists of a convection-diffusion problem coupled with a Poisson equation. The velocity field $\underline{v} = (v_1, v_2)$ is to be determined as part of the problem.

Example 2: Convection-Diffusion-Reaction Equations. Under the constant density hypothesis, the steady combustion of a laminar, premixed flame is governed by a system of convection-diffusion-reaction equations of the form:

$$-\frac{1}{(Le)_j} \Delta u_j + \frac{\partial}{\partial x} f_j(\vec{u}) = q_j(x, \vec{u}), \quad j = 1, \ldots, J.$$

1174

The general combustion equations for a premixed gas consist of a nonlinear convection-diffusion-reaction system for conservation of energy and the various chemical species in the exothermic reaction. These are coupled to the Navier-Stokes equations (Example 1) for the gas in which the reaction takes place. Under the constant density hypothesis, we can uncouple the two systems. For example, with constant velocity, in 1-D, with the constant density hypothesis, for a simple A → B reaction for a steady premixed laminar flame, we have the system [32]:

$$\rho v \frac{d}{dx} Y_A = \frac{1}{Le_A} \frac{d^2 Y_A}{dx^2} - K_A \Omega_A,$$

$$\rho v \frac{dT}{dx} = \frac{d^2 T}{dx^2} + Q_A K_A \Omega_A,$$

$$\Omega_A = \rho Y_A \exp(-N_A/T),$$

where ρ is the density, v the gas velocity, Y_A the mass fraction of the reactant A, Le_A the Lewis number of A, T the temperature, Q_A the heat released, N_A the activation energy and Q_A a preexponential factor. These equations have been nondimensionalized.

Even in the simplest type of reaction this problem already gives various numerical schemes great difficulties. This is especially true when the velocity, v, is not known exactly and may contain spurious oscillations.

1.3. Goals of Numerical Methods for the Model Problems. For a numerical method to show promise on the types of applied problems that occur in, e.g., the previous two examples, it must have several characteristics when used on the model problems (1.1), (1.2). These goals include (but are not restricted to) the following:

(1)   It must be extendable to systems, nonlinear problems, etc., without severely degrading its performance.

(2)   The stability properties must be independent of the mesh geometry, size and orientation, with respect to ε and $\underline{v}$.

(3)   The method should have high accuracy at least in smooth regions. This is especially important for problems in high dimensions as high accuracy reduces the number of points necessary to achieve a certain percent error and thus reduces the total storage needed.

(4)   It should give oscillation free solutions without exces-
      sive artificial viscosity. In combustion problems,
      oscillations in the velocity field can trigger premature
      detonation while excessive artificial viscosity can
      lead to flame squelching.

(5)   Storage requirements must be minimized if the method is
      to be useful for 3-D problems.

(6)   The method should be easy to implement using, as far as
      possible, readily available software.

     In the next section we will introduce a method that
we feel satisfies the above six criteria, namely the modified
defect-correction method. In Section 3 we give some sample
numerical results for the method. We also compare it experi-
mentally to a variant of the streamline diffusion method.
We summarize in Section 4 the theoretical results that have
been proved by the authors.

     1.4. Other Approaches. Exponentially fitted methods
have been studied intensively for the problem (1.2) in 1-D,
see Allen and Southwell [1]. For extensions of this work
see, e.g., Kellogg and Tsan [17], de Groen and Hemker [6],
El-Mistikawy and Werle [9] and Berger, Solomon and Ciment [3].
These exponentially weighted schemes are frequently the best
method for (1.2) but can prove to be costly, in particular
for nonlinear problems in 2 and 3-D.

     Much less work has been done on numerical methods for
the 2-D problem. Kellogg [16] has shown convergence of the
2-D Allen and Southwell scheme under assumptions on the vector
field $\underline{v}$. Raithby [24] has considered a skew upwind scheme
and Hemker [29] considers a scheme based upon a convex com-
bination of the usual finite difference and finite element
approximations. Streamline upwinding has been considered by
a number of people, Brooks and Hughs [5], Kelly, Nakazawa and
Zienkiewicz [18], and Nävert [23], who analyzed a finite
element streamline diffusion scheme. With the exception of
the 2-D Allen and Southwell scheme, these others do not, in
general, converge uniformly in ϵ in 2-D, Roos [25], and most
do not even possess a discrete maximum principle! A satis-
factory, high accuracy method for the 2-D problems remains to
be found.

     2.   ITERATED DEFECT CORRECTION.
     2.1. Basic Method. We now present the method in its
simplest context. Consider (1.1), let Ω be, e.g., the unit
square, $h = 1/N < 1$, $x_j = jh$, $y_j = jh$ ($j = 0,1,...,N$),
$\underline{v}_{ij} = \underline{v}(x_i,y_i)$, $u_{ij} = u(x_i,y_i)$, etc. Define the operator
D_x^+, D_x^-, D_x^0, by

1176

$$D_x^+ u_{ij} \equiv h^{-1}(u_{i+1j} - u_{ij}), \quad D_x^- u_{ij} \equiv h^{-1}(u_{ij} - u_{i-1j})$$

$$D_x^0 \equiv (2h)^{-1}(u_{i+1j} - u_{i-1j}).$$

The operators D_y^+, D_y^-, D_y^0, are similarly defined. Let us further introduce the notation:

$$\Delta^h u_{ij} = (D_x^+ D_x^- + D_y^+ D_y^-)u_{ij}, \quad \nabla^h u_{ij} = (D_x^0 u_{ij}, D_y^0 u_{ij})u_{ij},$$

$$\Omega^h \equiv \text{interior nodes} = \{(x_i, y_j): 1 < i < N, \ 1 < j < N\}$$

$$\Gamma^h \equiv \text{boundary nodes} = \{(0, y_j): 0 \le j \le N\} \cup \{(1, y_i): 0 \le j \le N\} \cup$$
$$\{(x_i, 0): 0 \le i \le N\} \cup \{(x_i, 1): 0 \le i \le N\}.$$

The first step is to calculate the usual artificial viscosity solution U_{ij}^1 given by

$$L_{\epsilon_0}^h U_{ij}^1 = -\epsilon_0 \Delta^h U_{ij}^1 + \underline{v}_{ij} \cdot \nabla^h U_{ij}^1 + g_{ij} U_{ij}^1 = q_{ij}, \quad (x_i, y_j) \epsilon \ \Omega^h,$$

$$U_{ij}^1 = 0, \quad (x_i, y_j) \epsilon \ \Gamma^h, \tag{2.1}$$

where

$$\epsilon_0 = h \max_{1 \le i,j \le N} \{|\underline{v}_{ij} \cdot (1,0)|, |\underline{v}_{ij} \cdot (0,1)|\}/2 \tag{2.2}$$

Next an updated approximation is calculated via

$$R_{ij}^1 = q_{ij} - L_{\epsilon}^h U_{ij}^1, \quad (x_i, y_j) \epsilon \ \Omega^h \tag{2.3}$$

$$\left. \begin{array}{l} L_{\epsilon_0}^h E_{ij}^1 = R_{ij}^1, \quad (x_i, y_j) \epsilon \ \Omega^h \\[2mm] E_{ij}^1 = 0, \quad (x_i, y_j) \epsilon \ \Gamma^h \end{array} \right\} \tag{2.4}$$

$$U_{ij}^2 = U_{ij}^1 + E_{ij}^1 \tag{2.5}$$

where

$$L_\varepsilon^h U_{ij}^1 \equiv -\varepsilon\Delta^h U_{ij}^1 + \underline{v}_{ij}\cdot\nabla^h U_{ij}^1 + g_{ij}U_{ij}^1$$

We stop after two or three iterations, otherwise, Hemker [30], if $U^j \to U^\infty$ then $L_\varepsilon^h U^\infty = q_{ij}$ and U^∞ is the highly oscillatory, unsatisfactory central difference approximation.

We will further refine the basic method in §2.3.

<u>Notes</u>

(1) The above method extends readily to non-uniform meshes and more complex geometries.

(2) Condition (2.2) ensures that the discrete coefficient matrix is a diagonally semi-dominant M-matrix [26], provided $g \geq 0$.

(3) There are a number of other possible and promising correctors to use in place of L_{ε_0}. For example, we may try the streamline diffusion corrector, $L_{\varepsilon,\delta}^h$, defined as follows: If

$$\frac{\partial}{\partial\underline{v}} = \underline{v}\circ\text{grad}, \quad \frac{\partial^2}{\partial\underline{v}^2} = \frac{\partial}{\partial\underline{v}}(\frac{\partial}{\partial\underline{v}}),$$

we define

$$L_{\varepsilon,\delta}w = -\varepsilon\Delta w - \delta\frac{\partial^2 w}{\partial\underline{v}^2} + \underline{v}\circ\nabla w + gw,$$

where $0 \leq \delta = O(h)$. When $\underline{v}$ is oriented with, or at 45° to, the mesh, $L_{\varepsilon,\delta}^h$ (the usual discretization of $L_{\varepsilon,\delta}$) works well, see Sections 3 and 4. Alternately, we may use the upwind approximation:

$$L_{\alpha,\beta}^h w_{ij} \equiv -\varepsilon\Delta^h w_{ij} + v_1(2h)^{-1}[(-1+\alpha)w_{i-1,j}-2\alpha w_{ij}+(1+\alpha)w_{i+1,j}]$$

$$+ v_2(2h)^{-1}[(1+\beta)w_{ij+1}-2\beta w_{ij}+(-1+\beta)w_{ij-1}] + gw_{ij},$$

where α,β are chosen to ensure that $L_{\alpha,\beta}^h$ is an M-matrix. We mention, in particular, the choice of "optimal upwind parameters" $(\rho = h/\varepsilon)$

$$\alpha = -2/\rho v_1 + \coth(\rho v_1/2),$$

$$\beta = -2/\rho v_2 + \coth(\rho v_2/2).$$

Another possibility is to use different amounts of viscosity in different directions. We replace $L_{\epsilon_0}^h$ by the central difference approximation to $L_{\epsilon_1,\epsilon_2} w \equiv -\epsilon_1 w_{xx} - \epsilon_2 w_{yy} + v_1 w_x + v_2 w_y + gw$, where

$$\epsilon_1 \geq \max\{\epsilon, \frac{hv_1}{2\epsilon}\}, \quad \epsilon_2 \geq \max\{\epsilon, \frac{hv_2}{2\epsilon}\}.$$

The use of $L_{\epsilon_0}^h$ is much cruder than these choices but it has the advantage of being independent of the velocity field $\underline{v}$.

Computational Complexity. The discrete system of equations (2.1) may equivalently be expressed in the form

$$A_{\epsilon_0} \underline{U}^1 = \underline{q} \tag{2.7}$$

where (i) A_{ϵ_0} is an $N^2 \times N^2$ banded matrix with main diagonal, and four off diagonals nonzero, and halfbandwidth equal to $N = h^{-1}$. (ii) $\underline{q} = (\ldots q_{ij} \ldots)^T$ is an $N \times 1$ vector. Likewise the updated solution (2.5) is given by

$$A_{\epsilon_0} U^{n+1} = \underline{q} + (A_{\epsilon_0} - A_\epsilon) U^n, \quad n = 1,2,\ldots \tag{2.8}$$

The "best" method of solution for (2.7) and (2.8) is dependent upon the size of the linear system and the facilities available to the researcher (e.g., scalar or vector computer). Solutions of (2.7) and (2.8) via (a) direct methods and (b) iterative methods (e.g., conjugate residuals, conjugate gradient, S.O.R. etc.) are briefly discussed below.

(a) Direct Method of Solution of (2.7) and (2.8). Using a direct method the system of equations (2.7) is factored initially as $A_{\epsilon_0} = LU$. Each iterate past the first requires only a residual calculation, a forwardsolve and a backsolve. Thus, the computational complexity involved in computing U^2, U^3, etc., is negligible w.r.t. the amount

required to compute U^1, the usual artificial viscosity approximation.

Since we only use A_ε for a residual calculation we may store it using only $O(N^2)$ locations, which is also negligible additional storage compared to the $O(N^3)$ locations needed to computer L-U decomposition of A_{ε_0}.

(b) Iterative Method of Solution of (2.7) and (2.8).
Iterative methods can be very efficient for the solution of (2.7) and (2.8) because we take U^n as an initial guess in calculating U^{n+1}. Typically, only a few iterations are needed to calculate U^{n+1} since we begin with a good initial guess. The coefficient matrix, A_{ε_0}, is a diagonally dominant M-matrix and thus such methods as conjugate residuals and (overrelaxed multilevel) Jacobi will converge nicely. More-over the storage requirements are reduced significantly as both A_{ε_0} and A_ε can be efficiently stored as five $N \times 1$ vectors respectively.

2.2. 4th Order Corrector. Hemker [29], [30], reports, based on "local mode" analysis, we can replace L_ε^h in (2.3) by a 4th order corrector $\tilde{L}_\varepsilon^h$, such as the nine point cross:

$$\tilde{L}_\varepsilon^h U_{ij} \equiv \begin{cases} -\varepsilon \Delta_4^h U_{ij} + \underline{v}_{ij} \cdot \nabla_4^h U_{ij} + g_{ij} U_{ij}, & 1 < i,j < N-1, \\ \\ -\varepsilon \Delta h U_{ij} + \underline{v}_{ij} \cdot \nabla^h U_{ij} + g_{ij} U_{ij}, & i \text{ or } j = 1 \text{ or } N-1, \end{cases}$$

$$\Delta_4^h U_{ij} = (12h^2)^{-1}(-U_{i+2j} + 16U_{i+1j} - 48U_{ij} + 16U_{i-1j} - U_{i-2j}$$

$$-U_{ij+2} + 16U_{ij+1} + 16U_{ij-1} - U_{ij-2})$$

$$\nabla_4^h U_{ij} = (4h)^{-1}(-U_{i+2j} + 4U_{i+1j} - 4U_{i-1j} + U_{i-2j},$$

$$-U_{ij+2} + 4U_{ij+1} - 4U_{ij-1} + U_{ij-2}),$$

and expect $\|u-U^4\| = O(h^4)$.

This is untested in practice but appears especially

promising as the implementation of this higher order corrector can be achieved with very little additional work or storage. Specifically, the associated corrector matrix $\tilde{A}_\epsilon$ has nine non-zeros diagonals (which may be efficiently stored as nine $N \times 1$ vectors since we only use $\tilde{A}_\epsilon$ for calculating residuals) and the computation of the RHS of (2.8) only requires an additional $4N$ scalar multiplications and additions.

2.3. Filter Step. We have shown that (2.1)-(2.5) gives a good approximation away from the layers (even on a coarse mesh). Experimentally, we observe that each iteration:

(i) Outside of the layer the approximation improves
(ii) The numerical boundary layer spreads.

To avoid the spread of the boundary layer and yet still iteratively improve the approximation we modify (2.1)-(2.5) to include a filter step. Specifically, replace R_{ij}^1 in (2.4) by $F(R_{ij}^1)$. The filter step can also be inserted in (2.5) and gives the Modified Defect Correction Method, Hemker [30]. We are currently investigating several different filters and comparing performance.

The order of the filter should be consistent with the accuracy of the method. For example, if we use a fourth order corrector we apply a fourth order filter. Specifically, for smooth $w(x,y)$, $w - F(w) = O(h^4)$. In other words, if $F(w_{ij}) = \sum_{\ell,m} \lambda_{\ell m} w_{i+\ell,j+m}$, then for $\underset{\sim}{\theta} = (\theta_1,\theta_2)$

$$f(\underset{\sim}{\theta}) \equiv \sum_{\ell,m} \lambda_{\ell,m} e^{i\ell\theta_1} e^{im\theta_2} = 1 + O(|\underset{\sim}{\theta}|^4)$$

as $|\underset{\sim}{\theta}| \to 0$. .

3. SAMPLE NUMERICAL RESULTS. We now give some sample numerical results for the defect correction algorithm and compare them with a variant, studied by Ervin and Layton [12], of the streamline diffusion method of Brooks and Hughs [5]. To summarize these findings: for simple model problems with constant coefficients or simple layer structure, etc., both methods yield good answers. The numerical experiments in these cases confirm the theoretical predictions as to their rates of convergence (see Section 4). On more complex problems with multiple and interior layers, attractive type turning points, etc., the defect correction approach proved to be more robust and gave excellent approximations -- both qualitatively and quantitatively. On these complex problems the modified streamline diffusion approach was sensitive to the precise choice of the diffusion parameter, δ. In some cases

oscillations appeared near turning points -- possibly due to a non-optimal choice of δ.

The modified defect correction method extends directly and easily to 2-D with excellent results. Extension of the streamline diffusion method to 2-D as a monotone type scheme is still an open problem and under investigations.

Before we give the numerical results we will introduce the "adjusted" streamline diffusion method and discuss some of the filter steps that have actually been used in (2.4).

3.1. Adjusted Streamline Diffusion Method.

Define, as above, $\partial/\partial \underline{v} = \underline{v} \circ \text{grad}$, $\partial^2/\partial \underline{v}^2 = \partial/\partial \underline{v}(\partial/\partial \underline{v})$. The continuous operator is defined via applying the operator $(I - \delta \frac{\partial}{\partial \underline{v}})$ to (4.1) and omitting the $O(\varepsilon\delta)$ (third order) term. This results in the B.V.P.

$$\widetilde{L}_{\varepsilon,\delta} w \equiv -\varepsilon\Delta w - \delta \frac{\partial^2}{\partial \underline{v}^2} w + \frac{\partial}{\partial \underline{v}}(w-\delta gw) + gw$$

$$= q - \delta \frac{\partial}{\partial \underline{v}} q \quad \text{in } \Omega,$$

$$w = 0 \quad \text{on } \partial\Omega. \tag{3.1}$$

Note that $\widetilde{L}_{\varepsilon,\delta}$ has the same form as L_ε, except for an added streamline diffusion term. Here $\delta = O(h)$ is picked to attempt to ensure that the central difference approximation to (3.1), $L^h_{\varepsilon,\delta}$, is of monotone type (see Layton and Morley [22]) once the boundary unknowns are eliminated from the linear system.

Define the usual $O(h^2)$ approximations to $\frac{\partial^2}{\partial x \partial y}$:

$$D^+_{xy} = \frac{1}{2}(D^+_x D^+_y + D^-_x D^-_y), \quad D^-_{xy} = \frac{1}{2}(D^+_x D^-_y + D^-_x D^+_y),$$

$$D^\theta_{xy} = \theta D^+_{xy} + (1-\theta)D^-_{xy} \quad (0 \le \theta \le 1),$$

where

$$D^+_x u(x,y) = h^{-1}(u(x+h,y) - u(x,y)),$$

etc. One $O(h^2)$ discretization of (3.1) proceeds as follows.

1182

The principle part, $L_{\varepsilon,\delta}^0$, of $\tilde{L}_{\varepsilon,\delta}$ is given by

$$L_{\varepsilon,\delta}^0 w \equiv -(\varepsilon+\delta v_1^2)w_{xx} - 2\delta v_1 v_2 w_{xy} - (\varepsilon+\delta v_2^2)u_{yy}$$

and is discretized by:

$$(L_{\varepsilon,\delta}^0)^h \equiv -(\varepsilon+\delta v_1^2)D_x^2 - 2\delta v_1 v_2 D_{xy}^\theta - (\varepsilon+\delta v_2^2)D_y^2 .$$

The remaining, lower order terms, in (3.1) are approximated by

(i) upwind differences for terms premultiplied by δ,
(ii) central differences for the remaining terms.

This yields the discrete operator $\tilde{L}_{\varepsilon,\delta}^h$. $\theta \in [0,1]$ and $\delta = O(h)$ can be chosen point by point. The "optimal" choice for stability of these two parameters is not clear at this moment, see Section 4 for more details.

3.2. Various Filter Steps. We have found that the method (2.1)-(2.5) is fairly robust as to the specific filter step used. For example, in the results which we are now reporting, we used a clipping-filter on the residual vector $\underline{R}^n = \underline{q}-L_\varepsilon^h U^n$:

Calculate $\bar{r}$ and r_σ (mean and standard deviation),

If $|r_j-\bar{r}| > r_\sigma$, set $r_j := 0$,   (3.2)

Otherwise $r_j := r_j$.

Hemker [30] proceeded by filtering U^{n+1} by applying one step of Jacobi iteration to U^{n+1} using the artificial viscosity matrix. Experiments are currently underway with other averaging operators, Tchebyscheff filters, etc.

3.3. Numerical Experiments. We now give some 2-D and 1-D examples for the methods discussed above.

Example 1.

$$L_\varepsilon u \equiv -\varepsilon\Delta u + u_x = g \text{ in } \Omega = (0,1) \times (0,1)$$
$$u = 0 \text{ on } \partial\Omega.$$ (3.3)

$g = g(x,y,\epsilon)$ was chosen so that the true solution was given by

$$u(x,y) = \sin(\pi x)\sin(\pi y)[e^{-\frac{1-x}{\epsilon}} + e^{-\frac{1-y}{\sqrt{\epsilon}}} + e^{\frac{-y}{\sqrt{\epsilon}}}].$$

Note that $u(x,y)$ has characteristic boundary layers of width $O(\sqrt{\epsilon})$ along the boundaries $y = 0$ and $y = 1$ and an outflow boundary layer of order ϵ along $x = 1$, Echaus [7]. We tested the methods with $h = \Delta x = \Delta y = N^{-1}$, $N = 8$, 15 and $\epsilon = 10^{-2}$, so that the mesh was extremely coarse with respect to the $O(\epsilon)$ outflow boundary layer.

With $L_{\epsilon_0}^h$ as the correction operator, we found that, away from the layers, the defect correction method gave perfect $O(h^2)$ convergence. In the figures that follow we give, respectively, the bilinear interpolant of the true solution and the approximate solution and error plot at the 4th iterate for $h = 1/7$ and $1/14$. It is interesting to note that the corners where the characteristic and outflow layers overlap appear to be the most difficult areas to approximate u. The global errors and decay exponents are given in Table 3.1.

Table 3.1. Global errors in defect-correction approximation to (3.3). Correction operator is taken to be artificial viscosity approximation to L_ϵ and no filter was used.

Iterate	Max-Norm Error		Decay Exponent	l_2-Error		Decay Exponent
	N = 8	N = 15		N = 8	N=15	
1	.098	.086	.2	.039	.035	.17
2	.087	.069	.37	.034	.027	.35
3	.079	.056	.53	.030	.022	.58
4	.075	.053	.56	.027	.017	.68

It is remarkable that the method seems to be converging (slowly) even in the layers and in the corners mentioned above. This is not even predicted by the estimates in 1-D.

We next tried to improve these results by using the streamline diffusion operator as the corrector, $L_{\epsilon,\delta}^h$, with $\delta = O(h)$. The improvement was dramatic and is summarized in Table 3.2.

Table 3.2. Approximation solution of (3.3) using streamline diffusion corrector, no filter is used.

Iterate	Max-Norm Error		Decay Exponent	l_2-Error		Decay Exponent
	N = 8	N = 15		N = 8	N=15	
1	.059	.033	.92	.025	.014	.9
2	.039	.016	1.44	.013	.0044	1.7
3	.033	.012	1.6	.0082	.0025	1.9
4	.030	.010	1.7	.0065	.0020	1.9

The example (3.3) is the best possible for the streamline diffusion correction: no grid orientation effects are present in the numerical model of (3.3) since the velocity field $\underline{v} = (1,0)$ is oriented with the mesh.

Many more experiments must be performed in 2-D to validate the method and to test various correctors.

The next example illustrates the advantage of applying a simple filter when using the defect-correction method.

Example 2.

$$-\varepsilon u'' + u' - u = 0$$

$$u(0) = 1, \quad u(1) = 1$$

(3.4)

The filter used for this and the following 1-D examples was the clipping filter given in (3.2).

With $\varepsilon = 10^{-2}$ and h = 1/10 and 1/20, the error and the decay exponents are given in Tables 3.3 and 3.4 for approximations computed with and without using the filter, respectively.

Moreover observe that after three iterations the approximate solution obtained simply by iterating is beginning to oscillate about the true solution. The filter step has virtually no effects on the approximate solution away from the layer. It actually provides a very good answer up to the edge of the boundary layer.

The following two examples illustrate the effectiveness of the defect-correction method (with filter) on problems involving complicated, multiple boundary layers, turning points and nonsmooth coefficients. The equations (3.5)-(3.6) were taken from Pearson [31] so that an asymptotic solution was available for comparison. A change of independent variable

Table 3.3. Error and decay exponents for (3.4) after 3 iterations without using clipping filter.

x_n	Error h = 0.1	Error h = 0.05	Decay Exponent
.1	.200E-3	.483E-4	2.05
.2	.444E-3	.107E-3	2.05
.3	.738E-3	.177E-3	2.06
.4	.109E-2	.262E-3	2.06
.5	.151E-2	.362E-3	2.06
.6	.201E-2	.481E-3	2.06
.7	-.161E+0	.620E-3	--
.8	.635E+0	.785E-3	--
.9	-.121E+1	.281E+0	--

Table 3.4. Error and decay exponents for (3.4) after 3 iterations using the simple clipping type filter.

x_n	Error h = 0.1	Error h = 0.05	Decay Exponent
.1	.200E-3	.483E-4	2.05
.2	.444E-3	.107E-3	2.05
.3	.738E-3	.177E-3	2.06
.4	.109E-2	.262E-3	2.06
.5	.151E-2	.362E-3	2.06
.6	.201E-2	.481E-3	2.06
.7	.859E-4	.620E-3	--
.8	.762E-2	.785E-3	3.28
.9	-.241E-2	.153E-2	0.66

was made to pose the problems on $0 \leq x \leq 1$ rather than $-1 \leq x \leq 1$. The boundary conditions for each equation was

$$u(0) = 1, \quad u(1) = 2.$$

Example 3.

$$\varepsilon u'' + |x-0.5|u' + (2x-1.5)^3 u = 0 \tag{3.5}$$

Example 4.

$$\varepsilon u'' + (2x^2-2x+0.25)u' + (2x-1)u = 0. \tag{3.6}$$

For the examples, $\varepsilon = 0.25E-4$ and $h = 0.01$.

In both cases the method yielded an approximate solution matching closely the asymptotic solutions found by Pearson [12] with the same qualitative behavior. Boundary layers occurring at the endpoints of the domain were contained to one mesh interval and internal boundary layers influenced no more than three mesh points. The approximate solutions to (3.5) and (3.6) are illustrated in the Figures 6 and 7.

Numerical approximations to the solutions of Examples 3 and 4 were also computed using the streamline diffusion method described in §3.1. This method also gave very good approximations but was more sensitive to the choice of h and δ. Specifically for Example 3 with h = 0.01 the numerical approximation contained a spurous oscillation, see Figure 9, however for h = 0.25E-2 the approximation's behavior agreed with that of the asymptotic solution, Figure 10.

4. ERROR ESTIMATES FOR THE METHODS.

In this section we summarize the error estimates that have been proven for the modified defect correction method and the adjusted streamline diffusion method. We consider the 2-D problem (1.1) and the 1-D problem (1.2). Here we assume that the coefficients of (1.1), (1.2) are smooth, $g(x) \geq 0$, the domain Ω is "meshlined" and ε is small w.r.t. acceptable (outer) meshwidths.

4.1. The Adjusted Streamline Diffusion Method.

We begin by considering the method defined in Section 3.1. Special questions are associated with this method in 2-D. To isolate these issues, we focus our attention briefly on the principle part of $\tilde{L}_{\varepsilon,\delta}$, $L^0_{\varepsilon,\delta}$ and its approximation.

Since the principle part of $L_{\epsilon,\delta}$ contains a cross-derivative term, we examine its approximation carefully. Recall that

$$L^0_{\epsilon,\delta}u \equiv (-\epsilon+\delta v_1^2)u_{xx} - 2\delta v_1 v_2 u_{xy} - (\epsilon+\delta v_2^2)u_{yy},$$

$$(L^0_{\epsilon,\delta})^h u_{ij} = -(\epsilon+\delta v_1^2)D_x^2 u - 2\delta v_1 v_2 D_{xy}^\theta u - (\epsilon+\delta v_2^2)D_y^2 u.$$

Here $D_x^2 = D_x^+ D_x^-$, $D_y^2 = D_y^+ D_y^-$ and $(L^0_{\epsilon,\delta})^h$ is defined on a uniform mesh on a meshlined domain Ω^h, see Section 3.1 for more details.

__Theorem 4.1. [see Ervin and Layton [12]].__

(i) For $\epsilon > 0$, $\delta \geq 0$, $L^0_{\epsilon,\delta}$ is an elliptic operator.

(ii) For $0 < \epsilon \ll 1$, $\delta \geq 0$ and general velocity fields $\underline{v} = (v_1,v_2)$, $(L^0_{\epsilon,\delta})^h$ is __not__ a positive type difference operator.

(iii) There does not exist a __consistent__, positive type approximation to $L^0_{\epsilon,\delta}$ under the assumptions of (ii). □

Nevertheless $(L^0_{\epsilon,\delta})^h$ does contain interesting mathematical structures -- the __interior__ discrete maximum principle holds for the operator:

__Theorem 4.2. [see Layton and Morley [22; Thm. 1]].__
Assume $\epsilon > 0$, $\delta \geq 0$. $(L^0_{\epsilon,\delta})^h$ is a (inverse) monotone operator on the interior nodes when the boundary conditions are homogeneous. Thus, the interior discrete maximum principle holds. □

We are currently working on extending Theorem 4.2 to the case of lower order terms, including a precise "prescription" for the proper choice of δ.

In 1-D the situation is much clearer. Define, for $x_n = h, 2h, \ldots, (N-1)h$, $(h = 1/N)$.

$$L^h_{\epsilon+\delta}U_n \equiv -(\epsilon+\delta_n f^2(x_n))D_x^2 U_n + f(x_n)(1-\delta_n f'(x_n)-\delta_n g(x_n))D_x U_n$$

$$+(g(x_n)-\delta_n f(x_n)g'(x_n))U_n = q-\delta_n f(x_n)q'(x_n) \qquad (4.1)$$

$$U_0 = \alpha, \quad U_N = \beta,$$

where $D_x U_n = (U_{n+1} - U_{n-1})/2h$. δ_n is chosen to satisfy the conditions:

$$\frac{h|f(x_n)||1-\delta_n(f'(x_n)+g(x_n))|}{2|\varepsilon+\delta_n f^2(x_n)|} \le 1, \qquad \delta = O(h) \qquad (4.2)$$

Theorem 4.3. [Ervin and Layton [12]]. The discrete maximum principle holds for $L_{\varepsilon+\delta}^h$ provided (4.2) holds and

$$g_n - \delta_n f_n g_n' \ge 0. \qquad (4.3)$$
□

We then have the global error estimate:

Theorem 4.4. [Ervin and Layton [12]]. Assume (4.2), (4.3) hold. Then the error in (4.1) satisfies

$$\max_{0 \le x_n \le 1} |u(x_n)-U_n| \le Ch(\varepsilon+\delta+h) \max_{0 \le x \le 1} |u'''(x)|. \qquad (4.4)$$

If ε is small w.r.t h and $f(x) \ge a > 0$ we have

$$\max_{0 \le x_n \le 1} |u(x_n)-U_n| \le Ch. \qquad (4.5)$$

In the above, C is independent of ε. □

(4.5) shows that the method converges linearly uniformly on [0,1] and (4.4) implies that when u is smooth in ε that the convergence is essentially quadratic $O(h^2+\varepsilon h)$. The next result shows that we obtain this high rate of convergence outside of the layers even when u is singular in ε. We assume that $f(x) > a > 0$ so that there is an $O(\varepsilon)$ outflow-type layer at $x = 1$. In this $u^{(j)}(1) = O(\varepsilon^{-j})$ as $\varepsilon \to 0$.

Theorem 4.5. [Ervin and Layton [12]]. In addition to the assumptions of Theorem 4.4, suppose $f(x) > a > 0$, and $f(x_n)-\delta_n f(x_n'g_n'(x_n) - \delta_n f(x_n)f'(x_n) \ge a$. Then, for $0 \le x_n < 1$:

$$|u(x_n)-U_n| \le Ch^2\{1+h^{-2} \exp[-b\frac{1-x_n}{\varepsilon_0}]\} + C\varepsilon h\{1+\varepsilon^{-2} \exp[-a\frac{1-x_n}{\varepsilon}]\},$$

where $c > 0$ is independent of ε, h, and u, $b = \min\{a, \ln 3\}$, $\varepsilon_0 = \max_n(\varepsilon+\delta_n f_n^2)$. □

We next consider the problem of estimating derivatives of solutions of singularly perturbed B.V.P.'s. This is the important problem as stress intensity factors, skin friction coefficients, etc. require knowledge of these terms.

Theorem 4.6. [Ervin and Layton [13]]. Under the assumptions of Theorem 4.5 we have:

$$|u'(x_n) - \frac{U_{n+1}-U_{n-1}}{2h}| \le Ch^2\{1+h^{-3}\exp[-b\,\frac{1-x_n}{\varepsilon_0}]\}$$
$$+ C\varepsilon h\{1+\varepsilon^{-3}\exp[-a\,\frac{1-x_n}{\varepsilon}]\}. \qquad \Box$$

Thus, away from the layer we can get good approximations via the usual methods. At the layer at $x = 1$ (in our case) we require an exponentially fitted difference quotient. Under the assumptions of Theorem 4.6 define

$$\tilde{D}U_N \equiv \varepsilon^{-1}(\frac{\beta - U_n}{1-\exp[-f(1)\,\frac{1-x_n}{\varepsilon}]})f(1), \quad x_n = 1-O(h).$$

We then have that the relative error is $O(h)$ (as $u'(1) = O(\varepsilon^{-1})$).

Theorem 4.7. [Ervin and Layton [13]]. Under the assumptions of Theorem 4.6. Suppose x_n is chosen so that $1-x_n = O(h)$ and $|u(x_n)-U_n| = O(\varepsilon h+h^2)$ uniformly in ε. Then

$$|\frac{u'(1)-\tilde{D}U_N}{\varepsilon^{-1}}| \le C(h+\varepsilon). \qquad \Box$$

The techniques used to establish these theorems involve a potpouri of arguments due to Gershgorin [15], barrier function arguments following Kellogg and Tsan [17], maximum principle arguments used to validate asymptotic expansions for B.V.P.'s, see Eckhaus [7], and the theory of monotone matrices developed by Bramble and Hubbard [4] and Varga [26].

We note that the "adjusted streamline diffusion method" we study is a finite difference interpretation of the finite element streamline diffusion method, proposed and studied by Wahlbin [27], [28] for scalar hyperbolic equations, examined for hyperbolic systems by Layton [19], [20], [21], and Du, Gunzburger and Layton [8]. The finite element implementation of this circle of ideas was analyzed for (1.1) by Nävert [23] and applied to fluid flow problems by Brooks and Hughs [5].

4.2.   The Defect-Correction Method.   Global error esti-
mates for the modified defect correction method can be
established in 2-D and in 1-D in a similar manner to Theorem
4.4.   These error estimates show that, when the true solution
is smooth uniformly in ε, the error after k iterations in
the basic method (artificial viscosity corrector, second order
defect operator, no filter step) is $O(h^2+(\varepsilon_0-\varepsilon)^k)$.   The local
error estimates in 1-D are more interesting as they give an
idea of the spread of the numerical boundary layer from one
iteration to the next.

    Theorem 4.8.   [Ervin and Layton [14]].   Let $U_n^k = u(x_n)$
be the kth defect correction approximation to the solution
of (1.1).   Suppose that $f(x) \geq a > 0$, $g(x) \geq 0$ and that
$F \equiv I$ (the filter step is omitted).   Then, for $n = 0,1,\ldots,N$

$$|u(x_n)-U_n^k| \leq Ch^2[1+\varepsilon_0^{-2}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ C(\varepsilon_0-\varepsilon)^k[1+\varepsilon\varepsilon_0^{-k}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ Ch^4[1+\varepsilon_0^{-4}\exp(-b\,\frac{1-x_n}{\varepsilon_0})],$$

for $n = 1,2,\ldots,N-1$,

$$|u'(x_n)-\frac{U_{n+1}^k-U_{n-1}^k}{2h}| \leq Ch^2[1+\varepsilon_0^{-3}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ C(\varepsilon_0-\varepsilon)^k[1+\varepsilon\varepsilon_0^{-k-1}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ Ch^4[1+\varepsilon_0^{-5}\exp(-b\,\frac{1-x_n}{\varepsilon_0})].$$

$$|u''(x_n)-\frac{U_{n+1}^k-2U_n^k+U_{n-1}^k}{h^2}| \leq Ch^2[1+\varepsilon_0^{-4}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ C(\varepsilon_0-\varepsilon)^k[1+\varepsilon\varepsilon_0^{-k-2}\exp(-a\,\frac{1-x_n}{\varepsilon_0})]$$

$$+ Ch^4[1+\varepsilon_0^{-6}\exp(-b\,\frac{1-x_n}{\varepsilon_0})]$$

where $b = \min\{a, \ln 3\}$.   □

    The modified defect correction iteration can also be

implemented via a finite element type procedure. This case
was studied in Axelsson and Layton [2] where global error
estimates in $L^2(\Omega)$ and $H^1(\Omega)$ were proven.

REFERENCES

1. Allen, D. N. de G. and Southwell, R. V., Relaxation methods to determine the motion in 2-D, of a viscous incompressible fluid past a fixed cylinder, J. Mech. Appl. Math., 8 (1955), 129-145.

2. Axelsson, O. and Layton, W., Defect correction methods for convection dominated convection-diffusion equations, Report 8335, Univ. of Nijmegen, 1983, submitted to Num. Math.

3. Berger, A. E., Solomon, J. M. and Ciment, M., Higher order accurate tridiagonal difference methods for diffusion convection equations, in: Proc. Third IMACS Conf. on Comp. Methods for P.D.E.'s, Lehigh University, 1979.

4. Bramble, J. H. and Hubbard, B. E., New monotone type approximations for elliptic problems, Math. Comp., 18 (1964), 349-367.

5. Brooks, A. N. and Hughs, T. J., Streamline-upwind Petrov Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier Stokes equations, Comp. Meth. Appl. Mech. Eng., 32 (1982), 199-259.

6. de Groen, P. P. N. and Hemker, P. W., Error bounds for exponentially fitted Galerkin methods applied to stiff two-point boundary value problems, in: Num. Anal. of Sing. Pert. Probs. (eds. P. W. Hemker and J. J. H. Miller) London, Academic Press, 1979.

7. Eckhaus, W., Asymptotic Analysis of Singular Perturbations, North Holland, Amsterdam, 1979.

8. Du, Q., Gunzburger, M. and Layton, W., Minimum dispersion finite element methods for multidimensional hyperbolic systems, preprint, 1986.

9. El-Mistikawy, T. M. and Werle, M. J., Numerical method for boundary layer with blowing, the exponential box scheme, Am. Inst., Astronaut. Aeronaut. J., 16 (1978), 749-751.

10. Ervin, V. and Layton, W., BDLAYER-software for convection-diffusion equations, 1984.

11. Ervin, V. and Layton, W., DEF-COR-a modified defect correction algorithm, submitted to A.C.M. Transactions on Math. Software, 1984.

12. Ervin, V. and Layton, W., A second order accurate, positive scheme for singularly perturbed boundary value problems, preprint, 1984, submitted to J. Appl. Comput. Mechanics.

13. Ervin, V. and Layton, W., On the approximation of derivatives of solutions of singularly perturbed boundary value problems, to appear in: SIAM J. Sci. Stat. Comp.

14. Ervin, V. and Layton, W., A study of defect-correction, finite difference methods for convection diffusion equations, submitted to SIAM J.N.A.

15. Gershgorin, S., Fehlerabschätzung für das Differenzenverfahren zur Lösung Partieller Differengleichungen, Z. Agnew Math. Phys., 10 (1930), 373-382.

16. Kellogg, R. B., Analysis of a difference approximation for a singularly perturbed problem in two dimensions, in: Proc. B.A.I.L. I, Dublin, 113-117, 1980.

17. Kellogg, R. B. and Tsan, A., Analysis of some difference approximations for a singular perturbation problem without turning points, Math. Comp., 32 (1978), 1025-1039.

18. Kelley, D. W., Nakazawa, S. and Zienkiewicz, O. C., A note on upwinding in finite element approximation to convection-diffusion problems, Int. J. Numer. Methods Eng., 15 (1980), 1705-1711.

19. Layton, W., Estimates away from a discontinuity for dissipative Galerkin methods for hyperbolic equations, Math. Comp., 36 (1981) 87-92.

20. Layton, W., Galerkin methods for two-point boundary value problems for first order systems, SIAM. J.N.A., 20 (1983), 161-171.

21. Layton, W., High accuracy finite element methods for positive symmetric systems, to appear in: Comp. and Math. w. Appls.

22. Layton, W. and Morley, T. D., On central difference approximations to general second order elliptic equations, preprint, 1986.

23. Nävert, U., A finite element method for convection-diffusion problems, C.S. Dept., Chalmers Inst. of Tech., Göteborg, Sweden, (Ph.D. Thesis), 1982.

24. Raithby, G., Skew upstream differencing for problems involving fluid flow, Comp. Meth. Appl. Mech. Eng., 9 (1976), 153-164.

25. Roos, H. G., Necessary convergence conditions for upwind schemes in the two-dimensional case, Int. J. Num. Meths. in Eng., 21 (1985), 1459-1469.

26. Varga, R. S., Matrix Iterative Analysis, Prentice Hall, Inc., Englewood Cliff, N.J., 1962.

27. Wahlbin, L., A dissipative Galerkin method applied to some quasilinear hyperbolic equations R.A.I.R.O., 8 (1974), 109-117.

28. Wahlbin, L., A dissipative Galerkin method for the numerical solution of first order hyperbolic equations, in: Math. Aspects of Finite Elts. in P.D.E.'s, Academic Press.

29. Hemker, P. W., An accurate method without directional bias for the numerical solution of a 2-D elliptic singular perturbation problem, in: Thy. and Appls. of Sing. Perts. (eds. W. Eckhaus and E. M. de Jaeger), Springer L.N.M., v. 942, Springer, Berlin, 1982.

30. Hemker, P. W., Mixed defect correction iteration for the accurate solution of the convection diffusion equation, in: Multigrid Methods (W. Hackbusch and U. Trottenberg, editors), Springer, Berlin, 1982.

31. Pearson, C. E., On a differential equation of boundary layer type, J. Math. Phys., 47 (1968), 134-154.

32. Glassman, I., Combustion, Academic Press, New York, 1977.

$u(x,y) = \sin \pi x \, \sin \pi y \left(e^{(x-1)/\varepsilon} + e^{-1/\varepsilon}, e^{x/\varepsilon} \right)$
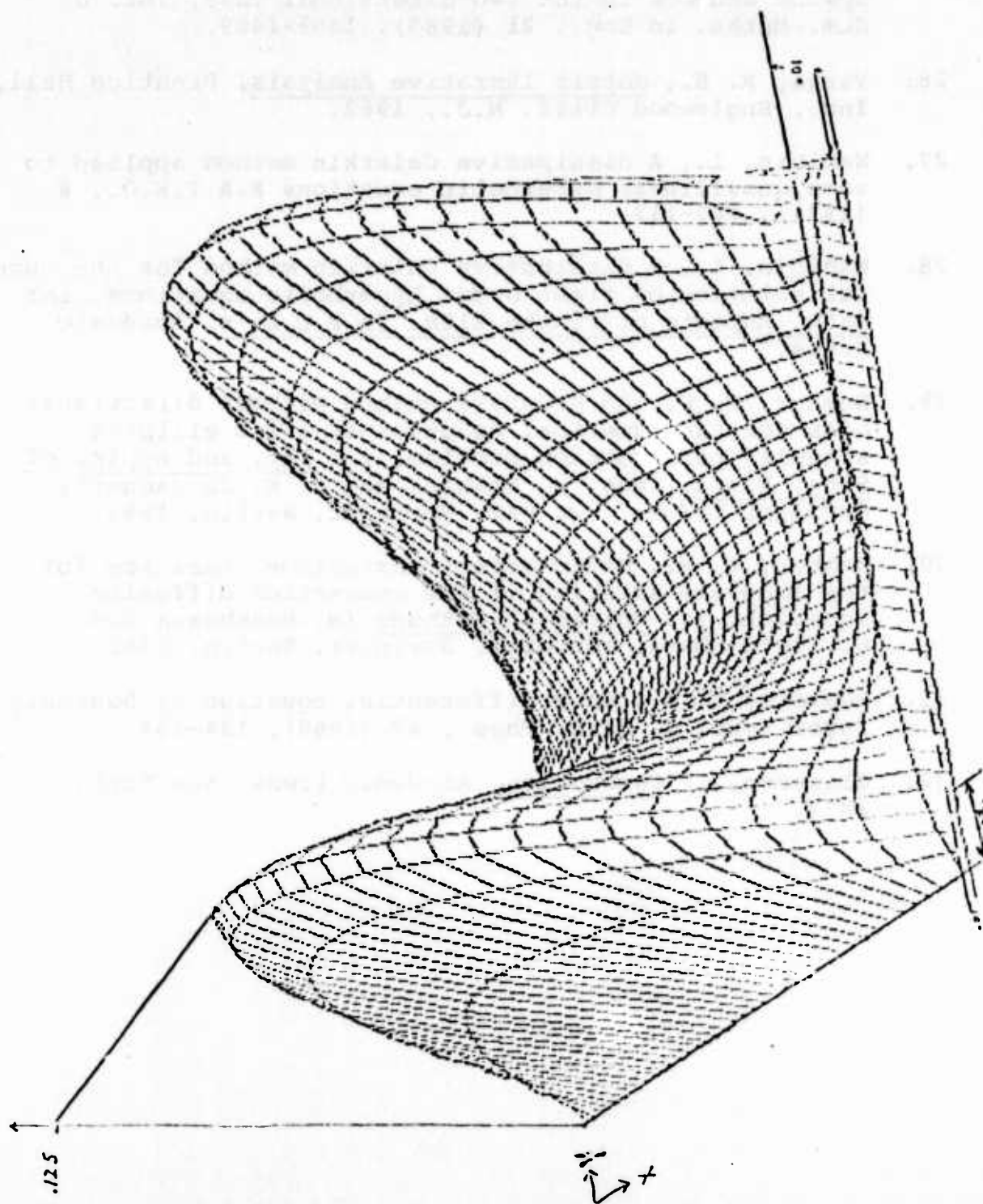
Fig 1: Bilinear interpolant of the true solution $u(x,y)$ to (3.3) with $\varepsilon = 0.01$.
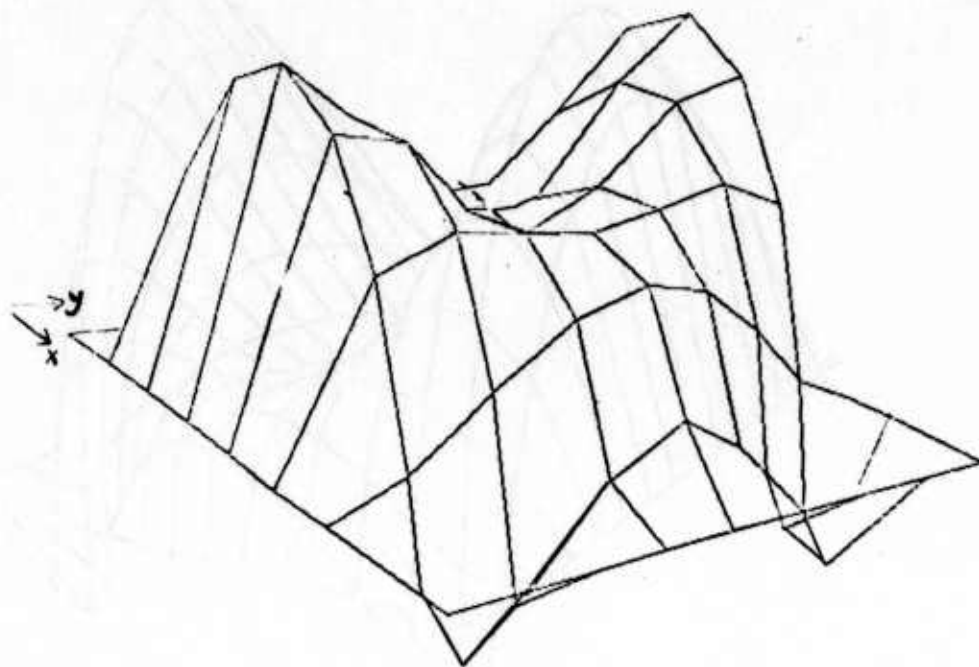
Fig 2. Approximate solution to (3.2) after 4th iterate: U_{ij}^4. Here $h = \frac{1}{7}$, $\varepsilon = 0.01$ and the bilinear interpolant is pictured.
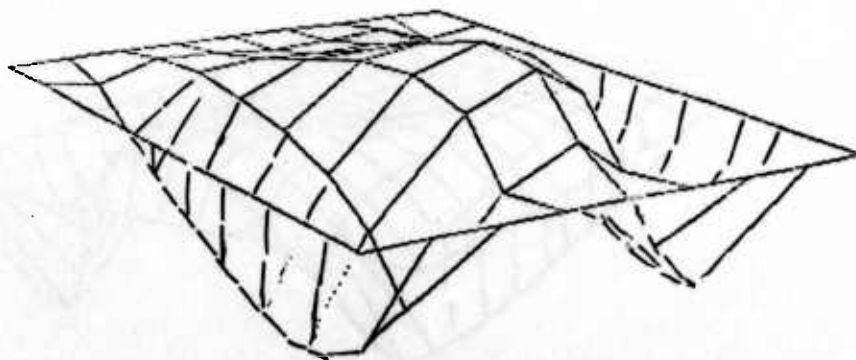
Fig 3. The error in the approximation to (3.3) after 4th iterate. Here $h = \frac{1}{7}$ and $\varepsilon = 0.01$.

1197
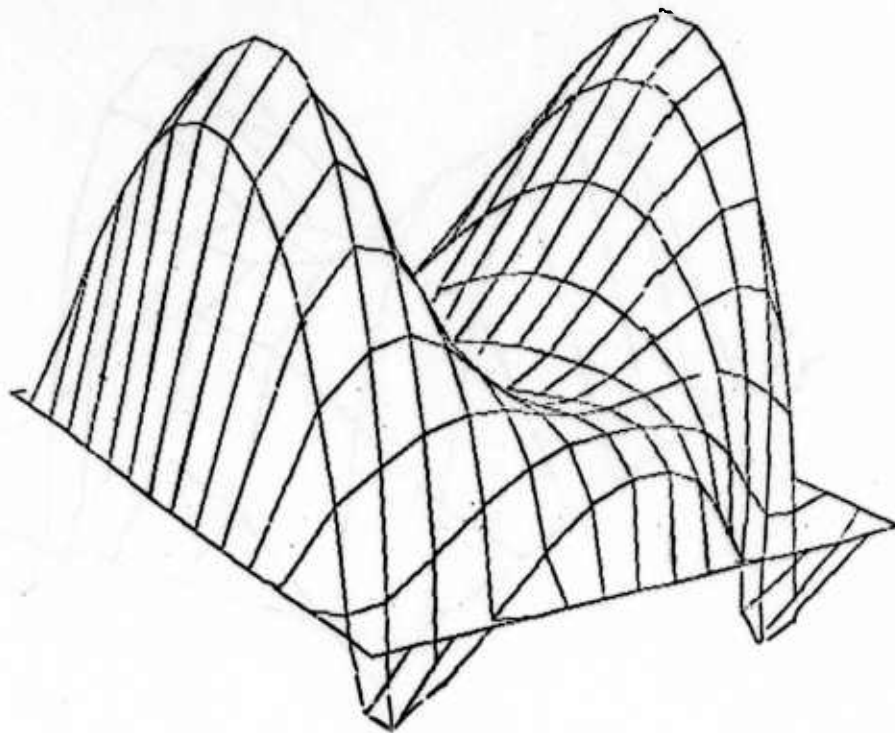
Fig. 4. Approximate solution to (3.2) after 4th iterate. Here $h = \frac{1}{14}$ and $\varepsilon = 0.01$.
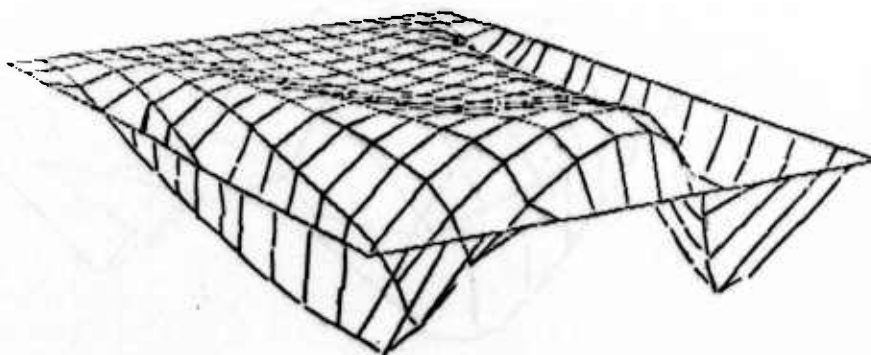
Fig. 5. The error in the approximation to (3.3) after 4th iterate. Here $h = \frac{1}{14}$ and $\varepsilon = 0.01$. 1198

Y ON THE 2 ITERATION :

X-AXIS

Y-AXIS

Fig 7: $\varepsilon y'' + |x-0.5|y' + (2x-1.5)^3 y = 0$
$y(0)=1, \quad y(1)=2$
$\varepsilon = 0.25E-4, \quad h = 0.01$

Y ON THE 2 ITERATION

X-AXIS

Y-AXIS
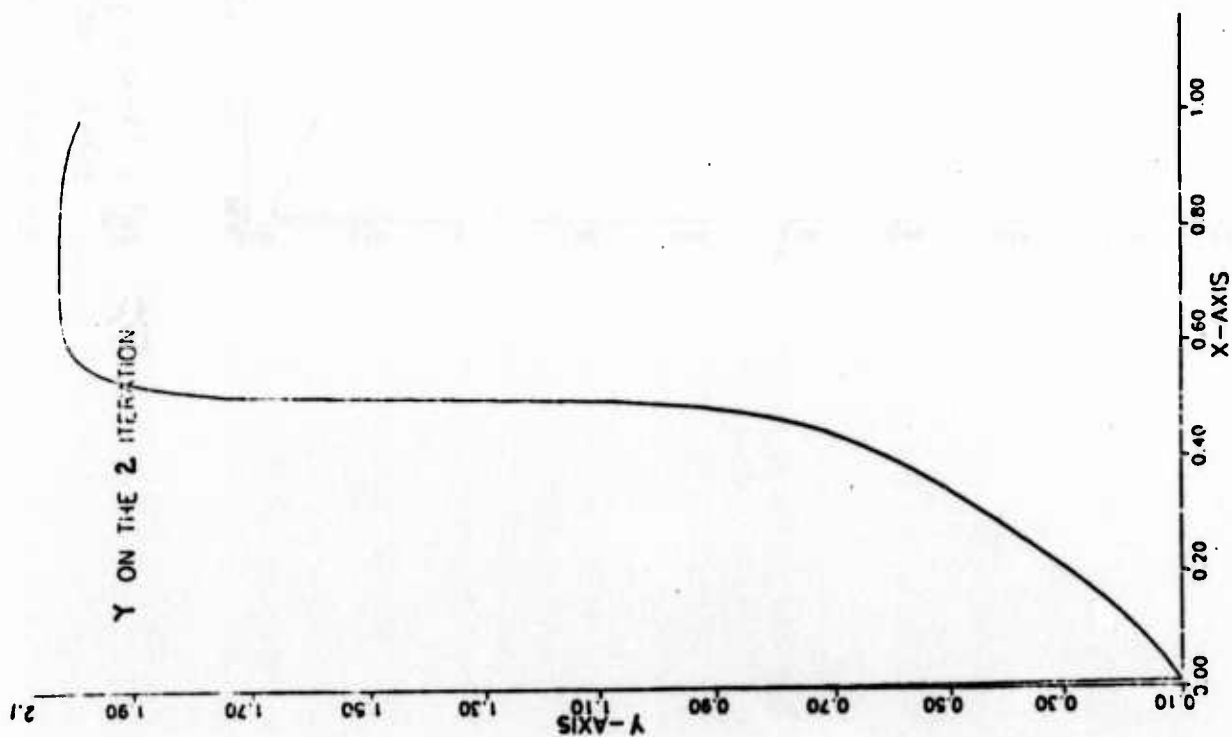
Fig 6: $\varepsilon y'' + (2x^2 - 2x + 0.25)y' + (2x-1)y = 0$
$y(0)=1, \quad y(1)=2$
$\varepsilon = 0.25E-1, \quad h = 0.01$

1199
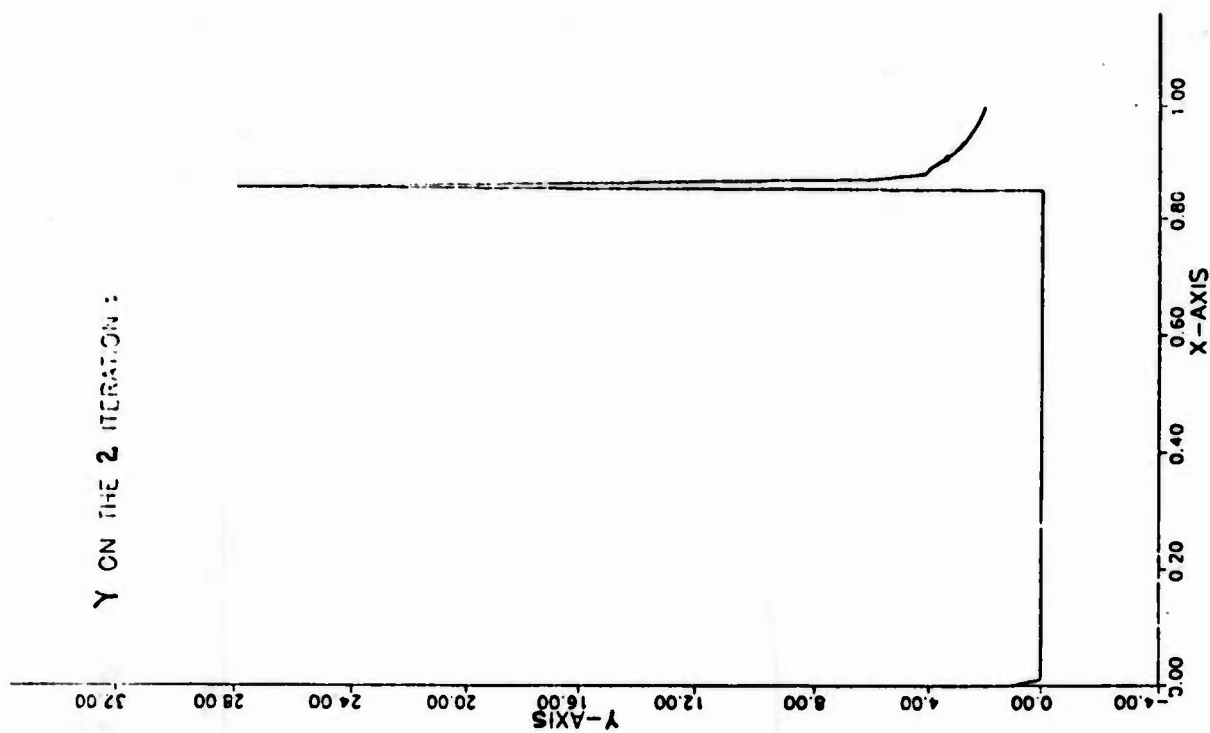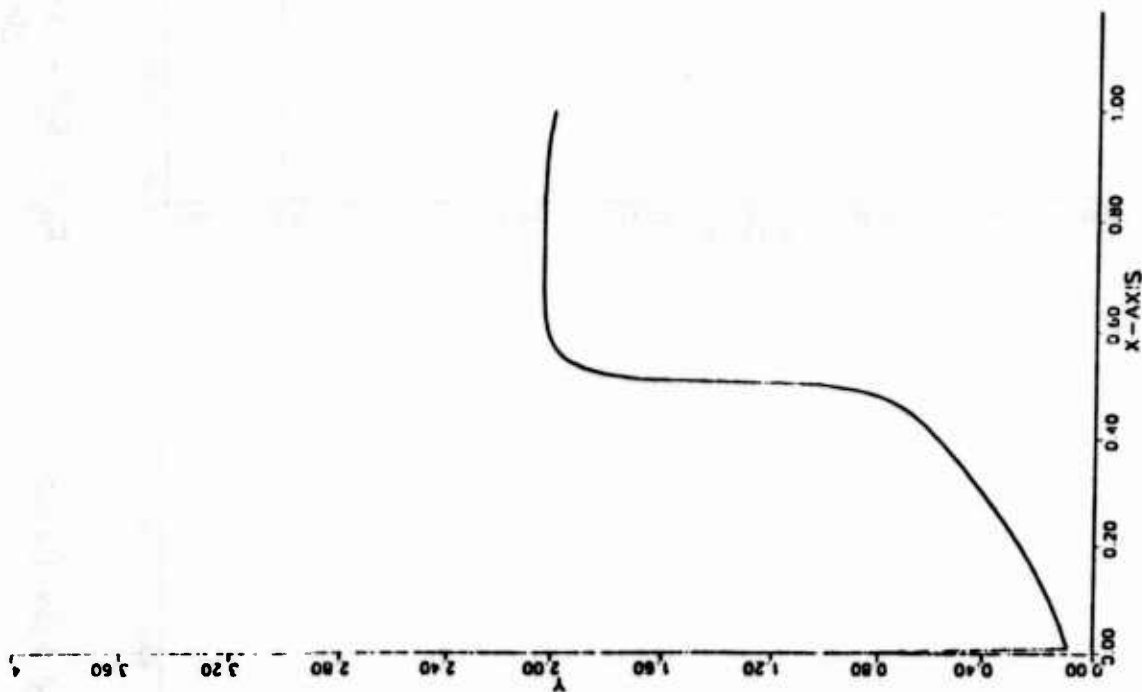
Fig.8 $\varepsilon y'' + |x-0.5| y' + (2x-1.5)^3 y = 0$
$y(0)=1$, $y(1)=2$
$\varepsilon = 0.25E\text{-}4$, $h = 0.01$.

Fig 10: $\varepsilon y'' + (2x^2 - 2x + 0.25)y' + (2x-1)y = 0$
$y(0) = 1$, $y(1) = 2$.
$\varepsilon = 0.25E-4$, $h = 0.25E-2$

Fig 9: $\varepsilon y'' + (2x^2 - 2x + 0.25)y' + (2x-1)y = 0$
$y(0) = 1$, $y(1) = 2$
$\varepsilon = 0.25E-4$, $h = 0.01$

1201

A PLANE PREMIXED FLAME PROBLEM WITH TWO-STEP KINETICS:

EXISTENCE AND STABILITY QUESTIONS

Cl. Schmidt-Lainé

CNRS - Département M.I.S. Ecole Centrale de Lyon

69131 ECULLY Cedex - France

Abstract : We consider a nonlinear differential system modelling a two-step reaction in a plane premixed flame. The unknowns are two functions u and v (temperature and mass fraction) and a parameter δ associated with the burning rate.

For the existence question, we introduce a normalized problem which is first studied on a bounded interval. Upper and lower solutions induce a priori estimates which enable us to pass to the limit of a doubly infinite interval. We obtain the existence of a solution, and we provide an explicit value for δ which is related to the L^2-norm of $w = u-v$.

Of special interest in the behaviour of the system in a neighbourhood of the space variable bound $+\infty$. An autonomous and 2nd order homogeneous system approximates it here, for which the boundary condition $0 \in \mathbb{R}^4$ appears to be a degenerate fixed point. The problem is embedded in the more general framework of the stability of the equilibrium point $0 \in \mathbb{R}^n$ for a second order homogeneous system of dimension n. The homogeneity property allows to reduce the dimension of the system by means of a change of the unknowns and variable. The stationary points of the reduced system are usually hyperbolic, and their asymptotic analysis can be lift back to get a stability theorem. These results are illustrated by the analysis of the combustion problem. For a special value of a physical real parameter, a bifurcation phenomenon occurs.

I. Physical framework

In a recent paper [4], we introduced a two-step irreversible reaction for a steady plane flame, with chain-branching /chain-breaking kinetics :

(1) $A + X \to 2 X$, $2 X + M \to 2 P + M$

Radical X is obtained in the production step, which has a very large activation energy θ, and provides product P in the recombination step for which the activation energy is taken to be zero ; A is the reactant and M a third body.

This two-step sheme is presented as an alternative to classical single-step kinetics and allows the description of a wider range of phenomena.

The equations are derived in the stretched flame zone, described by the one-dimensional space variable η , $-\infty < \eta < +\infty$, by assuming a fast recombination, i.e. that both production and recombination of radicals take place in the same thin zone. The system reads (see [4 p. 423]).

(2) $\begin{cases} u'' = q_1 \; r \; \delta(u-v)v \; e^{-u} + \delta(u-v)^2 \\ v'' = r \; \delta(u-v)v \; e^{-u} \end{cases}$

and the associated boundary conditions are :

(3) $\begin{cases} u = -\eta + o(1) \;, \; v = -\eta + o(1) \text{ as } \eta \rightarrow -\infty \\ u = o(1) \;, \; v = o(1) \text{ as } \eta \rightarrow +\infty \end{cases}$

The unknowns are positive functions u and v, representing temperature and mass fraction of the reactant and a positive constant δ, representing the burning rate. The parameters q_1 and q_2 are the proportions of the total heat released in the first and second steps of the reaction, so that $q_1 + q_2 = 1$. Physical considerations require the recombination step to be exothermic, so that $q_2 > 0$.
Finally r is a positive parameter, corresponding to the ratio of the two reaction rates. The boundary conditions (3) are obtained by matching with expansions on either side of the flame sheet.

The mathematical problem is the following : q_1, q_2 and $r > 0$ being given, find two functions $u > 0$, $v > 0$ and the constant $\delta > 0$ satisfying (2) (3). We refer to [4] for a numerical treatment of this problem, leading to curves in the (r,δ) - plane.

It is particularly convenient for our study to deal with an equivalent formulation of the system involving the radical mass fraction $w = u - v > 0$.

It consists of both systems :

(4) $\begin{cases} v'' = r \; \delta \; v \; we^{-v} \; e^{-w} \\ w'' = - q_2 \; r \; \delta \; v \; w \; e^{-v} \; e^{-w} + \delta \; w^2 \end{cases}$

and

(5) $\begin{cases} u'' = q_1 \; r \; \delta(u-w)w \; e^{-u} + \delta w^2 \\ w'' = - q_2 \; r \; \delta(u-w)w \; e^{-u} + \delta \; w^2 \end{cases}$

together with the boundary conditions :

(6) $\begin{cases} v = -\eta + o(1) \;, \; w = o(1) \text{ as } \eta \rightarrow -\infty \\ v = o(1) \;, \; w = o(1) \text{ as } \eta \rightarrow +\infty \end{cases}$

and

(7) $\begin{cases} u = -\eta + o(1) \;, \; w = o(1) \text{ as } \eta \rightarrow -\infty \\ u = o(1) \;, \; w = o(1) \text{ as } \eta \rightarrow +\infty \end{cases}$

II. The existence question [1], [2], [5]

Let us introduce the following problem in the x-variable, which is obtained from (4) by taking δ equal to 1 :

$$(8) \quad \begin{cases} \dfrac{d^2 v}{dx^2} = r v \, w \, e^{-v} \, e^{-w} \\[2mm] \dfrac{d^2 w}{dx^2} = - q_2 \, r v \, w \, e^{-v} \, e^{-w} + w^2 \end{cases}$$

We consider this nonlinear system subject to the following boundary conditions :

$$(9) \quad \begin{cases} v' \to 0 \quad \text{as} \quad x \to + \infty \\ w \to 0 \quad \text{as} \quad x \to \pm \infty \end{cases}$$

together with the normalization condition

$$(10) \quad v(0) = 1.$$

We prove the following theorem :

Theorem 1 : For any $r > 0$ fixed, there exists a solution $(v,w) \in (C^\infty(\mathbb{R}))^2$ to problem (8), (9),(10) such that

$$(11) \quad 0 < w < M_0 = q_2 \, r/e$$

$$(12) \quad v > 0 \; ; \; v' < 0 \; ; \; v \to + \infty \quad \text{as} \quad x \to - \infty \, , \, v \to 0 \quad \text{as} \quad x \to + \infty \, .$$

Moreover, $w \in H^2(\mathbb{R})$ and, as $x \to - \infty$, $v(x) = - \ell(x - x_0) + \text{E.S.T.}$, where $x_0 \in \mathbb{R}$ and

$$(13) \quad \ell = \frac{1}{q_2} \int_{-\infty}^{+\infty} w^2(x) \, dx = \frac{1}{q_2} \, ||w||^2_{L^2(\mathbb{R})} \, . \quad \blacksquare$$

To return to the initial problem (4) (6), we make the transformation

$$(14) \quad \eta = \ell (x - x_0)$$

Then (8) becomes

$$(15) \quad \begin{cases} \dfrac{d^2 v}{d\eta^2} = \dfrac{1}{\ell^2} \, r v \, w \, e^{-v} \, e^{-w} \\[2mm] \dfrac{d^2 w}{d\eta^2} = - \dfrac{1}{\ell^2} \, q_2 \, r v \, w \, e^{-v} \, e^{-w} + \dfrac{1}{\ell^2} \, w^2 \end{cases}$$

and, as $\eta \to - \infty$, $v(\eta) = - \eta + \text{E.S.T.}$ Thus it is clear that (4) (6) is solved by the pair $(v(\eta), w(\eta))$ with

$$(16) \quad \delta = 1/\ell^2$$

Therefore, we have determined

$$(17) \quad \delta = q_2 \, / \int_{-\infty}^{+\infty} w^2(\eta) \, d\eta \, .$$

In [2], we give a detailed demonstration of Theorem 1. A sketch of the proof is the following : First, we exhibit positive upper and lower solutions of the problem in a formal way. Next, we consider the system (8) on a bounded interval $[- a, b]$ with suitable boundary conditions and we prove the exis-tence of a solution (v,w) in a closed convex set K, by a fixed point argument (the convex K involves the upper and lower solutions). We derive a priori es-timates in the C^1 - norm as well as in the H^1 - norm. Finally, we let a and b

tend to $+ \infty$, and prove the existence of a limit which solves (8) (9) (10). Properties (12), (13) appear as a "spin off" of the existence proof. An alternative proof by a topological shooting method has been presented by S.P. Hastings, C. Lu and Y.H. Wan [3].

III. The stability question [5], [6]

System (5) can be rewritten in the canonical form :

(18) $\begin{cases} u' = p \\ p' = q_1 \ r \ \delta(u-w)we^{-u} + \delta w^2 \\ w' = q \\ q' = - q_2 \ r\delta \ (u-w)we^{-u} + \delta w^2 \end{cases}$

for which $0 \in \mathbb{R}^4$ appears to be a fixed point, derived from the boundary conditions (7). By linearizing (18) about this point, it comes 4 zero eigenvalues ; such a degenerate fixed point requires more sophisticated treatment. Let us then consider the second order approximation of (18) near this point :

(19) $\begin{cases} u' = p \\ p' = q_1 \ r \ \delta(u-w) w + \delta w^2 \\ w' = q \\ q' = - q_2 \ r \ \delta(u-w) w + \delta w^2 \end{cases}$

This system only contains quadratic terms and can be considered in the general form of second order homogeneous problem.

So, we consider the problem

(20) $\dfrac{d^2 X}{d\eta^2} = X'' = F(X)$

where $X \in \mathbb{R}^n$ and F is a second order homogeneous function. More precisely, we study the stability of the degenerate fixed point $0 \in \mathbb{R}^{2n}$ of the autonomous dynamical system :

(21) $\begin{cases} X' = Y \\ Y' = F(X) \end{cases}$

The homogeneity property allows to reduce the dimension of problem (21). By means of the change of functions :

(22) $x = \dfrac{X}{||X||} \ ; \ y = \dfrac{Y}{||X||^{3/2}}$

such that $(X,Y) \longrightarrow (x,y)$
$\in \mathbb{R}^n \times \mathbb{R}^n \qquad \in S^{n-1} \times \mathbb{R}^n$

where S^{n-1} is the unit sphere of $\mathbb{R}^n$, together with the change of variable defined by

(23) $\dfrac{ds}{d\eta} = ||x||^{1/2}$

1206

the reduced autonomous dynamical system is the following :

$$(24) \quad \begin{cases} \dfrac{dx}{ds} = y - x(x.y) \\ \dfrac{dy}{ds} = F(x) - \dfrac{3}{2} y(x.y) \end{cases}$$

where (.) denotes the euclidian scalar product in $\mathbb{R}^n$, and $||.||$ the associated norm.

Then we prove existence of at least a couple of symmetric fixed points of (24) : $P_i = (x_i, y_i) \in S^{n-1} \times \mathbb{R}^n$ and exhibit the mapping between trajectories of (21) and trajectories of (24) ; an essential remark is that a trajectory of (24) provides by lifting a <u>family</u> of trajectories of (21), due to an arbitrary integration parameter ; the stability result is the application of the former property to invariant manifolds :

Theorem 2 : Let $P_o = (x_o, y_o) \in S^{n-1} \times \mathbb{R}^n$, $n \geq 1$, be a fixed point of (24), such that $(x_o . y_o) < 0$. Then

(i)   The stable manifold $W^s(P_o)$ of problem (24) lifts to the stable manifold $W^s(0)$ of homogeneous problem (21).

(ii) $\dim W^s(0) = \dim W^s(P_o) + 1$. ∎

The main interest of this result is that investigation of the stable manifold $W^s(P_o)$ of problem (24) is often accessible by classical tool, namely by linearization, because the fixed points of (24) are usually hyperbolic. A complete stability analysis of fixed points P_o of (24) is presented in [6], leading to results on the dimension of the stable manifold $W^s(0)$ of (21). These results are summarized as follows :

Theorem 3 : Let us suppose that $F'(x_o)$ is similar to a diagonal matrix and that its eigenvalues λ_j are such that $\lambda_j = a_j + ib_j$:

$$(25) \quad \begin{cases} 6b_j^2 \neq 25 \, \rho \, (\rho - a_j) \\ \rho = \dfrac{3}{2} (x_o . y_o)^2 \end{cases}$$

Then by lifting of $W^s(P_o)$ to $W^s(0)$

$$(26) \quad \dim W^s(0) = \sum_{j \in J} m(\lambda_j)$$

where

$$J = \{ j \in \mathbb{N} \, / \, 6b_j^2 > 25 \, \rho \, (\rho - a_j) \}$$

and $m(\lambda_j)$ is the geometrical multiplicity of the eigenvalue λ_j . ∎

Applying this general method to (19) leads to a reduced system for which two fixed points P_o and P_1 are obtained, verifying $(x_o . y_o) < 0$, $(x_1 . y_1) < 0$. As $r = 1$, the two fixed points are no more distinct and this unique fixed point is degenerate. For $r \neq 1$, we obtain by linearization and lifting to (19) :

$$(27) \quad \dim W^s(0) = 2$$

The transcritical bifurcation case $r = 1$ is solved by means of the Center Manifold Theorem.

1207

References

[1] C.M. Brauner, Cl. Schmidt-Lainé,
Existence d'une solution pour le problème de la flamme plane prémé-
langée avec cinétique à deux pas, C.R. Acad. Sc. Paris, Série I, 301,
(1985), 667–670.

[2] C.M. Brauner, Cl. Schmidt-Lainé,
Existence of a solution to a certain plane premixed flame problem
with two-step kinetics. Subm. to SIAM J. Math. Anal.

[3] S.P. Hastings, C. Lu, Y.H. Wan,
A three dimensional shooting method as applied to a problem in com-
bustion theory. To appear in Physica D.

[4] G. Joulin, A. Liñan, G.S.S. Ludford, N. Peters, Cl. Schmidt-Lainé,
Flames with chain Branching/chain breaking kinetics, SIAM J. Appl.
Math, 45 (1985), 420–434.

[5] Cl. Schmidt-Lainé,
Sur quelques problèmes non linéaires en Mécanique des Fluides, Chimie,
et Combustion, Thèse d'Etat, Université Lyon I, 1985.

[6] Cl. Schmidt-Lainé, D. Serré,
Etude de stabilité d'un système non linéaire de dimension 4 en combus-
tion et généralisation à une classe de problèmes homogènes de degré 2,
to appear in Physica D.

Controlling Thermal Runaway in Catalytic Pellets

Jagdish Chandra
U.S. Army Research Office
Box 12211
Research Triangle Park, NC 27709

Paul Davis
Mathematical Sciences Dept.
Worcester Polytechnic Inst.
Worcester, MA 01609

Bistable response is common in many situations; one is the equlibrium temperature in a catalyst particle [5], as illustrated schematically in figure 1. The surface temperature of the pellet is the control parameter which chooses between the high and low temperature branches. Numerical calculations [4,5] and formal asymptotic [2] studies of a simplified model have shown that oscillating this control parameter at a sufficiently high frequency permits the pellet temperature to remain on the lower branch, even in the face of perturbations that otherwise would cause an undesireable jump to the higher branch.

We derive similar conclusions from a rigorous differential inequality analysis. This analysis also reveals that periodic oscillations are not necessary; any sort of oscillation suffices, provided its time integral is sufficiently small. These results are briefly sketched here. Complete details and their extension to a general class of problems will appear in [1].
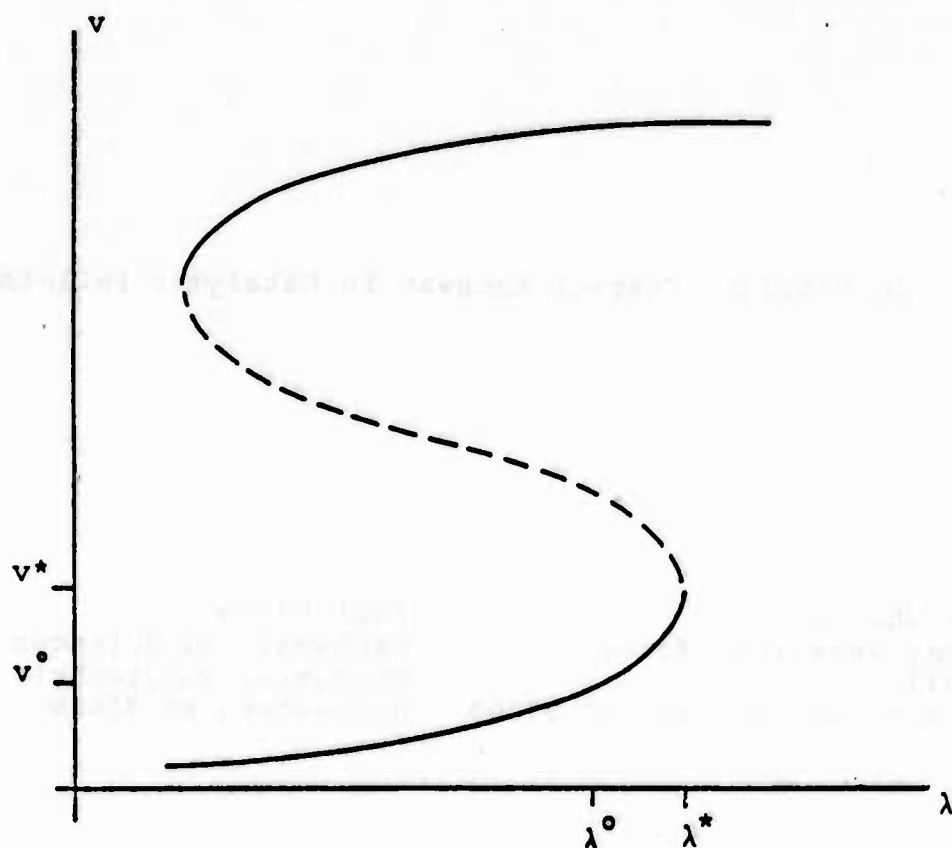
Figure 1: Equilibrium temperature v in a catalyst particle vs. pellet surface temperature λ. The solid lines are stable, the dashed unstable. The point (v^o, λ^o) is a stable low temperature operating point, and (v^*, λ^*) is the maximum low temperature operating point.

A simplified model of the catalyst particle is

$$v'(t) = \lambda - g(v),$$

where v is the spatially uniform temperature of the pellet's interior and λ is its surface temperature; see [2,5]. The graph of g is sketched in figure 2. Its form reflects the multiplicity of states possible from the nonlinear interaction of pellet temperature and reactant kinetics.
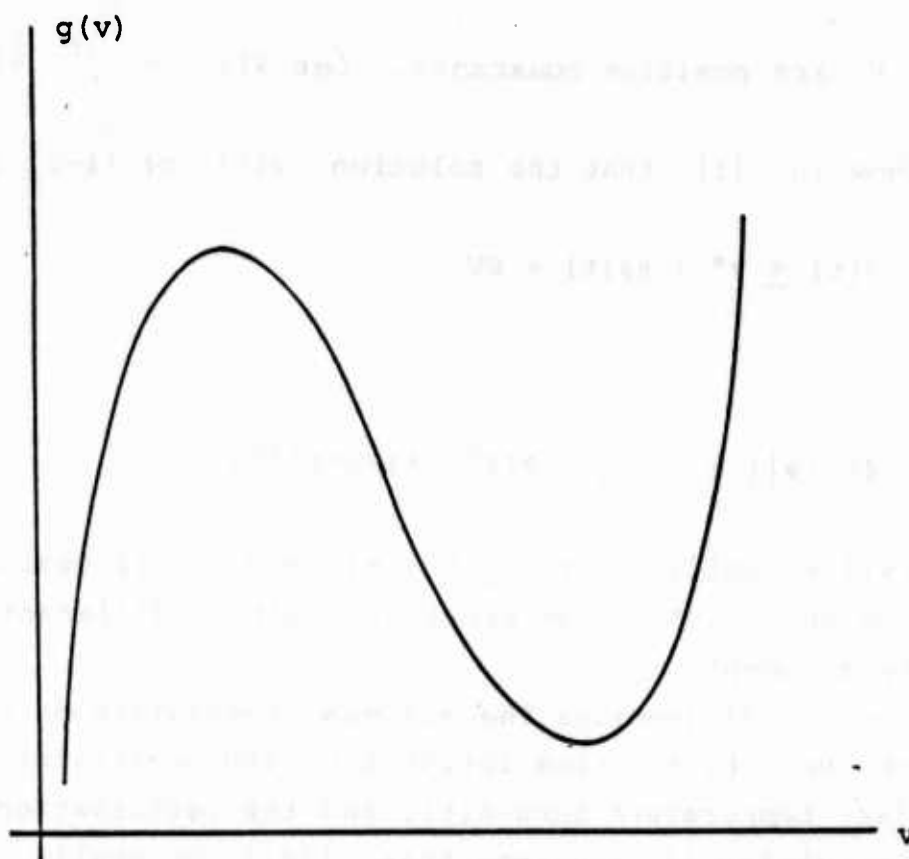
Figure 2: The nonlinear term $g(v)$ in (1) vs. v.

The operating point (v°, λ°) shown in figure 1 is a low temperature state; (v^*, λ^*) is the maximum low temperature operating point. The primary result of our analysis is that substantial perturbations of the lower equilibrium temperature v° can be bounded away from the corresponding higher equilibrium value v_o shown in figure 1 if the boundary temperature λ is oscillated about λ°.

Specifically, define $\Lambda(t) = \lambda^\circ + \beta\phi(t)$ for some arbitrary function ϕ. Let $v(t)$ denote the solution of the perturbed system

(1) $v'(t) = \Lambda(t) - g(v)$,

(2) $v(0) = v^\circ + \beta V$,

1211

where β, V are positive constants. Let $\psi(t) = \int^t \phi(s) \, ds$.

Then we show in [1] that the solution $v(t)$ of (1-2) satisfies

(3) $v(t) \le v^* + \beta\psi(t) + \beta V$

provided

(4) $\beta(||\psi|| + V) \le [2(\lambda^o - \lambda^*)/g'']^{1/2}$.

Here, $||\psi|| = \sup\{|\psi(t)| : 0 \le t < \infty\}$ and g'' is evaluated at its maximum on $[v, v^*]$. The proof is a simple differential inequality argument.

The bound (3) involves the maximum temperature on the low temperature branch, the time integral of the oscillatory part of the boundary temperature term $\Lambda(t)$, and the perturbation βV.

The condition (4) requires that $||\psi||$ be small. If the oscillations in the surface temperature are sinusoidal, e.g., if $\phi(t) = \sin \omega t$, then $||\psi|| \sim 1/\omega$ and (4) requires that the frequency be large.

This suggestion that there is a critical lowest frequency which stabilizes the low temperature branch is consistent with the numerical and asymptotic evidence [5,2]. Indeed, for $0 < \beta \ll 1$, Cohen and Matkowsky [2] found an expression for this critical frequency that is similar in form to (4).

However, (4) can certainly be satisfied when ϕ is other than sinusoidal; ϕ need not even be periodic. Any sort of oscillatory strategy will suffice so long as (4) holds.

These ideas are extended in [1] to a general class of problems of the form

$$v' = f(v, \lambda),$$

where f is only required to exhibit a local equilibrium that loses stability for λ sufficiently large. Complete proofs and applications to other physical problems, such as stabilizing phase transitions [6,7], appear in [1] as well.

REFERENCES

1.   J. Chandra and P. W. Davis, Stabilizing spatially homogeneous steady states, to appear

2.   D. S. Cohen and B. J. Matkowsky, On inhibiting runaway in catalytic reactors, SIAM J. Appl. Math. <u>35</u> (1978), 307-314

3.   V. Lakshmikantham and S. Leela, Differential Inequalities, v. 1, Academic Press, New York, 19

4.   C. McGreavy and J. M. Thornton, Generalized Criteria for the stability of catalytic reactors, Can. J. Chem. Eng. <u>48</u> (1970), 187-191

5.   C. McGreavy and J. M. Thornton, Stability studies of single catalyst particles, Chem. Eng. J. <u>1</u> (1970), 296-301

6.   P. H. Richter, I. Procaccia, and J. Ross, Chemical Instabilities, Adv. Chem. Phys. <u>48</u> (1980), 217-268

7.   F. Schlogl, Chemical reaction models for non-equilibrium phase transitions, Z. Physik <u>253</u> (1972), 147-161

PROPAGATION OF A PLANE, ADIABATIC FLAME THROUGH A MIXTURE

WITH A TEMPORAL ENTHALPY GRADIENT

A.K. Kapila and G. Ledder
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

ABSTRACT. The unsteady response of a plane, adiabatic flame to a temporal gradient in the enthalpy of the reacting medium is studied. The characteristic time of the gradient is taken to be of the same order as the natural time scale of the flame. The mathematical formulation leads to a moving boundary problem which must be treated numerically. The results show how the burning rate responds to variations in Lewis number, amplitude of the gradient, and characteristic time of the gradient.

I. INTRODUCTION. When a plane flame propagates through a combustible medium whose state is uniform, it does so at a constant speed. In many practical applications, however, the state of the fresh mixture exhibits spatial and/or temporal nonuniformities. These nonuniformities may occur, for example, in temperature, reactant concentration, or both. The flame will then propagate in an unsteady fashion.

The flame response depends crucially upon how the natural scales of the flame (i.e. the diffusion length and time scales) compare with the characteristic scales of the nonuniformity. If the scales of the nonuniformity are relatively long, the framework of Slowly-Varying Flames (SVFs) applies (see [1], Chap. 3). That problem is the subject of recent work by Bissett and Reuss [2], who have undertaken an analysis in the limit of large activation energy ($\theta \to \infty$). They assumed that enthalpy variations in the fresh mixture have a characteristic length θ times larger than the flame thickness, and amplitude $O(1/\theta)$ relative to that of the undisturbed state; it is well-known that $O(1/\theta)$ fluctuations in the flame temperature can lead to $O(1)$ fluctuations in the burning rate. Their analysis leads to an ordinary differential equation for the time variation of the burning rate. A study of this equation for Lewis numbers less than unity (the stable regime for planar flames subject to planar perturbations) reveals that the flame exhibits a delayed response to the enthalpy fluctuations, but that this sluggishness disappears as the Lewis number approahces unity. Bissett and Reuss also considered the effect of heat loss, since they were particularly concerned with the unsteady behavior of the flame near extinction.

The lively response of the flame for near-unity Lewis numbers is also confirmed by the analysis of Mikolaitis [3], who treats $O(1)$ variations in upstream enthalpy, and hence exponentially large fluctuations in burning rate. Only positive enthalpy gradients are considered, and the

characteristic scale of the nonuniformity is taken to be the same as the initial flame thickness. However, the problem quickly falls into the SVF mold, since the flame thickness shrinks exponentially as the flame encounters an O(1) increase in upstream enthalpy. Mikolaitis finds that the flame adjusts 'instantaneously' to the local state ahead of the preheat zone.

In the SVF limit, the flame retains its quasisteady structure, and that is what makes the problem analytically tractable. This is no longer the case when one considers $O(1/\theta)$ perturbations with a characteristic scale comparable to the scale of the flame. Now the burning rate variation is only O(1), so that Mikolaitis' analysis does not apply. In fact, O(1) time variations intrude into the preheat zone, and the flame becomes genuinely unsteady. Further analytical progress is not possible and one must resort to numerics.

This paper is concerned with purely temporal nonuniformities in the state of the medium, occuring on a time scale comparable to the time the flame takes in travelling a distance equal to its thickness. Spatial stratification is not considered here; it requires a different treatment, in contrast to the SVF framework where spatial and temporal fluctuations are equivalent. The aim here is to determine the burning-rate response of the flame as a function of Lewis number, fluctuation amplitude, and fluctuation time scale. The mathematical problem, it turns out, involves a moving boundary, but it can be solved by standard numerical techniques.

2. GOVERNING EQUATIONS. It is convenient to adopt the Near-Equidiffusional Flame (NEF) formulation (see [1], Chap. 3) which can be derived from the full combustion equations with Arrhenius kinetics under the assumptions of large activation energy ($\theta \to \infty$), near-unity Lewis number and nearly-uniform enthalpy, i.e.,

$$L^{-1} = 1 - \epsilon \ell/\alpha, \quad H \equiv T + \alpha Y = 1 + \alpha + \epsilon h,$$

$$\epsilon = (1 + \alpha)^2/\theta \to 0. \tag{1}$$

It is also convenient to employ a density-weighted spatial coordinate travelling with the flame. Then, to leading order in ϵ, the governing equations are

$$T_t + M T_x = T_{xx} \quad \text{for} \quad x < 0, \quad T = 1 + \alpha \quad \text{for} \quad x > 0, \tag{2a}$$

$$h_t + M h_x = h_{xx} + \ell T_{xx} + S(t) \quad \text{for} \quad x \gtrless 0, \tag{2b}$$

$$S(t) = \dot{h}_f(t), \tag{2c}$$

with boundary conditions

$$T \to 1, \quad h \to h_f(t) \quad \text{as} \quad x \to -\infty, \quad h_x \to 0 \quad \text{as} \quad x \to \infty, \tag{3}$$

jump conditions

1216

$$\delta T = \delta h = 0, \quad \delta(h_x + \ell T_x) = 0,$$

$$\delta T_x = -\alpha \exp(h/2) \quad \text{at} \quad x = 0, \tag{4}$$

and initial conditions

$$T = 1 + \alpha e^x, \quad h = -\ell x e^x \quad \text{for} \quad x < 0,$$

$$T = 1 + \alpha, \quad h = 0 \quad \text{for} \quad x > 0. \tag{5}$$

In equations (4) above δF is defined as follows:

$$\delta F = F(0+,t) - F(0-,t).$$

Several remarks about the governing equations are in order.

(i) The symbols T, H, M, L and α denote, respectively, the temperature, enthaply, burning rate, Lewis number (ratio of thermal diffusivity to mass diffusivity) and the heat-release parameter. The symbols h and ℓ, defined by equations (1) above, denote a reduced enthalpy and a reduced Lewis number respectively, and represent small departures, measured on the ϵ-scale, from constant values.

(ii) The zero-fluctuation state of the fresh mixture, also the initial state, is chosen as the reference for nondimensionalization, and is given by

$$T = 1, \quad h = 0, \quad M = 1.$$

As already mentioned, the spatial coordinate is density-weighted, and the dimensionless thermal diffusion coefficient is allowed to vary according to the prescription

$$\lambda/C_p = T,$$

where λ is the thermal conductivity and C_p the specific heat. For the sake of brevity details of the nondimensionalization process are omitted here, but are quite standard and can be found, for example, in [4].

(iii) It is assumed that enthalpy of the medium undergoes purely temporal fluctuations with $O(\epsilon)$ amplitude, i.e., far ahead of the flame,

$$H = 1 + \alpha + \epsilon h_f(t).$$

The source term $S(t)$ in equation (2b), defined by (2c), is included to ensure that equation (2b) balances at $x = -\infty$. The enthalpy nonuniformity may be due to fluctuations in temperature T, reactant concentration Y, or both. Also, it is worth noting that since the amplitude of the nonuniformity is $O(\varepsilon)$, the upstream boundary condition for temperature in equation (3) is unperturbed.

In the computations presented below, the enthalpy variation is assumed to be

$$h_f(t) = h_\infty [1-\exp(t/t_f)]^2, \quad t > 0,$$

i.e. the reduced enthalpy of the fresh mixture varies smoothly and monotonically from the value

$$h_f(0) = 0$$

to the value

$$h_f(\infty) = h_\infty .$$

The time constant t_f determines the rate of variation.

It is convenient to define a modifed enthalpy variable according to the prescription

$$\psi = h - h_f$$

which leaves (2a) unchanged and allows the remainder of the system (2) - (4) to be rewritten as

$$\psi_t + M \psi_x = \psi_{xx} + \ell T_{xx} \quad \text{for } x \gtrless 0, \tag{6a}$$

$$T \to 1, \quad \psi \to 0 \quad \text{as } x \to -\infty, \quad \psi_x \to 0 \quad \text{as } x \to \infty, \tag{6b}$$

$$\delta T = \delta\psi = \delta(\psi_x + \ell T_x) = 0, \quad \delta T_x = -\alpha \exp\left[(\psi + h_f)/2\right] \text{ at } x = 0. \tag{6c}$$

The initial conditions (5) remain unchanged, with ψ replacing h.

3. NUMERICS. The goal of this work is to study the behavior of the burning rate $M(t)$ as a function of the parameters ℓ, h_∞ and t_f. The above equations define a moving boundary problem which was treated numerically as follows. The doubly infinite problem was discretized on the

finite interval $[-A,B]$, A, $B > 0$, according to the Crank-Nicholson scheme. At each time step a provisional value of M was assumed and the governing differential equations for T and ψ integrated by using all the jump conditions in (6c) except the last. The remaining condition was then incorporated into a Muller root finder to iterate to the correct value of M.

For each choice of the parameter set (ℓ, h_∞), some experimentation was found to be necessary to determine the numerical boundary locations A and B, but the largest range ever used corresponded to $A + B = 18$. Almost all the calculations employed $\Delta t = 0.20$ and $\Delta x = 0.04$, and needed no more than three Muller iterations for convergence. The accuracy of the numerical scheme was tested by comparing the numerical results with asymptotic analytical results obtained in the limit $t \to 0$, and separately, in the limit $h_\infty \to 0$. In each limiting case the problem linearizes and can be solved analytically by using Laplace transformation.

.4. RESULTS. The numerical results are displayed in Figures 1-3. Each figure displays the effect on M of a single parameter in the triad (ℓ, h_∞, t_f), while the other two are kept fixed. All runs were computed at $\alpha = 4$.

(i) Effect of ℓ

Fig. 1 reveals the effect of variation of ℓ upon $M(t)$. In this figure t_f is fixed at the value unity, and h_∞ at $\ell n\, 4$, so that M varies from 1 to 2 as time increases from 0 to ∞. The dotted curve corresponds to the quasisteady, or instantaneous response. For large and negative values of ℓ the actual flame response lags behind the quasisteady response over most of the time interval. However, the flame becomes livelier as ℓ increases, and eventually, for ℓ sufficiently large and positive, the burning rate overshoots the ultimate value of 2 and a decaying oscillation appears. (The appearance of the oscillation is the precursor to eventual instability of the steady flame in favor of pulsatile motion.)

(ii) Effect of h_∞.

Fig. 2 (a,b,c) are drawn at $t_f = 1$ and $\ell = 0$, and display the variation of flame response with h_∞.

(iii) Effect of t_f.

In Fig. 3, h_∞ and ℓ are set at respective values $\ell n\, 4$ and zero, while t_f is changed from 1 to 1/4. For the shorter value of t_f the burning rate shows an overshoot, indicating that larger enthalpy gradients provoke a stronger response.

REFERENCES

1. J.D. Buckmaster and G.S.S. Ludford, "Theory of Laminar Flames," Cambridge University Press, 1982.

2. E.J. Bissett and D.L. Reuss, "Analysis of a slowly-varying nonadiabatic flame propagating through gradients of fuel or temperature, presented at the 24th International Symposium on Combustion, Munich, August 1986.

3. D.W. Mikolaitis, "The unsteady propagation of premized flames through nonhomogeneous mixtures and thermal gradients", Combustion and Flame, 57, pp. 87-94 (1984).

4. N. Peters and G.S.S. Ludford, "The effect of pressure variation on premixed flames", Combustion Science and Technology, 34, pp. 331-344 (1983).

FIG. 1

$l = 0 \quad t_f = 1 \quad h_\infty = LN(4)$

M_{STEADY}

FIG. 2a

$l = 0 \quad t_f = 1 \quad h_\infty = .2$

FIG. 2b

1223

$\ell = 0 \quad t_f = 1 \quad \hbar_\infty = -LN(4)$

M STEADY
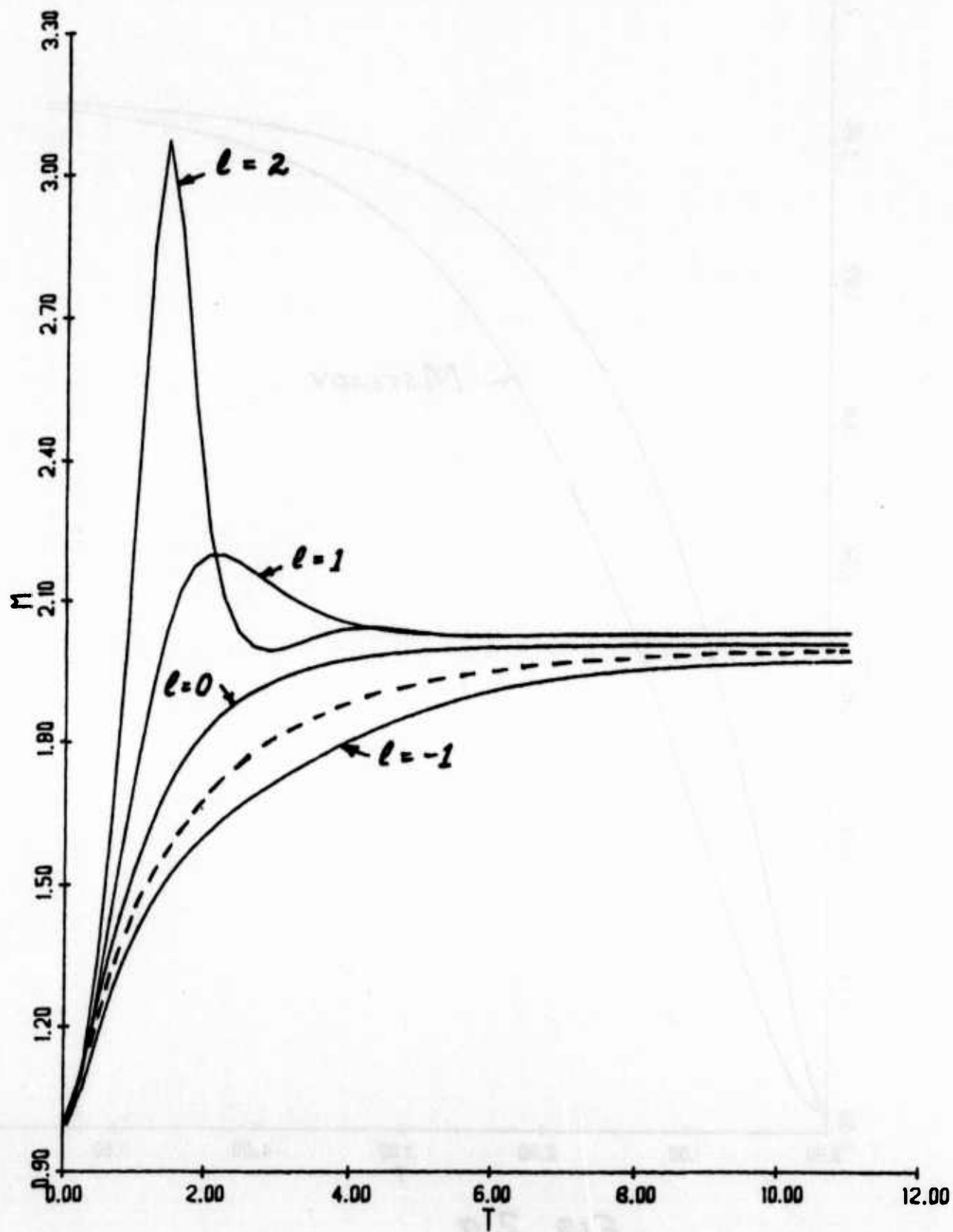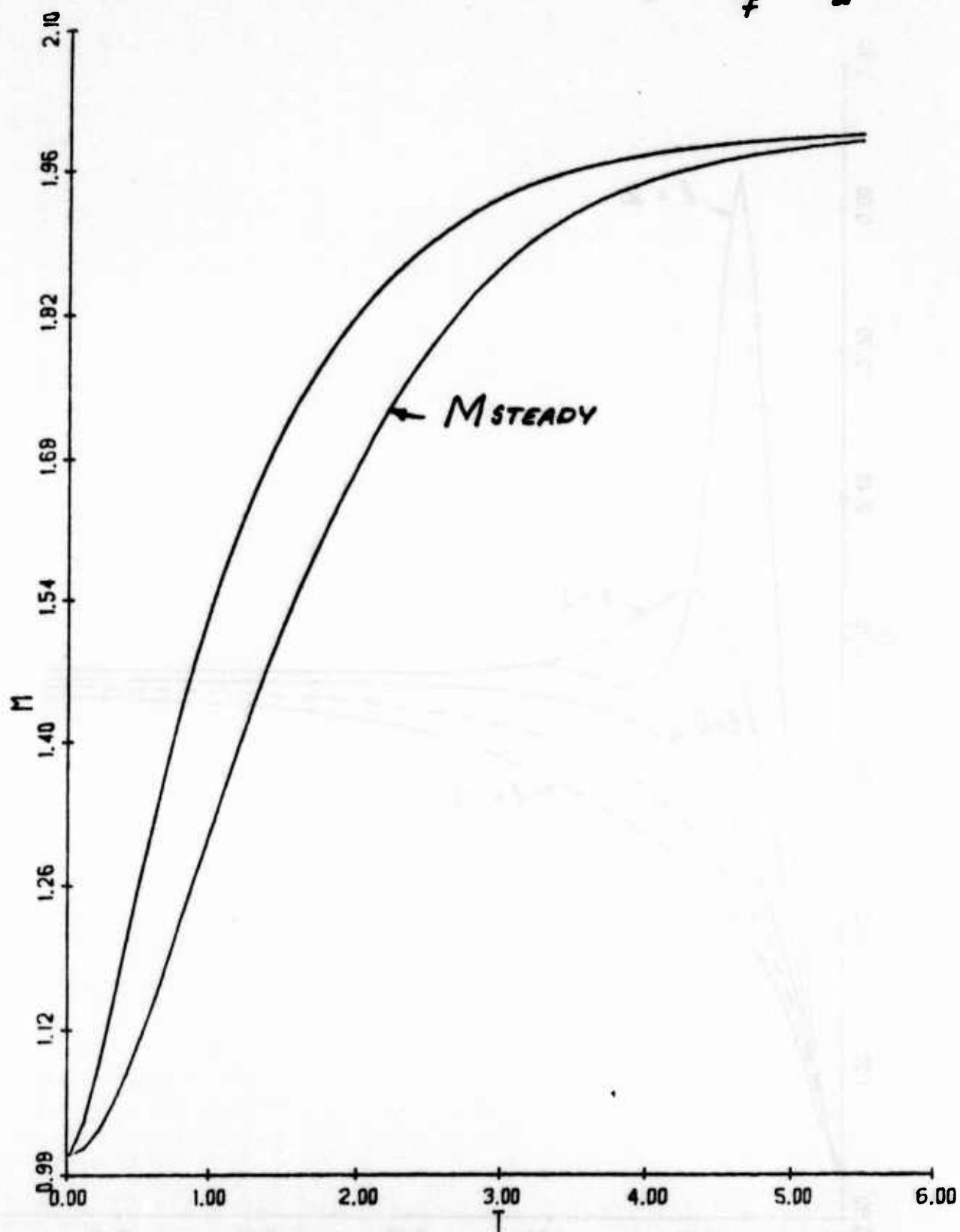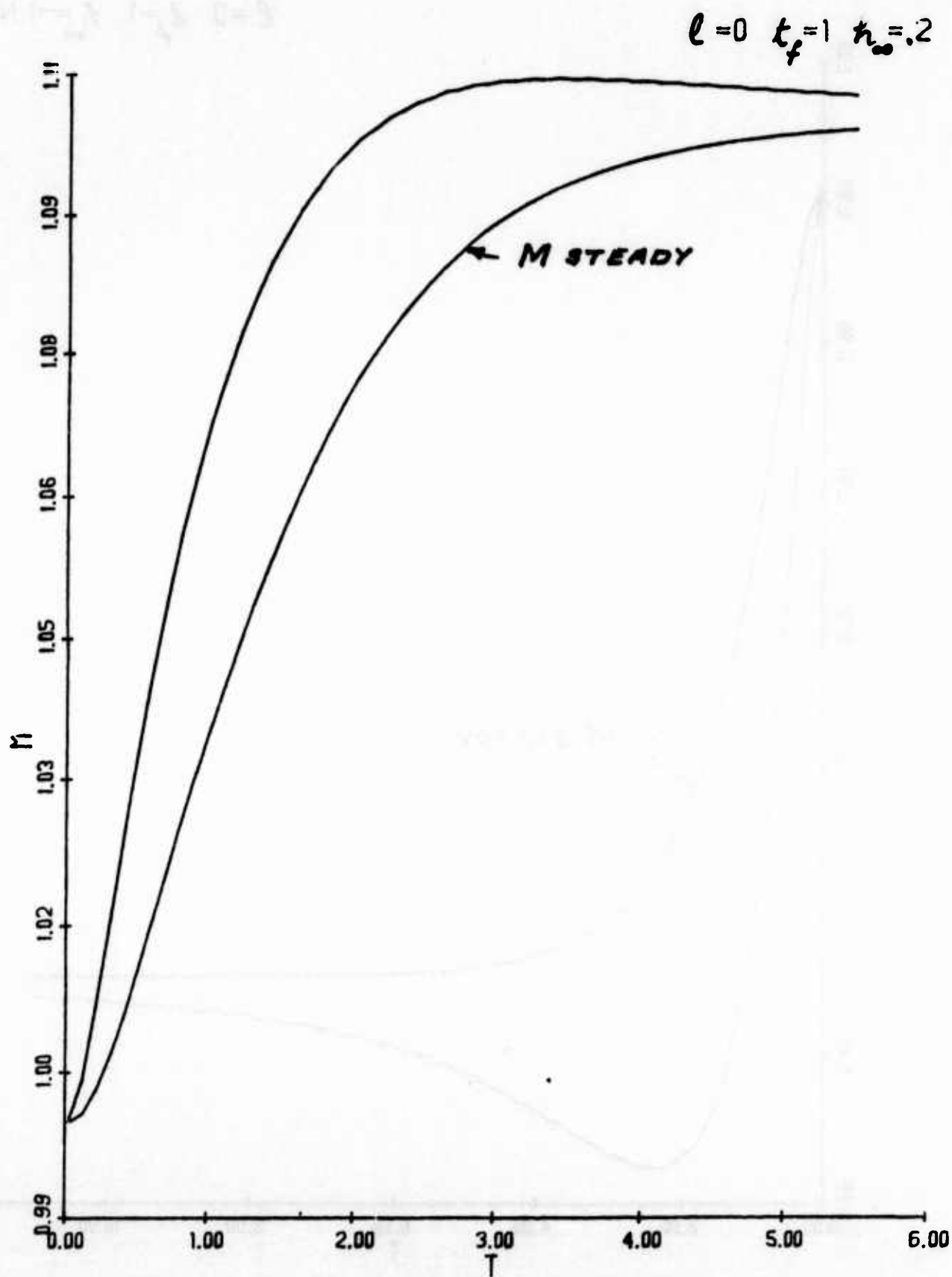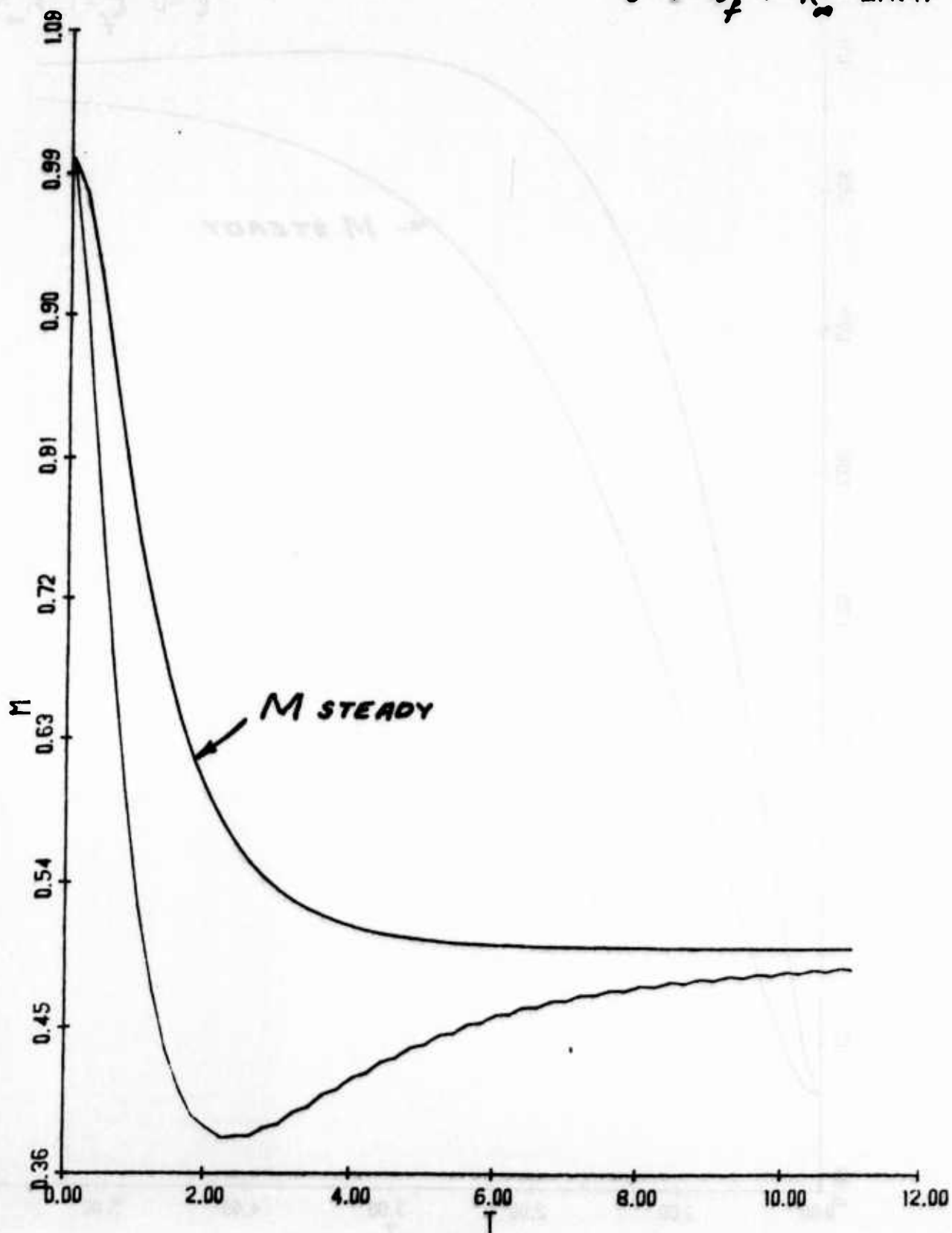
FIG. 2 C

1224
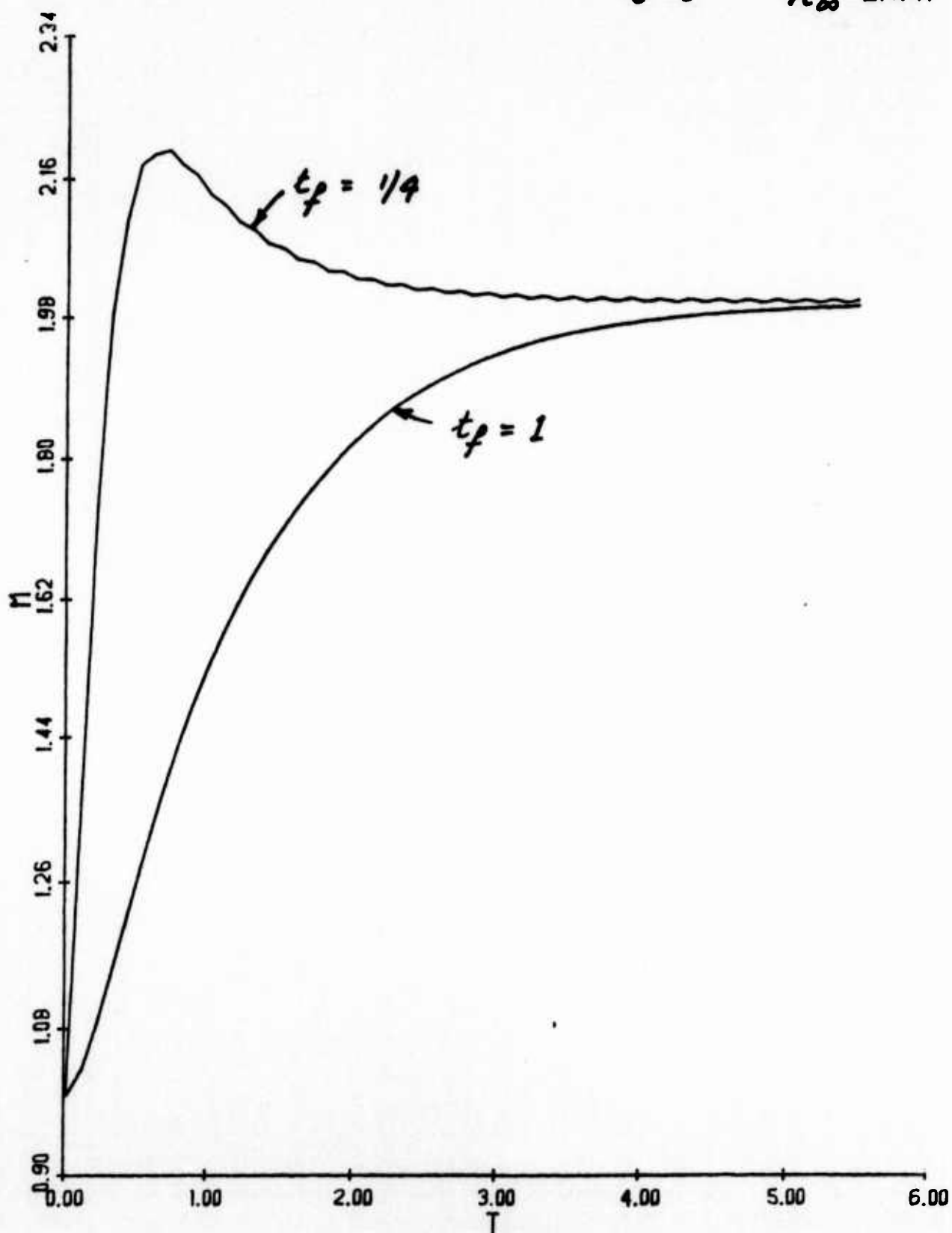
FIG. 3

A 2-Dimensional Scalar Chandrasekhar Filter

for Image Restoration

A. K. Mahalanabis and Kefu Xue

Department of Electrical Engineering

The Pennsylvania State University

University Park, Pa 16802

Abstract — Based on the previous work on the 2-dimensional (2D) strip Chandrasekhar filter (CF) algorithm by Mahalanabis and Xue [1], a more efficient and accurate scalar format 2D CF algorithm is described in this paper. The filtering algorithm is developed for the image modeled by an Non Symmetric Half Plane (NSHP) model. Unlike the conventional Kalman filtering (KF) algorithm which uses the Riccati- type difference equations, this algorithm is based on the Chandrasekhar-type difference equations which gives the algorithm better numerical properties and computational efficiency. The computational requirements of scalar CF algorithm, scalar KF algorithm and the suboptimal reduced update Kalman filtering (RUKF) algorithm developed by Woods and Radewan [2]-[3] are evaluated and compared. The comparison shows that the scalar CF algorithm costs less than 10% of the computational effort that the scalar KF algorithm needs and less than 30% of that the RUKF algorithm needs. The experiment on a simulated image demonstrates the great noise reduction and numerical stability of the algorithm.

i. Introduction

In the previous work [1], we have developed a relatively efficient recursive suboptimal filtering algorithm for reducing the noise in the image data which is so called strip Chandrasekhar filtering (CF) algorithm. Since the CF algorithm will reduce the computational requirement to the maximum extent when the observation quantity is a scalar. The aim of this paper is to develop a 2-dimensional (2D) CF algorithm which processes and restores one image pixel at a time. This is so called scalar 2D CF algorithm. The 2D scalar CF algorithm not only cuts down more than 90% computational effort comparing with the conventional KF algorithm, but also yields the optimal filtering result. A 2D version of the scalar CF algorithm for the image modeled by Non Symmetric Half Plane (NSHP) model is developed and analysed in the following sections.

In section ii, the noise reduction filtering problem is analytically formulated. An $M \times M$ order NSHP model is considered for the noise free data. The observed image data is corrupted with zero mean white noise. The global state space model proposed by Woods and Radewan [2]-[3] is adopted.

Section iii is devoted to the derivation of the 2D scalar CF algorithm and analysis of its computational requirement. In section iv, the computational requirements of scalar KF algorithm and the suboptimal reduced update Kalman filter (RUKF) developed by Woods and Radewan is evaluated. The computational requirements are expressed in term of the order of the NSHP model M as well as the image size index N. These algorithms are scalar processor, therefore the computational requirement can be compared with respect of the number of operations per pixel restoration. Since the number of memory assess operations is strongly computer structure dependent, the investigation of this requirement is out of the focus of this paper. Comparing with the operation time of multiplication and addition,

the logic operation time is ignorable and also the algorithms involve in very limited number of logic operations. Therefore the computational requirement comparison will focus only on the number of multiplications and additions per pixel restoration. section v contains the results of simulation studies and section vi serves as the conclusion.

ii. Problem Formulation

The image to be processed consists of $N \times N$ equally spaced gray level pixels. The noise free image is expressed by an array $\{g(i,j); 1 \leq i,j \leq N\}$ where i and j are vertical and horizontal pixel location indices respectively. The observed image array $\{z(i,j); 1 \leq i,j \leq N\}$ is corrupted with additive noise array $\{v(i,j); 1 \leq i,j \leq N\}$ which is a white zero mean random field with variance σ_v^2. It is assumed that the noise free image $\{g(i,j); 1 \leq i,j \leq N\}$ can be represented by a zero mean discrete Markov random field which is modeled by a autoregressive type NSHP predictive model. Because almost all images have only limited correlation distance, this assumption is reasonable. For an $M \times M$ order NSHP model, the present pixel value can be linearly related to its specified neighboring pixels.

$$
\begin{aligned}
g(i,j) = \sum_{n=1}^{M} \alpha(0,n)g(i,j-n)+ \\
\sum_{m=1}^{M} \sum_{n=-M}^{M} \alpha(m,n)g(i-m,j-n) + w(i,j),
\end{aligned}
\tag{1}
$$

where $1 \leq i,j \leq N$ and $\alpha(m,n)$'s are the coefficients of NSHP model. $\{w(i,j); 1 \leq i,j \leq N\}$ is a white zero mean random field with variance σ_w^2.

The observed image can be expressed as follows:

$$
z(i,j) = g(i,j) + v(i,j), \qquad 1 \leq i,j \leq N.
\tag{2}
$$

1229

Adopting the global state vector developed by Woods and Radewan [2]-[3], for the raster scanned image, i.e., left to right, advance one line, then repeat, the $(NM + M) \times 1$ state vector $x(i, j)$ is defined as follows:

$$x^T(i,j) = [g(i,j), g(i,j-1), \ldots, g(i,1); g(i-1,N), g(i-1,N-1), \ldots, g(i-1,1);$$

$$\ldots; g(i-M+1,N), g(i-M+1,N-1), \ldots, g(i-M+1,1); \tag{3}$$

$$g(i-M,N), g(i-M,N-1), \ldots, g(i-M,j-M+1)],$$

where N is the image size index and M is the order of the NSHP model of noise free image, the dimension of state vector is $(MN + M) \times 1$. Note that the elements of the state vector are the pixel value of the raster scanned noisefree image data.

Based on the definition of state vector $x(i, j)$, the following state equations are derived from the NSHP model.

$$x(i, j + 1) = Fx(i, j) + dw(i, j). \tag{4}$$

$$z(i, j) = hx(i, j) + v(i, j), \tag{5}$$

for $1 \leq i, j \leq N$. $z(i, j)$ is the scalar observed image data. $v(i, j)$ and $w(i, j)$ are scalar zero mean white noise field as defined in (1) and (2). It is assumed that the system noise $w(i.j)$ and additive observed noise $v(i, j)$ are uncorrelated.

The $(MN + M) \times (MN + M)$ transition matrix F consists of the coefficients of the NSHP model in a companion matrix form.

$$F = \begin{pmatrix} f(1,1) & f(1,2) & f(1,3) & \ldots & f(1, MN+M-1) & f(1, MN+M) \\ 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}, \tag{6}$$

where the elements of the first row of state transition matrix $f(i,j)$'s are the coefficients of the original NSHP model (1) which are assigned as follows:

$$f(1,k) = \alpha(0,n) \qquad n = 1,\ldots,M; \qquad k = 1,\ldots,M;$$

$$f(1,k) = \alpha(1,n) \qquad n = -M,\ldots,M; \qquad k = N - M,\ldots,N + M;$$

$$f(1,k) = \alpha(2,n) \qquad n = -M,\ldots,M; \qquad k = 2N - M,\ldots,2N + M;$$

$$\vdots$$

$$f(1,k) = \alpha(M-1,n) \quad n = -M,\ldots,M; \quad k = (M-1)N - M,\ldots,(M-1)N + M;$$

$$f(1,k) = \alpha(M,n) \qquad n = -M,\ldots,M; \qquad k = MN - M,\ldots,MN + M;$$

$$f(1,k) = 0; \qquad k = others.$$

The $(MN + M) \times 1$ column vector d is as follows:

$$d^T = (1 \quad 0 \quad \ldots \quad 0). \tag{7}$$

The $1 \times (MN + M)$ row vector h has the form as follows:

$$h = (1 \quad 0 \quad 0 \quad \ldots \quad 0). \tag{8}$$

The filtering problem can be stated as follows. From the NSHP image model (1) and noisy image data (2), the finite dimensional, discrete time, linear system (4) and (5) are defined for $1 \leq i,j \leq N$ and the estimates

$$\hat{x}_a = E\{x(i,j) \mid z(i,j), z(i,j-1),\ldots,z(1,1)\} \tag{9}$$

and

$$\hat{x}_b = E\{x(i,j) \mid z(i,j-1), z(i,j-2), \ldots, z(1,1)\} \tag{10}$$

should be determined by CF algorithm where subscript b represents "before update" which means the one step prediction and subscript a represents "after update" which means the filtering estimate.

iii. CF Algorithm

The Chandrasekhar algorithm for the solution to the minimum variance filtering problem for the linear discrete time system has been derived by Morf, Sidhu and Kailath [4]. This CF algorithm can be directly applied to the state space model (4) and (5) to yield following vector format equations for the scalar CF algorithm.

Prediction Equations:

$$\hat{x}_b(i,j+1) = F\hat{x}_a(i,j). \tag{11}$$

Filtering Equations:

$$\hat{x}_a(i,j) = \hat{x}_b(i,j) + K(i,j)[z(i,j) - \hat{x}_b(i,j)], \tag{12}$$

where z is the observed image pixel value along the raster scanned noisy image data.

Equations to Update the Kalman Gain Matrix:

$$R(i,j+1) = R(i,j) + hY(i,j)S(i,j)Y^T(i,j)h^T, \tag{13}$$

$$K(i,j+1) = [K(i,j)R(i,j) + Y(i,j)S(i,j)Y^T(i,j)h^T]R^{-1}(i,j+1), \tag{14}$$

$$Y(i,j+1) = F[Y(i,j) - K(i,j+1)hY(i,j)], \tag{15}$$

$$S(i, j+1) = S(i,j) + S(i,j)Y^T(i,j)h^T R^{-1}(i,j)hY(i,j)S(i,j). \tag{16}$$

Taking advantage of the companion matrix format of F matrix, some manipulation involving matrix algebra yields the following set of scalar CF equations.

Prediction Equations:

$$\hat{x}_{b1}(i, j+1) = \sum_{k=1}^{M} f_{(1,k)}\hat{x}_{ak}(i,j) + \sum_{k=M+N-2M}^{N+M} f_{(1,k)}\hat{x}_{ak}(i,j) +$$
$$\cdots + \sum_{k=NM+M-2M}^{NM+M} f_{(1,k)}\hat{x}_{ak}(i,j) \tag{17}$$

$$\hat{x}_{bk}(i, j+1) = \hat{x}_{a(k-1)}(i,j), \qquad k = 2, 3, \ldots, NM + M. \tag{18}$$

Filtering Equations:

$$\hat{x}_{ak}(i,j) = \hat{x}_{bk}(i,j) + K_k(i,j)[z(i,j) - \hat{x}_{b1}(i,j)]. \tag{19}$$

where $k = 1, 2, \ldots, NM + M$.

Equations for updating the $(MN + M) \times 1$ column vector K and Y and scalar R and S are easily obtained as follows:

$$R(i, j+1) = R(i,j) + S(i,j)Y_1^2(i,j). \tag{20}$$

$$K_k(i, j+1) = [K_k(i,j)R(i,j) + Y_k(i,j)S(i,j)Y_1(i,j)]R^{-1}(i, j+1), \tag{21}$$

where $k = 1, 2, \ldots, NM + M$.

$$Y_1(i, j+1) = \sum_{k=1}^{M} f_{(1,k)}[Y_k(i,j) - K_k(i,j)Y_1(i,j)] +$$
$$\sum_{k=M+N-2M}^{N+M} f_{(1,k)}[Y_k(i,j) - K_k(i,j)Y_1(i,j)] + \tag{22}$$
$$\cdots + \sum_{k=MN+M-2M}^{NM+M} f_{(1,k)}[Y_k(i,j) - K_k(i,j)Y_1(i,j)]$$

$$Y_k(i, j+1) = Y_{(k-1)}(i,j) + K_{(k-1)}(i,j)Y_1(i,j). \tag{23}$$

where $k = 2, 3, \ldots, NM + M$.

$$S(i, j+1) = S(i,j) + S^2(i,j)Y_1^2(i,j)R^{-1}(i,j). \tag{24}$$

The equations can be processed recursively starting with the initial conditions.

$$R(1,1) = \sigma_v^2. \tag{25}$$

$$K(1,1) = 0. \tag{26}$$

$$Y(1,1) = d. \tag{27}$$

$$S(1,1) = \sigma_w^2. \tag{28}$$

In the evaluation of the computational requirement of the CF equations (17) (19) (20) (25) (22) (23) and (24), the quantity $S(i,j)Y_1(i,j)$ and $S(i,j)Y_1^2(i,j)$ are only calculated once and stored. The resulting number of multiplication and addition for each iteration will be in terms of the order of NSHP model M as well as image size index N. Since this scalar algorithm restores one pixel at each iteration, the number is equal to the computation requirement of each pixel restoration.

The numbers of multiplication $N_{m_{()}F}$ and addition $N_{a_{()}F}$ per pixel restoration can be calculated by using equation (29) and (30) respectively.

$$N_{m_{()}F} = 5NM + 6M^2 + 11M + 8. \tag{29}$$

$$N_{a_{()}F} = 3NM + 6M^2 + 9M + 2. \tag{30}$$

iv. The Comparison of Computational requirements

The computational requirement of the scalar CF algorithm will be compared with the 2D Riccati-type KF algorithm as well as the suboptimal RUKF algorithm which is developed by Woods and Radewan [2]-[3]. First the number of the multiplication and addition of each iteration of the 2D KF algorithm will be evaluated. In order to compare the computational requirement properly, the 2D scalar KF algorithm which utilizes companion form of F matrix and symmetric property of covariance matrix P should be taken into consideration. With the same state equations as well as the prediction equations (17), (18) and filtering equation (19), only the scalar format updating equations of the Kalman gain matrix K and error covariance matrices P_a and P_b should be derived and analysed. A careful investigation and calculation yield following results.

The number of multiplications involved into the KF equations is

$$N_{m_{KF}} = N^2(M^2) + N(4M^3 + 6M^2 + 2M) + (4M^3 + 7M^2 + 4M). \tag{31}$$

The number of additions involved into the KF equations is

$$N_{a_{KF}} = N^2(M^2) + N(4M^3 + 6M^2 - 2M) + 4M^3 + 7M^2 + 3. \tag{32}$$

Then the computational effort of the RUKF is evaluated. In this work, the global state vector (1) will be partitioned into two parts. One is the local supporting part of elements which will join the updating recursion computation and the other part consists of the rest of elements which will not be updated at each iteration. Consequently, the Kalman gain matrix K and the error covariance matrices P_a and P_b are partitioned accordingly. Only part of the matrix K and matrix P_a corresponding to the local supporting part of the state

space vector are updated. This approximation approach reduces the computational burden significantly, but the RUKF algorithm only gives the suboptimal filtering result and also has poor numerical stability. Simply following the scalar format of the RUKF equations, the number of multiplications and additions are carefully evaluated. Taking into account of the companion matrix format and symmetric property, the evaluation results can be expressed as follows.

The number of multiplications involved into the RUKF equations is

$$N_{m_{RUKF}} = N(6M^3 + 2M^2D + 6M^2 + DM + M) \\ + (6M^3 + 2M^2D + 12M^2 + 5MD + 7M + 2D + 3) \tag{33}$$

The number of additions involved into the RUKF equations is

$$N_{a_{RUKF}} = N(6M^3 + 2M^2D + 6M^2 + DM - 2M) \\ + (6M^3 + 2M^2D + 10M^2 + 3MD + 2M + D + 4) \tag{34}$$

In order to make the RUKF algorithm more numerical stable and the filtering result more accurate, the local supporting area is often enlarged by adding $2D \times M$ pixels into the NSHP supporting pixel area. This is why the variable D appears in the equation (33) and (34).

Comparing the numbers of major computer operations of $N_{m_{KF}}$ (31) and $N_{a_{KF}}$ (32) of the scalar KF algorithm and $N_{m_{RUKF}}$ (33) and $N_{a_{RUKF}}$ (34) of the RUKF algorithm with $N_{m_{CF}}$ (29) and $N_{a_{CF}}$ (30) of the scalar CF algorithm, we can see that the numbers of the major computer operations per pixel restoration of the scalar KF algorithm will be of $O(N^2M^2)$, that of the RUKF will be of $O(NM^3)$ and that of the scalar CF will be only of $O(NM)$. The improvement of computational expense is obvious. Since the computational burden of the KF and RUKF algorithm will increase much faster than the

scalar CF algorithm does as the M or N increasing, a simple numerical example, N=64, M=1 and D=2, can give us an idea how much computational burden has been saved when the scalar CF algorithm is used. In this example, the KF algorithm needs about 4879 multiplications and 4622 additions per pixel restoration; the RUKF algorithm needs about 1257 multiplications and 1058 additions per pixel restoration; and the scalar CF algorithm only needs 345 multiplications and 209 additions per pixel restoration. It concludes that the scalar CF needs only less than 7% of the multiplications and less than 4.5% of the additions that the KF algorithm does and less than 27% of the multiplications and less than 20% of the additions that the suboptimal RUKF algorithm does.

v. Simulation Results

A random field is generated by a 1×1 order NSHP model which represents the noise free image shown on Fig. 1. White noise is added into this generated image to produce the noise contaminated image with $SNR = 3_{dB}$ which is shown on Fig. 2. An estimated image is then computed using the developed scalar CF algorithm which yields $SNR = 12.3_{dB}$. Fig. 3 displays the estimated image. The experiment also shows that the algorithm converges fast and possesses good numerical properties.

vi. Conclusion

The 2D optimal scalar CF algorithm has been derived and implemented in this paper. The computational requirements of this new algorithm is reduced significantly comparing with the 2D KF algorithm and the suboptimal RUKF algorithm. The effectiveness of this algorithm is verified by processing a simulated image. The experiment also shows that the numerical properties of the CF algorithm is better than the conventional KF algorithm.

References

[1]. Mahalanabis, A. K. and Xue, Kefu, "A study of 2-Dimensional strip Chandrasekhar filter for image restoration.", will appear in System Science IX, Poland, 1986.

[2]. Woods, J. W. and Radewan, C. H., "Kalman filtering in two dimensions", IEEE Trans. Info. Th., Vol. IT-23, 1977. pp. 473-482.

[3]. Woods, J. W. and Radewan, C. H., "Correction to Kalman filtering in two dimensions", IEEE Trans. Info. Th., Vol. IT-25, 1979. pp. 628-629.

[4]. Morf, M., Sidhu, G. S. and Kailath, T., "Some new algorithms for recursive estimation in constant, linear, discrete-time systems", IEEE Trans. Auto. Contr., Vol. AC-19, 1974. pp. 315-323.

Figures:

Fig. 1. The noise free image $g(i,j)$ generated by an 1×1 NSHP model
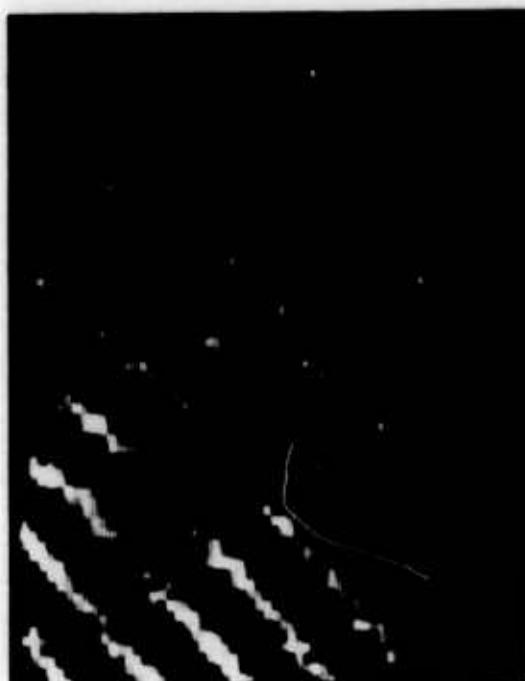
Fig. 2. The white noise corrupted image $z(i,j)$ with $SNR = 3\ dB$

Fig. 3. The filtered version of the image in Fig. 2 obtained with scalar CF.

OBJECT TRACKING USING SENSOR FUSION

Firooz A. Sadjadi & Michael E. Bazakos
Honeywell Systems & Research Center
3660 Technology Drive
Minneapolis, Minnesota 55418

ABSTRACT. Motion is an important cue for extracting moving targets. This is especially so when the targets are camouflaged in the background such that segmentation of the scene does not reveal any information about the targets. There are three different approaches for using passive sensors for obtaining optical flow fields, which are the projected velocity vectors of the points on the moving objects on a plane perpendicular to the line of sight. These approaches are matching methods, spatio-temporal gradient based techniques and biologically based methods. Problems common to these approaches are the sparseness of the optical flow fields and the existence of sensor and algorithm generated false vectors.

To alleviate these problems we are using a novel multi-sensor technique. In this paper we have investigated a gradient based approach to obtain optical flow fields from sequences of multi-sensor images. Due to the particular nature of each sensor, the obtained optical flow fields originated from each sensor, produce overlapping and complementary vectors and this point is exploited in our approach. A joint multidimensional histogram of the sensors' optical flow fields in terms of their salient features such as magnitude and direction are created. The highest peaks in this multidimensional space correspond to the different moving targets. This information can then be used to segment the scene and to separate the moving targets from the background.

This technique is potentially powerful, simple to implement and is relatively insensitive to background noise.

I. INTRODUCTION. Motion is an important cue for extracting information from moving objects. In numerous situations, an otherwise non-detectable target, camouflaged in the background, can be detected and recognized due only to its motion. There are three basically different approaches in the literature for image based motion detection namely matching techniques, spatio-temporal gradient methods and biologically based techniques [1-11]. In all of these approaches sequences of images containing the moving targets are used to obtain optional flow fields. Optical flow fields are the projected velocity vectors associated to each point on the scene on a plane perpendicular to the line of sight. One of the main problems common with all these approaches is that the optical flow fields are sparse due to the fact that textural variations on the target as viewed by a sensor are usually small. The other problems are the presence in the generated optical flow fields, of the background and algorithm induced false vectors that can adversely affect the entire detection/recognition system.

In this paper, we present a multi-sensor approach for object tracking. The sensors are assumed to be imaging and relatively collocated. There is a trend toward multi-sensor approach in many industrial and military applications [11]. This trend is justified and encouraged by the need for more reliable information and the potential robustness in performance that is usually associated with multi-sensor systems. The existence of multi-

sensors brings with it the problem of sensor fusion: how to synergistically combine the information that is available through individual sensors.

Our approach addresses these problems by presenting a novel object tracking technique that synergistically fuses the motion information available from the single sensors.

II. OBTAINING OPTICAL FLOW FIELDS. Optical flow can be obtained by matching, spatio-temporal and biologically based methods. In the matching methods attempt is made to establish correspondence between the successive image frames of the same scene. This correspondence is achieved by scene matching technique. Usually the images are needed to be segmented before they can be used by these methods.

Spatio-temporal methods on the other hand work on the raw images, do not need the solution to the correspondence problem and are good for determining the optical flow of multi-target scenes.

The basis for the spatio-temporal gradient technique is the so called gradient constraint equation which relates the changes in the brightness of images in successive frames to the temporal changes in the scene. For an object of constant brightness u[x, y, t], the following equation is derived:

$$C_1 \equiv V \cdot \nabla u + \frac{\partial u}{\partial t} = 0 \tag{1}$$

where

$$V \equiv \left[\frac{\partial x}{\partial t} \quad \frac{\partial y}{\partial t} \right]^t \equiv [v_x \quad v_y]^t \tag{2}$$

$$\nabla u \equiv \left[\frac{\partial u}{\partial x} \quad \frac{\partial u}{\partial y} \right] \equiv [u_x \quad u_y] \tag{3}$$

There are several approaches to determining optical flow V using Equation 1. These techniques differ in their stating of the second constraint equation and the expression that is to be minimized [3,4,5]. Figure 1 shows two sequential frames from a moving scale model car, obtained by video camera at the rate of 30 frames per second. The resultant optical flow fields shown in Figure 2 are obtained by using equation (1) and a second constraint relations

$$C_2 \equiv \left(\frac{\partial v_x}{\partial x} \right)^2 + \left(\frac{\partial v_y}{\partial y} \right)^2 + \left(\frac{\partial v_y}{\partial x} \right)^2 + \left(\frac{\partial v_x}{\partial y} \right)^2 = 0 \tag{4}$$

In the following minimization relation:

$$\text{Minimize error} = E = \int \int (a^2 C_2 + C_1) \, dx \, dy \tag{5}$$

where samples are taken at discrete points in space and time and quantized in brightness. The partial derivates $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ are estimated by averages using eight measurements in two image frames. As can be seen optical flow conveys information about the outer boundaries of the car. This information can be used as an aid in scene segmentation.

1242
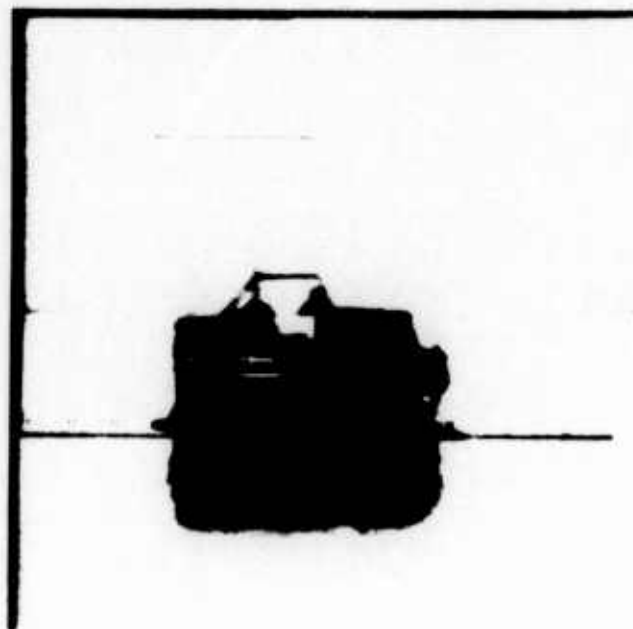
(a)
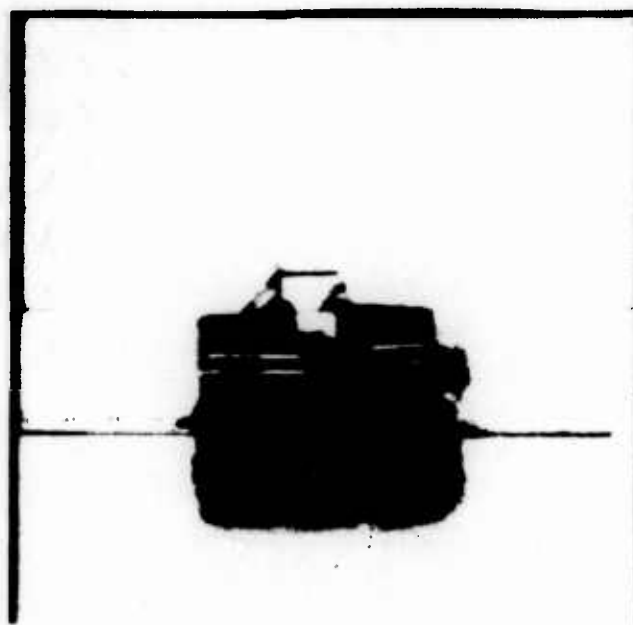
(b)

Figure 1.   The two frames of a sequence of a
            moving car. (a) First frame;
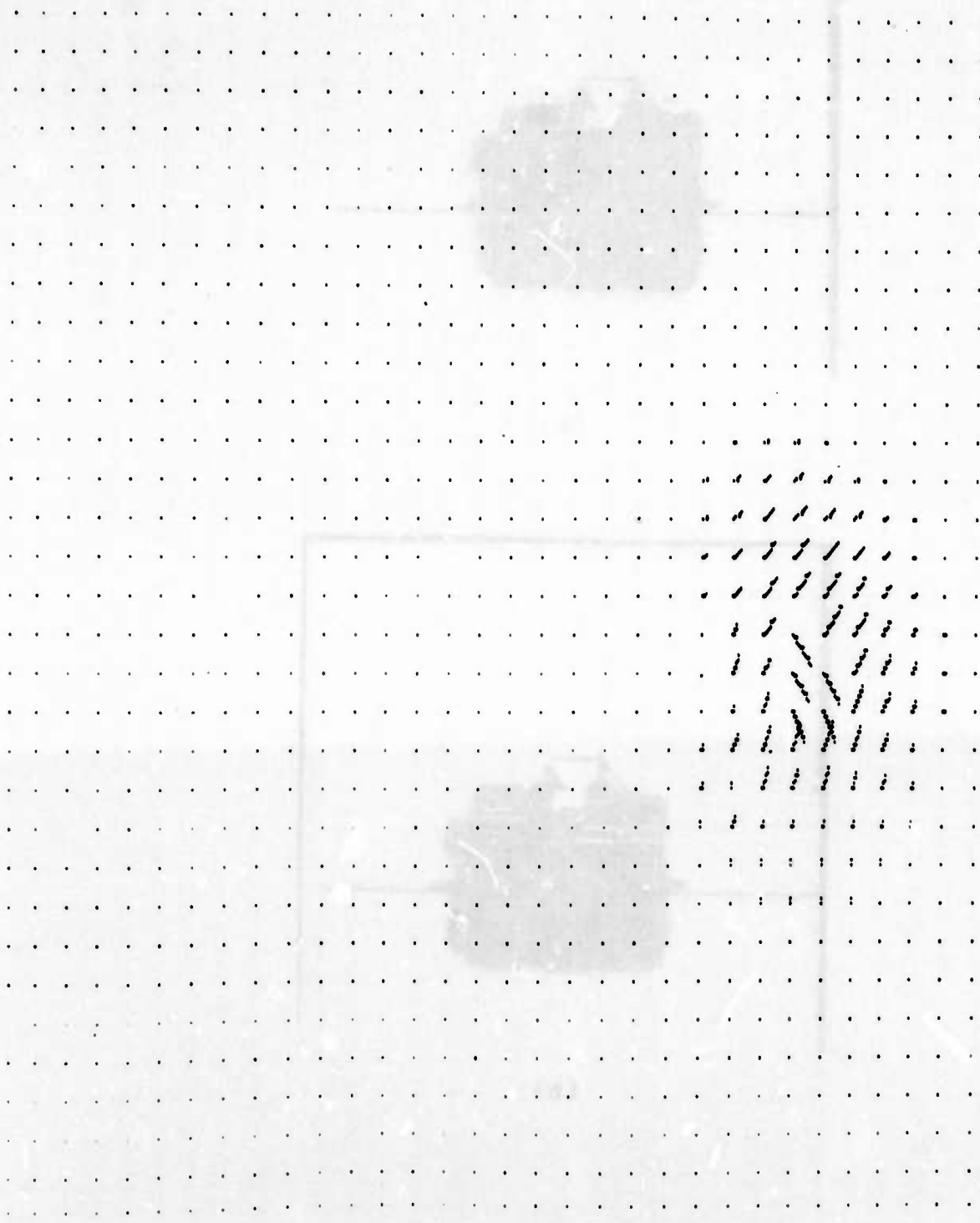   (b) Second frame.

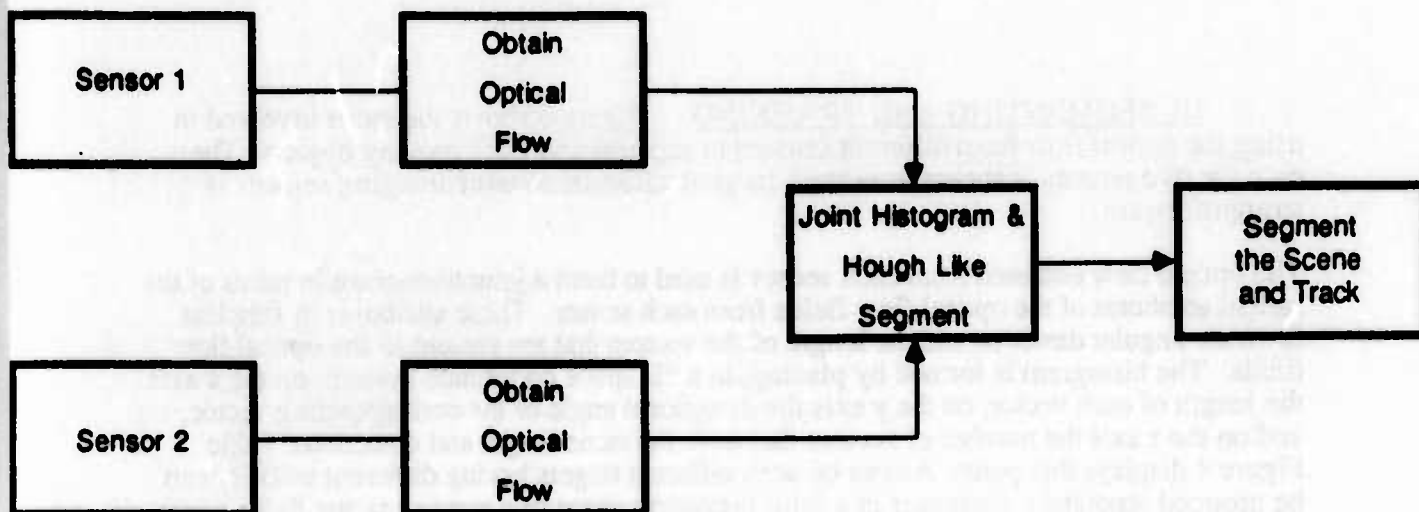Figure 2. Optical Flow Field of Car Using the First Two Frames of Pictures.
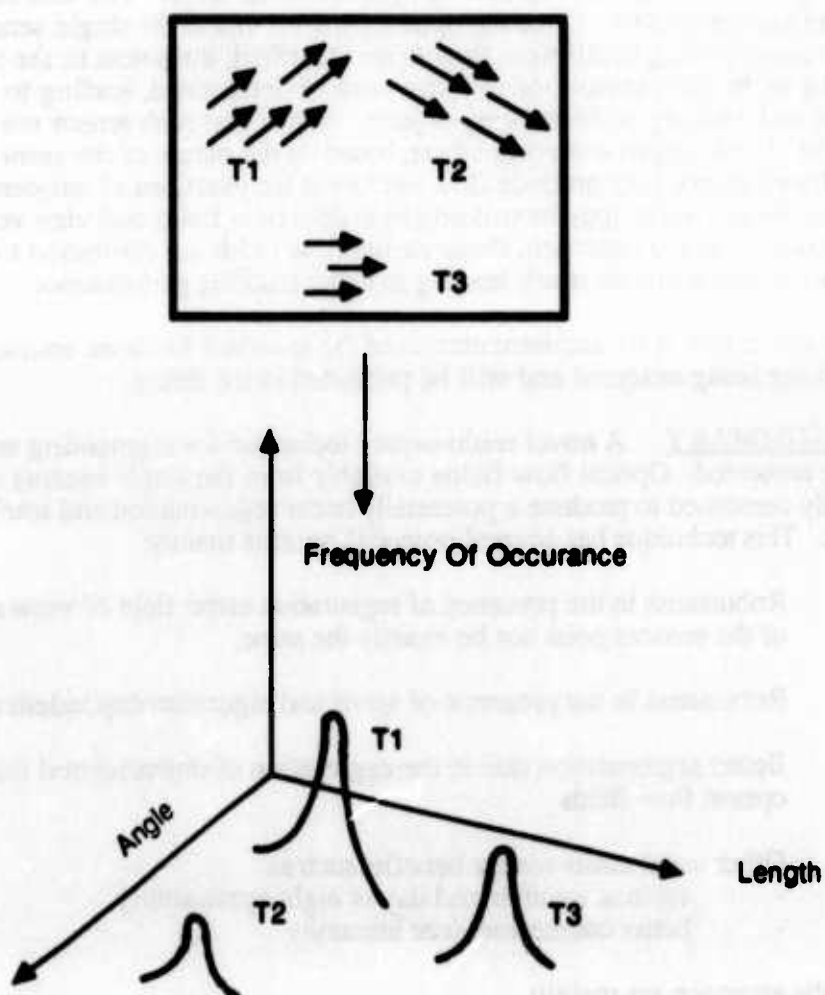
Figure 3. Multi-sensor Tracking System.

Figure 4. Segmentation of moving targets T_1, T_2, and T_3 using joint optical flow's histogram space.

1245

III. SEGMENTING AND TRACKING.   Figure 3 shows the steps involved in using the optical flow from different sensors to segment and track moving objects. The case for two sensors is shown, however, its generalization to more imaging sensors is straightforward.

The optical flow obtained from each sensor is used to form a joint histogram in terms of the salient attributes of the optical flow fields from each sensor. These attributes in simplest form are angular direction and the length of the vectors that are present in the optical flow fields. The histogram is formed by placing, in a 3D space coordinate system, on the x axis the length of each vector, on the y axis the directional angle of the corresponding vector, and on the z axis the number of vectors that have the same length and directional angle. Figure 4 displays this point. As can be seen different targets having different velocity can be grouped separately, moreover in a joint histogram space true moving target fields reinforce each other, leading to a higher peak, however the false vector, that are present in one sensor's optical flow field but are not present in the other, will have a diminishing effect. From the peaks in the histogram domain the corresponding optical flow vectors in the single sensor fields are identified. Due to noise not all of the optical flow vectors corresponding to a target have exactly the same direction and length, for this reason the peaks in the histogram space are not going to be sharp vertical lines. To tolerate these variations, one needs to choose a window centered at the peaks and consider all of the vectors falling inside the window as belonging to the same target. The size of the window can be chosen experimentally. Once the optical flow vectors in the single sensor optical flow fields, corresponding to different targets are identified, the points in the image domain corresponding to the joint sensor optical flows can be determined, leading to the segmentation and tracking of the moving objects. Notice that each sensor may produce partial flow fields relating to a moving target, based on the nature of the sensor used; for example, infrared sensor may produce flow vectors at the junctions of temperature variations, but these vectors may be missing in visible flow fields and vice versa. However, in our proposed approach, these partial flow fields are combined to produce a potentially better segmentation result leading to better tracking performance.

The preliminary results of the implementation of the approach has been encouraging. These results are being analyzed and will be published in the future.

IV. SUMMARY.   A novel multi-sensor technique for segmenting and tracking of objects were presented. Optical flow fields available from the single sensors are synergistically combined to produce a potentially better segmentation and tracking performance. This technique has several potential benefits mainly:

o   Robustness in the presence of registration error; field of view and resolution of the sensors need not be exactly the same.

o   Robustness in the presence of scene and algorithm dependent noise

o   Better segmentation due to the aggregation of object related features from optical flow fields

o   Other usual multi-sensor benefits such as
        -       various weather and day or night applicability
        -       better countermeasure immunity

The cost of the approach are mainly
        -       physical complexity of using multi-sensor

-        computational cost

Future experimentation will show whether the potential benefits of this approach will justify the cost that it entails.

REFERENCES

[1]. R. Jain, W.N. Martin, J.K. Aggarwal, "Segmentation Through the Detection of Changes Due to Motion, "Computer Graphics & Image Processing, Vol. II, Sept. 1979.

[2]. R. Jain, "Dynamic Scene Analysis Using Pixel-Based Processes," Computer, Vol. 14, No. 8, August 1981.

[3] . R. Paquir, E. Dubois, "A Spatio-Temporal Gradient Method for Estimating the Displacement Field in Time-Varying Imagery," Computer Vision, Graphics, and Image Processing, Vol. 21, No. 2, 1983.

[4]. B.K.P. Horn, B.G. Schunck, "Determining Optical Flow," Proceedings of SPIE: Techniques and Applications of Image Understanding, Vol. 281, 1981.

[5]. J.K. Kearney, W.B. Thompson, "Gradient-Based Estimation of Optical Flow with Global Optimization," Proceedings of 1984 IEEE Conference on Artificial Intelligence Applications, Denver, CO, Dec. 84.

[6]. K. Prazdny, "On the Information in Optical Flow," Computer Vision, Graphics, and Image Processing, Vol. 22, 1983.

[7]. E.C. Hildreth, The Measurement of Visual Motion, MIT Press, Cambridge, Mass, 1984.

[8]. A.B. Watson, A.J. Ahumada, "Model of Human Visual Motion Sensing," Journal of the Optical Society of America, Vol. 2, No. 2, February 1985.j

[9]. E.H. Adelson, J.K. Bergen, "Motion Channels Based on Spatio-temporal Energy," Invest. Ophthalmol. Vis. Sci. Suppl., Vol. 25, No. 14, 1984.

[10]. G. Sperling, J.P.H. Van Santen, "Temporal Vovariance Model of Human Motion Perception," Journal of the Optical Society of America, Vol. 1, 1984.

[11]. IEEE Fourteenth Workshop on Applied Imagery Pattern Recognition: Multi-Sensor Fusion, Baltimore Maryland, October 1985.

Millu Rosenblatt-Roth
Center for Automation Research, University of Maryland
College Park, MD 20742

1. Introduction.

In what follows we consider the following problem: Given a sample, determine the random field that generated it. In order to make the problem reasonable, it is necessary to assume that the field is not arbitrary but belongs to some specific class. Making such an assumption means in reality that we are not considering the problem of finding the field that generated the given sample, but some other field that belongs to the given class and approximates the field that generated the given sample. In this paper we will consider that the approximating field to be found is stationary, composed of independent random variables, so that it may be considered one-dimensional.

In this paper we will not be interested in the problem of evaluating how good this approximation is, because this aspect is treated in the author's papers [5], [6].

In order to be able to deal with digitized data and at the same time to reduce the complexity of the problem, we will consider only the case that each of the random variables takes only a finite set of values. In order to extend these results to the continuous case, we would have to consider some process of approximation, such as that used by the author [2]-[4].

Numerical examples are given, showing that good approximations can be obtained based on relatively small sample sizes. In particular, this approach can be used to find random field models that generate given samples of image texture, and so can be applied to texture classification or segmentation. Similar results were obtained already by the author for simple stationary Markov chains [7] as well as for unilateral Markov two-dimensional fields, and will be presented with other occasions.

The author thanks Prof. Azriel Rosenfeld for suggesting the problem, for the many substantial discussions of this subject as well as for his great interest in and sponsoring of this research.

2. The direct theorem.

2.1. Generalities

Let us consider a sequence of independent trials with possible outcomes A_i $(1 \leq i \leq n)$ and corresponding probabilities $p_i > 0$ $(1 \leq i \leq n)$ adding up to 1. Each possible result of a series of s consecutive trials can be written as a sequence

$$C_s = (A_{k_1}, A_{k_2}, \ldots, A_{k_s}) \qquad (2.1)$$

where each k_r $(1 \leq r \leq s)$ can take any value i $(1 \leq i \leq n)$. Because of stationarity, the probability of occurrence of the sequence C_s does not depend on the moment when the trials begin; taking into consideration the independence of the trials, this probability can be written as

$$P(C_s) = \prod_{r=1}^{s} P(A_{k_r}) \qquad (2.2)$$

Let us denote by m_i $(1 \leq i \leq n)$ the number of times the outcome A_i appears in the sequence C_s, so that

$$\sum_{i=1}^{n} m_i = s . \qquad (2.3)$$

The equality (2.2) can be written

$$P(C_s) = \prod_{i=1}^{n} p_i^{m_i} \qquad (2.4)$$

In what follows we denote by

$$H = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} \qquad (2.5)$$

the entropy of the random field characterized by the probabilities p_i $(1 \leq i \leq n)$, and

$$\rho = \sum_{i=1}^{n} \log \frac{1}{p_i} \qquad (2.6)$$

Obviously

$$0 < \rho < \infty \qquad (2.7)$$

2.2 The theorem

Let us denote by Γ_s the class of all sequences C_s. For given $\delta > 0$, $s > 0$ we denote by $\Gamma'_{\delta,s}$ the set of all sequences $C_s \in \Gamma_s$ such that

$$|m_i - sp_i| < s^\delta \qquad (2.8)$$

for all i $(1 \leq i \leq n)$, and by $\Gamma''_{\delta,s}$ its complement with respect to Γ_s.

Definition. Sequences $C_s \in \Gamma'_{\delta,s}$ will be called (δ,s)-standard sequences or simple standard sequences.

Let us consider the equation

$$\frac{1}{\sqrt{2\pi}} \int_0^u \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{2}\left(1 - \frac{c}{n}\right) \qquad (2.9)$$

and let us denote by $u(c)$ its solution.

Definition. Given $c > 0$, $\delta > 0$, $s > n$, condition A holds if

$$4\delta^2 c s > n \qquad (2.10)$$

and condition B holds if

$$4\delta^2 s > u^2(c) \qquad (2.11)$$

Let us denote by $N(\cdot)$ the cardinality of a set.

Theorem 1. Let us suppose that at least one of the conditions A, B holds. Then
(a) If C_s is a (δ,s)-standard sequence, it follows that

$$\left|\frac{1}{s} \log \frac{1}{P(C_s)} - H\right| < \delta\rho \qquad (2.12)$$

(b) $P(\Gamma'_{\delta,s}) \geq 1 - c$ $\qquad (2.13)$

(c) $\lim\limits_{\substack{s\to\infty\\ \delta\to 0}} \frac{1}{s}\log N(\Gamma'_{\delta,s}) = H$ (2.14)

Remark 1. The relation (2.12) is equivalent to

$$2^{-s(H+\delta\rho)} < P(C_s) < 2^{-s(H-\delta\rho)} \quad (2.15)$$

i.e. to

$$P(C_s) = 2^{-sH+s\delta\rho\theta}, \quad |\theta| < 1 \quad (2.16)$$

Remark 2. The relation (2.13) is equivalent to

$$P(\Gamma''_{\delta,s}) < \varepsilon \quad (2.17)$$

Remark 3. From (2.14) it follows that

$$\lim_{s\to\infty}\frac{N(\Gamma'_{\delta,s})}{N(\Gamma_s)} = 0, \quad \lim_{s\to\infty}\frac{N(\Gamma''_{\delta,s})}{N(\Gamma_s)} = 1 \quad (2.18)$$

Indeed, from (2.14) we obtain the relation

$$\log N(\Gamma'_{\delta,s}) = s\cdot\big(H+o(1)\big) \quad (2.19)$$

i.e.,

$$N(\Gamma'_{\delta,s}) = 2^{s\big(H+o(1)\big)} \quad (2.20)$$

Taking into consideration that

$$N(\Gamma_s) = n^s = 2^{s\log n} \quad (2.21)$$

and because

$$H < \log n , \quad (2.22)$$

it follows that

$$\frac{N(\Gamma'_{\delta,s})}{N(\Gamma_s)} = 2^{-s\big(\log n - H + o(1)\big)} = o(1) \quad (2.23)$$

which is equivalent to the first equality in (2.18), and

$$\frac{N(\Gamma''_{\delta,s})}{N(\Gamma_s)} = \frac{N(\Gamma_s)-N(\Gamma'_{\delta,s})}{N(\Gamma_s)} = 1 - \frac{N(\Gamma'_{\delta,s})}{N(\Gamma_s)}$$
$$= 1 + o(1) \quad (2.24)$$

which is equivalent to the second equality in (2.18).

Remark 4. Our Theorem 1 is closely related to some results which go back to Shannon [9] and received a mathematically acceptable form from Khinchine [1].

Our Theorem 1 (a), (b) refers to independent random variables, while that in [3] refers to ergodic simple Markov chains, but our result is not a particular case of that in [3]. Indeed, the results in [3] are existence theorems, considering that δ, ε can be taken as small and s as large as desired, while our results give effective relations between δ, ε, s in order that the results hold.

Our Theorem 1 (c) refers to the set $\Gamma'_{\delta,s}$ of all standard sequences C_s, while the result in [1], Th. 3) refers to another set of sequences C_s; our result contains a limit for $\delta \to 0$, $s \to \infty$, while the result in ([1], Th. 3) contains a limit for $s \to \infty$.

2.3. Proof

(a) Let us consider a sequence $C_s \in \Gamma'_{\delta,s}$. From (2.8) it follows that

$$m_i = sP_i + s\delta\theta_i , \quad |\theta_i| < 1 \quad (1\le i\le n) \quad (2.25)$$

From (2.4) there follows the relation

$$\log P(C_s) = \sum_{i=1}^{n} m_i \log P_i \quad (2.26)$$

and taking into consideration (2.25), there follows the equality

$$\log P(C_s) = \sum_{i=1}^{n} (sP_i + s\delta\theta_i) \log P_i$$
$$\quad (2.27)$$
$$= s\sum_{i=1}^{n} P_i \log P_i + s\delta\cdot\sum_{i=1}^{n} \theta_i \log P_i$$

which can also be written as

$$\log\frac{1}{P(C_s)} = sH + s\delta\cdot\sum_{i=1}^{n} \theta_i \log\frac{1}{P_i} \quad (2.28)$$

From (2.28) we obtain the result (a):

$$\left|\frac{1}{s}\log\frac{1}{P(C_s)} - H\right| < \delta\cdot\sum_{i=1}^{n} |\theta_i| \log\frac{1}{P_i} \quad (2.29)$$
$$\le \delta\cdot\sum_{i=1}^{n} \log\frac{1}{P_i} = \delta\rho$$

(b) Instead of proving inequality (2.13) we will prove (2.17). In order that a sequence $C_s \in \Gamma_s$ belong to $\Gamma''_{\delta,s}$, it is necessary that for at least some value of i $(1 \le i \le n)$ the inequality (2.8) does not hold, i.e.,

$$\Gamma''_{\delta,s} = \bigcup_{i=1}^{n} \left\{|m_i - sP_i| > s\delta\right\} \quad (2.30)$$

so that

$$P(\Gamma''_{\delta,s}) = P\left\{\bigcup_{i=1}^{n}\left\{|m_i-sP_i| > s\delta\right\}\right\} \le \sum_{i=1}^{n} P\left\{|m_i-sP_i| > s\delta\right\}$$
$$\quad (2.31)$$

(b1) Let us assume that condition A holds. It is known from the elements of the theory of probability that

$$P\left\{|m_i-sP_i| > s\delta\right\} \le \frac{P_i(1-P_i)}{s\delta^2} \quad (2.32)$$

But for $0 \le x \le 1$, we have the inequalities

$$0 \le x(1-x) \le \frac{1}{4} \quad (2.33)$$

where the maximum value is reached for $x = \frac{1}{2}$, so that from (2.32) it follows that

$$P\left\{|m_i-sP_i| > s\delta\right\} \le \frac{1}{4s\delta^2} \quad (1\le i\le n) \quad (2.34)$$

Consequently, from (2.31) there follows the inequality

$$P(\Gamma''_{\delta,s}) \le \frac{n}{4s\delta^2} \quad (2.35)$$

and because of (2.10), it follows that (2.17) holds.

(b2) Let us assume that condition B holds. From the central limit theorem in the Moivre-Laplace form, it is known that

$$P\left\{\left|\frac{m_i}{s} - P_i\right| \le \delta\right\} = P\left\{\left|\frac{m_i - sP_i}{\sqrt{sP_i(1-P_i)}}\right| \le \delta\sqrt{\frac{s}{P_i(1-P_i)}}\right\}$$
$$\quad (2.36)$$
$$\sim \frac{2}{\sqrt{2\pi}} \int_{0}^{\delta\sqrt{\frac{s}{P_i(1-P_i)}}} e^{-\frac{x^2}{2}} dx \quad (1\le i\le n)$$

so that

$$P\left\{|m_i - sp_i| > \delta s\right\} \sim 1 - \frac{2}{\sqrt{2\pi}} \int_0^{\delta\sqrt{\frac{s}{P_i(1-P_i)}}} e^{-\frac{x^2}{2}} \, dx \, ,$$

$$(1 \le i \le n) \qquad (2.37)$$

In order to obtain the relation (2.17) it is sufficient to take

$$1 - \frac{2}{\sqrt{2\pi}} \int_0^{\delta\sqrt{\frac{s}{P_i(1-P_i)}}} e^{-\frac{x^2}{2}} \, dx < \frac{\epsilon}{n} \quad (1 \le i \le n) \qquad (2.38)$$

i.e.,

$$\frac{1}{\sqrt{2\pi}} \int_0^{\delta\sqrt{\frac{s}{P_i(1-P_i)}}} e^{-\frac{x^2}{2}} \, dx > \frac{1}{2} \cdot \left(1 - \frac{\epsilon}{n}\right)$$

$$(1 \le i \le n) \qquad (2.39)$$

which is equivalent to the inequality

$$\delta \sqrt{\frac{s}{P_i(1-P_i)}} > u(\epsilon) \qquad (1 \le i \le n) \qquad (2.40)$$

Because of (2.33), we have the inequality

$$\delta \sqrt{\frac{s}{P_i(1-P_i)}} \ge 2\delta\sqrt{s} \qquad (1 \le i \le n) \qquad (2.41)$$

so that in order to satisfy (2.40) it is sufficient to take into consideration condibion B (2.11), i.e.

$$2\delta\sqrt{s} > u(\epsilon) \, . \qquad (2.42)$$

(c) If $C_s \in \Gamma'_{\delta,s}$, then (2.15) holds, so that

$$N(\Gamma'_{\delta,s}) \, 2^{-s(H+\delta\rho)} < \sum P(C_s) = P(\Gamma'_{\delta,s}) < 1 \quad (2.43)$$

where the summation is for all $C_s \in \Gamma'_{\delta,s}$. From (2.43) there follows the relation

$$\frac{1}{s} \cdot \log N(\Gamma'_{\delta,s}) < H + \delta\rho \qquad (2.44)$$

In a similar way, from (2.13), (2.15) there follow the relations

$$1 - \epsilon < P(\Gamma'_{\delta,s}) = \sum P(C_s) < N(\Gamma'_{\delta,s}) \, 2^{-s(H-\delta\rho)} \qquad (2.45)$$

where the summation is also for all $C_s \in \Gamma'_{\delta,s}$. From (2.45) we obtain the relation

$$H - \delta\rho < \frac{1}{s} \cdot \log N(\Gamma'_{\delta,s}) + \frac{1}{s} \log \frac{1}{1-\epsilon} \qquad (2.46)$$

From (2.44), (2.46) it follows that

$$H - \delta\rho - \frac{1}{s} \cdot \log \frac{1}{1-\epsilon} < \frac{1}{s} \log N(\Gamma'_{\delta,s}) < H + \delta\rho \qquad (2.47)$$

For ϵ given, arbitrary, δ as small as we want, and s as large as we want, because of (2.7) it follows that (2.14) holds.

3. The inverse theorem.

3.1. Generalities

Let $\delta > 0$, $\epsilon > 0$, $s > 1$, and let C_s^0 be an arbitrary specific sequence, belonging to Γ_s. Let us assume that one of the conditions A or B holds.

In what follows we assume that C_s^0 is generated by a sequence of independent trials, with possible outcomes A_i $(1 \le i \le n)$ with unknown probabilities P_i $(1 \le i \le n)$, and we will try to determine some intervals in which these probabilities can take values. Let us denote

$$m_i^0 = m_i(C_s^0) \qquad (1 \le i \le n) \qquad (3.1)$$

and by $W(S)$ the confidence of statement S.

3.2. The theorem

Because we have proved that

$$P(\Gamma'_{\delta,s}) > 1 - \epsilon \, , \qquad P(\Gamma''_{\delta,s}) < \epsilon \qquad (3.2)$$

it follows that with confidence larger than $1 - \epsilon$, $C_s^0 \in \Gamma'_{\delta,s}$, i.e.,

$$W\left\{|m_i^0 - sp_i| < \delta s \, , \quad (1 \le i \le n)\right\} > 1 - \epsilon \quad (3.3)$$

i.e.,

$$W\left\{\frac{m_i^0}{s} - \delta < p_i < \frac{m_i^0}{s} + \delta, \quad (1 \le i \le n)\right\} > 1 - \epsilon \qquad (3.4)$$

Let L_n be the Banach space of all vectors

$$q = (q_1, \ldots, q_n) \qquad (3.5)$$

with q_i real numbers of any sign, with norm

$$\|q\| = \sup\left\{|q_i| \, ; \, 1 \le i \le n\right\} \qquad (3.6)$$

Let Π_n be the totality of probability measures

$$p = (p_1, \ldots, p_n) \qquad (3.7)$$

with $p_i > 0$ $(1 \le i \le n)$, and

$$\sum_{i=1}^n p_i = 1 \qquad (3.8)$$

This is a metric space with distance

$$\|p - p'\| = \sup\left\{|p_i - p_i'| \, ; \, 1 \le i \le n\right\} \qquad (3.9)$$

where $p, p' \in \Pi_n$, $p - p' \in L_n$. If $p, p' \in \Pi_n$ are two different solutions, satisfying the inequalities in (3.4), it follows that

$$|p_i - p_i'| < 2\delta \qquad (1 \le i \le n) \qquad (3.10)$$

so that from (3.9) it follows that

$$\|p - p'\| < 2\delta \qquad (3.11)$$

We have thus proved

Theorem 2. Let us assume that

(1) ϵ, δ, s satisfy one of conditions A, B;
(2) the arbitrary sequence $C_s^0 \in \Gamma_s$ is generated by an independent identically distributed sequence of trials, with unknown probabilities p_i $(1 \le i \le n)$.
Then
(a) The relation (3.4) holds.
(b) If p, p' are two different solutions, their distance in Π_n is less than 2δ.

Remark 4. Let L' be the Banach space of all vectors (3.5) with norm the total variation

$$\||q\|| = \sum_{i=1}^{n} |q_i| \qquad (3.12)$$

Then Π_n is a metric space with distance

$$\||p - p'\|| = \sum_{i=1}^{n} |p_i - p_i'| \qquad (3.13)$$

where $p, p' \in \Pi_n$.

If $p, p' \in \Pi_n$ are two different solutions, satisfying (3.4), it follows from (3.13) that

$$\||p - p'\|| < 2n\delta \qquad (3.14)$$

It is easy to see that

$$\| p - p' \| \leq \||p - p'\|| \leq n \||p - p'\|| \qquad (3.15)$$

We remark also that if L^n is the Euclidean space of all vectors (3.5) with norm

$$((q)) = \left(\sum_{i=1}^{n} q_i^2 \right)^{1/2} \qquad (3.16)$$

then Π_n is a Euclidean space with distance

$$((p - p')) = \left(\sum_{i=1}^{n} |p_i - p_i'|^2 \right)^{1/2} \qquad (3.17)$$

It is easy to see that

$$\|p - p'\| \leq ((p - p')) \leq \sqrt{n} \|p - p'\| \qquad (3.18)$$

4. Examples.

4.1. Examples under Condition A

Example 1. Let C_s^0 be a sequence with $n = 2$, $s = 10^4$, $\varepsilon = 2^{-3} = 0.125$, $\delta > 0.02$, so that condition A holds. Let $m_1^0 = 3 \times 10^3$, $m_2^0 = 7 \times 10^3$. From (3.4) it follows that

$$W\left\{ 0.28 < p_1 < 0.32; \ 0.68 < p_2 < 0.72 \right\} > 0.875 \qquad (4.1)$$

and from (3.11) we obtain

$$\|p - p'\| < 0.04 \qquad (4.2)$$

Example 2. Let C_s^0 be a sequence with $n = 2$, $s = 10^6$, $\varepsilon = 2^{-3} = 0.125$, $\delta > 0.002$, so that condition A holds. Let $m_1^0 = 3 \times 10^5$, $m_2^0 = 7 \times 10^5$. From (3.4) it follows that

$$W\left\{ 0.298 < p_1 < 0.302; \ 0.698 < p_2 < 0.702 \right\} > 0.875 \qquad (4.3)$$

and from (3.11) we obtain

$$\| p - p' \| < 0.004 \qquad (4.4)$$

4.2. Examples under Condition B

Example 3. Let C_s^0 be a sequence with $n = 2$, $\varepsilon = 2^{-3} = 0.125$, $s = 10^4$, $\delta > 0.009$, $m_1^0 = 3 \times 10^3$, $m_2^0 = 7 \times 10^3$, so that

$$\frac{1}{2}\left(1 - \frac{\varepsilon}{2} \right) = 0.46875 \qquad (4.5)$$

and relation (2.39) takes the form

$$\frac{1}{\sqrt{2\pi}} \int_0^{u(\varepsilon)} e^{-\frac{x^2}{2}} \, dx > 0.46875 \qquad (4.6)$$

which holds for

$$u(\varepsilon) > 1.8 \qquad (4.7)$$

Considering Condition B in form (1.42) it is easy to see that it holds. From (3.4) it follows that

$$W\left\{ 0.291 < p_1 < 0.309; \ 0.691 < p_2 < 0.709 \right\} > 0.875 \qquad (4.8)$$

and from (3.11) it follows that

$$\| p - p' \| < 0.018 \qquad (4.9)$$

Example 4. Let C_s^0 be a sequence with $n = 2$, $\varepsilon = 2^{-3} = 0.125$, $s = 10^6$, $\delta > 0.0009$, $m_1^0 = 3 \times 10^5$, $m_2^0 = 7 \times 10^5$; in this case, relations (4.5)-(4.8) hold, so that Condition B holds. From (3.4) it follows that

$$W\left\{ 0.2991 < p_1 < 0.3009; \ 0.6991 < p_2 < 0.7009 \right\} > 0.875 \qquad (4.10)$$

and from (3.11) it follows that

$$\| p - p' \| < 0.0018 \qquad (4.11)$$

4.3. Examples involving images that satisfy Condition A or B

Example 5. Let us consider a digital television picture, i.e., an array of 500^2 points, where each point can have 256 levels of gray. Here $n = 256$, $s = 500^2 = 250{,}000$; let $\varepsilon = 1/256 = 0.00390625$. Taking these values, if we want condition A satisfied it is sufficient that

$$4\delta^2 \times 250{,}000 \times \frac{1}{256} > 256 \qquad (4.12)$$

or

$$10^6 \, \delta^2 > 256^2 , \qquad (4.13)$$

i.e.,

$$\delta > 0.256 . \qquad (4.14)$$

Consequently

$$W\left\{ \left| \frac{m_i^0}{s} - p_i \right| < 0.256; \ (1 \leq i \leq 256) \right\} > 0.9960937 \qquad (4.15)$$

with

$$\| p - p' \| = \max\left\{ |p_i - p_i'|, \ 1 \leq i \leq 256 \right\} < 0.512 \qquad (4.16)$$

Example 6. With the same basic data as in Example 5, we take $n = 256$, $s = 500^2$, $\varepsilon = 1/256 = 0.00390625$, and we consider that condition B holds, i.e.,

$$2\delta\sqrt{s} > u(\varepsilon) \qquad (4.17)$$

Here

$$\frac{1}{2}\left(1 - \frac{\varepsilon}{n} \right) = \frac{1}{2}\left(1 - \frac{1}{256^2} \right) = \frac{1}{2}\left(1 - \frac{1}{65{,}536} \right)$$
$$\sim \frac{1}{2}\left(1 - \frac{1}{60{,}000} \right) \sim \frac{1}{2}(1 - 0.16667)$$
$$= \frac{1}{2} \times 0.83334 = 0.41667 \qquad (4.18)$$

so that from tables it follows that

$$u(\varepsilon) \sim 1.30 \qquad (4.19)$$

Thus

$$2\delta \times 500 > 1.30 \qquad (4.20)$$

i.e.,

$$\delta > 0.0013 \qquad (4.21)$$

So

$$W\left\{ \left| \frac{m_i^0}{s} - p_i \right| < 0.0013; \ (1 \leq i \leq 256) \right\} > 0.9960937 \qquad (4.22)$$

and

$$\| p - p' \| < 0.0026 \qquad (4.23)$$

Example 7. Let us take $n = 256$, $s = 500^2$, $\varepsilon = 1/16 = 0.0625$, and let us assume that Condition A holds. Then

$$4\delta^2 \times 250,000 \times \frac{1}{16} > 256 \qquad (4.24)$$

i.e.,

$$10^6 \, \delta^2 > 2^{12} \qquad (4.25)$$

or

$$\delta > 0.064 \qquad (4.26)$$

so that

$$W\left\{\left|\frac{m_i^0}{s} - p_i\right| < 0.064; \; 1 \le i \le 256\right\} > 0.9375 \quad (4.27)$$

$$\| p - p' \| < 0.128 \qquad (4.28)$$

Example 8. Let $n = 256$, $s = 250,000$, $\varepsilon = 1/16 = 0.0625$ and let us assume that Condition B holds. Then

$$\frac{1}{2}\left(1 - \frac{\varepsilon}{n}\right) = \frac{1}{2}\left(1 - \frac{1}{16} \cdot \frac{1}{256}\right) = \frac{1}{2}\left(1 - \frac{1}{4096}\right) \qquad (4.29)$$

$$\sim \frac{1}{2}\left(1 - \frac{1}{4000}\right) = \frac{1}{2}(1 - 0.00025) = \frac{1}{2} \times 0.9975 = 0.49987$$

so that

$$u(\varepsilon) \sim 3.8 \qquad (4.30)$$

i.e.,

$$2\delta \times 500 > 3.8 \qquad (4.31)$$

or

$$\delta > 0.0038 \qquad (4.32)$$

Thus

$$W\left\{\left|\frac{m_i^0}{s} - p_i\right| < 0.0038; \; 1 \le i \le 256\right\} > 0.09375 \qquad (4.33)$$

$$\| p - p' \| < 0.0076 \qquad (4.34)$$

Example 9. Let us assume that we have a 30-minute sequence of TV pictures. If we have 32 pictures in each second, we have a total of

$$32 \times 60 \times 30 = 2^4 \times 60^2 \qquad (4.35)$$

pictures, succeeding each other in time. Assuming independence between the pictures, we have $n = 256$, $s = 500^2 \times 2^4 \times 60^2$, and let $\varepsilon = 1/256 = 0.00390625$. Assuming that Condition A holds, the value of   is given by

$$4\delta^2(250,000) \times 2^4 \times 60^2 \times \frac{1}{256} > 256 \qquad (4.36)$$

or

$$10^6 \, \delta^2 \times 2^4 \times 60^2 > 256^2 \qquad (4.37)$$

i.e.,

$$10^3 \, \delta \times 2^2 \times 60 > 256 \qquad (4.38)$$

Then

$$\delta > \frac{256}{10^2 \times 240} > 0.001 \qquad (4.39)$$

Consequently

$$W\left\{\left|\frac{m_i^0}{s} - p_i\right| < 0.001; \; (1 \le i \le 256)\right\} > 0.9960937 \qquad (4.40)$$

and

$$\| p - p' \| < 0.002 \qquad (4.41)$$

Example 10. Let us consider the same problem as in Example 9, with the supposition that Condition B holds. In this case

$$2\delta(500 \times 2^2 \times 60) > 1.30 \qquad (4.42)$$

i.e.,

$$\delta > \frac{13}{2,400,000} \sim 0.0000054 \qquad (4.43)$$

so that

$$W\left\{\left|\frac{m_i^0}{s} - p_i\right| < 0.0000054; \; (1 \le i \le 256)\right\} > 0.9960937 \qquad (4.44)$$

and

$$\| p - p' \| < 0.0000108 \qquad (4.45)$$

References

1. A. I. Khinchine, The entropy concept in probability theory, *Uspekhi Matematiceskih Nauk*, V. 8, N. 3, 1953, pp. 3-20 (Russian). English translation in A. I. Khinchine, *Mathematical Foundations of Information Theory*, Dover, New York, 1957, pp. 1-28.
2. M. Rosenblatt-Roth, The normed ε-entropy of sets and the theory of information. *Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes.* Publishing House of the Czechoslovak Academy of Sciences, Prague, 1960, pp. 569-577 (Russian).
3. M. Rosenblatt-Roth, Normed ε-entropy of sets and transmission of information from continuous sources through continuous channels, *Doklady Akademii Nauk SSSR*, 1960, V. 130, N. 2, pp. 265-268 (Russian). English translation in "Soviet mathematics", V. I, N. 1, 1960, pp. 48-50, published by the American Mathematical Society.
4. M. Rosenblatt-Roth, Approximations in information theory, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability Theory*, edited by L.M. Lecam and Jerry Neyman, University of California Press, Berkeley and Los Angeles, V. I, 1967, pp. 545-564.
5. M. Rosenblatt-Roth, On the best approximation of random sources of information with Markov chains, *Proceedings of the Seventeenth Annual Conference on Information Sciences and Systems*, Department of Electrical Engineering and Computer Science, Johns Hopkins University, 1983, pp. 255-260.
6. M. Rosenblatt-Roth, Random field identification from a sample. I. The independent case. Technical Report N. 1583, Center for Automation Research, Univ. of Maryland, College Park, 1985.
7. M. Rosenblatt-Roth, Random field identification from a sample. II. The simple Markov case. Technical Report N. 1599, Center for Automation Research, Univ. of Maryland, College Park, 1986.
8. M. Rosenblatt-Roth, The approximation of random vectors and information functionals, *Second International Symposium on Probability and Information Theory*, McMaster University, Hamilton, Canada, 1985 (to appear).
10. C.E. Shannon, The mathematical theory of communications, *Bell System Technical Journal*, V. 27, 1948, pp. 379-423 and 623-656.

Millu Rosenblatt-Roth
Center for Automation Research
University of Maryland
College Park, Maryland 20742

1. Introduction.

(a) In some problems of pattern recognition and image modeling, as well as in some aspects of statistical physics e.g. the Ising theory of ferromagnetism, the phenomena are described with the help of random fields. Because of the large number of random variables involved, as well as of the complexity of their interdependence, such random field representations cannot be handled in a straightforward manner.

Because none of the random variables can be deleted, the only way to simplify the study is to restrict their interdependence, i.e. to consider that the probabilities referring to each random variable depend only on the values of the random variables situated in some neighborhood of it, i.e., to consider random fields of a Markovian character.

Obviously, each problem may ask for some specific kind of neighborhood, so that it is necessary to use various classes of such particular random fields.

Sometimes even such simplifications are not sufficient and it is necessary to use various subclasses, containing Markovian random fields especially fit to describe unilateral developing phenomena.

(b) At this moment there does not exist a coherent theory of such random fields, but only isolated results, the most outstanding being contained in the oldest paper [1] dedicated to such problems, and unjustly forgotten today, due to the fact that those interested now in such studies are physicists, while this paper is published in a journal of information theory.

We mention that Chapter 2 of the present paper is a continuation and development of ideas and results contained in [1].

(c) It is of theoretical and practical importance to evaluate the error committed in the best approximation of the initial random field with such particular random fields of a given class, and to find that random field of this class which produces this minimization.

This paper presents the solution of this problem for two-dimensional rectangular random fields with arbitrary sets of states in each point.

We use the relative entropy as a measure of discrepancy and we determine explicitly for each class of unilateral Markov fields under discussion

(a) the Markov field which is the best approximation

(b) the expression of the error committed in this best approximation, which is a functional depending on the probability measure of the given random field.

We remark that in this study some approximation operators appear which are nonlinear projection operators in some Banach spaces.

2. The Concept of a Unilateral Markov Field.

Let σ be an array of points (i,j) $(1 \le i \le m, 1 \le j \le n)$ in a plane and τ some subset of it. Let ξ_{ij} be some random variable attached to the point (i,j), taking values in a measurable space

$$(X_{ij}, S_{ij}) , \qquad (i,j) \in \sigma \qquad (2.1)$$

A random field ξ is an array of arbitrarily dependent random variables

$$\xi_{ij} , \qquad (i,j) \in \sigma \qquad (2.2)$$

taking values in the measureable space

$$(X,S) = \underset{(i,j) \in \sigma}{X} (X_{ij}, S_{ij}) \qquad (2.3)$$

with joint probability measure

$$P_\xi(T) = P(\xi \in T) , \qquad T \in S \qquad (2.4)$$

For any set $\tau \subset \sigma$,

$$\xi^\tau = \{\xi_{ij}, \ (i,j) \in x^\tau\} \qquad (2.5)$$

is a random vector attached to the set τ and taking values in the measurable space

$$(X^\tau, S^\tau) = \underset{(i,j) \in \tau}{X} (X_{ij}, S_{ij}) \qquad (2.6)$$

with probability

$$P_{\xi^\tau}(T^\tau) = P(\xi^\tau \in T^\tau) , \qquad T^\tau \in S^\tau \qquad (2.7)$$

Obviously P_{ξ^τ} is a marginal of P_ξ.

Considering conditional marginals of P_ξ, we denote

$$P(\xi_{k\ell} \in T_{k\ell} \mid \xi_{ij} = x_{ij}, \ (i,j) \in \lambda)$$

$$= P_{\xi_{k\ell} \mid \xi_{ij}, \ (i,j) \in \lambda}(T_{k\ell} \mid x_{ij}, \ (i,j) \in \lambda) \qquad (2.8)$$

for any subset λ of σ, such that λ does not contain the point (k,ℓ).

Let us consider that $\alpha(k,\ell)$ is the set of all points (i,j) in the rectangle σ, such that $i \leq k$, $j \leq \ell$, with the exception of the point (k,ℓ), i.e.

$$\alpha(k,\ell) = \{(i,j); \; i \leq k, \; j \leq \ell\} - (k,\ell) \qquad (2.9)$$

In what follows we will denote by $\omega(k,\ell)$ some subset of $\alpha(k,\ell)$, and by $\beta(k,\ell)$ the set of all points (i,j) in σ such that either $i \leq k$ or $j \leq \ell$ or both, with the exception of the point (k,ℓ), i.e.

$$\beta(k,\ell) = \{(i,j); \; i \leq k, \; 1 \leq j \leq n\}$$
$$\cup \{(i,j); \; 1 \leq i \leq m, \; j \leq \ell\} - (k,\ell) \qquad (2.10)$$

<u>Definition.</u> *The random field ζ is unilateral of class ω if for each $(k,\ell) \in \sigma$ and any $T_{k\ell} \in S_{k\ell}$*

$$P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \beta(k,\ell)}(T_{k\ell}|x_{ij}, \; (i,j) \in \beta_{k\ell})$$
$$= P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \omega(k,\ell)}(T_{k\ell}|x_{ij}, \; (i,j) \in \omega(k,\ell))$$
$$(2.11)$$

LEMMA 1. *The unilateral random field ζ of class ω has its joint probability measure defined for any $T \in S$ by the expression*

$$P_{\zeta}(T) = \int_{T} \prod_{(k,\ell) \in \sigma} P_{\zeta_{k,\ell}|\zeta_{i,j}, \, (i,j) \in \omega(k,\ell)}(dx_{k\ell}|$$
$$| x_{ij}, \; (i,j) \in \omega(k,\ell))$$
$$(2.12)$$

Let $\omega(k,\ell)$ contain the points $A_{k,\ell}^{(s)}$ $(1 \leq s \leq N)$; let us consider a given value for s $(1 \leq s \leq N)$. Given an arbitrary point $(k,\ell) \in \sigma$, let us denote by

$$\theta[(u,v); \; (k,\ell); \; s] \qquad (2.13)$$

the set of points $(u,v) \in \sigma$ such that $\omega(u,v)$ contains the points

$$A_{u,v}^{(t)} \qquad (1 \leq t \leq N) \qquad (2.14)$$

with

$$A_{u,v}^{(s)} = (k,\ell) \qquad (2.15)$$

Let us denote

$$\omega'(k,\ell) = \bigcup_{s=1}^{N} \theta[(u,v); \; (k,\ell); \; s] - (k,\ell) \quad (2.16)$$

LEMMA 2. *Let us consider that (k,ℓ) is an arbitrary point in σ, and $T_{k\ell} \in S_{k,\ell}$, $(k,\ell) \in \sigma$. Then for each set*

$$\omega(k,\ell) \subset \alpha(k,\ell) \qquad (2.17)$$

there exists another set $\omega'(k,\ell) \subset \sigma$ such that

(a) $$\omega(k,\ell) \subset \omega'(k,\ell) \qquad (2.18)$$

(b) *for any unilateral random field of class ω, the following relation takes place,*

$$P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \sigma - (k,\ell)}(T_{k\ell}|x_{ij}, \; (i,j) \in \sigma - (k,\ell))$$

$$= P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \omega'(k,\ell)}(T_{k\ell}|x_{ij}, \, (i,j) \in \omega'(k,\ell))$$
$$(2.19)$$

for any point $(k,\ell) \in \sigma$.

(c) *the set $\omega'(k,\ell)$ is given by relation (2.16).*

Let r be a natural number. We denote by $\gamma_r(k,\ell)$ the set of points (i,j) in σ such that

$$|k-i| \leq r, \qquad |j-\ell| \leq r \qquad (2.20)$$

with the exception of (k,ℓ), i.e.

$$\gamma_r(k,\ell) = \{(i,j); \; k-r \leq i \leq k+r, \; \ell-r \leq j \leq \ell+r\} - (k,\ell)$$
$$(2.21)$$

<u>Definition.</u> **The random field ζ is an r-Markov field, if for any point $(k,\ell) \in \sigma$, and any set $T_{k\ell} \in S_{k\ell}$ the following relation takes place,**

$$P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \sigma - (k,\ell)}(T_{k\ell}|x_{ij}, \; (i,j) \in \sigma - (k,\ell))$$

$$= P_{\zeta_{k\ell}|\zeta_{ij}, \, (i,j) \in \gamma_r(k,\ell)}(T_{k\ell}|x_{ij}, \; (i,j) \in \gamma_r(k,\ell))$$
$$(2.22)$$

Let us denote now

$$\alpha_r(k,\ell) = \gamma_r(k,\ell) \cap \alpha(k,\ell) \qquad (2.23)$$

LEMMA 3. *For any unilateral random field ζ and any point $(k,\ell) \in \sigma$ the relation*

$$\omega(k,\ell) \subset \alpha_r(k,\ell) \qquad (2.24)$$

implies the relation

$$\omega'(k,\ell) \subset \gamma_r(k,\ell) \qquad (2.25)$$

LEMMA 4. *A unilateral random field ζ of any class ω is a Markov random field.*

EXAMPLES.

1. If

$$\omega(k,\ell) = \alpha_r(k,\ell) \qquad (2.26)$$

then

$$\omega'(k,\ell) = \gamma_r(k,\ell) \qquad (2.27)$$

so that the corresponding random field is an r-Markov random field.

In particular, let $r = 1$ and $\qquad (2.28)$

$$\omega(k,\ell) = \{(k-1,\ell),(k-1,\ell-1),(k,\ell-1)\} \quad (2.29)$$

Then

$$\omega'(k,\ell) = \{(k-\rho_1, \ell-\rho_2); \; \rho_1, \rho_2 \in (-k, 0, +1)\} - (k,\ell)$$

$$= \gamma_1(k,\ell) \qquad (2.30)$$

2. If

$$\omega(k,\ell) = \{(k-1,\ell),(k,\ell-1)\} \qquad (2.31)$$

it follows that

$$\omega'(k,\ell) = \{(k-\rho_1, \ell-\rho_2); \ \rho_1, \rho_2 \in (-1,0,+1), \ \rho_1 \neq \rho_2\} \tag{2.32}$$

3. The ω-amount of Information Determined by a Unilateral Random Field.

Given an arbitrary random field ζ with joint probability measure P_ζ, we can obviously calculate its marginal conditional probability measure

$$P_{\zeta_{k\ell}|\zeta_{ij}, (i,j) \in \omega(k,\ell)}(T_{k\ell}|x_{ij}, (i,j) \in \omega(k,\ell)) \tag{3.1}$$

for any given set $\omega(k,\ell)$ and any point $(k,\ell) \in \sigma$.

Consequently, for each given random field ζ, we can define a new random field $\zeta(\omega)$ with joint probability measure given by (2.12), with marginal conditional probability measures given by (3.1).

In what follows we will make essential use of the concept of relative entropy. Given two probability measures P_1, P_2 over some measurable space (X,S) the quantity $h(P_1:P_2)$ which takes the value

$$\int_X P_1(dx) \ \log \ \frac{P_1(dx)}{P_2(dx)} \tag{3.2}$$

if P_1 is absolutely continuous with respect to P_2, and $+\infty$ in the contrary case, is the relative entropy of P_1 with respect to P_2.

__Definition.__ *The quantity*

$$I_\omega(\zeta) = h(\zeta:\zeta(\omega)) \tag{3.3}$$

is the ω-amount of information determined by the arbitrary random field ζ.

THEOREM 1.
$$I_\omega(\zeta) \geq 0 \tag{3.4}$$

with equality iff

$$P_\zeta = P_{\zeta(\omega)} \tag{3.5}$$

i.e. *iff relation (2.11) holds for any $(k,\ell) \in \sigma$.*

Let us denote by L the totality of random fields defined over (X,S) and by $L(\omega)$ its subset containing the totality of random fields of class ω.

Let us consider an arbitrary random field $\eta(\omega) \in L(\omega)$ defined with the help of the conditional probability measures

$$P_{\eta_{k,\ell}|\eta_{ij}, (i,j) \in \omega(k,\ell)}(T_{k\ell}|x_{ij}, (i,j) \in \omega(k,\ell)) \tag{3.6}$$

so that its joint probability measure $P_{\eta(\omega)}$ is given by the expression

$$P_{\eta(\omega)}(T) = \int_T \prod_{(k,\ell) \in \sigma} P_{\eta_{k\ell}|\eta_{ij}, (i,j) \in \omega(k,\ell)}(T_{k\ell}|$$
$$|x_{ij}, (i,j) \in \omega(k,\ell)) \tag{3.7}$$

In what follows we will consider the expression

$$h(P_\zeta : P_{\eta(\omega)}) \tag{3.8}$$

representing the relative entropy of P_ζ with respect to $P_{\eta(\omega)}$.

THEOREM 2 (of best approximation). *Let ζ be an arbitrary random field. If $I_\omega(\zeta)$ is finite, then*

$$\min\{h(P_\zeta : P_{\eta(\omega)}); \ \eta(\omega) \in L(\omega)\} = I_\omega(\zeta) \tag{3.9}$$

This minimum is reached iff

$$P_{\eta(\omega)} = P_{\zeta(\omega)} \tag{3.10}$$

i.e. *iff*

$$P_{\eta_{k\ell}|\eta_{ij}, (i,j) \in \omega(k,\ell)} = P_{\zeta_{k\ell}|\zeta_{ij}, (i,j) \in \omega(k,\ell)} \tag{3.11}$$

for all points $(k,\ell) \in \sigma$.

LEMMA 5. *If for all points $(k,\ell) \in \sigma$,*

$$\tilde{\omega}(k,\ell) \subset \omega(k,\ell) \tag{3.12}$$

then, for any random field ζ,

$$I_{\tilde{\omega}}(\zeta) \leq I_\omega(\zeta) . \tag{3.13}$$

4. Two Measures of Discrepancy.

In order to evaluate the discrepancy between the given random field $\zeta \in L$ and the approximating random field $\zeta(\omega) \in L(\omega)$ we may also use the distance in variation between their corresponding probability measures,

$$\| P_\zeta - P_{\zeta(\omega)} \| \tag{4.1}$$

where the norm is the total variation of the signed measure

$$P_\zeta - P_{\zeta(\omega)} . \tag{4.2}$$

The relationship between (4.1) and (3.3) is given by the following

LEMMA 6.

$$\| P_\zeta - P_{\zeta(\omega)} \|^2 \leq 2 \ I_\omega(\zeta) \tag{4.3}$$

5. The Projection Operator.

Let us denote by $Z(\omega)$ the mapping of L into $L(\omega)$ such that

$$Z(\omega)\zeta = \zeta(\omega) , \quad \zeta \in L , \quad \zeta(\omega) \in L(\omega) \tag{5.1}$$

for $\omega \subset \sigma$.

Considering L imbedded in the Banach space of signed measures over (X,S), with norm the total variation, it follows that the norm of any probability measure is one, so that in particular

$$\| P_\zeta \| = \| P_{\zeta(\omega)} \| = 1 , \quad \omega \subset \sigma \tag{5.2}$$

1257

Consequently, the operator $Z(\omega)$ defined over L has norm one. It is easy to see that it is a *nonlinear idempotent* operator and its restriction on the set $L(\omega)$ is the identical operator, so that $Z(\omega)$ *is the projection* operator on $L(\omega)$. For these reasons it makes sense to introduce the following

Definition.

(a) $Z(\omega)$ is *the class ω projection operator.*

(b) For a given random field $\xi \in L$, the element

$$\xi(\omega) = Z(\omega)\xi \in L(\omega)$$

is its *class ω projection*, i.e., *its projection on $L(\omega)$*.

(c) In particular if

$$\omega(k,\ell) = \gamma_r(k,\ell) , \qquad (k,\ell) \in \sigma \qquad (5.4)$$

from (a), (b) we obtain the definitions of the r-Markov projection operator and of the r-Markov projection of a random field.

So the result (3.10) in Theorem 2 does express the fact that the minimum in (3.9) is reached iff $P_{\eta(\omega)}$ is the projection of $\xi \in L$ on $L(\omega)$.

6. On the Expression of $I_\omega(\xi)$.

Let us suppose that

(a) the probability measure P_ξ admits a probability density

$$p_\xi(x) , \quad x \in X \qquad (6.1)$$

(b) the probability measure $P_{\xi(\omega)}$ admits a probability density

$$p_{\xi(\omega)}(x) , \quad x \in X \qquad (6.2)$$

(c) the probability measure (2.11) admits a probability density

$$p_{\xi_{k\ell}|\xi_{ij}, (i,j)\in\omega(k,\ell)}(x_{k\ell}|x_{ij}, (i,j)\in\omega(k,\ell)),$$
$$(k,\ell) \in \sigma \qquad (6.3)$$

Let us denote

(a) $\qquad h(\xi) = -\int_X P_\xi(dx) \log p_\xi(x) \qquad (6.4)$

the (differential) entropy of the random field ξ

(b) $h(\xi_{k,\ell} | \xi_{ij}, (i,j)\in\omega(k,\ell))$

$= -\int_{X^{\omega(k,\ell)}} P_{\xi}^{\omega(k,\ell)}(dx^{\omega(k,\ell)}) h(\xi_{k\ell}|x_{ij},$
$(i,j) \in \omega(k,\ell)) \qquad (6.5)$

the conditional (differential) entropy of the random variable $\xi_{k,\ell}$ with respect to the random vector $\{\xi_{ij}, (i,j)\in\omega(k,\ell)\}$.

LEMMA 7. *If all quantities in the relation*

$$I_\omega(\xi) = \sum_{(k,\ell)\in\sigma} h(\xi_{k\ell}|\xi_{ij}, (i,j)\in\omega(k,\ell)) - h(\xi) \qquad (6.6)$$

exist and are finite, this equation holds.

From Theorem 1 and Lemma 7 there follows the

LEMMA 8. *For any random field ξ, and any class ω,*

$$h(\xi) \leq \sum_{(k,\ell)\in\sigma} h(\xi_{k\ell}|\xi_{ij}, (i,j) \in \omega(k,\ell))$$
$$(6.7)$$

with equality iff

$$I_\omega(\xi) = 0 \qquad (6.8)$$

i.e. ξ is a random field of class ω.

References

1. K. Abend, T. J. Harley, L.N. Kanal, Classification of binary random patterns, IEEE Trans. Info. Theory, v. IT-11, N. 4, pp. 538-542, 1965.
2. M.S. Bartlett, J.E. Besag, Correlation properties of some nearest-neighbor models, Bull. Intl. Stat. Inst., Proc. of 37th Session, London, v. 43, Book 2, pp. 191-193, 1969.
3. J. E. Besag, Nearest neighbor systems and the autologistic model for binary data, J. Royal Stat. Soc., Series B, v. 34, pp. 75-83, 1972.
4. J.E. Besag, Spatial interaction and the statistical analysis of lattice systems, J. Royal Stat. Soc., Series B, v. 36, pp. 192-235, 1974.
5. I.G. Enting, Crystal growth models and Ising models: disorder points, J. Phys. C: Solid State Physics, v. 10, pp. 1379-1388, 1977.
6. L. N. Kanal, Markov mesh models, in "Image Modeling," Ed. A. Rosenfeld, pp. 239-242, Academic Press, 1981.
7. P.A. Moran, Necessary conditions for markovian processes on a lattice, J. Appl. Prob., v. 10, pp. 605-612, 1973.
8. P.A. Moran, A gaussian markovian process on a square lattice, J. Appl. Prob., v. 10, pp. 54-62, 1973.
9. D.K. Pickard, A curious binary lattice process, J. Appl. Prob., v. 14, pp. 717-731, 1977.
10. D. K. Pickard, Unilateral Ising models, in "Spatial Patterns and Processes," Suppl. Adv. Appl. Prob., v. 10, pp. 58-64, 1978.
11. D. K. Pickard, Unilateral Markov Fields, Adv. Appl. Prob. v. 12, pp. 655-671, 1980.
12. M. Rosenblatt-Roth, On the best approximation of random sources of information with Markov chains, Proc. Seventeenth Annual Conference on Information Sciences and Systems, Dept. of Electr. Eng. and Computer Sci., The Johns Hopkins University, pp. 255-260, 1963.
13. M. Rosenblatt-Roth, The relative entropy of a random vector with respect to another random vector, Technical Report No. 85-35, Center for Multivariate Analysis, University of Pittsburgh, pp. 1-68, 1985.
14. M. Rosenblatt-Roth, The approximation of random vectors and information functionals, Second International Symposium on Probability and Information Theory, McMaster University, Hamilton, Canada, 1985 (to appear).
15. A. Rosenfeld, A. C. Kak, Digital picture processing, Academic Press, 1976.
16. A.M.W. Verhagen, An exactly soluble case of the triangular Ising model in a magnetic field, J. Stat. Physics, v. 15, N. 3, pp. 219-231, 1976.
17. A.M.W. Verhagen, A three parameter isotropic distribution of atoms and the hard-core square lattice gas, J. Chemical Physics, v. 67, N. 11, pp. 5060-5065, 1977.

18. T. R. Welberry, R. Galbraith, A two-dimensional model of crystal-growth disorder, J. Appl. Cryst. v. 6, pp. 87-96, 1973.
19. T. R. Welberry, R. Galbraith, The effect of nonlinearity on a two-dimensional model of crystal growth disorder, J. Appl. Cryst., v. 8, pp. 636-644, 1975.
20. T. R. Welberry, G. H. Miller, A phase transition in 3D growth-disorder model, Acta Cryst., v. A.34, pp. 120-123, 1978.

INTERPOLATION BY BIVARIATE QUADRATIC SPLINES
ON A NON-UNIFORM RECTANGULAR GRID

Charles Chui[1]
Department of Mathematics
Texas A & M University
College Station, TX 77843

Harvey Diamond[2]
Department of Mathematics
West Virginia University
Morgantown, WV 26506

Louise Raphael[2]
Department of Mathematics
Howard University
Washington, D.C. 20059

ABSTRACT. We study interpolation by bivariate quadratic splines on a criss-cross triangulated non-uniform rectangular grid. The data points are at the corners of the rectangles. We develop a quasi-interpolation formula giving optimal order of approximation $O(h^3)$ and provide an interpolation scheme whose order of approximation is $O(h^2)$. The results apply to functions in C^3 and for bounded mesh ratios.

I. INTRODUCTION. Given data $\{f(x_i,y_j)\}$, $i=1,..,m$; $j=1,..,n$ representing function values at the nodes $\{(x_i,y_j)\}$ of a non-uniform rectangular grid, we wish to produce a C^1 function interpolating the data and approximating the function f on the rectangle $[x_1,x_m] \times [y_1,y_n]$, where we assume $x_1 < x_2 \ldots < x_m$ and $y_1 < y_2 \ldots < y_n$. The class of functions from which the interpolant will be chosen is the space of C^1 bivariate quadratic splines developed by Chui and Wang [CW]. These splines are the C^1 functions which are quadratic polynomials on each triangle of a criss-cross triangulated rectangular grid. Our data points are assumed to be at the nodes of this rectangular grid, as show in Figure 1.

Fig.1
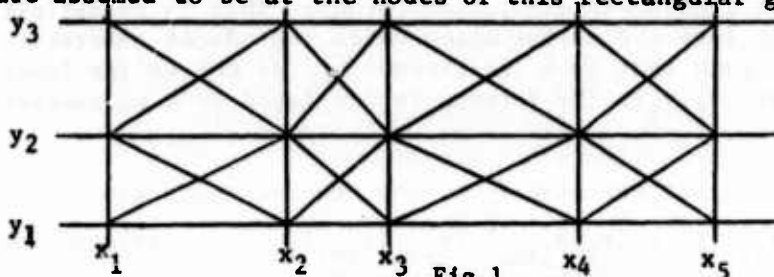
We remark that while it might be more desirable for the data points to be located at the centers of the rectangles of the grid defining the spline space, it is easy to construct examples for which this is not possible, e.g. $x_1=-3$, $x_2=-2$, $x_3=2$, $x_4=3$.

The more standard choice of interpolating space employs quadratic tensor splines. One drawback of this choice is the the high degree (four) results in poor shape preserving properties. While we do not investigate here the shape preserving properties of our interpolants, their lower degree would seem to warrant the development of basic results concerning interpolation and approximation.

The main results of this paper are as follows:

a) We provide a simple quasi-interpolation formula which (by definition) reproduces quadratic polynomials from their values at the data points, and which, when applied to data from arbitrary C^3 functions f, produces a spline having optimal order of approximation $O(h^3)$ where h is the maximum grid spacing.

b) We develop an explicit interpolation formula which produces an interpolant with approximation order $O(h^2)$ provided f is in C^3 and the global mesh ratio is bounded as $h \to 0$. The approximation order $O(h^2)$ for interpolation is optimal since it is known for one dimensional quadratic splines.

II. RESULTS. As developed in [CW], the spline space is spanned by the set of B-splines $B_{ij}(x,y)$ where the octagonal support of B_{ij} and its function values at the data points are shown below in Figure 2.
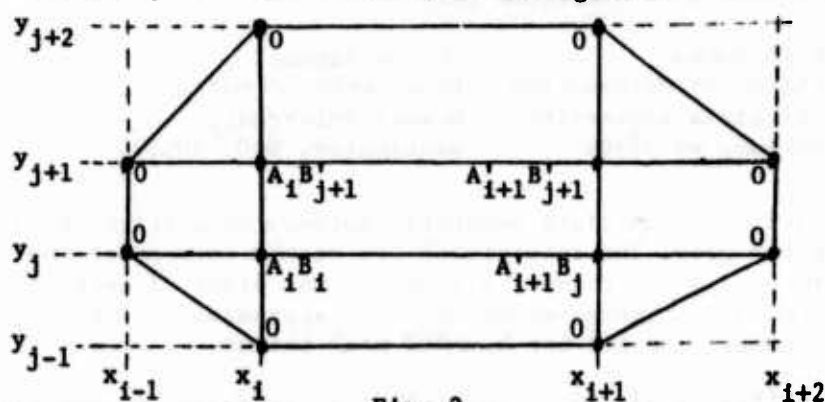
Fig. 2

A_i, A_i', B_j, B_j' are defined in terms of mesh ratios as follows:

$$A_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} \; ; \; A_i' = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}$$

$$B_j = \frac{y_j - y_{j-1}}{y_{j+1} - y_{j-1}} \; ; \; B_j' = \frac{y_{j+1} - y_j}{y_{j+1} - y_{j-1}}$$

A spline function takes the form $\sum \beta_{ij} B_{ij}(x,y)$, where the coefficients β_{ij} are to be determined. The interpolation equations are easily obtained as

$$L(\beta)_{ij} \equiv (A_i B_j) \beta_{ij} + (A_i' B_j) \beta_{i-1,j} + (A_i B_j') \beta_{i,j-1} + (A_i' B_j') \beta_{i-1,j-1} = f_{ij} \quad (1)$$

where $f_{ij} \equiv f(x_i, y_j)$ and $i=1,..,m; \; j=1,..,n$.

We proceed to construct a quasi-interpolation formula. This is by definition a linear mapping from the data into the spline space which reproduces quadratic polynomials from their data and such that each coefficient β_{ij} of the spline function depends locally on data near (x_i, y_j). The formula is developed by a successive guessing approach.

The B-spline $B_{ij}(x,y)$ has, as the corners of the central rectangle in its support, the data points (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}), (x_{i+1}, y_{j+1}). It is perhaps then reasonable to guess that its corresponding coefficient β_{ij} can be conveniently expressed in the form

$$\beta_{ij} = \frac{f_{ij} + f_{i+1,j} + f_{i,j+1} + f_{i+1,j+1}}{4} + e_{ij} \quad (2)$$

where e_{ij} represents an error term. Plugging the formula (2) into the interpolation equations (1) gives an equation for the e_{ij}:

$$L(e)_{ij} = - \frac{(x_i - x_{i-1})(x_{i+1} - x_i)}{2} n_{ij}^{(1)} - \frac{(y_j - y_{j-1})(y_{j+1} - y_j)}{2} n_{ij}^{(2)} \quad (3)$$

where $n_{ij}^{(1)}$ and $n_{ij}^{(2)}$ are combinations of second divided differences of the data in the x and y directions respectively:

$$n_{ij}^{(1)} = \frac{B_j'}{4} f([x_{i-1}, x_i, x_{i+1}], y_{j-1}) + \frac{3}{4} f([x_{i-1}, x_i, x_{i+1}], y_j) + \frac{B_j}{4} f([x_{i-1}, x_i, x_{i+1}], y_{j+1})$$

$$n_{ij}^{(2)} = \frac{A_i'}{4} f(x_{i-1}, [y_{j-1}, y_j, y_{j+1}]) + \frac{3}{4} f(x_i, [y_{j-1}, y_j, y_{j+1}]) + \frac{A_i}{4} f(x_{i+1}, [y_{j-1}, y_j, y_{j+1}]) \quad (4)$$

1262

We note that both $\eta_{ij}^{(1)}$ and $\eta_{ij}^{(2)}$ are constants if f is a quadratic. If this is the case, it is a simple matter of substitution to show that (3) is satisfied if $e_{ij} = -\frac{1}{2}(x_{i+1}-x_i)^2\eta_{ij}^{(1)} - \frac{1}{2}(y_{j+1}-y_j)^2\eta_{ij}^{(2)}$. We can now write our quasi-interpolation formula.

Theorem 1: If f(x,y) is a quadratic polynomial and the coefficients β_{ij} are given by

$$\beta_{ij} = \frac{f_{ij}+f_{i+1,j}+f_{i,j+1}+f_{i+1,j+1}}{4} - \frac{1}{2}(x_{i+1}-x_i)^2\eta_{ij}^{(1)} - \frac{1}{2}(y_{j+1}-y_j)^2\eta_{ij}^{(2)} \equiv q(f)_{ij} \qquad (5)$$

where the rectangular grid with nodes $\{(x_i,y_j)\}$ covers the plane, then

$$Q(f)(x,y) \equiv \sum q(f)_{ij}B_{ij}(x,y) \qquad (6)$$

is equal to f(x,y) for all (x,y).

Proof: The previous arguments have shown that if $\beta_{ij} = q(f)_{ij}$ then $\sum\beta_{ij}B_{ij}(x,y)$ interpolates f(x,y) at the nodes of the grid. It remains to show only that Q(f)(x,y) is in fact a quadratic polynomial, in which case it must coincide with f(x,y). Unfortunately, this task must apparently be carried out by brute force calculation. One approach is to calculate the second derivatives of $B_{ij}(x,y)$ (which are constant) and then successively choosing $f(x,y) = 1,x,y,x^2,y^2,xy,$ to show that the second derivatives of Q(f) are constant. Of course various symmetries will reduce the amount of actual calculation. We omit the details here.

Henceforth, Q(f) as defined by (6) (via (4) and (5)) will be referred to as the quasi-interpolant of f. Standard arguments show that if f is in C^3 on the closed rectangle $[x_1,x_m] \times [y_1,y_n]$ then

$$|f(x,y) - Q(f)(x,y)| = O(h^3) \qquad (7)$$

for (x,y) in the interior of the rectangle.

We now develop our interpolation scheme and investigate its order of approximation. The interpolation scheme takes the form

$$\beta_{ij} = q(f)_{ij} + \epsilon_{ij} \qquad (8)$$

where ϵ_{ij} satisfies

$$L(\epsilon_{ij}) = f_{ij} - L(q(f))_{ij} \equiv g_{ij} . \qquad (9)$$

Since g_{ij} depends locally and linearly and is identically zero for quadratic data, it follows from a standard Taylor polynomial approximation argument that if $f \in C^3$ on the rectangle defined by the grid points, then $g_{ij} = O(h^3)$ in the interior of the rectangle. If we could show that ϵ_{ij} satisfying (9) was also $O(h^3)$ then the interpolant would differ from the quasi-interpolant by $O(h^3)$ and so provide the same order of approximation. Unfortunately, $\epsilon_{ij} = O(h^3)$ is not necessarily true.

It is possible to write down an explicit solution for ε_{ij}, since there is a simple formula for the fundamental function $\tilde{\beta}_{ij}^{rs}$ satisfying

$$L(\tilde{\beta}_{ij}^{rs}) = \delta_{ir}\delta_{js} \tag{10}$$

namely

$$\tilde{\beta}_{ij}^{rs} = \frac{(x_{r+1}-x_{r-1})(y_{s+1}-y_{s-1})}{(x_{r+1}-x_r)(x_r-x_{r-1})(y_{s+1}-y_s)(y_s-y_{s-1})}(x_{i+1}-x_i)(y_{j+1}-y_j)\sigma_{ij}^{rs} \tag{11}$$

where σ_{ij}^{rs} as a function of i and j for a given r and s, is ± 1 according to the following pattern in Figure 3:

```
+ - + - - + - + - +
- + - + + - + - + -
+ - + - - + - + - +   The circled + is situated at the (r,s) position.
- + - +⊕ - + - + -
- + - + + - + - + -
+ - + - - + - + - +   Fig. 3
```

We can then write

$$\varepsilon_{ij} = \sum_{r,s} \tilde{\beta}_{ij}^{rs} g_{rs} \tag{12}$$

where, for fixed i and j, $\tilde{\beta}_{ij}^{rs}$ has the same sign pattern as in Figure 3, with (i,j) and (r,s) interchanged. If the mesh ratios are bounded then $\tilde{\beta}_{ij}^{rs} = O(1)$ and $\varepsilon_{ij} = O(mn)\|g_{rs}\| = O(h)$. It is in fact possible that ε_{ij} has the same order of magnitude as h, for the g_{rs} may alternate in sign due to oscillations in the mesh spacing. While this would suggest that $|f(x,y) - \sum \varepsilon_{ij}B_{ij}(x,y)| = O(h)$ is best possible for the interpolant, in fact $O(h^2)$ can be obtained for it turns out that $\sum \varepsilon_{ij}B_{ij}(x,y) = O(h^2)$ even if the ε_{ij} as given by (12) are $O(h)$. We have

$$\sum_{i,j} \varepsilon_{ij}B_{ij}(x,y) = \sum_{i,j}\sum_{r,s} \tilde{\beta}_{ij}^{rs}g_{rs}B_{ij}(x,y) = \sum_{r,s}g_{rs}\left(\sum_{i,j}\tilde{\beta}_{ij}^{rs}B_{ij}(x,y)\right) \tag{13}$$

Next we use the linear dependence of the B_{ij} as shown in [CW]:

$$\sum_{i,j}(-1)^{i+j}(x_{i+1}-x_i)(y_{j+1}-y_j)B_{ij}(x,y) \equiv 0. \tag{14}$$

It is not difficult to see now that $\sum_{i,j}\tilde{\beta}_{ij}^{rs}B_{ij}(x,y)$ by virtue of (11) and (14) is, for a fixed (x,y), zero outside of a cross-shaped region centered at (x,y) whose arms are only a bounded number of nodes in width and which consequently contains $O(m+n)$ nodes. It follows that $\sum_{r,s}g_{rs}\left(\sum_{i,j}\tilde{\beta}_{ij}^{rs}B_{ij}(x,y)\right)$ has only $O(m+n)$ non-zero terms so that $\sum_{i,j}\varepsilon_{ij}B_{ij}(x,y) = O(m+n)\,g_{rs} = O(h^2)$ using (13).

The approximation order of the interpolant can then be estimated as

$$|f(x,y) - \sum \beta_{ij} B_{ij}(x,y)| \leq |f(x,y) - \sum q(f)_{ij} B_{ij}(x,y)| + |\sum \epsilon_{ij} B_{ij}(x,y)|$$
$$= O(h^3) + O(h^2) = O(h^2).$$

The above discussion requires some formalizing, especially as regards the lack of data beyond the boundary of $[x_1, x_m] \times [y_1, y_n]$. The precise statement of the approximation result is given in the following Theorem:

Theorem 2: Suppose the following hold:

a) $f \in C^3(\Omega)$, where Ω is an open set containing the rectangle $E = [a,c] \times [b,d]$.

b) (x_i, y_j), $i=1,..,m$; $j=1,..,n$ is any set of grid nodes such that

i) $(x_1, y_j) \in E$

ii) $\max \{|\frac{h}{x_{i+1}-x_i}|, |\frac{h}{y_{j+1}-y_j}|\} < M$, where M is a constant independent of

m and n and $h = \max \{|x_{i+1}-x_i|, |y_{j+1}-y_j|\}$.

c) The underlined augmented data set $\{(x_i, y_j, f_{ij})\}$ is defined as follows:

i) $f_{ij} = f(x_i, y_j)$, $i=1,..,m$; $j=1,..,n$

ii) $x_{i+1}-x_i = h$ for $i \geq m$ and $i \leq 0$

$y_{j+1}-y_j = h$ for $j \geq n$ and $j \leq 0$

f_{ij} is defined for all $(i,j) \notin \{1,..,m\} \times \{1,..,n\}$

using a local extrapolation formula exact for quadratics.

d) The coefficients β_{ij} of the spline function $\sum \beta_{ij} B_{ij}(x,y)$, $i=0,..,m+1$; $j=0,..,n+1$ are chosen as follows:

$$\beta_{ij} = q(f)_{ij} + \epsilon_{ij}$$

where $q(f)$ is defined in (5) (via (4)) and ϵ_{ij} is defined by

$$\epsilon_{ij} = \sum_{\substack{r=1,..,m \\ s=1,..,n}} \tilde{\beta}_{ij}^{rs} g_{rs}$$

where g_{rs} is defined in (9) and $\tilde{\beta}_{ij}^{rs}$ is defined in (11).

Then $\sum \beta_{ij} B_{ij}(x,y)$ interpolates $f(x,y)$ on the set $\{(x_i, y_j)\}$, $i=1,..,m$; $j=1,..,n$ and $|f(x,y) - \sum \beta_{ij} B_{ij}(x,y)| = O(h^2)$ uniformly for $(x,y) \in [x_1, x_m] \times [y_1, y_n]$.

1265

Proof: The extrapolated values of f_{ij} differ from the actual values $f(x_i, y_j)$ by $O(h^3)$ for (x_i, y_j) within $O(h)$ of the rectangle $[x_1, x_m] \times [y_1, y_n]$. If follows that if $q(f)_{ij}$ is calculated from the augmented data set, we still have $|f(x,y) - q(f)_{ij} B_{ij}(x,y)| = O(h^3)$ for (x,y) in the rectangle, uniformly. Similarly, when calculated using the augmented data set, $g_{ij} = O(h^3)$ for (i,j) in the original data set. The sum in (11) is restricted to (r,s) in the original data set since this insures interpolation on the original data points and minimizes the number of terms in the sum. The rest of the proof follows from previous arguments.

We remark that there is no essential requirement that the data be given on an entire rectangular grid $\{x_i\}_{i=1}^m \times \{y_j\}_{j=1}^n$. Our quasi-interpolant and interpolation scheme apply just as well if say, we wish to interpolate at the nodes of a rectangular grid lying inside a prespecified open region.

References

[CW] C.K. Chui and R.H. Wang, Bivariate B-splines on triangulated rectangles, in *Approximation Theory IV*, ed. by C.K. Chui, L.L. Schumaker, and J.D. Ward, Academic Press, N.Y., 1983, pp. 413-418

ON THE C² CONTINUITY OF PIECEWISE CUBIC HERMITE POLYNOMIALS
WITH UNEQUAL INTERVALS

C. N. Shen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Close Combat Armaments Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

ABSTRACT. Cubic hermite polynomials are usually C² continuous. With the introduction of smoothing within the intervals, the second derivatives can be made continuous. This may be applied to the autonomous vehicle problem with unequal laser scanning.

In using a laser range finder to measure the range, the direction of these laser rays can be subjected to angular errors. These errors, in the direction of the elevation angle, affect the determination of in-path slopes for navigation of autonomous vehicles. Nonuniform grid may be employed in computation by the spline function method with cubic hermite polynomials. For the purpose of smoothing, it is essential to obtain continuous second derivatives at the grid point from both sides.

I. INTRODUCTION. The smoothing of gradients can be obtained by using an optimization method for approximation involving spline functions. Nonuniform grid may be employed in computation by the spline function method with cubic hermite polynomials. Continuous second derivatives at the grid point from both sides are essential for the purpose of smoothing. This method can be applied to solve the following problems: Whether the platform can climb on the estimated in-path slope or whether it will tip over the estimated cross-path slope.

II. RECURSIVE FILTERING AND SMOOTHING PROCEDURE. A spline function $s(\xi)$ is a solution to the optimization problem

$$J^* = \min_{h \in C^2} \left\{ \sum_{i=1}^{N} [h(\beta_i) - m_i]^T R_i^{-1} [h(\beta_i) - m_i] + \rho \sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{h}]^2 d\xi \right\} \tag{1}$$

where for clarity and simplicity in discussion we only consider the cubic spline case. Higher order polynomial spline can also be treated in a similar manner with more complicated computation.

A cubic spline, s, is a piecewise polynomial of class C² which has many good properties, such as the minimum norm property and local base property [1,2]. From the approximation theory we know that for each set $A = \{a_1,...,a_N, a'_1, a'_N\}$, there exists a unique cubic spline $s(\xi;A)$ such that

$$s(\beta_i;A) = a_i , \quad i = 1,2,...,N \tag{2}$$

$$\dot{s}(\beta_i;A) = a'_i, \quad i = 1,N \tag{3}$$

where $\dot{s}$ is the first derivative of the function s. The above equations can be thought of as boundary conditions for the piecewise cubic spline interpolation given a set of data (β_i, a_i), for $i = 1, 2, \ldots, N$. Thus, solving the problem in Eq. (1) is equivalent to determining a set of constraints A for the optimization problem:

$$J^\star = \underset{A}{\text{Min}} \left\{ \sum_{i=1}^{N} [s(\xi_i;A) - m_i]^T R_i^{-1} [s(\xi_i;A) - m_i] + \rho \sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{s}(\xi;A)]^2 d\xi \right\} \tag{4}$$

Instead of taking a direct approach to find an optimal set of constraints for the problem above, it is proposed to further transform this problem into a form which is convenient to be solved. From the theory of numerical analysis [3], it is well known that a piecewise cubic Hermite polynomial $p(\xi)$ is in the family of C^1. For each set $B = A \cup A^C$, where A^C is a complement of A, i.e., $A^C = \{a'_i, i = 2, 3, \ldots, N-1\}$, then $B = \{a_i, a'_i, i = 1, 2, \ldots, N\}$, there exists a unique piecewise cubic Hermite polynomial $p(\xi;A)$ such that

$$p(\beta_i;B) = a_i \quad , \quad i = 1, 2, \ldots, N \tag{5}$$

$$\dot{p}(\beta_i;B) = a'_i \quad , \quad i = 2, \ldots, N \tag{6}$$

where $\dot{p}$ is the first derivative of p.

It should also be noted that for each set A, there are an infinite number of piecewise Hermite polynomials $p(\xi;A)$ such that

$$p(\beta_i;A) = a_i \quad , \quad i = 1, 2, \ldots, N \tag{7}$$

$$\dot{p}(\beta_i;A) = a'_i \quad , \quad i = 1, N \tag{8}$$

Let a set of $p(\xi;A)$ which satisfies the constraints in the equations above be P, i.e.,

$$P = \{p(\xi;A):(5),(6) \text{ satisfied}\} \tag{9}$$

With reference to the paper by de Boor [4], it is noted that there exists a unique cubic spline $s(\xi;A)$ in the set P. Also from the minimum norm property of a cubic spline we have the following relation

$$\sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{s}(\xi;A)]^2 d\xi \leq \sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{p}(\xi;A)]^2 d\xi \tag{9}$$

That is

$$\sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{s}(\xi;A)]^2 d\xi = \underset{p \in P}{\inf} \ J_p(p) \tag{10}$$

where

$$J_p = \sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{p}(\xi;A)] d\xi \tag{11}$$

1268

Since a cubic spline $s(\xi;A)$ is unique, a piecewise cubic Hermite polynomial $p(\xi;A)$ which minimizes the smoothing integral J_p in the above equation with respect to A^C becomes a cubic spline $s(\xi;A)$. To be more precise, we have the following theorem.

THEOREM: Let P represent a set of piecewise cubic Hermite polynomial p which satisfies the constraints below

$$p(\beta_i;A^C) = a_i \quad , \quad i = 1,2,\ldots N \tag{12}$$

$$\dot{p}(\beta_i;A^C) = a'_i \quad , \quad i = 1,N \tag{13}$$

where $p \in C^1$, A, and A^C are the same as mentioned before. Then there exists a unique cubic spline $s(\xi)$ such that

$$\sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{s}(\xi)]^2 d\xi = \operatorname*{Min}_{A^C} \sum_{i=2}^{N} \int_{\beta_{i-1}}^{\beta_i} [\ddot{p}(\xi,A^C)]^2 d\xi \tag{14}$$

where $\ddot{s}$ and $\ddot{p}$ are the second derivatives of functions s and p and $s \in C^2$. A simple example with N = 3 is given next.

III. EXAMPLE FOR C^2 CONTINUITY. For convenience and simplicity, we only consider a special case with N = 3. The node points are given as β_1, β_2, and β_3. The intervals are not equal, i.e.,

$$(\beta_2-\beta_1) \neq (\beta_3-\beta_2) \tag{15}$$

Let a set of piecewise cubic Hermite polynomial $p(t)$ be

$$P = [p(t;A^C) \quad , \quad p \in C^1 [t_1,t_3], \dot{p}(t_2) = a, a \in A^C] \tag{16}$$

which satisfies the constraints in the equations below

$$p(t_i;A^C) = a_i \quad , \quad \text{for } i = 1,2,3$$

$$\dot{p}(t_i;A^C) = a'_i \quad , \quad \text{for } i = 1,3 \tag{17}$$

In this special case, a set $A^C = a'_2 = a$.

We want to show here that the cubic Hermite polynomial $p(t;A^C)$ which is obtained by minimizing the smoothing integral will become a cubic spline function $s(t) \in C^2[t_1,t_3]$

$$J^* = \operatorname*{Min}_{A^C} \left\{ \int_{t_1}^{t_2^-} [\ddot{p}(t;A^2)]^2 dt + \int_{t_2^+}^{t_3} [\ddot{p}(t;A^C)]^2 dt \right\}$$

$$= \operatorname*{Min}_{a} \left\{ \int_{t_1}^{t_2^-} [\ddot{p}(t;a)]^2 dt + \int_{t_2^+}^{t_3} [\ddot{p}(t;a)]^2 dt \right\} \tag{18}$$

From Eq. (A14) of the Appendix, the smoothing integral above can be written as

$$J(a) = (x_2 - A_1 x_1)^T B_1^{-1} (x_2 - A_1 x_1) + (x_3 - A_2 x_2)^T B_2^{-1} (x_3 - A_2 x_2) \tag{19}$$

where A_i, B_i^{-1}, and x_i are defined in the Appendix, and

$$x_i = (a_i, a'_i)^T, \quad \text{with } a'_2 = a, \quad i = 1,2,3 \tag{20}$$

$$\Delta_{i-1} = d_{i-1} = t_i - t_{i-1} \tag{21}$$

Using Eqs. (A11) and (A12), the functional $J(a)$ is written as

$$\left[\begin{bmatrix} a_2 \\ a \end{bmatrix} - \begin{bmatrix} 1 & d_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a'_1 \end{bmatrix} \right]^T \begin{bmatrix} 12d_1^{-3} & -6d_1^{-2} \\ -6d_1^{-2} & 4d_1^{-1} \end{bmatrix} \left[\begin{bmatrix} a_2 \\ a \end{bmatrix} - \begin{bmatrix} 1 & d_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a \end{bmatrix} \right]$$

$$+ \left[\begin{bmatrix} a_3 \\ a'_3 \end{bmatrix} - \begin{bmatrix} 1 & d_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_2 \\ a \end{bmatrix} \right]^T \begin{bmatrix} 12d_2^{-3} & -6d_2^{-2} \\ -6d_2^{-2} & 4d_2^{-1} \end{bmatrix} \left[\begin{bmatrix} a_3 \\ a'_3 \end{bmatrix} - \begin{bmatrix} 1 & d_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_2 \\ a \end{bmatrix} \right]$$

$$J(a) = 12d_1^{-3} (a_2 - a_1 - d_1 a'_1)^2 - 12d_1^{-2} (a_2 - a_1 - d_1 a'_1)(a - a'_1)$$

$$+ 4d_1^{-1} (a - a'_1)^2 + 12d_2^{-3} (a_3 - a_2 - d_2 a)^2$$

$$- 12d_2^{-2} (a_3 - a_2 - d_2 a)(a'_3 - a) + 4d_2^{-1} (a'_3 - a)^2 \tag{22}$$

Taking the partial derivative with respect to a yields

$$\frac{\partial J}{\partial a} = -12d_1^{-2} (a_2 - a_1 - d_1 a'_1) + 8d_1^{-1} (a - a'_1)$$

$$+ 24d_2^{-3} (a_3 - a_2 - d_2 a)(-d_2) - 12d_2^{-3} (-d_2)(a'_3 - a)$$

$$- 12d_2^{-2} (-1)(a_3 - a_2 - d_2 a) - 8d_2^{-1} (a'_3 - a) = 0 \tag{23}$$

Solving the equation above for a, one obtains

$$a^* = [3d_1^{-2}(a_2-a_1)-d_1^{-1}a'_1+3d_2^{-2}a_3-d_2^{-2}3a_2-d_2^{-1}a'_3]/[2(d_1^{-1}+d_2^{-1})] \qquad (24)$$

To show that $p(t;a^*)\epsilon C^2[t_1,t_3]$, we only need to show that

$$\lim_{t\to t_2^-} p(t;a^*) = \lim_{t\to t_2^+} \ddot{p}(t;a^*) \qquad (25)$$

That is, for piecewise cubic Hermite polynomial $p(t)$,

$$\ddot{p}_{1,2}(t_2;a^*) = \ddot{p}_{2,3}(t_2;a^*) \qquad (26)$$

where p_{12} is the cubic Hermite polynomial within the interval β_1 and β_2, and p_{21} is the cubic Hermite polynomial within the interval β_2 and β_3.

Now from the definition of piecewise cubic Hermite polynomial in the Appendix, we have

$$\ddot{p}_{1,2}(t_2;a^*) = 6d_1^{-2}(a_1-a_2) + 2d_1^{-1}a'_1 + 4_1^{-1}a^* \qquad (27)$$

By using Eq. (24), the above equation can be expressed as

$$\ddot{p}_{1,2}(t_2;a^*) = [-6a_2(d_1^{-1}+d_2^{-1}) + 6(a_1d_1^{-1}+a_3d_2^{-1} + 2(a'_1-a'_3)]/(d_1+d_2) \qquad (28)$$

In a like manner, omitting the detailed derivation, we obtain easily

$$\ddot{p}_{2,3}(t_2;a^*) = [-6a_2(d_1^{-1}+d_2^{-1}) + 6(a_1d_1^{-1}+a_3d_2^{-1} + 2(a'_1-a'_3)]/(d_1+d_2) \qquad (29)$$

Thus, Eq. (26) is always true, that is, the conclusion in the Theorem is valid. It is proved that the C^2 continuity exists in the optimization procedure for piecewise cubic Hermite polynomials with unequal intervals.

IV. CONCLUSION. For scanning in the direction of elevation angle from the top of a mast where a laser is located, the intervals needed in angles are small for far away targets, while the same are large for close-by objects. The smoothing algorithm discussed in this paper indicates that cubic Hermite polynomials can be used for unequal intervals or nonuniform grids.

REFERENCES

1. Ahlberg, J. H., Nilson, E. N., and Walsh, J. L., The Theory of Splines and Their Applications, Academic Press, Inc., 1967.

2. Schumaker, L. L., Spline Functions: Basic Theory, John Wiley & Sons, 1981.

3. Burden, R. L. et al, Numerical Analysis, Prindle, Weber, & Schmidt, 1978.

4. de Boor, C., "Bicubid Spline Interpolation," J. Math Phys., Vol. 41, 1962, pp. 212-218.

APPENDIX

EVALUATION OF THE SMOOTHING INTEGRAL

A piecewise cubic Hermite polynomial in the interval $[\beta_{i-1},\beta_i]$ is represented in terms of the basis functions and the state vectors x_i, x_{i-1}, where the state vectors are defined as in Eq. (20). By changing the independent variable below,

$$t = \xi - \beta_{i-1} \tag{A1}$$

Then the smoothing integral in the interval $[\beta_{i-1},\beta_i]$ becomes

$$I_{i-1,i} = \int_0^{\Delta_{i-1}} [\ddot{p}_{i-1,i}(t)]^2 dt \tag{A2}$$

where $\Delta_{i-1} = t_i - t_{i-1} = \beta_i - \beta_{i-1}$, $\Delta_{i-1} \neq \Delta_i$.

With the change of the variable above, the second derivative of the Hermite polynomial can be written as

$$p_{i-1,i}(t) = \begin{bmatrix} \ddot{\phi}_{i,1}(t) \\ \ddot{\psi}_{i,1}(t) \\ \ddot{\phi}_{i,0}(t) \\ \ddot{\psi}_{i,0}(t) \end{bmatrix}^T \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} \tag{A3}$$

$$\tag{A4}$$

where the second derivatives of the basis functions can be derived as follows.

Using the change of variables, we rewrite the basis functions as

$$\phi_{i,1}(t) = t^2(3\Delta_{i-1} - 2t)/\Delta_{i-1}^3$$

$$\psi_{i,1}(t) = t^2(t - \Delta_{i-1})/\Delta_{i-1}^2$$

$$\phi_{i,0}(t) = (\Delta_{i-1} - t)^2(\Delta_{i-1} + 2t)/\Delta_{i-1}^3$$

$$\psi_{i,0}(t) = t(\Delta_{i-1} - t)^2 \Delta_{i-1}^2 \tag{A5}$$

Then, taking the second derivative with respect to t yields

$$\ddot{\phi}_{i,1}(t) = 6(\Delta_{i-1}-2t)/\Delta_{i-1}^3$$

$$\ddot{\psi}_{i,1} = (6t-2\Delta_{i-1})/\Delta_{i-1}^2$$

$$\ddot{\phi}_{i,0} = 6(2t-\Delta_{i-1})/\Delta_{i-1}^3$$

$$\ddot{\psi}_{i,0} = (6t-4\Delta_{i-1})/\Delta_{i-1}^2 \tag{A6}$$

Therefore, the integrand of the smoothing integral is expressed as

$$[\ddot{p}_{i-1,i}(t)]^2 = \begin{bmatrix} x_i \\ \\ x_{i-1} \end{bmatrix}^T \begin{bmatrix} K_{i-1,i}(t) \end{bmatrix} \begin{bmatrix} x_i \\ \\ x_{i-1} \end{bmatrix} \tag{A7}$$

where $K_{i-1,i}$ is defined as

$$K_{i-1,i}(\mu) \triangleq$$

$$\begin{bmatrix} \ddot{\phi}_{i,1}(\mu)\ddot{\phi}_{i,1}(\mu) & , & \ddot{\phi}_{i,1}(\mu)\ddot{\psi}_{i,1}(\mu) & , & \ddot{\phi}_{i,1}(\mu)\ddot{\phi}_{i,0}(\mu) & , & \ddot{\phi}_{i,1}(\mu)\ddot{\psi}_{i,0}(\mu) \\ \ddot{\psi}_{i,1}(\mu)\ddot{\phi}_{i,1}(\mu) & , & \ddot{\psi}_{i,1}(\mu)\ddot{\psi}_{i,1}(\mu) & , & \ddot{\psi}_{i,1}(\mu)\ddot{\phi}_{i,0}(\mu) & , & \ddot{\psi}_{i,1}(\mu)\ddot{\psi}_{i,0}(\mu) \\ \ddot{\phi}_{i,0}(\mu)\ddot{\phi}_{i,1}(\mu) & , & \ddot{\psi}_{1,0}(\mu)\ddot{\psi}_{i,1}(\mu) & , & \ddot{\psi}_{i,0}(\mu)\ddot{\phi}_{i,0}(\mu) & , & \ddot{\psi}_{i,0}(\mu)\ddot{\psi}_{i,0}(\mu) \\ \ddot{\psi}_{i,0}(\mu)\ddot{\phi}_{i,1}(\mu) & , & \ddot{\psi}_{1,0}(\mu)\ddot{\psi}_{i,1}(\mu) & , & \ddot{\psi}_{i,0}(\mu)\ddot{\phi}_{i,0}(\mu) & , & \ddot{\psi}_{i,0}(\mu)\ddot{\psi}_{i,0}(\mu) \end{bmatrix}$$

$$\tag{A8}$$

By utilizing the above equation, the smoothing integral becomes

$$I_{i-1,i} = \begin{bmatrix} x_i \\ \\ x_{i-1} \end{bmatrix} \begin{bmatrix} \int_0^{\Delta_{i-1}} K_{i-1,i}(t)dt \end{bmatrix} \begin{bmatrix} x_i \\ \\ x_{i-1} \end{bmatrix} \tag{A9}$$

Evaluating the above integral, we obtain

1273

$$\int_0^{\Delta_{i-1}} K_{i-1,i}(t)dt =$$

$$\begin{bmatrix} 12/\Delta_{i-1}^3 & -6/\Delta_{i-1}^2 & -12/\Delta_{i-1}^3 & -6/\Delta_{i-1}^2 \\ -6/\Delta_{i-1}^2 & 4/\Delta_{i-1} & 6/\Delta_{i-1}^2 & 2/\Delta_{i-1} \\ -12/\Delta_{i-1}^3 & 6/\Delta_{i-1}^2 & 12/\Delta_{i-1}^3 & 6/\Delta_{i-1}^2 \\ -6/\Delta_{i-1}^2 & 2/\Delta_{i-1} & 6/\Delta_{i-1}^2 & 4/\Delta_{i-1} \end{bmatrix} \qquad \text{(A10)}$$

By defining matrices B_{i-1}^{-1} and A_{i-1} as follows

$$A_{i-1} = \begin{bmatrix} 1 & \Delta_{i-1} \\ 0 & 1 \end{bmatrix} \qquad \text{(A11)}$$

$$B_{i-1}^{-1} = \begin{bmatrix} 12\Delta_{i-1}^{-3} & -6\Delta_{i-1}^{-2} \\ -6\Delta_{i-1}^{-2} & 4\Delta_{i-1}^{-1} \end{bmatrix} \qquad \text{(A12)}$$

where B_{i-1}^{-1} is a symmetric matrix.

Equation (A10) can be expressed as

$$\begin{bmatrix} B_{i-1}^{-1} & -B_{i-1}^{-1}A_{i-1} \\ (-B_{i-1}^{-1}A_{i-1})^T & A_{i-1}^T B_{i-1}^{-1} A_{i-1} \end{bmatrix} \qquad \text{(A13)}$$

where B_{i-1}^{-1} and A_{i-1} are functions of the variable Δ_{i-1}. By using the above notation, Eq. (A9) is rewritten as

$$
\begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}^T
\begin{bmatrix} B_{i-1}^{-1} & -B_{i-1}^{-1}A_{i-1} \\ -A_{i-1}^T B_{i-1}^{-1} & A_{i-1}^T B_{i-1}^{-1} A_{i-1} \end{bmatrix}
\begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}
$$

$$
=
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}^T
\begin{bmatrix} A_{i-1}^T B_{i-1}^{-1} A_{i-1} & -A_{i-1}^T B_{i-1}^{-1} \\ -B_{i-1}^{-1} A_{i-1} & B_{i-1}^{-1} \end{bmatrix}
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}
$$

$$
=
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}^T
\begin{bmatrix} -A_{i-1}^T \\ I \end{bmatrix}
B_{i-1}
\begin{bmatrix} -A_{i-1} & I \end{bmatrix}
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}
$$

$$
= (x_i - A_{i-1}x_{i-1})^T B_{i-1}^{-1} (x_i - A_{i-1}x_{i-1}) \tag{A14}
$$

or

$$
I_{i-1,i} =
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}
\begin{bmatrix} C_{i-1} & D_{i-1} \\ D_{i-1}^T & E_{i-1} \end{bmatrix}
\begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix} \tag{A15}
$$

where

$$
C_{i-1} = \rho A_{i-1}^T B_{i-1}^{-1} A_{i-1} \tag{A16}
$$

$$
D_{i-1} = -\rho A_{i-1}^T B_{i-1}^{-1} \tag{A17}
$$

$$
E_{i-1} = \rho B_{i-1}^{-1} \tag{A18}
$$

Thus, the smoothing integral is transformed into the above quadratic form.

VIEWS ON THE WEIERSTRASS
AND GENERALIZED WEIERSTRASS FUNCTIONS

M.F. Shlesinger*, M.A. Hussain, and J.T. Bendler
Corporate Research and Development
General Electric Company
Schenectady, New York

1. INTRODUCTION

Weierstrass provided the first example of a function which is continuous but nowhere differentiable.[1] Hardy[2] later established that under certain conditions this function even fails to have a well-defined infinite derivative, i.e., a vertical tangent. Weierstrass never published his result but only read it to the Berlin Academy on July 18, 1872. It was first published in 1875 by Paul DuBois Reymond[3] who wrote that Weierstrass' result was "equally too strange for immediate perception as well as for actual understanding" and "will lead to the limit of our intellect."

Singh[4] has written a treatise on nondifferentiable functions and provides some examples of infinite series as well as geometric constructions such as Koch curves which are nondifferentiable. Rather than being hailed as a great mathematical discovery, the Weierstrass function was presented more as a pathological curiosity and used as counterexample to warn freewheeling users of mathematics that care must sometimes be exercised when attempting to manipulate mathematical functions. It took the genius of Mandelbrot[1] to bestow honor on a large collection of "path is logical" mathematical objects, such as the Weierstrass function. This function is nondifferentiable because it has wiggles on all scales. This absence of a characteristic scale is the paradigm of Mandelbrot's fractals.

We review in the next section how the Weierstrass function appears naturally in a 1-D random walk context.[5] In Section 3 we review, through the random walk framework, a manner in which to generalize the Weierstrass function.[6] Although we do not prove that the novel mathematical functions observed are nondifferentiable, we do provide suggestive numerical evidence. The proof or disproof of being everywhere nondifferentiable is an open question left to those with a stronger mathematics background than the authors.

A cautionary example is the function of Rieman, $\sum_{k=1}^{\infty} \frac{\sin{(k^2 x)}}{k^2}$, which was assumed to be nowhere differentiable until Gerver in 1970 proved that it possessed a derivative equal to $-\frac{1}{2}$ at values of x which were in lowest order ratios of odd numbers.

2. THE WEIERSTRASS FUNCTION AND RANDOM WALK

Consider a random walker on an infinite 1-D lattice of unit spacing whose initial position is the origin. The probability of being at site l after n steps is denoted by $P_n(l)$. From the Chapman-Kolmogorov equation

$$P_{n+1}(l) = \sum_{l'} P_n(l - l')p(l') \tag{1}$$

where $p(l)$ is the probability of making a jump of length l. Equation 1 is in a convolution form and is easily handled in Fourier space. Define

* Office of Naval Research - Code 412, 800 North Quincy St., Arlington, VA.

$$\hat{P}_n(k) = \sum_l e^{ikl} P_n(l) \tag{2a}$$

$$\hat{p}(k) = \sum_l e^{ikl} p(l) \tag{2b}$$

and

$$\langle l^n \rangle = \sum_l l^n p(l) \tag{2c}$$

Then from Equation 1

$$\hat{P}_n(k) = [p(k)]^n$$

or inverse Fourier transforming

$$P_n(l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikl} [p(k)]^n dk \tag{3}$$

For an unbiased walk, $\langle l \rangle = 0$. If $\langle l^2 \rangle$ is finite, then for $k \to 0$

$$\hat{p}(k) \sim 1 - \tfrac{1}{2} k^2 \langle l^2 \rangle + o(k^2) \tag{4}$$
$$\sim e^{-\frac{1}{2} k^2 \langle l^2 \rangle}$$

and

$$P_n(l) \sim \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikl} e^{-\frac{1}{2} n k^2 \langle l^2 \rangle} \tag{5}$$

$$= \frac{1}{\sqrt{2\pi n \langle l^2 \rangle}} e^{-\frac{l^2}{2n \langle l^2 \rangle}} \tag{6}$$

which is the standard Gaussian behavior insured by the Central Limit Theorem.

However, if $\langle l^2 \rangle$ is infinite, then the different behavior will result as first described in general by P. Levy in the 1920's. Essentially, if $p(l) \sim l^{-1-\beta}$ with $0 < \beta < 2$, then $\langle l^\beta \rangle$ is the lowest moment which diverges and

$$\hat{p}(k) \sim 1 - const. \, |k|^\beta \tag{7}$$

and

$$\hat{P}_n(k) \sim e^{-n|k|^\beta} \tag{8}$$

Equation 8 is called the Fourier transform of a Levy stable flow. They have a somewhat more general form but this need not concern us here. The Levy stable laws are best characterized in Fourier space, Equation 8, and cannot in general be calculated in a useful closed form analytic solution.

The analysis can easily be given for a D dimensional cubic lattice, but instead we now connect it to the Weierstrass function by choosing a particular form for $p(l)$.

Let

$$p(l) = \frac{a-1}{2a} \sum_{n=0}^{\infty} \left[\delta_{l+b^n} + \delta_{l-b^n} \right] a^{-n} , \quad (a, b > 1) \tag{9}$$

This allows jumps of all orders of magnitude on base b, with each order of magnitude longer jump occurring with an order of magnitude smaller probability in base a. Then

$$\hat{p}(k) = \frac{a-1}{a} \sum_{n=0}^{\infty} a^{-n} \cos(b^n k) \tag{10}$$

which is the Weierstrass function. It satisfies the scaling equation

$$\hat{p}(k) = a^{-1} \hat{p}(bk) + \frac{a-1}{a} \cos k \tag{11}$$

Any singular behavior of $\hat{p}(k)$ must satisfy

$$\hat{p}_{sing}(k) = a^{-1} \hat{p}_{sing}(bk) \tag{12}$$

which has the solution

$$\hat{p}_{sing}(k) = k^{\beta} Q(k) \tag{13}$$

where

$$\beta = \ln a / \ln b \qquad 0 < \beta < 2 \tag{14}$$

and $Q(k)$ is a function periodic on $\ln k$ with period $\ln b$. It can be shown that[5] the full solution to Equation 11 is

$$\hat{p}(k) = k^{\beta} Q(k) + \frac{a-1}{a} \sum_{n} \frac{(-1)^n k^{2n}}{(2n)! \, (1 - a^{-1} b^{2n})}$$

with

$$Q(k) = \frac{a-1}{a \ln b} \sum_{n=-\infty}^{\infty} \Gamma\left[-\beta + \frac{2n\pi i}{\ln b}\right] \cos\left[\frac{\pi}{2}\left[-\beta + \frac{2n\pi i}{\ln b}\right]\right] \times \exp\left[-\frac{2n\pi \ln k}{\ln b}\right] . \tag{15}$$

It was the Weierstrass function of Equations 11 and 15 that Hardy established had a finite derivative at no value of k when $b \geqslant a$.

In this regime we see numerically how the wiggles grow upon wiggles until the function becomes nondifferentiable. See Figures 1 and 2.

3. THE SPHERICALLY SYMMETRIC CONTINUUM WEIERSTRASS RANDOM WALK

In D dimensions let the random walker take spherically symmetric jumps of variable length

$$p(\vec{x}) = \frac{p_1(|\vec{x}|)}{S_D |\vec{x}|^{D-1}} \qquad 0 < x < \infty \tag{16}$$

where $S_D = 2\pi^{D/2} / \Gamma(D/S)$ is the surface area of a unit hypersphere.

Fourier transforming the radial function $p(\vec{x})$, one obtains

$$\hat{p}(\vec{k}) = \Gamma\left[\frac{D}{2}\right] \int_0^{\infty} \left(\frac{1}{2}|k|\lambda\right)^{1 - \frac{D}{2}} J_{\frac{D}{2}-1}(|k|\lambda) p_1(\lambda) \, d\lambda \tag{17}$$

In analogy to the Weierstrass walk we choose

$$p_1(|\vec{x}|) = \frac{a-1}{a} \sum_{n=0}^{\infty} a^{-n} \delta(|\vec{x}| - b^n) , \quad b > 1 \tag{18}$$

which when substituted into Equation 17 yields

$$\hat{p}(\vec{k}) = \frac{a-1}{a} \sum_{n=1}^{\infty} a^{-n} \Gamma\left[\frac{D}{2}\right] \left(\frac{1}{2}|\vec{k}|b^n\right)^{1 - \frac{D}{2}} J_{\frac{D}{2}-1}(|k|b^n) \tag{19}$$

which satisfies the scaling equation

1279

$$\hat{p}(\overline{k}) = a^{-1}\hat{p}(b\overline{k}) + \frac{a-1}{a}\Gamma\left(\frac{D}{2}\right)\left(\frac{1}{2}|\overline{k}|\right)^{1-\frac{D}{2}} J_{\frac{D}{2}-1}(|\overline{k}|)$$

which can be shown to have the solution

$$\hat{p}(\overline{k}) = \frac{a-1}{2\,a\,\ln b}(\frac{1}{2}k)^{\beta}\Gamma\left(\frac{D}{2}\right)\sum_{m=-\infty}^{m=\infty}\frac{\Gamma(-\frac{1}{2}\beta + m\pi i/\ln b)}{\Gamma(\frac{1}{2}D + \frac{1}{2}\beta - m\pi i/\ln b)}\exp\left(-\frac{2m\pi\,i\ln\left(\frac{k}{2}\right)}{\ln b}\right)$$

$$+ \frac{a-1}{a}\sum_{n=0}^{\infty}\frac{\Gamma\left(\frac{D}{2}\right)(-1)^n(\frac{1}{2}k)^{2n}}{n!\,\Gamma(n+\frac{1}{2}D)(1-a^{-1}b^{2n})}, \quad (0 < \beta < 2) \tag{20}$$

It is Equation 19 which is hypothesized to be a proper generalization of the Weierstrass function. Let us now consider its differentiability. Since $\hat{p}(\overline{k})^- 1 - O(k^{\beta})$, differentiability holds at $k = 0$ if $\beta > 1$. It can be shown that if $\beta > \frac{1}{2}(3-D)$, then $p(\overline{k})$ is differentiable with respect to k for *all* $k > 0$. This corresponds to $\beta > 1$ in one dimension, $\beta > \frac{1}{2}$ in two dimensions, and $\beta > 0$ in three dimensions. As the one-dimensional case reduces to the Weierstrass function, we focus on the two-dimensional case

$$\hat{p}(k) = \frac{a-1}{a}\sum_{n=0}^{\infty}a^{-n}J_0(kb^n) \tag{21}$$

For the above equation we again see numerically that for $b > a$ wiggles grow upon wiggles until the function becomes nondifferentiable. See Figures 3, 4, and 5.

REFERENCES
1. B.B. Mandelbrot, *The Fractal Geometry of Nature*, (Freeman, San Francisco) 1982, p. 420.
2. G.H. Hardy, *Trans. Amer. Math. Soc. 17*, 301 (1916).
3. P. DuBois Reymond, *J. für die reine und angewandte Math. (Crelle) 79*, 21 (1875).
4. A.N. Singh, *The Theory and Construction of Nondifferentiable Functions*, Lucknow University Press (1935); reprinted in E.W. Hobson et al., *Squaring the Circle and Other Monographs* (Chelsea, NY, 1953).
5. B.D. Hughes, M.F. Shlesinger, and E.W. Montroll, *Proc. Nat. Acad. Sci. (USA) 78*, 3287 (1981).
6. B.D. Hughes, E.W. Montroll, and M.F. Shlesinger, *J. Statistical Phys. 28*, 111 (1982).
7. J. Gerver *Am. J. Math. 92*, 33 (1970). See also A. Smith, *Proc. Amer. Math Soc. 34*, 463 (1972).
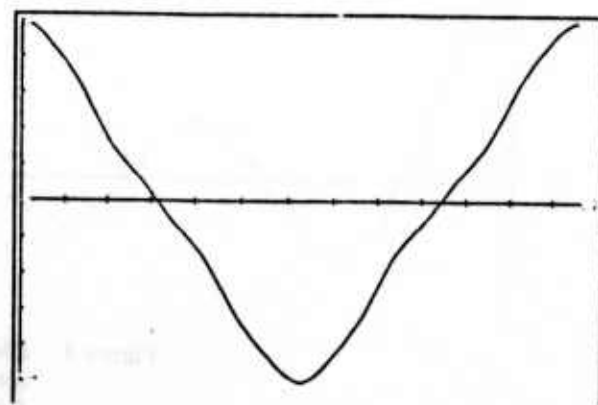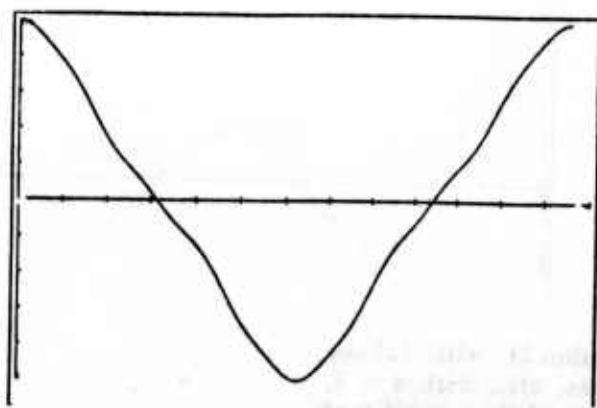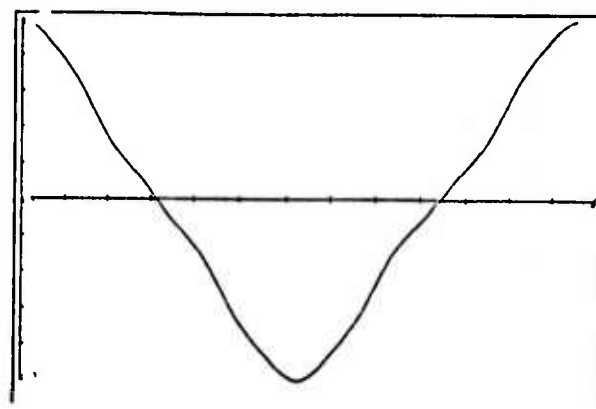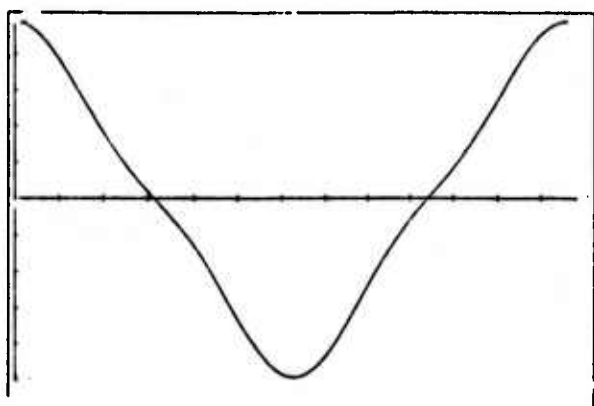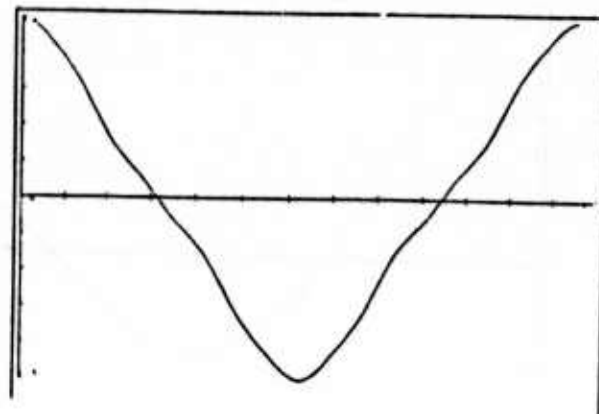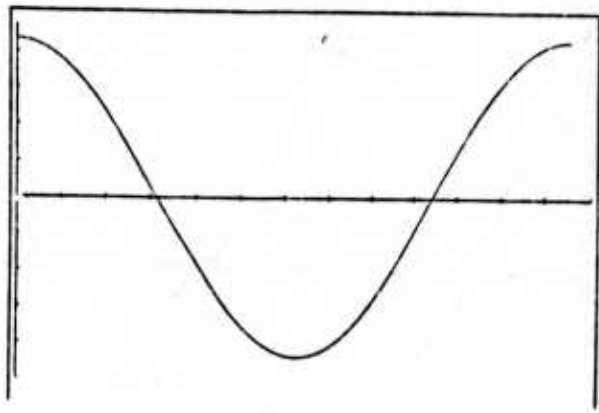
Figure 2. Plots from Equation 10 with (a) one term, (b) two terms, etc., with $a = 8$ and $b = 3$ showing the lack of growth of wiggles.
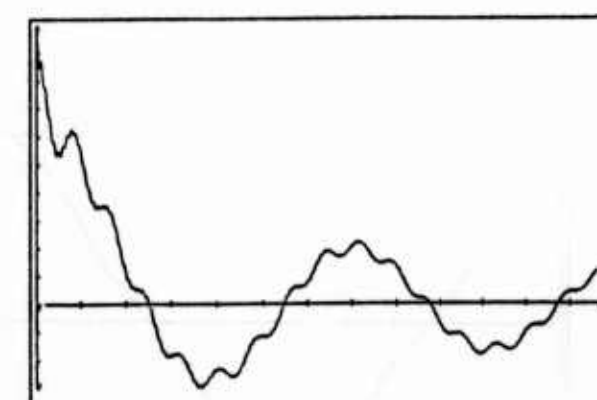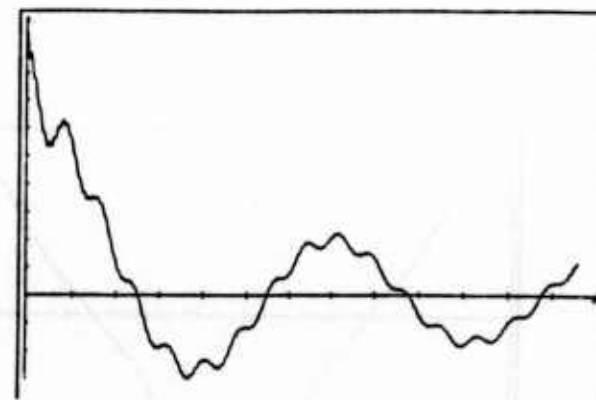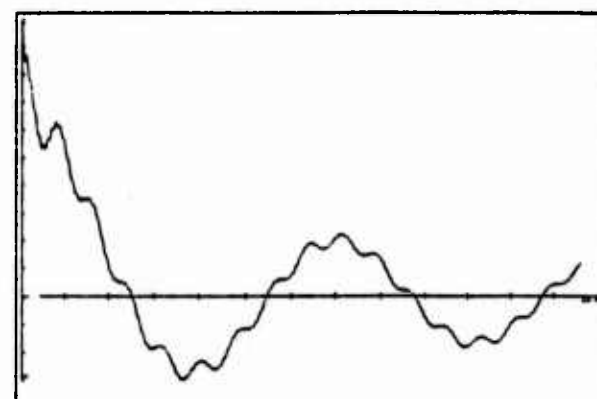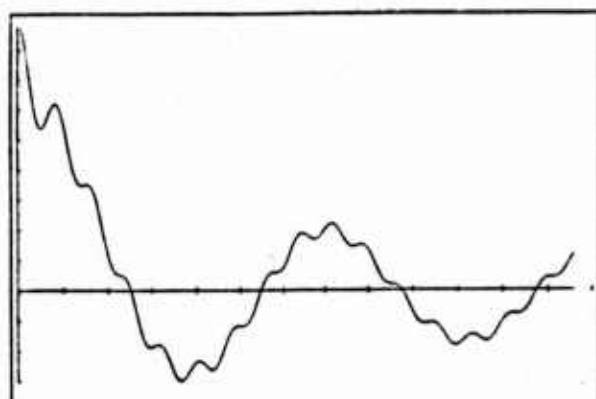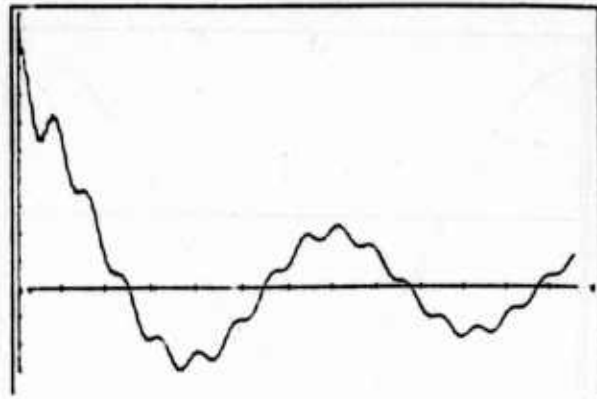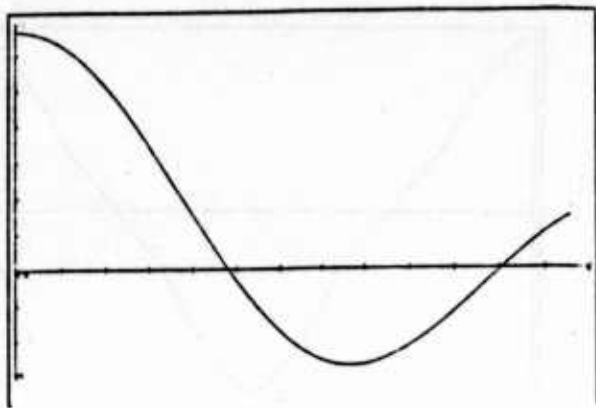
Figure 3. Plots form Equation 21 with (a) one term, (b) two terms, etc., with $a = 3$, $b = 8$ showing the nondifferentiability of the three-dimensional random walk case.
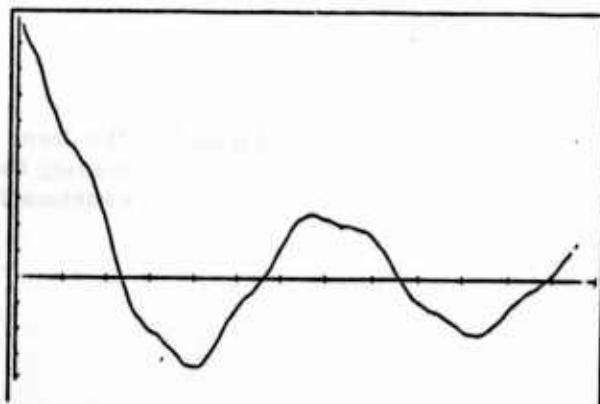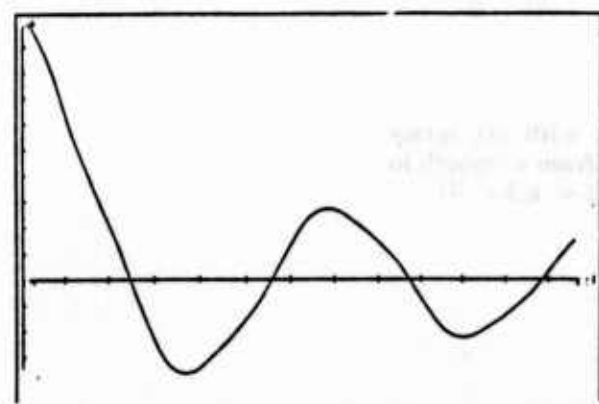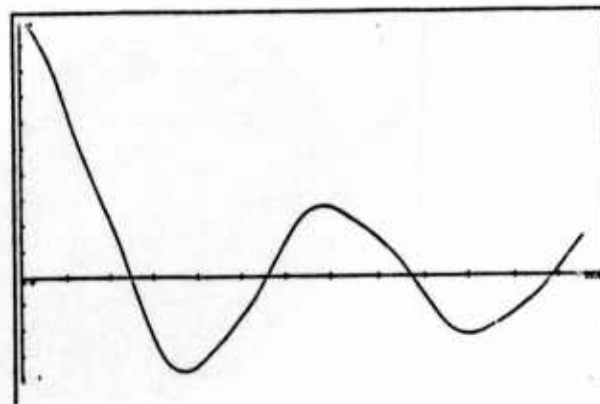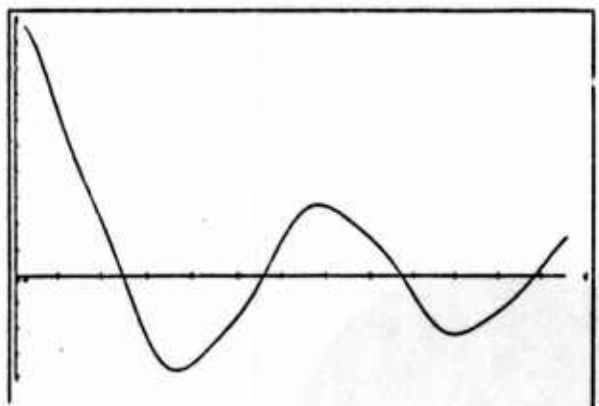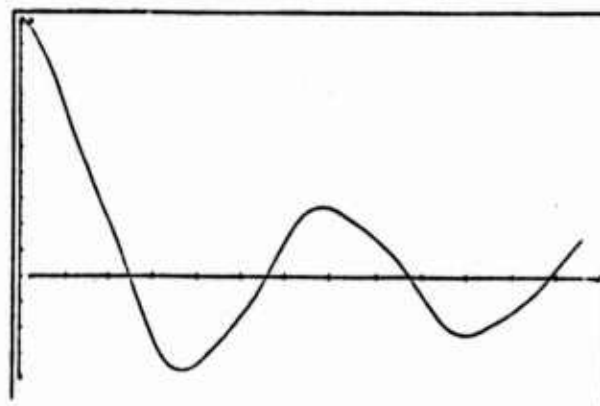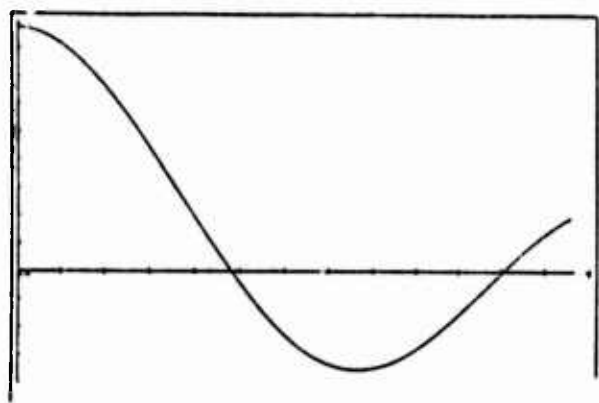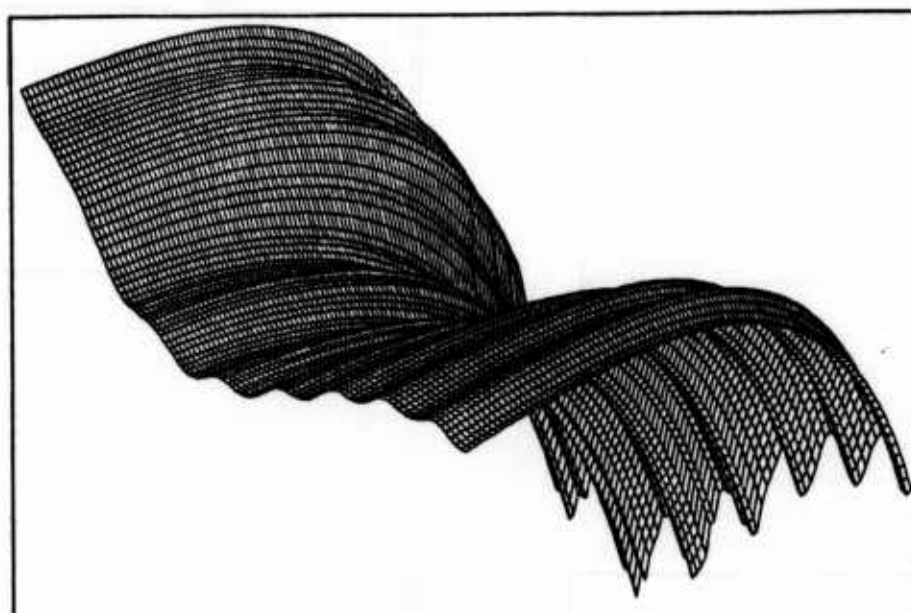
1282

Figure 4. Plots from Equation 21 with (a) one term, (b) two terms, etc., for $a = 8$ and $b = 3$ showing the lack of growth of wiggles.

1283

Figure 5. Plot from Equation 21 with six terms showing the transition from a smooth to a nonsmooth function $(2 < a, b < 7)$.

A FAST ALGORITHM FOR THE MULTIPLICATION OF GENERALIZED HILBERT MATRICES WITH VECTORS†

Apostolos Gerasoulis

Department of Computer Science

Rutgers University

New Brunswick, NJ 08903

Abstract. *We describe an algorithm with an $O_A(n(\log n)^2)$ time complexity for the multiplication of generalized Hilbert matrices with vectors. These matrices are defined by $(B_p)_{i,j} = 1/(t_i - s_j)^p$, $i, j = 1, ..., n$ and $p = 1, 2$, where t_i and s_i are distinct elements and $t_i \neq s_i$, $i = 1, ..., n$. An implementation of the algorithm for the Chebyshev points, which arise in the numerical approximation of Cauchy singular integral equations, is presented. The time complexity of the algorithm, for this special set of points, reduces to $O_A(n \log n)$.*

1. Introduction:

Let us define the matrices $\mathbf{B}_p$, $p = 1, 2$, by

$$(B_p)_{i,j} = \frac{1}{(t_i - s_j)^p}, \qquad i, j = 1, ..., n \tag{1}$$

where t_i and s_i, are distinct elements and $t_i \neq s_i$, $i = 1, ..., n$. The special case, $p = 1$, $t_i = i$ and $s_j = -j+1$, is the well known Hilbert matrix

$$(B_1)_{i,j} = \frac{1}{i+j-1}, \quad i, j = 1, ..., n. \tag{2}$$

In [17], the following question was considered:

"Given a vector $\mathbf{y}$. Does there exist an algorithm for computing the product $\mathbf{Ty}$ in less than $O_A(n^2)$ operations?"

where the matrix $\mathbf{T}$ is given by

$$T_{i,j} = \begin{cases} \dfrac{1}{c_i - c_j} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}, \quad i, j = 1, ... n \tag{3}$$

with distinct c_i, and where the time or space complexity $O_A(f(n))$ is defined in (Aho et. al. [1], pp. 19-22). This problem was initially posed by Golub in [18] and [19] and it is known as Trummer's problem. It has generated great interest because of various applications in the computation of conformal mappings (Trummer et. al. [27], [32]), the zeta function (Odlyzko and Schönhage [30]), and the numerical evaluation of singular integrals (this paper).

In [17], we have proposed an $O_A(n(\log n)^2)$ algorithm for Trummer's problem, henceforth the

GGS algorithm. The GGS algorithm uses Fast Fourier Transform (FFT) polynomial multiplication, polynomial interpolation and polynomial evaluation at n distinct points.

In Section 2, we show that the GGS algorithm can be extended to include the matrices defined in (1). The time complexity of the extended algorithm is the same as the GGS. In Section 3, we present examples of generalized Hilbert matrices similar to (1)-(3), which arise in the numerical approximation of singular integrals. In Section 4, we implement the extended algorithm for the points $s_j = \cos(j\pi/n)$, $j = 1, ..., n-1$ and $t_i = \cos((2i-1)\pi/2n)$, $i = 1, ..., n$, which arise in the numerical approximation of Cauchy singular integral equations. The time complexity of the algorithm, for this special set of points, reduces to $O_A(n \log n)$. Finally, in the Appendix, we present a collection of problems for which the new fast algorithm could be used to speed up computations.

2. An extended GGS algorithm

In this section, we briefly describe an extension to the GGS algorithm for the multiplication of $\mathbf{B}_p$ with a vector.

We first notice that the problem of multiplying $\mathbf{B}_p \mathbf{y}$ is equivalent to evaluating the function $f_p(x)$ at the points t_i, $i = 1, ..., n$, where

$$f_p(x) = \sum_{j=1}^{n} \frac{y_j}{(x - s_j)^p}, \quad p = 1, 2.$$ (4)

Since $f_2(x) = -f_1'(x)$, we only need to consider $f_1(x)$.

We follow Gastinel [13] and express $f_1(x)$ as the ratio of two polynomials $h(x)$ and $g(x)$, where $g(x)$ is an n-th degree polynomial defined by

$$g(x) = \prod_{j=1}^{n} (x - s_j)$$ (5)

and $h(x)$ is a polynomial. of degree at the most $n-1$, determined from

$$f_1(x) = \frac{h(x)}{g(x)} = \sum_{j=1}^{n} \frac{y_j}{x - s_j}.$$ (6)

By setting $x = s_i$, $i = 1, ..., n$, in (6), we derive

$$h(s_i) = y_i \, g'(s_i), \quad i = 1, ..., n$$ (7)

which implies that $h(x)$ is the interpolation polynomial at the points $(s_i, h(s_i))$, $i = 1, ..., n$.

It is clear now that the matrix multiplication problem, $\mathbf{B}_p \mathbf{y}$, is equivalent to evaluating the functions

$$f_1(t_i) = \frac{h(t_i)}{g(t_i)}, \quad f_2(t_i) = \frac{h'(t_i)}{g(t_i)} - \frac{h(t_i) \, g'(t_i)}{g^2(t_i)}, \quad i = 1, ..., n$$ (8)

while Trummer's problem, $\mathbf{T}\mathbf{y}$, is equivalent to evaluating (Gerasoulis et. al. [17])

1286

$$z_j = \frac{h'(c_j) - \frac{1}{2} y_j\, g''(c_j)}{g'(c_j)}, \qquad j = 1, ..., n. \tag{9}$$

It should be mentioned here that for $p \geq 3$, the multiplication of $\mathbf{B}_p y$ is equivalent to relations similar to (8). These relations may be derived from the identity $f_p(z) = f'_{p-1}(z)/(1-p)$, $p = 3, 4, ...$.

We now describe an efficient algorithm for the evaluation of $f_1(t_i)$ and $f_2(t_i)$, $i = 1, ..., n$.

Procedure $FAST$(n, t, s, y); return $\mathbf{f_1}$, $\mathbf{f_2}$;

1. Compute the coefficients of $g(z)$ in its power form, by using FFT polynomial multiplication, in $O_A(n\,(\log n)^2)$ time (e.g. Horowitz [22], Aho et. al. [1], Theorem 8.14, p. 299);

2. Compute the coefficients of $g'(z)$ in $O_A(n)$ time;

3. Evaluate $g(t_i)$, $g'(t_i)$, $i = 1, ..., n$, and $g'(s_j)$, $j = 1, ..., n$, in $O_A(n\,(\log n)^2)$ time (Aho et. al. [1], Corollary 2, p. 294);

4. Compute $h(s_j) = y_j g'(t_j)$, $j = 1, ..., n$, in $O_A(n)$ time;

5. Find the interpolation polynomial $h(z)$ at the points $(s_j, h(s_j))$, $j = 1, ..., n$, in $O_A(n\,(\log n)^2)$ time (Aho et. al. [1], Theorem 8.14, p. 299);

6. Compute the coefficients of $h'(z)$ in $O_A(n)$ time, and evaluate $h(t_i)$ and $h'(t_i)$, $i = 1, ..., n$, in $O_A(n\,(\log n)^2)$ time, by following the same technique as in steps 2 and 3;

7. Compute $f_1(t_i) = h(t_i)/g(t_i)$ and $f_2(t_i) = h'(t_i)/g(t_i) - h(t_i) g'(t_i))/g^2(t_i)$, $i = 1, ..., n$, in $O_A(n)$ time;

end FAST;

The space and time complexity of $FAST$ are $O_A(n \log n)$ and $O_A(n\,(\log n)^2)$, respectively. In Section 4, we consider two important special cases for which the time complexity of $FAST$ reduces to $O_A(n \log n)$.

3. Generalized Hilbert matrices

In this section, we present generalized Hilbert matrices which arise in the quadrature approximation of Cauchy singular integrals. Additional problems for which $FAST$ is applicable are given in the Appendix.

We consider the Cauchy principal value integral

$$I(y; s) = \int_{-1}^{1} w(t) \frac{y(t)}{t-s} dt, \qquad |s| < 1 \tag{10}$$

where $w(t)$ is a weight function defined by

$$I_H(y\,;\,s) = \frac{d\,I(y\,;\,s)}{ds} = \fint_{-1}^{1} w(t)\frac{y(t)}{(t-s)^2}dt, \qquad |s| < 1 \tag{12}$$

where it is assumed that $w(t)$ and $y(t)$ are such that the derivative of $I(y\,;\,s)$ exists. The singular integrals defined by (10) and (12) arise in fields such as aerodynamics, wave-guide theory, scattering, fracture mechanics and others (see Appendix). For example, in fracture mechanics the solution of the equation

$$\fint_{-1}^{1} (1-t^2)^{-1/2}\frac{y(t)}{t-s}dt = f(s), \qquad |s| < 1 \tag{13}$$

represents the derivative of the crack opening under a given pressure distribution $f(s)$ along $(-1, 1)$.

We will now derive quadratures for the singular integrals (10) and (12). These quadratures give rise to matrices similar to $\mathbf{B}_p$ and $\mathbf{T}$. We only need to consider the quadrature approximation of (10). Quadratures for $I_H(y\,;\,s)$ and matrices similar to $\mathbf{B}_2$ can be obtained from the quadratures for $I(y\,;\,s)$ via (12).

By rewriting (10) as

$$I(y\,;\,s) = \int_{-1}^{1} w(t)\frac{y(t)-y(s)}{t-s}dt \; + \; y(s)\int_{-1}^{1} \frac{w(t)}{t-s}dt \tag{14}$$

we see that classical quadratures may be used to approximate $I(y\,;\,s)$, provided that the second integral in (14) can be computed to within any given tolerance. This computation may be performed once via a very fast convergent series of the Hypergeometric function. Then, the first integral can be computed via a quadrature for different functions $y(t)$. These computations may be performed by using procedure *FAST*.

For simplicity, we consider below only two cases of the weight function $w(t)$. We will use the trapezoidal and Gauss-Chebyshev quadratures for the approximation of $I(y\,;\,s)$. The analysis can easily be extended to the general weight function $w(t)$ in (11).

The case $\alpha = \beta = 0$:

Here, the weight function $w(t) = 1$. By using the trapezoidal rule for the approximation of (14), we derive

$$I_n(y\,;\,s) = \sum_{i=0}^{n} w_i \frac{y(t_i)-y(s)}{t_i-s} + y(s)\log\frac{|1-s|}{|1+s|}, \qquad |s| < 1 \tag{15}$$

where $t_i = -1 + ih$, $i = 0, 1, ..., n-1$, $h = 2/n$, $w_0 = w_n = h/2$ and $w_i = h$, $i = 1, ..., n-1$. Trummer's matrix $\mathbf{T}$, can be derived by setting $s = t_j$, $j = 1, ..., n-1$ in (15)

$$I_n(y; t_j) = \sum_{\substack{i=0 \\ i \neq j}}^{n} \frac{w_i \, y(t_i)}{t_i - t_j} + y(t_j) \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{w_i}{t_i - t_j} + w_j y'(t_j) + y(t_j) \log \frac{|1 - t_j|}{|1 + t_j|}, \quad j = 1, \ldots, n-1 \qquad (16)$$

where we have assumed that the $y'(t_j)$ exists.

The case $\alpha = \beta = 1/2$:

For this special case the weight function becomes $w(t) = (1 - t^2)^{-1/2}$. By using the Gauss-Chebyshev quadrature and the identities (Erdogan et. al. [12])

$$n^{-1} \sum_{i=1}^{n} \frac{1}{t_i - s} = -\frac{U_{n-1}(s)}{T_n(s)}, \qquad \int_{-1}^{1} \frac{(1-t^2)^{-1/2}}{t - s} dt = 0, \qquad |s| < 1 \qquad (17)$$

we see that $I(y; s)$ is approximated by

$$I_n(y; s) = n^{-1} \sum_{i=1}^{n} \frac{y(t_i) - y(s)}{t_i - s} = n^{-1} \sum_{i=1}^{n} \frac{y(t_i)}{t_i - s} + \frac{U_{n-1}(s)}{T_n(s)} y(s) \qquad (18)$$

where $t_i = \cos((2i-1)\pi/2n)$, $i = 1, \ldots, n$, are the zeros of $T_n(t)$ and $s_j = \cos(j\pi/n)$, $j = 1, \ldots, n-1$, are the zeros of $U_{n-1}(s)$, and where $T_n(t)$ and $U_{n-1}(s)$ are the Chebyshev polynomials of the first and second kind, respectively.

Now, by setting $s = s_j$, $j = 1, \ldots, n-1$, in (18) we obtain a matrix similar to $\mathbf{B}_1$

$$I_n(y; s_j) = n^{-1} \sum_{i=1}^{n} \frac{y(t_i)}{t_i - s_j}, \qquad j = 1, \ldots, n-1 \qquad (19)$$

while by setting $s = t_i$, $i = 1, \ldots, n$, we obtain $\mathbf{T}$ and summations similar to (16).

Next, we use (19) and *FAST* to obtain a numerical solution for equation (13).

4. An application of FAST

In this section, we present an $O_A(n \log n^{\frac{1}{2}})$ implementation of *FAST* for the points $t_i = \cos((2i-1)\pi/2n)$, $i = 1, \ldots, n$, and $s_j = \cos(j\pi/n)$, $j = 1, \ldots, n-1$, which arise in (18). For the points $t_i = -1 + ih$, $i = 1, \ldots, n-1$, we do not need to use *FAST*, because $t_i - t_j = (i - j)h$, and therefore the sums in (16) can be computed directly via FFT convolutions in $O_A(n \log n)$ time (Brigham [6], Henrici [21]).

We now consider the numerical solution of (13). By using (19) to approximate (13), we obtain the $(n-1) \times n$ algebraic system

$$\mathbf{Ay} = \mathbf{f}, \qquad A_{j,i} = \frac{1}{n(t_i - s_j)}, \qquad j = 1, \ldots, n-1, \quad i = 1, \ldots, n \qquad (20)$$

where $\mathbf{y} = [y(t_1), y(t_2), \ldots, y(t_n)]^T$ and $\mathbf{f} = [f(s_1), f(s_2), \ldots, f(s_{n-1})]^T$. Since $\mathbf{A}$ possesses a right inverse $\mathbf{A}^I$ (Gerasoulis [14]), the solution of (20) can be obtained from

$$y = \mathbf{A}^I \mathbf{f} + b_n \mathbf{u}_n, \qquad (A^I)_{i,j} = \frac{1 - s_j^2}{n(t_i - s_j)}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n-1 \qquad (21)$$

where b_n is an arbitrary constant determined by an additional condition imposed on the solution $y(t)$ and $\mathbf{u}_n$ is a vector with all elements equal to one. We are now ready to apply *FAST* for the computation of

$$\mathbf{A}^I \mathbf{f} = n^{-1} \sum_{j=1}^{n-1} \frac{(1 - s_j^2) f(s_j)}{t_i - s_j}, \quad i = 1, \ldots, n. \qquad (22)$$

The algorithm presented below is a modification of *FAST* for the INPUT:
$t_i = \cos((2i-1)\pi/2n)$, $i = 1, \ldots, n$, $s_j = \cos(j\pi/n)$ and $f(s_j)$, $j = 1, \ldots, n-1$.

1. Instead of computing $g(x)$ in its power form, we will use its product form directly. We have $g(x) = \prod_{j=1}^{n-1}(x - s_j) = c_n U_{n-1}(x) = c_n \sin(n\theta)/\sin(\theta)$, where $c_n = 2^{-(n-1)}$ and $\cos(\theta) = x$.

2. Similarly, $g'(x) = -c_n [n T_n(x) - x U_{n-1}(x)]/(1-x^2)$.

3. Since $g'(s_j) = c_n (-1)^{j+1} n/(1-s_j^2)$, the computational complexity for this step reduces to $O_A(n)$ time.

4. Equation (7) and step 3 above imply that $h(s_j) = c_n (-1)^{j+1} y_j$, $j = 1, \ldots, n-1$.

5. We find $h(x)$ by using orthogonal polynomial interpolation. We set

$$h(x) = \sum_{k=1}^{n-1} a_k U_{k-1}(x) \qquad (23)$$

and use the orthogonality identities, which hold for all integers l, m such that $l+m \leq 2n-1$,

$$\int_{-1}^{1} (1-x^2)^{1/2} U_l(x) U_m(x)\, dx = \pi n^{-1} \sum_{k=1}^{n-1} (1 - s_k^2) U_l(s_k) U_m(s_k) = \begin{cases} \dfrac{\pi}{2} & \text{if } l = m \\ 0 & \text{if } l \neq m \end{cases} \qquad (24)$$

to obtain

$$a_k = 2 n^{-1} \sum_{j=1}^{n-1} (1 - s_j^2) h(s_j) U_{k-1}(s_j) = \sum_{j=1}^{n-1} 2 z_j \sin\left(k \frac{j\pi}{n}\right), \qquad z_j = n^{-1} h(s_j) \sin\left(\frac{j\pi}{n}\right) \qquad (25)$$

for $k = 1, \ldots, n-1$. The coefficients a_k, $k = 1, \ldots, n-1$, can be computed via FFT in $O_A(n \log n)$ time (Aho et. al. [1]).

6. From step 1 and (23), we obtain

$$g(t_i) = \frac{c_n (-1)^{i+1}}{\sin((2i-1)\pi/2n)}, \qquad h(t_i) = \sum_{k=1}^{n-1} a_k \sin\left(k \frac{(2i-1)\pi}{2n}\right) / \sin\left(\frac{(2i-1)\pi}{2n}\right) \qquad (26)$$

for $i = 1, \ldots, n$.

1290

7. Finally, $\mathbf{A}^I \mathbf{f}$ is computed from

$$f_1(t_i) = \frac{h(t_i)}{g(t_i)} = \sum_{k=1}^{n-1} 2 z_k \sin\left(k \frac{(2i-1)\pi}{2n}\right), \quad z_k = \frac{(-1)^{i+1} a_k}{2} \tag{27}$$

for $i = 1, ..., n$, by using FFT in $O_A(n \log n)$ time.

The total cost of the above implementation of *FAST* is $O_A(n \log n)$, since it only requires the application of FFT twice. In Table 4.1, we present our computational experience with *FAST* for the function $f(s) = 1$. For this case the summations can be obtained exactly from the identity (e.g. Erdogan et. al. [12])

$$n^{-1} \sum_{j=1}^{n-1} \frac{1-s_j^2}{t_i-s_j} = t_i, \quad i = 1, ..., n. \tag{28}$$

The computations have been performed by using the subroutines *SINT* and *SINQF* from *FFTPACK* of *NETLIB* (Swarztrauber [31]), which are most efficient whenever $n = 2^k$. As expected, *FAST* outperforms the $O_A(n^2)$ algorithm for all $n \geq 32$, (Brigham [6], p. 152). The Table also shows that in addition to its performance, *FAST* attains better accuracy than the $O_A(n^2)$ algorithm. Similar results have been obtained for several other choices of $f(s)$.

Note. All computations were performed on a DECSYSTEM/2060T using FORTRAN 77 with a single precision floating point arithmetic (with a mantissa of approximately 8 decimals and with an exponent in the range 0.14×10^{-38} to 1.7×10^{38}).

5. Concluding remarks

In Section 4, we have described an efficient and *stable* implementation of *FAST* for a special set of points which arise in numerous applications (see Appendix). The implementation and stability properties of *FAST* for an arbitrary set of points still remain to be addressed.

The computational advantages of *FAST*, over direct multiplications algorithms, become apparent if n is very large (see Table 4.1) or if the product $\mathbf{B}_p \mathbf{y}$ is repeatedly computed. In the case of repeated computation of $\mathbf{B}_p \mathbf{y}$ the advantages, in terms of actual CPU execution time, are even more pronounced. For example, in the quadrature approximation of two dimensional singular integrals, using the Chebyshev points, the complexity via *FAST* is $O_A(n^2 \log n)$ as opposed to $O_A(n^3)$ via the direct algorithm. Thus, the times given in Table 4.1 will have to be multiplied by n.

6. Acknowledgments

1291

		$O_A(n \log n)$ Algorithm			$O_A(n^2)$ Algorithm	
		Time in	Max. Error		Time in	Max. Error
	$n = 2^k$	DEC-20 sec.			DEC-20 sec.	
k						
2	4	0.0013	0.372×10^{-7}		0.0004	0.223×10^{-7}
3	8	0.0031	0.447×10^{-7}		0.0014	0.968×10^{-7}
4	16	0.0054	0.447×10^{-7}		0.0055	0.194×10^{-6}
5	32	0.0122	0.671×10^{-7}		0.0222	0.484×10^{-6}
6	64	0.0248	0.596×10^{-7}		0.0889	0.789×10^{-6}
7	128	0.0524	0.104×10^{-6}		0.3563	0.200×10^{-5}
8	256	0.1082	0.372×10^{-6}		1.4264	0.400×10^{-5}
9	512	0.2335	0.738×10^{-6}		5.7057	0.814×10^{-5}
10	1024	0.4831	0.114×10^{-5}		22.8265	0.167×10^{-4}
11	2048	0.9975	0.627×10^{-5}		91.3278	0.336×10^{-4}
12	4096	2.0592	0.743×10^{-5}		365.2306	0.713×10^{-4}

Table 4-1: The performance of *FAST* for $f(s) = 1$

7. References:

1. A. Aho, J.E. Hopcroft and J.D. Ullman, *The design and analysis of computer algorithms*, Addison-Wesley, 1974.

2. C. Atkinson and Leppington, "The asymptotic solution of some integral equations", *IMA J. of Applied Mathematics*, 31 (1983).

3. D. Atkinson, "Study of singular integral equations of Calogero", *Il Nuovo Cimento*, 60 B (1980), N.2, 143-155.

4. J. L. Bassani and F. Erdogan, " Stress intensity factors in bonded half planes containing inclined cracks and subjected to antiplane shear loading", *Int. J. Fracture*, 15 (1979), 145-158

5. A. V. Boiko & L. N. Karpenko, "On some numerical methods for the solution of the plane elasticity problem for bodies with cracks by means of singular integral equations", *International Journal of Fracture*, 17 (1981), 381-388.

6. E. Brigham, *The fast fourier transform*, Prentice Hall, 1974.

7. P. Constantin, P. D. Lax and A. Majda, "A simple one dimensional model for the three dimensional vorticity equation", preprint, 1985.

8. M. Comninou, "The interface crack", *J. of Applied Mechanics*, Transactions of ASME, (1977), 631-636.

9. P. Concus and G. Golub, "A generalized conjugate gradient method for non-symmetrick systems of equations", Lectures notes in Economical and Mathematical Systems, Springer Verlag, Berlin, 1976.

10. D. Elliott, "The numerical treatment of singular integral equations- A review", *Treatment of Integral Equations by Numerical Methods*, Ed. C. Baker & G. Miller, Academic Press, 1982.

11. F. Erdogan and T. S. Cook, " Antiplane shear crack terminating at and going through a bimaterial interface", *Int. J. Fracture*, 10 (1974), 227-240.

12. F. Erdogan and G. D. Gupta, "On the numerical solution of singular integral equations", *Quart. Appl. Math.*, 29 (1972), 525-534.

13. N. Gastinel, "Inversion d'une matrice generalisant la matrice de Hilbert", *Chiffres* 3 (1960), 149-152.

14. A. Gerasoulis, "On the existence of approximate solutions for singular integral equations of Cauchy type discretized by Gauss-Chebyshev quadrature formulae", *BIT*, 21(1981), 377-380.

15. A. Gerasoulis, "Singular Integral Equations: Direct and Iterative Variant Methods", *Numerical Solution of Singular Integral Equations*, Eds. Gerasoulis & Vichnevetsky, IMACS publication, (1984), 133-141.

16. A. Gerasoulis, "Nyström's Iterative Variant methods in the solution of Cauchy singular integral equations", (In preparation)

17. A. Gerasoulis, M. D. Grigoriadis and Liping Sun, "A fast algorithm for Trummer's problem", *SIAM J. Sci. Stat. Comp.*, (1986).

18. G. Golub, "Trummer's Problem", Communication to NA.DIS@SU-SCORE, July 22, 1985.

19. G. Golub, "Trummer problem", *SIGACT News* 17 (1985), No. 2, ACM Special Interest Group on Automata and Computability Theory, p. 17.2-12.

20. M. Hashimoto, "A method for solving large matrix equations reduced from Fredholm integral equations of the second kind", *JACM*, 17 (1970), 629-636.

21. P. Henrici, "Fast Fourier methods in computational complex analysis", *SIAM Review*, 21 (1979), 481-527.

22. E. Horowitz, "A unified view of the complexity of evaluation and interpolation", *Acta Informatica*, 3 (1974), 123-133.

23. N. Ioakimidis, "On the numerical evaluation of derivatives of Cauchy principal value integrals", *Computing*, 27 (1981), 81-88.

24. N. Ioakimidis, "Three iterative methods for the numerical determination of stress intensity factors", *Engineering Fracture Mechanics*, 14 (1981), 557-564.

25. A. Kaya, "Applications of integral equations with strong singularities in Fracture Mechanics", Ph. D. thesis, Lehigh University, 1984.

26. C. T. Kelley and T. W. Mullikin, "Collocation Methods for some singular integral equations in linear transport theory", *Journal of Integral Equations*, 4 (1982), 77-88.

27. N. Kerzman and M. Trummer, "Numerical conformal mapping via the Szegö kernel", J. Computational & Applied Mathematics, 14 (1986), 111-123.

28. R. Krasny, "A numerical study of Kelvin-Helmholtz instability by the point vortex method", Ph. D. Thesis, Laurence Berkley Laboratory, University of California, LBL-17092, December 1983.

29. G. Monegato, "Convergence of product formulas for the numerical evaluation of certain two-dimensional Cauchy principal value integrals", *Numer. Math.* 43 (1984), 161-173.

30. A. M. Odlyzko and A. Schönhage, *Fast algorithms for multiple evaluations of the Reimann zeta function*, Technical Report, AT&T Bell Laboratories, Murray Hill, NJ, 1986.

31. P. Swarztrauber, FFTPACK, NETLIB@anl-mcs.ARPA, Argone National Laboratory

32. M. Trummer, "An efficient implementation of a conformal mapping method using the Szegö kernel", to appear in *SIAM J. Numerical Analysis*.

33. G. Tsamashpyros and P. S. Theocaris, "A recurrence formula for the direct solution of singular integral equations", *Computer Methods in Applied Mechanics and Engineering*, 31 (1982), 79-89.

I. Appendix

In this Appendix, we briefly discuss iterative algorithms for the solution of singular integral equations, for which *FAST* could be used to speed up computations.

1. Consider the Cauchy singular integral equation

$$\frac{1}{\pi}\int_{-1}^{1} w(t)\frac{y(t)}{t-s}dt + \lambda\int_{-1}^{1} w(t)K(t,s)y(t)dt = f(s), \quad |s|<1 \tag{29}$$

where $w(t)$ is a weight function, $K(s,t)$ and $f(s)$ are given input function.

By approximating the last equation via a quadrature similar to the one described for equation (13), we obtain the algebraic system

$$(\mathbf{A} + \lambda\mathbf{C})\mathbf{y} = \mathbf{f}, \quad (A)_{j,i} = \frac{w_i}{t_i - s_j}, \quad (C)_{j,i} = \pi w_i K(t_i, s_j), \quad j = 1, ..., m, \ i = 1, ..., n \tag{30}$$

where w_i are the quadrature weights (e.g. Gerasoulis [14]). The last algebraic system can be solved by using either a direct or an iterative method. Iterative methods such as the generalized conjugate gradient (Concus and Golub [9], Trummer [32]), Nyström's iterative variants (Gerasoulis [15], [16]), residual correction (Hashimoto [20]) and successive approximation (Tsamasphyros and Theocaris [33], Ioakimidis [24]), could be used to solve (30), particularly if the dimension of the system is large. As an example, we present the following iterative method for (29), (Tsamasphyros and Theocaris [33]),

$$F_l(t) = f(t) + \lambda\pi n^{-2}\sum_{i=1}^{n} K(t_i,t)\sum_{j=1}^{n-1}\frac{(1-s_j^2)}{s_j-t_i}F_{l-1}(t_i), \quad l = 1, .., \quad F_0(t) = f(t) \tag{31}$$

where $t_i = \cos((2i-1)\pi/2n)$, $i = 1, ..., n$, $s_j = \cos(j\pi/n)$, $j = 1, ..., n-1$, and $w(t) = (1-t^2)^{-1/2}$. The solution $y(t)$, for $|t| \leq 1$, is obtained from $F_l(t)$ via a summation, e.g. $y(1) \approx y_M(1) = n^{-1}F_M(1) + n^{-1}\sum_{j=1}^{n-1}(s_j+1)F_M(s_j)$, for M sufficiently large.

Now, if $K(t,s)$ is such that the resulting matrix $\mathbf{C}$ is similar to $\mathbf{B}_p$, then we can apply *FAST*. This is indeed the case for the following singular integral equations.

(a) *The Interface Crack equation:*

The weight function is given by $w(t) = (1-t^2)^{1/2}$ and the kernel by

$$K(t,s) = b\frac{(1-\gamma^2 t^2)^{1/2}(1-\gamma^2 s^2)^{-1/2}-1}{t-s}$$

where b and γ are given material constants. This equation has received considerable attention in the literature. Comninou [8] has obtained a numerical solution for $1-10^{-4} \leq \gamma < 1-10^{-7}$, by using direct methods for the linear algebraic system (30). Because the dimension of the algebraic system becomes very large for $1 > \gamma \geq 1-10^{-7}$, direct methods fail and the solution is not known in this region. As a result, certain important physical quantities, such as the stress

intensity factor, had to be conjectured (Comninou [8], pp. 633-634). We refer the reader to C. Atkinson et. al. [2] for a recent discussion regarding this problem. An iterative method, such as the ones discussed above, together with the $O_A(n \log n)$ implementation of *FAST*, presented in Section 4, may be found useful in solving this equation.

(b) *Equations with generalized kernels*

These equations arise in wave-guide theory, elasticity and dislocations in metallurgy. The kernel is given by $K(t,s) = 1/(t+s-2)$ and the weight function by $w(t) = (1-t)^\alpha(1+t)^{-1/2}$.

Numerical results reported in the literature differ somewhat, e.g. in Erdogan et. al. [12] the solution is $y(1) = 13.13$, while in Bassani et. al. [4] $y(1) = 15.863$. This is because, the quadrature method converges very slowly and large algebraic systems must be solved to attain a reasonable accuracy. Here, a general stable implementation of *FAST* along with an iterative technique could also be found useful.

2. Additional examples for which *FAST* is applicable can be found in the following references: The linear transport equation (Kelley et. al. [26], eq. 43), the antiplane shear crack (Tsamasphyros & Theocaris [33], eq. 46), edge crack (Boiko & Karpenko [5], eq. 20), multidimensional singular integrals (e.g. Monegatto [29]), Hadamard's singular integrals (e.g. Kaya [25], Ioakimidis [23]), Calogero's equation for the asymptotic density of the zeros of Hermite polynomials (D. Atkinson [3]), the H-function of Chandrasekhar, Neutron transport equation, Prandtl's equation for thin airfoil, the sail equation (e.g. Elliott [10]), Vorticity equation (Constantin et. al. [7], Krasny [28]).

A detail analysis of iterative methods for singular integral equations and their applications discused above is beyond the scope of this paper. Some of these methods have already been described in the literature (Tsamasphyros and Theocaris [33], Ioakimidis [24], Gerasoulis [15]), while others are currently under investigation and will be reported in a future communication (Gerasoulis [16]).

EFFECT OF ROTATION ON THE LATERAL STABILITY OF A FREE-FLYING COLUMN SUBJECTED TO AN AXIAL THRUST WITH DIRECTIONAL CONTROL

J. D. Vasilakis
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armament Center
Benet Weapons Laboratory
Watervliet, NY 12189-4050

J. J. Wu
U.S. Army European Research Office
London, England

ABSTRACT. This paper discusses some aspects of the stability problems of a free-flying column subjected to axial thrusts. In an age of spacecrafts and missiles, the stability of unsupported flying structures is obviously of great importance. Suprisingly though, there has not been a great deal of work addressing this type of problem. In this paper, first the brief history of the lateral stability of a column is reviewed, and then the basic characteristic features of the stability problem of a free-free column are discussed. The mathematical techniques developed to solve these problems depend on a particular problem considered. The most general case requires the solution of a nonself-adjoint differential equation/boundary condition system, which is homogeneous and with zero eigenvalues. Numerical procedures for such a system appear to work well, although theoretical proof of convergence is still lacking. Results of these procedures are discussed.

I. INTRODUCTION. In this paper, a long free-free slender beam is used as a model for a flexible missile or rocket. The beam behaves as a Bernoulli-Euler column, and in this case is assumed to be rotating about its longitudinal axis and subject to an end thrust (Figure 1). Of prime interest is the effect of the rotation on the lateral stability of the beam. The motion is assumed to be planar.

Different phases of the problem have been investigated in the past. A summary of the previous work is given in Reference 1. Silverberg [2] was the first to include thrust on the flying column. The differential equation for a free-flying beam was given earlier as shown in Reference 3. Beal [4] and Feodos'ev [5] obtained results with pulsating thrust. In 1972, Matsomato and Mote [6] treated a similar problem with directional thrust. In this case, however, feedback control was included and a time delay was applied to the control. The next contribution to understanding the problem was given by Peters and Wu [1]. They concentrated on mode shape solutions at zero frequency for different thrusts. A comprehensive description is also given in Reference [1] for the eigenvalues and mode shape near zero thrust and with a thrust direction close to a that of a follower force. Recently, Park and Mote [7] included a concentrated mass and feedback control. The feedback control included was allowed to be from different points along the beam.

As stated above, in this paper the effect of rotation on the stability of a free-free beam is assessed. The next section is a description of the problem. In Section III the variational statement used for the solution is described. Section IV shows how the variational statement is used with finite elements to solve the problem, and Section V discusses the results of the investigation.

II. **PROBLEM STATEMENT**. The geometry of the problem is shown in Figure 1. The beam has a constant cross-section of area A, density ρ, Young's modulus E, and moment of inertia I. It shows a free-flying column subject to axial thrust with directional control and rotating about its axis. The differential equation for the beam is given by

$$EIu^{iv} + P(\frac{x}{\ell} u')' + \rho A\ddot{u} + \rho A\Omega^2 u = 0 \qquad (1)$$

The first three terms represent the column as treated in Reference 1. The last term on the left hand side shows the effect of the rotation. The boundary conditions are given by

$$u''(0) = 0 \quad , \quad u''(\ell) = 0$$

$$u'''(0) = 0 \quad , \quad EIu'''(\ell) - K_Q Pu'(\ell) = 0 \qquad (2)$$

In dimensionless form with

$$\bar{u} = u/\ell \quad , \quad \bar{x} = x/\ell \quad , \quad \tau = t/T$$

$$T^2 = \frac{\rho A\ell^4}{EI} \quad , \quad Q = \frac{P\ell^2}{EI} \quad , \quad \omega = \frac{\Omega}{T} \qquad (3)$$

and writing

$$\bar{u}(x,t) = \bar{u}(x)e^{\lambda t} \qquad (4)$$

the differential equation then becomes

$$\bar{u}'''' + Q(\bar{x}\bar{u}')' + \lambda^2\bar{u} + \omega^2\bar{u} = 0 \qquad (5)$$

with the boundary conditions

$$\bar{u}''(0) = 0$$

$$\bar{u}'''(0) = 0$$

$$\bar{u}''(1) = 0$$

$$\bar{u}'''(1) - K_\theta Q[\bar{u}'(1)] = 0 \qquad (6)$$

1298

Rewriting Eq. (5) as (and dropping hats)

$$u'''' + Q(xu') + (\lambda^2 + \omega^2)u = 0 \qquad (7)$$

it appears that the addition of rotation simply shifts the frequency of vibration of the system. The boundary conditions, Eq. (6), become

$$u''(0) = 0$$

$$u'''(0) = 0$$

$$u''(1) = 0$$

$$u'''(1) - K_\theta Qu'(1) = 0 \qquad (8)$$

The spacial variables are made dimensionless by dividing through by the beam's length ℓ and time is made dimensionless by dividing through by a constant $T = (\rho A \ell^4 / EI)^{\frac{1}{2}}$ which has the units of time.

The parameter λ is a complex number in general

$$\lambda = \lambda_R + i\lambda_I$$

where both λ_R and λ_I are real numbers.

III. VARIATIONAL STATEMENT. To find the form of the variational statement, the differential equation is multiplied by an arbitrary variation of the adjoint field variable, $\delta v(x)$, and integrated over the beam length. Integration-by-parts indicates the form of the variational statement and the natural boundary conditions. The variational statement is given by

$$\delta J = 0 \qquad (9)$$

where

$$J = \int_0^1 [u''v'' - Qxu'v' + (\lambda^2 + \omega^2)uv]dx + Q(1+K_\theta)u'(1)v(1) \qquad (10)$$

Performing the variation of J with respect to u and v, one can arrive at the original boundary value problem as well as the adjoint. Equation (10) is the basis for a finite element solution to the described problem.

IV. FINITE ELEMENT AND NUMERICAL FORMULATION. The procedure begins by taking the variation of Eq. (10) and allowing the variations in the problem variable, $\delta u(x)$, to be zero, i.e., varying adjoint variable $v(x)$ only for now,

$$\int_0^1 [u''\delta v'' - Qxu'\delta v' + \Lambda^2 u\delta v]dx - Q(1+K_\theta)u'(1)\delta v(1) = 0 \qquad (11)$$

where $\Lambda^2 = \lambda^2 + \omega^2$. To discretize, the beam is divided into L elements, letting

$$\xi = L\left(x - \frac{i-1}{L}\right) \qquad i = 1,2,3....,L \qquad (12)$$

be the running coordinate in each element. Substituting Eq. (12) into Eq. (11)

$$\sum_{i=1}^{L} \int_0^1 [L^3 i^{(i)"} \delta v^{(i)"} - Q\{\xi + (i-1)\}u^{(i)'} \delta v^{(i)'} + \frac{\Lambda^2}{L} u^{(i)} \delta v^{(i)}]ds$$

$$- Q(1+K_\theta)u^{(L)'}(1)\delta v^{(L)}(1) = 0 \qquad (13)$$

In order that the displacements and their derivatives within an element be expressed in terms of their nodal values, the coordinate vectors are introduced.

$$\bar{U}^{(i)T} = \{U_1^{(i)} \quad U_2^{(i)} \quad U_3^{(i)} \quad U_4^{(i)}\}$$

$$\bar{V}^{(i)T} = \{V_1^{(i)} \quad V_2^{(i)} \quad V_3^{(i)} \quad V_4^{(i)}\} \qquad (14)$$

$U_1^{(i)}$, $U_2^{(i)}$ represent the displacement and slope at the left end of the ith element, and $U_3^{(i)}$ and $U_4^{(i)}$ represent deflection and slope at the right end. A similar interpretation is applied to the adjoint coordinate vector $\bar{V}^{(i)}$. The transform is indicated by T.

Hermitian polynomials are used to relate the displacements within an element to its nodal values, hence, the following shape function is assumed

$$\bar{a}^T(\xi) = \{1 - 3\xi^2 + 2\xi^3 \quad , \quad \xi - 2\xi^3 + \xi^3 \quad , \quad 3\xi^2 - 2\xi^3 \quad , \quad -\xi^2 + \xi^3\} \qquad (15)$$

so that

$$u^{(i)}(\xi) = \bar{a}^T(\xi)\bar{U}^{(i)}$$

$$v^{(i)}(\xi) = \bar{a}^T(\xi)\bar{V}^{(i)} \qquad (16)$$

Substituting Eq. (16) into Eq. (13)

$$\sum_{i=1}^{L} \bar{U}^{(i)T}\{L^3\bar{C} - Q[\bar{D}+(i-1)\bar{B}] + \frac{\Lambda^2}{L}\bar{A}\}\delta\bar{V}^{(i)} - Q[1+K_\theta]\bar{U}^{(L)T}\bar{E}\delta\bar{V}^{(L)} = 0 \qquad (17)$$

with

$$\bar{A} = \int_0^1 \bar{a}\bar{a}^T d\xi \quad , \quad \bar{B} = \int_0^1 \bar{a}'\bar{a}'^T d\xi \quad , \quad \bar{C} = \int_0^1 \bar{a}''\bar{a}''^T d\xi$$

$$\bar{D} = \int_0^1 \xi\bar{a}'\bar{a}'^T d\xi \quad , \quad \bar{E} = \bar{a}'(L)\bar{a}^T(L) \qquad (18)$$

Rewriting Eq. (17),

$$\sum_{i=1}^{L} \bar{U}^{(i)T}\{\Lambda^2 P^{(i)} + S^{(i)}\}\delta\bar{V}^{(i)} = 0 \qquad (19)$$

1300

where

$$P^{(i)} = \bar{A}/L \qquad\qquad i = 1,2,\ldots,L$$

$$S^{(i)} = L^3\bar{C} - Q[\bar{D} + (i-1)\bar{B}] \qquad i = 1,2,\ldots,L-1$$

$$S^{(L)} = L^3\bar{C} - Q[\bar{D} + (L-1)\bar{B}] - Q(1+K_\theta)\bar{E} \tag{20}$$

Using certain continuity conditions between the element nodal values

$$U_1^{(i)} = U_3^{(i-1)} \qquad V_1^{(i)} = V_3^{(i-1)}$$

$$U_2^{(i)} = U_4^{(i-1)} \qquad V_2^{(i)} = V_4^{(i-1)} \tag{21}$$

One can write

$$\bar{U}^T = \{U_1^{(1)} \quad U_2^{(1)} \quad U_3^{(1)} \quad U_4^{(1)} \quad U_3^{(2)} \quad U_4^{(2)} \quad \ldots \quad U_3^{(L)} \quad U_4^{(L)}\}$$

$$\bar{V}^T = \{V_1^{(1)} \quad V_2^{(1)} \quad V_3^{(1)} \quad V_4^{(1)} \quad V_3^{(2)} \quad V_4^{(2)} \quad \ldots \quad V_3^{(L)} \quad V_4^{(L)}\} \tag{22}$$

Finally, [P] and [S] are NxN matrices with N = 2L+2. Since δv is arbitrary, the eigenvalue problem reduces to

$$\bar{U}^T\{\Lambda^2[P] + [S]\} = 0 \tag{23}$$

which is solved for the eigenvalues.

V. CONCLUSIONS AND DISCUSSION. In this paper, we have included rotation about the longitudinal axis in the dynamic stability study of a free-flying missile subjected to axial thrusts. It is assumed that the motions of bending and the thrust are in the same plane. In the differential equation, the only difference resulting from the introduction of rotation is a change in the frequency parameter Λ^2 to

$$\Lambda^2 = \lambda^2 + \omega^2 \tag{24}$$

where ω is the rotation. Consequently, all the stability curves obtained previously [1] can be used with some simple modifications. It should be noted that in Reference [1], we have written (with $\omega = 0$)

$$\Lambda = \lambda = \lambda_R + i\lambda_I \tag{25}$$

and the stability character of the problem is indicated by: (1) stable vibrations = $\lambda_I \neq 0$, $\lambda_R = 0$; (2) unstable by buckling (divergence) = $\lambda_R \neq 0$, $\lambda_I = 0$; (3) unstable by flutter = $\lambda_R \neq 0$, $\lambda_I \neq 0$; and (4) marginally stable = $\lambda_I = \lambda_R = 0$.

For the present case, the stability behavior is indicated as above, but with Λ_I and Λ_R replacing λ_I and λ_R in the previous stability curves

$$\Lambda = \Lambda_R + i\Lambda_I \tag{26}$$

and

$$\Lambda^2 = (\Lambda_R + i\Lambda_I)^2 = \lambda^2 + \omega^2 = (\lambda_R + i\lambda_I)^2 + \omega^2 \tag{27}$$

or

$$\lambda^2 = (\lambda_R + i\lambda_I)^2 = \Lambda^2 - \omega^2 = (\Lambda_R + i\Lambda_I)^2 - \omega^2 \tag{28}$$

From Eq. (28), when $\Lambda_R = 0$, $\lambda^2 = -\Lambda_I^2 - \omega^2$, hence $\lambda_R = 0$ and $\lambda_I^2 = \Lambda_I^2 + \omega^2$. Thus, originally stable vibrations will remain stable with higher vibration frequency. On the other hand, when $\Lambda_I = 0$, $\lambda^2 = \Lambda_R^2 - \omega^2$, hence $\lambda^2 = \Lambda_R^2 - \omega^2$. Thus, originally divergent motions will become stable vibrations when $\Lambda_R^2 < \omega^2$. In the case of marginal stability $\Lambda = 0$ will certainly be stabilized since $\lambda_I^2 = \omega^2$.

In the case of flutter instability, Eq. (28) states that λ is complex ($\lambda_I \neq 0$, $\lambda_R \neq 0$) if and only if Λ is complex ($\Lambda_I \neq 0$, $\Lambda_R \neq 0$). Therefore, the flutter instability is not effected by the introduction of the rotation, which is an interesting observation.

Several demonstrative stability curves with λ^2 (and Λ^2) versus Q/π^2 are shown in Figures 2 through 5. Only the lowest eigenvalue's branches are shown, since they are the ones which dictate the stability behavior. Figure 2 shows the two lowest stable vibration modes and two rigid body modes on the $\Lambda^2 = 0$ axis. This is the case of a free-flying missile with a follower thrust ($K_\theta = 0$) and with a dimensionless rotation of $\omega^2 = 500$. The two fluxural modes coalesce at load $Q/\pi^2 = 11.18$ beyond which flutter instability begins. The rigid body modes without rotation indicate marginal stability. Due to the rotation ω, the axis is shifted from $\Lambda^2 = 0$ to $\lambda^2 = 0$, therefore, these previously rigid body modes are now stable modes of vibrations. The thrust that is controlled with a small negative tangency ($K_\theta = -0.05$) is shown in Figure 3. It is noted in this figure that the divergence instability without rotation is stabilized by $\omega^2 = 500$. However, the new critical load is lowered from $Q/\pi^2 = 11.18$ to 5.30, not because of ω^2, but due to the negative control parameter K_θ. Figure 4 shows the case of $K_\theta = -1$ or that the thrust has a fixed direction of the inertia axis. It is clear that the divergence instability of the lowest branch is stabilized so that the critical load has been raised from zero to $Q_{CR} = 1.50\,\pi^2$. Finally, the case for a small positive tangency control parameter ($K_\theta = 0.05$) is shown in Figure 5. In this figure, the original divergence instability at $Q/\pi^2 = 3.00$ is stabilized by ω^2. However, the original critical load of flutter instability at $Q/\pi^2 = 9.90$ is not changed by the rotation. Hence, the critical load in this case is raised from 3.00 to 9.90 due to the rotation of $\omega^2 = 500$.

REFERENCES

1. Peters, D. A. and Wu, J. J., "Asymptotic Solutions to a Stability Problem," _Journal of Sound and Vibration_, Vol. 59, No. 4, 1978, pp. 591-610.

2. Silverberg, S., "The Effect of Longitudinal Acceleration Upon the Natural Modes of Vibration of a Beam," Technical Report TR-59-0000-00791, Space Technology Laboratories, 1959.

3. Timoshenko. S. and Young, D. H., _Vibration Problems in Engineering_, D. von Nostrand Inc.. New York, 1955, pp. 297-303.

4. Beal, T. R., "Dynamic Stability of a Flexible Missile Under Constant and Pulsating Thrusts," _American Institute of Aeronautics and Astronautics Journal_, Vol. 3, 1965, pp. 486-494.

5. Feodos'ev, V. I., "On a Stability Problem," _Prikladnaia Matematika I Mekhanika_, Vol. 29, 1965, pp. 391-392 (Translated from Russian).

6. Matsumoto, G. Y. and Mote, C. D., Jr., "Time Delay Instabilities in Large Order Systems With Controlled Follower Forces," _Journal of Dynamic Systems, Measurement, and Control_, December 1972, pp. 330-334.

7. Park, Y. P. and Mote, C. D., Jr., "The Maximum Controlled Follower Force on a Free-Free Beam Carrying a Concentrated Mass," _Journal of Sound and Vibration_, 1985, Vol. 98, No. 2, pp. 247-256.
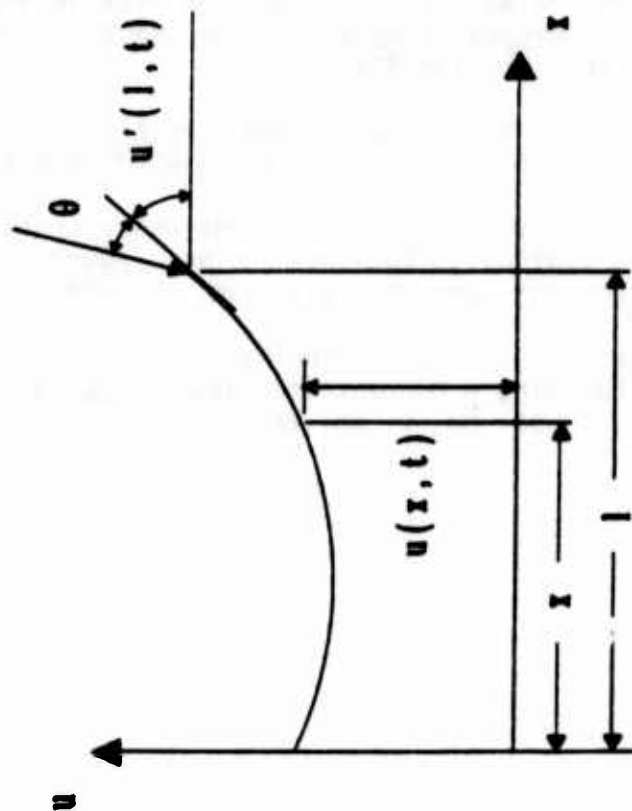
FIGURE 1. GEOMETRY OF THE PROBLEM
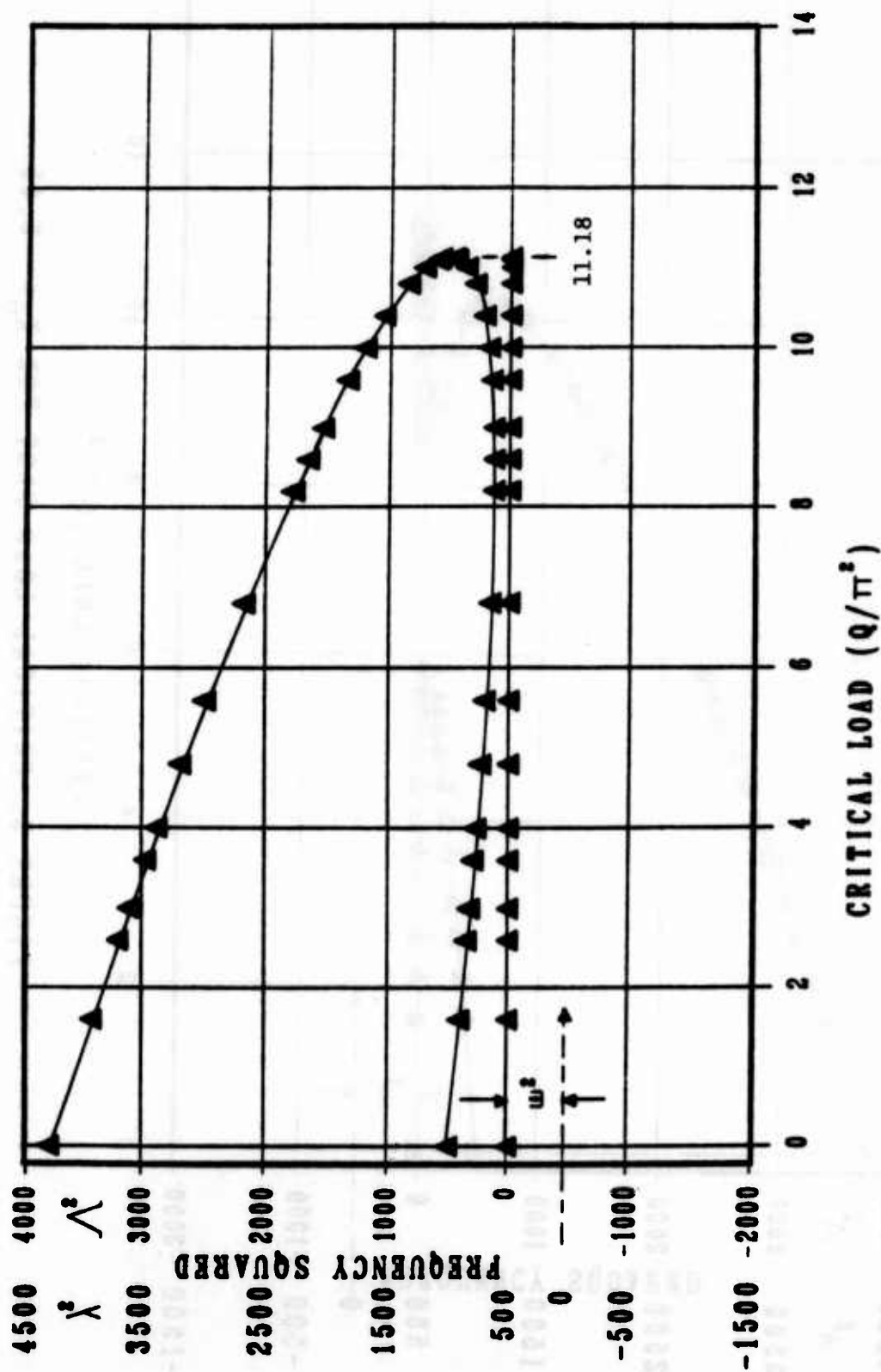
CRITICAL LOAD (Q/π^2)

FIGURE 2. CRITICAL LOAD PLOT FOR $K_\theta = 0.00$

(FOLLOWER FORCE)
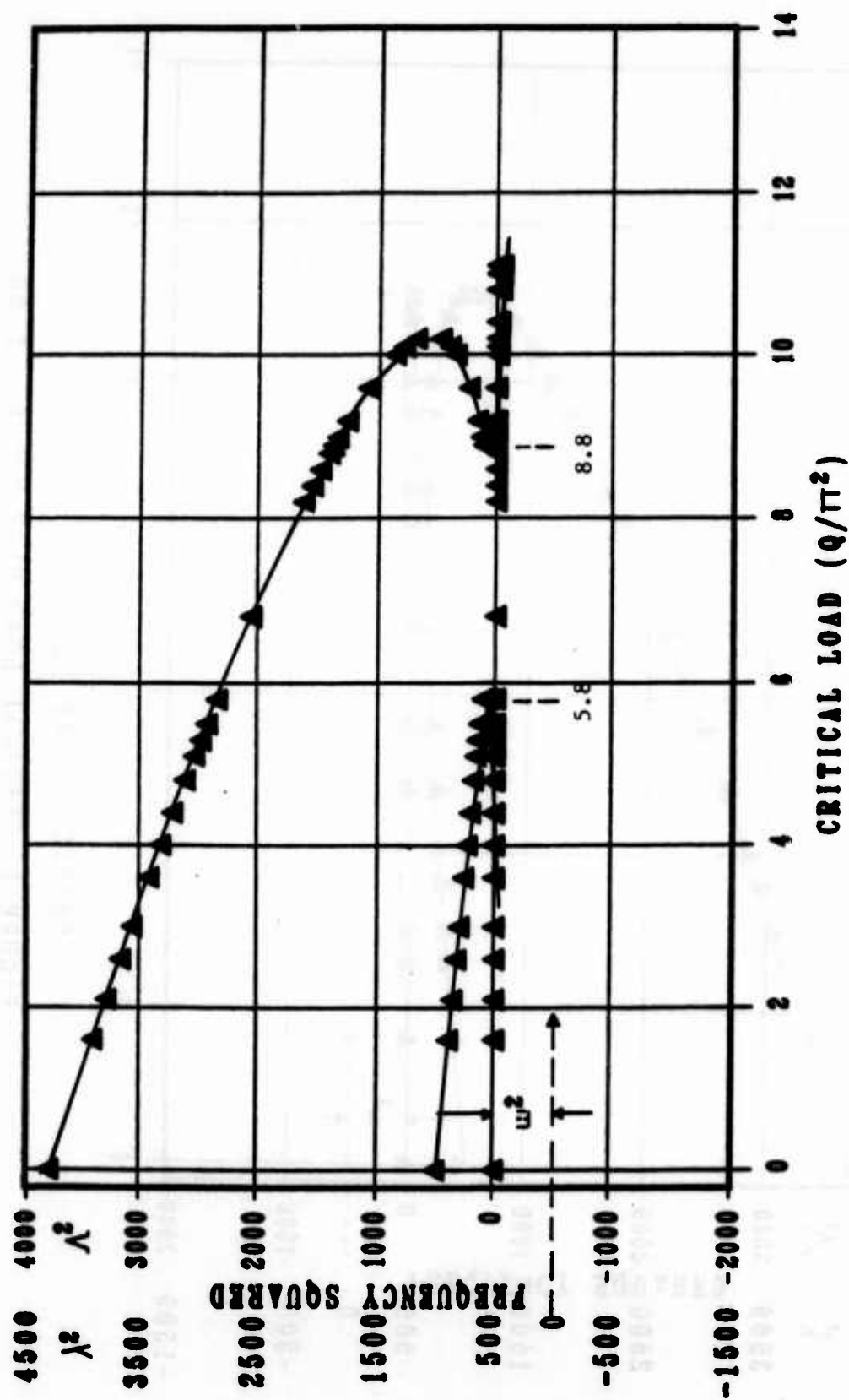
FIGURE 3. CRITICAL LOAD PLOT FOR $K_\theta = -0.05$
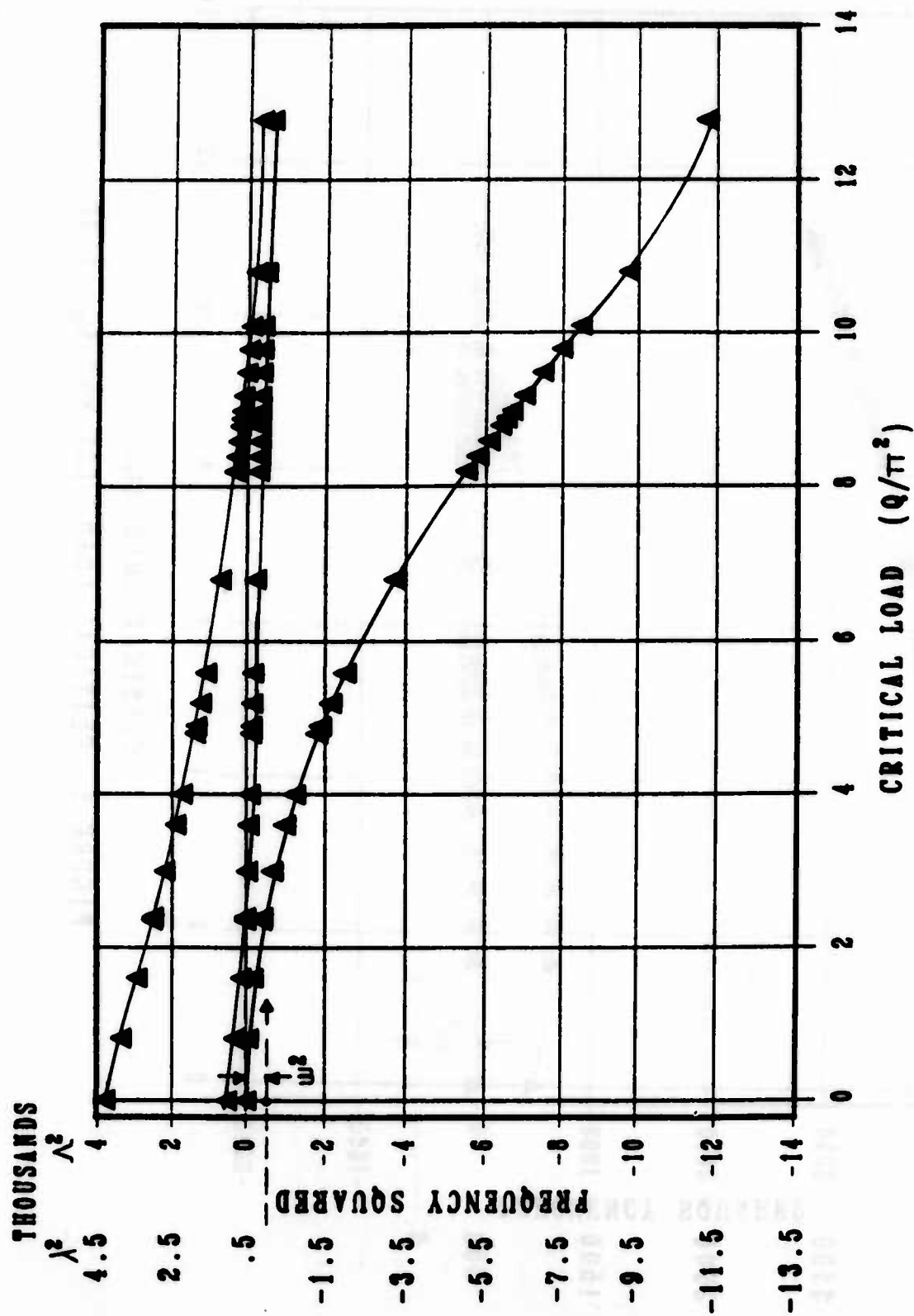
FIGURE 4. CRITICAL LOAD PLOT FOR $K_\theta = -1.00$
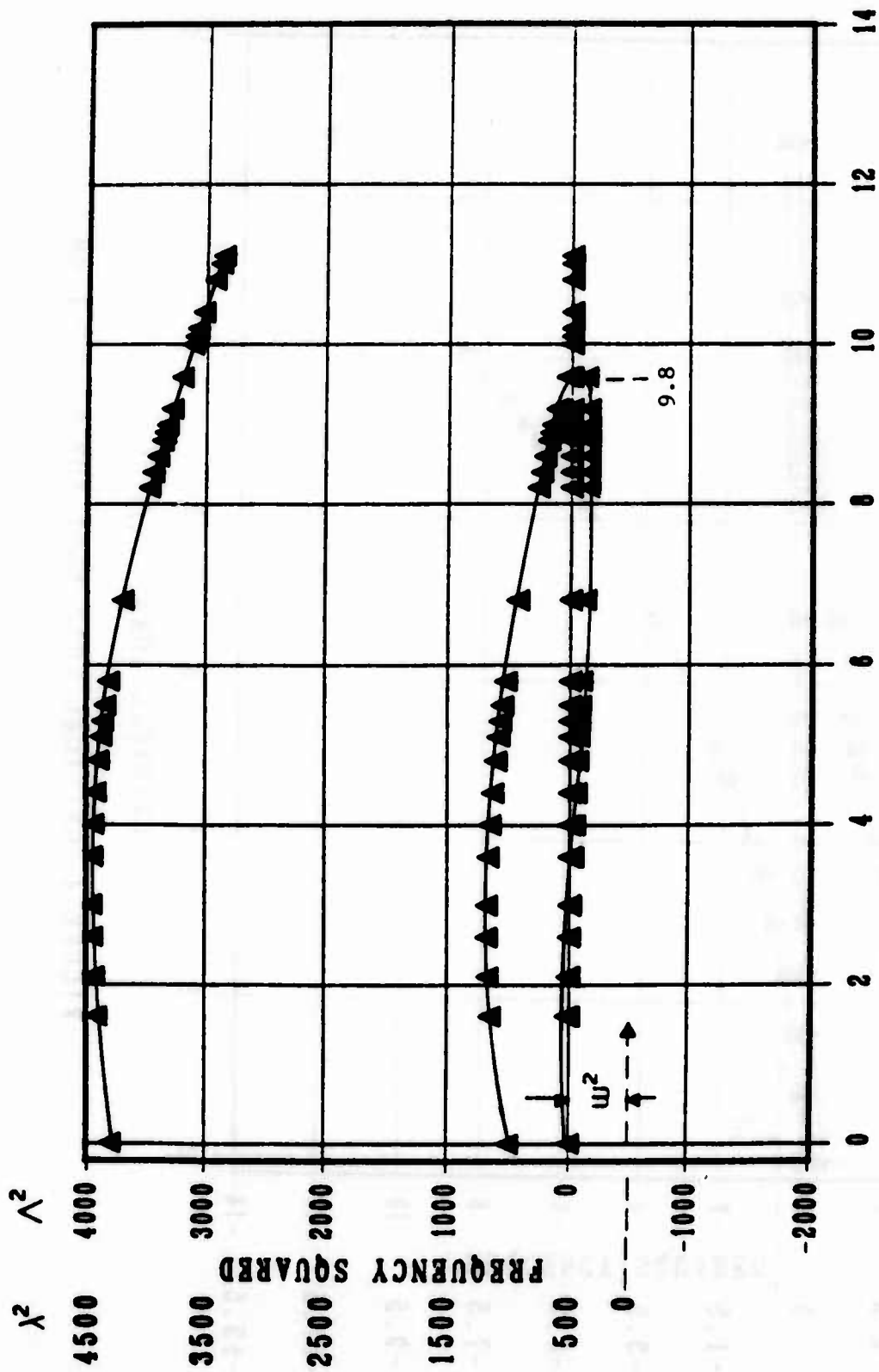
CRITICAL LOAD (Q/π^2)

FIGURE 5. CRITICAL LOAD PLOT FOR $K_\theta = 1.00$

DETONATION WAVE INITIATION BY RAPID ENERGY DEPOSITION AT A CONFINING BOUNDARY

D.R. Kassoy, Mechanical Engineering Department
University of Colorado, Boulder 80309

J.F. Clarke, College of Aeronautics
Cranfield Institute of Technology
Cranfield, England

N. Riley, School of Mathematics and Physics
University of East Anglia
Norwich, England

ABSTRACT. A study is made of planar detonation wave birth and evolution in a reacting gas mixture. The chemistry is described by the high activation energy global reaction A→B. A prescribed heat flux, applied at a planar boundary, is used to initiate the thermomechanical processes which result in detonation. Finite difference methods are used to solve the one-dimensional, compressible, unsteady describing equations which include reaction effects and transport terms. Early power deposition at the boundary heats an adjacent thin layer of gas in which significant chemical heat release occurs. The total power deposited generates thermomechanical effects which cause a fully resolved shock to propagate away from the boundary. The shock conditions unreacted gas and thereby initiates a reaction process that propagates with a speed similar in magnitude to the shock. The resulting chemical power deposition enhances the shock strength, which in turn accelerates the reaction process further. The total rate of chemical energy release increases relentlessly evolving suddenly into a power pulse nearly 100-times larger that the initial boundary heat flux. This explosive process leads to the formation of a coupled shock-reaction zone structure that propagates as an identifiable entity.

1. *INTRODUCTION*. The work presented at the Fourth Army Conference on Applied Mathematics and Computing was derived from Clarke et al. (1). A complete copy of this manuscript is available upon request to the first author. A summary of the key results is given in the next section.

II. *SUMMARY*. An extension of Clarke et al. (2) to a reactive gas mixture provides a basis for studying the transient development of a planar detonation consisting of a fully resolved lead shock followed by a reaction zone. The detonation propagates away from a planar confining boundary leaving behind a hot, reactant depleted, zone in which a variety of weak gasdynamic waves can be observed.

The mathematical model is based on the equations for a reactive compressible, perfect gas with transport effects. Initially the gas is at rest, at a temperature of 300K and a pressure of 1.01×10^5Pa. A heat flux of 10^6 W/m² is applied on the planar boundary during a period of

$O(10^{-9}s)$. Solutions obtained from an implicit finite difference calculation show that on a time-scale of $O(10^{-8}s)$ there is a ten-fold increase in global power deposition associated with chemical reaction induced by shock passage. During an ensuing transitional period of $O(10^{-7}s)$ the reaction rate is enhanced gradually and the shock is strengthened accordingly. The propagation speed of the shock is larger than that of the clearly indentifiable reaction zone although both move at supersonic speed relative to the undisturbed gas. The transition period is terminated dramatically by a rapidly developing burst of power deposition of magnitude 10^{10} W/m². This rapid reaction process develops during a period of about $2 \times 10^{-9}s$. and in a region 5×10^{-4} cm in extent. The explosion is localized sufficiently in space and time to ensure that an inertially dominated heat-addition process occurs. It follows that the temperature rise is associated with an enormous pressure increase, some 50 times the initial value of 1 atmosphere. Localized reactant depletion terminates the explosive process and the global power deposition declines. Further shock strengthening, resulting from compression waves generated by the pressure buildup, leads to a significant reduction in the ignition delay time. As a result, the reaction zone Mach number accelerates rapidly up to that of the shock and the two structures propagate away together like a ZND-wave. The reaction zone structure itself is not unlike that described by Kassoy and Clarke (3).

In conclusion, our calculations show that direct initiation of detonation requires sufficent power input to first of all generate a suitable strong precursor shock wave, which then becomes the trigger to switch on vigorous chemical activity in its wake. The hallmark of this vigor is its capacity to exploit the inertia of the fluid by raising local pressures and temperatures, with little diminution in local density; the pressure waves so formed propagate and increase precursor shock strength, which therefore lifts overall density levels, as well as those of pressure and temperature. All of these processes interlock in a continuously accelerated sequence that progresses toward a steady state in the shape of a ZND detonation.

Although we have restricted our attentions here to "direct initiation", as it is called, it must be said that it is difficult to imagine any other sequence of events over longer times provided that the precursor shock strength is not allowed, or forced, to decay in the transitional time domain. It is here that intial input energy, as opposed to power, can have an important part to play in preventing shock decay, by geometric attenuation for example in two and three dimensions. In the one-dimensional geometry of the present study calculations carried out for finite switch-off times σ_Q of boundary heat flux show that the initiation process is only delayed by reductions of σ_Q, and hence, of input energy. There is no suggestion from our calculations that the "formation" and "ZND-like" events will not always eventually follow the "transitional" ones. Thus, in one-dimension our calculations imply that initiation of detonation will always follow deposition of power, no matter how little the energy. Of course, this implies in its turn that unlimited distance is available for the precursor shock to travel. In reality this distance is always limited and so, as power diminishes, we would expect energy to rise in monotone fashion for detonation initiation to take place within a

given configuration. This is just what Dabora (4) finds in his experiments with hydrogen-oxygen-nitrogen mixtures in a shock tube. We remark that Dabora's boundary input power levels range from 2×10^2 to 10^3 MW/m^2, which is precisely in the range of values that we have studied.

BIBLIOGRAPHY

1. J. F. Clark, D. R. Kassoy and N. Riley, "On the Direct Initiation of a Plane Detonation Wave," in press, Proc. Roy. Soc. London (1986).

2. J. F. Clark, D. R. Kassoy and N. Riley, "Shocks Generated in a Confined Gas Due to Rapid Heat Addition at the Boundary. II, Strong Shock Waves". Proc. Roy. Soc., *A*, 393, 331-351 (1984).

3. D. R. Kassoy, and J. F. Clarke, "The Structure of a Steady High-Speed Deflagration with a Finite Origin". J. Fluid Mech., *150*, 253-280 (1985).

4. E. K. Dabora, "The Relation Between Energy and Power for Direct Initiation of Hydrogen-Air Detonations." Paper presented at the Second International Workshop on the Impact of Hydrogen on Water Reactor Safety, Albuquerque, N.M. (1982).

INTERACTION OF ROTATING BAND AND RIFLING*
IN ARTILLERY PROJECTILES

S. Handgud and H.P. Chen
School of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30332

T. Tsui
Army Materials & Mechanics Research Center
Watertown, MA 02172

ABSTRACT

The interaction between the rotating band and the rifling is usually used
to provide a desired angular acceleration or spin to a projectile as the
projectile accelerates along the length of the barrel of a gun. Evaluations
and improvements of the design of rotating bands need an understanding of
the stresses and deformations of the band resulting from the interaction. An
accurate knowledge of these requires a three dimensional solution of the problem.
However, significant amount of the mechanics of interaction can be understood
by studying an idealized two dimensional problem that considers only the
circumferential flow of the rotating band material into the rifling groove.
In this report, a numerical solution procedure has been discussed for a two
dimensional problem. The circumferential flow has been discussed for specific
cases.

*This paper was presented at the Third Army conference on Applied Mathematics
and Computing.

I. INTRODUCTION.

The interaction, between the rotating band and the rifling, is usually
used to provide a desired spin or angular acceleration to a projectile as the
projectile accelerates along the length of the barrel of a gun. This interac-
tion takes place in the following way. Before entering the barrel, a fired
projectile enters a "forcing cone". The outer diameter of the projectile is
usually designed to be smaller than the minor diameter of the bore. However,
the outer diameter of the rotating band which is located on the outer hard
surface of the projectile is larger than the minor diameter of the bore (See
Fig. 1) As a result, the radial dimensions of the rotating band are reduced
at first in the forcing cone and then in the barrel.

Inside the forcing cone, the reduction in the radial dimensions of the
rotating band is accompanied by an axial flow along the length of the projec-
tile. The deformation process is axisymmetric since the forcing cone has a
smooth wall. Upon encountering the rifling in the barrel, radial flow, axial
flow along the length of the projectile and circumferential flow into the
rifling groove occur. The deformation becomes non-axisymmetric and hence a
three dimensional description is necessary. It is to be noted here that the
rifling has a twist along the length of the barrel. It is the combination of
the aforementioned deformation of the rotating band and the designed twist of
the rifling along the length of the barrel that result in an angular accelera-
tion being imparted to the projectile as the projectile travels along the
length of the barrel.

The functions of a rotating band impose certain basic requirements on the
band material and geometry. The rotating band must have sufficient "plastic
flow characteristics" to deform from its initial configuration to a shape
dictated by the forcing cone and the rifling in order to reduce the wear of the
barrel. At the same time, must also have sufficient strength to

  (a)   transmit the torque to the projectile,
  (b)   withstand the propellant gas pressure, and
  (c)   provide the required attachment characteristics to
    the projectile.

As a consequence of these requirements, evaluations and improvements of the design of rotating bands need an understanding of the stresses and deformations of the band which should be obtained by including in the analysis the following parameters:

(a) flow strength of the rotating band material,
(b) breathing strength of the projectile wall,
(c) depth of the rotating band, and
(d) the interference.

The results of a preliminary study of the stresses and deformations in the rotating band of an artillery projectile caused by the interaction of the band and the rifling is presented in this report. The results of the interaction of the rotating band and the forcing cone have been examined by other investigators as a one dimensional problem [1].

II. DESCRIPTION OF ANALYTICAL MODEL.

As discussed before, the interaction of the rotating band and rifling causes the band material to flow in three directions. Accurate determination of the corresponding states of stresses and deformations in the band requires a three dimensional, elastic-plastic, large deformation dynamic analysis. In the present study, a simplified two dimensional problem with only radial and circumferential flow is analyzed. Since the distance between two neighboring rifling grooves is sufficiently small compared to the radius of the projectile, the effect of the curvature can be neglected. As a result, the simplified problem can be considered as a two dimensional plane strain problem as shown in Fig. 2. The gun barrel is assumed to move along the y direction into the rotating band. The initial velocity of the gun barrel is determined on the basis of the longitudinal velocity of the projectile and the amount of interference between rotating band and the gun barrel.

In general, the rotating band material is much softer than that of the gun barrel and the projectile resulting in deformations. The following assumptions are made in the first stage of the analysis:

(a) Gun barrel and projectile are both rigid,
(b) the rotating band material is modeled by an elastic-perfect plastic material and strain rate effect is absent,
(c) perfect bonding exists between the rotating band and the projectile.

III. GOVERNING EQUATIONS

In the analysis, ξ, η are used to denote the undeformed coordinates. After the body deforms, x,y, (ξ,η,t) are used to denote the deformed configuration of the particle at time t whose initial coordinates are (ξ, η). The governing equations are then written in terms of Cauchy stresses and Lagrangian coordinates. These equations, for a two-dimensional plane strain problem, are as follows:

Kinematics equations:

$$\dot{x} = u, \quad \dot{y} = v \tag{1}$$

Equations of motion:

$$\dot{u} = \frac{1}{\rho} \frac{\partial \sigma_{xx}}{\partial x} + \frac{1}{\rho} \frac{\partial \tau_{xy}}{\partial y}$$

$$\dot{v} = \frac{1}{\rho} \frac{\partial \tau_{xy}}{\partial x} + \frac{1}{\rho} \frac{\partial \sigma_{yy}}{\partial y} \tag{2}$$

Constitutive equations:

$$\dot{\sigma}_{xx} = (\rho \frac{dp}{d\rho} + \frac{4}{3} G)\frac{\partial u}{\partial x} + (\rho \frac{dp}{d\rho} - \frac{2}{3}G) \frac{\partial v}{\partial y} + \tau_{xy}(\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}) \tag{3}$$

$$\dot{\sigma}_{yy} = (\rho \frac{dp}{d\rho} - \frac{2}{3} G) \frac{\partial u}{\partial x} + (\rho \frac{dp}{d\rho} + \frac{4}{3} G) \frac{\partial v}{\partial y} + \tau_{xy}(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}) \tag{4}$$

$$\dot{\tau}_{xy} = G \frac{\partial v}{\partial x} + G \frac{\partial u}{\partial y} + \frac{1}{2} (\sigma_{xx} - \sigma_{yy})(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}) \tag{5}$$

$$\dot{\alpha}_{zz} = \nu(\sigma_{xx} + \sigma_{yy}) \tag{6}$$

Continuity equation:

$$\dot{\rho} = - \rho \frac{\partial u}{\partial x} - \rho \frac{\partial v}{\partial y} \tag{7}$$

It is assumed that stress deviators and pressure can be defined as follows:

$$S_{xx} = p + \sigma_{xx}$$

$$S_{yy} = p + \sigma_{yy} \tag{8}$$

$$S_{zz} = -S_{xx} - S_{yy}$$

where

$$p = - \frac{1}{3}(\sigma_{xx} + \sigma_{yy} + \sigma_{zz}) \tag{9}$$

The von Mises yield condition for elastic-perfect plastic material is given by:

$$S_{xx}^2 + S_{yy}^2 + S_{zz}^2 + 2 \tau_{xy}^2 \leq \frac{2}{3} Y^2 \tag{10}$$

1316

u: velocity in x-direction

v: velocity in y-direction

σ_{xx}:   normal stress in x-direction

σ_{yy}:   normal stress in y-direction

σ_{zz}:   normal stress in z-direction

τ_{xy}:   shear stress

S_{xx}:   normal stress deviator in x-direction

S_{yy}:   normal stress deviator in y-direction

S_{zz}:   normal stress deviator in z-direction

ρ: density

p: hydrostatic pressure

G: modulus of rigidity

Y: yield stress in simple tension

Equations (1) to (5) and (7) are written in the following compact form

$$\{\dot{u}\} = [A]\,\{u\}_{,x} + [B]\,\{u\}_{,y} \tag{11}$$

where

$$\{u\}^T = \{u,\ v,\ \sigma_{xx},\ \sigma_{yy},\ \tau_{xy},\ \rho\}.^T$$

The matrices A and B are not constants in the finite deformation problem of elastic-plastic materials and are given as follows:

$$[A] = \begin{bmatrix} 0 & 0 & 1/\rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/\rho & 0 \\ \rho\,\dfrac{dp}{d\rho} + \dfrac{4}{3}G & -\tau_{xy} & 0 & 0 & 0 & 0 \\ \rho\,\dfrac{dp}{d\rho} - \dfrac{2}{3}G & \tau_{xy} & 0 & 0 & 0 & 0 \\ 0 & G + \dfrac{1}{2}(\sigma_{xx} - \sigma_{yy}) & 0 & 0 & 0 & 0 \\ -\rho & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1317

$$[B] = \begin{bmatrix} 0 & 0 & 0 & 0 & 1/\rho & 0 \\ 0 & 0 & 0 & 1/\rho & 0 & 0 \\ \tau_{xy} & \rho\frac{dp}{d\rho} - \frac{2}{3}G & 0 & 0 & 0 & 0 \\ \cdot\tau_{xy} & \frac{dp}{d\rho}\rho + \frac{4}{3}G & 0 & 0 & 0 & 0 \\ G - \frac{1}{2}(\sigma_{xx} - \sigma_{yy}) & 0 & 0 & 0 & 0 & 0 \\ 0 & -\rho & 0 & 0 & 0 & 0 \end{bmatrix}$$

From the assumed two dimensional plane strain case, the boundary conditions of a typical part of rotating band as shown in Figure 3 are as follows:

On the boundary

$$\xi = 0, \qquad\qquad 0 \leq \eta \leq R,$$
$$u = 0, \qquad\qquad \tau_{xy} = 0.$$

On the boundary

$$\xi = L, \qquad\qquad 0 \leq \eta \leq R,$$
$$u = 0, \qquad\qquad \tau_{xy} = 0.$$

On the boundary

$$0 \leq \xi \leq L, \qquad\qquad \eta = 0$$
$$u = v = 0$$

On the boundary $0 \leq \xi \leq L$ and $\eta = R$, if $x(\xi, \eta, t) \leq L_1$,

$$v = at \quad, \quad \tau_{xy} = 0$$

where a is the constant acceleration. If $x(\xi, \eta, t) > L_1$,

$$\sigma_{xx} n_1 + \tau_{xy} n_2 = 0$$

and

$$\tau_{xy} n_1 + \sigma_{yy} n_2 = 0$$

where

$$\vec{n} = n_1 \vec{i} + n_2 \vec{j}$$

is the unit normal to the deformed surface, on the boundary $L_1 \leq \xi \leq L$, $\eta = \bar{R}$,

if $L_1 \leq x (\xi, \eta, t) \leq L$ and $y (\xi, \eta, t) \leq \bar{R} - 1/2$ at^2

$$\sigma_{xy} n_1 + \tau_{xy} n_2 = 0$$

$$\tau_{xy} n_1 + \sigma_{yy} n_2 = 0$$

IV. SOLUTION TECHNIQUE

The problem described in Section III includes material nonlinearity and geometric nonlinearity. It is easily seen that analytical solutions, for such a transient dynamic response problem with finite deformation and elastic--plastic material are very difficult to obtain. Therefore, suitable numerical methods have to be used to solve the problem. Finite difference techniques based on the Lax-Wendroff scheme [2] and the modified version of Strang's method [3,4] due to Morris and Gottlieb [5, 6] are used for the study of the transient dynamic response of elastic-plastic solids under conditions of finite deformations. Since the finite deformation formulation used in the present problem is based on a Cauchy stress formulation and updated Lagrangian approach, the finite difference meshes may distort with increasing time. Thus, the conventional finite difference schemes for spatial derivatives in which the meshes are fixed for all time are no longer suitable. A second order accurate numerical technique based on the Lax-Wendroff scheme, the modified Strang method and the contour MacCormack two-step procedure has been developed for use with deformable Lagrangian meshes at Georgia Institute of Technology [7,8]. This modified scheme is efficient and also is suitable for deformed meshes. Therefore, it has been used to solve the rotating band and rifling interaction problem described in Section III. The detailed formulation is given in [7,8].

The modified Strang's method along with the Lax-Wendroff scheme enables one to write the solution of Eq. (11) as follows:

$$\{u\}^{t + \Delta t} = L_y L_x L_x L_y \{u\}^t \tag{12}$$

where L_x, L_y are the one-dimensional Lax-Wendroff operators, which are defined as:

$$L_x \{u\}^t = \{u\}^t + \Delta t [A]\{u\}^t_{,x} + \frac{1}{2} \Delta t^2 \Big\{ [A]^2 \{u\}^t_{,xx} \tag{13}$$

$$+ [A][A]_{,u} \{u\}^t_{,x} + [A]_{,u} [A]\{u\}^t_{,x} \Big\} \{u\}^t_{,x} \tag{14}$$

$$L_y \{u\}^t = \{u\}^t + \Delta t [B]\{u\}^t_{,y} + \frac{1}{2} \Delta t^2 \Big\{ [B]^2 \{u\}^t_{,yy}$$

$$+ [B][B]_{,u} \{u\}^t_{,y} + [B]_{,u} [B]\{u\}^t_{,y} \Big\} \{u\}^t_{,y}.$$

In using the procedure described in equations (12) to (14), it is necessary to compute the products of matrices, derivatives of matrices, and finite difference approximations for second order partial derivatives at each step of numerical integration. These computations can be time consuming when the matrices become large. But they can be eliminated by using a MacCormack two-step procedure [9] which consists of the successive application of two first order accurate scheme to achieve a second order accuracy. The steps involved in calculating L_x are as follows:

$$\{u\}^* = \{u\}^t + \Delta t \, [A]\{u\}^t, x \tag{15}$$

$$L_x\{u\}^t = 1/2 \, (\{u\}^t + \{u\}^*) + \Delta t/2 \, [A]^* \, \{u\}^*, x \tag{16}$$

where $[A]^*$ is evaluated by using the value of $\{u\}^*$. A similar expression can be written for L_y in the y direction. For the purpose of stability and accuracy, it is necessary to calculate the spatial derivatives in (15) by a forward difference in the predicted step and that in (16) by a backward difference in the corrected step or vice versa. As discussed earlier, in a problem with finite deformations, the initially regular meshes distort with increasing time, thus the contour difference forms for calculating the spatial derivatives are necessary. Specifically, for the present two-step method, this requires the use of contour finite differences of backward and forward types with a second order accuracy. Such a numerical technique has been developed [7,8]. By using the MacCormack's two-step procedure (15), (16) and the modified Strang's method (12), a finite difference scheme for the solution of equation (11) can be written as follows:

$$\{V\}_{(1)} = \{u\}^t + \Delta t [B] \Delta y \{u\}^t$$

$$L_y\{u\}^t = \{V\}_{(2)} = \frac{1}{2} (\{u\}^t + \{V\}_{(1)}) + \frac{\Delta t}{2} [B]_{(1)} \nabla_y \{V\}_{(1)}$$

$$\{V\}_{(3)} = \{V\}_{(2)} + \Delta t [A]_{(2)} \, \Delta_x \{V\}_{(2)} \tag{17}$$

$$L_x L_y\{u\}^t = \{V\}_{(4)} = \frac{1}{2} (\{V\}_{(2)} + \{V\}_{(3)}) + \frac{\Delta t}{2} [A]_{(3)} \, \Delta_x (V)_{(3)}$$

$$\{V\}_{(5)} = \{V\}_{(4)} + \Delta t [A]_{(4)} \, \Delta_x \{V\}_{(4)}$$

$$L_x L_x L_y\{u\}^t = \{V\}_{(6)} = \frac{1}{2} (\{V\}_{(4)} + \{V\}_{(5)}) + \frac{\Delta t}{2} [A]_{(5)} \nabla_x (V)_{(5)}$$

$$\{V\}_{(7)} = \{V\}_{(6)} + \Delta t [B]_{(6)} \, \Delta_y \{V\}_{(6)}$$

$$L_y L_x L_x L_y\{u\}^t = \{V\}_{(8)} = \frac{1}{2} (\{V\}_{(6)} + \{V\}_{(7)}) + \frac{\Delta t}{2} [B]_{(1)} \, \Delta_y (V)_{(7)}$$

$$\{u\}^{t+2\Delta t} = \{V\}_{(8)}$$

The operators Δ_x, Δ_y are forward contour differences. Similarly, ∇_x and ∇_y are backward contour differences. The form of $\Delta_x \{u\}_{i,j} \equiv \nabla_x \{u(\xi_i, \eta_j)\}$ is written in the following way.

$$\Delta_x\{u\}_{i,j} = \Big[(\{u\}_{i+1,j} - \{u\}_{i,j})(y_{i,j+1} - y_{i,j-1}) - (\{u\}_{i,j+1} - \{\ddot{u}\}_{i,j-1}).$$

$$(y_{i+1,j} - y_{i,j})\Big] \Big/ \Big[(x_{i+1,j} - x_{i,j})(y_{i,j+1} - y_{i,j-1}) - (x_{i,j+1} - x_{i,j-1}).$$

$$(y_{i+1,j} - y_{i,j})\Big] \tag{18}$$

The set of finite difference equations (17) are valid for any interior point. A slightly different version is needed for points on the boundaries [10].

V.   NUMERICAL RESULTS AND DISCUSSIONS.

The numerical procedure described in Section IV has been employed to solve a specific two-dimensional plane strain problem simulating the radial and circumferential flow of the band material into rifling grooves as a result of its interaction with the rifling. The geometry of the undeformed rotating band and the physical properties of the band material are listed below:

(a) geometry (ref. to Fig. 3)

L = 0.2 in (0.508 cm)
L_1 = 0.075 in (2.032 cm)
R_1 = 0.1 in (0.254 cm)
$\bar{R}$ = 0.15 in (0.381 cm)

(b) physical properties (copper)

Density, ρ=8.941 grms/cc
Young's Modulus, E = 1.6×10^7 psi (110320 MPa)
Poisson's Ratio, ν = 0.32
Yield stress in simple tension, Y=4.5×10^4 psi
(310275MPa)

The initial configuration of grid points employed in the finite difference discretization is shown in Fig. 4. It contains the region bounded by the edges at x=0, L and y=0,R. There are 200 cells in the region and the dimensions of these cells are x=y=0.01 in (.0254 cm).

In the numerical analysis, two distinct cases have been considered. In the first case, the gun barrel is assumed to move into the rotating band at a constant velocity of 0.01 in/μ sec while in the second case at a constant acceleration of 5.22×10^5 in/μ sec. In the numerical computations the CFL (Courant-Friedrichs-Levy) number is chosen to be 0.98 for the finite difference scheme.

The results of the first case are depicted in Figures 5 and 6 with plots of the deformed grid configuration, the velocity field, and the stress field at times 2.08 μ sec and 2.79 μ sec. The results indicate a strong circumferential flow of the band material into the rifling groove. Both the flow and the stress fields indicate that the flow initiates from the top surface of the band. Similar results for the second case are shown in Figs. 7 and 8.

The velocity and stress fields in the deformed configurations obtained in the two cases exhibit the behavior or trend expected in the interaction between rotating band and the rifling in the gun barrel. The preliminary results in the study demonstrate that the numerical scheme described in this paper can be used to analyze the rotating band and rifling interaction problem involving plastic flow and large deformations. Future efforts will be devoted to studying the influence of the parameters such as the flow strength of the rotating band material, the stiffness of the gun barrel and the projectile, the dimensions of the rotating band, the frictions at the contact surface of the rotating band and the gun barrel on the stress and deformations of the band.

Since the numerical procedure used in the present analysis is based on an explicit finite difference scheme, the selection of the spatial and temporal increment sizes must satisfy the stability condition $\max| (A)| < |$. This leads to a restriction on the size of the time steps. The computational efficiency of the present procedure can be improved by considering the use of a hybrid "explicit-implicit method". Such a combination has the benefits of both methods and will be investigated in future studies.

VI. REFERENCES.

1. Wagner, M.H. and Kreyenhagen, K.N. "Stress Analysis of Plastic Rotating Bands," Technical Report AFML-TR-76-12, Air Force Materials Laboratory, Ohio, February 1976.

2. Lax, P. and Wendroff, B. "Difference Schemes for Hyperbolic Equations with High Order of Accuracy," Comm. Pure Apply. Math., 17, pp. 381-390, 1964.

3. Strang, G. "Accurate Partial Difference Methods II, Nonlinear Problems," Number. Math., 6, pp. 37-46, 1964.

4. Strang, G. "On the Construction and Comparison of Difference Schemes," Siam. J. Number. Anal., 5, pp. 506-517, 1968.

5. Morris, J.L., "Splitting Methods for Parabolic and Hyperbolic Partial Differential Equations," Paper Presented at Conference on Numerical Solution of Partial Differential Equations, Oberwolfach, 1972.

6. Gottlieb, D., "Strang-Type Difference Schemes for Multidimensional Problems," SIAM J. Numer. Anal., 9, pp. 650-661, 1972.

7. Chen, H.P., "A New Second Order Accurate Finite Difference Method for Dynamic Response of Elastic-Plastic Finite Deformation Problems," Ph.D. Thesis, Georgia Institute of Technology, December 1983.

8. Chen, H.P. and Hanagud, S., "A Two-Step Procedure for Numerical Solution of Hyperbolic Differential Equations Involving Deformable Lagrangian Meshes," paper presented at the U.S. Conference on Numerical Methods in Computing, R.P.I., Troy, New York, 1984.

9. MacCormack, R.W., "The Effect of Viscosity in Hypervelocity Impact Cratering", AIAA Paper No. 69-354, 1969.

hanagud.115/rm
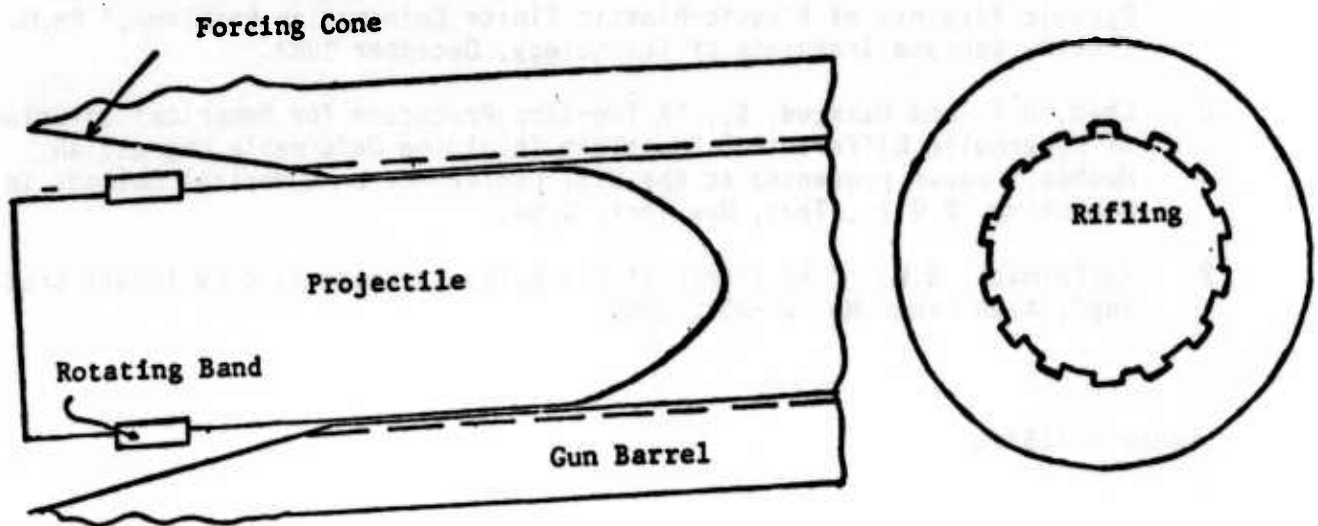
Figure 1. Schematic of an Artillery Projectile in the Gun Barrel
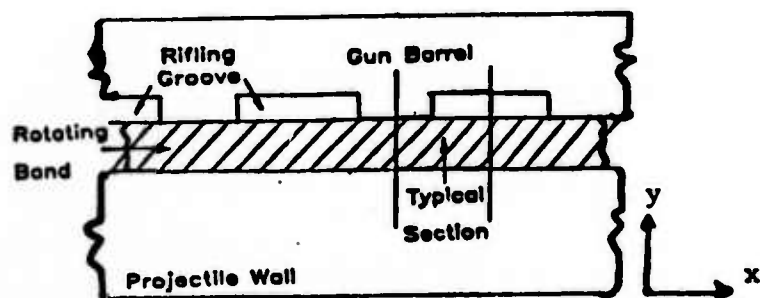
Figure 2. A Typical Section
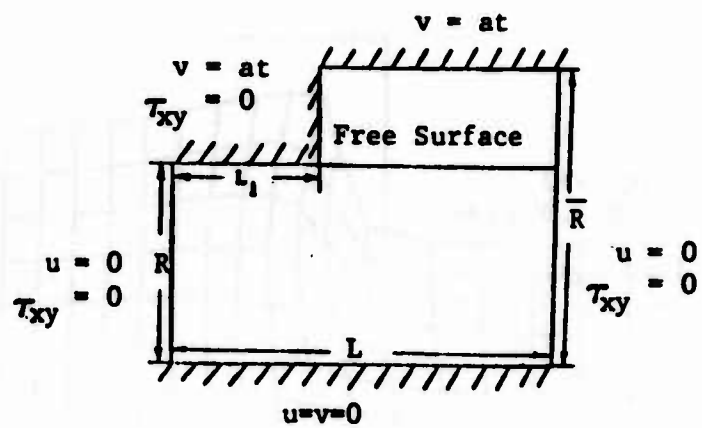
Figure 3. Boundary Conditions on Typical Section
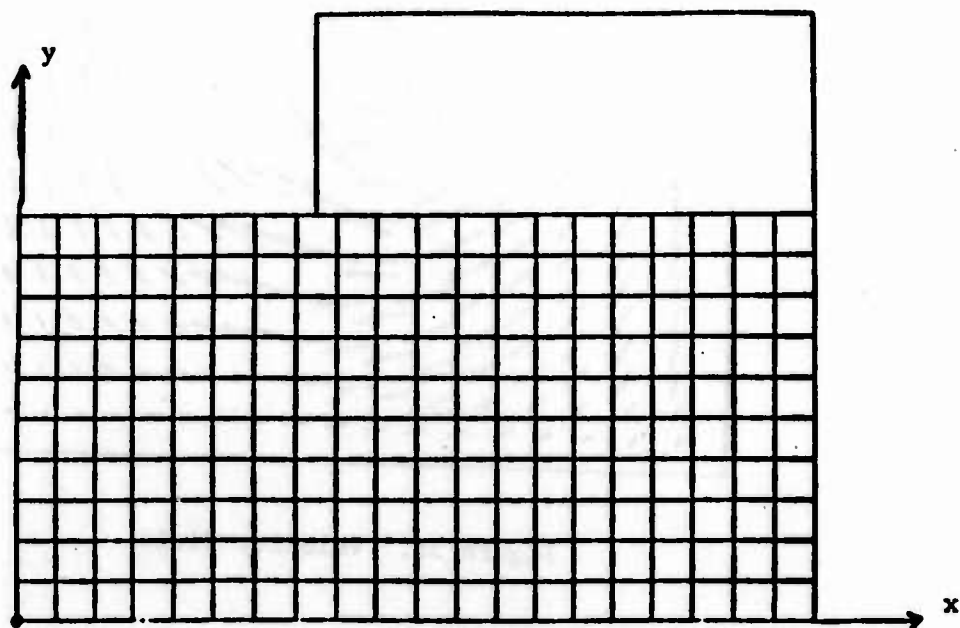
Figure 4. Initial Configuration of Grid Joints
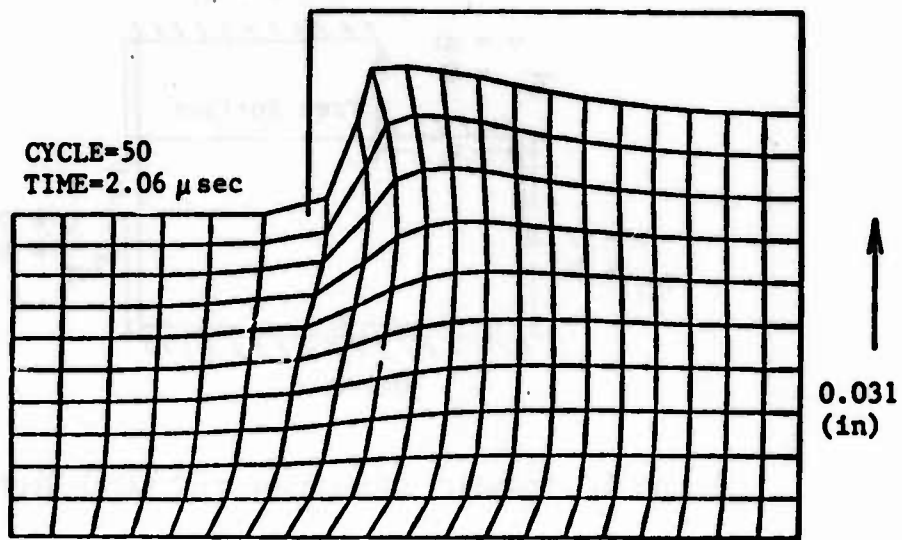
Figure 5a.   Deformed Configuration
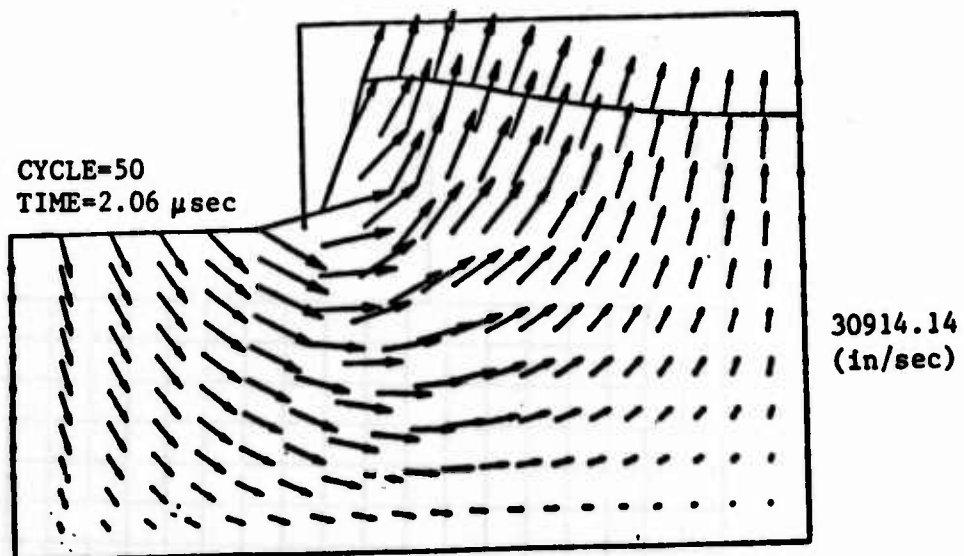
Figure 5b.   Velocity Field

CYCLE=50
TIME=2.06 μ sec

σ_{yy} τ_{xy}

σ_{xx}

2151519.70
(psi)

Figure 5c.   Stress Field

CYCLE=100
TIME=2.79 μ sec

0.031
(in)
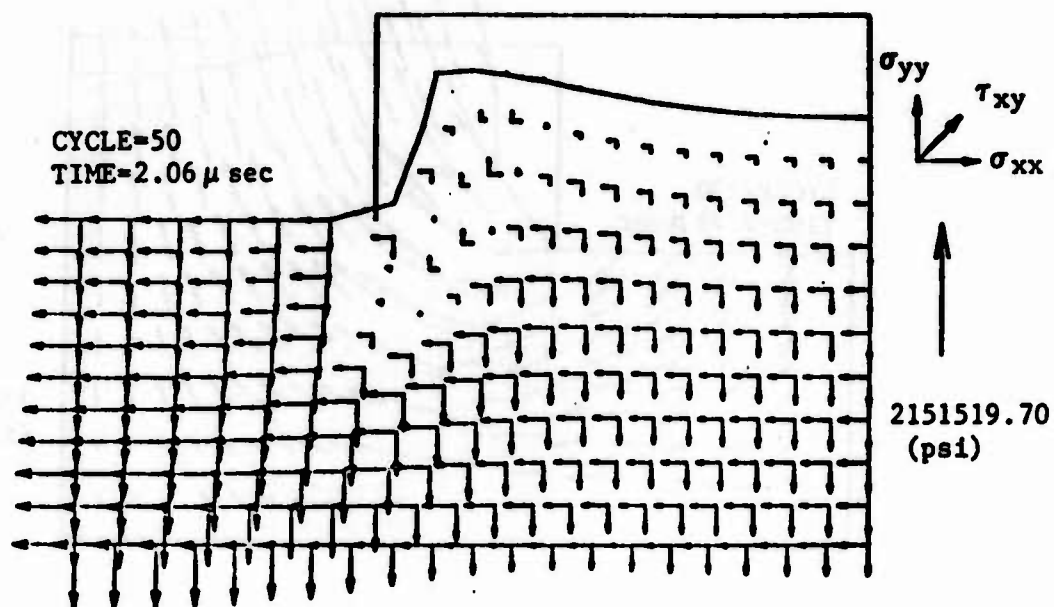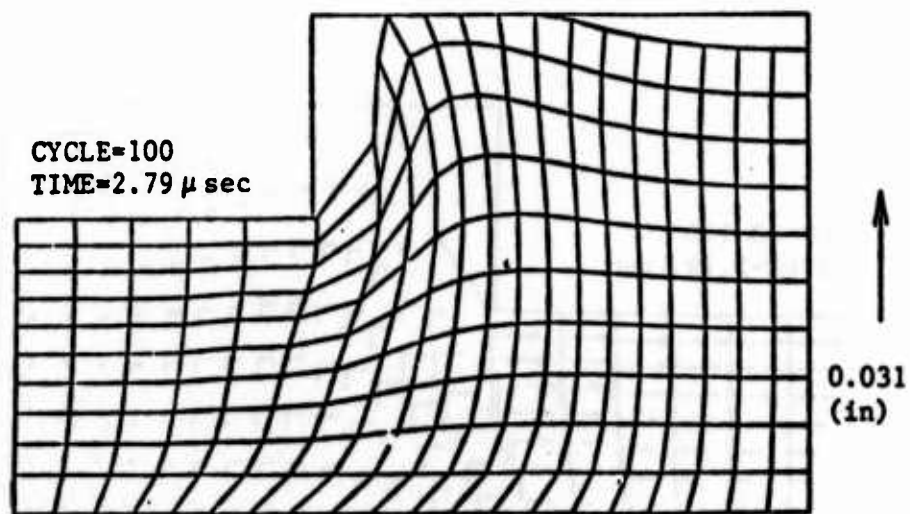
Figure 6a.   Deformed Configuration

CYCLE=100
TIME=2.79 μ sec

33298.95
(in/sec)
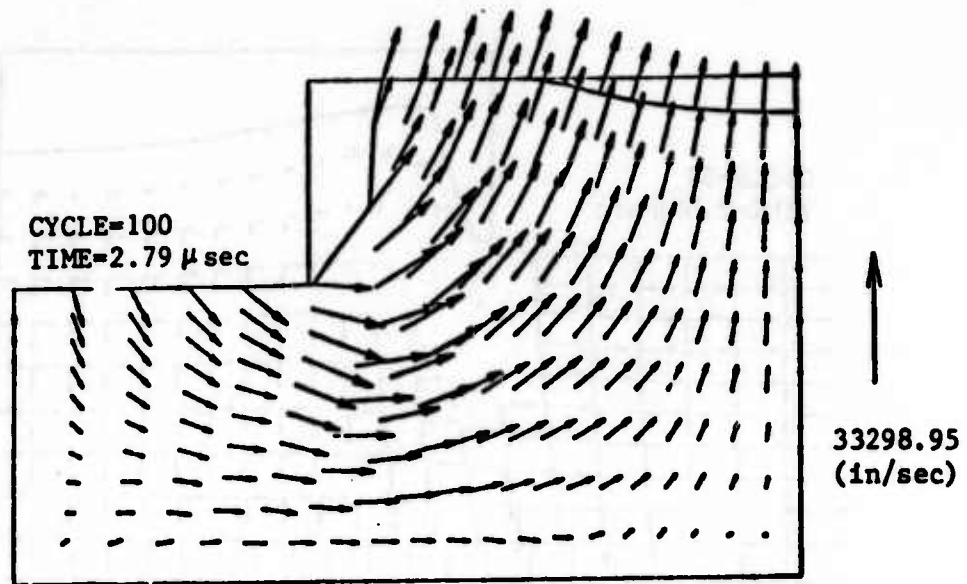
Figure 6b.   Velocity Field

CYCLE=100
TIME=2.79 μ sec

σ_{yy}

τ_{xy}

σ_{xx}

2076390.56
(psi)

Figure 6c.   Stress Field

CYCLE=200
TIME=11.45 μ sec

0.031
(in)
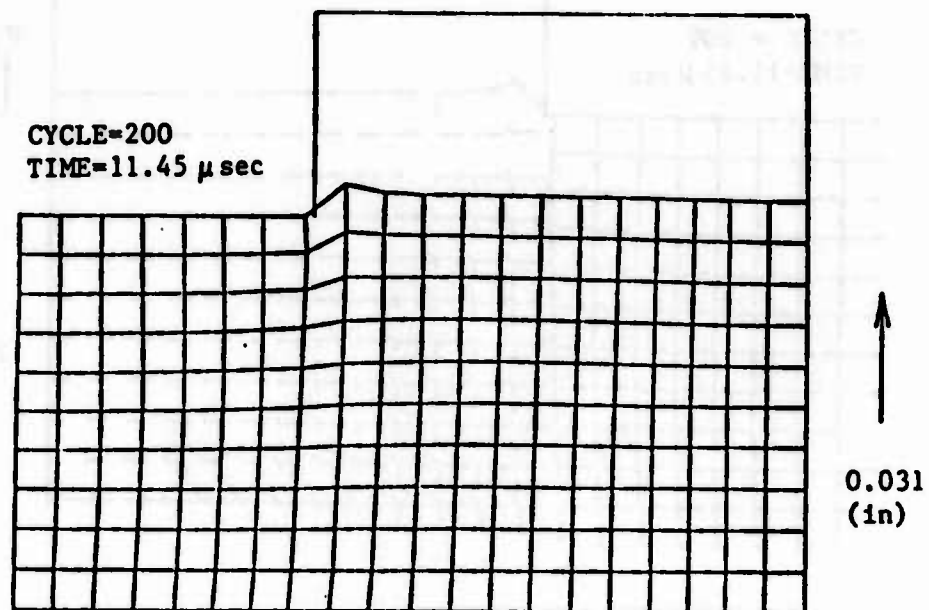
Figure 7a.   Deformed Configuration

CYCLE=200
TIME=11.45 μ sec

1597.44
(in/sec)
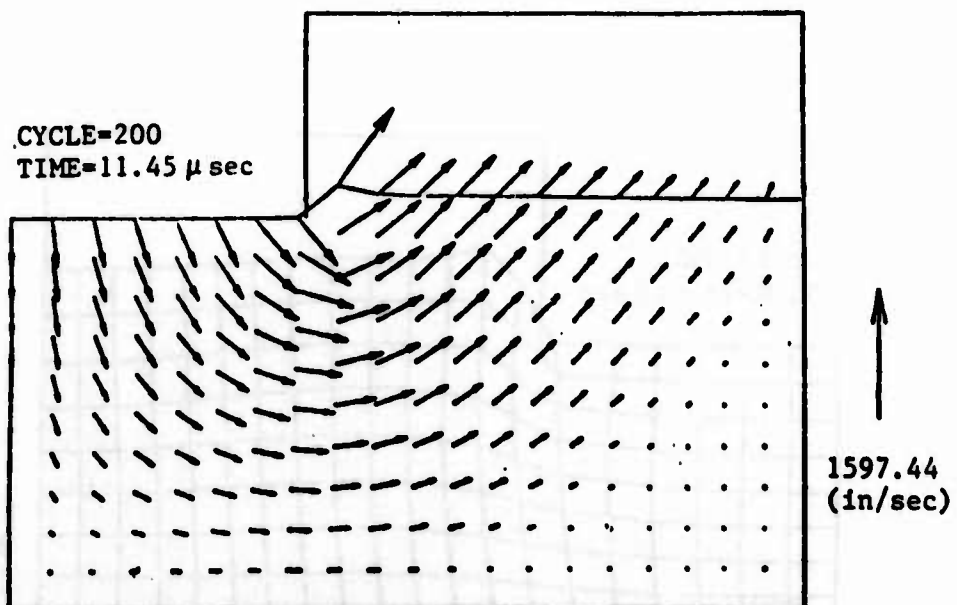
Figure 7b.   Velocity Field

1329
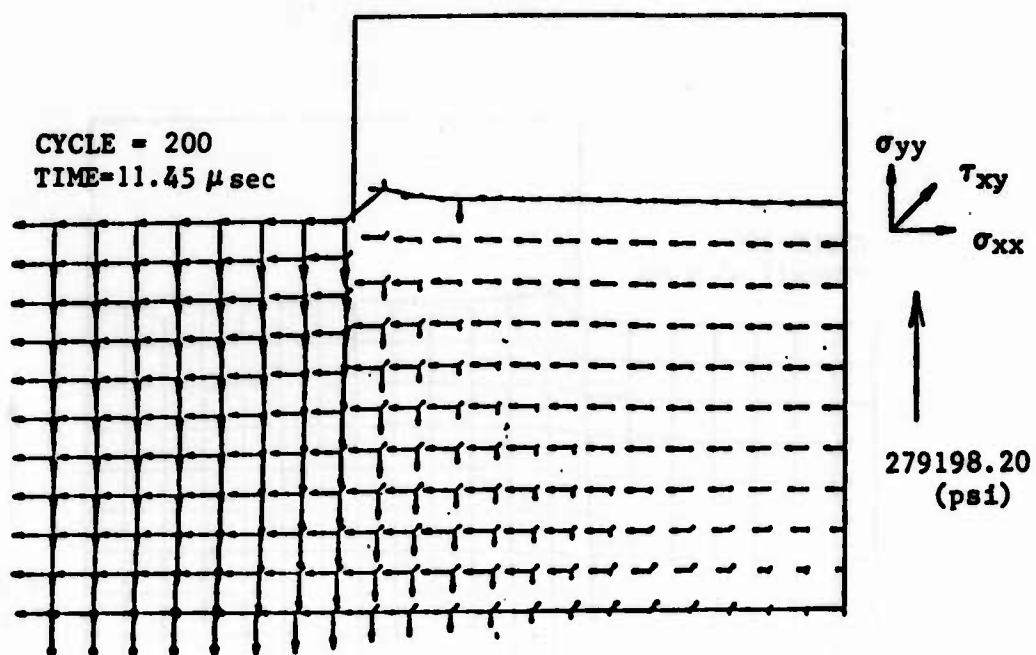
Figure 7c. Stress Field
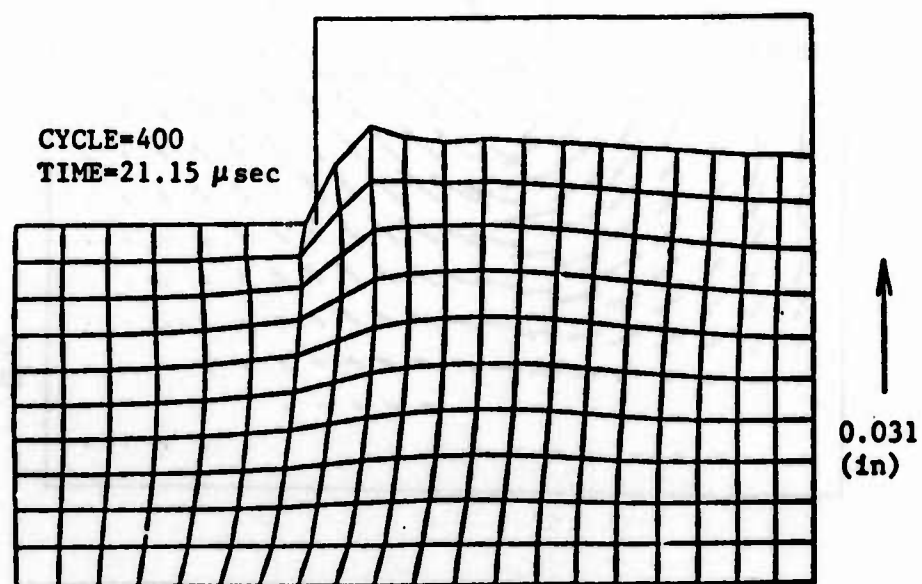
Figure 8a. Deformed Configuration

1330

CYCLE=400
TIME=21.15 μ sec

2695.54
(in/σ..c)

Figure 8b. Velocity Field

CYCLE=400
TIME=21.15 μ sec

σyy

τxy

σxx

460899.72
(psi)
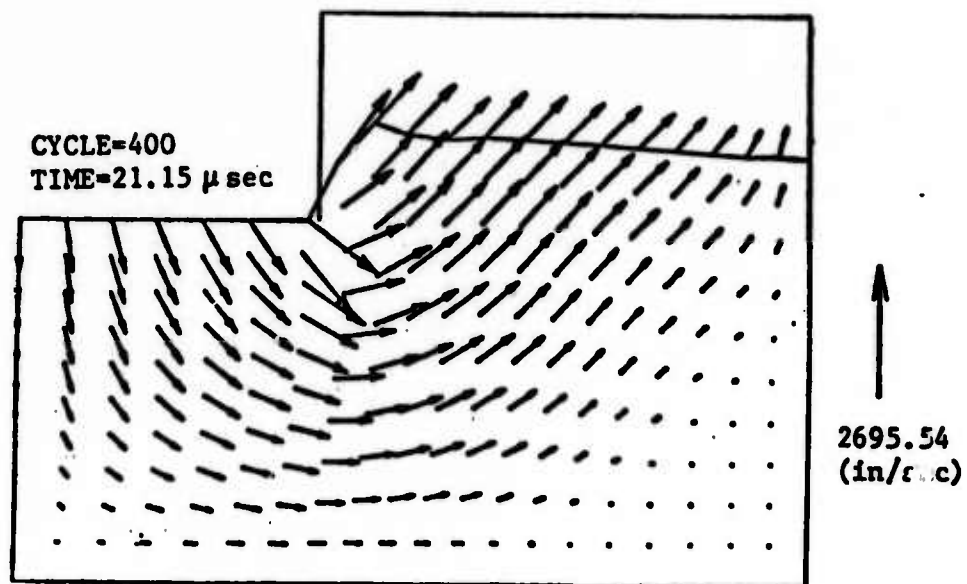
Figure 8c. Stress Field

USING SUPERCOMPUTERS TODAY AND TOMORROW*

John R. Rice

Computer Science Department

Purdue University

West Lafayette, Indiana 47907

ABSTRACT. The past and future growth of supercomputer power is summarized along with the changes in modes of accessing and using supercomputers. Three particular applications are considered from 1983, 1985 and 1995 (hypothetical). It is argued that the software and peripheral support for supercomputers has fallen far behind the increase in power. Solutions to the access and software support are discussed; it is concluded that the access problem is both difficult and very expensive to solve while the software support problem is difficult and only moderately expensive to solve.

I. SUPERCOMPUTER POWER.

First, we briefly review the growth in supercomputer power. The peak performance grew slightly less than exponentially over the 1965 to 1980 period. This growth masks the fact that the typical scientist and engineer outside a few weapons laboratory experienced very little growth in the power of the computers available to them. The prices came down but the power did not grow dramatically. Supercomputer power growth has accelerated since 1980 and even more acceleration is forecast for the next 10 years. This growth is summarized in Table 1.

Table 1. Some trends in scientific supercomputing

Year/Machine	Speed	Speed Increase	
		10-year	20-year
1966/CDC6600	1 MFLOPS	–	–
1975/CDC7600	4 MFLOPS	4	–
1980/Cray 1	10 MFLOPS	5	–
1985/Cyber 205	100 MFLOPS	25	100
1990/–	2 GFLOPS	200	1000
1995/–	200 GFLOPS	2000	50,000

The projected 1995 machine has 1000 processors with a 2 nanosecond cycle time.

During the period 1965-1985 there has been a significant change in access to computers. Batch processing has been replaced by some sort of terminal access. The types

---

*The author of this paper presented it at the Third Army Conference on Applied Mathematics and Computing.

of access vary considerably and are listed below in what we think is decreasing frequency:

1. Terminal attached to front end machine connected to network connected to supercomputer

2. Terminal attached to front end machine connected to supercomputer

3. Workstation attached to network attached to front end machine attached to network attached to supercomputer

4. Terminal attached to supercomputer

There are other types of access, but the key point here is that most access is via one or more layers of intermediate machines and networks.

One of the barriers in effective use of supercomputers is the disparate speeds within the intermediate machines and networks. Table 2 presents data on the transfer rates of these facilities.

Table 2. Peak and effective transfer rates of various facilities
measured in bits per second (K = 1000, M = 1,000,000)

Facility	Peak Rate	Effective Rate
Telephone	300	300
2400 baud line	2400	2400
9600 baud line	9600	9600
ARPANET	57K	20K
Bus on VAX 11/780	1M	160K
Ethernet	10M	1.5M
CDC LCN	50M	3M
Cyber 205 channel	200M	100M

It is easy to see that current supercomputers produce results at rates that completely swamp more the capacity of most user's access facilities.

Figure 1 shows the current configuration of the Cyber 205 facility at Purdue University. Most users access the Cyber 205 through the CDC6000 systems or by long haul networks attached to a VAX 11/780.

While progress in access to supercomputers has been significant (yet modest compared to the progress in supercomputer speed), the progress in programming has been uneven and even in reverse for some areas. Editors, on-line file systems, program update systems, libraries, etc. have greatly improved the environment for writing programs. Programming itself has gone downhill. In 1966 Fortran IV was well established. In 1986 supercomputer users can use Fortran 77 (a small improvement), but, if you want to get real supercomputer speeds, you must use machine specific Fortan statements, tricks and generally be rather knowledgeable about the whole machine organization. I believe that the programming task itself is perhaps 40% as efficient on the current Cray and CDC
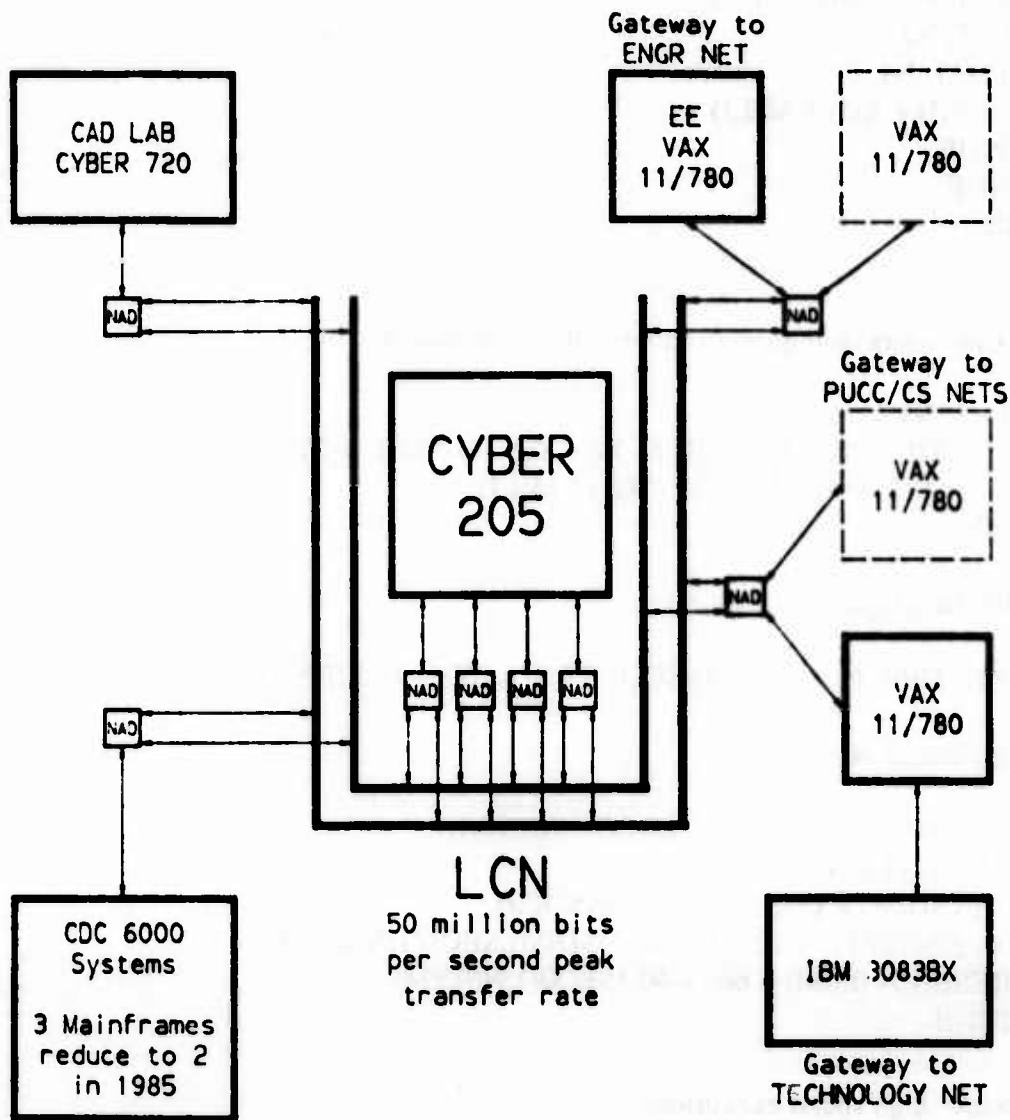
Figure 1. Configuration of the Purdue high speed file transfer network to support the Cyber 205 supercomputer.

supercomputers as it was in 1965 (again, this is for achieving something close to the potential of the machines). Automatic vectorizers are very worthwhile, but they also fall far short of providing the vectorization possible.

I illustrate this development in programming with two randomly selected examples. In [DoEi84] we see the subprogram

1335

```
      SUBROUTINE SMXPY (N1, Y, N2, LDM, X, M)
      REAL Y(*), X(*), M(LDM, *)
      DO 20 J = 1, N2
        DO 10 I = 1, N1
          Y(I) = Y(I) + X(J) * M(I,J)
10    CONTINUE
20    CONTINUE
      RETURN
      END
```

and learn that Cray users should learn to write the inner loop as

```
      DO 10 I = 1, N1
        Y(I) = (((((Y(I) + X(J -3) * M(I, J - 3)) + X(J - 2) * M(I, J -2))
     $            + X(J -1) *M(I, J -1)) + X(J) * M(I,J)
10    CONTINUE.
```

On a Cyber 205 the simple computation

FORALL (I=1:NTDIM, J=1:NSDIM) SCORES(I,J) = 60. + 40.*SIN(I*J*63.21)

should be programmed as

```
      SEQ(1;NSDIM) = Q8VINTL(1,1;SEQ(1;NSDIM))
      DO 100 I = 1,NTDIM
        ARG(1;NSDIM) = I*63.21*SEQ(1;NSDIM)
        SEQ0(1;NSDIM) = VSIN(ARG(1;NSDIM);SEQ0(1;NSDIM))
        SCORES(I,1;NSDIM) = 60. + 40.*SEQ0(1;NSDIM)
100   CONTINUE
```

in order to achieve high speed execution.

We conclude that the software and peripheral support for supercomputers has fallen far behind the increase in supercomputer computational power.

II. THE SIZE OF SUPERCOMPUTER ANSWERS. Everyone in the supercomputer area and most outside it visualize that supercomputer applications use enormous amounts of computation, millions and billions and trillions of arithmetic steps. Much less widely known is that the answers produced in a typical supercomputer applications are also huge. By answer, we mean the information the user needs in order to understand the computed solution; we do not mean the total set of numerical results computed, which is usually very much larger. We illustrate this with three sample applications, two real ones from 1983 and 1985 and one hypothetical one from 1995.

1983 APPLICATION: *The high speed impact of two steel cubes into a block of aluminum.* This computation was performed at Los Alamos [Los83] on a Cray 1 and used 30 mintues to cover 2.5 microseconds of real time. The problem is eight-dimensional with 3 space variables, time and 4 dependent variable (temperature,

pressure, density of steel and density of aluminum).

Thirty minutes of Cray time represents about 150 billion instructions (12 nanosecond cycle time) including about 18 billion arithmetic operations (10 MFLOPS). The answer can be represented by data on a 100 by 80 by 80 special grid for 150 time steps; each of the 96 million grid points has 4 values (64 bits long). Thus only 400 million numbers represent the result of 18 billion computed numbers, or the answer requires only 4.5% of the numbers computed. Some nice color plots are given in [Los83] and illustrate the effectiveness of this medium for presenting information about computed results. Note that the answer is 3 gigabytes in size which is close to the size of the entire disk memory space on many large scale computer systems.

1985 APPLICATION: *Accretion of material into a black hole (2D model).* This computation [Sm85] shows the evolution of a black hole over a period of millions of years. It assumes axial symmetry to reduce the problem to a feasible size. The answer is 1.25 billion numbers (10 gigabytes). The author discusses how to view the results using color movies. He notes that his computation only provides moderate resolution in time and space and that a good quality movie would require considerably more computation and produce a considerably larger answer.

In this same issue of Science magazine there is a discussion by Joy and Gage [JoGa85] which analyzes the information flow required to produce color movies. Modest resolution, slow motion requires 250 Kbytes/sec while high resolution, normal motion requires about 20 Mbytes/sec. The author argues that color movies are the only way to really assimilate the results of many supercomputer computations.

I estimate that a 3D black hole model giving comparable accuracy would have about 1.5 terabytes in the answer. This would produce a 100 hour movie with normal motion and modest resolution.

1995 APPLICATION: *Tank battle simulation.* In this hypothetical application we assume there are six tanks and the study focuses on the weapons system, the targeting system, the armor and the defensive systems. Thus, intensive, detailed computational analysis is made of a special event such as a shell hit, laser strike or mine explosion. In one of these special events, the physics is followed at the level of the shell explosion, shell case fragmentation and attempted penetration of the armor by blast pressure and heat. Other aspects such as mechanics of the tanks or terrain is simulated at a much coarser level.

A summary of the computation is as follows:

Independent variables:	3 space, time, input of 6 tank drivers, input of 6 tank gunners
Dependent variables:	tank positions
	state of all weapons systems
	state of all defensive systems
	effects of all special events

Computation:	1 hour
	2 mega-giga instructions (2 nanosecond cycle, 1000 processors)
	700 MFLOPS (200 teraFLOPS machine)
Answer:	A. General Scene: 200 by 200 by 50 grid
	B. One tank geometry: 100 by 100 by 25
	C. One tank weapon system: 10,000 variables
	D. One tank defensive system: 10,000 variables
	E. Tank mechanics: 2000 variables
	F. High level battle scene: 5000 variables, 5000 time steps
	G. Special Events: 200 events with 200 x 100 x 100 x 200 grid

The size of the answer is then (in megawords)

$$A + 6B + 6C + 6D + 6E + F + G$$
$$= 2 + 6(.25 + .01 + .01 + .002) + 25 + 200 * 400$$
$$\sim 1,000,000$$

Thus the size of this answer is about 8 terabytes. This answer could be shown, in full, as a color movie with normal motion and high resolution that lasts about 100-120 hours. We visualize that the study of this application would involve several people viewing different parts of the answer over a period of time.

We now pose the question: *Suppose the answer has been computed and resides in the supercomputer system, how long will it take to move the answer to the user's location?* We use the effective transfer rates from Table 2 plus the size of answers to produce the results of Table 3. It is obvious from Table 3 that systems which separate the user from the supercomputer by 2 ethernets and a VAX are totally unable to provide reasonable service for many supercomputer applications. Few will put up with waiting a week to see the results of a 30 minute computation. And once he gets the answer "locally" the user neither has a place to put it nor adequate means to view it.

III. SUPERENVIRONMENTS. We draw three conclusions from the above material:

1. Today's peripherals/workstations/networks are grossly inadequate even for todays supercomputations.

2. Today's programming environments/languages for supercomputers are grossly inadequate, even antiquated.

3. Raw computing power will increase dramatically in the next decade.

The peripheral/workstation/network problem is not easily solved. Fiber optics networks provide a great deal of capacity for networks, but are not yet very cheap. Moderately priced terabit memory systems are not now on the horizon. Workstations with high quality color graphics and movie capabilities are quite expensive and probably

Table 3. Times to transfer the answers of the three applications using
various facilities.

	Application		
Facility	*1983*	*1985*	*1995*
Telephone	3 years	9 years	6 millennia
9600 baud line	1 month	3 months	2 centuries
ARPANET	2 weeks	7 weeks	1 century
VAX 11/780 bus	2 days	6 days	7 years
Ethernet	5 hours	15 hours	16 months
Cyber 205 channel	4 min	13 min	1 week
Run time	30 min	1 hour	1 hour

will remain so for some time. For the next decade it may well be that large organizations will have a few superworkstations which are shared by a large community of users.

The programming environment/language problem has many technical difficulties to be overcome, but the initial problem is simply lack of effort. The software support for supercomputers is very meager. We have senior scientists and engineers using facilities that would be instantly rejected by travel agents, junior high math students, secretaries and the general public. It is incredible to see one of the nation's scarest human resources wasted due to the lack of a modest investment (compared to the other aspects of super-computing) in software support.

One key software area is very high level languages appropriate for scientific compu-tations. Figures 2-4 show three examples of the kind of things we should expect. We do not discuss the ELLPACK [RiBo85], DEQSOL [Ume83] or PROTRAN [AiRi83] sys-tems in any detail but do note they have the following characteristics:

1. They dramatically improve programming productivity.

2. They were implemented with moderate efforts (2-4 man years).

3. They improve execution time efficiency.

Each of these languages has short-commings that one would not expect in production quality systems for supercomputers, yet they represent a great advance over the software currently supplied with supercomputers.

The other key software area is how to map computations onto complex supercom-puter architectures so as to produce high efficiency execution. This is a challenging technical problem where many approaches are being actively pursued. However, there is still little indication that the existing and future techniques will be embodied into good user-oriented tools or systems.

We close with Figure 5 which shows the schematic of a superworkstation which is appropriate for supercomputers. It will cost 10-50 times as much as the current good workstation. It needs supersoftware that will also cost 4-50 times as much as current software for supercomputers. The superworkstation and supersoftware are equally important, but the total investment for the software will be an order of magnitude less. Then one might have the superenvironment to take full advantage of the supercomputer power that will appear in the next decade.

V. REFERENCES.

[AiRi83] T.J. Aird and J.R. Rice, PROTRAN: Problem solving software. Adv. Engin. Software, 5 (1983) 202-206.

[JoGa85] W. Joy and J. Gage, Workstations in science. Science, **228** (1985) 467-470.

[Los83] Computers and Computing. Los Alamos Science, Winter/Spring (1983), 140-141.

[RiBo85] J.R. Rice and R.F. Boisvert, Solving Elliptic Problems using *ELLPACK*. Springer-Verlag, New York (1985).

[Sm85] L.L. Smarr, An approach to complexity: Numerical computations. Science, **228** (1985) 403-408.

[Ume83] Y. Umetani, M. Tsuji, K. Iwasawa and H. Hirayama, DEQSOL: A numerical simulation language for vector/parallel processors. Hitachi report (1983).

```
*                    THE PLATEAU PROBLEM
EQUATION. (1.+UY(X,Y)**2) UXX  +  (1.+UX(X,Y)**2) UYY          &
                     - 2.*UX(X,Y)*UY(X,Y)   UXY                &
         + 2.*(UX(X,Y)*UYY(X,Y) - UY(X,Y)*UXY(X,Y)) UX         &
         + 2.*(UY(X,Y)*UXX(X,Y) - UX(X,Y)*UXY(X,Y)) UY         &
    =      2.*(UX(X,Y)*UYY(X,Y) - UY(X,Y)*UXY(X,Y))*UX(X,Y)    &
         + 2.*(UY(X,Y)*UXX(X,Y) - UX(X,Y)*UXY(X,Y))*UY(X,Y)
*
BOUNDARY. U = BOUND(X,Y) ON Y = 0.0  $ U = BOUND(X,Y) ON Y = 1.0
          U = BOUND(X,Y) ON X = 1.0  $ U = BOUND(X,Y) ON X = 0.0
*
GRID.           5 X POINTS  $  5 Y POINTS
TRIPLE.         SET ( U = ZERO )
fORTRAN.
                DO 100 IT = 1,5
DISCRETIZATION.     HERMITE COLLOCATION
SOLUTION.          BAND GE
FORTRAN.
  100           CONTINUE
OUTPUT.         PLOT(U)
SUBPROGRAMS.
      FUNCTION BOUND(X,Y)
      BOUND = SIN(X+AMAX1(.66,.1+Y**2))*EXP(X-Y)
      RETURN
      END
END.
```

Figure 2. An ELLPACK program that solves the Platean problem (the soap film problem) using Newton iteration combined with Hermite-cubic collocation for the linearized problem.

```
        PARAMETER ( N = 16 )
        REAL MATRIX HILBERT(N,N), X(N,4), B(N,4), RESID(N,4)
        REAL VECTOR RNORM(4)
C               CREATE HILBERT MATRIX
        ASSIGN HILBERT(I,J) = 1/(I+J-1.)
C               DEFINE HILBERT MATRIX, FIRST 3 RIGHT SIDES
        ASSIGN   B(I,1) = 0.0
                 B(I,2) = 1.0
                 B(I,3) = 1.0 + .01*SIN(100.*I)
        B(1,1) = 1.0
C               COMPUTE 4TH SIDE TO MAKE SOLUTION =1.
        DO 20 I = 1,N
   20       SUM HILBERT(I,J); FOR (J=1,N); IS B(I,4)
C               SOLVE THE 4 SYSTEMS AND COMPUTE NORM OF RESIDUALS
        LINSYS HILBERT*X = B ; HIGHACCURACY ; SAVE HILBERT
        PRINT B,X
        ASSIGN   RESID = HILBERT*X - B
                 RNORM   = RESID'*RESID
                 RNORM(K) = SQRT(RNORM(K))
        PRINT RNORM
        END
```

Figure 3. A PROTRAN program that creates a Hilbert matrix and four right sides, solves these linear systems and prints the least squares residual for each system.

```
VAR      TT;
DOM      x = [0:1],
         y = [0:1.2]  ;
TDOM     t = [0:1]
MESH     x = [0.0:1.0:0.2]  ,
         y = [0.0:1.2:0.2]  ,
         t = [0.0:1.0:0.02]  ;
CONST    A = 0.62  ;
REGION   R = [*,*]  ,
         L = [0,*0  ,
         R1 = [1,*]  ,
         D = [*,0]  ,
         U = [*,1.2]  ;
EQU          dt[TT] = A*[lapl [TT] ]  ;
INIT         TT = 100   AT R  ;
BOUND        TT = 200   AT D,
             TT = 200   AT U,
             dx[TT] = 0   AT L+R1;
SCHEME  ;
   ITER        NT UNTIL   NT GT 4  ;
             TT<+1>=TT+DLT*A*lapl [TT]
   PRINT       TT    AT R ;
   DISP        TT    AT R ;
   END ITER  ;
END SCHEME  ;
END ;
```

Figure 4. A DEQSOL program that solves a time dependent partial differential equation.
The solution obtained this way ran three times as fast as the same method
implemented in ordinary Fortran and then hand optimized for vector speeds.

1342
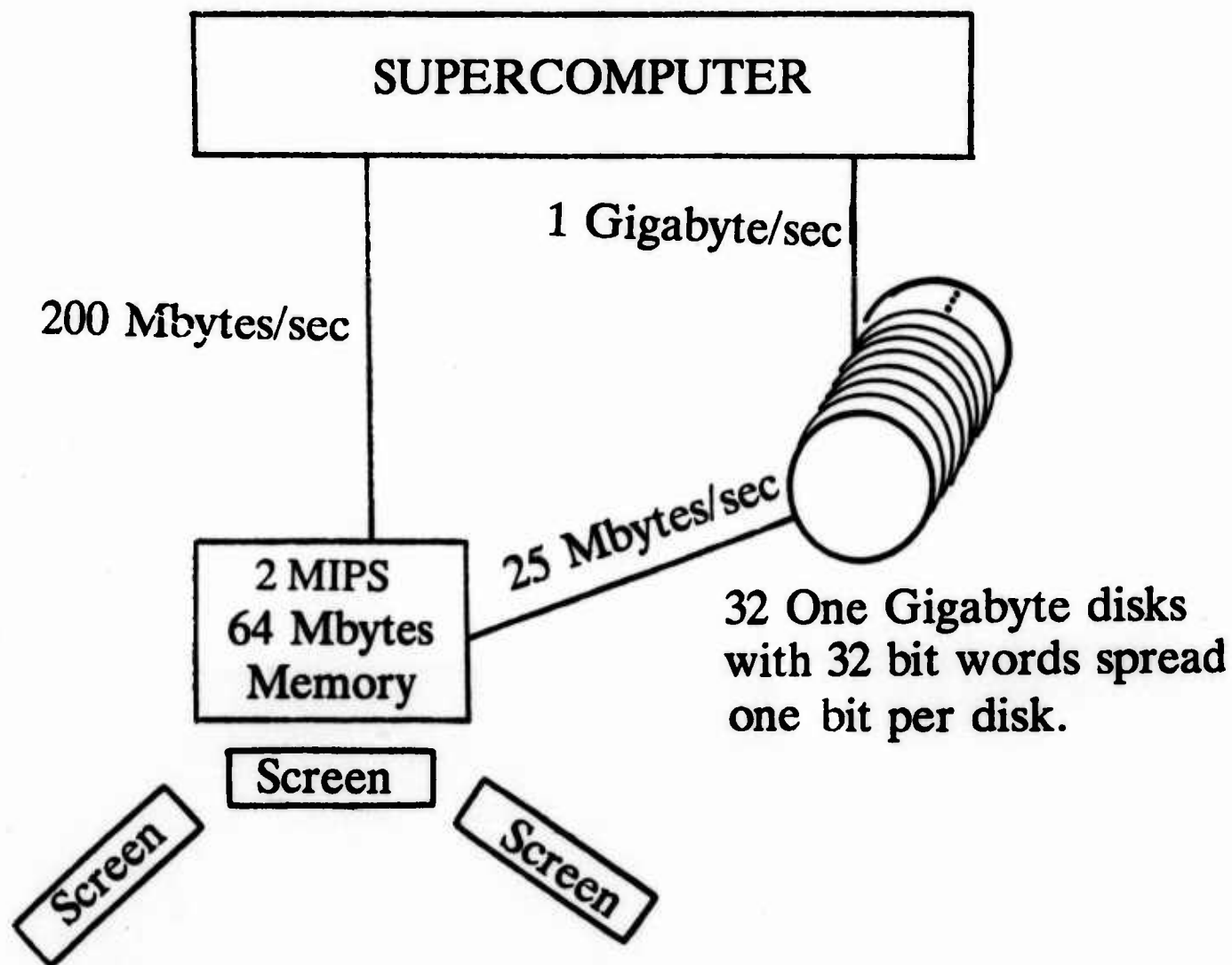
Figure 5. Schematic diagram of a superworkstation appropriate for the supercomputers of the 1990's.

4th ANNUAL ARMY CONFERENCE

List of Registrants

Adjerid, Soimane	Grigoriadis, Michael	Rosenblath-Roth, Millu
Andersen, Gerald	Grossberg, Steven	Roytburd, Victor
Ball, Michael	Gurtin, Morton	Sacks, Paul
Basu, Sankar	Hajek, Bruce	Sadhal, Satwindar
Bebernes, Jerry	Harrison, Ralph	Sahu, Jubaraj
Biermann	Heaton, Kenneth	Sambandham, M.
Bissel	Herbert, Thorwald	Sedney, Raymond
Brown, Frank	Howison	Seren
Buckmaster, John	Jackson, William	Sham, T.L.
Buklet, Bruce	Johnson, Arthur	Shearer, Michael
Burger, Marc	Kapila, A.K.	Shieh
Burman, Simeon	Kassoy, David	Shilepsky
Carofano, Garry	Khalsa	Siedlecki, W.
Carpenter, Gail	Kushner, Harold	Sileo
Chandra, Jagdish	Laine-Schmidt, Claudine	Skylansky, Jack
Chen, Peter	Lehnick, Siegfried	Soanes, Royce
Chen, Tsu-Fen	Levey	Srivastav, Ram P.
Chow, Pao-Liu	Lippman, Alan	Steen
Chui, Charles	Loveland, Donald	Steeves, Earl
Chu, Shih	Mahalabais, A.K.	Stewart, Bradley
Cohen, C.	Majda, Andrew	Stewart, Scott
Cohen, Herb	Marroquin, Jose	Stienberg, Stanly
Coyle, J.M.	Melander	Strikwerda, John
Cummings, Ben	Menyuk, Curtis	Stuempfle, Arthur
Delchamps, David	Miller, Miles	Takagi
Diamond, Harvey	Mitter	Tessler, Alexander
Doss	Morris, Phillip	Thomas
Drew, Donald	Moysey, Brio	Tracy, Fred
Egerland, Walter	Nohel, John	Wahlquist
Elmagrabhy, S.E.	Patel, Nisheeth	Weiss, Richard
Esogbue, Augustine	Payne	Whinston, Andrew
Ewing, Dick	Pezeshki, Charles	Whiteman, J.R.
Fishman, Louis	Phillipson, Paul	White, Chelsea
Flaherty, Joseph	Piscitelle, Louis	Wouk, Arthur
Fried, Isaac	Please, Colin Peter	Wright, Thomas
Gerasoulis, A.	Polk, John	Yip, Pak T.
Ghoniem, Ahmed	Prabu, N.U.	Zabausky
Gidas, Basilis	Pumir, Alain	Zemanian, Armen
Glimm, James	Pu, San Li	Zemanian, Thomas
Glynn, Peter	Rall, L.B.	Zhifen, Zhang
Golshani, Forouzan	Raphael, Louis	

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 87-1	2. GOVT ACCESSION NO. AD-A183 544	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Transactions on the Fourth Army Conference on Applied Mathematics and Computing		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on behalf of the Chief of Research, Development and Acquisition		12. REPORT DATE February 1987
		13. NUMBER OF PAGES 1345
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) U.S. Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709-2211		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

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

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is a technical report resulting from the Fourth Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat various Army applied mathematical problems.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

computer vision	data analysis
imaging processing	combustion
probabilistic methods	multiphase flow
fluid mechanics	numerical PDE
solid mechanics	optimization
shock waves	vortex flow
electromagnetics	reaction-diffusion
nonlinear analysis	approximation
control theory	computational complexity
stochastic analysis	bifurcation
transonic flow	viscoelastic material
expert systems	hyperbolic conservation laws
artificial intelligence	fractals
statistical analysis	finite elements

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